VerbaNexAI at CheckThat! 2025: Fine-Tuning DeBERTa for Multi-Label Scientific Discourse Detection in Tweets*

Notebook for the CheckThat! Lab at CLEF 2025

Mervin Jesus Sosa Borrero[1,*,†], Jairo Enrique Serrano Castañeda[1,†], Juan Carlos Martinez Santos[1,†] and Edwin Alexander Puertas Del Castillo[1,*,†]

[1]Universidad Tecnológica de Bolívar, School of Digital Transformation, Cartagena de Indias, 130010, Colombia

Abstract

This paper presents VerbaNexAI's submission to Task 4a of the CheckThat! 2025 Lab, which focuses on the identification of scientific discourse in English-language tweets. We propose a multi-label classification approach based on a fine-tuned DeBERTa-v3 model, optimized through stratified cross-validation, threshold calibration using precision-recall curves, and ensemble prediction with soft-voting. Our system ranked 2nd overall in the official leaderboard with a macro-averaged F1 score of 0.7983 and achieved the top F1 score (0.8133) in Category 1 (scientific claims), demonstrating strong performance in detecting verifiable assertions in noisy social media contexts. To address class imbalance and label sparsity, we employed class-specific weighting and threshold tuning strategies. The final system combines predictions from multiple folds and loss configurations, resulting in robust generalization. Code, models, and evaluation scripts are publicly available to promote reproducibility and further research on trustworthy scientific information detection in social platforms.

Keywords

Scientific Discourse Detection, Multi-label Classification, Transformer Models, Twitter Data, Threshold Optimization, CEUR-WS

1. Introduction

In recent years, the widespread use of social media platforms has transformed the way scientific information is produced, shared, and consumed. Platforms such as Twitter play a central role in shaping public discourse on critical scientific issues, including public health, climate change, and technology policy. However, this rapid dissemination of information also amplifies challenges around the reliability, credibility, and traceability of scientific claims online. In response, the computational community has increasingly focused on developing systems to automatically detect, verify, and classify scientific discourse across social platforms.

The CLEF CheckThat! Lab [1] has emerged as a pivotal benchmark for evaluating such systems. In particular, Task 4a of the CheckThat! 2025 Lab [2] addresses the detection of scientific discourse in tweets, grounded in an annotation framework established in prior work such as SciTweets. This framework introduces a nuanced categorization of science-related content, distinguishing between: (1) scientifically verifiable claims, (2) references to scientific knowledge, and (3) mentions of scientific research or context. This task is a multi-label classification challenge complicated by linguistic variability, domain ambiguity, and class imbalance—factors that mirror the real-world complexity of scientific communication.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

sosam@utb.edu.co (M. J. Sosa Borrero); jserrano@utb.edu.co (J. E. Serrano Castañeda); jcmartinezs@utb.edu.co (J. C. Martinez Santos); epuerta@utb.edu.co (E. A. Puertas Del Castillo)

0009-0003-4095-3332 (M. J. Sosa Borrero); 0000-0001-8165-7343 (J. E. Serrano Castañeda); 0000-0003-2755-0718 (J. C. Martinez Santos); 0000-0002-0758-1851 (E. A. Puertas Del Castillo)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4038/paper_97.pdf

Based on this foundation, we present a multi-stage pipeline designed to detect scientific discourse in tweets. Our approach integrates transformer-based architectures, threshold calibration techniques, and ensemble learning strategies to maximize performance across all three categories. Specifically, we fine-tuned the microsoft/deberta-v3-base model using stratified 5-fold cross-validation, optimizing hyperparameters such as learning rate and training epochs. To handle label imbalance, we introduced class-specific weighting using the BCEWithLogitsLoss function and applied precision-recall-based threshold tuning per label. We produced final predictions through a soft-voting ensemble of the top-performing models. Unlike previous approaches that relied on heuristic sampling or simpler classifiers, our system emphasizes careful optimization and generalization.

We trained solely on the official ct-train.tsv dataset, without external corpora or retrieval augmentation, thereby validating its robustness in a constrained but realistic setting. Importantly, we did not employ additional pretraining, zero-shot methods, or rule-based heuristics. It highlights the strength of focused fine-tuning and calibration.

