Automatic Question classification in Portuguese: A Large-Scale Dataset and Comparative Evaluation of Classification Strategies

Murilo Boccardo and Valéria D. Feltrim
Departamento de Informática,
Universidade Estadual de Maringá, Brazil
muriloboccardo@gmail.com and vdfeltrim@uem.br

Abstract

This paper presents a comparative evaluation of automatic classification strategies for Brazilian university entrance exam questions by subject and fine-grained topic. A central contribution of this study is the creation and curation of a large-scale Portuguese-language dataset comprising approximately 17,000 questions collected from the Agatha.edu platform, carefully cleaned and normalized. We investigated two alternative classification strategies: a single-step approach that directly predicts fine-grained topics and a two-stage approach in which an initial model predicts the subject, followed by specialized topic classifiers. These strategies were evaluated using both classical machine learning methods, such as Support Vector Machines, Naive Bayes, and Random Forest, and transformer-based language models pre-trained for Portuguese. Experimental results show the feasibility of large-scale automatic question classification and highlight the potential of NLP-based classification strategies to support the curation, analysis, and organization of educational question banks.

1 Introduction

Digital educational platforms have increased the volume of assessment content, motivating automated methods to organize and analyze large question banks. Within this context, the automatic classification of educational texts emerges as a relevant challenge. This task involves assigning labels to statements and instructional materials according to predefined criteria. Question categorization is particularly important, as it enables the organization of large assessment item banks by subject, topic, or difficulty level. Unlike document-level classification, question classification must handle short and information-sparse texts, which increases the ambiguity and complexity of the task (Silva et al., 2019). In educational applications, this type of classification has been explored in adaptive learning systems, automated test generation, and content recommendation, supporting more strategic use of question repositories and more targeted learning experiences (Urdaneta-Ponte et al., 2021).

In this work, we present a comparative evaluation of automatic classification strategies for Brazilian university entrance exam questions by subject and fine-grained topic. The focus of the study was to analyze how different strategies for structuring the label space affect performance, scalability, and robustness in large-scale question categorization. In particular, we investigated whether explicitly modeling the hierarchical relationship between subjects and topics provides advantages over direct fine-grained classification. We explored two classification strategies: a single-step approach that directly predicts fine-grained topics and a two-stage approach in which an initial model predicts the subject, followed by specialized topic classifiers. Both strategies were evaluated within a common experimental pipeline to ensure a systematic comparison.

To support this study, we compiled a dataset of approximately 17,000 questions extracted from the Agatha.edu platform[1]. The data underwent collection, cleaning, and category normalization. To our knowledge, this is the first large-scale dataset of its kind in Brazilian Portuguese and constitutes an important contribution of this work.

Our experimental analysis included classical machine learning (ML) models such as Support Vector Machines (SVM), Naive Bayes, and Random Forest, as well as deep learning models such as BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019), with a focus on models pre-trained for Portuguese. For classical models, the preprocessing pipeline included stopword removal, lemmatization, tokenization, and vectorization using TF-IDF. Model evaluation was based on accuracy, precision,

[1]Agatha Project: https://projetoagathaedu.com.br/banco-de-questoes.php. Accessed: November 7, 2025.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 436–445

recall, F1-score, and confusion matrix analyses. We did not evaluate large language models (LLMs), as the primary objective of this work is to introduce the dataset and establish baselines using discriminative models tailored for closed taxonomies.

The results show the feasibility of large-scale automatic question classification and highlight the practical relevance of comparing alternative classification strategies for educational data.

2 Related Work

Automatic question classification has been investigated using educational taxonomies and thematic or semantic criteria. However, research focusing on Portuguese-language datasets and Brazilian exam questions remains limited (Silva et al., 2019). In the context of the Brazilian National Student Performance Examination (ENADE), for instance, Araújo (2025) observe that fully automated NLP-based pipelines for question categorization are still relatively rare.

Among the studies most closely aligned with ours, Araújo (2025) address the automatic classification of ENADE Computer Science questions into 17 fine-grained topics using OCR, standard preprocessing techniques, TF-IDF and Word2Vec representations, and classical ML classifiers such as Naive Bayes, SVM, and Random Forest. Results indicate moderate performance and strong sensitivity to dataset size and class imbalance, highlighting challenges inherent to the automatic classification of short and information-sparse educational texts.

