An Axiological Matrix for Brazilian Portuguese Bluesky Discourse on AI in Education

Marcela Y. A. Barbosa[1,2] **, Bryan K. S. Barbosa**[1,2,3]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)
São Carlos, SP – Brazil

[2]Programa de Pós-Graduação em Linguística,
Universidade Federal de São Carlos (UFSCar)
São Carlos, SP – Brazil

[3]Instituto Atlântico
Fortaleza, CE – Brazil

marcela.barbosa@estudante.ufscar.br, bryankhelven@ieee.org

Abstract. *Public discussion of AI in education is poorly captured by topical retrieval or generic sentiment polarity. We present an axiological matrix, annotation protocol, Brazilian Portuguese Bluesky corpus, classifiers, and temporal projection. The framework filters applicability and maps applicable posts by acceptance and evaluative intensity. Of 1,497 posts, 734 were adjudicated as applicable; pre-adjudication agreement reached 88.4% and Cohen's* $\kappa = .770$*. Repeated cross-validation shows BERTimbau strongest for applicability (macro-F1* $= .903$*) and three-class intensity, while sparse models remain competitive for acceptance and strongest on five-class tasks. Linear and hybrid pipelines project 989 posts from 2024 into complementary 3×3 and 5×5 matrices.*

1. Introduction

Generative artificial intelligence is a disputed object in educational discourse. Public discussion addresses authorship, plagiarism, assessment, teacher authority, institutional policy, unequal access, automation, and the future of learning. Research and policy documents pair pedagogical opportunities with concerns about reliability, dependency, transparency, and human oversight [UNESCO 2023, Kasneci et al. 2023, Chan and Hu 2023, Tlili et al. 2023]. These disputes circulate on social media, where teachers, students, researchers, institutions, news pages, and other users evaluate AI through compressed language, headlines, irony, affect, and references to current controversies.

This paper studies such evaluations in Brazilian Portuguese Bluesky posts. The task begins before polarity classification. Search queries retrieve posts that mention AI and education but do not necessarily evaluate AI in an educational context. A course announcement, a headline, a metaphorical use of *teacher*, or a link with little visible framing may be topically related yet unsuitable for axiological interpretation. Applicable posts introduce another difficulty: orientation and force are distinct. A post can reject AI with restrained wording, accept it conditionally while stressing risk, or report opposed positions without adopting either one.

We address these problems through a hierarchical axiological matrix. An applicability gate decides whether the visible text provides enough evaluative framing about AI in education. Applicable posts receive two ordinal labels. *Acceptance* represents orientation toward AI in education, ranging from strong rejection to strong acceptance. *Intensity* represents the rhetorical, affective, or emphatic force with which that orientation is constructed. The matrix was adapted from a discourse-analytic proposal developed for interactions about AI and education on social media [Barbosa 2025] and operationalized here as a reproducible annotation and classification task.

The design draws on a view of utterances as socially situated and responsive, in which verbal signs carry evaluative accents in relation to objects, interlocutors, and prior discourse [Bakhtin 1986, Voloshinov 1973]. Work on appraisal provides a complementary account of evaluative attitude and graduation [Martin and White 2005]. These perspectives motivate a task that preserves direction and force as separate dimensions rather than reducing them to a single sentiment score.

Bluesky is used as an accessible source of contemporary user-generated content under the AT Protocol. It is not treated as a representative sample of Brazilian public opinion. Its methodological value lies in enabling corpus construction at a time when research access to Twitter/X data has become more restricted and unstable [Davidson et al. 2023, Kleppmann et al. 2024].

The study makes six contributions. First, it releases the design and analysis of a 1,497-post corpus with a formally adjudicated applicability layer. Second, it provides a detailed manual that separates applicability, acceptance, and intensity and covers recurring edge cases. Third, it quantifies raw agreement, calibration differences, consensus-subset selection, and disagreement patterns. Fourth, it compares all predefined TF-IDF configurations, BERTimbau, majority, and stratified-random baselines under shared repeated cross-validation. Fifth, it evaluates disagreement-aware training over all applicable items rather than discarding interpretive variation. Sixth, it projects two pipelines onto a 2024 corpus, audits the raw collection linguistically, retains both matrix granularities, and conducts a blind, adjudicated diagnostic review of 50 cases.

2. Related Work

2.1. User-generated content, platform access, and Bluesky

User-generated content provides naturally occurring language outside controlled experimental settings. It also contains informal spelling, hashtags, URLs, mentions, ellipsis, irony, code-switching, affective punctuation, and platform-specific conventions [Owoputi et al. 2013, Eisenstein 2013]. These features matter here because evaluative force may be carried by punctuation, capitalization, emojis, or rhetorical compression rather than by isolated sentiment words.