Our system achieved competitive results in the official evaluation of Task 4a. It ranked 2nd overall in macro-averaged F1-score (0.7983) across all categories. Notably, it obtained the highest F1-score (0.8133) in Category 1, which focuses on scientifically verifiable claims—arguably the most impactful category for downstream fact-checking and credibility analysis. These results demonstrate that our approach not only generalizes well but also excels in the most challenging and socially relevant facet of the task. This paper documents our system, from data preprocessing and model selection to experimental design and result analysis. We also reflect on the implications of our design choices and propose avenues for future research, including multilingual adaptation, external knowledge integration, and explainability in scientific discourse classification.

2. Tasks and Objectives

The primary objective of our participation in Task 4a of the CheckThat! 2025 Lab was to design a robust and reproducible system capable of accurately identifying different types of scientific discourse in tweets. This involved addressing the multi-label nature of the task, optimizing performance across all three predefined categories: scientifically verifiable claims (C1), references to scientific knowledge (C2), and mentions of scientific research or context (C3) Table 1 summarizes these three categories.

Table 1
Categories of scientific discourse used in Task 4a.

Category	Description
1 - Scientific Claim	Scientifically verifiable statements or questions.
2 - Reference	Mentions or links to scientific knowledge or publications.
3 - Research Context	Mentions of researchers, institutions, or ongoing studies without claims.

We formulated the classification problem as a multi-label task, where each tweet may belong to zero, one, two, or all three categories simultaneously. The dataset, provided in tabular format with tweets and corresponding binary vectors (e.g., ‘[0.0, 1.0, 0.0]‘), reflects the challenges inherent in analyzing social media content, including informal language, abbreviations, hashtags, hyperlinks, emojis, and occasional sarcasm or irony. A significant class imbalance further complicates the task, with a majority of tweets falling into the “no-category” case (‘[0.0, 0.0, 0.0]‘). We evaluated performance using macro-averaged F1-score across the three categories.

2.1. Main Objectives of Experiments

Our primary goal was to develop a scalable system that achieves high predictive performance across all three categories while ensuring consistency and reproducibility. Table 2 summarizes the guiding design principles that structured our approach. Through these objectives, we aim to demonstrate that careful

Table 2
Principles guiding system design.

Step	Description
1	Fine-tune a transformer model to capture scientific language patterns.
2	Handle class imbalance via weighting and threshold tuning.
3	Validate with stratified 5-fold cross-validation.
4	Aggregate predictions using ensemble (soft-voting).
5	Limit resources to official data and reproducible methods.

model tuning and architectural design can yield high-quality results in challenging multi-label settings, even in the absence of auxiliary data or external knowledge sources.

3. Methodology

We built our system for Task 4a of CheckThat! 2025 upon a transformer-based architecture designed for multi-label classification of scientific discourse in tweets. The overall methodology consisted of four key stages: data preprocessing, model fine-tuning, threshold calibration, and ensemble integration. All experiments were conducted exclusively on the English-language training set provided by the organizers, without incorporating external corpora, retrieval systems, or manually crafted rules. Our pipeline is illustrated in Figure 1, which outlines the main processing stages.

