CLASSIFICAÇÃO SUPERVISIONADA DE OPINIÃO EM RESPOSTAS ABERTAS DA EDUCAÇÃO A DISTÂNCIA

Rafael Ribeiro Brant Nobre[1] - rafael.rbn.gb@gmail.com
Lucas Rocha Aguilar[1] - rochaaguilar07@gmail.com
Guilherme Lopes Antunes[1] - guilhermelantunes123@gmail.com
Alysson Steve Mota Lacerda[1] - steve.lacerda@unimontes.br
Patrícia Takaki[1] - patricia.takaki@unimontes.br
[1]Universidade Estadual de Montes Claros, UNIMONTES - Montes Claros, MG, Brasil

Resumo. *A Educação a Distância (EaD) tem gerado grandes volumes de respostas abertas em questionários de pesquisas de satisfação, cuja análise manual em larga escala pode ser inviável. Este trabalho aplica uma metodologia de classificação supervisionada de opinião a um corpus de 1.167 respostas de acadêmicos da EaD de uma instituição pública brasileira, anotadas manualmente em três classes de polaridade (negativo, neutro, positivo). Foram avaliadas cinco variantes de pré-processamento textual e quatro classificadores clássicos (Naive Bayes, SVM, Regressão Logística e Random Forest) sobre representação TF-IDF, com validação cruzada agrupada por texto para evitar vazamento de dados. A melhor configuração, Regressão Logística com stoplist customizada preservando negações e intensificadores, alcançou acurácia de 75,66% e macro-F1 de 0,7305, muito acima do baseline de classe majoritária (macro-F1 = 0,2027). A classe neutra apresentou o menor desempenho (F1 = 0,5844) e concentrou cerca de 45% dos erros envolvendo respostas verdadeiramente neutras. A análise qualitativa dessas respostas mal classificadas mostra que grande parte delas são mensagens curtas e protocolares, afirmações factuais sem termos avaliativos explícitos, ou sugestões corteses com carga emocional ambígua, evidenciando os limites de um esquema de três classes.*

Palavras-Chave: *Mineração de textos, Análise de sentimentos, Classificação supervisionada, Anotação manual, Educação a distância*

1. INTRODUÇÃO

A Educação a Distância (EaD) tem ampliado o acesso ao ensino superior no Brasil, gerando grandes volumes de respostas abertas em questionários de pesquisas de satisfação, aplicados periodicamente pelas instituições de ensino superior aos seus acadêmicos, fonte relevante para

compreender a percepção dos estudantes, mas cuja análise manual em larga escala pode se mostrar inviável. Essa limitação tem motivado o uso de mineração de textos (Feldman & Sanger, 2007) e análise de sentimentos (Liu, 2012; Pang & Lee, 2008), abordagem já consolidada na avaliação em ensino superior (Grimalt-Álvaro & Usart, 2024). Na educação superior a distância, classificação de textos e análise de sentimentos estão entre as tarefas de mineração de textos mais frequentes (Takaki & Dutra, 2024), e as aplicações voltadas à satisfação discente concentram-se na avaliação da atividade docente e no retorno para o redesenho de estratégias didáticas (Chamorro-Atalaya et al., 2025). Na literatura, predominam métodos baseados em léxico, que dispensam anotação manual, e métodos supervisionados, que exigem dados rotulados e que capturam melhor as nuances semânticas do domínio (Jayasudha & Thilagu, 2022). Em cenários de textos curtos e poucos dados, como respostas abertas de pesquisas de satisfação, modelos clássicos com TF-IDF (*Term Frequency-Inverse Document Frequency*) são competitivos com, ou superiores a, aprendizagem profunda e métodos de dicionário (Sunar & Khalid, 2023; Shaik et al., 2022). Os autores Sunar e Khalid (2023) mostram que a anotação manual predomina em estudos supervisionados de feedback educacional (82,6% dos casos), e a literatura converge para a classificação em três classes (positivo, negativo, neutro) como padrão, embora escalas mais granulares também sejam relatadas (Grimalt-Álvaro & Usart, 2024). Diante desse cenário, este trabalho aplica e avalia uma metodologia de classificação supervisionada de opinião, com anotação manual em três classes, sobre respostas abertas de estudantes da EaD, comparando estratégias de pré-processamento e classificadores clássicos, e investiga em profundidade os fatores que dificultam a identificação automática de respostas neutras, limitação recorrente e pouco discutida na literatura.

