Linguistic Bias in Teacher Evaluations: Sentiment Analysis and Counterfactual Explanations Reveal Systematic Inequities Correlated with Perceived Course Difficulty

Sesgo Lingüístico en Evaluaciones Docentes: Análisis de Sentimientos y Explicaciones Contrafactuales Revelan Inequidades Sistemáticas Correlacionadas con la Dificultad Percibida

Jorge Calvo-Martin[1,2] **Jesús Serrano-Guerrero**[1] **Jared D.T. Guerrero-Sosa**[1]
Andrés Montoro-Montarroso[1] **Francisco P. Romero**[1] **José A. Olivas**[1]
[1]Department of Information Technologies and Systems,
University of Castilla-La Mancha (UCLM), Ciudad Real, Spain
[2]Universidad Alfonso X El Sabio (UAX), Madrid, Spain
JorgeMiguel.Calvo1@alu.uclm.es, Jesus.Serrano@uclm.es, JaredDavid.Guerrero@uclm.es,
Andres.Montoro@uclm.es, FranciscoP.Romero@uclm.es, JoseAngel.Olivas@uclm.es

Abstract: Teacher evaluation platforms such as RateMyProfessor increasingly inform educational decisions, yet their textual reviews may carry biases unrelated to teaching quality. We investigate whether perceived course difficulty introduces linguistic bias in student comments, analysing 45,835 evaluations (from 49,923 after length filtering) of 260 professors. Using VADER and TextBlob, comments for high-difficulty courses show significantly lower sentiment ($\Delta = 0.499$, Cohen's $d = 0.893$, $p < 0.001$) and a fourfold rise in negative language (39.6% vs. 9.3%); an intra-professor analysis confirms the effect is contextual (99.2% of professors, Wilcoxon $p < 0.001$). A convergent-validity check against a transformer model (Twitter-RoBERTa) corroborates these results ($\rho = 0.745$). Finally, a bias classifier (F1 = 0.80) with Diverse Counterfactual Explanations (DiCE) reveals which linguistic changes would reverse penalization, contributing to explainable provider fairness in educational text.
Keywords: sentiment analysis, counterfactual explanations, teacher evaluation bias, provider fairness.

Resumen: Las plataformas de evaluación docente como RateMyProfessor influyen cada vez más en decisiones educativas, pero sus reseñas pueden contener sesgos ajenos a la calidad docente. Investigamos si la dificultad percibida introduce sesgo lingüístico en los comentarios, analizando 45.835 evaluaciones (de 49.923 tras filtrar por longitud) de 260 profesores. Con VADER y TextBlob, los comentarios de cursos de alta dificultad muestran un sentimiento significativamente menor (Δ = 0,499, d = 0,893, $p < 0{,}001$) y un incremento cuádruple del lenguaje negativo (39,6% vs. 9,3%); un análisis intra-profesor confirma que el efecto es contextual (99,2% de los profesores, Wilcoxon $p < 0{,}001$). Una validación convergente frente a un modelo transformer (Twitter-RoBERTa) corrobora los resultados ($\rho = 0{,}745$). Finalmente, un clasificador de sesgo (F1 = 0,80) con Explicaciones Contrafactuales Diversas (DiCE) revela qué cambios lingüísticos revertirían la penalización, contribuyendo a la equidad explicable del proveedor en texto educativo.
Palabras clave: análisis de sentimientos, explicaciones contrafactuales, sesgo en evaluación docente, equidad del proveedor.

1 Introduction

Online teacher evaluation platforms such as RateMyProfessor have become influential sources of information for students making enrollment decisions and, increasingly, for institutional assessment of teaching quality

(Felton et al., 2004; Boring, 2017). These platforms collect both numerical ratings and free-text comments from students, creating large-scale corpora of educational opinion text. While numerical biases in student evaluations of teaching (SET) have been extensively documented—including effects of gender (Boring, 2017), ethnicity (Reid, 2010), and course characteristics (Uttl, White, and Gonzalez, 2017)—the linguistic dimension of these biases remains largely unexplored.

The intersection of Natural Language Processing (NLP) and educational fairness presents a critical research gap. As institutions increasingly employ automated text analysis tools to process student feedback (Acheampong, Wenyu, and Nunoo-Mensah, 2020), understanding how systematic biases manifest in textual evaluations becomes essential for ensuring equitable assessment. If student comments carry sentiment patterns that correlate with contextual factors rather than teaching quality, NLP-based assessment systems may inadvertently perpetuate or amplify these inequities.