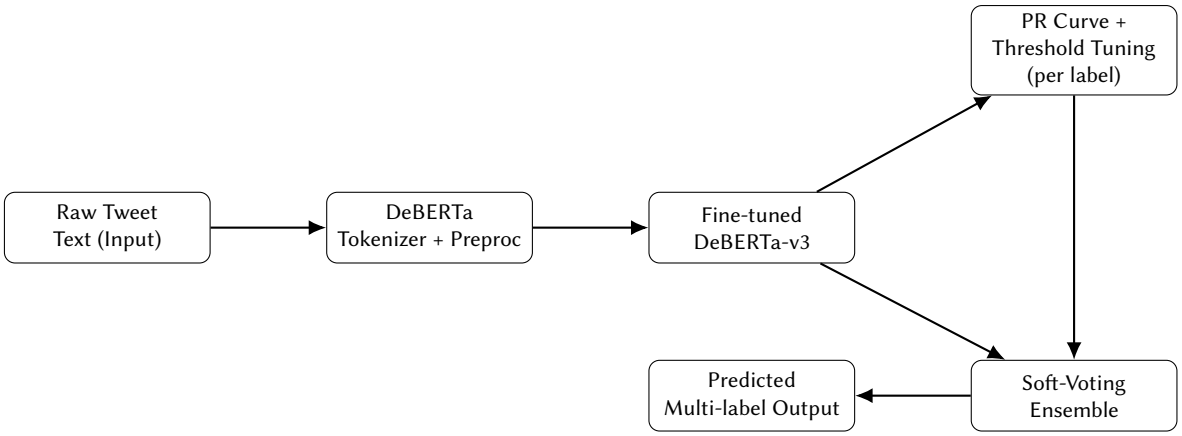

Figure 1: System architecture of the pipeline for multi-label scientific discourse detection in tweets. We tokenized raw tweet text using the DeBERTa tokenizer and fed it into a fine-tuned DeBERTa-v3 model. We passed outputs through a threshold tuning module (based on precision-recall curves) and a soft-voting ensemble that combines multiple model predictions. The final output consists of predicted multi-label categories for each tweet.

3.1. Model Architectures

We selected `microsoft/deberta-v3-base` as the backbone model for our system based on its empirical performance and architectural advantages over other widely used transformers such as BERT, RoBERTa, and XLNet. DeBERTa (Decoding-enhanced BERT with disentangled attention) improves upon BERT-based models by decoupling positional and content-based attention, enabling the model to capture longer-range dependencies and nuanced phrase structures more effectively. It is particularly beneficial for scientific discourse, where key information may appear in complex or indirect expressions [3].

Compared to RoBERTa, which enhances BERT with more robust pretraining, DeBERTa additionally integrates a relative position bias and enhanced mask decoder that improves fine-tuning performance in sentence-level classification tasks. XLNet, while powerful in autoregressive settings, introduces increased complexity and instability in multi-label fine-tuning for short-form noisy text such as tweets.

We conducted preliminary ablation experiments with RoBERTa-base and BERT-base on a validation fold. We observed that DeBERTa-v3 consistently achieved 2–3 F1 points higher on Category 1 (scientific claims) and showed better calibration across all thresholds. For these reasons, we adopted DeBERTa-v3-base as the most appropriate foundation for our multi-label tweet classifier.

3.2. Preprocessing

The input data consists of tweets accompanied by binary vectors that indicate the presence or absence of each of the three target categories. We tokenized the tweets using the Hugging Face implementation of the DeBERTa tokenizer, with a maximum sequence length of 128 tokens. Special tokens, hashtags, emojis, and URLs were preserved during tokenization, as these often provide important context in scientific discourse on social media. No additional text normalization or truncation was applied beyond the tokenizer's built-in handling.

To optimize classification performance, we followed a multi-stage training and ensembling strategy summarized in Algorithm 1. This pipeline involves training two model variants separately: one using the `BCEWithLogitsLoss` function with class weights (via the `pos_weight` parameter) to mitigate label imbalance, and another using the same loss function without weights. Both variants are trained using 5-fold stratified cross-validation, and their best-performing checkpoints (based on macro F1-score) are retained. Final predictions are obtained by applying soft-voting over the outputs of the top models from each variant.

Algorithm 1 Multi-Stage Fine-Tuning Strategy for Multi-Label Tweet Classification. The pipeline includes two model variants trained independently—with and without class weights—and combines their predictions via soft-voting.

Initialize DeBERTa-v3-base model with pretrained weights.
Set training hyperparameters: learning rate, epochs, batch size.
Tokenize tweets using DeBERTa tokenizer (max 128 tokens).
Prepare optimizer (AdamW) and loss function (BCEWithLogitsLoss).
for each fold in 5-fold cross-validation **do**
    Train on four folds and validate on the 5th.
    Tune thresholds using precision-recall curve.
    Save the best model based on macro F1.
end for
Load top 2 models (with and without class weights).
Compute soft-voting over prediction probabilities.
Return final predictions.

