Annales Mathematicae et Informaticae
63 (2026) pp. 18–31
DOI: 10.33039/ami.2026.06.003
URL: https://ami.uni-eszterhazy.hu

Adaptive sentiment evaluation in social media analysis

Anikó Apró[a b], **Levente Sasi**[c]

[a]University of Debrecen, Doctoral School of Informatics,
apro.aniko@inf.unideb.hu

[b]Eszterházy Károly Catholic University
apro.aniko@uni-eszterhazy.hu

[c]Eszterházy Károly Catholic University
sasi.levente@gmail.com

Abstract. Social media sentiment analysis faces a persistent aggregation problem: lexicon-based and transformer-based models often produce inconsistent outputs for the same short, informal, and stylistically heterogeneous texts. This paper introduces ADRTW (Adaptive Dynamic Reliability-Triggered Weighting), an interpretable sentiment fusion framework that combines heterogeneous sentiment estimators using rule-guided reliability weights derived from textual cues, inter-model disagreement, and consistency patterns [5, 8].

The framework is evaluated on a Reddit dataset containing 1,577 posts, 354,050 comments, and 187,666 authors collected between 2017 and 2025, together with a controlled synthetic benchmark for aggregation comparison. The results show that ADRTW remains competitive with static averaging in controlled settings while preserving context-sensitive local variation in large-scale discourse analysis.

Beyond sentiment fusion, the ADRTW-derived signal supports complementary analyses of online discussions, including temporal trend inspection, toxicity-aware interpretation, and participation-based clustering. Overall, the proposed framework provides a transparent and reusable basis for examining emotional dynamics in social media discourse.

Keywords: sentiment analysis, social media, Reddit, adaptive weighting, toxicity detection, clustering

Submitted: March 31, 2026
Accepted: June 30, 2026
Published online: July 17, 2026

1. Introduction

The rapid growth of social media has resulted in an unprecedented volume of user-generated text. Platforms such as Reddit provide rich sources of opinions, discussions, and emotional expressions across diverse topics. Sentiment analysis has become a fundamental tool for extracting meaningful insights from such data [11, 18].

However, existing sentiment analysis approaches often rely on individual models, such as lexicon-based or deep learning-based methods. These models exhibit varying performance depending on linguistic context, domain, and textual characteristics. As a result, their outputs may differ significantly when applied to the same input. Benchmark studies show that model choice strongly affects sentiment estimation quality, especially in short, informal texts [11, 13, 18].

This inconsistency challenges reliable sentiment interpretation, especially in dynamic environments such as Reddit discussions.

To address this limitation, this paper proposes an adaptive sentiment evaluation framework called ADRTW (Adaptive Dynamic Reliability-Triggered Weighting). The method dynamically combines multiple sentiment models by assigning context-dependent reliability weights.

The main contribution of this paper is not a new sentiment classifier, but an interpretable adaptive fusion mechanism that reweights heterogeneous sentiment signals according to contextual reliability cues. In contrast to static averaging, ADRTW adjusts model influence using disagreement patterns, surface-level text characteristics, and inter-model consistency cues. This makes the final sentiment signal more suitable for discourse-level analysis in heterogeneous social media environments.

The contributions of this paper are as follows:

- We propose ADRTW, an interpretable adaptive late-fusion mechanism for combining lexicon-based and transformer-based sentiment estimators in short social media texts.
- We provide an explicit rule-guided weighting formulation based on contextual text features, inter-model agreement, and consistency penalties.
- We evaluate ADRTW against individual base models and static aggregation on a controlled benchmark, and we examine its large-scale behavior on Reddit through discourse-level and alignment-based analyses.
- We show that ADRTW-derived sentiment can also serve as a useful discourse-level signal in complementary analyses such as temporal trend inspection, toxicity-aware interpretation, and participation clustering.

Thus, ADRTW is positioned as a transparent late-fusion strategy rather than as a replacement for existing sentiment classifiers, adaptively reweighting heterogeneous sentiment signals using textual and inter-model consistency cues.

2. Related work

Sentiment analysis methods can be broadly categorized into lexicon-based and machine learning-based approaches [11, 18]. Lexicon-based methods, such as VADER and TextBlob, rely on predefined sentiment dictionaries and rule-based heuristics [2, 6]. These approaches are computationally efficient but often struggle with context-dependent language, sarcasm, and domain-specific expressions.