This paper addresses this gap by investigating the relationship between perceived course difficulty and the linguistic characteristics of student evaluations. Our work builds upon a broader provider fairness framework for teacher evaluation systems in which we have previously established, through multivariate stratification analysis, that perceived difficulty introduces systematic numerical biases in professor ratings ($\Delta = 1.250$ points on a 1–5 scale, Cohen's $d = 1.390$, $p < 0.001$), with intra-stratum disparities reaching up to 3.11 points between professors evaluated by students with identical difficulty perceptions, and a Disparate Impact Ratio of 0.710 that violates the EEOC legal threshold of 0.80 (Equal Employment Opportunity Commission, 1978; Feldman et al., 2015). Having established these numerical biases through non-parametric statistical validation (Kruskal-Wallis $H = 13{,}790$, $\eta^2 = 0.276$), the natural question arises: *do student comments carry the same systematic biases as their numerical ratings?*

We focus on *provider fairness* (Burke, Sonboli, and Ordonez-Gauger, 2018)—the protection of professors (service providers) from systematic evaluation biases—rather than the more commonly studied consumer fairness perspective that protects students. This framing aligns with emerging multi-stakeholder fairness frameworks in recommendation systems (Deldjoo et al., 2024). Crucially, the combination of numerical and linguistic bias analysis provides a more complete picture of evaluation inequity than either channel alone.

Our research questions are:

- **RQ1:** Does perceived course difficulty systematically affect the sentiment expressed in student comments?
- **RQ2:** Do these linguistic biases persist at the intra-professor level, i.e., does the same professor name group receive different textual feedback depending on student context?
- **RQ3:** How do linguistic biases relate to numerical rating disparities, and do they constitute an independent channel of inequity?
- **RQ4:** Can counterfactual explanations identify what linguistic features would need to change for a penalized professor to receive equitable textual feedback?

Our contributions are fivefold: (1) we provide empirical evidence that perceived difficulty introduces systematic linguistic bias in teacher evaluations, with large effect sizes ($d = 0.893$); (2) we provide intra-professor evidence that these biases are contextual rather than professor-specific, with 99.2% of professors affected; (3) we quantify the *dual channel of inequity*—numerical and linguistic—that compounds unfairness against professors teaching challenging courses; (4) we connect previously established stratification-based fairness metrics with NLP-derived features, demonstrating that both channels of analysis converge on the same systematic inequities; and (5) we apply Diverse Counterfactual Explanations (DiCE) (Mothilal, Sharma, and Tan, 2020) to NLP features, generating actionable explanations of what linguistic patterns would need to change for a professor not to be classified as unfairly penalized—the first application of counterfactual explainability to linguistic fairness in educational evaluation.

Throughout this paper, we use the term *bias* in the statistical sense of a systematic association between perceived course difficulty and evaluation outcomes, rather than in a

strict causal sense. Our analyses are observational; while the intra-professor design substantially reduces confounding from stable professor-level quality differences, we do not claim full causal identification, and residual confounders (e.g., actual course difficulty) cannot be entirely ruled out.

2 Related Work

2.1 Bias in Student Evaluations of Teaching

Research on Student Evaluation of Teaching (SET) biases spans decades. Uttl, White, and Gonzalez (2017) conducted a meta-analysis of 97 studies, finding no correlation between teaching evaluations and student learning, but a strong negative correlation with perceived difficulty. Felton, Mitchell, and Stinson (2004) demonstrated that gender, difficulty perception, and physical attractiveness significantly predict RateMyProfessor ratings. Boring (2017) provided causal evidence of gender biases in evaluations using randomized experiments. Reid (2010) documented racial bias patterns on RateMyProfessor.

These studies focus primarily on numerical ratings. Recent stratification-based analyses have extended these findings by demonstrating that perceived difficulty explains up to 27.6% of variance in professor ratings ($\eta^2 = 0.276$) and produces Disparate Impact Ratios below the 0.80 legal threshold (Equal Employment Opportunity Commission, 1978), with intra-stratum disparities reaching 3.11 points between professors evaluated by students with identical contextual profiles. However, the textual dimension of evaluation bias—how student language varies with these same contextual factors—has received comparatively little attention, despite its growing importance in automated assessment systems.

2.2 Sentiment Analysis in Educational Contexts

Sentiment analysis has been applied to educational text in various forms. Acheampong, Wenyu, and Nunoo-Mensah (2020) surveyed text-based emotion detection methods, including applications in educational feedback analysis. Studies have examined sentiment in course reviews (Pang and Lee, 2008), but typically focus on predicting ratings from text rather than investigating how contextual variables bias the sentiment itself.

VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014) has become a standard lexicon-based tool for social media sentiment, designed to handle the informal language common in online reviews. TextBlob provides complementary polarity and subjectivity measures based on pattern analysis.