Most short-text NLP resources have historically been built from Twitter/X, including influential sentiment and stance benchmarks [Rosenthal et al. 2017, Mohammad et al. 2016]. Changes in platform governance and API access now limit independent data collection and reproducibility [Davidson et al. 2023]. Bluesky offers a different technical environment. Its AT Protocol architecture supports an open social web with portable identity and decentralized components [Kleppmann et al. 2024]. We use the platform as an empirical site where personal opinions, institutional communication, news circulation, and promotional discourse coexist.

2.2. AI in education as evaluative discourse

Generative AI in education is discussed simultaneously as pedagogical support, institutional risk, infrastructure, shortcut, threat to authorship, and condition for future participation. UNESCO emphasizes regulation, capacity building, and human-centered use [UNESCO 2023]. Kasneci et al. review opportunities and risks for learning, assessment, misinformation, and teacher roles [Kasneci et al. 2023]. Student perceptions likewise combine expected benefits with concerns about reliability, skills, and ethics [Chan and Hu 2023]. Tlili et al. show how educational chatbot use brings together promise and pedagogical or ethical risk [Tlili et al. 2023]. Broader work on AI in education also treats the field through its implications for teaching and learning [Holmes et al. 2019].

These debates form a public evaluative field rather than a single technical topic.

A binary positive/negative label cannot represent conditional acceptance, neutral but applicable framing, or the difference between mild criticism and highly emphatic rejection. The matrix therefore models orientation and force independently.

2.3. Axiological evaluation and adjacent NLP tasks

Bakhtin's account of speech genres emphasizes the situated and responsive character of utterances [Bakhtin 1986]. Voloshinov foregrounds the social orientation and evaluative dimension of verbal signs [Voloshinov 1973]. Martin and White describe evaluation through attitude, engagement, and graduation [Martin and White 2005]. These approaches are useful for posts in which AI may be framed as a tool, threat, innovation, teacher substitute, policy object, or future condition of schooling.

The computational task is adjacent to sentiment analysis, opinion mining, and stance detection [Pang and Lee 2008, Rosenthal et al. 2017, Mohammad et al. 2016, Küçük and Can 2021], but differs in three respects. First, it filters topical mentions through an axiological applicability decision. It then treats acceptance as an ordinal orientation rather than a fixed favor/against benchmark label. Finally, it represents intensity as a separate axis, allowing the same orientation to occur under different degrees of textual force.

2.4. Reliability, disagreement, and modeling

Subjective annotation requires explicit guidelines and reliability analysis. Cohen's kappa remains a standard categorical measure [Cohen 1960], while computational-linguistic surveys stress that agreement must be interpreted in relation to task design and label structure [Artstein and Poesio 2008]. Disagreement may also reveal properties of the phenomenon rather than simple annotation error [Passonneau and Carpenter 2014]. This matters especially for intensity: annotators may agree on direction while calibrating force differently.

For modeling, we compare sparse TF-IDF representations and contextual transformer fine-tuning. Sparse linear models remain strong for high-dimensional text classification [Salton and Buckley 1988, Manning et al. 2008, Joachims 1998]. BERT provides contextual bidirectional representations [Devlin et al. 2019]; BERTimbau adapts this architecture to Brazilian Portuguese [Souza et al. 2020]. Repeated stratified evaluation and class weighting address the small and imbalanced label distributions [Kohavi 1995, He and Garcia 2009].

3. Corpus and Annotation Scheme

3.1. Corpora and unit of analysis

The main corpus contains 1,497 public Bluesky posts collected with lexical combinations related to artificial intelligence and education. AI terms included *inteligência artificial*, *IA*, *ChatGPT*, and related expressions; educational terms included *educação*, *ensino*, *professor*, *aluno*, *estudante*, *escola*, and *faculdade*. The collection was designed as a focused discourse corpus, not as a representative survey of platform users.

The annotation unit is the visible collected post. Visible quoted or preview text is considered when it is present in the record. Annotators do not open external links or inspect profiles. This restriction prevents unequal access to context and makes the decision reproducible from the stored item. User handles are omitted from the paper, and corpus examples are translated and lightly paraphrased.

A separate external corpus contains 989 records from 2024. It supports temporal projection and diagnostic review, not retraining. A global audit found 973 unique text hashes, nine duplicate groups, and 124 clearly non-Portuguese records. Both applicability pipelines excluded all 124 from the projected matrices.

3.2. Applicability and the final decision flow

Applicability is a gate rather than a matrix axis. A post is applicable when the visible text provides enough evaluative framing about AI in education to support acceptance and intensity labels. Topical co-occurrence alone is insufficient. Neutral applicable posts discuss an evaluation, controversy, benefit, risk, or transformation without adopting a clear pole; non-applicable posts merely mention or circulate the topic.

The two annotators first labeled applicability independently, yielding 173 raw conflicts. A coordinated anonymized revision and calibration stage then reviewed systematic disagreements and boundary cases against the manual, using coded annotator identifiers. When the effective values converged, that value entered `final_applicability_score`; remaining conflicts were resolved in the same coordinated anonymized review and recorded as adjudicated item-level labels. This process yielded final decisions for all items: 734 applicable and 763 excluded. Final 100% coverage is therefore adjudication coverage, not raw inter-annotator agreement.