In contrast, transformer-based models such as RoBERTa and related BERT-style architectures provide context-aware representations and often achieve stronger performance in complex linguistic scenarios. Related Hungarian NLP results also suggest that monolingual transformer models may outperform multilingual alternatives, although such models remain computationally expensive and may still exhibit inconsistencies across domains [3, 9, 16].

Several studies have explored ensemble approaches that combine multiple sentiment models [8, 12]. While these methods improve robustness, they typically rely on static weighting schemes, which do not adapt to the characteristics of individual texts.

The present work addresses this limitation by introducing a dynamic, context-sensitive weighting mechanism for fusing heterogeneous sentiment signals.

Recent work has also emphasized that social media analysis should not be reduced to polarity estimation alone. Toxicity detection, discourse quality assessment, and behavior-aware clustering provide complementary perspectives for understanding online communities, especially in environments where emotional intensity, disagreement, and participation patterns interact [10, 15, 17].

Taken together, these studies suggest that the main unresolved challenge is not merely sentiment classification accuracy, but the context-sensitive fusion of heterogeneous sentiment signals in socially complex environments. This is the specific problem addressed by ADRTW in the present work.

Although prior studies have explored sentiment ensembles, most of them optimize predictive robustness rather than interpretable context-sensitive fusion at the individual-text level. The present work addresses this gap by proposing a rule-guided adaptive weighting mechanism designed for heterogeneous social media discourse.

3. Dataset and data collection

Data were collected through the official Reddit API (PRAW) from multiple topic-oriented communities, including *love*, *politics*, *gaming*, *war*, *technology*, *nature*, *science*, *health*, *education*, *climate*, *family*, and *religion*. The resulting dataset consists of 1,577 posts, 354,050 comments, and 187,666 authors collected between 2017 and 2025. Each record includes textual content, metadata, and relational links between posts and comments.

Preprocessing included text normalization, noise removal, and filtering of non-English content. To reduce topical bias, the selected subreddits were intention-

ally balanced to include both emotionally positive (e.g., *love*, *family*) and conflict-oriented contexts (e.g., *war*, *politics*).

Data collection was performed using multiple listing strategies (*hot*, *rising*, *new*, and *top*) to capture diverse engagement patterns. Duplicate handling was implemented based on post identifiers and comment-level matching (post ID, text content, timestamp). Deleted and empty content were excluded, and non-English texts were filtered using rule-based language detection.

The data were stored in a structured SQLite database preserving post–comment relationships and author-level aggregation statistics. This sampling strategy may introduce topical and community-specific biases, which should be considered when interpreting the results.

Only publicly available Reddit content was processed, and the analysis was conducted at aggregate level without attempting to identify individual users.

4. Methodology

The sentiment analysis pipeline includes data preprocessing, model evaluation, and aggregation.

Three sentiment analysis models were applied to each text: VADER, TextBlob, and a RoBERTa-based sentiment model fine-tuned for social media text.

VADER and TextBlob represent lexicon-based sentiment estimation, while the transformer-based component relies on a RoBERTa-style contextual language model [2, 3, 6, 9].

In addition to simple averaging and fixed-weight aggregation, we examine the proposed ADRTW framework, which dynamically adjusts model weights based on reliability indicators.

4.1. Operational definition of ADRTW

Let x denote an input text and let $s_i(x) \in [-1, 1]$ be the sentiment score produced by model $i \in \{1, \dots, N\}$. In this work, $N = 3$ and the models are VADER, TextBlob, and a RoBERTa-based transformer model. The final ADRTW score is defined as

$$S_{\mathrm{ADRTW}}(x) = \sum_{i=1}^{N} w_i(x) s_i(x), \qquad \sum_{i=1}^{N} w_i(x) = 1, \quad w_i(x) \geq 0.$$

The weight computation consists of three steps: context-sensitive base weighting, consistency correction, and normalization [12].