2.3 Fairness in NLP and Recommendation Systems

Algorithmic fairness has become a central concern in NLP research (Blodgett et al., 2020). In recommendation systems, Burke, Sonboli, and Ordonez-Gauger (2018) introduced multi-stakeholder fairness, distinguishing consumer-side from provider-side fairness, later extended by Wu et al. (2021) to jointly optimize both perspectives. Gómez et al. (2022) applied provider fairness to educational recommender systems across geographic contexts, though without examining the textual dimension. Deldjoo et al. (2024) provide a comprehensive survey of fairness in recommender systems, mapping the research landscape and future directions.

The concept of *disparate impact*—originating from the EEOC Four-Fifths Rule (Equal Employment Opportunity Commission, 1978)—has been formalized for algorithmic auditing by Feldman et al. (2015). We adapt this framework to evaluate whether linguistic patterns in educational text produce disparate outcomes for professors.

2.4 Counterfactual Explanations for Fairness

Counterfactual explanations answer the question "what would need to change for a different outcome?" (Wachter, Mittelstadt, and Russell, 2018). Mothilal, Sharma, and Tan (2020) introduced DiCE (Diverse Counterfactual Explanations), which generates multiple diverse counterfactuals to provide richer explanations. Counterfactual reasoning has been applied to algorithmic fairness (Kusner et al., 2017) to define individual-level fairness through causal reasoning: an outcome is fair if it would remain the same in a counterfactual world where a protected attribute were different. However, the application of counterfactual explanations to NLP-

derived features in educational contexts remains unexplored. We bridge this gap by applying DiCE to linguistic features extracted from teacher evaluations, generating explanations of what textual patterns would need to change for a professor not to be classified as unfairly penalized.

3 Methodology

3.1 Dataset

We analyze a strategically filtered subset of RateMyProfessor data comprising 49,923 evaluations from 260 professors, extracted from 9.5 million original records through a quality-focused filtering algorithm developed as part of a broader provider fairness research program. An initial exploration on a 20,000-record sample revealed insufficient variability for meaningful bias detection (0% disparities detected), motivating the strategic extraction. The filtering criteria require each professor to simultaneously satisfy: (a) high rating variability (mean range = 3.90 on a 1–5 scale), (b) confirmed presence of both low (1–2) and high (4–5) star ratings, and (c) sufficient evaluation volume (mean = 192 evaluations per professor). All 260 professors meet 100% of quality criteria. It should be noted that, as a publicly available dataset, RateMyProfessor does not provide unique professor identifiers; the analysis unit is therefore the *professor name group*, which may aggregate evaluations of homonymous instructors across different institutions. This is a known limitation of this platform's data and is discussed further in Section 5.5.

Each evaluation contains a numerical star rating (1–5), a perceived difficulty score (1–5), and a free-text comment (mean length: 39.8 words). After removing comments with fewer than 5 words, 45,835 evaluations remain for NLP analysis (91.8% retention). Students are stratified by perceived difficulty into three groups: Low (≤ 2), Medium ($= 3$), and High (≥ 4), yielding a balanced distribution (37.8%, 25.5%, 36.7% respectively). This stratification variable was selected because it achieves 92.1% professor coverage, compared to 46.3% for attendance and only 10.1% for academic performance.

Dataset composition and coverage. The corpus spans 290 academic departments at institutions across the United States; the most represented disciplines are English (5,061 evaluations), Mathematics (4,528), History (3,325), Psychology (2,410), and Biology (2,220). All comments are written in English, and the evaluations were posted between 1999 and 2018 (the bulk concentrated in 2003–2018). Two attributes relevant to interpreting our results are *not* available in the public data: (i) student demographic information (e.g., gender, age), so we make no claims about demographic bias; and (ii) structured course-level metadata such as a course code or academic level (undergraduate vs. graduate). Consequently, perceived difficulty is measured at the level of the individual evaluation rather than the course, and the evaluations grouped under a given professor name may correspond to different courses (Section 5.5).

3.2 Sentiment Analysis Pipeline

We employ a dual-method sentiment analysis approach for robustness:

VADER Sentiment Analysis. VADER (Hutto and Gilbert, 2014) produces a compound score in $[-1, +1]$ calibrated for social media text. Comments are classified as Positive (> 0.05), Neutral ($[-0.05, 0.05]$), or Negative (< -0.05). VADER is particularly suited to our corpus given its handling of emphasis (capitalization, exclamation marks), slang, and emoticons common in student reviews.

TextBlob Analysis. TextBlob[1] provides complementary polarity ($[-1, +1]$) and subjectivity ($[0, 1]$) scores based on a pattern-based lexicon, offering a second perspective for triangulation.