3.3. Acceptance, intensity, and operational examples

Acceptance uses five ordered values: -2 strong rejection, -1 critical rejection or concern, 0 neutral or ambivalent orientation, $+1$ prudent or conditional acceptance, and $+2$ strong acceptance. Intensity uses values 1-5 and captures textual force rather than the social importance of the topic. Punctuation, capitalization, irony, affective vocabulary, intensifiers, and rhetorical framing inform intensity.

Table 1. Operational distinctions from the annotation manual. Examples are translated and paraphrased from corpus items.

Decision	Operational criterion	Paraphrased example
Not applicable	AI and education co-occur, but no evaluative framing is available in the visible text.	"How should the AI generation be prepared? See what changes in education."
Applicable, 0	A benefit, risk, controversy, or impact is framed, but the post remains balanced or does not adopt a pole.	"Experts discuss benefits and challenges of AI in digital education."
Acceptance -2	Categorical rejection, catastrophic prediction, or unqualified condemnation.	"Artificial intelligence will destroy the future of education."
Acceptance -1	Critical concern, risk framing, or rejection without maximal categorical force.	"Commentary on the risks of using AI in education."
Acceptance $+1$	Conditional support, guarded endorsement, or acceptance tied to ethics, training, or limits.	"AI can support education when its use is ethical and guided."
Acceptance $+2$	Unqualified endorsement or transformation framed as clearly beneficial.	"AI is revolutionizing education and improving learning."
Intensity 1	Informational wording with little evaluative marking.	A report announces a survey on AI's effects in public education.
Intensity 4-5	Strong affect, capitalization, repeated punctuation, categorical lexis, or emphatic irony.	"It is extremely terrifying that everything in education is now AI!"

The axes are independent. A negative post can be calm, and a conditionally positive post can be forceful. Irony is annotated from the evaluation constructed by the utterance, not from literal positive words. Journalistic wording is applicable when the visible headline or summary itself frames risk, benefit, replacement, revolution, or another evaluative category.

3.4. Granularity and mapping audit

The original labels form a 5×5 matrix. We also derive a 3×3 representation. Acceptance maps $-2, -1$ to negative, 0 to neutral, and $+1, +2$ to positive. Intensity maps 1 to low, $2, 3$ to medium, and $4, 5$ to high.

All contiguous three-bin intensity mappings were audited. The adopted mapping reached macro-F1 $= .614$, balanced accuracy $= .666$, QWK $= .593$, and MAE $= .379$ between annotators. The mapping with the marginally highest QWK ($.601$) grouped levels 1-2 as low and level 3 as medium, but reduced macro-F1 to $.574$ and increased MAE to $.452$. Alternatives with higher macro-F1 collapsed the high class to level 5 alone. We retain $1/(2, 3)/(4, 5)$ because it preserves the manual's distinction between weak, moderate, and strongly marked formulations while maintaining balanced empirical behavior.

3.5. Annotation procedure and consensus subsets

Two annotators with linguistic training used a detailed manual covering the decision hierarchy, labels, irony, mixed targets, journalistic calls, insufficient context, and confidence. Calibration focused on topical relevance versus axiological applicability and on the difference between topic strength and textual intensity.

After final applicability was established, agreement for the axes was computed on the 733 applicable posts with paired valid axis labels. One applicable item had only the second annotator's axis values and was excluded from pairwise agreement calculations. Disagreements on acceptance and intensity were preserved. For supervised hard-label experiments, we extracted 395 items with exact 3-class consensus and 244 with exact 5-class consensus. These high-confidence subsets do not estimate the full distribution of applicable posts.

3.6. Ethics and reproducibility

The study uses public posts and does not reproduce usernames. Examples in the paper are short paraphrases. The annotation manuals, aggregate tables, seeds, hyperparameters, environment snapshot, and analysis scripts are available in the anonymous artifact repository.[1] The recorded environment used for the original training runs was Python 3.11.13, scikit-learn 1.7.0, PyTorch 2.5.1, Transformers 4.41.0, CUDA 12.1, and an NVIDIA RTX 4070 Laptop GPU.

4. Experimental Setup

4.1. Tasks and data splits

We model five tasks: applicability over 1,497 items; acceptance and intensity in three classes over 395 consensus items; and acceptance and intensity in five classes over 244 consensus items. The 3-class axis tasks are the main supervised setting. The 5-class tasks preserve the full analytical matrix and are reported as fine-grained experiments.

All experiments use repeated stratified cross-validation with five folds and seeds 42, 2026, and 7. Out-of-fold predictions are pooled once per item within each seed. Metrics are computed on the complete OOF set for that seed, then summarized by mean and standard deviation across the three seeds. This aggregation is applied identically to sparse models, BERTimbau, and baselines.

