Weak Ensemble Learning from Multiple Annotators for Subjective Text Classification

Ziyi Huang[1], **N. R. Abeynayake**[2], **Xia Cui**[2]

[1]Hubei University, Wuhan, China.
ziyihuang@hubu.edu.cn
[2]Manchester Metropolitan University, Manchester, UK.
{n.abeynayake, x.cui}@mmu.ac.uk

Abstract

With the rise of online platforms, moderating harmful or offensive user-generated content has become increasingly critical. As manual moderation is infeasible at scale, machine learning models are widely used to support this process. However, subjective tasks, such as offensive language detection, often suffer from annotator disagreement, resulting in noisy supervision that hinders training and evaluation. We propose *Weak Ensemble Learning* (WEL), a novel framework that explicitly models annotator disagreement by constructing and aggregating weak predictors derived from diverse annotator perspectives. WEL enables robust learning from subjective and inconsistent labels without requiring annotator metadata. Experiments on four benchmark datasets show that WEL outperforms strong baselines across multiple metrics, demonstrating its effectiveness and flexibility across domains and annotation conditions.

1 Introduction

Harmful information, such as offensive and abusive language, has been known as one of the main threats on social media platforms. Typically, the moderation of online harmful information is conducted manually. With an increasing amount of information, manual moderation is expensive and insufficient. There is a growing demand for developing a Natural Language Processing (NLP) tool to support the detection and mitigation of harmful content on online platforms. Addressing this challenge requires high-quality annotated data to train accurate and reliable machine learning models. In recent years, social media has become a popular source for data collection, and crowdsourcing has emerged as a widely used solution for large-scale data annotation. However, concerns have been raised about the reliability of crowdworkers, particularly in complex linguistic tasks where annotators often lack domain-specific training (Uma et al., 2022). Furthermore, the incentive structures of crowdsourcing platforms can encourage rapid completion of tasks with careful judgment, potentially compromising label quality (Daniel et al., 2018; Leonardelli et al., 2021; Leonardellli et al., 2023). A particularly challenging issue in this context is human label variation (Plank, 2022), which arises when annotators assign different labels to the same instance. This is especially common in subjective tasks such as emotion detection (Buechel and Hahn, 2018) and offensive language detection (Leonardellli et al., 2023), where annotation involves personal interpretation, contextual nuance, and cultural perspective. Unlike objective tasks with clearly defined ground truth, subjective annotations inherently invite disagreement. Such variation introduces noise into training data, complicates evaluation, and challenges the assumption of a single "correct" label (Uma et al., 2022; Cabitza et al., 2023). Understanding and modelling this variability is critical for developing NLP systems that are more robust, interpretable, and aligned with the diversity of human judgement.

Previously, several methods have been proposed to address this issue by estimating and incorporating annotator reliability into the modelling process (Sheng et al., 2008; Cui, 2023; Fleisig et al., 2023; Xu et al., 2024). These approaches typically assign higher weights to labels provided by more consistent or trustworthy annotators, aiming to reduce the influence of noisy or unreliable inputs on the final model. However, their effectiveness is often limited by the composition of the annotator pool. They require sufficient diversity among annotators to model reliability accurately and may risk overfitting when such diversity is lacking or when the model overrelies on a small subset of annotators (Räbiger et al., 2018; Cui, 2023).

We aim to develop a method applicable to more general multi-annotation settings. Specifically, the proposed approach is designed to function effec-

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 87–99

tively when annotators are shared across the entire dataset or when there is a heterogeneous distribution of annotator workload (i.e., some annotators contribute more than others).

While prior approaches often rely on a single loss function, such as cross-entropy (CE) (Uma et al., 2020), to train classification models, this may be insufficient for subjective tasks where both hard and soft supervisory signals are informative. In such settings, different loss components capture complementary aspects of learning: CE supports probabilistic calibration, F1 loss promotes classification accuracy on hard labels, and distributional losses like mean absolute error or Manhattan distance (MD) (Rizzi et al., 2024) help align predictions with the soft label distributions reflecting annotator disagreement. By jointly optimising these objectives, we can balance predictive accuracy with nuanced representation of label uncertainty, leading to more robust and interpretable models.

Our contributions can be summarised as follows:

- We propose *Weak Ensemble Learning*[1] (WEL), a novel ensemble-based framework for learning from multiple annotations in subjective tasks.
- We introduce two variants: *WEL-Random*, which builds weak predictors from randomly sampled labels to capture annotator variation without metadata, and *WEL-TopAnn*, which trains per-annotator models for the top-ranked annotators.
- We present a systematic study of selection strategies, aggregation methods and loss functions for optimising the ensemble.
- Experiments on four datasets from Le-Wi-Di 2023 shared task show that WEL consistently outperforms two strong baselines across multiple metrics.

2 Related Work

Subjective NLP tasks such as offensive language detection, hate speech classification and emotion analysis often suffer from high variability in human annotations. Annotators may interpret linguistic cues differently based on their personal, cultural or contextual backgrounds (Aroyo and Welty, 2015; Uma et al., 2022). This subjectivity introduces label noise and inconsistency, making it challenging to define a single ground truth (i.e., a hard label). In particular, datasets annotated via crowdsourcing tend to reflect these disagreements, raising questions about how best to represent and learn from multiple perspectives (Leonardelli et al., 2021; Davani et al., 2022).

A common approach to address label disagreement is to replace hard labels with soft targets, usually probability distributions over classes derived from annotator votes, and train models using probabilistic loss functions. The most prevalent is the cross-entropy loss, which treats soft distributions as targets, encouraging models to reflect label uncertainty rather than force a single decision (Uma et al., 2020; Zheng et al., 2021). More recent methods have proposed alternative loss formulations, such as Kullback-Leibler divergence, expected calibration error (Uma et al., 2020) and Manhattan distance (Rizzi et al., 2024). These techniques aim to improve robustness to noisy or subjective labels by preserving the signal in disagreement rather than collapsing it through majority vote.