Beyond the Brazilian educational context, several studies have explored short-text classification using neural architectures. Wang et al. (2020) propose a CNN-based model tailored for short texts, combining n-gram features, nonlinear windows, and attention mechanisms, and report improvements over traditional methods on English and Chinese datasets. More recently, Tong et al. (2022) introduced a lightweight DistilBERT–BiLSTM architecture designed for small and imbalanced datasets, achieving competitive performance while reducing computational cost. These works illustrate the progression from feature-based to contextualized neural models for short-text classification.

In contrast to previous studies, we contribute a large-scale Portuguese dataset spanning multiple subjects and a systematic comparison of direct versus two-stage hierarchical classification with classical ML and Portuguese pre-trained transformers. This positioning aligns our contribution with recent advances in NLP while addressing the specific challenges of educational question classification in the Brazilian context.

3 Dataset

We compiled our dataset by collecting Brazilian university entrance exam questions from the Agatha.edu platform, which organizes assessment items by subject and thematic lists. Prior to data collection, we verified the platform's *robots.txt* file[2], which explicitly permits automated crawling without restrictions. Then, using a Python-based web scraping pipeline built with Requests and BeautifulSoup, we extracted the main fields associated with each question, namely: subject, topic, statement, answer options, and, when available, the originating institution. The raw data collection resulted in a total of 21,634 questions.

3.1 Initial Structure

Agatha.edu groups questions into broad subject areas such as Mathematics, Biology, Chemistry, Physics, History, and Languages. Within each subject, the platform provides thematic exercise lists. The titles of these lists were used as initial topic labels (e.g., *Mathematics* → *Plane Geometry*). However, the raw taxonomy exhibited very high granularity, comprising 1,689 distinct topics across all subjects, many of which were semantically overlapping or excessively specific (e.g., *Brasil República I/II/III* in History or *Geometria Plana I/II* in Mathematics). In addition, several topics contained fewer than five instances, which compromise their suitability for supervised learning.

3.2 Cleaning, Normalization, and Topic Consolidation

We performed a multi-stage data cleaning and normalization process composed of:

(i) removal of scraping artifacts (e.g., header fragments and contextual texts not corresponding to actual questions); (ii) text normalization, including whitespace cleanup and removal of formatting characters such as `\n` and `\t`; and (iii) deduplication. A total of 6,596 duplicated statements were identified: 5,173 duplicates within the same class were removed, while cross-topic duplicates (1,423 instances) were retained due to their multidisciplinary relevance.

[2]Agatha robots.txt: https://projetoagathaedu.com.br/robots.txt. Accessed: March 6, 2026.

To reduce fragmentation and improve label coherence, we unified semantically equivalent or pedagogically redundant topics. Examples include merging *Geometria Plana I/II* into *Geometria Plana*, consolidating solid geometry sublists under *Geometria Espacial*, and grouping several variants of *Brasil República* into a single category. Also, topics with fewer than 25 questions were discarded, a threshold chosen to ensure minimal representativeness for the training of supervised models (Tänzer et al., 2022).

After consolidation, the number of topics was reduced from 1,689 to 125, corresponding to a reduction of approximately 92.6%. This process resulted in more coherent topic definitions and more balanced class distributions across subjects.

3.3 Text Length Analysis

An important characteristic of the dataset is the variability in question length across different disciplines. Analyzing text length is particularly relevant for transformer-based models such as BERT, which impose a maximum input sequence length. Figure 1 shows substantial variation in question length across subjects: Humanities tend to include longer texts (often requiring interpretation of excerpts), whereas STEM subjects typically present shorter statements.

3.4 Final Dataset Distribution

After cleaning, consolidation, and filtering, the final dataset comprises 17,556 questions distributed across 12 subjects and 125 topics. Table 1 summarizes the dataset composition by subject.

Subject	Topics	Questions
History	19	3,340
Biology	21	3,207
Mathematics	14	2,566
Chemistry	11	2,055
Physics	14	1,782
Literature	10	1,155
Philosophy	5	935
Arts	9	733
Portuguese	7	728
Geography	8	565
Sociology	5	289
Foreign Languages	2	201
Total	**125**	**17,556**

Table 1: Distribution of questions and topics per subject in the final dataset.