4.2. Models and hyperparameters

Sparse models use raw text or a stopword-filtered variant. Word TF-IDF uses 1-2 grams, `min_df=2`, and at most 30,000 features. Character TF-IDF uses `char_wb` 3-5 grams, `min_df=2`, and at most 50,000 features. Logistic Regression uses $C = 1$, L2 regularization, the `lbfgs` solver, balanced class weights, and 3,000 iterations. LinearSVC uses $C = 1$, L2 regularization, squared-hinge loss, balanced class weights, and 1,000 iterations.

BERTimbau uses `neuralmind/bert-base-portuguese-cased`, a maximum length of 160 tokens, dynamic padding, three epochs, batch size 8, learning rate 2×10^{-5}, weight decay .01, and fold-specific class-weighted cross-entropy. Training uses no early stopping or checkpoint-based selection. Raw text is retained because preliminary comparisons showed that aggressive filtering removes function words, punctuation, hashtags, and other cues relevant to axiological interpretation.

4.3. Metrics, baselines, and paired comparisons

Macro-F1 is the primary classification metric. We also report weighted-F1, accuracy, balanced accuracy, Cohen's kappa, QWK, and MAE where appropriate. Majority and stratified-random classifiers provide trivial baselines. Per-class precision, recall, and F1

[1] `https://anonymous.4open.science/r/axiological_matrix_repo`

are computed for each seed and summarized across seeds.

For the three tasks with BERTimbau, we compare the contextual model with the strongest sparse configuration using paired item-level bootstrap confidence intervals and paired permutation tests. These tests use the median OOF prediction across the three seeds for each item. They complement, rather than replace, the seed-level repeated-CV summaries.

4.4. Disagreement-aware experiment

Consensus-only training can select easier items. We therefore run a complementary experiment over all applicable posts with available labels. Each item receives the empirical label distribution of the two annotators. We compare unweighted soft labels, class-balanced soft labels, consensus-only training, a corrected hard compromise, and models trained separately on each annotator. Unweighted soft labels target probabilistic fidelity; balanced soft labels emphasize minority-class recovery. We report cross-entropy against the human distribution, Brier score, macro-F1, QWK, and performance on disagreement items.

4.5. External 2024 projection and diagnostic review

Two pipelines are applied to 989 external records from 2024. The linear pipeline uses the selected sparse models. The hybrid pipeline uses BERTimbau for applicability and 3-class intensity, where contextual modeling was strongest, while retaining the selected sparse components for acceptance and the fine-grained tasks. Thus, applicability and 3-class intensity provide the independent linear-versus-hybrid comparison; shared components are not interpreted as independent agreement.

The raw 2024 collection is audited for language, exact duplicate texts, length, URLs, hashtags, mentions, emojis, punctuation, and possible truncation. Both 3×3 and 5×5 projections are retained because they answer different analytical questions. Since the tasks are trained independently, disagreement across granularities is reported as cross-granularity predictive disagreement rather than treated as an annotation inconsistency.

A blind 50-item challenge set was stratified across five regions: both pipelines excluded; both included with the same 3×3 cell; linear-only inclusion; hybrid-only inclusion; and both included with different cells. The same items were reviewed by the annotators and adjudicated. Because the sample deliberately enriches difficult cases, it is used for diagnostic error analysis rather than as an unbiased performance estimate for the 989 records.

5. Results

5.1. Corpus composition, agreement, and consensus selection

Table 2. Final corpus composition.

Subset	n
Total	1,497
Applicable	734
Excluded	763
3-class consensus	395
5-class consensus	244

Table 3. Raw applicability and paired axis agreement.

Task	n	Exact	κ	QWK	MAE
Applicability (raw)	1,497	0.884	0.770	–	–
Acceptance 5	733	0.630	0.515	0.800	0.434
Intensity 5	733	0.442	0.282	0.667	0.694
Acceptance 3	733	0.779	0.652	0.764	0.259
Intensity 3	733	0.630	0.399	0.594	0.378

Raw applicability reached .884 exact agreement (95% bootstrap CI .868-.900) and $\kappa = .770$ (95% CI .739-.800). The 173 conflicts were resolved through the coordinated anonymized review and adjudication flow described in Section 3. Table 3 therefore separates raw reliability from final decision coverage.

Acceptance is more stable than intensity at both granularities. In three classes, acceptance reaches exact agreement .779 and $\kappa = .652$, while intensity reaches .630 and

$\kappa = .399$. The 5-class QWK values remain substantially higher than unweighted kappa, showing that many disagreements are local on the ordinal scale.

Table 4. Selection effect of exact-consensus subsets. Rates compare retained and removed applicable items.

