Does Modelling Annotator Distributions Help? Evidence and a Metric Caveat from Enthymeme Detection

Moez Ur Rehman[1], Haiyan Alam[2]

[1, 2] *National University of Computing and Emerging Sciences (FAST-NUCES)*

Abstract

The MediaEval 2026 Missing Pieces and Misinformation dataset is constructed from five independent annotators' classifications of tweets using a three-label scheme for an enthymeme detection task. The multi-annotator data is typically treated to compute a majority-vote label, while minority votes are discarded, but in this way, information about the task's realistic uncertainty is lost. This is particularly the case with the rare conclusion class, where the annotators' consensus is statistically low. This paper addresses the Quest for Insight question 2, i.e. whether an improved performance on borderline cases when compared to majority-vote labels is witnessed when modelling the full distribution of human judgments? By training two BERTweet-base classifiers under similar conditions, (Run 1) used weighted cross-entropy as a loss function on majority vote-labels and (Run 2) with a loss function that was an combination of a mixed cross-entropy and KL-divergence loss calculated against the full annotator distribution. On the blind test set, soft label supervision causes an improvement in three class label macro F1 scores from 0.4865 to 0.5087; this clearly is a 47% relative gain on the rare conclusion class, as F1 goes up to 0.27, rising from 0.18. On the binary classification task, these two are identical. We report a third observation complicating the scenario, where in both tasks, cross-entropy against one-hot gold predictions favors Run 1, while Run 2 produced more confident classification and was more faithful to annotator agreement. We recommend that label distribution learning methods should be assessed in comparison to annotator distributions and they should not rely on the majority vote argmax, based on our reasoning that this is an issue with the evaluation metric itself.

1 INTRODUCTION

Enthymeme detection in social media is a task where different people often disagree. The MediaEval 2026 dataset is noted to achieve a Fleiss κ of around 0.574 across three label categories. The rare category, which is the conclusion class, has a large share of these disagreements. This suggests that the disagreement isn't just because people are making mistakes or misunderstanding the guidelines. Instead, it shows how human labeling varies, as discussed in earlier research [1]. People's judgment on difficult cases leads to different label choices and makes it hard to reach a single correct answer, which is why annotators can't always agree.

The primary method in supervised machine learning views it as a problem that can be tackled by obtaining the label based on majority voting and training the model to reproduce it. This is computationally efficient and corresponds with the usual descriptions of evaluation metrics like accuracy, F1 and cross-entropy. It is informationally problematic due to the associated loss. A tweet identified as premise by three of five annotators and considered none by the other two is not the same data point as one where all five annotators agreed, hence the majority vote framing renders them nearly identical.

A different approach employs the experimental distribution of the annotator votes as a soft training objective [2,3]. The classifier receives a divergence penalty if it fails to align with this distribution. The divergence penalty is founded on the idea that calibrated uncertainty provides more accurate information and may be more beneficial than simply replicating a confident majority vote. The MediaEval 2026 organizers have clearly posed this question in their Quest for Insight prompts: Is there an observed enhancement in performance on borderline cases when compared to majority-vote labels by modelling the complete distribution of human judgments?

This document offers an empirical examination and solution to this. We trained two BERTweet-base classifiers with the same hyperparameters, splits and a single seed; they differed only in the formulation of the loss. They are assessed on the blind test set using three criteria: macro-F1 (both overall and per-class), accuracy and cross-entropy compared to one-hot gold labels. We present three findings: a significant improvement on the rare conclusion class in the three-class label task, while there was no impact on the more challenging binary task and the soft label model is hindered by a contradictory cross-entropy loss, thereby establishing a connection between its training goal and the evaluation metric.

The contribution is twofold. Empirically, the situations in which soft label training leads to noticeable improvements are determined. From a methodological standpoint, we argue that the observation of cross entropy should not be viewed as a failure stemming from the calibration of the soft label model rather, it serves as a characteristic that generates distributional predictions in relation to hard targets.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
khawajamoez07@outlook.com (M. Ur Rehman); haiyan.alamshykh@gmail.com (H. Alam)

2 RELATED WORK

Learning from annotators' disagreement spans label distribution learning [2], soft label distillation [3] and broader frameworks with integrated annotator level signal during the training phase [4,5]. The literature on these methods depicts gains in areas where uncertainty is inherent. Subjective text classification, perceptual judgmental tasks and the tasks where there is no single correct label that exists.