Step 1: Context-sensitive base weighting. For each text, we extract a small set of surface features:

$$L(x) = |x|, \quad d_{\mathrm{emoji}}(x) = \frac{n_{\mathrm{emoji}}(x)}{\max(|x|, 1)}, \quad n_{\mathrm{sent}}(x) = \text{sentence-fragment count}.$$

Each model receives a prior π_i and a compatibility score $g_i(x)$:

$$w_i^{(\text{base})}(x) = \pi_i \cdot g_i(x).$$

The compatibility score is defined heuristically:

$$g_i(x) = \begin{cases} 1 + \lambda_{\text{short}} & \text{for lexicon models if } L(x) < \tau_L, \\ 1 + \lambda_{\text{emoji}} & \text{for VADER if } d_{\text{emoji}}(x) > \tau_E, \\ 1 + \lambda_{\text{long}} & \text{for RoBERTa if } L(x) \geq \tau_L \text{ and } n_{\text{sent}}(x) \geq \tau_S, \\ 1 & \text{otherwise.} \end{cases}$$

In practice, short and emoji-rich texts favor lexicon-based models, while longer and structurally richer texts increase the influence of the transformer-based model.

Step 2: Consistency correction. Let

$$m(x) = \text{median}\{s_1(x), \ldots, s_N(x)\}$$

and

$$\delta_i(x) = |s_i(x) - m(x)|.$$

We define the global disagreement level as

$$D(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_i(x).$$

The adaptive sensitivity parameter is

$$\alpha_{\text{eff}}(x) = \alpha_0 + \alpha_1 D(x),$$

where $\alpha_0 > 0$ and $\alpha_1 \geq 0$ control the base and disagreement-dependent penalty strength.

The consistency term is then

$$w_i^{(\text{cons})}(x) = \exp\big(-\alpha_{\text{eff}}(x)\delta_i(x)\big).$$

If $\delta_i(x) > \tau_{\text{out}}$, an additional outlier penalty $\rho_{\text{out}} \in (0, 1)$ is applied. If the polarity sign of $s_i(x)$ contradicts a sufficiently strong consensus ($|m(x)| > \tau_{\text{sign}}$), a contradiction penalty $\rho_{\text{sign}} \in (0, 1)$ is introduced. These corrections reduce the influence of unstable or conflicting model predictions.

Step 3: Final weight and normalization. The unnormalized weight is

$$\widetilde{w}_i(x) = w_i^{(\text{base})}(x) \cdot \big(w_i^{(\text{cons})}(x)\big)^{\gamma},$$

where $\gamma > 0$ controls the strength of the consistency correction. The final normalized weight is

$$w_i(x) = \frac{\widetilde{w}_i(x)}{\sum_{j=1}^{N} \widetilde{w}_j(x)}.$$

The ADRTW score is thus obtained by combining transparent priors, text-dependent compatibility, and disagreement-aware consistency correction. Since all components are explicit, the method remains interpretable and can be reproduced by following the described weighting procedure and parameter settings.

Parameterization and practical behavior. The ADRTW weighting mechanism is rule-guided and interpretable rather than end-to-end learned. Its parameters control the relative influence of surface-level textual cues, model priors, and inter-model consistency penalties. Short and emoji-rich texts increase the relative influence of lexicon-based estimators, whereas longer and structurally richer texts increase the relative influence of the transformer-based component. Two key hyperparameters were selected on the synthetic benchmark: the consistency correction strength and the prior weight assigned to the RoBERTa-based estimator. After this tuning step, the selected parameters were kept fixed for all Reddit-based analyses. Consequently, the Reddit corpus was not used for supervised parameter optimization, but only for exploratory discourse-level analysis.

5. Experiments and results

The primary focus of the evaluation is the ADRTW aggregation mechanism. The toxicity and clustering analyses are included as complementary interpretive layers demonstrating how ADRTW-derived sentiment signals can support higher-level discourse interpretation.

5.1. Experimental setup

The empirical evaluation was conducted in two settings: a controlled synthetic benchmark for aggregation comparison and the full Reddit dataset for discourse-level analysis. Temporal sentiment trajectories were smoothed using a 60-day moving average. Toxicity detection was performed using a TF–IDF + logistic regression baseline with a rule-augmented extension, and clustering was applied to interaction and sentiment features with silhouette-based evaluation and PCA visualization.

5.2. Synthetic benchmark construction

To provide a controlled reference setting for evaluating aggregation behavior, we used a synthetic benchmark designed to simulate sentiment estimation inconsistencies commonly observed in short social media texts. The benchmark consists of 500 sentiment instances with reference polarity scores distributed across $[-1, 1]$,

covering negative, neutral, positive, borderline, and ambiguous sentiment conditions.