Subset	Retained	Removed	Acceptance agree	Intensity agree
3-class	395 (53.8%)	339	100% / 52.1%	100% / 19.8%
5-class	244 (33.2%)	490	100% / 44.6%	100% / 16.4%

Table 4 confirms that consensus extraction selects easier cases. The removed items contain most of the interpretive uncertainty, especially for intensity. The annotators' divergence is also directional: on the 5-point intensity scale, annotator 2 assigns a higher value in 368 cases and a lower value in 41, with a mean shift of $+0.58$. Acceptance shows no comparable global shift. These patterns motivate both the consensus and disagreement-aware evaluations.

5.2. Human matrix distributions

Figures 1 and 2 show the human consensus distributions at the two analytical resolutions.

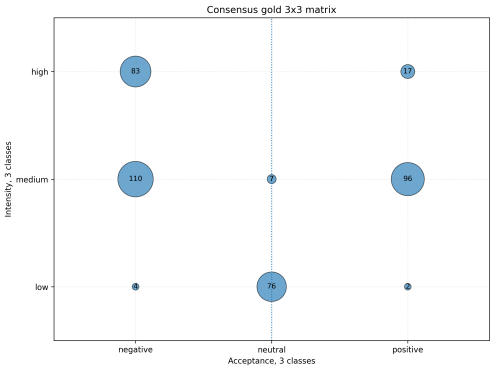


Figure 1. Human consensus 3×3 matrix ($n = 395$).

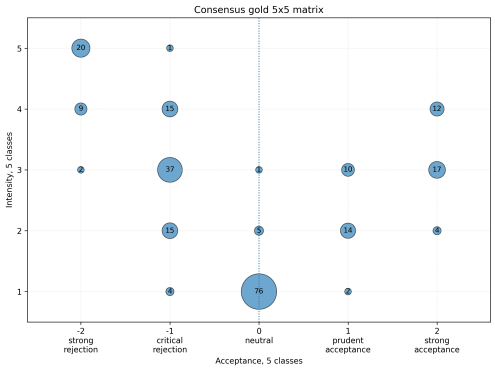


Figure 2. Human consensus 5×5 matrix ($n = 244$).

The 3×3 matrix contains 197 negative, 83 neutral, and 115 positive items. Neutral cases concentrate at low intensity: 76 of 83 appear in that region. Negative and positive items occur mainly at medium and high intensity. The 5×5 view refines this structure. Strong rejection (-2) is concentrated at intensity 4-5, while prudent acceptance ($+1$) occurs mainly at intensity 2-3. The two matrices therefore represent the same annotation scheme at different analytical resolutions.

5.3. Classification results

Table 5. Complete macro-F1 comparison; bold marks the best mean per task.

Model	Applic.	Accept. 3	Intens. 3	Accept. 5	Intens. 5
Word LogReg (raw)	0.861	0.744	0.605	0.556	0.435
Char LogReg (raw)	0.859	0.731	0.612	0.514	**0.472**
Word LinearSVC (raw)	0.872	**0.746**	0.609	**0.561**	0.460
Word LogReg (filtered)	0.857	0.685	0.590	0.525	0.370
Char LogReg (filtered)	0.856	0.685	0.603	0.495	0.419
Word LinearSVC (filtered)	0.858	0.680	0.573	0.518	0.401
BERTimbau weighted	**0.903**	**0.746**	**0.664**	–	–
Majority baseline	0.338	0.222	0.234	0.101	0.101
Stratified random	0.496	0.328	0.326	0.192	0.209

Table 6. Selected results after common seed-level aggregation.

Task	Model	Macro-F1	QWK	MAE
Applicability	BERTimbau weighted	0.903 ± 0.003	–	–
Acceptance 3	BERTimbau weighted	0.746 ± 0.012	0.669 ± 0.013	0.309 ± 0.016
Intensity 3	BERTimbau weighted	0.664 ± 0.009	0.659 ± 0.007	0.346 ± 0.014
Acceptance 5	Word LinearSVC	0.561 ± 0.010	0.625 ± 0.049	0.534 ± 0.045
Intensity 5	Char LogReg	0.472 ± 0.037	0.621 ± 0.003	0.702 ± 0.023

All learned models substantially exceed the trivial baselines. BERTimbau is strongest for applicability and 3-class intensity. For 3-class acceptance, BERTimbau (.7463) and raw word LinearSVC (.7460) are practically tied. Raw-text variants consistently outperform their filtered counterparts, particularly for acceptance, showing that function words, punctuation, hashtags, and related surface cues inform axiological classification.

BERTimbau's applicability gain over the strongest sparse model is robust: the paired macro-F1 difference is $+.031$ with a 95% bootstrap interval $[+.017, +.047]$ and $p = .0005$. Acceptance shows no macro-F1 difference ($-.002$, $p = .953$); BERTimbau's QWK advantage of $+.058$ is suggestive but its interval includes zero. For intensity, the macro-F1 interval also includes zero, while the QWK improvement is clear ($+.144$, 95% CI $[+.063, +.227]$, $p = .0015$). Contextual modeling therefore contributes most reliably to the ordinal structure of evaluative force.