Convergent validity check. To address the concern that lexicon-based tools may not be adapted to the educational domain, we validate our VADER measurements against a transformer-based sentiment model, Twitter-RoBERTa (Loureiro et al., 2022), trained on ∼124M tweets and fine-tuned for sentiment. On a stratified random sample of 1,800 evaluations (600 per difficulty stratum, fixed seed), we compute a continuous transformer sentiment score as $P(\text{positive}) - P(\text{negative})$ and assess (i) its rank correlation with the VADER compound score and (ii) whether the difficulty effect replicates under the transformer (Section 4).

[1] `https://textblob.readthedocs.io`

3.3 Linguistic Feature Extraction

Beyond sentiment scores, we extract the following features:

- **Lexical features:** Counts of positive/negative domain-specific words from a manually curated educational lexicon (18 positive terms, e.g., *clear, helpful, engaging*; 21 negative terms, e.g., *boring, confusing, unfair*), as well as hedging words and intensifiers.
- **Stylistic features:** Exclamation mark frequency, question mark frequency, capitalization ratio, and average word length.
- **Diversity measures:** Type-token ratio (TTR) as a measure of lexical diversity.

3.4 Statistical Framework

All statistical tests are non-parametric, as Likert-scale data and sentiment scores violate normality assumptions (Cohen, 1988):

- **Mann-Whitney U test**: For two-group comparisons (Low vs. High difficulty).
- **Kruskal-Wallis H test**: For three-group comparisons across all difficulty levels.
- **Cohen's** d: Effect size magnitude (small > 0.2, medium > 0.5, large > 0.8).
- **Bootstrap confidence intervals**: 10,000 replications for robust estimation.
- **Wilcoxon signed-rank test**: For paired intra-professor comparisons.
- **Spearman correlation** (ρ): For monotonic associations between variables.
- **Partial correlation**: Sentiment–rating association controlling for difficulty.

3.5 Intra-Professor Analysis

To distinguish contextual bias from professor quality differences, we conduct an intra-professor analysis. For the 260 professors with evaluations in both Low and High difficulty strata, we compute the sentiment shift:

$$\Delta S_p = \bar{S}_{p,\text{Low}} - \bar{S}_{p,\text{High}} \tag{1}$$

where $\bar{S}_{p,s}$ is the mean VADER compound score for professor p in stratum s. A positive ΔS_p means that professor p's comments are, on average, more positive under low difficulty than under high difficulty (equivalently, more negative under high difficulty); the sign alone does not indicate the magnitude or the absolute sentiment level in either stratum.

3.6 Counterfactual Explanation Pipeline

To move beyond bias detection toward explainable fairness, we apply Diverse Counterfactual Explanations (DiCE) (Mothilal, Sharma, and Tan, 2020) to NLP-derived features.

Penalization labeling. Using the intra-professor sentiment shift ΔS_p, we define professors in the top quartile ($\Delta S_p \geq P_{75}$) as *severely penalized*—those whose textual feedback degrades most dramatically with difficulty perception. This yields a binary classification problem: 65 severely penalized vs. 195 non-penalized professors.

Feature matrix. For each professor, we aggregate 14 NLP features from their high-difficulty evaluations (the context where bias manifests): VADER compound, positive, and negative scores; TextBlob polarity and subjectivity; counts of positive, negative, hedging, and intensifier words; exclamation and question mark frequency; capitalization ratio; type-token ratio (TTR); and average word length.

Bias classifier. We train a Random Forest classifier (200 trees, max depth 8, balanced class weights) evaluated via stratified 5-fold cross-validation. Feature importance is extracted to identify which linguistic dimensions best predict penalization.

DiCE generation. For each penalized professor, DiCE generates 4 diverse counterfactual instances (Mothilal, Sharma, and Tan, 2020) that flip the prediction from "penalized" to "not penalized," revealing which NLP features would need to change—and in what direction—to achieve equitable classification. This provides actionable insights: rather than merely detecting bias, we identify specific linguistic patterns that drive the penalization.

4 Results

4.1 RQ1: Sentiment Bias by Perceived Difficulty

Table 1 presents the VADER compound sentiment scores by difficulty stratum. Students

who perceive the course as highly difficult produce comments with substantially lower sentiment than those who perceive low difficulty, with a large effect size.

Diff.	n	Mean	Std	95% CI
Low	17,304	0.650	0.437	[.644, .657]
Medium	11,691	0.524	0.536	[.515, .534]
High	16,840	0.151	0.660	[.141, .161]

Table 1: VADER Compound Sentiment by Difficulty Level.
$\Delta = 0.499$, $d = 0.893$, $p < 0.001$.
Bootstrap 95% CI: [0.487, 0.511].
Kruskal-Wallis $H = 5{,}796.1$, $\eta^2 = 0.126$.