The reference sentiment score of each synthetic instance was treated as the controlled ground-truth value. The constructed examples were designed to include short, stylistically marked, and potentially ambiguous cases in which lexicon-based and transformer-based models may react differently. To emulate heterogeneous model behavior, controlled perturbations were introduced into the sentiment estimates, including polarity shifts, disagreement amplification, and context-sensitive noise.

The actual sentiment estimators were then applied to the synthetic text instances, allowing the benchmark to compare VADER, TextBlob, RoBERTa, static averaging, and ADRTW against the same controlled reference labels. The purpose of this benchmark is not to replace manually annotated social media data, but to isolate the aggregation problem under controlled and reproducible disagreement conditions.

5.3. Controlled benchmark evaluation

Hyperparameter tuning. Two ADRTW hyperparameters were tuned on the synthetic benchmark using Gaussian process-based Bayesian optimization. The optimized parameters were the consistency correction strength and the prior weight assigned to the RoBERTa-based sentiment estimator. The objective function minimized the mean absolute error (MAE) between the ADRTW score and the synthetic reference sentiment labels. The search space was defined as $[0.2, 1.0]$ for the consistency strength and $[1.0, 1.3]$ for the RoBERTa prior. The optimization was run for 30 function evaluations with 8 initial random points and a fixed random seed for reproducibility. The selected parameters were then kept fixed for all subsequent Reddit-based analyses.

Table 1 reports the performance of the individual sentiment models, the static averaging baseline, and ADRTW on the synthetic benchmark after Bayesian optimization of the ADRTW hyperparameters.

On the synthetic benchmark, ADRTW achieves the lowest MAE among all compared methods. Relative to static averaging, the MAE decreases from 0.2122 to 0.2003, corresponding to an improvement of approximately 5.6%. ADRTW also yields the highest correlation with the benchmark labels (0.9201), slightly exceeding both the strongest individual base model (RoBERTa: 0.9105) and the static averaging baseline (0.9176).

The MSE of ADRTW is marginally higher than that of static averaging (0.0633 vs. 0.0618). This suggests that the proposed weighting scheme is slightly more responsive to disagreement between models and therefore less conservative than simple averaging in some cases. In the context of this paper, this behavior is acceptable because the goal of ADRTW is not only error minimization, but also the preservation of context-sensitive variation in aggregated sentiment.

Taken together, the synthetic benchmark indicates that ADRTW remains competitive with static averaging in overall predictive quality while offering a more

adaptive aggregation behavior.

Table 1. Benchmark results on the synthetic sentiment dataset ($n = 500$). MAE = mean absolute error; MSE = mean squared error; Corr = Pearson correlation with ground-truth labels. ADRTW hyperparameters selected by Bayesian optimization within the predefined search space: consistency strength = 0.200, RoBERTa prior = 1.300.

Model / Aggregation	**MAE↓**	**MSE↓**	**Corr↑**
VADER	0.3708	0.2344	0.7311
TextBlob	0.3394	0.1636	0.7698
RoBERTa	0.3086	0.1436	0.9105
Static avg (avg3)	0.2122	**0.0618**	0.9176
ADRTW	**0.2003**	0.0633	**0.9201**

Alignment analysis on the Reddit corpus. Since no manually annotated sentiment ground truth was available for the Reddit corpus, we do not interpret the large-scale Reddit evaluation as supervised performance validation. Instead, we report an auxiliary alignment analysis relative to the transformer-based model, which showed the strongest individual benchmark performance among the base estimators.

On comments, ADRTW reduces the mean absolute deviation from the RoBERTa scores from 0.315 (static averaging) to 0.261, while increasing the correlation from 0.848 to 0.894. On posts, the corresponding values improve from 0.258 to 0.202 for mean absolute deviation and from 0.848 to 0.915 for correlation.

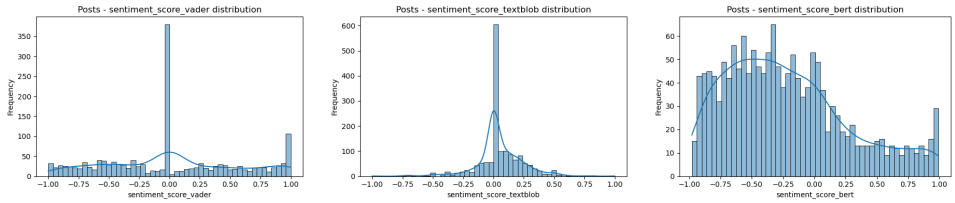
These results do not establish superiority with respect to ground truth on the Reddit corpus; rather, they indicate that ADRTW produces a more coherent aggregation relative to the strongest individual context-aware component while still retaining the contribution of the lexicon-based models.

