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*

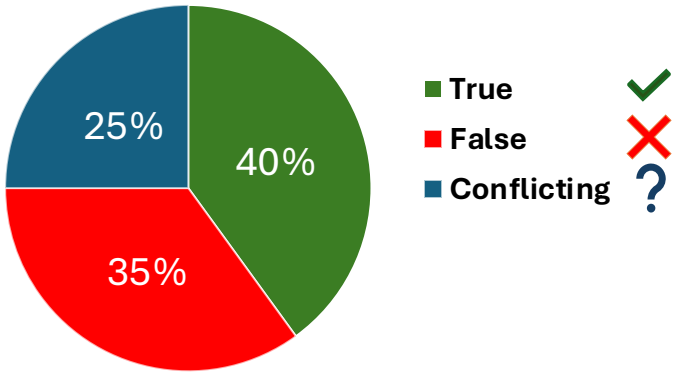
Introduction – Challenges in Numerical Claim Verification

Finance domain

"The unemployment rate fell from 5.2 % to 5.0 % last quarter."
Original label **Conflicting**
Predicted label **False**

Health domain

"Vaccination coverage exceeded 75 % by mid-year, up from 74.8 %"
Original label: **Conffficting**
Predicted label **False**



High-stakes importance: Automated verification supports domains such as finance and public health, because decisions and investments are guided by precise data.

Three-way classification task: The CLEF 2025 task 3 requires assigning each claim to one of three labels—True, False or Conflicting—based on retrieved evidence.

Sensitive to small errors: Small quantitative inconsistencies (e.g., 5.2 % vs 5.0 %) can invert a claim's truth value, making robust models necessary.

Imbalanced dataset: Approximately 40 % True, 35 % False, 25 % Conflicting, challenging for training and evaluation.

Need for fine-grained numeric reasoning: Models must capture subtle numerical differences and handle imbalance to achieve reliable performance.

Task Description – CLEF 2025 Task 3: Numerical Claim Verification

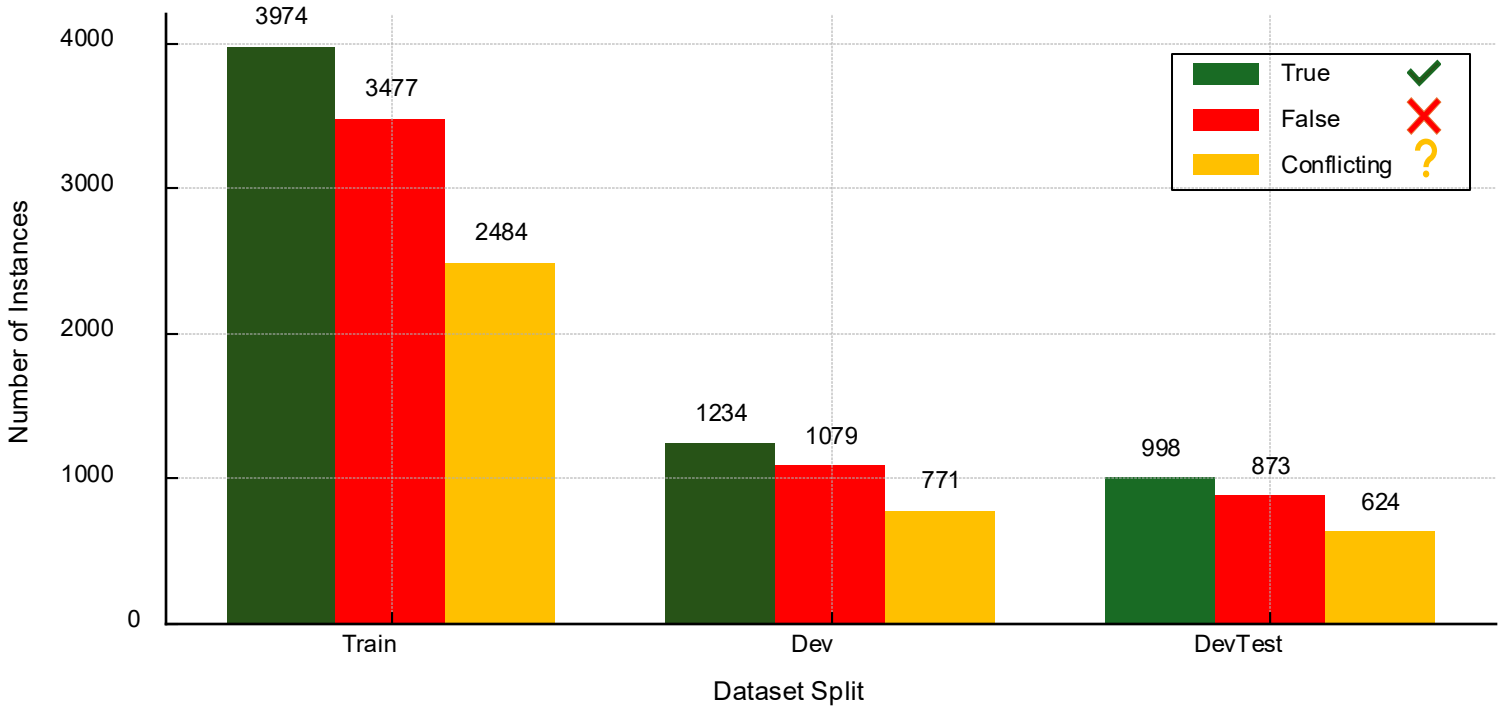
Task: Classifying numerical claims as True, False, or Conflicting based on retrieved evidence.