Ensemble methods have also been explored as a way to leverage annotator disagreement rather than suppress it. Instead of aggregating labels before training, several works train separate models for each annotator and combine their predictions during inference (Akhtar et al., 2021; Gordon et al., 2021; Xu et al., 2024). This strategy captures the full range of annotator perspectives and has shown promise in capturing subjective variation in tasks like emotion classification and hate speech detection. However, these models may suffer from scalability issues, especially when the number of annotators is large or unbalanced. Other work has approached the problem from a probabilistic modelling perspective, estimating annotator reliability as a latent variable during training (Paun et al., 2018a,b; Xu et al., 2020). These approaches often combine annotator-specific models with global learning signals, aiming to balance personalised and consensus-based predictions. In addition, instance weighting has been used as a practical solution to reduce the influence of unreliable or biased supervision. For instance, Zhang et al. (2020) apply instance reweighting to mitigate demographic bias in toxicity detection, while Liu et al. (2021) introduce dynamic instance weighting to adapt to concept drift in evolving datasets. Cui (2023) and Fleisig et al. (2023) proposed to compute individual annotator ratings and combine this information to better capture the subjectivity inherent. These methods adjust the learning signal based on

[1]Codebase for WEL and Evaluation: https://github.com/YhzyY/Weak-Ensemble-Learning

example-level characteristics, enabling models to better generalise under noisy or imbalanced conditions.

Our work unifies ensemble-based disagreement modelling. We extend ensemble methods that capture annotator disagreement by randomly sampling weak predictors to simulate diverse viewpoints, and by embedding annotator-specific models whose ensemble selection is learned end-to-end rather than relying on a fixed set as in Xu et al. (2024). In contrast to probabilistic reliability estimation techniques that depend on annotator metadata (Paun et al., 2018a; Xu et al., 2020), our framework requires no such information, broadening its applicability. At the ensemble level, we adapt instance-weighting strategies to emphasise predictor utility and mitigate dataset bias (Zhang et al., 2020; Liu et al., 2021). Drawing on label distribution modelling, our loss function blends soft and hard supervision to achieve both nuanced learning and interpretability (Tian et al., 2024). These elements yield a scalable and flexible approach for managing noisy subjective annotations.

3 Methods

Given a dataset annotated by multiple annotators, the goal is to learn a predictive function that accounts for the variability and potential noise introduced by differing annotator judgments. Let $\mathcal{D} = \{(x_i, y_i^{(1)}, \ldots, y_i^{(A_i)})\}_{i=1}^{N}$ denote a dataset of N instances, where $x_i \in \mathcal{X}$ is the input (e.g., a text sample), and $\{y_i^{(1)}, \ldots, y_i^{(A_i)}\}$ are the labels provided by A_i annotators for instance x_i, with $y_i^{(j)} \in \mathcal{Y}$ representing the label from the j-th annotator, which $j \in \{1, \ldots, J\}$ and J is the total number of annotators in $\mathcal{D}$. The objective is to learn a predictive function $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterised by θ, that approximates the underlying true label distribution y_i^*, which is unobserved due to annotator disagreement.

To address this challenge, we propose a three-stage method, named *Weak Ensemble Learning* (WEL), designed to learn from multiple annotators while accounting for disagreement and annotator variability. First, we construct a set of weak predictors by employing a *random sampling* and a *top-ranked annotators* selection strategies (Section 3.1). Second, we aggregate the outputs of weak predictors using a weighted ensemble, where the weights will be tuned to balance contributions in the next stage, enabling the model to leverage diverse annotator perspectives effectively (Section 3.2). Finally,

Algorithm 1 Weak Ensemble Learning (WEL)

Input: Dataset $\mathcal{D} = \{(x_i, \{y_i^{(j)}\}_{j=1}^{A_i})\}_{i=1}^{N}$; Loss coefficients α, β, γ; Regularisation weight λ; Maximum number of weak predictors M_{max}
Output: Final predictive ensemble model $f(x) = \sum_{m=1}^{M^*} w_m f_{\theta_m}(x)$

Stage 1: Construct Weak Predictors

Strategy 1: Random Sampling

for $m = 1$ *to* M_{max} **do**
  $\mathcal{D}^{(m)} = \{(x_i, y_i^{(j)}) \mid j \sim \text{Uniform}(1, A_i)\}$
  Train f_{θ_m} on $\mathcal{D}^{(m)}$
end

Strategy 2: Top-Ranked Annotators

Compute the annotation counts of a set of annotators $\{A_1, \ldots, A_{M_{\text{max}}}\}$
for $m = 1$ *to* M_{max} **do**
  $\mathcal{D}^{(m)} = \{(x_i, y_i^{(A_m)}) \mid A_m \text{ annotated } x_i\}$
  Train f_{θ_m} on $\mathcal{D}^{(m)}$
end

Stage 2: Define Aggregated Supervision

foreach *instance* x_i **do**
  Compute hard-aggregated label:
$$\bar{y}_i^{\text{hard}} = \arg\max \sum_{j=1}^{A_i} (y_i^{(j)} = c),$$
  Compute soft-aggregated label:
$$\bar{y}_i^{\text{soft}}[c] = \frac{1}{A_i} \sum_{j=1}^{A_i} (y_i^{(j)} = c)$$
end

Stage 3: Joint Optimisation

Initialize ensemble size $M \leftarrow M_{\text{max}}$ and weights $\mathbf{W} = [w_1, ..., w_M]$
repeat
  foreach *instance* x_i **do**
    Compute ensemble output:
$$\hat{y}_i = \sum_{m=1}^{M} w_m f_{\theta_m}(x_i)$$
  end
  Compute total loss:
$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{F1}} + \beta \cdot \mathcal{L}_{\text{CE}} + \gamma \cdot \mathcal{L}_{\text{MD}} + \lambda \cdot \Omega(\mathbf{W})$$
  Update ensemble weights $\mathbf{W}$
  Prune predictors: retain only f_{θ_m} such that $w_m > \epsilon$, for $m = 1, \ldots, M$
  Reinitialise weights: $\mathbf{W} \leftarrow \mathbf{W} / \sum_{m=1}^{M} w_m$
until *convergence*;
Set $M^* \leftarrow M$ and return final ensemble: $f(x) = \sum_{m=1}^{M^*} w_m f_{\theta_m}(x)$

we jointly optimise the weak predictors' ensemble weights by minimising a multi-objective loss over soft and hard aggregated labels, balancing cross-entropy, distributional similarity, and F1-score performance (Section 3.3). The complete procedure of WEL is described in Algorithm 1.