3.3. Training Configuration

We trained the model using the `BCEWithLogitsLoss` function [4] to accommodate the multi-label nature of the task. To address class imbalance—particularly the prevalence of negative-only examples ([0, 0, 0])—we applied label-specific positive class weighting via the pos-weight parameter computed from the training label distribution. We utilized the Paged AdamW optimizer (32-bit) from Hugging Face Accelerate, incorporating weight decay, and explored three learning rates (1e-5, 2e-5, and 3e-5) along with training durations of 6, 10, and 12 epochs. The use of stratified cross-validation ensured that the training folds reflected the class distribution of the original dataset, a common strategy in multi-label learning tasks [5]. We initially conducted a grid search over three learning rates (1e-5, 2e-5, 3e-5) and epoch counts (6, 10, 12) using the standard AdamW optimizer. However, due to memory constraints in our Colab environment and instability observed with longer training cycles, we subsequently switched to the Paged AdamW optimizer provided by Hugging Face Accelerate, which supports gradient accumulation and mixed precision training. Under this configuration, we observed

that a slightly higher learning rate (2e-4) combined with a shorter cycle of 3 epochs yielded better generalization performance on validation folds and faster convergence. We empirically validated the change through multiple runs. We selected the final hyperparameters based on the best macro-averaged F1 Score obtained during cross-validation. Table 3 reflects these final settings, which diverge slightly from the initial grid due to optimizer-related changes and runtime constraints. We conducted training and evaluation using stratified 5-fold cross-validation [5] to ensure balanced label distributions across folds. Table 3 provides a summary of the final configuration and training parameters used in the best-performing model. We evaluated model checkpoints using macro-averaged F1-score across the three categories, consistent with the official task metric.

Table 3
Configuration and training parameters.

Parameter	Value
Epochs	3
Training batch size	2
Gradient accumulation steps	2 (to simulate larger batch size under Colab constraints)
Optimizer	Paged AdamW 32bit
Learning rate	2e-4

3.4. Threshold Calibration and Ensemble

We found that default sigmoid thresholds of 0.5 are suboptimal due to class imbalance and skewed prediction probabilities. To address this, we applied threshold tuning [6] per class using the precision-recall-curve function from sci-kit-learn, selecting the threshold that maximized the F1-score on each validation fold for each label independently. Finally, we constructed an ensemble [7] by combining the two best-performing models—one trained with class weighting and one without—using soft-voting over the predicted probabilities. This approach stabilized predictions and improved generalization on the test set. We implemented all model components using PyTorch and Hugging Face Transformers.

3.5. Implementation Details

All experiments were implemented in Python using Hugging Face's Transformers and Accelerate libraries along with PyTorch 2.0. The training was conducted in Google Colab using a single NVIDIA Tesla T4 GPU with 16 GB of VRAM. Each fold in the 5-fold cross-validation required approximately 35 minutes to complete, including both the training and validation phases. We used mixed-precision training (`fp16`) to reduce memory usage and accelerate computation. We executed the full training pipeline within Linux-based virtual environments, utilizing CUDA 11.8 support.

4. Results

We evaluated the model on the official test set (ct-test.tsv) provided by Task 4a in CheckThat! 2025. Final predictions were submitted via Codalab and assessed using the task's primary evaluation metric: macro-averaged F1-score across the three target categories [6].

4.1. Development Results

Table 4 presents the cross-validation performance on the training set using 5 folds. The ensemble model showed consistently higher F1-scores, especially for Category 1.

Table 4
5-Fold Cross-Validation Results on ct-train.tsv

Model Variant	C1 F1	C2 F1	C3 F1
DeBERTa (weighted)	0.8014	0.7812	0.7880
DeBERTa (non-weighted)	0.7923	0.7691	0.7755
Ensemble (soft voting)	**0.8133**	**0.7908**	**0.8009**

4.2. Official Evaluation Results

The final predictions on the test set submitted to the Codalab platform yielded the following macro-F1 scores:

- **Overall macro-F1:** 0.7983
- **Category 1 (C1):** 0.8133 (Ranked 1st among all participants)
- **Category 2 (C2):** 0.7841
- **Category 3 (C3):** 0.7976

Our system ranked **2nd overall** and achieved the highest score in the most critical category for scientific verification: scientifically verifiable claims (C1). These results demonstrate the effectiveness of our multi-stage fine-tuning and ensemble strategy, particularly in distinguishing scientific claims (Category 1), where our system achieved the highest ranking. All code, model checkpoints, and training scripts used in this study are publicly available at our GitHub repository [8]

4.3. Comparison to Baseline

Compared to the official baseline provided by the organizers, which achieved an overall F1 of 0.718, our approach improved macro-F1 by approximately 8%. Key factors contributing to this improvement include threshold calibration and ensemble integration.