As shown in Table 1, History and Biology are the most represented disciplines, reflecting their predominance in Brazilian entrance exams. Although some degree of class imbalance remains inherent to the domain, the consolidation process substantially reduced extreme class fragmentation compared to the raw data. As a result, even minority classes retain a sufficient number of instances to allow models to learn meaningful decision boundaries, mitigating issues associated with extremely sparse categories.

Each instance in the dataset includes the fields: `subject`, `topic`, `text`, and `options`. All preprocessing scripts and mappings used for topic consolidation, as well as the raw questions and final dataset, are publicly available in the project's repository[3].

4 Classification Strategies

We explored alternative strategies for classifying entrance exam questions by subject and topic using the full text of each question, including the statement and answer options. To support this evaluation, we adopted a modular experimental pipeline composed of three main stages: preprocessing, classification, and output generation.

Two classification strategies were implemented and evaluated: (i) *direct classification* and (ii) *two-stage hierarchical classification*. These strategies allowed us to investigate different ways of structuring the label space and to analyze trade-offs between simplicity and granularity.

4.1 Direct Classification

In the direct classification strategy, a single model predicts fine-grained topic labels without explicitly separating subjects from topics. The model receives the preprocessed and vectorized question text as input and outputs a single label corresponding to the final topic.

This strategy offers a simpler pipeline and lower inference time, as no intermediate predictions are required. However, the large number of target classes, even after the dataset consolidation, and the substantial class imbalance increase task difficulty, especially in the presence of semantically similar topics. As a result, direct classification is more sensitive to data sparsity and overfitting.

[3]Available in: https://github.com/murilob03/automatic-question-classification-pt

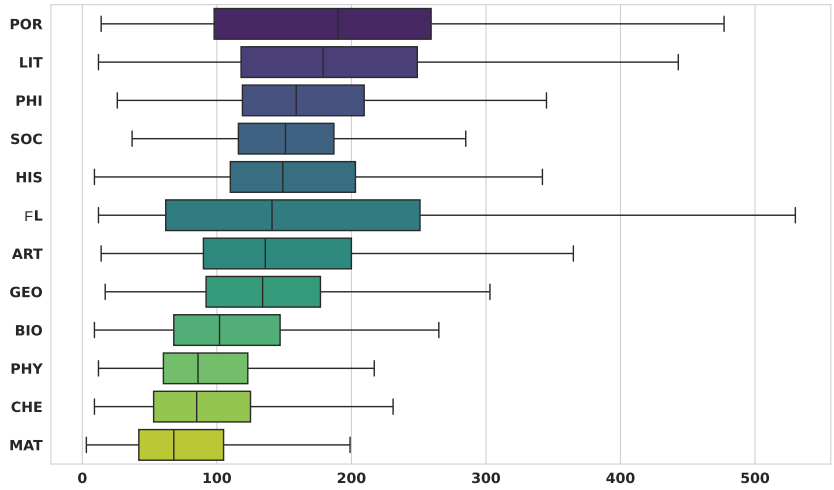

Figure 1: Distribution of question lengths (word count) across subjects: Arts (ART), Biology (BIO), Philosophy (PHI), Physics (PHY), Geography (GEO), History (HIS), Foreign Languages (FL), Literature (LIT), Mathematics (MAT), Portuguese (POR), Chemistry (CHE), and Sociology (SOC).

4.2 Two-Stage Hierarchical Classification

The second strategy follows an explicit hierarchical structure consisting of two sequential prediction levels:

1. Subject classification: in this stage, a single model identifies the main subject of a question (e.g., Physics);

2. Topic classification: for each subject, a specialized classifier distinguishes among its associated fine-grained topics (e.g., Mechanics, Thermodynamics, Optics).

This decomposition reduces the complexity of the decision space by restricting each secondary model to a smaller and semantically coherent subset of topics, enabling more precise discrimination. However, the hierarchical strategy introduces additional computational cost and depends heavily on the accuracy of the first stage classifier, since misclassifications at the subject level propagate directly to the topic prediction.

5 Models and Experimental Setup

This section describes the preprocessing pipeline, text representations, classification models, and evaluation procedures used in our experimental study.

5.1 Preprocessing

The preprocessing pipeline combined the question statement and answer options into a single text field. Unicode normalization (NFKC/NFD) was applied, and typographic variants were standardized. Mathematical symbols with semantic value (e.g., π, Ω) were mapped to canonical tokens such as `<pi>` and `<ohm>`. URLs and numerical expressions were replaced with `<url>` and `<num>`.