2. METODOLOGIA

A metodologia possui caráter descritivo e quantitativo, estruturada em três etapas: construção do corpus anotado manualmente, pré-processamento textual e classificação supervisionada com validação cruzada.

2.1 Corpus e anotação manual

O corpus é composto por 6.697 respostas abertas de pesquisas de satisfação aplicadas em cursos de graduação e pós-graduação a distância da Universidade Estadual de Montes Claros (Unimontes), via Sistema Universidade Aberta do Brasil (UAB). A Tabela 1 detalha as etapas sucessivas de limpeza aplicadas até a obtenção do corpus final.

Tabela 1: Etapas de limpeza do corpus, do total bruto ao corpus final

Etapa	Removidas	Restantes
Total bruto coletado	–	6.697
Remoção de respostas vazias (”sem resposta”)	5.081	1.616
Remoção de respostas em escala Likert (concordo/discordo)	382	1.234
Remoção de respostas com apenas símbolos/pontuação	27	1.207
Remoção de respostas vazias após normalização	4	1.203
Exclusão manual de texto aleatório/sem sentido na anotação	36	**1.167**

As respostas remanescentes após a limpeza automática (aprox. 1.200) foram ordenadas

por contagem de palavras, para priorizar a rotulação das mais curtas, e agrupadas por texto normalizado, o que reduziu em 18% o esforço de rotulação ao evitar reclassificar textos idênticos. Cada resposta foi então classificada manualmente por um único anotador em três classes de polaridade: negativo (-1), neutro (0) e positivo (+1), esquema consistente com o padrão dominante na literatura (Grimalt-Álvaro & Usart, 2024; Sunar & Khalid, 2023). A distribuição de classes do corpus final é apresentada na Tabela 2.

Tabela 2: Distribuição das classes no corpus anotado manualmente

Classe	n	%
Negativo	510	43,70
Neutro	282	24,16
Positivo	375	32,13
Total	**1.167**	**100,00**

A razão entre a classe majoritária e a minoritária é de 1,81, um desbalanceamento moderado, tratado na etapa de classificação por meio de ponderação de classes. Como limitação, destaca-se que a anotação foi realizada por um único anotador, estando sujeita à variabilidade interpretativa típica de respostas com sentimento misto, ponto retomado na discussão da Seção 3.2.

2.2 Pré-processamento textual

Todas as variantes compartilham de uma base comum: conversão para minúsculas, remoção de acentos, pontuação, números e tokens com menos de dois caracteres. Sobre ela, foram criadas cinco versões cumulativas: v1 (apenas base); v2 (base + *stopwords* do NLTK, `https://www.nltk.org`); v3 (base + *stopwords* customizadas); v4 (v3 com *stemming* RSLP); e v5 (base + lematização via spaCy, `https://spacy.io` + *stopwords* da v3). A lista customizada parte dos 200 termos da lista padrão e preserva os 6 que carregam polaridade (”não”, ”nem”, ”sem”, ”só”, ”muito” e ”mais”), ou seja, retira-os do conjunto de termos a serem eliminados, de modo que permaneçam no texto. A ela somam-se 7 termos de ruído do domínio (”ead”, ”uab”, ”unimontes”, ”curso”, ”disciplina”, ”aula” e ”aulas”), totalizando 201 termos. A hipótese era testar se evitar a eliminação indiscriminada de *stopwords*, preservando negações e intensificadores, produziria ganhos na classificação.

2.3 Representação vetorial e classificação

Os textos foram representados por TF-IDF (escala sublinear), em duas configurações de *n*-gramas (unigramas; unigramas+bigramas). Avaliaram-se quatro classificadores: Naive Bayes multinomial, SVM (*Support Vector Machine*) linear (*LinearSVC*), Regressão Logística e Random Forest (400 árvores), com ponderação de classes balanceada (exceto Naive Bayes). Para evitar vazamento de dados entre treino e teste (respostas curtas idênticas, 19% das linhas), a validação cruzada usou *StratifiedGroupKFold* com 5 *folds*, agrupando por texto normalizado. A métrica de seleção foi o macro-F1, mais robusto que a acurácia ao desbalanceamento de classes, desafio comum em mineração de textos aplicada à educação que tende a enviesar o desempenho de classificação (Shaik et al., 2022). Como os baselines reportados na área costumam ser modelos de ML clássicos, não triviais por frequência de classe (Sunar & Khalid, 2023), adotou-se aqui um classificador trivial de classe majoritária, implementado com o `DummyClassifier` do scikit-learn (`https://scikit-learn.org`), que atribui a cada

documento a classe mais frequente do conjunto de treino e é avaliado sob o mesmo protocolo de validação cruzada dos demais modelos. Sendo N o tamanho do corpus, n_{maj} o número de respostas da classe majoritária e C o número de classes, esse classificador atribui F1 nulo às classes minoritárias, resultando em:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \cdot \frac{2\, n_{\mathrm{maj}}}{N + n_{\mathrm{maj}}} \tag{1}$$

