

Building a Compact Math Corpus

Andrea Ferreira

Independent Researcher

andreafer.uni@gmail.com

Abstract

This paper introduces the Compact Math Corpus (CMC), a preliminary resource for natural language processing in the mathematics domain. We process three open-access undergraduate textbooks from distinct mathematical areas and annotate them in the CoNLL-U format using a lightweight pipeline based on the spaCy Small model. The structured output enables the extraction of syntactic bigrams and TF-IDF scores, supporting a syntactic-semantic analysis of mathematical sentences.

From the annotated data, we construct a classification dataset comprising bigrams potentially representing mathematical concepts, along with representative example sentences. We combine CMC with the conversational corpus UD English EWT and train a logistic regression model with K-fold cross-validation, achieving a minimum macro-F1 score of 0.989. These results indicate the feasibility of automatic concept identification in mathematical texts.

The study is designed for easy replication in low-resource settings and to promote sustainable research practices. Our approach offers a viable path to tasks such as parser adaptation, terminology extraction, multiword expression modeling, and improved analysis of mathematical language structures.

1 Introduction

Mathematical textbooks, though precise and structured, present unique challenges to standard Natural Language Processing (NLP) tools. Their language differs significantly from general-domain English, incorporating symbolic notation, diagrams, and domain-specific terminology. Consequently, models trained primarily on non-technical corpora often underperform on this type of texts.

Recent benchmarks such as MATHVISTA (Lu et al., 2024) illustrate these challenges. Even advanced vision-language models, for example: GPT-4V, achieve accuracy in the range 50%

when tasked with understanding mathematical content (see Figure 3, Appendix A). Meanwhile, datasets including GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) focus on mathematical problem solving, providing question-answer pairs, but lack the linguistic annotations necessary for syntactic or semantic analysis.

Other resources have been introduced for the processing of the mathematical language. An example is NaturalProofs (Welleck et al., 2021), which focuses on theorem proving and alignment of formal and informal proofs, but does not address the broader expository writing found in instructional or pedagogical texts. This scarcity limits the development and evaluation of NLP tools tailored for mathematical language. The **Compact Math Corpus (CMC)** aims to help bridge this gap by offering a preliminary, automatically annotated resource, not a gold standard, but a practical starting point for linguistic processing in this domain.

The CMC is built from three open-access undergraduate textbooks: *Abstract Algebra: Theory and Applications* (Judson, 2022), *Linear Algebra* (Hefner, 2022), and *Discrete Mathematics: An Open Introduction* (Levin, 2024), all sourced from the Open Math Textbook Initiative¹. Using a computationally lean NLP pipeline based on the spaCy Small model (Honnibal et al., 2020), we automatically annotate these texts in the CoNLL-U format², capturing both morphological and syntactic features.

Designed for low-resource and sustainable settings (Luccioni et al., 2023), our pipeline prioritizes accessibility and replicability. To assess its utility, we extract syntactic bigrams from CMC, combine them with the conversational UD English EWT corpus (Silveira et al., 2014), and pair each bigram with a representative sentence from its source cor-

¹<https://textbooks.aimath.org/>

²<https://universaldependencies.org/>

pus. We then train a logistic regression model using scikit-learn’s `LogisticRegression` with K-fold cross-validation. The model achieves a minimum macro-F1 score of 0.989, indicating that cost-effective methods can effectively support the detection of mathematical concepts.

We release the annotated corpus and supporting materials on GitHub³.

2 Corpus Preprocessing and Annotation

The Compact Math Corpus (CMC) was constructed from the three undergraduate-level textbooks introduced in the previous section, each covering a distinct area of mathematics. All texts are openly licensed and available in PDF format.

Although \LaTeX is generally a more suitable format for mathematical texts due to its richer structural markup (Collard et al., 2024; de Paiva et al., 2023), most educational materials are distributed in PDF format, which aligns better with our goal of scalability.

To better understand the trade-offs involved, we processed a matched section from the *Linear Algebra* textbook in both \LaTeX and PDF formats. The source \LaTeX was converted to JSON using `pylatexenc`⁴, and the PDF using `PyMuPDF`⁵. Both outputs were then passed through the same annotation pipeline, and we extracted compounds⁶ from the resulting CoNLL-U files.



Figure 1: Overlap of Compounds Between \LaTeX and PDF.