Dataset Splits

Train: 9935 instances

Dev: 3084 instances.

DevTest: 2495 instances.



Evaluation Metrics: Macro-F1 (averaged across all labels), F1-OBJ (True/False), F1-SUBJ (Conflicting).

Three-Stage Meta-Ensemble Pipeline for Numerical Claim Verification

Stage 1 - Conflict Detector

Goal: Isolate conflicting claims.

Model: RoBERTa binary classifier (threshold-tuned).

Hugging Face RoBERTa

Techniques:

- Threshold tuning.
- Class-weighted loss.
- Light text augmentation.

Impact: High recall ensures most conflicting claims are captured early, reducing noise for later classification.

Output:
Conflicting (→ Stage 3)
Non-conflicting (→ Stage 2)

Stage 2 - Sequence Classifier

Goal: Classify non-conflicting claims as True/False.

Model: RoBERTa sequence classifier (MNLI pre-trained, fine-tuned on pooled True/False data).

Hugging Face RoBERTa MNLI

Techniques:

- Threshold tuning.
- Class-weighted loss.

Impact: Improves precision/recall balance and provides calibrated scores for Stage 3.

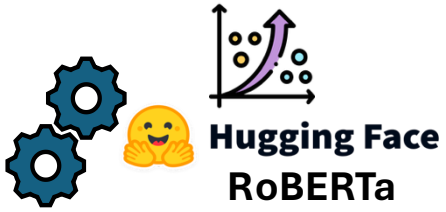
Input: Non-conflicting claims from Stage 1.

Output:
Softmax (True/False) and hard label (→ Stage 3) .

Stage 3 - Meta-classifier ensemble

Goal: Fuse predictions from stages 1, Stage 2 and additional models.

Model: Logistic Regression Meta-Classifer (ensemble fusion).



Techniques:

- Combines softmax and hard labels.
- Inputs: Stage 1 (Conflict) and Stage 2 (True/False).
- Threshold-tuned multi-class RoBERTa.
- Majority-voting ensemble of RoBERTa variants.