3.1 Weak Predictor Construction

To capture diverse annotator perspectives, the first stage of WEL constructs M weak predictors $\{f_{\theta_1}, \ldots, f_{\theta_M}\}$, each trained on a different slice of the annotation space. We propose two selection strategies:

Random Annotator Selection. For each training instance x_i with A_i annotations $\{y_i^{(1)}, \ldots, y_i^{(A_i)}\}$, we sample one label $y_i^{(j)}$ uniformly at random:

$$j \sim \text{Uniform}\{1, \ldots, A_i\} \tag{1}$$

Repeating this process M times produces M datasets $\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(M)}\}$, each reflecting a single-annotator view.

Top-Ranked Annotator Selection. We identify the M annotators with the largest label contributions and train one weak predictor per annotator using only their labels. Unlike Xu et al. (2024), which assumes a fixed set of annotators, our M is treated as a *learnable parameter* in the optimisation stage, allowing the ensemble size to adapt to the dataset.

Model Architecture. Each weak predictor f_{θ_m} consists of a Transformer encoder (BERT or AraBERT) followed by a linear classification head mapping the `[CLS]` representation to class logits:

$$z = W h_{\texttt{[CLS]}} + b, \tag{2}$$

where $h_{\texttt{[CLS]}} \in \mathbb{R}^d$ is the encoder output, $W \in \mathbb{R}^{C \times d}$, $b \in \mathbb{R}^C$, and C is the number of classes. The logits are passed through a softmax layer to produce probability distributions over classes:

$$\hat{y} = \text{softmax}(z) \tag{3}$$

This stage yields a diverse pool of predictors that differ in training data and potentially in decision boundaries, forming the foundation for weighted ensemble learning in Section 3.2.

3.2 Weighted Ensemble Learning

In the second stage, we aggregate the probability outputs from the M weak predictors $\{f_{\theta_1}, \ldots, f_{\theta_M}\}$ into a single ensemble prediction. Let $\hat{y}_i^{(m)} \in [0,1]^C$ denote the predicted probability distribution over C classes for instance x_i from the m-th weak predictor, computed via the softmax output of its linear classification head (Section 3.1).

We adopt a *weighted ensemble* strategy, where each predictor is assigned a learnable non-negative weight w_m subject to the constraint $\sum_{m=1}^{M} w_m = 1$. The ensemble prediction is then:

$$\hat{y}_i = \sum_{m=1}^{M} w_m \, \hat{y}_i^{(m)} \tag{4}$$

Here, $\mathbf{W} = [w_1, \ldots, w_M] \in \mathbb{R}^M_{\geq 0}$ encodes the contribution of each weak predictor to the final decision.

While up to M_{max} predictors can be initially constructed, the optimisation process (Section 3.3) automatically determines an effective subset $M^* \leq M_{\text{max}}$. Predictors with $w_m < \epsilon$ (e.g., $\epsilon = 10^{-3}$) are pruned to improve computational efficiency and reduce noise from low-utility models.

By combining multiple probability distributions, this ensemble mechanism captures complementary information from diverse annotator views, improving robustness and mitigating the bias of any single weak predictor.

3.3 Optimisation

In the third stage, we optimise the ensemble weights $\{w_m\}_{m=1}^M$ (and optionally other parameters) to improve predictive performance. Given the ensemble prediction $\hat{y}_i$ from Eq. (4), computed as the weighted sum of individual predictor outputs $\hat{y}_i^{(m)}$, our goal is to minimise a multi-objective loss that balances classification accuracy, calibration, and distributional alignment.

To accommodate the uncertainty introduced by annotator disagreement, we investigate learning from both *soft-aggregated* and *hard-aggregated* labels, and explore separate and joint optimisation strategies based on multiple objective functions.

3.3.1 Aggregated Supervision

Let $\mathcal{D} = \{(x_i, \{y_i^{(j)}\}_{j=1}^{A_i})\}_{i=1}^N$ be a dataset annotated by multiple annotators. We derive two forms of supervision:

- **Hard Aggregated Label** $\bar{y}_i^{\text{hard}} \in \mathcal{Y}$: computed via majority vote over annotator labels.

- **Soft Aggregated Label** $\bar{y}_i^{\text{soft}} \in [0,1]^C$: a normalised label distribution over C classes, reflecting the empirical frequency of annotators' choices.

3.3.2 Objectives

To robustly train the ensemble model under varying supervision signals, we define the following optimisation targets of the loss function $\mathcal{L}$:

(1) F1-Score (F1): A discrete metric evaluated using $\bar{y}_i^{\text{hard}}$, which we aim to maximise:

$$\mathcal{L}_{\text{F1}} = -\text{F1}(\arg\max(\hat{y}_i), \bar{y}_i^{\text{hard}}), \tag{5}$$

where the negative sign denotes that the F1-score is being maximised during training.

(2) Cross-Entropy (CE) (Uma et al., 2020; Leonardellli et al., 2023): A soft objective used when training with $\bar{y}_i^{\text{soft}}$, minimising:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \bar{y}_i^{\text{soft}}[c] \cdot \log \hat{y}_i[c], \tag{6}$$

where N is the number of training instances, C the number of classes, $\bar{y}_i^{\text{soft}}[c]$ the soft target (i.e., annotator-derived label distribution), and $\hat{y}_i[c]$ the predicted probability for class c on instance x_i.