Of the total compounds detected, 23 appeared in both formats, 6 were exclusive to \LaTeX , and 15 were exclusive to PDF (see Figure 1). These results

³<https://github.com/andreafer-uni/Compact-Math-Corpus>

⁴<https://pylatexenc.readthedocs.io/>

⁵<https://pymupdf.readthedocs.io/>

⁶In UD, a *compound* refers to a noun–noun construction where one noun modifies another.

indicate that, despite some discrepancies, PDF-based processing yields comparable compound extraction quality, an encouraging outcome given the prevalence and accessibility of PDF materials in educational settings.

2.1 Preprocessing

Following best practices in corpus development (Collard et al., 2024), PDF textbooks were converted to structured JSON to support consistent downstream analysis.

Linguistic content was extracted using the `PyMuPDF` library, which provides raw text and layout information. Since PDFs prioritize visual formatting over semantic structure, extraction introduced common issues, including token merging, incorrect sentence segmentation, hyphenation artifacts, and irregular or misplaced line breaks.

To address these concerns, we applied a preprocessing pipeline with anomaly detection and cleaning steps, including sentence filtering, non-ASCII character removal, and format normalization. The cleaned text was stored in JSON and parsed using `spaCy` `Small` to generate part-of-speech and dependency annotations in CoNLL-U format⁷.

This setup enables us to evaluate the performance of general-purpose NLP tools on math-heavy texts and points to challenges such as symbolic content, domain-specific terminology, and structural noise, areas where parser adaptation or multimodal integration may improve outcomes.

2.2 Linguistic Annotation

To assess the performance–efficiency balance of our method, we compared `spaCy`’s `small` model (`en_core_web_sm`) with its transformer-based counterpart (`en_core_web_trf`) on a section of the *Linear Algebra* textbook.

As shown in Figure 2, both models produced nearly identical counts of tokens, lemmas, and unique words. The main difference was in sentence segmentation, where the transformer model generated more sentence boundaries. However, this did not affect the performance of our downstream classification task, supporting the adequacy of the compact model given its streamlined computational requirements.

⁷See (Nivre et al., 2016) for a complete overview of the UD framework and CoNLL-U structure.

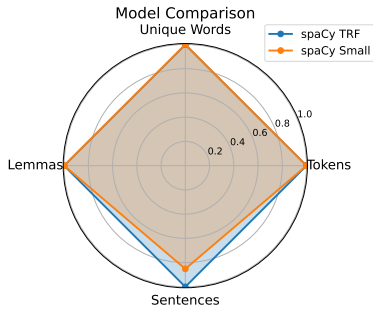


Figure 2: Sentence segmentation differences between transformer-based and small models.

To enhance the linguistic value of the annotation process, the use of the CoNLL-U format provides substantial linguistic benefits. In contrast to unstructured text, syntactically annotated corpora enable the systematic recognition of domain-specific constructions such as compounds and multiword expressions (MWEs), which act as indicators of domain-specific language (Collard et al., 2022).

As highlighted in Table 1, the integration of syntactic information, specifically the dependency relations captured by Universal Dependencies (UD), increased the prominence of mathematically relevant bigrams in the Compact Math Corpus (CMC). For example, the terms *vector space*, *linear combination*, and *closed formula* not only received high TF-IDF scores after annotation, but also aligned well with the core mathematical concepts.

Before CoNLL-U		After CoNLL-U	
Bigram	TF-IDF	Bigram	TF-IDF
vector space	1328.53	vector space	1957.41
closed formula	834.34	closed formula	1105.03
linear combination	726.46	linear combination	1008.42
recurrence relation	668.48	recurrence relation	831.47
bit string	585.78	bit string	804.58

Table 1: Top 5 TF-IDF bigrams in the Compact Math Corpus before and after CoNLL-U annotation.

The syntactic layer, though often optional in modern NLP pipelines, remains crucial for applications that prioritize interpretability and robustness. As noted by Sag et al. (Sag et al., 2002), the combination of symbolic and statistical methods produces a more complete view of language, particularly in detecting MWEs that define mathematical vocabulary.

Although the CMC is not manually annotated,

its syntactic enrichment via spaCy and CoNLL-U enables a hybrid pipeline in which frequency-based methods such as TF-IDF are grounded in structural patterns. This becomes evident when we compare the results before and after annotation. After syntactic parsing, high-ranking terms became more semantically coherent, whereas noisy entries (e.g. *many way*) were relatively de-emphasized.

