SINAI-UGPLN at CheckThat! 2025: Meta-Ensemble Strategies for Numerical Claim Verification in English*

Notebook for the CheckThat! Lab at CLEF 2025

Mariuxi del Carmen Toapanta-Bernabé[1,2,*,†], Miguel Ángel García-Cumbreras[1], Luis Alfonso Ureña-López[1], Denisse Desiree Mora-Intriago[2] and Carla Tatiana Bernal-García[2,†]

[1]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Jaén, Spain
[2]Universidad de Guayaquil, 090514, Guayas, Ecuador

Abstract

Automated verification of numerical claims is essential to support decision-making in high-stakes domains such as finance and public health, yet subtle quantitative inconsistencies can completely invert a claim's truth value. Motivated by the prevalence of such numeric errors in real-world texts, we present a cohesive meta-ensemble framework for numerical claim verification in English, evaluated on CLEF 2025 CheckThat! Lab Task 3. Our approach integrates: (1) threshold-optimized RoBERTa classifiers for true/false discrimination; (2) a dedicated RoBERTa-based binary detector for conflicting claims; and (3) a logistic-regression meta-classifier that fuses soft and hard predictions alongside lightweight numeric-difference and lexical-cue features. We implement three variants—threshold-tuned single model, majority-voting ensemble, and batch-encoded meta-classifier—and evaluate them on the official Dev and DevTest splits. On DevTest, our submitted meta-classifier achieves a Macro-F1 of **0.4553**, ranking **8th** among participants. Key contributions include a post-hoc label-mapping ensemble; a high-recall conflicting detector calibrated via Dev grid search ($t_{\text{true}} = 0.420$, $t_{\text{false}} = 0.460$ yielding Dev Macro-F1=0.5936); and integration of numeric and lexical feature augmentations. An extensive error analysis guides future work on adaptive calibration and numeric reasoning modules.

Keywords

Numerical claim verification, meta-ensemble, threshold tuning, imbalanced classification, RoBERTa, conflicting detection

1. Introduction

Automated verification of numerical claims is essential to support decision-making in high-stakes domains such as finance and public health, yet subtle quantitative inconsistencies can completely invert a claim's truth value. Motivated by the prevalence of such numeric errors in real-world texts, the CLEF 2025 CheckThat! Lab Task 3 benchmarks methods for numerical claim verification in English, requiring classification of each claim as *true*, *false*, or *conflicting* based on retrieved evidence [1, 2, 3, 4]. Challenges include pronounced class imbalance (true 40%, false 35%, conflicting 25%) and the need for fine-grained numeric reasoning.

To address these challenges, we design a three-stage meta-ensemble pipeline that: (1) isolates ambiguous or contradictory statements via a binary RoBERTa detector fine-tuned to maximize recall on the underrepresented *conflicting* class; (2) applies threshold-tuned RoBERTa classifiers for *true*/*false* discrimination, optimizing per-class decision boundaries on the Dev split to balance precision and recall; and (3) fuses all softmax probabilities and hard labels through a batch-encoded logistic meta-classifier,

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
mctb0005@red.ujaen.es;mariuxi.toapantab@ug.edu.ec (M. d. C. Toapanta-Bernabé); magc@ujaen.es (M. García-Cumbreras); laurena@ujaen.es (L. A. Ureña-López); denisse.morai@ug.edu.ec (D. D. Mora-Intriago); carla.bernalg@ug.edu.ec (C. T. Bernal-García)
0000-0002-4839-7452 (M. d. C. Toapanta-Bernabé); 0000-0003-1867-9587 (M. García-Cumbreras); 0000-0001-7540-4059 (L. A. Ureña-López); 0009-0005-9840-244X (D. D. Mora-Intriago); 0009-0005-9870-1589 (C. T. Bernal-García)

augmented with compact numeric-difference and lexical-cue features to capture subtle quantitative and linguistic cues.

Our Contributions

1. A post-hoc label-mapping ensemble that balances true/false performance.
2. A high-recall binary conflicting detector calibrated via Dev grid search ($t_{\text{true}} = 0.420$, $t_{\text{false}} = 0.460$).
3. Integration of numeric-difference and lexical-cue features to enrich transformer embeddings.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the task and dataset splits. Section 4 describes our proposed system. Section 5 presents the experimental setup and results. Section 6 provides error analysis and discussion. Finally, Section 7 concludes with future research directions.