Alternative markers (e.g., `A)`, `B.`) were removed using regex, and punctuation cleanup preserved symbols relevant to scientific notation (e.g., %, /, <, >). Two text variants were produced: a cased version for transformer-based models and a normalized lowercase version for classical ML models.

Stopwords were removed using spaCy's Portuguese list, except for negation words and domain-relevant terms (e.g., *mínimo*, *ponto*). Lemmatization (spaCy) and stemming (RSLP) were also evaluated, however, stemming was used only for comparison purposes due to its over-aggressive conflation behavior.

5.2 Text Representations

Classical models used TF–IDF vectors (generated using scikit-learn) with configurable n-gram ranges and minimum document frequency thresholds. Extremely frequent terms were removed (`max_df = 0.8`), and a supervised χ^2 feature selection step retained the K most informative attributes. Transformer-based models relied on contextual embeddings learned internally during fine-tuning and therefore required no external vectorization step.

5.3 Handling Class Imbalance

The dataset exhibits substantial class imbalance, particularly at the topic level. As an initial mitiga-

tion step, extremely fine-grained or sparsely populated topics were consolidated during dataset preparation, as described earlier. During model training, all classifiers used class-weighted loss functions. For classical models, class weights were handled using `class_weight='balanced'` in scikit-learn. For transformers, class weights were computed with `compute_class_weight` and incorporated directly into the loss function during fine-tuning.

5.4 Classification Models

Four classical models were evaluated: Linear SVC, Logistic Regression, Multinomial Naive Bayes, and Random Forest. Hyperparameters were selected via grid search, as detailed in Section 5.5.

Deep learning experiments used two Portuguese transformer models: BERTimbau (Souza et al., 2020) and DistilBERTimbau (Adalberto Junior, 2024). Based on the text length analysis, which revealed that over 95% of the questions contain fewer than 365 tokens, the maximum sequence length was set to 365 to ensure maximum information retention.

5.5 Training Protocol and Evaluation

To ensure a robust evaluation, the dataset was stratified into training (70%), validation (15%), and test (15%) sets. We adopted a *refit* strategy to maximize the data available for the final models, as follows:

Classical Models: Hyperparameters were tuned using grid search with stratified 4-fold cross-validation on the training and validation sets. Once optimal hyperparameters were identified, the models were evaluated on the held-out test set.

Transformers: Models were fine-tuned on the training set (batch size 12, AdamW optimizer) with performance monitored on the validation set using early stopping. The epoch that yielded the best macro-F1 score was recorded. Subsequently, the models were retrained from scratch on the combined training and validation sets for that specific number of epochs (without further validation) and evaluated on the test set.

Pipelines and Scenarios: Experiments followed two setups: (i) **Direct Prediction**, using a single model to predict topics directly; and (ii) **Two-Stage Prediction**, where a subject classifier is followed by specialized topic classifiers.

For the two-stage pipeline, the second stage (topic prediction) was evaluated in two scenarios:

1. **Gold Standard:** Specialized topic models predicted topics given the ground-truth subject label. This assesses the topic classifiers in isolation.

2. **Cascaded:** Specialized topic models predicted topics based on the subject inferred by the first-stage model (specifically, the best-performing BERT-based subject classifier). This propagates errors from the first stage, reflecting real-world usage.

Evaluation metrics included accuracy, precision, recall, and macro-F1 score. Given the imbalanced label distribution, macro-F1 was considered the primary metric.

6 Results and Discussion

In this section, we present the experimental results. The analysis was structured to assess the effectiveness of both the direct single-step strategy and the hierarchical two-stage strategy. In addition, we provide a comparative discussion of these strategies and a qualitative analysis of error patterns using confusion matrices.

6.1 Direct Classification

Table 2 summarizes model performance under the direct classification strategy, in which a single classifier predicts one of the 125 topics. Among classical models, Linear SVC and Logistic Regression achieved the best results (macro-F1 = 0.58), confirming the effectiveness of linear decision boundaries combined with TF–IDF representations. Naive Bayes performed reasonably well given its simplicity, whereas Random Forest exhibited limitations when handling high-dimensional sparse text, obtaining the lowest scores across all metrics.

