JU_NLP at CheckThat! 2025: Leveraging Hybrid Embeddings for Multi-Label Classification in Scientific Social Media Discourse*

Arpan Majumdar[1], Dipankar Das[2] and Pritam Pal[2]

[1]University of Kalyani, West Bengal, India
[2]Department of Computer Science and Engineering, Jadavpur University, West Bengal, India

Abstract

With the abundance of scientific information and misinformation propagated on social media, especially on Twitter, there is an urgent need for automated systems to classify science-related information. The CLEF-2025 Task 4a [1] of the CheckThat! Lab is intended to detect three types of scientific discourse found in tweets: scientific claims, mentions of scientific studies, and references to scientific entities. In this paper, we present a hybrid transformer-based solution which utilizes SciBERT, a model of scientific-text training, and Twitter-RoBERTa, a specifically-tuned protocol for tweets. By taking advantage of the strengths of both models we capture the formal, technical aspects of scientific language and the informal, noisy conventions of social media discourse. The pooled embeddings from both encoders are concatenated and fed through a multilayer classification head that incorporates dropout as regularization and sigmoid activation for multilabel prediction. The model was trained using binary cross-entropy loss and incorporated early stopping and adaptive learning rate scheduling. Our model achieved a macro-averaged F1 score of 0.8262 on the development set, as well as a minimum validation loss of 0.1744. These results provide evidence for the benefit of hybrid pretraining for scientific tweet classification and provide a foundation for future extensions, which could incorporate relationships between labels and multilingual aspects.

1. Introduction

As the shift of scientific communication continues to shift to social media, the ability to automatically find and analyze science-related content has become a necessary tool for researchers, journalists, and fact-checkers (simply referred as ”users” in this task). Within Twitter, there is a unique mix of public conversations involving scientific claims, and informal commentary, news, and misinformation, creating a complex environment. The CheckThat! Lab at CLEF-2025 could see that demand for tools that detect particular forms of discourse and introduced Task 4a - an attempt to detect three forms of science web discourse in individual tweets. The three forms of science web discourse consist of scientific claims, references to scientific studies, and mentions of scientific entities like researchers and institutions.

Due to the interplay between two very different forms of language - the formal and precise language of scientific publications and the informal and context-dependent language of tweets - we acknowledge it will be difficult for classification models, since the amount of variety of any form of language is beyond the best of traditional text classification models. So, in this experiment, we introduce a hybrid solution that exploits the complementary strengths of two pretrained transformer models - SciBERT [2], pretrained on a large collection of scientific publications, and Twitter-RoBERTa [3], fine-tuned for social media text.

We predict that an appropriate combination of these two embeddings will allow the model to better navigate both the protocols present in scientific claims as well as the distinctive language of Twitter. We propose a very simple and effective architecture that gets the pooled embeddings from each encoder and simply concatenates them as inputs to a classification head for predicting the multilabel outputs.

CLEF 2025 Working Notes, 9–12 September 2025, Madrid, Spain
*Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, 9–12 September 2025, Madrid, Spain.
arpanmajumdar952@gmail.com (A. Majumdar); dipankar.dipnil2005@gmail.com (D. Das); pritampal522@gmail.com (P. Pal)

The system is trained end-to-end using binary cross-entropy loss and optimized using early stopping based on the validation performance. This manuscript describes the details of the dataset, the model architecture, the training strategy, the evaluation results, and qualitative observations, and culminates with a limitations and future work section.

2. Dataset Description

The data used for CLEF-2025 Task 4a is an annotated version of the SciTweets corpus, whose tweets have been annotated for scientific discourse. The data is provided in three subsets: a training set with 1,229 labeled tweets, a development set (137 tweets), and a test set (240 tweets). Within each tweet are three separate binary labels, which identify three independent tasks: labeling whether the tweet contains a scientific claim, refers to a scientific study or publication, and identifies a scientific entity (such as a university or scientist).