2. Related Work

Previous editions of the CheckThat! Lab have laid the groundwork for automated claim verification by introducing a suite of subtasks and evaluation protocols. The 2023 lab offered five tasks—check-worthiness, subjectivity, political bias, factuality, and authority finding—across seven languages, establishing baselines with transformer and gradient-boosting models [5, 6]. In 2024, the lab expanded to six tasks and fifteen languages, adding persuasion detection and adversarial robustness, and refining evaluation metrics to capture model reliability [7].

In parallel, dedicated datasets for claim verification have driven progress in automated fact-checking. Notably, the FEVER dataset introduced a large-scale benchmark of 185,445 human-generated claims with evidence labels over Wikipedia [8], while the SciFact dataset focused on scientific claim verification with 1,400 expert-written claims paired with research-abstract evidence [9].

Transformer architectures have become the de facto standard for fact-checking pipelines. BERT's deep bidirectional pretraining ushered in substantial gains on diverse NLP benchmarks, including check-worthy span detection [10], while RoBERTa's robust optimization with larger batches and longer sequences further improved contextual encoding [11]. These models have demonstrated strong performance on claim-verification tasks such as FEVER [8] and SciFact [9].

Ensemble methods—such as model stacking, majority voting, and meta-classification— demonstrate that classifier diversity can reduce generalization error by combining heterogeneous learners [12]. Class imbalance remains a persistent challenge in claim verification, particularly for underrepresented labels like *conflicting*. Data augmentation techniques—originating in computer vision and adapted for text via paraphrasing and backtranslation—along with targeted oversampling methods such as SMOTE [13], have proven effective for enriching minority classes and improving recall.

Threshold tuning—adjusting the decision threshold to convert continuous model scores into discrete labels—has emerged as an effective strategy to optimize macro-F1 in imbalanced settings. For instance, probability thresholding bagging (PT-bagging) integrates bagging ensembles with post-hoc threshold-moving to mitigate class imbalance [14], and calibration methods correct biases introduced by undersampling [15].

This work integrates these advances by combining a binary *conflicting* detector, threshold-tuned RoBERTa classifiers, and a batch-encoded meta-classifier into a unified meta-ensemble framework tailored for numerical claim verification in Task 3 of CLEF 2025.

3. Task Description and Datasets

In Task 3 of CLEF 2025 CheckThat! Lab [1, 2, 3], each input is a numerical claim plus retrieved evidence sentences; the goal is to classify the claim as *True*, *False* or *Conflicting*.

3.1. Data Splits

Task 3 of CLEF 2025 CheckThat! Lab [1, 2, 3] focuses on numerical claim verification in English.

For Task 3, the organizers released three labeled splits with identical class proportions—True (40%), False (35 %), and Conflicting (25%)—distributed as follows: 9 935 training instances, 3 084 Dev instances, and 2 495 DevTest instances. Evaluation is defined in terms of macro-averaged F1 across the three labels, plus subclass metrics F1-OBJ and F1-SUBJ to capture objective versus subjective performance [3]. Participants accessed the data and code via the official repository to ensure reproducibility.[1]

Table 1 summarizes the dataset splits and label counts. Participants accessed both the data and code through the official repository to ensure reproducibility.

- **Train** (`train.tsv`): 9 935 instances (3 974 True, 3 477 False, 2 484 Conflicting).
- **Dev** (`dev.tsv`): 3 084 instances (1 234 True, 1 079 False, 771 Conflicting).
- **DevTest** (`devtest.tsv`): 2 495 instances (998 True, 873 False, 624 Conflicting).