Model	Acc.	Prec.	Rec.	F1-m	F1-w
L-SVC	0.653	0.568	0.63	0.58	0.647
LR	0.649	0.586	0.593	0.58	0.647
RF	0.457	0.413	0.489	0.413	0.443
NB	0.624	0.593	0.505	0.517	0.604
BERT	**0.666**	**0.612**	**0.601**	**0.600**	**0.664**
D-BERT	0.663	0.616	0.61	0.602	0.662

Table 2: Direct classification performance. Acc=Accuracy; Prec=Precision; Rec=Recall; F1-m=Macro-F1; F1-w=Weighted-F1. L-SVC=Linear SVC; LR=Logistic Regression; RF=Random Forest; NB=Naive Bayes; D-BERT=DistilBERT.

Transformer-based models outperformed the classical baselines. BERT and D-BERT achieved

the best overall performance (macro-F1 $\approx$ 0.6). Despite these gains, even transformer-based models were affected by the high granularity of the label space and the long-tailed class imbalance, with notable degradation in low-frequency topics. These observations motivated the evaluation of a hierarchical alternative.

6.2 Two-Stage Classification

We now report the results obtained with the two-stage classification strategy. We first analyze the performance of the subject classifier (Stage 1), followed by the topic-level classification results (Stage 2), in order to assess the impact of hierarchical decomposition across different model families.

6.2.1 Stage 1: Subject Classification

Table 3 reports model performance in the first stage, where the goal is to predict the subject of each question. The results indicate high separability between subjects. Except for Random Forest, all classical models achieved a macro-F1 > 0.83, while BERT obtained the best overall performance (macro-F1 = 0.867). This outcome was expected given the relatively small number of subjects (12) and the clearer linguistic distinctions between broad disciplines compared to fine-grained topics.

Model	Acc.	Prec.	Rec.	F1-m	F1-w
L-SVC	0.896	0.851	0.849	0.85	0.896
LR	0.894	0.854	0.849	0.851	0.894
RF	0.816	0.772	0.747	0.753	0.813
NB	0.893	0.858	0.825	0.836	0.891
BERT	**0.917**	**0.888**	**0.867**	**0.877**	**0.916**
D-BERT	0.907	0.87	0.855	0.861	0.907

Table 3: Subject classification performance (Stage 1).

Because errors at this stage propagate directly to the second stage, high subject-level accuracy is essential for reliable hierarchical classification. To minimize error propagation effects, we selected the fine-tuned BERT model (Accuracy = 0.917) as a fixed subject predictor for all subsequent topic-level experiments. This choice ensured that second-stage classifiers operate under the most reliable subject predictions available.

6.2.2 Stage 2: Topic Classification

In the second stage, topic classification was performed conditioned on the subject predicted in Stage 1. Table 4 presents results averaged across all subjects. The results show that hierarchical decomposition affects model families differently. Simpler models, such as Naive Bayes and Random Forest, benefited from the reduced and semantically constrained label space. Random Forest improved from a macro-F1 of 0.413 in the direct strategy to 0.542 in the hierarchical setting.

Model	Acc.	Prec.	Rec.	F1-m	F1-w
L-SVC	0.667	0.598	0.626	0.596	0.662
LR	0.658	0.601	0.603	0.592	0.656
RF	0.608	0.572	0.550	0.542	0.602
NB	0.648	0.613	0.568	0.572	0.640
BERT	**0.679**	**0.643**	**0.634**	**0.621**	**0.673**
D-BERT	0.663	0.597	0.611	0.590	0.659

Table 4: Topic classification performance (Stage 2).

In contrast, more robust models, such as Linear SVC, Logistic Regression, BERT, and DistilBERT, exhibited marginal performance drops or stagnation. This suggests that these models can handle the full topic space directly without explicit hierarchical decomposition. Overall performance in Stage 2 remains constrained by residual error propagation from Stage 1, particularly for subjects that are harder to distinguish based solely on surface-level linguistic cues.

6.3 Direct vs. Two-Stage Strategies

The comparison in Table 5 reveals distinct behaviors between classical and transformer-based models. Classical models consistently benefited from the hierarchical approach, with Random Forest and Naive Bayes showing the most substantial gains.