5.4. Distributional analysis of sentiment estimators

We examine the distributional characteristics of the individual models and the ADRTW aggregation.

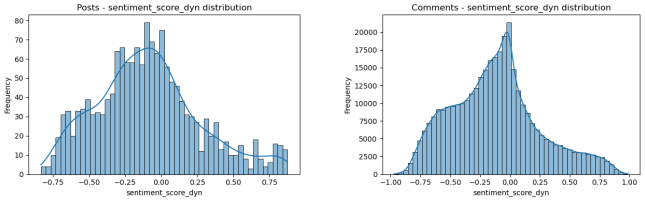
Figure 1 highlights systematic differences between the underlying sentiment estimators. The lexicon-based models (VADER and TextBlob) tend to produce distributions centered closer to neutral or slightly positive values, whereas the transformer-based model exhibits a noticeable shift toward negative sentiment. This suggests that ADRTW reduces the dominance of divergent individual outputs while remaining more aligned with the strongest context-aware component.

The ADRTW aggregation produces a more expressive and context-sensitive sentiment representation at both post and comment level (Figure 2). The resulting



(a) VADER **(b)** TextBlob **(c)** RoBERTa

Figure 1. Distribution of sentiment scores produced by the individual sentiment models on the dataset.



(a) Posts **(b)** Comments

Figure 2. Distribution of ADRTW sentiment scores at post and comment level. The aggregated scores remain expressive while reducing the dominance of any single underlying model.

distributions preserve meaningful variation while reducing the dominance of individual models and mitigating systematic biases. This indicates adaptive integration rather than simple arithmetic smoothing.

Table 2 further summarizes the distributional properties of the individual models and aggregation methods on the full Reddit comment corpus.

Table 2. Distributional statistics of sentiment estimators on the full Reddit comment corpus ($n = 354{,}050$).

Statistic	**VADER**	**TextBlob**	**RoBERTa**	**Avg**	**ADRTW**
Mean	+0.039	+0.059	−0.314	−0.072	−0.119
Std	0.509	0.264	0.517	0.348	0.369
Median	0.000	0.000	−0.417	−0.075	−0.124
Q1	−0.340	−0.007	−0.743	−0.321	−0.384
Q3	+0.440	+0.182	−0.028	+0.139	+0.079

The statistical results confirm the visual observations. While the lexicon-based models exhibit slightly positive central tendencies, the transformer-based model is shifted toward negative values. ADRTW remains closer to the transformer-based signal while preserving contributions from the lexicon-based models, resulting in a

more balanced aggregated distribution.

5.5. Temporal dynamics and collective rebalancing

Figure 3 presents the smoothed sentiment trajectories (60-day moving average) over time. While both methods follow similar global trends, ADRTW exhibits sharper local variations, indicating higher sensitivity to contextual changes.

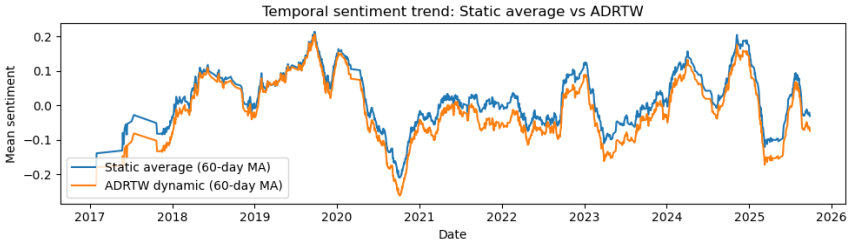


Figure 3. Temporal comparison of static averaging and ADRTW-based sentiment using moving averages. While both methods follow similar global trends, ADRTW reacts more sensitively to local contextual changes.

This suggests that ADRTW preserves context-sensitive variation rather than acting as a simple smoothing mechanism.

Figure 4 shows the temporal evolution of ADRTW sentiment at post and comment level. Posts tend to exhibit stronger emotional amplification, whereas comments display a moderating effect over time. This suggests that collective discussion may partially rebalance initially polarized emotional responses.