(3) Average Manhattan Distance (MD) (Rizzi et al., 2024): A distributional similarity measure minimising the L_1 distance between predicted and soft labels:

$$\mathcal{L}_{\text{MD}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} |\hat{y}_i[c] - \bar{y}_i^{\text{soft}}[c]|_1 \tag{7}$$

3.3.3 Separate and Joint Optimisation

We explore two optimisation paradigms:

- **Separate Optimisation :** Each objective is minimised independently in different optimising regimes. For example, cross-entropy is minimised on soft labels, while F1-score is optimised using hard labels during model aggregation.

- **Joint Optimisation:** A combined loss function integrates all objectives, Eq. (5), (6) and (7), to guide the model jointly. We define:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{F1}} + \beta \cdot \mathcal{L}_{\text{CE}} + \gamma \cdot \mathcal{L}_{\text{MD}} + \lambda \cdot \Omega(\mathbf{W}), \tag{8}$$

where $\alpha, \beta, \gamma \geq 0$ are loss balancing coefficients, and $\Omega(\mathbf{W})$ is an ℓ_2-norm regularisation term to encourage balanced weight distributions to prevent overfitting. The regularisation weight λ controls the degree of smoothing across predictors.

By jointly optimising prediction objectives and ensemble composition, our model leverages annotator disagreement as a source of diversity, improving both robustness and alignment with subjective supervision.

3.3.4 Implementation Details

During the optimisation stage, we employ two derivative-free optimisation algorithms: OPTUNA (Akiba et al., 2019) and the `SciPy`[2] differential evolution algorithm (Storn and Price, 1997). Both are well-suited to searching continuous, bounded parameter spaces without requiring gradient information. In our setting, the optimiser iteratively updates the ensemble weights $\mathbf{W}$ to minimise the chosen objective(s) (either a single loss or the joint formulation in Eq. (8)), subject to the simplex constraint:

$$w_m \geq 0, \quad \sum_{m=1}^{M} w_m = 1. \tag{9}$$

For each optimisation step, the selected subset of weak predictors is reinitialised to reduce sensitivity to specific model subsets. We run each optimiser for up to 100 trials or steps, and both methods yield comparable results. OPTUNA is generally faster due to GPU support and efficient sampling strategies, while differential evolution offers robust CPU-based parallelism, making it preferable in non-GPU environments. The framework remains agnostic to the choice of optimiser, allowing other search strategies to be integrated as needed.

4 Experiments

To evaluate the performance of our method across diverse domains and text genres, we use

[2] https://docs.scipy.org/

Table 1: Data statistics for the four textual datasets. #Train, #Dev, and #Test denote the number of instances in the training, development, and test splits, respectively. #TotalAnn indicates the total number of annotators in each dataset, while #Ann represents the minimum and maximum number of annotators per instance.

Dataset	#Train	#Dev	#Test	#TotalAnn	#Ann
ArMIS	657	141	145	3	3
ConvAbuse	2398	812	840	8	2-7
HS-Brexit	784	168	168	6	6
MD-Agreement	6592	1104	3057	670	5

four publicly available datasets from the Le-Wi-Di shared task at SemEval 2023 (Leonardelli et al., 2023): **ArMIS** (Almanea and Poesio, 2022), **ConvAbuse** (Cercas Curry et al., 2021), **HS-Brexit** (Akhtar et al., 2021), and **MD-Agreement** (Leonardelli et al., 2021). Each dataset includes multiple annotations per instance, with at least two annotators per instance sample.

To maintain the generalisability of our approach, we exclude domain information and annotator metadata during training. All models are trained solely on input text and its associated labels. Summary statistics are provided in Table 1, while Table 2 presents dataset meta-information. In particular, we distinguish between *Fixed Ann.* datasets, where each instance is labelled by the same group of annotators, and *Mixture* datasets, where annotators vary across instances. Further details on the datasets and preprocessing steps are provided in Appendix A.

Table 2: Dataset metadata covering annotator contribution, diversity, language and genre.

Dataset	Contribution	Diversity	Language	Genre
ArMIS	Fixed Ann.	Low	Arabic	Short Text
ConvAbuse	Mixture	Low	English	Conversation
HS-Brexit	Fixed Ann.	Low	English	Short Text
MD-Agreement	Mixture	High	English	Short Text

4.1 Training

While the proposed framework is model-agnostic and compatible with various machine learning architectures, we employ BERT (Devlin et al., 2019) for English datasets (ConvAbuse, HS-Brexit, MD-Agreement) and AraBERTv2 (Antoun et al., 2020) for ArMIS, using the `base` checkpoints from HuggingFace. We train $M_{\max}$=10 weak predictors using different selection strategies (Section 3.1). Hyperparameters for Transformers are tuned on development sets (Appendix B). The predictors are fixed before the joint optimisation of ensemble weights.

4.2 Evaluation Metrics

We evaluate model performance using three complementary metrics: (a) micro-averaged F1 score (F1), which assesses classification accuracy on hard-aggregated labels; (b) cross-entropy loss (CE); and (c) average Manhattan distance (MD) between predicted and target label distributions. The latter two metrics are used to evaluate how well the model captures soft supervision signals arising from annotator disagreement (Leonardellli et al., 2023; Rizzi et al., 2024).

4.3 Label Selection Strategies

First, we experiment with two label selection strategies for constructing weak predictors: *random sampling* (*Random*), which selects one annotation per instance uniformly at random, and *top-ranked annotators* (*TopAnn*), which trains one model per annotator using data from the most frequent annotators. For simplicity and fair comparison, we fix all loss coefficients and the regularisation weight to 1.

Table 3 shows results across the four datasets. The *Random* strategy consistently achieves higher F1 and better CE than *TopAnn*. We attribute this to the greater diversity introduced by random sampling: each weak predictor is trained on a unique stochastic projection of the label space, encouraging the ensemble to learn decision boundaries that generalise across annotator-specific biases. This is especially beneficial when F1 is the main objective, as it rewards consistent hard-label predictions on majority-vote labels, which Random sampling implicitly approximates over many diverse predictors.