Para o corpus deste trabalho ($N = 1.167$, $n_{\mathrm{maj}} = 510$, $C = 3$), a Eq. (1) resulta em macro-F1 de 0,2027 e acurácia de 0,4370 (n_{maj}/N), valores que coincidem com os obtidos empiricamente sob validação cruzada.

3. RESULTADOS E DISCUSSÃO

3.1 Desempenho dos modelos

A Tabela 3 apresenta o melhor macro-F1 por variante de pré-processamento.

Tabela 3: Melhor macro-F1 por variante de pré-processamento

Variante	Macro-F1
v1 – tokenização bruta	0,7025
v2 – *stopwords* padrão	0,7082
v3 – *stopwords* customizada	0,7295
v4 – customizada + *stemming*	0,7305
v5 – lematização	0,7144

A preservação de negações e intensificadores (v3) elevou o macro-F1 de 0,7082, obtido com a lista padrão (v2), para 0,7295, confirmando a hipótese de que remover indiscriminadamente esses termos prejudica a classificação de sentimentos. O ganho é modesto, o que sugere que a customização adotada teve alcance limitado sobre o vocabulário do corpus. Verificar se uma lista de preservação mais abrangente amplia esse ganho fica como trabalho futuro. A adição de *stemming* (v4) produziu o melhor resultado geral: Regressão Logística, com unigramas mais bigramas, alcançou acurácia de 75,66% e macro-F1 de 0,7305, um ganho de mais de 0,52 ponto sobre o baseline trivial. A Tabela 4 detalha o desempenho por classe.

Tabela 4: Relatório de classificação do melhor modelo (Regressão Logística)

Classe	Precisão	Revocação	F1	n
Negativo	0,7640	0,8824	0,8189	510
Neutro	0,6286	0,5461	0,5844	282
Positivo	0,8378	0,7440	0,7881	375
Total / Macro média	**0,7435**	**0,7242**	**0,7305**	**1.167**

As classes negativo e positivo, com marcadores lexicais mais claros, alcançaram F1 acima de 0,78. A classe neutra, no entanto, apresentou desempenho substancialmente inferior (F1 = 0,5844), tornando-se o principal fator limitante do desempenho geral do modelo, fenômeno investigado em detalhe na próxima seção.

3.2 O desafio da classificação da classe neutra

Do total de 282 respostas verdadeiramente neutras, 89 (31,6%) foram classificadas como negativas e 39 (13,8%) como positivas, totalizando 45,4% de erro, taxa muito superior à das demais classes. A Tabela 5 apresenta uma amostra qualitativa dessas respostas mal classificadas. Os exemplos revelam quatro padrões de ambiguidade: (i) respostas curtas e protocolares; (ii)

Tabela 5: Exemplos de respostas neutras classificadas incorretamente

Resposta original	Predição	Possível fonte de ambiguidade
”Ok”	Positivo	Encerramento protocolar
”Nota cinco”	Negativo	Número sem contexto; sentido depende da escala
”Não tivemos tutores presenciais.”	Positivo	Afirmação factual, sem polaridade explícita
”Gostaria de ter o gabarito das minhas avaliações.”	Positivo	Sugestão cortês; elogio ou queixa velada
”eu queria que as webinar deveria explicar mais a matéria que irá cair na prova”	Negativo	Sugestão com tom de insatisfação implícita
”Difícil avaliar de maneira geral...”	Negativo	Recusa a opinar, mas contém ”difícil”