Most prior empirical analysis regimes: aggregate metrics, accuracy and macro F1 score are compared for the assumed improvements in the calibration of the model. Less attention has been paid to what may occur if the evaluation protocol heavily relies on the majority vote label, where gold labels are one-hot, while the model's distributional output depends on argmax for scoring, the evaluation will then fail to differentiate between strong fidelity and calibrated uncertainty when one appears to be more truthful than the other. To preserve annotator level signal, Plank [1] and others have argued for evaluation protocols, but in practice, most shared task evaluations score against the majority vote gold.

Flaccavento et al. [6] specifically in Enthymeme detection introduce the dataset and annotation protocols that form the basis for MediaEval 2026 and document the high frequency of disagreement on minority class labels. To our knowledge, no prior work has directly compared both supervision, i.e. hard vs soft, on this dataset.

3 APPROACH

3.1 Data and Disagreement Patterns

The merged_annotations_v2.csv release contains a total of 1,333 tweets, each of which is independently annotated by five annotators. Each tweet gets one of the three labels: none, premise, or conclusion. Across the whole dataset, these labels are distributed as 66.2%, 29.6% and 4.2% respectively. We have used a structured 89/11 train/ validation split with a seed of 42, that results in 1,185 training tweets and the remaining 148 validation tweets. The testing is done on the blind test set held by the organizer.

The Fleiss $\kappa \approx 0.574$, which is the inter annotator agreement on the full three label class task and this value is 0.629 on the binary class enthymeme vs none. The entropy of the per-class annotated vote label vector is highest on the conclusion class, with approximately 38% of the test instances having at least one annotator who disagrees with the majority and the rate is twice as high among occurrences where the majority of the annotator votes for conclusion as where there is none. The inspection of the individual annotator further reveals that more differences occurs in label use as one of the annotator was assigning the conclusion label 6.6 times more frequently than other annotators, this is a divergence which highlights two different conceivable frameworks for which cause an enthymeme to either have a missing conclusion or a missing premise, which concludes it to be unlikely a random noise.

We built soft labels by normalizing votes from the five annotator vote labels and per tweet, a probability distribution over the three class labels is established. A tweet with two of the none votes and three annotators voting them as a premise has a soft label [p_none=0.4, p_premise=0.6, p_conclusion=0.0]. A tweet where the general consensus of the annotator is the same without disagreement has a one-hot label and the KL term there converges to standard cross entropy.

3.2 Loss Formulations

BERTweet-base [7] with similar sequence classification serves as the backbone of both runs. The setup uses a BPE tokenizer with a maximum sequence length of 128 tokens, applies a dropout rate of 0.1 after pooling and includes a linear projection layer that produces three output logits. The model is trained for 10 epochs using the AdamW optimizer with (β_1=0.9, β_2=0.999 and weight decay = 0.01). The learning rate is set to = $2x10^{-5}$ with a linear warm-up schedule of 0.1, a batch size of 16, FP16 mixed precision and a gradient clipping threshold of 1.0. Model checkpoints are chosen based on the highest validation Macro F1 score. All experiments are run with a random seed of 42 on dual Tesla T4 GPUs and the two runs are only different in how the loss function is set up.

Run 1: weighted cross-entropy on majority labels.

$$L_1 = -\Sigma_i w_{yi} \log p\_\theta(y_i \mid x_i) \quad (1)$$

Where, y_i is the majority vote label for a tweet x_i and w_k denotes the inverse class frequency of class k in the training set. The weight of the rare conclusion class is 7.795. So this treats the majority vote as the ground truth and the weighted loss is attributed to class imbalance.

Run 2: cross-entropy plus KL against the soft label.

$$L_2 = \alpha \cdot CE\,(hard) + (1 - \alpha) \cdot KL\,(soft_i \mid model) \quad (2)$$

Here the $soft_i$ is the distribution of the annotators vote. The sharp gradient signal form the majority vote is because of the CE term, The KL term then normalizes predicted distribution toward the experimental annotator's one.