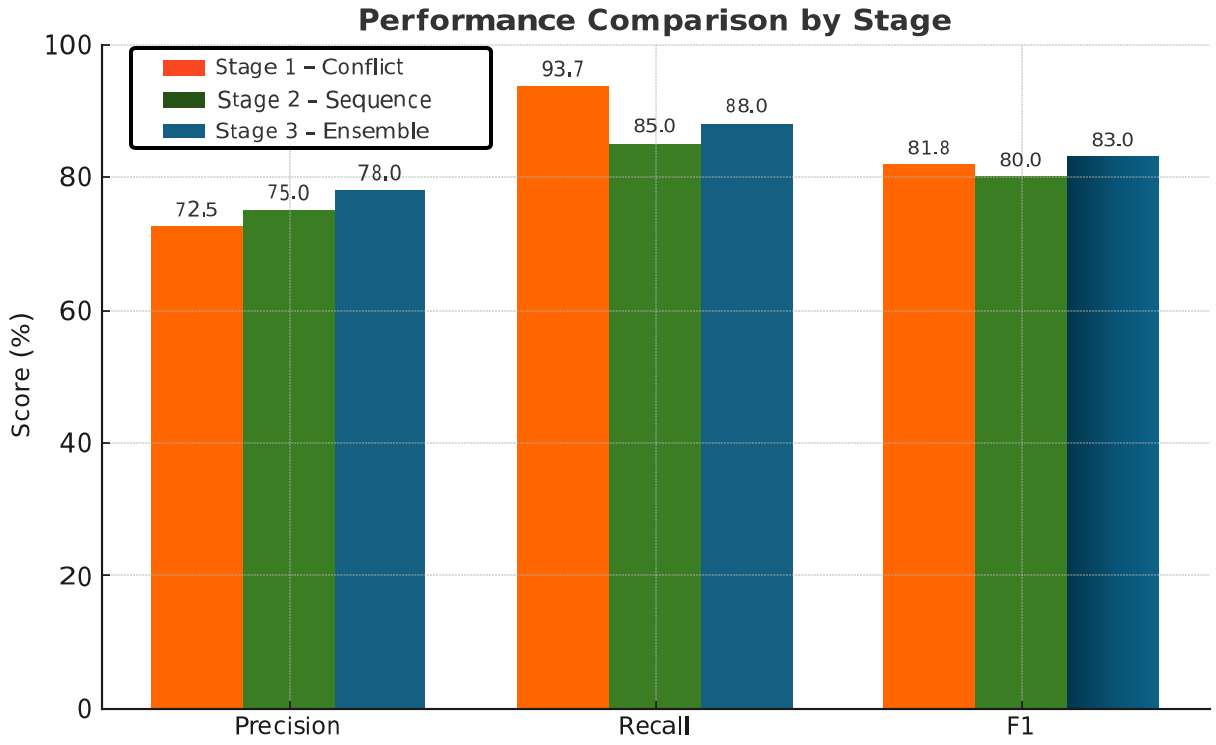
Impact: Robust and balanced predictions across classes.

Output:
Final classification (Conflicting / True / False).

Performance by Pipeline Stage for Numerical Claim Verification

Table 1 – Stage-wise performance

Stage	Precision	Recall	F1
1 – Conflict Detector	72.54%	93.65%	81.76%
2 – Sequence Classifier	≈ Balanced	≈ Balanced	Improved vs Stage 1
3 – Meta-Classifier Ensemble	Optimized	Optimized	Best overall



Conflict Detector (Stage 1): High recall (**93.65%**), ensuring most conflicting claims captured.
Sequence Classifier (Stage 2): Improved precision/recall balance with threshold tuning and MNLI fine-tuning.
Meta-Classifier Ensemble (Stage 3): Best overall performance, leveraging softmax and hard labels and diverse model fusion.
Robust performance across imbalanced classes; ensemble **improved Macro-F1** significantly compared to individual models.