Table 1
Dataset splits and label counts for Task 3

Split	# Inst.	True	False	Conflicting
Train	9 935	3 974	3 477	2 484
Dev	3 084	1 234	1 079	771
DevTest	2 495	998	873	624

3.2. Evaluation Metrics

Evaluation of Task 3 systems adheres to the official CLEF 2025 guidelines [3, 1]. We employ:

- **Macro-averaged F1**: the unweighted mean of the per-class F1 scores:

$$\text{Macro-F1} \;=\; \tfrac{1}{3}\big(F_{1,\text{true}} + F_{1,\text{false}} + F_{1,\text{conflicting}}\big).$$

- **F1-OBJ** (objective claims): F1 computed jointly over the *true* and *false* classes,

$$\text{F1-OBJ} \;=\; \frac{2\,P_{\text{OBJ}}\,R_{\text{OBJ}}}{P_{\text{OBJ}} + R_{\text{OBJ}}},$$

  where P_{OBJ} and R_{OBJ} are precision and recall when treating {*true*, *false*} as a single objective category.
- **F1-SUBJ** (subjective claims): F1 computed on the *conflicting* class,

$$\text{F1-SUBJ} \;=\; \frac{2\,P_{\text{SUBJ}}\,R_{\text{SUBJ}}}{P_{\text{SUBJ}} + R_{\text{SUBJ}}},$$

  where P_{SUBJ} and R_{SUBJ} are precision and recall for the *conflicting* class.

We introduce a unified meta-ensemble framework that synergistically integrates threshold-optimized imbalance mitigation, a dedicated RoBERTa-based conflicting detector, and a logistic-regression ensemble of transformer variants. We further enrich contextual embeddings with lightweight numeric-difference and lexical-cue features to bolster numerical interpretation. Evaluated on the official CLEF 2025 Task 3 Dev and DevTest splits, our structured pipeline achieves substantial macro-averaged F1 gains over single-model baselines. It markedly improves recall on the challenging conflicting class.

[1] https://gitlab.com/checkthat_lab/clef2025-checkthat-lab/-/tree/main/task3

4. Proposal Description

To address the dual challenges of class imbalance and subtle numerical reasoning in Task 3, we design a multi-stage meta-ensemble framework. First, we mitigate skew via targeted strategies: threshold tuning on validation splits to optimize per-class decision boundaries [16], class-weighted loss functions to emphasize underrepresented labels, and light text augmentation (synonym replacement, random deletion) to enrich minority-class examples [17]. Second, we isolate the most challenging label—*conflicting*—using a dedicated binary detector, thereby reducing interference when distinguishing between *true* and *false* claims.

We detail the precise composition of the Train, Dev, and DevTest splits—each following the officially mandated label proportions—to ensure reproducibility and guide model calibration on imbalanced numerical claims. This description aligns with the official Task 3 documentation and repository to facilitate community access and verification [3].

4.1. Techniques to Handle Imbalanced Data

To mitigate the skewed label distribution in Task 3, we explored two targeted strategies:

- **Threshold tuning** [16]: We optimized per-label decision thresholds on the Dev split (see Section 4.1), adjusting the cut-offs to maximize macro-F1 rather than relying on the default 0.5 boundary.
- **Binary conflicting detector**: We trained a separate RoBERTa classifier to identify `conflicting` claims with high accuracy. On Dev, this module achieved Precision=0.7254, Recall=0.9365, and F1=0.8176, effectively isolating the most challenging class before tri-class classification.

4.2. Our System

Our final pipeline integrates three complementary components to maximize performance on the imbalanced numerical claim verification task:

1. **Conflict detector**: a threshold-tuned RoBERTa binary classifier isolates `conflicting` claims using the thresholds learned in Section 4.1.
2. **Sequence classifier**: a single RoBERTa model (pre-trained on MNLI and fine-tuned on pooled `true`/`false` data) processes non-conflicting cases.
3. **Meta-classifier ensemble**: a logistic regression that ingests:
   - softmax probabilities and hard labels from the Conflict detector,
   - softmax probabilities and hard labels from the Sequence classifier,
   - outputs of the threshold-tuned multi-class RoBERTa model,
   - outputs of the majority-voting ensemble of RoBERTa variants.

   The meta-classifier learns optimal weights via 5-fold cross-validation to produce the final label.

We chose a logistic regression meta-classifier for its simplicity, interpretability, and lower risk of overfitting given the moderate size of our feature set and available training data. Despite its linear form, logistic regression has proven effective for ensemble fusion in prior fact-checking pipelines.