The mixing coefficient α is used as a critical hyperparameter. For equal weighting initially this is set to 0.5 but due to the KL term becoming significant over the imbalanced CE on hard cases where we noticed the conclusion class for its poor performance this resulted in deviating the model towards being the marginal class distribution which itself assigns it to either

none or premise. This resulted in rare class to lose it weighted CE protection. Setting α to 0.7 restored stable training and it retained the signal of the rare class gradient function and these are the results we are reporting which were produced when setting the value of α = 0.7. So no further investigation of finely tuning its value was performed and the transition from 0.5 to 0.7 was sufficient to recover the balance and promote stable results.

4 RESULTS AND ANALYSIS

4.1 Aggregate Results

Table 1 illustrates macro F1 score, accuracy and cross entropy on the blind test set used by the organizers, it is summarized for both runs and both classification

Table 1: **Task 1 evaluation on the blind test set**

Task	Run	Macro-F1	Accuracy	Cross-entropy
1A (binary)	Run 1	0.6679	0.6824	**0.7021**
1A (binary)	Run 2	0.6673	**0.6892**	0.8391
1B (three-class)	Run 1	0.4865	0.6486	**0.8691**
1B (three-class)	Run 2	**0.5087**	0.6486	1.0071

Best per column are highlighted in bold per task. Two patterns revealed here are that Run 2 wins Macro F1 on 1B but it loses on cross entropy on both of the tasks.

4.2 Per-Class Results on Task 1B

Table 2 depicts Task 1B macro F1 score by class. Due to the rare conclusion class the overall macro F1 is concentrated around it.

Table 2: **Per-class F1 on Task 1B**

Class	Support	Run 1	Run 2	Δ
none	97	0.7444	0.7487	+0.004
premise	45	0.5333	0.5106	−0.023
conclusion	6	0.1818	**0.2667**	**+0.085**

The conclusion class F1 score nearly doubles when using soft labels, but this comes with a small trade-off, which is a slight drop in the premise class. This change helps identify the minority class better improving the F1 score from 0.17 (1 out of 6 cases recovered) to 0.33 (2 out of 6). This small improvement shows the difference between a model that rarely predicts the minority class and one that can detect it more often. In a binary task where annotators have very low agreement, the distinction between premise and conclusion class labels becomes one category called an enthymeme and the advantage of soft labels disappears completely(Δ macro- F1 = −0.0006).

4.3 The Cross-Entropy Paradox

Run 2's higher cross entropy than Run 1 is recorded to be approximately 0.14 on both tasks and without further investigation this would be an evidence for the soft label, that it is well calibrated and is exactly the opposite of which its training target should produce. Cross Entropy is calculated against one hot majority vote labels in this shared task evaluation

$$CE = -\Sigma_i \log p\_\theta(y_i\_majority \mid x_i) \quad (3)$$

You can consider an instance of the hard case where the annotators are realistically split, where three voted for premise and two voted for none. Because of the majority vote label the label becomes premise. Run 2 is trained to match [0.4, 0.6, 0.0], predicts something close to [0.4, 0.6, 0.0] and incurs a loss of –log(0.6) ≈ 0.51 on the majority class. Run 1, trained against the one-hot [0.0, 1.0, 0.0], this predicts more sharply compared to [0.10, 0.85, 0.05] this incurs only –log(0.85) ≈ 0.16 on the same

item. Run 1 wins cross entropy because it has learnt to over-state confidence and it is not because that it is better in calibration and the state confidence of Run 1 is comparable to one-hot gold rewards.

The process is a generalization, whenever the gold is one hot and model output is distributional cross entropy so we get sharper and more confident predictions even when experimental annotator distribution is inherent. Run 1's training target being one hot causes it to learn sharper distributions. Run 2 because of the KL term learns to converge to outputs that reflects annotators' disagreements on the label. Under one hot evaluation the model less truthful to the annotator distributed scores higher. This is not reported as the calibration failure of Run 2. This is either an evaluation measure mismatch. The annotator distribution itself is the correct cross entropy comparison for a soft label, this may be called the soft cross entropy or the KL divergence against the experimental annotator vote which is an n-dimension vector. Without an access to a test occurrence of each of the annotator votes we are unable to calculate this on the blind test set but we report this as a recommendation being methodological in nature for future evaluations of these label distribution method of learning.