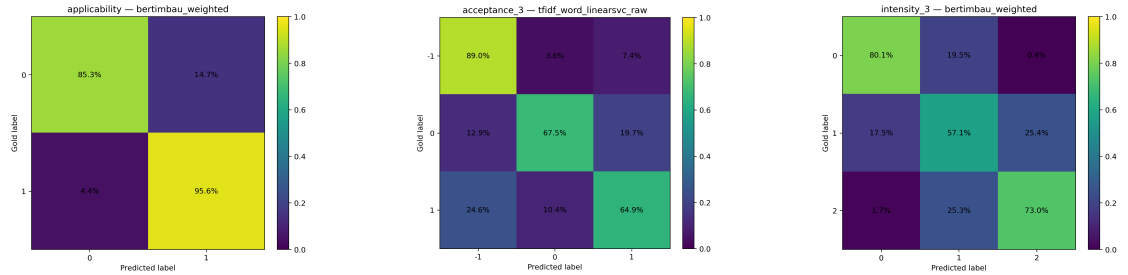
Table 7. Paired contextual-minus-sparse comparisons over median OOF predictions.

Task	Metric	Difference	95% bootstrap CI	p_{perm}
Applicability	Macro-F1	+0.031	[+0.017, +0.047]	0.0005
Acceptance 3	Macro-F1	-0.002	[-0.049, +0.051]	0.9530
Acceptance 3	QWK	+0.058	[-0.003, +0.127]	0.0860
Acceptance 3	MAE	-0.030	[-0.089, +0.023]	0.3363
Intensity 3	Macro-F1	+0.053	[-0.005, +0.112]	0.0765
Intensity 3	QWK	+0.144	[+0.063, +0.227]	0.0015
Intensity 3	MAE	-0.061	[-0.124, +0.003]	0.0700

Table 8. Per-class performance of selected models; means across seeds, with F1 standard deviation.

Task	Class	Precision	Recall	F1	Support
Applicability	Excluded	0.953	0.853	0.900 ± 0.004	763
	Applicable	0.862	0.956	0.907 ± 0.003	734
Acceptance 3	Negative	0.818	0.890	0.853 ± 0.008	197
	Neutral	0.747	0.675	0.709 ± 0.024	83
	Positive	0.707	0.649	0.677 ± 0.005	115
Intensity 3	Low	0.632	0.801	0.704 ± 0.016	82
	Medium	0.748	0.571	0.647 ± 0.022	213
	High	0.574	0.730	0.642 ± 0.011	100
Acceptance 5	-2	0.522	0.333	0.401 ± 0.055	31
	-1	0.627	0.787	0.698 ± 0.006	72
	0	0.795	0.789	0.792 ± 0.024	82
	+1	0.409	0.205	0.272 ± 0.061	26
	+2	0.596	0.697	0.642 ± 0.034	33
Intensity 5	1	0.683	0.858	0.760 ± 0.006	82
	2	0.533	0.298	0.381 ± 0.084	38
	3	0.428	0.478	0.452 ± 0.022	67
	4	0.363	0.287	0.319 ± 0.012	36
	5	0.521	0.397	0.448 ± 0.081	21

Table 8 and Figure 3 localize the remaining difficulty. Applicability prioritizes recall for applicable posts (.956) at the cost of lower precision (.862), which is appropriate for a first-stage gate intended to avoid losing evaluative material. Negative acceptance is easiest in the 3-class task. In the 5-class task, prudent acceptance ($+1$) has F1 .272, reflecting its combination of endorsement, conditions, and critical vocabulary. Intensity level 4 is the hardest fine-grained class (F1 .319), and most errors occur at adjacent levels.



Applicability: BERTimbau | Acceptance 3: LinearSVC | Intensity 3: BERTimbau

Figure 3. Row-normalized confusion matrices averaged across three seeds.

5.4. Disagreement-aware results

Table 9. Corrected disagreement-aware experiment over all applicable items. CE is cross-entropy against the empirical human label distribution.

Task	Regime	Global CE	Disagreement CE	Mean macro-F1	Mean QWK
Acceptance	Soft, unweighted	0.746	1.124	0.658	0.599
Acceptance	Soft, balanced	0.762	1.075	0.663	0.589
Acceptance	Consensus only	0.755	1.105	0.660	0.580
Acceptance	Hard compromise	0.786	1.078	0.649	0.581
Intensity	Soft, unweighted	0.845	0.978	0.464	0.316
Intensity	Soft, balanced	0.884	0.979	0.550	0.488
Intensity	Consensus only	0.870	0.982	0.534	0.432
Intensity	Hard compromise	0.847	0.995	0.507	0.381

For acceptance, unweighted soft labels provide the best global probabilistic fit, while balanced soft labels perform best on disagreement cases and yield the highest mean

macro-F1. For intensity, unweighted soft labels again minimize global cross-entropy and slightly improve the disagreement subset; balanced soft labels produce the best hard-label macro-F1 and QWK. The distinction is expected: class weighting improves minority recovery but changes probability calibration. Disagreement-aware learning is therefore most informative when its objective is explicit, not reduced to one score.