By contrast, *TopAnn* tends to produce more similar decision boundaries within the ensemble because each predictor is tied to a single annotator's style. This can be beneficial for modelling annotator-specific distributions, but under fixed coefficients, it can limit the ensemble's ability to optimise for F1, which benefits from capturing the aggregate rather than individual perspectives.

Nevertheless, *TopAnn* achieves lower MD on ConvAbuse and HS-Brexit, likely because these datasets have annotators with high internal consistency. In such cases, modelling them individually yields predictions more aligned with the soft label distribution.

Table 3: The selection strategies for constructing weak predictors on AraBERT and BERT.

Dataset	Selection	F1	CE	MD
ArMIS	Random	**0.7310**	**0.6390**	**0.5301**
	TopAnn	**0.7310**	0.6536	0.5487
ConvAbuse	Random	**0.9333**	**0.5559**	0.1749
	TopAnn	0.9310	0.5652	**0.1645**
HS-Brexit	Random	**0.9107**	**0.5842**	0.2733
	TopAnn	0.8929	0.6140	**0.2379**
MD-Agreement	Random	**0.8162**	**0.6246**	**0.3648**
	TopAnn	0.7668	0.6695	0.4156

4.4 Ensemble Optimisation Paradigms

We conduct an ablation study to assess the individual and combined contributions of the loss components in Eq. (8): $\mathcal{L}_{\text{F1}}$, $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{MD}}$. For clarity, we fix the selection strategy to *Random* and activate specific losses by setting their corresponding coefficients (α, β, γ) to 1 while setting the others to 0. In each setting, we optimise both the ensemble weights $\mathbf{W}$ and the number of members M.

Tables 4 and 5 show results for the ArMIS and MD-Agreement datasets. Across both datasets, $\mathcal{L}_{\text{MD}}$ consistently achieves the lowest MD values, confirming its role in aligning predictions with annotator label distributions. Similarly, configurations including $\mathcal{L}_{\text{CE}}$ tend to improve calibration (lower CE), while $\mathcal{L}_{\text{F1}}$ boosts classification accuracy when paired with $\mathcal{L}_{\text{MD}}$. However, using all three objectives together does not yield additional gains, and in some cases slightly reduces performance, likely due to competing optimisation signals. Overall, these results suggest that each loss serves a distinct purpose: $\mathcal{L}_{\text{F1}}$ strengthens hard-label accuracy, $\mathcal{L}_{\text{CE}}$ improves probabilistic calibration, and $\mathcal{L}_{\text{MD}}$ enhances alignment with annotator distributions. Effective combinations emerge when the selected losses complement rather than compete, even without tuning the loss coefficients, underscoring the value of a flexible and modular objective in ensemble optimisation. The ConvAbuse and HS-Brexit datasets corroborate these findings, with further analysis provided in Appendix D. Similar results were also found using *TopAnn*.

4.5 Loss Coefficients and Regularisation Term

The joint objective in Eq. (8) balances four components through parameters (α, β, γ and λ) with the regularisation term $\Omega(\mathbf{W})$ demonstrating three key effects. Due to the page limit, we present the impact of $\Omega(\mathbf{W})$ on **MD-Agreement** in Table 6: (1) F1 improvement (up to +0.0056), (2) CE reduction (max -0.0005 for $\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$) and (3) MD gains in soft supervision (-0.0008) with limited degradation ($\leq$+0.0020 for $\mathcal{L}_{\text{MD}}$ alone).

Table 4: Ablation study of loss optimisation paradigms on ArMIS dataset. In each setting, one loss component is activated (associated scaler set to 1), while the remaining components are deactivated (set to 0).

Case	F1	CE	MD
$\mathcal{L}_{\text{F1}}$ only	0.7448	0.6395	0.5048
$\mathcal{L}_{\text{CE}}$ only	0.7379	**0.6385**	0.5252
$\mathcal{L}_{\text{MD}}$ only	0.7379	0.6505	**0.4900**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	0.7448	0.6412	0.5179
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	0.7172	0.6505	0.5111
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	**0.7517**	0.6406	0.5294
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.6897	0.6468	0.5243

Table 5: Ablation study of loss optimisation paradigms on MD-Agreement dataset.

Case	F1	CE	MD
$\mathcal{L}_{\text{F1}}$ only	0.8132	0.6249	0.3672
$\mathcal{L}_{\text{CE}}$ only	0.8119	0.6246	0.3633
$\mathcal{L}_{\text{MD}}$ only	0.8165	0.6250	**0.3626**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	0.8145	**0.6245**	0.3660
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	**0.8175**	**0.6245**	0.3670
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.8109	0.6247	**0.3626**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.8142	**0.6245**	0.3647

We conduct a Spearman correlation analysis (Kendall and Stuart, 1969) over four parameters in the objective function, each sampled from the range $[0, 0.001, 0.01, 0.1, 1]$, resulting in 1,295 unique combinations per dataset (excluding 0s for all). The F1 coefficient α significantly improves F1 ($\geq$+0.9) while degrading MD ($\geq$+0.7), with similar but weaker trade-offs for β (CE-focused) and γ (MD-focused). The regularisation strength λ shows model-dependent effects, enhancing F1 on BERT ($\approx$+1.0) but reducing performance on AraBERT ($\leq$-0.9). Finally, the optimised number of weak predictors M strongly correlates with both improved F1 ($\geq$+0.9) and reduced CE ($\leq$-0.9), though typically at the cost of MD degradation (ΔMD $\geq$+0.5) in BERT implementations.

4.6 Model Aggregation Strategies

Table 9 presents results on four datasets using three aggregation strategies for combining weak predictors: (a) *Voting*, which applies majority voting over class labels; (b) *Averaging*, which computes the unweighted mean of probabilistic outputs; and (c) *Optimised*, which learns weighted combinations through loss-minimising ensemble optimisation. In binary classification settings, *Voting* and *Averag-*

Table 6: Improvements when adding regularisation term $\Omega(\mathbf{W})$, Δ = with $\Omega(\mathbf{W})$ - without $\Omega(\mathbf{W})$.