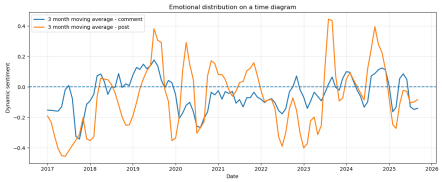


Figure 4. Temporal evolution of post- and comment-level ADRTW sentiment using moving averages. Posts often show stronger emotional amplification, whereas comments tend to moderate the discourse over time.

5.6. Toxicity as a complementary behavioral signal

Negative sentiment and toxic discourse are not equivalent phenomena. A comment may express dissatisfaction, criticism, or emotional tension without targeting another participant in a hostile or abusive manner [4, 10, 15]. To evaluate whether

ADRTW-based sentiment remains meaningful beyond polarity estimation alone, we include toxicity analysis as a complementary discourse-quality layer. Toxicity is not treated as an alternative sentiment model, but as an orthogonal signal that helps distinguish emotional negativity from socially harmful interaction.

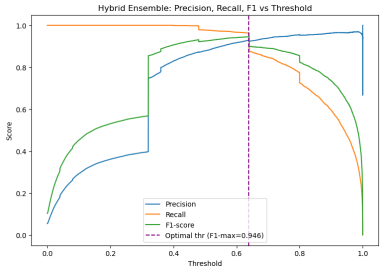


Figure 5. Precision, recall, and F1-score of the hybrid toxicity model across classification thresholds.

The toxicity baseline model represents comments using TF–IDF features and applies logistic regression to estimate the probability of toxic discourse. A hybrid ensemble extends this baseline with a rule-based component tailored to community-platform language patterns, including explicitly insulting, threatening, or personalized hostile expressions. The final toxicity score is adjusted by combining statistical prediction with rule-based signals, allowing the system to remain sensitive to both lexical regularities and explicit toxic markers.

The evaluation results indicate that the hybrid model does not primarily improve the overall F1-score, but rather shifts the precision–recall balance toward higher sensitivity. In toxicity-related classification tasks, such threshold-dependent trade-offs are especially important because aggregate metrics alone may hide practically relevant differences in model behavior [1]. This supports the methodological claim that toxicity should be modeled as a distinct discourse-quality dimension rather than inferred directly from negative sentiment alone [4, 15]. The threshold-dependent trade-off between precision, recall, and F1-score is illustrated in Figure 5.

5.7. Clustering analysis and participation archetypes

We further examined whether ADRTW-based sentiment supports higher-level discourse interpretation through clustering of posts and comments. Here, clustering serves not as a predictive task but as an interpretive layer to identify recurring participation patterns and behavioral archetypes.

For posts, clusters were formed using interaction and sentiment-related features (e.g., score, comment score, aggregated sentiment), while for comments we used like score, text length, and sentiment. Although smaller cluster numbers yielded higher silhouette scores, a four-cluster solution proved more informative for interpretation. Silhouette analysis assessed cluster quality [14], while PCA was used only for visualization [7].

The resulting clusters suggest four recurring roles: critical/debating participants, constructive contributors, short positive acknowledgers, and low-visibility passive participants. Rather than simply summarizing engagement, clustering reveals how emotional polarity, cognitive effort, and interaction patterns jointly structure community discourse.

Overall, the findings support the view that large-scale social media discussions are not emotionally uniform, but organized into recurring activity profiles and topic-dependent emotional dynamics.

6. Discussion and limitations

The joint analysis of sentiment, toxicity, and clustering suggests that social media interaction is better understood as a layered process rather than a simple positive–negative continuum. Short, highly polarized reactions often reflect immediate affective responses, whereas longer comments tend to be less extreme and more argumentative. This indicates that discourse structure is shaped not only by emotional polarity, but also by cognitive effort and interactional intent.

The synthetic benchmark and the Reddit corpus serve different evaluative purposes. The benchmark provides a controlled setting for comparing aggregation behavior against known reference scores, whereas the Reddit corpus demonstrates the applicability of the resulting ADRTW signal in large-scale exploratory discourse analysis. Therefore, conclusions about predictive accuracy are drawn only from the benchmark, while Reddit-based findings are interpreted in terms of coherence, alignment, and interpretability rather than supervised sentiment validity.