Given that this is a multilabel dataset, any given tweet may contain none, one, or multiple labels. The training set has a fairly balanced distribution of the three labels, which means the model cannot simply learn to identify an arbitrary label based on counting approaches. The annotation developed for the study relies on the SciTweets framework and provides clear instructions to labelers to achieve consistency in labeling each following examples. Thus, this dataset presents a realistic and challenging scenario that models labeled for multilabel classification [4] can be developed to navigate nuanced discourse in a scientific context across social media.

3. Related Work

Recent advances in transformer models have significantly improved scientific text analysis and discourse classification. SciBERT [5], a BERT variant pretrained on scientific publications, demonstrates superior performance in capturing scientific language nuances for downstream scholarly NLP tasks. This success extends to other domain-specific models like BioBERT and LegalBERT, establishing transformer-based architectures with specialized pretraining as the standard approach in scientific NLP.

For social media discourse analysis, specialized models have emerged to address platform-specific challenges. BERTweet [6] targets Twitter-specific language patterns, while the TweetEval benchmark consolidates seven heterogeneous tweet classification tasks, reporting strong performance from transformer models further pretrained on Twitter corpora. The SciTweets dataset [7] specifically addresses scientific discourse on Twitter, introducing annotation guidelines for science-related tweets and achieving approximately 89% F1-score in distinguishing scientific content using domain-specific features.

Recent shared tasks provide additional context for scientific discourse detection across platforms. The CLEF CheckThat! lab [1] featured tasks on identifying check-worthy scientific claims in tweets and news articles. Participants predominantly employed transformer architectures including multilingual RoBERTa, XLM-RoBERTa, and GPT variants, consistently outperforming heuristic baselines. These results reinforce that pretrained transformers constitute robust foundations for scientific discourse classification tasks.

Our approach to model efficiency draws inspiration from O'Neill et al.'s layer fusion technique [8], which identifies and merges similar layers in pretrained networks to achieve structured compression. Their experiments demonstrated that transformers could be reduced to approximately 20% of their original size while maintaining performance, with only minimal perplexity increases. We adapt this methodology by aligning and fusing analogous transformer layers, preserving inter-layer information while creating more compact models. Although fusion-based compression remains underexplored in discourse classification, it offers a principled framework for deploying large transformer models efficiently in scientific discourse detection tasks.

4. Methodology

In order to be able to pragmatically address the issue of scientific discourse classesinee on Twitter, we are proposing a hybrid transformer-based architecture that exploits domain-specific and social-contextual language to inform its understanding. Further, this framework aims to leverage the two materials - SciBERT and the Twitter-RoBERTa - by separately combining via late fusion layer[9], followed by a deeper classification head for multi-label predictions.

At the core of our proposal lies a dual-encoder structure. We first preprocess the tweets using the typical NLP processes such as tokenization, lowercasing, truncation etc., to create the input required for the transformer models. Each input tweet results in two representations, where the tweets pass through two encoder models separately. SciBERT is models trained on scientific literature and therefore offers coherent representations of domain-specific vocabularies as well as formal discourse. In parallel, Twitter-RoBERTa parses a large volume of tweets capturing informal, noisy, and user generated text identified to social media contexts.

From both models, we obtain the contextualized [CLS] token embedding which is a compressed semantic summary of the input text. We then concatenate these two 768-dimensional vectors into a 1536-dimensional combined representation. This vector serves as the fusion layer combining the benefits of the two encoders to account for the domain relevance and social context.

We then send the fused representation through a classification head which consists of two fully connected (FC) layers separated by dropout layers to mitigate overfitting. The first FC layer reduces the dimension from 1536 to 512 which is subsequently followed by a ReLU activation and a dropout layer (p = 0.3). The second FC layer then reduces the representation to 128 dimensions, followed by ReLU and dropout. The last layer is the output layer, which applies activation functions Sigmoid, to produce three probabilities associated with the labels, scientific, non-scientific, and uncertain. The sigmoid allows for multi-label outputs which is a necessity in this task, where a tweet can fall into multiple categories.