However, it is crucial to interpret these gains in the context of the experimental setup. In the realistic two-stage scenario, the first classification step (subject prediction) was performed by the best-performing model (BERT). Consequently, the improved metrics for weaker classifiers like Random Forest and Naive Bayes are not solely due to the reduced classification space within each subject. They also stem from the hybrid nature of the pipeline: these models effectively "inherited" the high accuracy of the transformer-based subject predictor, shielding them from inter-subject confusion errors they would likely commit if operating as standalone subject classifiers.

In contrast, the results for transformer-based models were inconclusive regarding a preferred strategy. While BERT achieved a slightly higher Macro-F1 in the two-stage setting (0.621 vs. 0.600), DistilBERT performed marginally better in the direct setting (0.602 vs. 0.590). Given the small magnitude of these differences and the

contradictory trends, these variations are likely attributable to stochastic variance rather than a structural advantage. This indicates that pre-trained language models are sufficiently robust to handle the high granularity of topics directly.

Model	Direct	Two-Stage
L-SVC	0.580	0.596
LR	0.580	0.592
RF	0.413	0.542
NB	0.517	0.572
BERT	0.600	**0.621**
D-BERT	**0.602**	0.590

Table 5: Macro-F1 comparison between direct and two-stage classification strategies.

Ultimately, despite the gains observed for classical algorithms, the two-stage strategy proves difficult to justify in this experimental context. The performance improvements were marginal and came at the expense of a disproportionate increase in architectural complexity and computational overhead. While the direct approach requires training a single model, the hierarchical pipeline necessitates training and managing dozens of specialized models, one for each subject, in addition to the initial router. More importantly, relying on a computationally expensive transformer as the first-stage router completely negates the inference-time advantages of using lightweight classical models, such as Naive Bayes or Random Forest, in the second stage. If the computational cost of a transformer is already being incurred, it is far more efficient to deploy a single transformer to predict the fine-grained topics directly. Consequently, the direct strategy emerges as the superior approach, offering a much more efficient balance between performance and resource utilization.

6.4 Classical vs. Transformer-based Models

Transformer-based models consistently achieved the highest macro-F1 scores under both strategies, demonstrating superior handling of semantic overlap and contextual dependencies. BERT delivered the best overall performance, while DistilBERT provided a competitive alternative with significantly lower computational cost.

Among classical approaches, Linear SVC and Logistic Regression emerged as strong baselines despite their lower complexity. Naive Bayes exceeded expectations, particularly in the hierarchical setting, reinforcing its suitability for large-scale deployments that require fast inference. Random Forest, on the other hand, consistently underperformed, likely due to its difficulty in modeling sparse high-dimensional text features.

To illustrate the performance differences between transformer-based and classical models, consider an Arts question (ID 15144) containing the poem excerpt "Já cantam vitória, Já meigos atendem à voz do cantor" alongside historical references to "Romantismo europeu" (European Romanticism). The Linear SVC, relying on TF-IDF representations, misclassified this instance into the fine-grained topic "Music and Dance", likely misled by isolated lexical triggers such as "cantam" (sing) and "cantor" (singer). In contrast, BERT correctly predicted "Art from Renaissance to Romanticism". This demonstrates the transformers ability to capture broader contextual dependencies, assigning greater weight to the overall historical and artistic context than to isolated and superficial features that easily confound classical models.

These findings highlight a clear trade-off: transformer-based models maximize predictive performance but require greater computational resources, whereas linear models remain attractive for real-time or resource-constrained applications despite their limited ability to capture broader contextual dependencies.

6.5 Qualitative Analysis and Error Patterns

Figure 2 presents the topic-level performance aggregated by subject using the best-performing model (BERT). It is important to note that these scores correspond to the *Gold Standard* evaluation scenario. Analyzing per-subject Macro-F1 in the cascaded, more realistic, pipeline is structurally inconsistent, as errors in the first stage (subject prediction) route questions to incorrect specialized classifiers, introducing out-of-domain topics that invalidate the calculation of intra-subject metrics.

The results reveal substantial variation across domains. Subjects with well-defined and semantically distinct topics, notably Foreign Language (1.0) and Sociology (0.828), achieved the highest scores. Conversely, disciplines such as Biology (0.611), Portuguese (0.547), and Arts (0.461) presented the lowest macro-F1 values. In these cases, topics exhibit strong lexical and conceptual overlap, making them harder to distinguish. This is particularly evident in Biology, where shared terminology across genetics, ecology, and physiology reduces topic separability despite the technical nature of the vocabulary.