Case	ΔF1	ΔCE	ΔMD
$\mathcal{L}_{\text{F1}}$ only	+0.0033	-0.0003	-0.0013
$\mathcal{L}_{\text{CE}}$ only	+0.0039	-0.0001	+0.0001
$\mathcal{L}_{\text{MD}}$ only	+0.0026	-0.0005	+0.0020
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	0.0000	0.0000	0.0000
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	-0.0036	+0.0001	+0.0010
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	+0.0056	-0.0005	-0.0008
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	+0.0020	+0.0001	+0.0001

Table 7: Correlation between parameter and evaluation metrics (F1, CE and MD) on the ArMIS and MD-Agreement datasets. * indicates statistical significance ($p < 0.05$). Green indicates improvement, red indicates degradation. For CE and MD, negative correlations are desirable.

Dataset	ArMIS			MD-Agreement		
Param	F1	CE	MD	F1	CE	MD
α	+0.9*	+0.7	+1.0*	+1.0*	+0.6	+1.0*
β	+0.2	-1.0*	+0.4	-0.7	-0.6	+0.3
γ	-0.9*	+1.0*	-1.0*	-0.7	-0.9*	-1.0*
λ	-1.0*	-1.0*	+0.7	+0.9*	-0.3	+1.0*
M	+0.01	-1.0*	-0.37	+0.98*	-0.97*	+0.82*

ing yield identical predictions under a shared 0.5 threshold (Hovy et al., 2013; Plank et al., 2014). Across all datasets, the *Optimised* strategy consistently achieves superior performance in F1 and MD, highlighting the benefit of learning ensemble weights tailored to the task and supervision signal. A slight performance drop is observed in CE on the ConvAbuse and HS-Brexit datasets. This may be due to the optimisation process prioritising improvements in classification accuracy (F1) and distributional alignment (MD), potentially at the expense of precise probabilistic calibration (CE).

4.7 Comparison with Baseline Models

To ensure a fair comparison, we use the same model backbone with identical hyperparameters (BERT for English datasets and AraBERT for ArMIS). We reimplement and evaluate two baseline approaches:

- **BERT-CE** (Uma et al., 2020): a non-ensemble single model optimised using a CE-focused soft loss function.
- **Top-5 Annotator Voting (Top-5 Voting)** (Xu et al., 2024): a majority-vote ensemble of per-annotator models, each trained on labels from one of the top 3 or 5 most frequent annotators (depending on availability). Unlike the original version, which used multiple BERT variants, we adopt a uniform model architecture across all predictors for consistency.

Table 8 shows the best results of our proposed method under two selection strategies: *random sampling* (WEL-Random) and *top-ranked annotator* (WEL-TopAnn). Results correspond to the optimal configurations found via ensemble optimisation and parameter tuning (Appendix C). Both WEL variants consistently outperform the baselines across most evaluation metrics, demonstrating the effectiveness of jointly optimising ensemble weights while capturing annotator diversity through weak predictors. The only exceptions occur in MD on the ConvAbuse and HS-Brexit datasets, where WEL-TopAnn outperforms both WEL-Random and the baselines. Additionally, in terms of F1, WEL-Random consistently exceeds the baselines, reinforcing the robustness of the ensemble approach even with random annotator selection. As noted in Section 4.3, the superior performance of WEL-TopAnn in MD likely reflects the influence of a few highly consistent annotators, which benefits the top-ranked selection strategy. However, WEL-Random remains competitive across other metrics (F1 and CE), suggesting that the ensemble framework is effective even without explicit annotator ranking.

5 Conclusions

In this paper, we introduced *Weak Ensemble Learning* (WEL), a flexible framework for subjective text classification that learns from multiple annotations by constructing diverse weak predictors and jointly optimising their contributions. We explored two variants: WEL-Random, which captures annotator variation through random label sampling, and WEL-TopAnn, which models the most frequent annotators individually. Experiments on four datasets showed that WEL consistently outperforms baselines, with WEL-Random excelling in hard-label classification and WEL-TopAnn offering advantages in distributional alignment when annotator consistency is high. Future work will integrate annotator profiles and reliability estimates into a unified neural architecture to improve performance and efficiency, and extend WEL to larger annotator pools and multilingual contexts.

Limitations

Although our method provides a general and scalable approach to learning from annotator disagreement, it has several limitations. First, we train weak

Table 8: Comparison with baseline models. * indicates a statistically significant difference ($p < 0.05$, t-test) from the BERT-CE baseline in terms of predicted labels (hard evaluation metric, F1) or soft distributions (soft evaluation metrics, CE and MD).

Dataset	ArMIS			ConvAbuse			HS-Brexit			MD-Agreement		
Metric	F1	CE	MD	F1	CE	MD	F1	CE	MD	F1	CE	MD
BERT-CE	0.6596	0.8039	0.7144	0.8362	0.9671	4.8068	0.7917	0.7652	0.7985	0.7880	0.9948	1.7574
Top-5 Voting	0.7310	0.6529	0.5498	0.9310	0.5651	0.1648	0.8929*	0.6154*	0.2394*	0.7808*	0.6629*	0.3995*
WEL-Random	**0.7793***	0.6385	**0.5028**	**0.9405**	**0.5577**	0.1709	**0.9167**	**0.5889***	0.2585*	**0.8214***	**0.6245***	**0.3632***
WEL-TopAnn	0.7448	**0.6362**	0.5143	0.9321	0.5662	**0.1586**	0.8929*	0.6237*	**0.2354***	0.7815*	0.6636*	0.4034*

Table 9: Aggregation strategies for the weak predictors.