Experimental Results – Performance Evaluation

Table 2 – Dev split results for all system variants

Variant	True F1	False F1	Conf F1	Macro-F1
Threshold tuning (t_{true}=0.420, t_{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

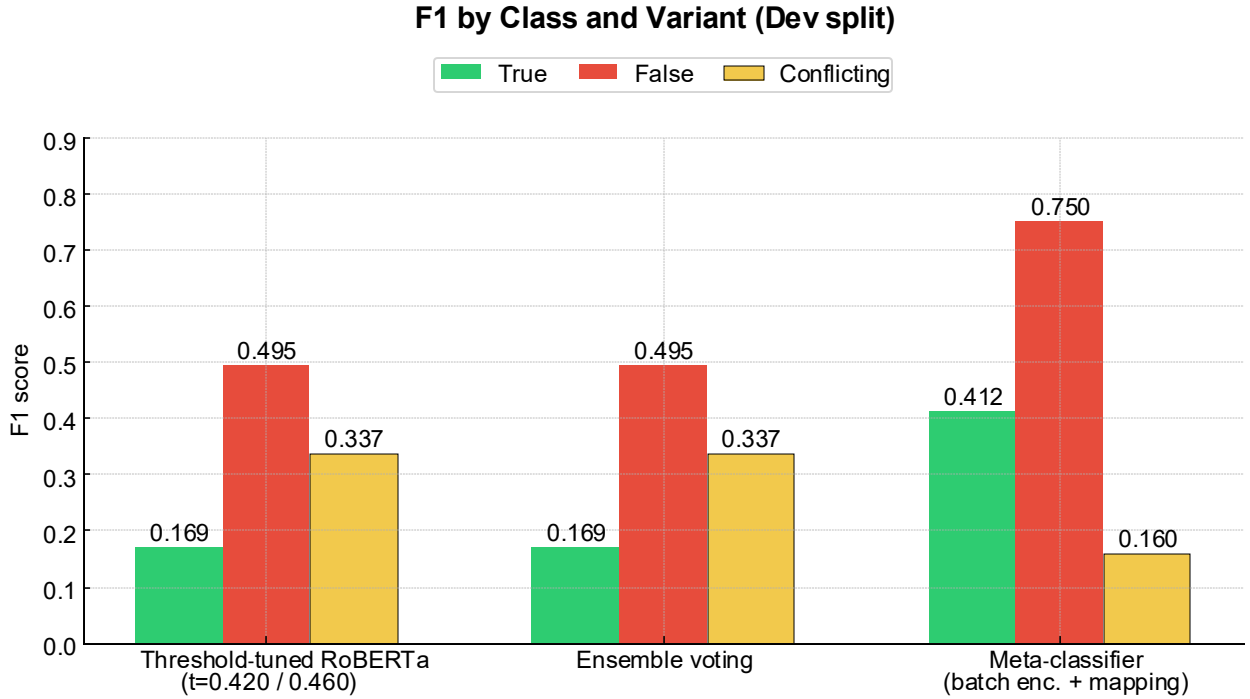
Table 3 – Top-5 systems on DevTest + UGPLN (ours)

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**

Meta-classifier ensemble achieved the best Macro-F1 on Dev, while our UGPLN system ranked 8th on DevTest (Macro-F1 = 0.4553).

Discussion

Stage 1 (Conflict Detector)	High recall (≈93.6%) ensures most conflicting claims are captured
Stage 2 (Sequence Classifier)	Threshold tuning + MNLI pre-training improved precision/recall balance.
Stage (Meta-classifier Ensemble)	Best Macro-F1, confirming the benefit of model fusion.
Variants	Meta-classifier excels in True/False but struggles with Conflicting class



The ensemble meta-classifier achieved the best overall balance, but handling Conflicting claims remains the main challenge.

Conclusions

- We proposed a three-stage pipeline consisting of conflict detection, true/false classification and ensemble fusion.
- The ensemble meta-classifier achieved the highest Macro-F1, clearly outperforming single models.
- In the official CLEF 2025 evaluation, our system SINAI and UGPLN ranked within the Top-10 on the DevTest leaderboard

Future Work

- Extend evaluation to other languages and domains.
- Explore lighter and interpretable models for real-time fact-checking (Project FCI-079-2023 Universidad de Guayaquil about Fake News in Ecuador)
- Optimize the ensemble strategy with more diverse model variants and adaptive thresholding.

Acknowledgements

CONSENSO
MODERATES
SocialTox

”

Thanks you