5.5. External 2024 projection

The 2024 corpus contains 989 records and 973 unique texts. The language audit identifies 122 English posts, one German post, one Norwegian post, one bilingual Portuguese-English post, and one very short Portuguese post. All 124 clearly non-Portuguese records are excluded by both applicability gates and therefore do not enter the matrices. The corpus also contains 147 posts with a URL or domain, 101 with emojis, 166 with ellipsis, and 167 with at least one all-caps word, confirming that platform-specific cues matter for intensity.

The linear gate predicts 773 applicable posts and the hybrid gate 752. Their exact agreement is .855 with $\kappa = .591$. Table 10 decomposes the comparison. The pipelines are completely stable, as both excluded or same 3×3 cell for 547 records (55.3%). The remaining records differ at the gate or intensity stage.

Table 10. Regions of the 2024 linear/hybrid comparison.

Region	n	%
Both excluded	155	15.67%
Both applicable, same 3x3 cell	392	39.64%
Both applicable, different 3x3 cell	299	30.23%
Linear applicable, hybrid excluded	82	8.29%
Linear excluded, hybrid applicable	61	6.17%

Figures 4-7 compare the 2024 projections across pipelines and granularities.

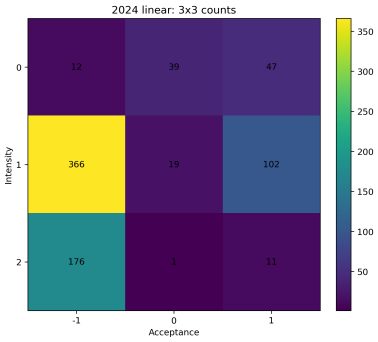


Figure 4. 2024 linear 3×3.

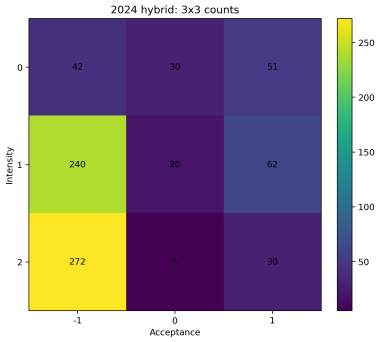


Figure 5. 2024 hybrid 3×3.

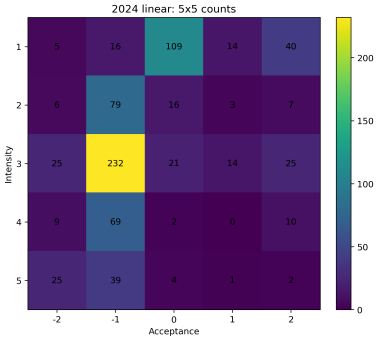


Figure 6. 2024 linear 5×5.

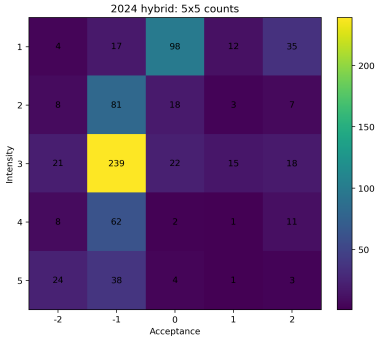


Figure 7. 2024 hybrid 5×5.

Among the 691 posts accepted by both gates, the pipelines assign the same 3-class intensity to 392 (56.7%). The hybrid model assigns higher intensity in 194 cases (28.1%) and lower intensity in 105 (15.2%). Its distribution shifts medium cases toward high intensity, especially in the negative region. This is the principal independent difference between the two pipelines; acceptance is produced by the same selected sparse component.

Both matrix granularities are retained. The 3-class view is the more robust aggregate representation; the 5-class view preserves distinctions such as critical versus strong rejection and prudent versus strong acceptance. Because coarse and fine classifiers are trained independently, their predictions need not be nested item by item. Cross-granularity disagreement occurs in 18.1% of linear acceptance and 28.6% of linear intensity predictions; the corresponding hybrid rates are 17.8% and 50.3%. These rates reflect independently trained predictors, not a defect in the human annotation scheme.

The adjudicated 50-item challenge set was used to inspect the five comparison regions. Its stratified design deliberately overrepresents disagreements, so it is not used to estimate population accuracy. The review supports the error categories already visible in the corpus audit: non-Portuguese leakage in the raw collection, off-topic lexical matches, insufficient visible context, and boundary decisions in applicability and intensity.

6. Discussion

The results support the hierarchical structure of the task. Topical retrieval alone would place 763 unsuitable posts into the matrix. Raw applicability agreement was already substantial, but the 173 conflicts show that the boundary requires explicit rules. Separating raw agreement from final coverage avoids treating adjudication as independent reliability.