5. Analysis of the Results

The results from the official evaluation reveal distinct performance patterns across the three classification categories. Our system achieved its highest F1-score in Category 1 (Scientific Claims), reaching 0.8133. This strong result likely stems from the model's capacity to identify explicit, verifiable assertions that align structurally and semantically with scientific discourse. Tweets in this category frequently feature recognizable linguistic markers—such as causal constructions, statistical phrasing, or study citations—that are effectively captured through contextual encoding by transformer architectures [5].

5.1. Performance per Category

The results obtained in the official evaluation reveal important distinctions in performance across the three target categories. We achieved the highest F1 score in Category 1 (Scientific Claims), with a system score of 0.8133. We can attribute this strong performance to the model's ability to capture explicit, verifiable assertions that are often more structurally and semantically aligned with scientific language. Many of these tweets contain indicative patterns—such as causal statements, statistical language, or references to studies—that benefit from contextual encoding in transformer architectures [5].

In contrast, Category 2 (References) yielded the lowest F1-score (0.7719) among the three. This category includes both direct and indirect references to scientific sources, which can be subtle, implicit, or dependent on external knowledge (e.g., recognizing that a link points to a scientific repository) [9]. Tweets in this category often present ambiguity or lack sufficient textual cues to signal a reference reliably. Since we trained the system solely on tweet text without leveraging metadata or external link resolution, it may have missed signals necessary to capture this label fully. The inclusion of explicit

features, such as DOIs, paper titles, or domain-specific cues, could improve classification precision in future iterations.

Category 3 (Research Context), with an F1-score of 0.8098, reflects a middle ground in difficulty. Tweets labeled under this class often refer to researchers, institutions, or ongoing studies without making verifiable claims or citing specific resources. The model's performance suggests it effectively learned to identify institutional or contextual phrases (e.g., "new study from," "Harvard researchers," "ongoing trials"), which likely served as discriminative features.

5.2. Effect of Calibration

The inclusion of threshold tuning per label, using precision-recall curves, had a notable positive effect, particularly in balancing recall and precision for the less-represented categories. Similarly, class-specific weighting via the pos-weight parameter helped reduce the bias toward the dominant no-label class ([0, 0, 0]), ensuring the model remained sensitive to minority classes.

5.3. Observed Limitations

While the system demonstrated overall robustness, several limitations persist. Tweets exhibiting sarcasm, fragmented phrasing, or lacking explicit linguistic cues proved especially difficult to classify reliably. In particular, short tweets that only contained links, emojis, or hashtags often lacked sufficient contextual signals to support confident categorization. These were especially problematic in distinguishing between Category 2 (Reference) and Category 3 (Context), as both may mention research implicitly but differ in intent.

Furthermore, Category 2 presented persistent ambiguity, as references to scientific knowledge were often made implicitly through hyperlinks or vague mentions (e.g., "see this" or "the study shows") without explicit citations or domain indicators. Since our system relied solely on tweet text and excluded metadata or link resolution, it struggled to infer whether such tweets truly referenced scientific sources. For example, the lack of URL domain analysis (e.g., arxiv.org, pubmed.ncbi.nlm.nih.gov) limited the model's ability to detect indirect references.

Additionally, our approach did not incorporate conversational or user-level features—such as quote tweets, thread continuity, or reply context—which could clarify meaning in multi-tweet exchanges. As a result, label predictions occasionally failed to capture implicit discourse continuity. Future iterations may benefit from integrating URL resolution, external knowledge bases, or discourse-aware modeling strategies to improve robustness in real-world social media scenarios.