The model is trained with binary cross-entropy loss with early stopping based on the validation macro F1-score, assuring convergence and generalization on unseen data.

5. Results

Our model trained cleanly and had a strong performance on the development set. The best model (by early stopping) had a validation loss of 0.1744 and a macro-averaged F1 score of 0.8262. The validation loss and F1 score on the development set conveys the model's generalization across all three label categories, without overfitting. Although the actual loss curve figure could not be included due to submission constraints, we observed a clear downward trajectory in both training and validation losses across epochs. The model began converging steadily after the third epoch, and early stopping was triggered around the sixth epoch when validation loss plateaued. The consistent correlation between training and validation losses indicates that the model learned meaningful features without overfitting. correlated until they converged.

Although our model performed well overall, the category-wise results reveal some variation in label-specific performance. For instance, on the development set, the model achieved F1-scores of 0.7451, 0.8302, and 0.9014 for Category 1 (Scientific Claim), Category 2 (Study Reference), and Category 3 (Scientific Entity), respectively. This indicates stronger performance in identifying entities compared to claims. Similarly, the test set shows a drop in Category 2 (Study Reference) with an F1 of 0.5965, highlighting this as a relatively harder category. These results suggest label-specific challenges that future work could address through additional training signals or class balancing strategies.

6. Observation

During training, we found that the model had stable convergence because both the training and validation losses steadily decreased, indicating that the model achieved some learning. The validation

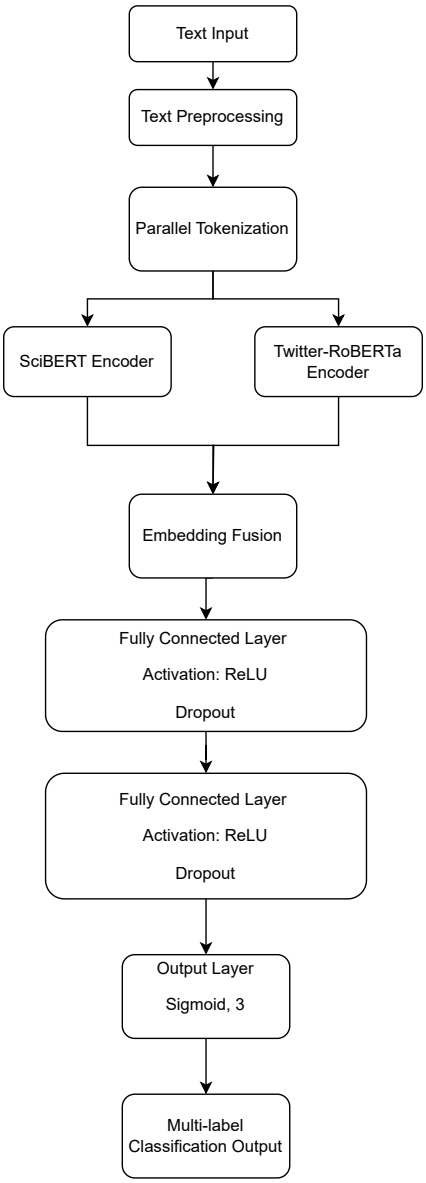

Figure 1: Model diagram illustrating the architecture.

Table 1
Training and validation performance at the best epoch

Train Loss	Validation Loss	Macro F1-Score
0.1136	0.1744	0.8262

loss plateaued mid-way through the training process, approximately after the 6th epoch, which was when our early stopping was triggered. With early stopping, we could ensure that the model did not overfit the training data.

It was also quite encouraging to see that the model's macro-F1 score remained above 0.80 during the early epochs and the final score was 0.8262. This further amplifies the importance of concatenation of embeddings between SciBERT and Twitter-RoBERTa, which provided sufficient features to the model that would allow it to learn language features across a much broader feature space.

While experimenting with the classification head, we also made adjustments to the structure of the classification head by performing tests with varying dropout probabilities and number of units in the

Table 2
Performance of our system on the development and test sets for CLEF-2025 Task 4a.