Acceptance and intensity behave differently under annotation and modeling. Acceptance is more stable, and sparse lexical models remain highly competitive. Intensity is harder and shows a systematic calibration shift between annotators. BERTimbau's strongest contribution is not simply a higher accuracy score; it produces substantially better ordinal agreement for 3-class intensity. This fits the linguistic character of the axis, which depends on interactions among lexis, punctuation, irony, and rhetorical framing.

The consensus subsets provide supervised targets, but Table 4 shows their selection effect: models evaluated only on consensus cases face a less ambiguous distribution. The disagreement-aware experiment complements this setting by preserving information discarded by consensus extraction. Balanced and unweighted soft-label variants serve different goals, which is more informative than arbitrarily forcing posts into a label.

The 3×3 and 5×5 matrices should be read together. The coarse matrix supports reliable classification and compact corpus comparison. The fine matrix retains the theoretical distinctions encoded by the annotation manual and exposes specific difficulties, especially prudent acceptance and upper-middle intensity. Because independently trained granularities can disagree, matrix-level results are preferable to deterministic nesting.

The 2024 projection demonstrates both the utility and the limits of the framework. The matrices make shifts between pipelines visible, especially the contextual model's upward movement of intensity. The linguistic audit shows that automatic collection can retain language and topical noise, while the gate prevents clearly non-Portuguese material from entering the analysis. Still, the projection describes the collected corpus and the behavior of the trained models; it is not a statistical estimate of Brazilian public opinion.

7. Limitations

Bluesky users are not representative of the Brazilian population, and search-based collection introduces lexical and platform biases. The main supervised axis experiments use high-confidence consensus subsets, which underrepresent ambiguous posts despite the complementary disagreement-aware analysis. The external 2024 corpus contains duplicate texts and is evaluated primarily through model projection; its manually reviewed challenge set is intentionally stratified and cannot provide an unbiased corpus-wide accuracy estimate. The 5-class tasks have small class counts, especially prudent acceptance and intensity level 5. Finally, the framework models the evaluation constructed in visible text and does not

recover author intention or external conversational context.

Because Bluesky relies on a decentralized architecture, future replications should preserve corpus snapshots and collection metadata, since platform access, indexing behavior, and data availability may change over time.

8. Conclusion

This paper presented an axiological matrix, annotation protocol, corpus, and computational evaluation for Brazilian Portuguese Bluesky discourse on AI in education. The framework separates an applicability gate from acceptance and intensity, preserving orientation and textual force as distinct dimensions. Raw applicability agreement reached .884 with $\kappa = .770$, and the adjudication process produced final decisions for all 1,497 posts.

The experiments show that BERTimbau is strongest for applicability and for the ordinal modeling of 3-class intensity, while sparse models remain competitive or superior for acceptance and the fine-grained tasks. Consensus extraction improves label certainty but selects easier items; soft-label experiments show how human divergence can be modeled directly. The 2024 projection, language audit, diagnostic human review, and paired $3\times3/5\times5$ matrices demonstrate how the framework can support transparent temporal analysis without treating model output as ground truth.

The resulting contribution is methodological as much as predictive. It provides a way to represent public evaluations of AI in education that preserves applicability, direction, force, granularity, and disagreement within one auditable pipeline.

References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bakhtin, M. M. (1986). *Speech Genres and Other Late Essays*. University of Texas Press, Austin.

Barbosa, M. Y. A. (2025). Inteligência artificial e educação: Uma análise de interações discursivas na rede social x. Undergraduate thesis, Universidade Federal de Campina Grande, Campina Grande, Brazil. In Portuguese.

Chan, C. K. Y. and Hu, W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1):43.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L., and Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, 7(12):2054–2057.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, Boston, MA.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML 1998)*, pages 137–142. Springer.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Kleppmann, M., Frazee, P., Gold, J., Graber, S., Jenson, S., Kamber, D., Little, D., Newbold, N., Proctor-Rogers, J., and Ripa, E. (2024). Bluesky and the AT protocol: Usable decentralized social media. In *Proceedings of the ACM on Human Factors in Computing Systems (CHI)*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145. Morgan Kaufmann.

Küçük, D. and Can, F. (2021). Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martin, J. R. and White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan, London.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. Association for Computational Linguistics.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*, volume 2 of *Foundations and Trends in Information Retrieval*. Now Publishers.

Passonneau, R. J. and Carpenter, B. (2014). The benefits of a model of annotation. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 311–326.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518. Association for Computational Linguistics.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*, pages 403–417. Springer.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., Agyemang, B., Adzraku, M., Ayanwale, M. A., and Kottahachchi, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15.

UNESCO (2023). Guidance for generative AI in education and research. Technical report, United Nations Educational, Scientific and Cultural Organization. Official institutional guidance document.

Voloshinov, V. N. (1973). *Marxism and the Philosophy of Language*. Seminar Press, New York.