Dataset	Method	F1	CE	MD
ArMIS	Voting	0.7172	0.6389	0.5216
	Averaging	0.7172	0.6389	0.5216
	Optimised	**0.7793**	**0.6385**	**0.5028**
ConvAbuse	Voting	0.9333	**0.5545**	0.1814
	Averaging	0.9333	**0.5545**	0.1814
	Optimised	**0.9405**	0.5577	**0.1709**
HS-Brexit	Voting	0.9107	**0.5845**	0.2874
	Averaging	0.9107	**0.5845**	0.2874
	Optimised	**0.9167**	0.5889	**0.2585**
MD-Agreement	Voting	0.8178	**0.6245**	0.3659
	Averaging	0.8178	**0.6245**	0.3659
	Optimised	**0.8214**	**0.6245**	**0.3632**

predictors independently and do not update their parameters during joint optimisation. Although this design improves computational efficiency, it can limit the capacity of the ensemble to adapt if individual predictors are poorly calibrated or sub-optimal. Second, while we evaluate across multiple data sources, our experiments are limited to mostly short social media texts (3/4), two languages and binary classification settings for simplicity. Additional evaluation of long-form text, multilingual corpora or structured annotation settings would help assess generalisability (Uma et al., 2022).

Ethical Statements

This research uses publicly available datasets from the SemEval-2023 Le-Wi-Di shared task, including user-generated content from social media and conversational agents. The datasets contain potentially sensitive language related to hate speech, offensive content, and abuse and were originally collected and annotated under ethical guidelines by their respective authors. We do not attempt to identify or profile any individual users or annotators. Our work focuses on improving the robustness and fairness of machine learning models in the presence of subjective disagreement and does not aim to make normative judgments about content or annotators. To support reproducibility and transparency, we use standard preprocessing, avoid introducing annotator-level biases, and refrain from incorporating demographic or personal information. All experiments are conducted following standard ethical practices for human-centred AI research, with a focus on minimising harm and respecting annotator diversity.

Acknowledges

We are deeply grateful to the reviewers for their thorough evaluation and insightful recommendations, which helped us enhance the clarity, rigour and impact of this paper.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. pages 9–15, Marseille, France. European Language Resource Association.

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

Sven Buechel and Udo Hahn. 2018. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings*

of the 27th International Conference on Computational Linguistics, pages 2892–2904, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xia Cui. 2023. xiacui at SemEval-2023 task 11: Learning a model in mixed-annotator datasets using annotator ranking scores as training weights. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1076–1084, Toronto, Canada. Association for Computational Linguistics.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. 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*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Maurice G Kendall and Alan Stuart. 1969. The advanced theory of statistics. vol. 3. *Biometrics*, 25(2):435.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elisa Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Anjin Liu, Jie Lu, and Guangquan Zhang. 2021. Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):293–307.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018a. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio. 2018b. A probabilistic annotation model for crowdsourcing coreference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1937, Brussels, Belgium. Association for Computational Linguistics.

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*, Abu Dhabi. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Stefan Räbiger, Myra Spiliopoulou, and Yücel Saygın. 2018. How do annotators label short texts? toward understanding the temporal dynamics of tweet labeling. *Information Sciences*, 457-458:29–47.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Rainer Storn and Kenneth Price. 1997. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359.

Xiaoyu Tian, Yongbin Qin, Ruizhang Huang, and Yanping Chen. 2024. A Label Information Aware Model for Multi-label Text Classification. *Neural Processing Letters*, 56(5):242.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. volume 72, page 1385–1470, El Segundo, CA, USA. AI Access Foundation.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. 2020. Automatic perturbation analysis for scalable certified robustness and beyond. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experimental Data Preprocessing and Implementation Details

Three of the datasets (**ArMIS**, **HS-Brexit**, and **MD-Agreement**) consist of tweets collected from X[3]. The **ArMIS** dataset comprises Arabic tweets labelled for misogyny detection, focusing on offensive language directed toward women. The **HS-Brexit** dataset includes English tweets annotated for hate speech related to Brexit. The **MD-Agreement** dataset contains English tweets labelled for offensive language across three domains: Black Lives Matter, Elections, and COVID-19. To maintain the generalisability of our approach, we do not use domain information from the MD-Agreement dataset during training. For these three Twitter-based datasets, we apply a standardised preprocessing pipeline that includes the removal of HTML tags, URLs, hashtags, user mentions (@names), punctuation, non-ASCII characters, digits and redundant whitespace.

The fourth dataset, **ConvAbuse**, differs from the others as it is not sourced from social media but consists of English dialogues between users and two conversational agents. We include it to assess the model's performance on a different text genre: conversational dialogue. The original annotations span five levels of abuse severity, from -3 (highly abusive) to 1 (non-abusive). We simplify this into a binary classification task, labelling utterances with severity < 0 as offensive and those with severity ≥ 0 as non-offensive. For processing, we concatenate each dialogue into a single text sequence.

B Hyperparameter Tuning for BERT

We supervise each weak predictor using a joint objective function as in Eq. (8) combining: (a) the F1 micro score computed with hard labels (F1), (b) the cross-entropy loss with soft target distributions (CE), and (c) the average Manhattan distance (MD). For hyperparameter tuning of a single BERT model, we fixed all loss coefficients and regularisation weight to 1 to simplify the optimisation landscape. To ensure consistency and fair comparison across datasets, we use the ConvAbuse dataset, which is moderate in size relative to the others, to tune hyperparameters for fine-tuning BERT. Hyperparameter optimisation is performed using OPTUNA (Akiba et al., 2019). The model is trained on the training set and validated on the development set, while the test set remains completely unseen during both training and tuning.

The search is performed over 10 trials. The Weight decay is held constant at 0.01 to enforce moderate parameter shrinkage, preventing overfitting while avoiding excessive bias in the learned weights. The following hyperparameters are optimised: the learning rate (lr, sampled logarithmically in the range $[10^{-6}, 10^{-4}]$), number of training epochs (n_ep $\in [2, 5]$), batch size (bs $\in \{4, 8, 16, 32, 64\}$), and the number of warm-up steps (w_steps $\in [1, 500]$). The detailed results are shown in Table 10. These optimal settings are then applied uniformly across all datasets to ensure a consistent training setup.

Table 10: Hyperparameter tuning results sorted by joint loss (ascending). Pruned trials (6 and 9) are excluded. Bold values indicate best performance per column: highest F1, and lowest CE, MD and Joint loss.