Together, the CoNLL-U annotation process enhances the linguistic utility of the CMC by facilitating the extraction of interpretable, conceptually grounded MWEs, despite being derived from non-pretrained model. This makes the resource not only computationally efficient but also linguistically rich enough to support downstream tasks including classification and terminology extraction.

3 Dataset Construction

To investigate whether syntactic bigrams can help distinguish mathematical language from general English, we constructed a binary classification dataset combining the Compact Math Corpus (CMC) with the UD English Web Treebank (EWT), leveraging their shared CoNLL-U format.

3.1 The UD-EWT

The UD English Web Treebank (EWT) is a gold-standard corpus that provides syntactic and morphological annotations for English as part of the Universal Dependencies (UD) project. It contains informal web-based texts from blogs, emails, forums, product reviews, and Q&A websites such as Yahoo! Answers.

Due to its conversational and general-domain nature, EWT serves as a useful baseline for comparison with domain-specific corpora such as CMC. Since both CMC and UD-EWT follow the same annotation format, we used this compatibility to extract and compare syntactic features between corpora.

3.2 Bigrams and Labeling

To assess the potential of syntactic bigrams for the recognition of mathematical concepts, we first extracted bigrams from UD-EWT using the same TF-IDF methodology applied to CMC. We then compared the resulting lists and removed all overlapping bigrams, retaining only the unique ones from each corpus.

This comparison produced two distinct sets of bigrams: one containing candidates for mathemat-

ical concepts (unique to CMC) and another reflecting general English patterns (unique to EWT). For each bigram, we retrieved representative sentences from their respective corpora in which the bigrams occurred. These sentences were labeled as `True` (mathematical concept) or `False` (non-mathematical), resulting in a dataset of 2,796 sentences evenly balanced between mathematical and general-domain content.

The resulting dataset forms the basis for the classification task described in the next section, where we evaluate the feasibility of concept recognition using a logistic regression model.

4 Identifying Mathematical Concepts

To evaluate whether syntactic bigrams can effectively signal mathematical content, we framed a binary classification task: Given a sentence containing a candidate bigram, predict whether it expresses a mathematical concept. The classifier operates solely on text-based features, without access to labels or metadata from the source corpora.

4.1 Model and Setup

We trained a logistic regression model using TF-IDF features over unigrams and bigrams (max features = 5,000), implemented with `scikit-learn` (Pedregosa et al., 2011). We opted for logistic regression due to its efficiency, interpretability, and reliable performance in sparse feature spaces, well-suited to our minimal NLP setup.

4.2 Performance and Evaluation

When we evaluated the logistic regression model using 10-fold cross-validation over the dataset, the classifier achieved a macro F1-score of 0.996 ± 0.003 , indicating consistent performance across all folds. Table 2 reports precision, recall, and F1-scores per class.

Although performance is reliable within the labeled dataset, it is important to note that the model relies on sparse, surface-level features and may struggle with ambiguous or out-of-distribution cases. However, it correctly recognized sentences containing previously unseen bigrams, indicating that the model generalizes based on sentence context rather than simple memorization.

Metric	Mean	Std
Accuracy	0.9964	0.0034
Precision_0	0.9934	0.0072
Recall_0	0.9993	0.0023
F1_0	0.9963	0.0035
Precision_1	0.9993	0.0022
Recall_1	0.9938	0.0069
F1_1	0.9965	0.0033
F1_Macro	0.9964	0.0034

Table 2: Cross-validation results (mean \pm std) for logistic regression classifier.

4.3 Error Analysis and Interpretability

The confusion matrix (see Appendix B, Figure 4) confirms the model’s consistent performance: only three mathematical examples were misclassified as non-mathematical, and no false positives were observed.

Zero-shot evaluations on held-out examples (Appendix C, Table 3) revealed a similar generalization capacity. The model correctly labeled unseen concepts such as *probability distribution* and *integral calculus* with high confidence, while remaining uncertain in borderline cases. An inspection of errors showed that school-related phrases such as *school project* and *homework folder* introduced ambiguity due to their lexical proximity to educational and mathematical contexts.