The sentiment difference of $\Delta = 0.499$ on the VADER compound scale represents a shift from clearly positive sentiment (0.650) to marginally positive (0.151) as difficulty perception increases. The effect size ($d = 0.893$) is classified as large by Cohen's conventions (Cohen, 1988). The Kruskal-Wallis test confirms significance across all three levels ($H = 5{,}796.1$, $p < 0.001$), with difficulty explaining 12.6% of sentiment variance ($\eta^2 = 0.126$).

The sentiment class distribution reveals a striking asymmetry (Table 2). For low-difficulty courses, 88.7% of comments are positive and only 9.3% negative. For high-difficulty courses, positive comments drop to 56.4% while negative comments increase to 39.6%—a fourfold increase.

Diff.	VADER	Pos.	Neu.	Neg.
Low	0.650	88.7	2.0	9.3
Medium	0.524	80.6	3.0	16.4
High	0.151	56.4	4.0	39.6

Table 2: Sentiment Class Distribution (%) by Difficulty, with mean VADER compound score per level.
VADER = mean compound score. $\chi^2 = 5{,}134.0$, $p < 0.001$, df = 4.

The TextBlob analysis corroborates these findings: polarity drops from 0.274 (Low) to 0.051 (High), with $\Delta = 0.222$ and $d = 0.786$ ($p < 0.001$). The convergence of two independent sentiment tools strengthens the validity of our findings.

4.2 Linguistic Feature Analysis

Table 3 presents the linguistic feature analysis across difficulty strata. Students in high-difficulty courses use significantly more negative vocabulary (2.6× the rate of low-difficulty), while positive word usage decreases. Subjectivity remains relatively stable across conditions, suggesting the bias affects valence rather than objectivity.

Feature	Low	Med	High	p
Neg. words	0.193	0.273	0.494	<.001
Pos. words	0.654	0.549	0.326	<.001
Excl. marks	1.321	1.198	1.016	<.001
Subjectivity	0.592	0.576	0.579	<.001
Hedging words	0.086	0.094	0.103	<.001

Table 3: Linguistic Features by Difficulty Level (Mean).

4.3 RQ2: Intra-Professor Sentiment Shift

The intra-professor analysis provides the strongest evidence that linguistic bias is contextual rather than quality-driven. Among the 260 professors with evaluations in both Low and High difficulty strata, **258 (99.2%)** show lower mean sentiment in the high-difficulty context.

The mean intra-professor sentiment shift is $\Delta S = 0.514$ ($p < 0.001$, Wilcoxon signed-rank test $W = 60.0$). This means that the *same professor name group*, evaluated by different groups of students, receives textual feedback that is systematically more negative when students perceive the course as difficult. This finding substantially reduces the influence of stable professor-level quality differences, strengthening the interpretation of perceived difficulty as a contextual correlate of textual sentiment.

4.4 Robustness to the Sentiment Tool

A potential concern is that our sentiment measurements rely on lexicon-based tools not specifically adapted to the educational domain. To address this, we validate VADER against a transformer-based model, Twitter-RoBERTa (Loureiro et al., 2022), on a stratified sample of 1,800 evaluations (Table 4). The transformer score correlates strongly with the VADER compound score (Spearman $\rho = 0.745$, $p < 0.001$; Pearson $r = 0.770$), indicating that our lexicon-based measurements capture the same sentiment signal as a domain-proximate transformer. Critically, the difficulty effect replicates under

	VADER	T-RoBERTa
Mean (Low)	0.641	0.605
Mean (High)	0.123	−0.188
Cohen's d	0.918	1.100
Spearman $\rho = 0.745$ ($p < 0.001$)		

Table 4: Convergent validity: VADER vs. Twitter-RoBERTa on a stratified sample of 1,800 evaluations (600 per stratum). Rows report mean sentiment by stratum and the Low-vs-High effect size.

Twitter-RoBERTa: mean transformer sentiment decreases monotonically from Low (0.605) through Medium (0.427) to High (−0.188) difficulty, with a large effect size (Cohen's $d = 1.100$, Mann-Whitney $p < 0.001$, Kruskal-Wallis $H = 293.9$, $p < 0.001$) that even exceeds the VADER-based estimate on the same sample ($d = 0.918$). This confirms that our central finding is not an artefact of the lexicon-based approach.

4.5 RQ3: Dual Channel of Inequity

The Spearman correlation between VADER compound sentiment and star rating is $\rho = 0.622$ ($p < 0.001$), confirming strong alignment between linguistic and numerical evaluations. Crucially, the partial correlation controlling for difficulty remains high ($\rho_{\text{partial}} = 0.615$), indicating that sentiment and ratings share variance beyond what difficulty explains.