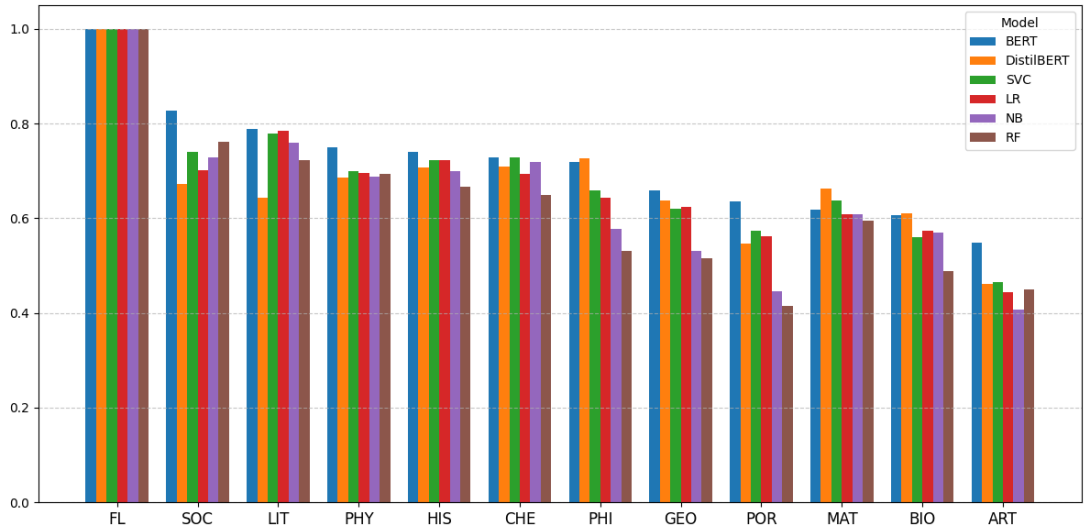

Figure 2: Topic-level macro-F1 by subject: Arts (ART), Biology (BIO), Philosophy (PHI), Physics (PHY), Geography (GEO), History (HIS), Foreign Languages (FL), Literature (LIT), Mathematics (MAT), Portuguese (POR), Chemistry (CHE), and Sociology (SOC).

An interesting pattern is that subject size does not correlate with performance. Smaller subjects (Foreign Language and Sociology) yielded some of the highest scores, whereas larger ones (e.g., History and Biology) did not. This observation reinforces that semantic separability, rather than dataset size, is the primary driver of topic-level performance.

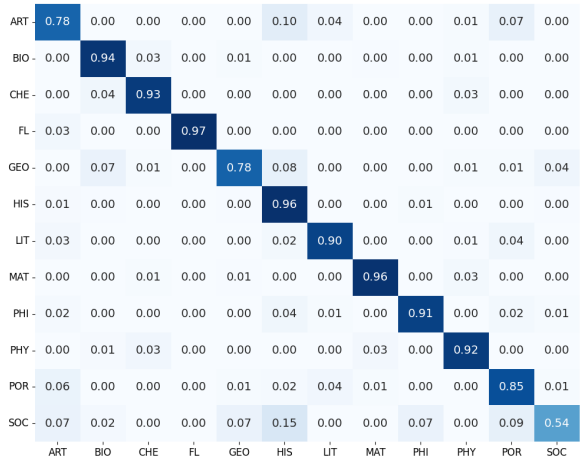

Figure 3: Subject-level confusion matrix (BERT).

Figure 3 presents the confusion matrix for the best-performing subject classifier (BERT). It highlights systematic confusions between discursive domains such as Arts, Portuguese, and Literature, as well as between History, Geography, and Sociology. For instance, a question (ID 15418) analyzing characters from Ariano Suassuna's play "Auto da Compadecida" is labeled in the ground truth as Arts (Theater), but BERT confidently classified it as Literature (Theater). In practice, this prediction is contextually valid, as the text legitimately spans both disciplines. This example clearly demonstrates that while the model successfully grasps the thematic core of the task, it is penalized by the rigid, single-label constraints of the dataset. In contrast, more formal domains like Philosophy and Chemistry exhibit cleaner boundaries, resulting in fewer confusions.