We further explored the behavior of the models through feature weight analysis (Appendix D, Figure 5). Some high-weight features aligned with intuitive mathematical concepts. However, others reflected corpus-specific statistical artifacts, terms *email* or document identifiers that lacked semantic relevance. This contrast illustrates how statistical models often rely on distributional regularities that may not align with human notions of meaning, underscoring the gap between symbolic interpretability and statistical association.

5 Conclusions and Future Directions

Summary of Findings. This paper presented the Compact Math Corpus (CMC), a syntactically annotated resource built from open-access instructional materials and processed using a lightweight NLP pipeline based on `spaCy Small` and the `CoNLL-U` format. Our goal was to enable structured linguistic analysis of mathematical language in a format that supports downstream applications

such as terminology extraction and concept classification.

Through a comparison of TF-IDF bigrams before and after annotation, we confirmed that syntactic information enhances the identification of multiword expressions aligned with core mathematical concepts. Furthermore, we constructed a balanced dataset by combining CMC with UD-EWT and found that a non-neural approach was able to effectively distinguish mathematical from general-domain content using only text-based features.

These findings indicate that even a non-pretrained, resource-efficient model can accurately detect domain-specific content, provided the input is linguistically enriched and well balanced. The approach is computationally efficient and interpretable, making it a suitable framework for educational or resource-constrained NLP scenarios. Although more powerful models may improve generalization on ambiguous input, our results provide evidence that structured features like those in CoNLL-U can support accurate concept classification in controlled settings.

Future Directions in Mathematical NLP Future work will explore expanding the CMC with additional textbooks across a broader range of mathematical topics, as well as refining the annotation pipeline to improve linguistic coverage.

Furthermore, the classification task can be extended by incorporating richer context (e.g., surrounding sentences or section-level cues) and testing generalization on unseen technical or scientific domains. Developing a small-scale gold-standard evaluation set with human-labeled concept phrases could further support benchmarking and model calibration.

An open direction for future research involves integrating symbolic representations, such as formulas or equation labels, with the current linguistic layer remaining an open direction. Despite being outside the scope of this initial study, such integration would enable a more comprehensive modeling of mathematical discourse, combining syntactic structure with symbolic reasoning.

6 Limitations

The main limitations of this study stem from its resource-constrained setup and reliance on general-purpose tools not specialized for mathematical language. The use of the `spaCy Small` model,

while efficient and accessible, introduces trade-offs in parsing accuracy, particularly for domain-specific terminology, symbolic expressions, and non-standard syntactic constructions typical of mathematical discourse.

Another concern involves the quality of the input data. Although the CoNLL-U format assumes clean, well-formed text, most of our source material originated from PDFs, which are optimized for visual layout rather than semantic structure. This introduces common issues such as token merging and sentence-boundary errors. Despite preprocessing steps mitigated some of these problems, they did not completely eliminate noise.

A further challenge lies in the lack of a domain-specific, human-annotated gold standard. Without a reliable reference for comparison, it is difficult to measure parsing quality or validate the accuracy of syntactic analyses. This restricts our ability to compare tools rigorously or perform fine-grained error analysis. Creating a gold-standard treebank for mathematical texts within the CoNLL-U framework remains an open direction for future work.

Although compact models promote reproducibility and sustainability, they may underperform in complex linguistic contexts where transformer-based alternatives would offer greater accuracy. In our case, the simplicity of the classification task helped offset this topic, but it remains a relevant factor when considering more advanced downstream applications.

Acknowledgments

This work was inspired by Valeria de Paiva’s leadership in the Network Mathematics project, which provided both direction and intellectual grounding.

We are also grateful to Jacob Collard for his earlier work on the mathematical topic, which offered valuable perspective throughout this study.

We acknowledge the American Institute of Mathematics for its exemplary initiatives to encourage the adoption of open-source and open-access textbooks in mathematics, an essential step toward equity and reproducibility in mathematical education and research.

Special thanks go to Diego Frassinelli, whose dedication to teaching and thoughtful guidance reflect the values of true academic mentorship.