Metric	Rating	Sentim.
Mean (Low diff.)	4.314	0.650
Mean (High diff.)	3.065	0.151
Δ	1.250	0.499
Cohen's d	1.390	0.893
η^2 (var. expl.)	27.6%	12.6%
p-value	<0.001	<0.001
Intra-prof. aff.	–	99.2%

Table 5: Dual Channel: Numerical vs. Linguistic Bias.
"–": the intra-professor rate for the numerical channel is not computed here; it is the focus of the companion stratification study.

Table 5 compares the numerical and linguistic channels. Both exhibit large effect sizes, though the numerical channel shows stronger effects ($d = 1.390$ vs. $d = 0.893$). This convergence is significant for two reasons. First, it validates that the numerical biases detected through our stratification framework are not statistical artifacts—they are independently confirmed by linguistic patterns in the same evaluations. Second, it reveals that difficulty perception creates a *dual bias*: professors teaching challenging courses are penalized both in their star ratings *and* in the language used to describe their teaching, with the two channels reinforcing each other.

The practical consequence is severe. Consider that our prior stratification analysis identified cases where professors with identical student difficulty profiles received ratings differing by up to 3.11 points. The NLP analysis now shows that these same professors also receive qualitatively different textual feedback, with 99.2% experiencing measurably more negative language in high-difficulty contexts. This dual penalization compounds the unfairness: a professor may receive both a low numerical score *and* a negative textual review, not because of poor teaching, but because of the perceived difficulty of their course.

4.6 RQ4: Counterfactual Explanations for Penalized Professors

Classifier performance. The Random Forest classifier achieves F1 $= 0.800 \pm 0.064$ and accuracy $= 0.892 \pm 0.036$ in 5-fold cross-validation, outperforming Gradient Boosting (F1 $= 0.756 \pm 0.098$). This confirms that NLP features from high-difficulty evaluations can reliably distinguish severely penalized professors.

Feature importance. Table 6 shows the top predictive features. VADER compound score dominates (33.0%), followed by VADER negative (18.1%) and positive (13.8%) components, and TextBlob polarity (9.4%). The dominance of sentiment-level features over lexical counts suggests that overall evaluative tone, rather than specific word choices, best predicts penalization severity.

Counterfactual analysis. DiCE generated counterfactual explanations for 40 penalized professors (160 counterfactual instances). Table 7 presents the mean feature changes required to flip a professor from "penalized" to "not penalized." The dominant change is an increase in VADER compound

Feature	Imp.
VADER compound	0.330
VADER negative	0.181
VADER positive	0.138
TextBlob polar.	0.094
Positive words	0.078
Negative words	0.036
Intensifiers	0.024
Excl. marks	0.024

Table 6: Feature Importances for Bias Classifier.

of +0.340 (modified in 90.6% of counterfactuals), indicating that the overall sentiment tone of high-difficulty evaluations is the primary lever. Secondary changes include increasing positive word count (+0.016, 6.2%), decreasing negative words (−0.011, 5.6%), and adding intensifiers (+0.015, 4.4%).

Feature	$\overline{\lvert\Delta\rvert}$	Dir.	%Ch.
VADER comp.	0.340	↑	90.6
Excl. marks	0.028	↓	6.9
Pos. words	0.016	↑	6.2
Intensifiers	0.015	↑	4.4
Neg. words	0.011	↓	5.6
TB polarity	0.009	↑	4.4
Hedging	0.006	↑	5.6

Table 7: DiCE Counterfactual Feature Changes.

A concrete example. To illustrate at the individual level, consider a penalized professor whose high-difficulty evaluations average a VADER compound of −0.134. DiCE produces four diverse counterfactuals that each flip the classification to "not penalized" via a different route: raising the compound to +0.144 ($\Delta = +0.279$) by lowering negative sentiment (VADER negative 0.149 → 0.046); to +0.494 ($\Delta = +0.629$) by increasing positive-word density (0.11 → 0.61); to +0.363 ($\Delta = +0.498$) by reducing exclamation use; and to +0.117 ($\Delta = +0.251$) via a combination of smaller shifts. In every case the compound sentiment rises, mirroring at the instance level the aggregate pattern of Table 7—for a severely penalized professor to receive equitable evaluations, student comments would need to shift in overall sentiment by roughly +0.34 on the VADER scale.

5 Discussion

5.1 Implications for NLP-Based Educational Assessment

Our findings have direct implications for systems that apply NLP techniques to teacher evaluation text. If automated systems extract sentiment from student comments to assess teaching quality—whether through lexicon-based methods, machine learning classifiers, or large language models—they will systematically disadvantage professors who teach courses perceived as difficult. This represents a form of *algorithmic bias amplification*: the human bias in text is captured and potentially magnified by NLP processing.