afirmações factuais sem marcador avaliativo; (iii) sugestões corteses cuja carga emocional depende de interpretação subjetiva; e (iv) recusas a opinar que contêm palavras isoladas de polaridade (como ”difícil”), que confundem o classificador. Esse padrão sugere que a dificuldade não é apenas do modelo, mas do próprio esquema de anotação em três classes: casos-limite, como uma sugestão cortês entre neutro e levemente negativo, dependem de uma escolha subjetiva do anotador. Essa leitura é reforçada pelos termos mais associados a cada classe pelo modelo vencedor (coeficientes da Regressão Logística, em formas radicalizadas). As classes negativo e positivo concentram termos de carga avaliativa clara, como ”prov”/prova e ”horari”/horário nas queixas e ”otim”/ótimo, ”excel”/excelente e ”parab”/parabéns nos elogios. Já a classe neutra é dominada por marcadores protocolares de ausência de opinião (”nad”, ”ok”, ”sem mais”), o mesmo vocabulário da Tabela 5, o que explica por que seus termos mais informativos carregam pouca carga semântica própria. Uma direção promissora é adotar anotação com granularidade mais fina, já que escalas mais detalhadas que a de três classes são relatadas na literatura (Grimalt-Álvaro & Usart, 2024). Dividir a polaridade em mais classes permitiria acomodar sugestões corteses sem forçá-las para ”neutro” ou ”negativo”, e fica como sugestão para trabalhos futuros.

4. CONCLUSÃO

Este trabalho aplicou classificação supervisionada de opinião, com anotação manual em três classes, a respostas abertas de estudantes da EaD, alcançando macro-F1 de 0,7305 com Regressão Logística sobre TF-IDF, bem acima do baseline trivial (0,2027); preservar negações e intensificadores no pré-processamento foi determinante para esse resultado. A principal contribuição é a análise da classe neutra, responsável pelo menor desempenho (F1 = 0,5844)

e por 45,4% de erro entre as respostas verdadeiramente neutras. A análise qualitativa e os termos mais informativos do modelo mostram que boa parte dessas respostas são mensagens curtas e protocolares ou sugestões corteses de carga ambígua, evidenciando que o desafio é também inerente à subjetividade da anotação manual em três classes. Escalas mais granulares, com graus leves de positividade e negatividade, são discutidas como estratégia futura. Os resultados sugerem benefícios potenciais do modelo proposto para a gestão educacional e, por consequência, para a qualidade da educação ofertada. A classificação converte texto disperso em métricas objetivas e comparáveis ao longo do tempo, oferecendo à gestão um panorama da satisfação discente que a leitura manual por amostragem dificilmente sustentaria. O modelo pode ainda ser empregado como recurso para limitar a necessidade de avaliação humana futura. Como limitações, destacam-se a anotação por um único anotador e o uso de dados de uma única instituição. Trabalhos futuros devem incluir múltiplos anotadores com cálculo de concordância, ampliar a lista de *stopwords* customizadas, cuja extensão reduzida ajuda a explicar o ganho modesto observado, e comparar com estratégias baseadas em léxico.

Agradecimentos

Os autores agradecem ao Programa Institucional de Iniciação Científica Voluntária (ICV), Edital PROINIC - PRP 06/2025 da Unimontes. Os autores declaram que ferramentas de Inteligência Artificial Generativa (Claude, Anthropic) foram utilizadas nas etapas de revisão bibliográfica, verificação de citações e dados, e redação e formatação do manuscrito, com a finalidade de apoiar a clareza da escrita e a conformidade com as normas do evento. Todo o conteúdo foi revisado criticamente e validado pelos autores, que assumem integral responsabilidade pela originalidade e veracidade do texto final, em conformidade com a Portaria CNPq nº 2.664/2026.

Referências

Chamorro-Atalaya, O. et al. (2025), "Identification of the satisfaction of university students through sentiment analysis: a systematic review", *International Journal of Evaluation and Research in Education (IJERE)*, vol 14, n. 1, 37-49.

Feldman, R. and Sanger, J. (2007), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge, 410 p.

Grimalt-Álvaro, C. and Usart, M. (2024), "Sentiment analysis for formative assessment in higher education: a systematic literature review", *Journal of Computing in Higher Education*, vol 36, n. 3, 647-682.

Jayasudha, J. and Thilagu, M. (2022), "A survey on sentimental analysis of student reviews using natural language processing (NLP) and text mining", in *International Conference on Innovations in Intelligent Computing and Communications*, Springer International Publishing, Cham, 365-378.

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, San Rafael, 167 p.

Pang, B. and Lee, L. (2008), "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, vol 2, n. 1-2, 1-135.

Shaik, T. et al. (2022), "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis", *IEEE Access*, vol 10, 56720-56739.

Sunar, A.S. and Khalid, M.S. (2023), "Natural language processing of student's feedback to instructors: a systematic review", *IEEE Transactions on Learning Technologies*, vol 17, 741-753.

Takaki, P. and Dutra, M.L. (2024), "Text mining applied to distance higher education: a systematic literature review", *Education and Information Technologies*, vol 29, n. 9, 10851-10878.