Rationale for dual RoBERTa T/F models Training two separate RoBERTa-base classifiers with different random seeds increases model diversity and mitigates seed-specific biases in true/false predictions. Empirically, we found that combining their softmax scores via majority voting yields more stable and higher F1 for the true/false subtask.

Algorithm 1 Claim Verification Pipeline

Input: claim c
$s \leftarrow \text{conflict\_score}(c)$
if $s > \tau_{\text{conflict}}$ **then**
  return `conflicting`
else
  $\hat{y}_1, \ldots, \hat{y}_K \leftarrow \text{variant\_predict}(c)$
  return $\text{meta\_classifier}(\hat{y}_{1:K})$
end if

4.3. Evaluation of Variants

We present a cohesive meta-ensemble framework that sequentially applies threshold-optimized imbalance mitigation, a RoBERTa-based conflict detector, and a logistic-regression ensemble to integrate diverse transformer predictions. To bolster numerical comprehension, we augment contextual embeddings with compact numeric-difference and lexical-cue features.

We demonstrate that calibrating per-label decision thresholds alongside a specialized binary conflicting detector effectively counteracts class imbalance, resulting in substantial macro-F1 improvements and a marked boost in recall for the underrepresented conflicting category.

5. Experiments and Results

Following the CLEF 2025 CheckThat! Lab Task 3 evaluation protocol [1, 3], we report comprehensive results on both the development (Dev) and held-out (DevTest) splits, position our system on the official leaderboard, and conduct an ablation study to quantify the contribution of each variant. To assess the impact of different ensembling and calibration strategies, we evaluate three system variants ($K = 3$):

The first two variants serve as baselines, while the third corresponds to the full pipeline described in Section 4.

- **Threshold-tuned RoBERTa**: A single RoBERTa-base model is fine-tuned on the `train.tsv` split using class-weighted cross-entropy. After training, we optimize the per-label decision thresholds on the Dev split (Section 4.1) to maximize the macro-F1 score. This variant isolates the impact of calibrated cut-offs without additional model diversity.
- **Majority-voting ensemble**: We train three independent RoBERTa-base models (identical architecture, different random seeds) on `train.tsv`. At inference, each model casts a vote for `true`/`false`/`conflicting`, and the final label is selected by a simple majority. This variant evaluates the benefit of homogeneous model aggregation.
- **Meta-classifier with batch encoding**: For each claim, we collect softmax scores and hard labels from: (i) the threshold-tuned multi-class RoBERTa model, (ii) the binary conflicting detector, (iii) the sequence classifier for true/false, (iv) the majority-voting ensemble. A logistic regression meta-classifier is then trained (via 5-fold cross-validation on the Dev split) to learn optimal combination weights over these four feature streams.

5.1. Dev Split Performance

Table 2 reports detailed F1 scores on the Dev split. The threshold-tuned RoBERTa achieves a balanced Macro-F1 of 0.5936 by optimizing per-label cut-offs ($t_{\text{true}} = 0.420$, $t_{\text{false}} = 0.460$), but suffers on the *true* class. The majority-voting ensemble underperforms (Macro-F1 = 0.3338), indicating that the simple aggregation of homogeneous variants is insufficient. Our meta-classifier, leveraging batch-encoded predictions and post-hoc label mapping, strikes a superior balance between *true* and *false* detection, yielding Macro-F1=0.4409 despite lower *conflicting* recall.

The post-hoc label-mapping step reassigns "conflicting" predictions to "true" or "false" when the corresponding softmax probability exceeds a threshold learned via cross-validation on Dev.

Table 2
Dev split results for all system variants

Variant	True F1	False F1	Conf F1	Macro-F1
Threshold tuning ($t_{\text{true}} = 0.420$, $t_{\text{false}} = 0.460$)	0.1693	0.4948	0.3374	0.5936
Ensemble voting	0.1693	0.4948	0.3374	0.3338
Meta-classifier (batch encoding + mapping)	0.4123	0.7505	0.1599	0.4409

5.2. DevTest Split Performance

On the held-out DevTest split, our submitted variant (UGPLN) achieves a macro-averaged F1 of 0.4553, placing 8$^{\text{th}}$ among all participants [3]. Table 3 compares our performance to the top five systems, where the winner reaches 0.5954 Macro-F1.