From a provider fairness perspective (Burke, Sonboli, and Ordonez-Gauger, 2018; Gómez et al., 2022), this creates an unjust evaluation landscape. Professors have limited control over students' difficulty perceptions, which are influenced by subject matter, institutional expectations, and student preparation (Uttl, White, and Gonzalez, 2017). Penalizing professors based on language patterns driven by these external factors violates principles of procedural fairness.

Critically, the linguistic evidence presented here corroborates our prior finding that numerical ratings exhibit a Disparate Impact Ratio of 0.710, below the EEOC threshold of 0.80 (Equal Employment Opportunity Commission, 1978; Feldman et al., 2015): any sentiment-based metric mined from these reviews will inherit—and potentially amplify—the same systematic bias, doubly disadvantaging the 54.5% of professor pairs that already show significant numerical disparity within the same difficulty stratum.

5.2 Methodological Contributions

The dual-tool approach (VADER + TextBlob) with non-parametric testing provides a robust methodology for bias detection in educational text. The intra-professor design—comparing the same professor name group across difficulty contexts—reduces confounding due to individual teaching style by holding the name group constant while varying the student difficulty context.

The linguistic feature analysis reveals that bias manifests across multiple dimensions:

not only in overall sentiment, but in specific lexical choices (more negative vocabulary, fewer positive expressions) and stylistic features (fewer exclamation marks, more hedging language in high-difficulty contexts).

5.3 From Detection to Explanation: The DiCE Contribution

The application of counterfactual explanations to NLP-derived fairness features represents a novel methodological contribution: while traditional bias detection identifies *that* bias exists, counterfactual explanations identify *what would need to change* to eliminate it. The dominant counterfactual change—a +0.34 shift in VADER compound—provides a concrete, measurable target for fairness-aware text processing systems, bridging bias detection (Blodgett et al., 2020) and actionable intervention. Combined with the classifier's strong performance (F1 = 0.80), these results open the door to automated fairness monitoring systems that flag potentially biased textual evaluations before they influence institutional decisions.

5.4 Lessons Learned for Future Analyses and Tools

Beyond the specific findings, our study yields several practical lessons for researchers and for developers of NLP-based assessment tools:

- **Stratify by difficulty before interpreting sentiment.** Aggregating sentiment from evaluation text without conditioning on perceived difficulty conflates teaching quality with course context; future analyses should stratify at the evaluation level and, where possible, adopt intra-instructor designs.
- **Audit and monitor text-derived metrics for disparate impact.** Any tool that derives teaching-quality signals from review text should be tested for difficulty-driven disparate impact before deployment, and the dominant counterfactual target (+0.34 in VADER compound) can be operationalised in fairness dashboards that flag reviews departing from difficulty-adjusted expectations.
- **Lightweight lexicon tools are defensible when validated.** The strong agreement between VADER and a domain-proximate transformer ($\rho = 0.745$) indicates that reproducible, low-cost lexicon pipelines are adequate for difficulty-bias auditing, provided they are spot-checked against a transformer model.
- **Prefer unique instructor identifiers.** Name-based grouping conflates homonymous instructors; datasets and tools intended for instructor-level fairness analysis should rely on stable unique identifiers.

5.5 Limitations

This study has several limitations. First, our corpus is English-language and from a single platform (RateMyProfessor), limiting generalizability to other languages and evaluation systems. Second, our primary sentiment analysis is lexicon-based (VADER, TextBlob); although we validated these measurements against a transformer model (Twitter-RoBERTa) and found strong convergence ($\rho = 0.745$; Section 4.4), fine-grained patterns such as sarcasm or domain-specific irony may still be imperfectly captured. Third, we cannot fully disentangle difficulty perception from actual course difficulty, though our intra-professor design mitigates this concern substantially. Fourth, the counterfactual explanations are generated from aggregated professor-level features; individual comment-level counterfactuals could provide finer-grained insights. Fifth, the penalization threshold (top 25% of sentiment shift) is a design choice that could be adjusted for different institutional contexts. Sixth, professor identification relies on the `professor_name` field, which does not provide unique identifiers; a single name may aggregate evaluations of homonymous instructors across institutions. The review-level findings (RQ1, differential sentiment by difficulty stratum) are unaffected, as they operate at the individual evaluation level. The intra-professor results (RQ2) reflect within-name-group consistency; a re-analysis with a deduplicated unit (name + institution + department) is reserved for future work.