The study has several limitations. First, the real-world analysis is restricted to Reddit, so the observed weighting behavior may partly reflect platform-specific discourse patterns. Second, the transformer component was originally fine-tuned on social media text from another platform, which may introduce domain-transfer bias. Third, the ADRTW weighting mechanism is rule-guided and interpretable rather than end-to-end learned; this improves transparency but may limit cross-domain adaptability. Finally, the toxicity and clustering analyses should be interpreted as complementary interpretive layers rather than universal downstream guarantees.

7. Conclusion and future work

This paper introduced ADRTW, an interpretable and context-sensitive sentiment aggregation framework for heterogeneous social media text. Instead of replacing existing sentiment models, ADRTW combines them through adaptive reliability weighting based on textual cues and inter-model consistency. The results show that ADRTW remains competitive with static averaging on a controlled benchmark while providing a flexible signal for large-scale discourse-level analysis.

Future work will focus on author-aware and predictive analysis, including behavioral profiling, recurring user roles, and early ADRTW-based signals for modeling

later engagement trajectories. Further evaluation on manually annotated datasets and platforms beyond Reddit is also needed.

References

[1] D. Borkan, L. Dixon, J. Sorensen, N. Thain, L. Vasserman: *Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification*, Proceedings of the 2019 World Wide Web Conference (2019), pp. 491–500, doi: 10.1145/3308560.3317593.

[2] T. De Smedt, W. Daelemans: *Pattern for Python*, Journal of Machine Learning Research 13 (2012), pp. 2063–2067.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: *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, 2019, doi: 10.18653/v1/N19-1423.

[4] P. Fortuna, S. Nunes: *A Survey on Automatic Detection of Hate Speech in Text*, ACM Computing Surveys 51.4 (2018), 85:1–85:30, doi: 10.1145/3232676.

[5] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger: *On Calibration of Modern Neural Networks*, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.

[6] C. J. Hutto, E. Gilbert: *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*, in: Proceedings of the International AAAI Conference on Web and Social Media, 2014, doi: 10.1609/icwsm.v8i1.14550.

[7] I. T. Jolliffe, J. Cadima: *Principal Component Analysis: A Review and Recent Developments*, Philosophical Transactions of the Royal Society A 374.2065 (2016), p. 20150202, doi: 10.1098/rsta.2015.0202.

[8] B. Lakshminarayanan, A. Pritzel, C. Blundell: *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 6402–6413.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, arXiv preprint arXiv:1907.11692 (2019), arXiv: 1907.11692.

[10] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder: *Hate Speech Detection: Challenges and Solutions*, PLoS ONE 14.8 (2019), e0221152, doi: 10.1371/journal.pone.0221152.

[11] W. Medhat, A. Hassan, H. Korashy: *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal 5.4 (2014), pp. 1093–1113, doi: 10.1016/j.asej.2014.04.011.

[12] R. Polikar: *Ensemble based systems in decision making*, IEEE Circuits and Systems Magazine 6.3 (2006), pp. 21–45, doi: 10.1109/MCAS.2006.1688199.

[13] S. Rosenthal, N. Farra, P. Nakov: *SemEval-2017 Task 4: Sentiment Analysis in Twitter*, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, 2017, doi: 10.18653/v1/S17-2088, url: https://aclanthology.org/S17-2088/.

[14] P. J. Rousseeuw: *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*, Journal of Computational and Applied Mathematics 20 (1987), pp. 53–65, doi: 10.1016/0377-0427(87)90125-7.

[15] Z. Waseem, T. Davidson, D. Warmsley, I. Weber: *Understanding Abuse: A Typology of Abusive Language Detection Subtasks*, in: Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 78–84, doi: 10.18653/v1/W17-3012.

[16] Z. G. Yang, Á. Agócs, G. Kusper, T. Váradi: *Abstractive text summarization for Hungarian*, Annales Mathematicae et Informaticae 53 (2021), pp. 299–316, doi: 10.33039/ami.2021.04.002.

[17] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis: *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media*, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1425–1447, doi: 10.18653/v1/2020.semeval-1.188, url: https://aclanthology.org/2020.semeval-1.188/.

[18] L. Zhang, S. Wang, B. Liu: *Deep Learning for Sentiment Analysis: A Survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2018), doi: 10.1002/widm.1253.