Set	Macro F1	Cat 1 F1	Cat 2 F1	Cat 3 F1	Rank
Development Set	0.8256	0.7451	0.8302	0.9014	9
Test Set	0.7347	0.7881	0.5965	0.8195	9

fully connected layers. We found that two fully connected layers (512 intermediate units) and a dropout probability of 0.5 provided the best trade-off between capacity and overall generalization.

7. Conclusion

This research introduced one possible hybrid model to detect scientific discourse in tweets, specifically for the challenges of the CLEF-2025 Task 4a. Our two-encoder model successfully integrates the domain knowledge common to nearly all forms of scientific texts in SciBERT, and the social-language knowledge of Twitter-RoBERTa to classify tweets into three different scientific subject categories. The metrics - especially the macro-F1 of 0.8262 and validation loss (0.1744) - mean that this dual-encoder model has a promising, promising direction for future scientific content classification work on social media.

The primary value of our model is its connection between two distinct language domains, formal language of science and informal language of social media. As a result, it is appropriate for this particular issue, but potentially could be applied to another challenge related to domain adaptation or mixed register text classification.

8. Limitations and Future Work

Although the performance of our approach is promising, we recognize multiple limitations. First, the model treats each of the three labels as independent. In practice, however, such categories are often inter-dependent; for example, a scientific claim will often come with a reference to a study. To improve performance, it would be beneficial to introduce label inter-dependencies in the approach via conditional random fields, or a graph.

Second, the model is currently trained on English tweets. As scientific discussions are becoming increasingly hybrid and multilingual, we could extend the approach to either develop multilingual or cross-lingual transformer models such as XLM-RoBERTa, or multilingual and cross-lingual research studies more generally, where we could approach the same task in a more language-agnostic manner.

A related improvement could include utilizing larger variants of models such as SciBERT-large, or Twitter-RoBERTa-large, if compute was not an issue and better performance could be achieved due solely to a model artifact. Lastly, the model could also be thought better if it was trained with additional auxiliary information- such as tweet metadata, timestamps, or even contents of the articles they link to- that may provide additional context that could improve classification.

In light of these limitations, it should be recognized that this paper provides a framework, and the experience and potential for continued research in the area of scientific discourse detection on social media is perspicuous.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools including **ChatGPT**, **Claude**, and **Grammarly** to support writing tasks such as grammar correction, sentence rephrasing, and improving clarity. Specifically, these tools were used in drafting the *Abstract*, *Related Work*, and *Conclusion* sections. All AI-generated content was carefully reviewed and edited by the authors. No AI tool was listed as an author, and the authors take full responsibility for the accuracy and integrity of the final manuscript.

References

[1] S. Hafid, Y. 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, ????

[2] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, 2019. URL: https://arxiv.org/abs/1903.10676. arXiv:1903.10676.

[3] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1644–1650. URL: https://aclanthology.org/2020.findings-emnlp.148. doi:10.18653/v1/2020.findings-emnlp.148.

[4] J. Bogatinovski, L. Todorovski, S. Džeroski, D. Kocev, Comprehensive comparative study of multi-label classification methods, Expert Systems with Applications 203 (2022) 117215. URL: https://www.sciencedirect.com/science/article/pii/S0957417422005991. doi:https://doi.org/10.1016/j.eswa.2022.117215.

[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), 2019, pp. 3615–3620.

[6] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[7] H. Hafid, N. Alghamdi, M. Alshehri, Scitweets: A dataset and classifier for detecting scientific discourse on twitter, arXiv preprint arXiv:2206.07360 (2022).

[8] J. O'Neill, S. J. Delany, B. MacNamee, Model compression for bert using layer fusion, arXiv preprint arXiv:2012.13136 (2020).

[9] J. O. Neill, G. V. Steeg, A. Galstyan, Compressing deep neural networks via layer fusion, 2020. URL: https://arxiv.org/abs/2007.14917. arXiv:2007.14917.