Trial	lr	n_ep	bs	w_steps	F1	CE	MD	Joint
2	4.56e-6	2	32	463	0.8362	0.9678	**1.1516**	**1.2832**
0	1.48e-6	2	16	41	0.8362	0.9718	1.2891	1.4246
7	2.65e-6	3	16	342	0.8362	0.9753	1.2992	1.4383
3	4.75e-6	5	64	221	0.8818	0.8161	1.5719	1.5062
4	7.24e-5	5	8	487	**0.9470**	**0.8082**	2.0893	1.9504
1	9.80e-5	3	32	12	0.9434	0.8617	3.2045	3.1228
5	9.57e-5	5	32	150	0.9360	0.8409	3.5261	3.4310
8	4.35e-5	5	4	94	0.9200	0.8737	4.8390	4.7928

C Best Parameters for WEL

We perform a grid search over four parameters in the objective function, each sampled from the range $[0, 0.001, 0.01, 0.1, 1]$, resulting in 1,295 unique combinations per dataset (excluding 0s for all). Table 11 reports the best-performing parameter configurations (α, β, γ and λ) for the two variants of our proposed method: WEL-Random and WEL-TopAnn. The optimal values vary across datasets and selection strategies, indicating that performance is sensitive to the interplay of loss components. Notably, there is no consistent trend suggesting that any single parameter dominates

Table 11: Best performing hyperparameters for WEL.

	Method	α	β	γ	λ
ArMIS	WEL-Random	1	0.0001	0.01	0.001
	WEL-TopAnn	0.001	0.0001	0.1	0
ConvAbuse	WEL-Random	0	0.1	1	0
	WEL-TopAnn	0	1	0.01	0.01
HS-Brexit	WEL-Random	1	0.001	0	0.001
	WEL-TopAnn	0.1	0.1	0	0.001
MD-Agreement	WEL-Random	0.001	0.0001	0	0.001
	WEL-TopAnn	1	0	0	0

[3] https://X.com/

Table 12: Ablation study of loss optimisation paradigms on ConvAbuse dataset.

Case	F1	CE	MD
$\mathcal{L}_{\text{F1}}$ only	0.9298	0.5573	0.1862
$\mathcal{L}_{\text{CE}}$ only	0.9286	**0.5536**	0.1801
$\mathcal{L}_{\text{MD}}$ only	0.9202	0.5737	**0.1680**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	**0.9333**	0.5540	0.1777
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	**0.9333**	0.5575	0.1689
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.9298	0.5556	0.1755
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	**0.9333**	0.5571	0.1707

Table 13: Ablation study of loss optimisation paradigms on HS-Brexit dataset.

Case	F1	CE	MD
$\mathcal{L}_{\text{F1}}$ only	0.8988	0.5868	0.2636
$\mathcal{L}_{\text{CE}}$ only	**0.9048**	**0.5851**	0.2859
$\mathcal{L}_{\text{MD}}$ only	0.8750	0.6066	0.2325
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	**0.9048**	0.5857	0.2789
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	0.8810	0.6182	0.2374
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.8929	0.6035	**0.2306**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.8690	0.6093	0.2342

performance across settings.

D Ensemble Optimisation Paradigms on ConvAbuse and HS-Brexit

Tables12 and 13 present ablation results for ConvAbuse and HS-Brexit using WEL-Random. Across both datasets, multiple configurations achieve similar scores, suggesting that when loss coefficients are fixed, different objectives can lead to comparable outcomes.

For **ConvAbuse**, combining $\mathcal{L}_{\text{F1}}$ with either $\mathcal{L}_{\text{CE}}$ or $\mathcal{L}_{\text{MD}}$ yields the highest F1 (0.9333), while $\mathcal{L}_{\text{MD}}$ alone achieves the lowest MD (0.1680). For **HS-Brexit**, $\mathcal{L}_{\text{CE}}$ alone gives the highest F1 (0.9048), and the $\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$ pairing yields the lowest MD (0.2306). Including all three losses does not consistently improve results and can slightly reduce F1, likely due to competing objectives without tuned coefficients.

Overall, these results indicate that when coefficients are fixed, several loss configurations can perform similarly, and gains from specific combinations are modest. The impact of loss balancing is explored further in the next subsection on parameter correlations.

E Parameter Impact on ConvAbuse and HS-Brexit

Table 14 illustrates how each control parameter balances the multi-objective trade-offs in the joint optimisation (Eq. (8)). The regularisation term λ demonstrates consistently strong performance on both ConvAbuse and HS-Brexit, achieving near-perfect correlations with F1 (+1.0*/+0.99*) and CE (-1.0*), though its effect on MD diverges from patterns observed on MD-Agreement (Table 7).

In contrast, γ consistently improves MD (-1.0*) but harms F1 (-0.9*/-1.0*), making it better suited for MD-focused objectives. The effect of α varies: it improves CE and MD on ConvAbuse but degrades them on HS-Brexit. Finally, β reliably improves CE (-1.0*) on both datasets, but at the cost of worse MD (+0.7/+0.9*). These differences likely reflect the distinct text genres and annotation distributions of the datasets, underscoring the need for task-specific parameter tuning.

These results align closely with MD-Agreement findings in the main paper.

Table 14: Correlation between parameter and evaluation metrics (F1, CE and MD) on the ConvAbuse and HS-Brexit datasets using WEL-Random with BERT.

Dataset	ConvAbuse			HS-Brexit		
Param	F1	CE	MD	F1	CE	MD
α	-0.21	-0.4	-1.0*	+0.82	+1.0*	+0.1
β	+0.8	-1.0*	+0.7	+0.6	-1.0*	+0.9*
γ	-0.9*	+1.0*	-1.0*	-1.0*	+1.0*	-1.0*
λ	+1.0*	-1.0*	+0.9*	+1.0*	-1.0*	+1.0*
M	+1.0*	-1.0*	+0.98*	+0.99*	-1.0*	+0.55