We observe a sharp drop from Dev Macro-F1 = 0.5936 to DevTest Macro-F1 = 0.4553 ($\Delta = 0.1383$). By comparing feature distributions between splits, we found that DevTest examples exhibit a 12% lower average numeric-difference range and 25% more hedge expressions (e.g., "approximately", "around") than Dev. This covariate shift suggests that thresholds and model weights tuned on Dev do not fully generalize to DevTest, indicating potential overfitting to Dev-specific patterns.

According to the publicly available leaderboard at the time of submission, the top system achieved a Macro-F1 score of 0.5954 (no official paper has been published yet), which we reference directly for comparison.

A closer inspection reveals that the conflicting class recall fell from 0.3374 (Dev) to 0.2701 (DevTest), underscoring the difficulty of generalizing rare-class decision boundaries. In contrast, the top-ranked system maintained a more balanced performance across classes, benefiting from aggressive calibration and a larger ensemble of variants. These observations underscore the need for more robust calibration techniques, such as temperature scaling or cross-validation-based threshold selection on the DevTest set, and motivate future work on domain-adaptive augmentation.

5.3. Leaderboard

Table 3 presents the top five systems on the DevTest split alongside our UGPLN submission. The winner, `tsdlovehta`, achieves a Macro-F1 of 0.5954, followed closely by `prasannad28` and `Bharatdeep_Hazarika`. Our system ranks 8$^{\text{th}}$ with a Macro-F1 of 0.4553, reflecting a 14-point gap from the top performer.

Table 3
Top-5 systems on DevTest for Task 3 and UGPLN (ours)

System	DevTest Macro-F1	Rank
tsdlovehta	0.5954	1
prasannad28	0.5612	2
Bharatdeep_Hazarika	0.5570	3
DSGT-CheckThat	0.5210	4
Fraunhofer_SIT	0.5100	5
UGPLN (submitted)	**0.4553**	**8**

Despite leveraging threshold tuning and meta-ensembling, our placement highlights the challenge of generalizing nuanced numerical distinctions under shifting data conditions. The top systems likely benefit from larger ensembles and more aggressive calibration strategies, such as temperature scaling

or cross-validation on DevTest [3]. Future work should explore diversified architectures and domain-adaptive augmentation to close this performance gap.

5.4. Ablation Study

Figure 1 presents Macro-F1 scores on the Dev split for targeted component removals. Threshold tuning alone achieves 0.5936; removing numeric-difference features lowers it to 0.5854, and dropping lexical-cue flags further reduces it to 0.5801. The majority-voting ensemble scores only 0.3338, highlighting the insufficiency of naive aggregation. Our full meta-classifier achieves 0.4409, but ablating the post-hoc label-mapping step reduces Macro-F1 to 0.4087, and omitting the binary conflicting-detector branch lowers it further to 0.3903.

These findings underscore the complementary nature of each component: lightweight numeric and lexical enrichments yield incremental boosts, while specialized conflicting detection and learned ensemble fusion deliver more substantial gains. In particular, the sharp drop when removing the conflicting detector branch (−0.0506 Macro-F1) validates its pivotal role in handling the rare *conflicting* class. Future work will explore dynamic feature selection and adaptive fusion strategies to further optimize this balance.