Generalizability and selection bias. Our strategic filtering—retaining professors with high rating variability and a guaranteed presence of both low and high ratings—deliberately departs from a random sample of the platform and therefore warrants

closer scrutiny. This choice was necessary: an initial random sample of 20,000 reviews exhibited insufficient rating variation to detect within-professor disparities (0% detected), rendering bias analysis impossible. Two considerations bound its impact. First, the selection operates at the *professor* level (overall rating spread), whereas our central effect is measured at the *evaluation* level and, for RQ2, *within* each professor across difficulty strata; cross-professor selection thus does not mechanically inflate the intra-professor sentiment shift. Second, the excluded professors are those with near-uniform ratings, for whom a difficulty-driven disparity is by construction small or undetectable; our results therefore generalize to the population of professors with non-trivial evaluation variation—precisely those for whom evaluation-based decisions carry the highest stakes. We nonetheless caution against extrapolating the absolute magnitudes (e.g., the 99.2% intra-professor figure) to the full platform population, and we identify replication on an unfiltered, adequately powered sample as important future work.

6 Conclusions and Future Work

This paper demonstrates that perceived course difficulty introduces systematic linguistic bias in teacher evaluation comments and provides explainable counterfactual pathways toward equitable evaluation. Through sentiment analysis of 45,835 evaluations and counterfactual explanation of 260 professor profiles, we show that:

1. Comments for high-difficulty courses carry significantly lower sentiment ($\Delta = 0.499$, $d = 0.893$) with a fourfold increase in negative language (39.6% vs. 9.3%).
2. This bias is contextual: 99.2% of professors receive more negative textual feedback in high-difficulty contexts (Wilcoxon $p < 0.001$).
3. Linguistic and numerical biases form a dual channel of inequity, with Spearman $\rho = 0.622$ between sentiment and ratings.
4. A bias classifier trained on NLP features (F1 = 0.80) reliably identifies severely penalized professors, and DiCE counterfactuals reveal that a sentiment increase of +0.34 is the primary change needed to reverse penalization.

These findings raise important concerns for the growing use of NLP in educational assessment: any system that mines student evaluation text without accounting for contextual biases risks systematically disadvantaging professors who teach challenging courses. The convergence between our numerical stratification framework (DIR = 0.710, disparities up to 3.11 points) and the present linguistic analysis ($d = 0.893$, 99.2% of professors affected) provides compelling dual-channel evidence of systematic inequity, while the counterfactual explanations provide actionable targets for fairness-aware evaluation systems.

Future work will explore several directions. First, while we validated our sentiment measurements against a pretrained transformer (Section 4.4), transformer models *fine-tuned* for educational text could capture more nuanced linguistic patterns such as sarcasm or domain-specific irony. Second, the Choquet integral, for which we have conducted preliminary interaction analysis between stratification variables, can serve as a multi-criteria aggregation mechanism that compensates for both numerical and linguistic biases simultaneously. Third, cross-lingual validation with Spanish-language university evaluation datasets would extend these findings to the Ibero-American educational context. Finally, the integration of counterfactual explanations into real-time fairness monitoring dashboards could enable institutions to identify and mitigate linguistic bias as evaluations are submitted.

Acknowledgements

This work was carried out within the doctoral programme in Advanced Computing Technologies at the University of Castilla-La Mancha. We thank the anonymous reviewers for their constructive feedback, which substantially improved the paper.

References

Acheampong, F. A., C. Wenyu, and H. Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Blodgett, S. L., S. Barocas, H. D. III, and H. Wallach. 2020. Language (technology) is power: A critical survey of bias in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5454–5476.

Boring, A. 2017. Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41.

Burke, R., N. Sonboli, and A. Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pages 202–214.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition.

Deldjoo, Y., D. Jannach, A. Bellogin, A. Difonzo, and D. Zanzonelli. 2024. Fairness in recommender systems: Research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34:59–108.

Equal Employment Opportunity Commission. 1978. Uniform guidelines on employee selection procedures. 29 CFR Part 1607.

Feldman, M., S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Felton, J., J. Mitchell, and M. Stinson. 2004. Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, 29(1):91–108.

Gómez, E., C. W. S. Zhang, L. Boratto, M. Salamó, and G. Ramos. 2022. Enabling cross-continent provider fairness in educational recommender systems. *Future Generation Computer Systems*, 127:435–447.

Hutto, C. J. and E. Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 216–225.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Loureiro, D., F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 251–260.

Mothilal, R. K., A. Sharma, and C. Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 607–617.

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

Reid, L. D. 2010. The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education*, 3(3):137–152.

Uttl, B., C. A. White, and D. W. Gonzalez. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54:22–42.

Wachter, S., B. Mittelstadt, and C. Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887.

Wu, Y., J. Cao, G. Xu, and Y. Tan. 2021. TFROM: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1022.