6. Perspectives for Future Work

The competitive performance achieved in Task 4a highlights clear directions for enhancing the system's capabilities. While the current approach relies solely on tweet text and supervised fine-tuning, future iterations can integrate architectural, contextual, and multilingual improvements. Below, we outline four specific enhancements, their rationale, and the implementation paths.

1. Integration of Resolved URLs and Domain Signals. Many tweets in Category 2 implicitly reference scientific content through shortened links. To address this, we plan to incorporate a URL resolution pipeline using tools such as `newspaper3k` or `tldextract`, which enables the extraction of domain names (e.g., `pubmed.ncbi.nlm.nih.gov`, `arxiv.org`) and page metadata. We can encode these features as auxiliary embeddings or categorical indicators within the model. Implementation will require API access and preprocessing routines for link expansion.

2. Discourse-Aware and Contextual Modeling. The current model treats each tweet independently, which limits its ability to capture implicit context. Future versions will utilize discourse structures, such as tweet threads, quote-retweets, and user reply chains, via the Twitter API. We aim to explore hierarchical models (e.g., HierBERT) and contrastive learning approaches to encode inter-tweet dependencies better.

3. Multilingual Adaptation. Given the global nature of science communication, extending the classifier to support additional languages is a priority. We plan to adapt multilingual transformer models such as XLM-R and mDeBERTa, with a focus on Spanish and Portuguese. It will involve collecting a parallel annotated corpus following the SciTweets schema, possibly through partnerships with fact-checking organizations in Latin America or through crowdsourcing platforms.

4. Explainability and Transparency. To foster model interpretability, we intend to integrate explainability tools such as BertViz, exBERT, SHAP, or Integrated Gradients. These tools can visualize attention weights or saliency scores, helping users understand model decisions—particularly in ambiguous or borderline cases. We release all code, model checkpoints, and training logs as open-source via GitHub and track them through Weights & Biases. These efforts aim not only to enhance classification performance but also to support reproducibility, trust, and community collaboration in scientific discourse detection.

7. Acknowledgments

We would like to thank the organizers of the CheckThat! Lab at CLEF 2025 for providing a well-structured and valuable evaluation framework. This work was supported in part by the Universidad Tecnológica de Bolívar and its School of Digital Transformation. We also acknowledge the collaborative efforts of the VerbaNexAI team for their dedication to advancing scientific discourse detection and reproducibility in social media NLP research. We extend special thanks to the Colombian Navy, through its Naval Education Command (Jefatura de Educación Naval) and the Naval Technological Development Center (Centro de Desarrollo Tecnológico Naval), for their financial support and institutional endorsement of this research. We will make the system and code described in this paper publicly available to promote reproducibility and community collaboration.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 and Grammarly for grammar and style revision. All outputs were carefully reviewed and edited by the authors, who take full responsibility for the final 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. ECIR 2025. Lecture Notes in Computer Science, volume 15576, Springer Nature Switzerland, Cham, 2025, pp. 467–478. doi:10.1007/978-3-031-88720-8_68.

[2] S. Hafid, S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! lab task 4 on scientific web discourse, in: Working Notes of CLEF 2025, 2025.

[3] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: Proceedings of the 9th International Conference on Learning Representations (ICLR), OpenReview.net, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[4] Z.-H. Zhang, M. R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in: Advances in Neural Information Processing Systems 31 (NeurIPS), Curran Associates, Inc., 2018, pp. 8792–8802.

[5] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.

[6] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLOS ONE 10 (2015) e0118432. doi:10.1371/journal.pone.0118432.

[7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems 32 (NeurIPS), Curran Associates, Inc., 2019, pp. 5754–5764. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[8] M. J. Sosa-Borrero, J. E. Serrano, J. C. Martinez-Santos, E. Puertas, Verbanexai lab at checkthat! 2025: Scientific discourse detection in tweets – code repository, https://github.com/VerbaNexAI/CLEF2025/tree/main/CheckThat, 2025. Accessed: 2025-05-29.

[9] S. Hafid, S. Schellhammer, S. Bringay, K. Todorov, S. Dietze, Scitweets: A dataset and annotation framework for detecting scientific online discourse, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM), ACM, 2022, pp. 3988–3992. doi:10.1145/3511808.3557516.