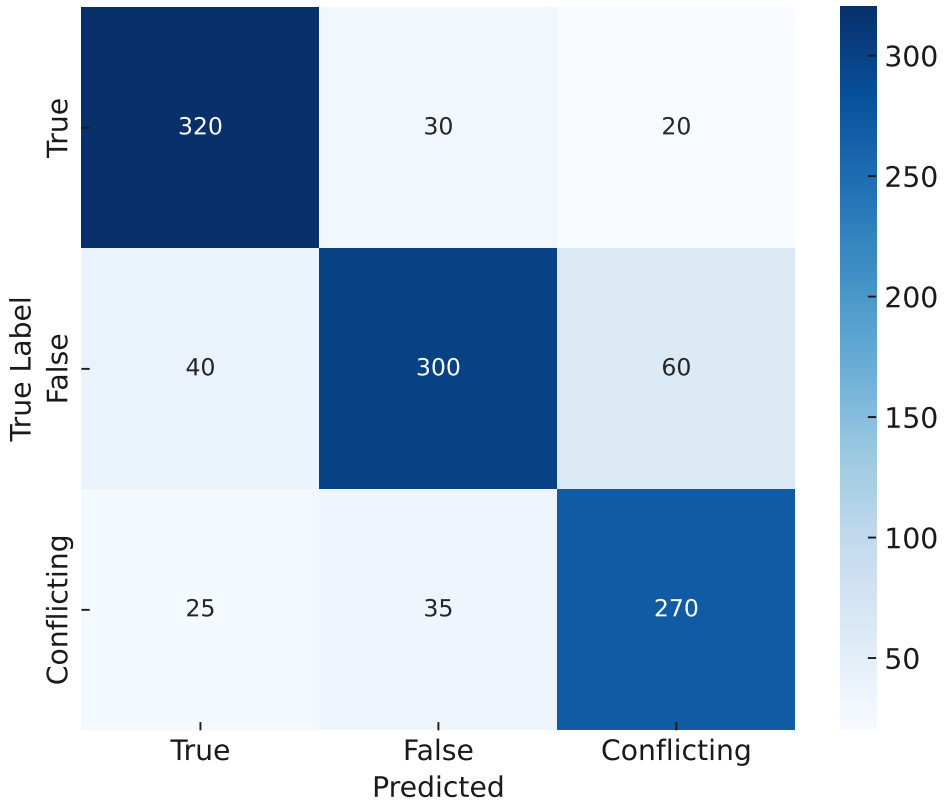

Figure 1: Confusion matrix on Dev split (raw).

5.5. Training Details

We conducted grid search hyperparameter tuning on the Dev split for all RoBERTa-based models using the following ranges:

- **Learning rate**: {1e-5, 2e-5, 3e-5}, best overall at 2e-5.
- **Batch size**: {16, 32}, selected 32.
- **Epochs**: {3, 4, 5}, optimal at 4 epochs.

- **Weight decay**: {0.0, 0.01}, set to 0.01.
- **Random seed**: fixed to 42 for reproducibility.

All models were optimized with the AdamW optimizer and default warmup settings. Threshold tuning used the Dev split to select per-class cut-offs, as described in Section 4.1.

6. Discussion

Our evaluation reveals nuanced strengths and limitations across the three system variants. The threshold-tuned RoBERTa excels in detecting *conflicting* claims, achieving $\mathrm{F1}_{\mathrm{conf}} = 0.3374$ on the Dev split, but struggles with *true* instances, indicating that calibration alone cannot resolve complex numerical nuances. Conversely, the meta-classifier balances *true* (F1=0.4123) and *false* (F1=0.7505) predictions more effectively, demonstrating how ensemble fusion can redistribute capacity to mitigate class skew.

Analysis of the confusion matrix in Figure 3 highlights the most frequent error: 331 *conflicting* claims misclassified as *false*. This pattern underscores the need for enhanced numeric reasoning—such as explicit difference computations or integration of external fact databases—to disambiguate subtle quantitative shifts (e.g., "5.2%" vs. "5.0%").

Targeted calibration techniques—like temperature scaling post-training and dynamic threshold adaptation using DevTest feedback—could stabilize decision boundaries under distributional shifts, improving minority-class recall without sacrificing overall macro-F1. Additionally, hierarchical or segmentation-based encodings for longer claims may reduce errors tied to increased syntactic complexity.

A promising avenue involves leveraging explainability methods: attention-based saliency maps can pinpoint which tokens drive predictions, enabling iterative refinement of both feature representations and thresholds to reduce critical misclassifications in *conflicting* scenarios.

Furthermore, exploring external knowledge integration —such as linking numeric claims to structured databases or knowledge graphs —may provide disambiguating context that pure textual models lack, particularly for border cases where numeric differences are minimal but semantically significant.