References

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- Jacob Collard, Valeria de Paiva, Brendan Fong, and Eswaran Subrahmanian. 2022. [Extracting mathematical concepts from text](#). *Preprint*, arXiv:2208.13830.
- Jacob Collard, Valeria de Paiva, and Eswaran Subrahmanian. 2024. [Mathematical entities: Corpora and benchmarks](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11080–11089, Torino, Italia. ELRA and ICCL.
- Valeria de Paiva, Qiyue Gao, Pavel Kovalev, and Lawrence S. Moss. 2023. [Extracting mathematical concepts with large language models](#). *Preprint*, arXiv:2309.00642.
- Jim Hefferon. 2022. *Linear Algebra*, 4 edition. Orthogonal Publishing L3C. Available online at <https://hefferon.net/linearalgebra/>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). *CoRR*, abs/2103.03874.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spacy: Industrial-strength natural language processing in python](#). Project homepage: <https://spacy.io>.
- Thomas W. Judson. 2022. *Abstract Algebra: Theory and Applications*, annual edition 2022 edition. Stephen F. Austin State University. Available online at <http://abstract.pugetsound.edu/download/aata-20220728-sage-9.6.pdf>.
- Oscar Levin. 2024. *Discrete Mathematics: An Open Introduction*, 4 edition. Self-published. Available online at <https://discrete.openmathbooks.org/>.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *International Conference on Learning Representations (ICLR)*.
- Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. [Estimating the carbon footprint of bloom, a 176b parameter language model](#). *Journal of Machine Learning Research*, 24(253):1–15.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. [Scikit-learn: Machine learning in python](#). *Journal of Machine Learning Research*, 12:2825–2830.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. [Multiword expressions: A pain in the neck for nlp](#). In *Conference on Intelligent Text Processing and Computational Linguistics*.
- Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. [A gold standard dependency corpus for english](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 2897–2904. European Language Resources Association (ELRA).
- Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. [Naturalproofs: Mathematical theorem proving in natural language](#). *Preprint*, arXiv:2104.01112.
-

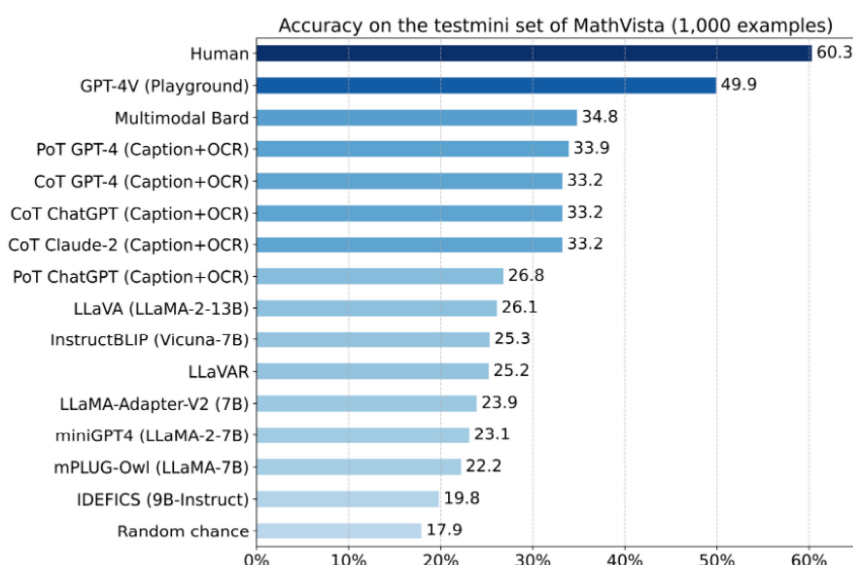
Appendix

This appendix provides supplementary material, including detailed linguistic annotations in the CoNLL-U style, as referenced in the main text.

A MathVista Experiment Results

The following graph comes from the MATHVISTA paper website at <https://mathvista.github.io/>.

Results on Existing Foundation Models




Accuracy scores of primary baselines on the testmini subset (1,000 examples) of  MATHVISTA. Both CoT GPT-4 and PoT GPT-4 are augmented with Bard captions and OCR text.

Figure 3: Foundation model performance on MathVista visual reasoning tasks.