These findings suggest two main directions for future refinement: (i) revisiting topic definitions in domains with high intra-subject overlap, potentially merging or re-labeling ambiguous categories, and (ii) incorporating richer linguistic signals, such as syntactic features, discourse markers, or subject-specific embeddings, to improve topic resolution in interpretative domains.

7 Conclusion

This work addressed the automatic classification of entrance exam questions, a task characterized by short texts and specific pedagogical nuances. Beyond the comparative evaluation of classification strategies, a central contribution of this study is the creation and curation of a large-scale dataset of entrance exam questions in Portuguese. By collecting, cleaning, and normalizing data from the Agatha.edu platform, we help address a gap in the literature regarding openly available resources for educational NLP in Portuguese. We are making this dataset publicly available to foster reproducibility

and to serve as a starting point for future research.

From a modeling perspective, this study establishes a comprehensive benchmark for the proposed dataset. We evaluated both a direct single-step strategy and a hierarchical two-stage strategy using classical ML models and transformer-based models. The results indicate that high-capacity models, particularly transformer-based models and linear classifiers, are better suited to the direct strategy, as they can effectively capture fine-grained distinctions without suffering from error propagation. In contrast, the hierarchical strategy proved beneficial mainly for lower-capacity models, such as Naive Bayes and Random Forest, which struggle with large and sparse label spaces.

The dataset reflects the inherent complexity of real-world educational assessment, including pronounced class imbalance and substantial lexical and conceptual overlap across topics. Rather than constituting limitations, these characteristics position the dataset as a challenging and realistic benchmark for studying class imbalance and semantic ambiguity in short educational texts.

Finally, the availability of this dataset opens several directions for future work. These include domain-adaptive pre-training on large educational corpora and data augmentation techniques to mitigate imbalance. Most notably, while this study established baselines using discriminative models tailored for closed taxonomies, a natural next step is to evaluate the performance of modern Large Language Models (LLMs) on this benchmark. Future research should investigate whether zero-shot, few-shot, or parameter-efficient fine-tuning (PEFT) strategies can successfully navigate the highly granular 125-topic space without outputting out-of-domain labels, and how the generative paradigm compares to the hierarchical strategies explored in this work.

Limitations

This study is subject to several important limitations that should be considered when interpreting the results. First, the dataset is highly imbalanced, with long-tailed topic distributions that remain difficult to model even with class weighting and topic consolidation. Second, the high granularity and the semantic overlap inherent to this problem created ambiguous boundaries in the ground-truth labels, which likely impacted and artificially limited model performance.

Certain subjects, such as Biology, Portuguese and Arts, exhibit substantial lexical and conceptual overlap between topics, reducing intrinsic separability and making fine-grained classification particularly challenging. Finally, transformer-based models were used without domain-adaptive pre-training and under computational constraints, limiting the exploration of larger architectures or more extensive hyperparameter tuning.

Acknowledgments

The authors thank the Pró-Reitoria de Ensino of the Universidade Estadual de Maringá for the financial support provided to the student involved in this project.

References

Adalberto Junior. 2024. distilbert-portuguese-cased (revision df1fa7a).

Lucas Roges de Araújo. 2025. Classificação automática de questões de provas: análise comparativa de algoritmos e aplicação ao enade. Bachelor's Thesis (B.Sc. in Computer Science), Universidade Federal de Santa Maria, Santa Maria, RS, Brazil.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Valtemir A. Silva, Ig Ibert Bittencourt, and José C. Maldonado. 2019. Automatic question classifiers: A systematic review. *IEEE Transactions on Learning Technologies*, 12(4):485–502.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil*, pages 403–417.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Jiajun Tong, Zhixiao Wang, and Xiaobin Rui. 2022. A multimodel-based deep learning framework for short text multiclass classification with the imbalanced and extremely small data set. *Computational Intelligence and Neuroscience*, 2022(1):7183207.

María Cora Urdaneta-Ponte, Amaia Mendez-Zorrilla, and Ibon Oleagordia-Ruiz. 2021. Recommendation systems for education: Systematic review. *Electronics*, 10(14).

Haitao Wang, Jie He, Xiaohong Zhang, and Shufen Liu. 2020. A short text classification method based on n-gram and cnn. *Chinese Journal of Electronics*, 29(2):248–254.