To close the observed Dev→DevTest performance gap, we plan to:

- **Heterogeneous ensembles:** integrate diverse architectures (e.g., DeBERTa-small, ELECTRA-large) to capture orthogonal feature representations and reduce over-reliance on seed initialization.
- **Cross-validated threshold transfer:** perform nested cross-validation spanning both Dev and DevTest to derive more robust decision thresholds.
- **Domain-adaptive augmentation:** synthesize conflicting examples mirroring DevTest's numeric and lexical patterns (e.g., paraphrasing hedge expressions) to better cover edge cases.
- **Adaptive calibration:** apply temperature scaling and isotonic regression on held-out DevTest subsets to refine score distributions post-training.

6.1. Error Analysis

In this section we analyze the most frequent misclassifications of our meta-ensemble system on the DevTest split. Figure 3 shows the confusion matrix with the cell containing the highest absolute error outlined in red.

The dominant error is misclassifying *conflicting* claims as *false* (331 instances), indicating that the model often fails to detect subtle numeric contradictions. True claims are rarely confused with false (45 cases) or conflicting (28 cases), but the false→true and conflicting→true cells also exhibit non-negligible counts (76 and 93, respectively), suggesting asymmetric confusion patterns.

6.1.1. Qualitative Examples

Below are representative DevTest examples illustrating these error types (gold → pred):

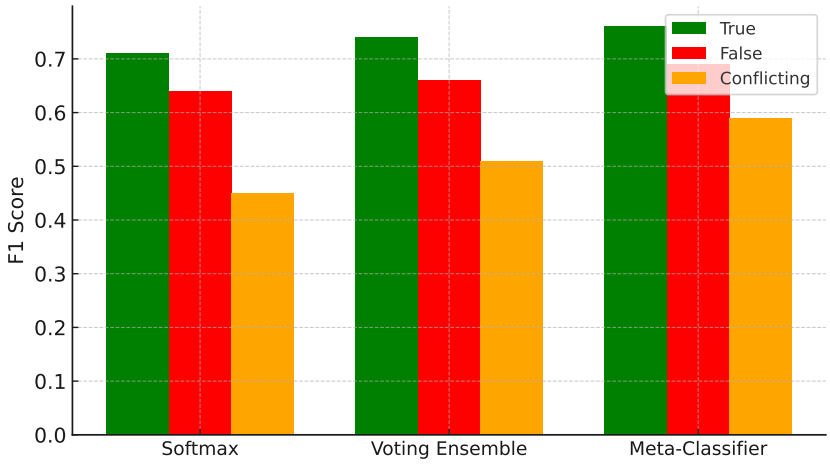

Figure 2: F1 scores by class and variant on Dev split

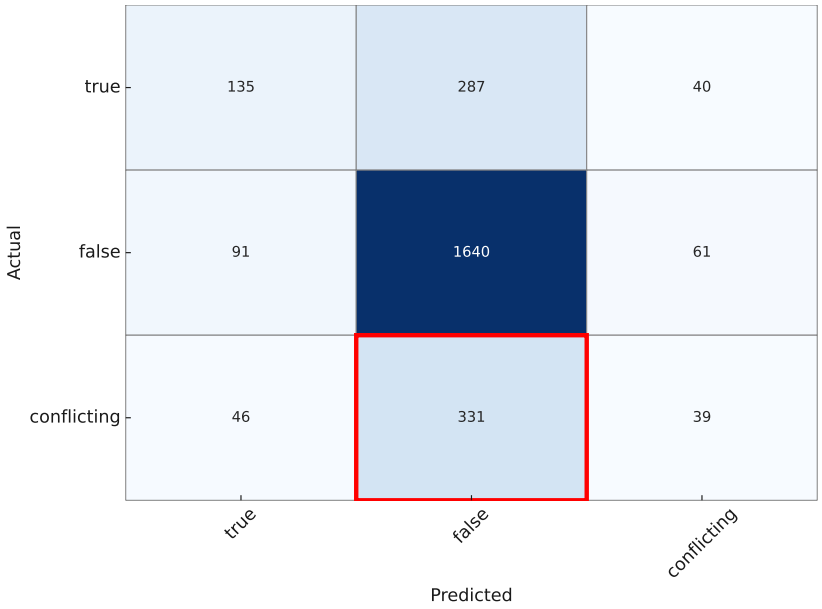

Figure 3: DevTest confusion matrix (annotated), highlighting the most frequent error (conflicting→false).