B Confusion Matrix

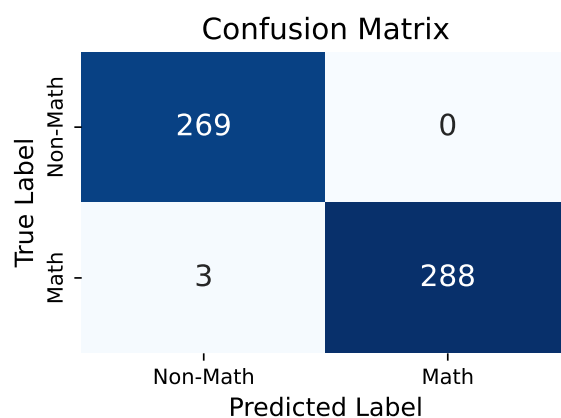


Figure 4: Confusion matrix for the classification task. The model correctly classifies most instances, with only a small number of false negatives, indicating effective performance in distinguishing mathematical from non-mathematical concepts.

C Zero Shot Predictions

Bigram	Sentence	Predicted Class	Probability
linear function	A linear function represents a straight line on a Cartesian plane and has a constant rate of change.	1 (Math)	92.6%
probability distribution	The shape of a probability distribution affects how likely specific outcomes are.	1 (Math)	81.6%
school project	She worked late on her school project about environmental science.	1 (Math)*	63.1%
integral calculus	Integral calculus deals with accumulation and the calculation of areas under curves.	1 (Math)	76.8%
homework folder	He forgot his homework folder on the bus.	0 (Non-Math)	49.3%

Table 3: Zero-shot predictions on bigrams not seen during training. *Incorrect prediction — *school project* is not a mathematical concept.

D Model Interpretation

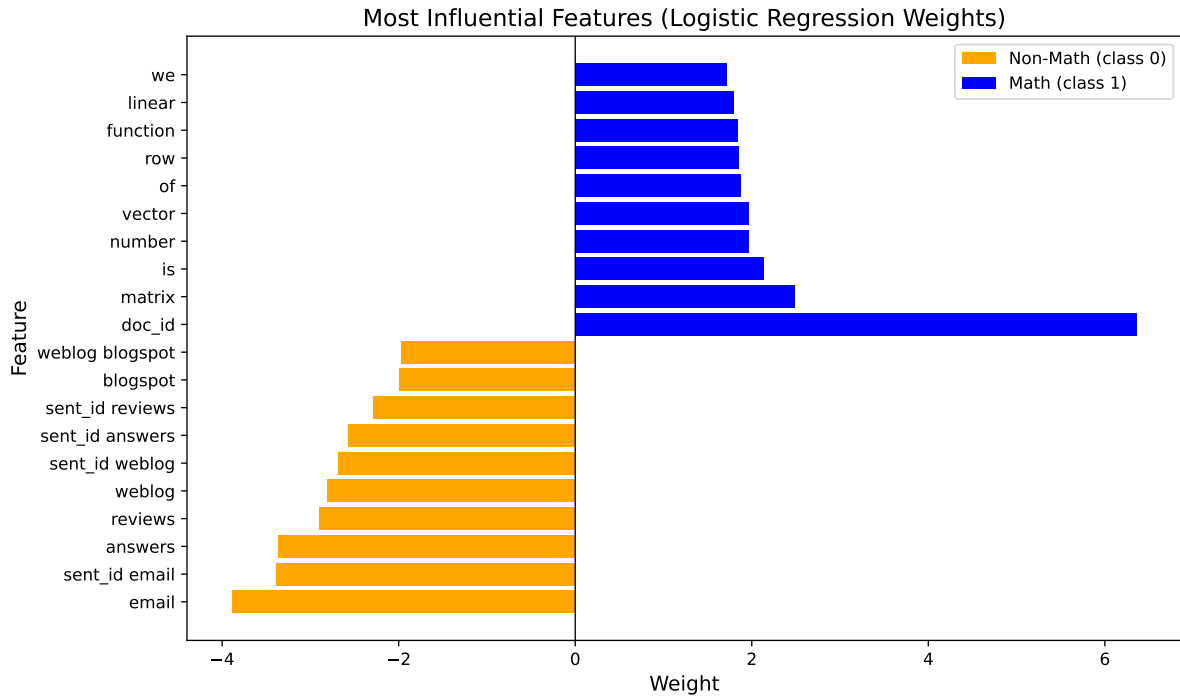


Figure 5: Top weighted features from the logistic regression model. Positive weights (blue) indicate strong association with mathematical concepts, while negative weights (orange) are associated with non-mathematical content. Some features may reflect structural tokens (e.g., *doc_id*, *email*) from the dataset.