4.4 When Does Soft-Label Supervision Help?

We identify three conditions under which soft label supervision could work on this dataset by combining the metric level observations stated above on per class and per task:

Condition 1:Fine-grained label space; Soft label supervision helps with Task 1B, which is a three-class task but it does not work well for Task 1A, which is a binary classification task. When the labels are reduced to just two options, the small differences that make annotators disagree are lost. This means there is not enough detailed information left for the soft labels to capture and they end up only showing a simple yes or no outcome.

Condition 2: Minority class with a non-trivial disagreement; The conclusion class meets both conditions as it has the highest annotator disagreement and is underrepresented in the training data (occupying only 4.2% of the dataset). On the other hand, the premise class shows the most annotators' agreement and also and has a larger number of examples, which leads to a small drop in performance rather than an improvement. This supports the idea that the KL divergence term shifts probability away from the more common, highly agreed-upon classes and toward the less common ones where there is less agreement among annotators.

Condition 3: Evaluation where no penalty is assigned to a calibration ambiguity; Macro-F1 is calculated from argmax predictions depicts Run 2's benefits clearly. However, cross entropy calculated against one hot gold inverses the ranking because it operates on rewarding predictions that are made with high confidence. As a result, the choice of evaluation metric has a big impact on the conclusions about how effective soft labels are.

5 CONCLUSIONS

We have directly addressed the second Quest for Insight question for the MediaEval 2026 Missing Pieces and Misinformation task: Does modeling the full distribution of human judgments improve performance on borderline cases compared to majority-vote labels? Our answer is conditional. Yes, in the three-class task where the labels are detailed and the rare conclusion class has a lot of disagreement among annotators, using the full distribution leads to better results. The Macro-F1 score increases from 0.4865 to 0.5087, which is a 47% improvement in the F1-score for the conclusion class. However, soft labels work worse in binary tasks where disagreement is reduced and also when using cross-entropy evaluation. This is because cross-entropy with one-hot labels rewards confident predictions, which means it doesn't reward models that predict the full distribution, even if those predictions better match the actual human judgments.

Therefore, label distribution learning methods should be assessed using the actual annotator distribution instead of relying on majority-vote targets. Ignoring this leads to missing the benefits of using soft labels especially when evaluation is based on one-hot predictions.

We suggest three future directions. First, using a soft cross-entropy metric based on the real annotator distribution would fix the current mismatch and allow clearer comparisons. Second, doing a more detailed search for the mixing coefficient α based on class behavior might help fine-tune the model for stability especially if dynamic class weighting is used. Finally, combining the sharp, confident predictions from Run 1 with the broader coverage from Run 2 could lead to better results especially for rare classes where both types of predictions are important.

ACKNOWLEDGMENTS

We want to express our appreciation and thanks to the task organizers for managing the shared task smoothly and for offering their support, assistance and encouragement whenever we contacted them.

Declaration on Generative AI

While working on this paper, the authors used Grammarly to check for grammar and spelling mistakes. Once they were done with the tools, they went through the text again, made any necessary changes and were fully responsible for the content that was published.

REFERENCES

[1] Barbara Plank. 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 (EMNLP 2022)*, 10671–10682, 2022.

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

[3] Geoffrey Hinton, Oriol Vinyals and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[4] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 1120–1130, 2013.

[5] Marie-Catherine de Marneffe, Christopher D. Manning and Christopher Potts. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38, 2 (2012), 301–333, 2012.

[6] Alessandro M. Flaccavento, Youen Peskine, Paolo Papotti, Riccardo Torlone and Raphaël Troncy. Automated detection of tropes in short texts. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 1420–1432, 2025.

[7] Dat Quoc Nguyen, Thanh Vu and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020)*, pages 9–14, 2020.

[8] Martial Pastor and Nelleke Oostdijk. A Resource for Enthymeme Detection in Controversial Political Discourse. *arXiv preprint arXiv:XXXX.XXXXX*, 2026.