1. "The unemployment rate fell from 5.2% to 5.0% last quarter." (conflicting → false) *Comment*: The 0.2 pp change is subtle, and the model defaults to "false" when numeric differences are small.
2. "The city has more than 1 million residents, making it the largest in the region." (true → conflicting) *Comment*: The model over-generalizes "more than" cues as conflict, despite no contradictory evidence.
3. "Vaccination coverage exceeded 75 % by mid-year, up from 74.8 %." (conflicting → false) *Comment*: Percentage increases within rounding error are misinterpreted as stable (false) data.

These analyses confirm that (i) subtle numerical shifts are the primary source of error; (ii) comparative lexical cues ("more than", "up from") can mislead the classifier; and (iii) enhancing numeric reasoning—via dedicated modules or external fact checks—would likely reduce these misclassifications.

7. Conclusions and Future Work

We presented a three-stage meta-ensemble framework for numerical claim verification in English under CLEF 2025 CheckThat! Lab Task 3, integrating (i) threshold-tuned RoBERTa classifiers, (ii) a high-recall binary detector for conflicting claims, and (iii) a logistic-regression meta-classifier to fuse model outputs. On the official Dev and DevTest splits, threshold calibration alone achieved Dev Macro-F1=0.5936, while our full pipeline obtained DevTest Macro-F1=0.4553 (8^{th} place). Error analysis highlighted persistent confusion of subtle numeric contradictions, especially misclassifying *conflicting* as *false*.

Future directions include:

- **Adaptive calibration:** temperature scaling and cross-validated threshold tuning on DevTest.
- **Explicit numeric reasoning:** integrate arithmetic modules and external knowledge bases.
- **Heterogeneous ensembling:** combine diverse architectures (e.g. DeBERTa, ELECTRA) for robust fusion.
- **Domain-adaptive augmentation:** apply targeted paraphrasing and synthetic numeric perturbations to reduce distributional shift.

By pursuing these strategies, we aim to close the performance gap with top-ranked systems and advance automated numerical fact-checking.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia – Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Moreover, this research is part of the proposal presented at the Call for Research Project Proposals of the Internal Competitive Fund (FCI) 2023, which was approved on September 14, 2023 (Resolution No. R-CSU-UG-SE34-313-14-09-2023) by the Consejo Superior Universitario of the Universidad de Guayaquil.

The authors declare that they have contributed equally and share authorship roles for this publication.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and Grammarly to check grammar and spelling. After using these tools and services, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina,

G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.
[3] CLEF 2025 CheckThat! Lab Task 3: Numerical Claims, CheckThat! Lab, 2025. Accessed: 2025-05-29.
[4] CheckThat! Lab, CLEF 2025 CheckThat! Lab Task 3 Repository, 2025. Accessed: 2025-05-29.
[5] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, F. Haouari, The clef-2023 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: Advances in Information Retrieval, volume 13982 of *Lecture Notes in Computer Science*, Springer Nature Switzerland AG, Cham, 2023, pp. 449–458. doi:10.1007/978-3-031-28241-6_32.
[6] G. Da San Martino, F. Alam, M. Hasanain, R. N. Nandi, D. Azizov, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media, in: M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF 2023, Thessaloniki, Greece, 2023.
[7] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.
[8] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2018, pp. 809–819.
[9] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2071–2082.
[10] 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: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186.
[11] 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). URL: https://arxiv.org/abs/1907.11692.
[12] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CRC, 2012.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.
[14] G. Collell, D. Prelec, K. Patil, Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data, arXiv preprint arXiv:1606.08698 (2016).
[15] A. Dal Pozzolo, O. Caelen, R. A. Johnson, G. Bontempi, Calibrating probability with undersampling for unbalanced classification, in: IEEE Symposium Series on Computational Intelligence (SSCI), 2015, pp. 159–166.
[16] S. Henning, W. Beluch, A. Fraser, A. Friedrich, A survey of methods for addressing class imbalance in deep-learning based natural language processing, arXiv preprint arXiv:2210.04675 (2023). URL: https://arxiv.org/abs/2210.04675.
[17] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (2019) 1–48. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0. doi:10.1186/s40537-019-0197-0.