UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
Bacharelado em CIÊNCIA DA COMPUTAÇÃO

Técnicas de Análise de Sentimento no Contexto Educacional: um Estudo Comparativo

Rafael Gomes Lopes

JUIZ DE FORA
JANEIRO, 2026

Técnicas de Análise de Sentimento no Contexto Educacional: um Estudo Comparativo

Rafael Gomes Lopes

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
Bacharelado em CIÊNCIA DA COMPUTAÇÃO

Orientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA
JANEIRO, 2026

Técnicas de Análise de Sentimento no Contexto Educacional: um Estudo Comparativo

Rafael Gomes Lopes

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação

Regina Mariel Maciel Braga Villela
Doutora em Engenharia de Sistemas e Computação - COPPE/UFRJ

José Maria Nazar David
Doutor em Engenharia de Sistemas e Computação - COPPE/UFRJ

JUIZ DE FORA

19 DE JANEIRO, 2026

Aos meus amigos e irmãos.

Aos pais, pelo apoio e sustento.

Resumo

A crescente adoção de ambientes virtuais de aprendizagem tem gerado um grande volume de interações textuais entre estudantes e docentes, tornando inviável o acompanhamento manual do feedback discente em larga escala. Nesse contexto, a análise de sentimentos surge como uma ferramenta promissora para apoiar intervenções pedagógicas proativas. Este trabalho realiza um estudo comparativo entre diferentes técnicas de análise de sentimento aplicadas ao domínio educacional, confrontando abordagens léxicas consolidadas (TextBlob e VADER), um modelo especialista treinado especificamente para esse domínio (Pred-inter) e um modelo discriminativo baseado em Transformers disponibilizado pela biblioteca Hugging Face. O estudo utiliza o dataset Stanford MOOCPosts, composto por mensagens de fóruns educacionais rotuladas manualmente em uma escala de sentimento. As saídas das ferramentas foram normalizadas para permitir comparações diretas, sendo avaliadas por meio de métricas quantitativas, como acurácia global e erro médio absoluto, além de uma análise qualitativa das divergências. Os resultados demonstram desempenho amplamente superior do modelo especialista, confirmando a relevância da adaptação de domínio. As abordagens léxicas apresentaram resultados intermediários, enquanto o modelo baseado em Transformers, apesar de seu sucesso em domínios gerais, mostrou desempenho insatisfatório no contexto educacional analisado. Por fim, discute-se a viabilidade do uso de grandes modelos de linguagem, destacando limitações práticas e desafios para sua aplicação efetiva.

Palavras-chave: Análise de Sentimentos; Educação; Processamento de Linguagem Natural; Learning Analytics; Adaptação de Domínio.

Abstract

The increasing adoption of virtual learning environments has generated a large volume of textual interactions between students and instructors, making manual monitoring of student feedback impractical at scale. In this context, sentiment analysis emerges as a promising tool to support proactive pedagogical interventions. This work presents a comparative study of different sentiment analysis techniques applied to the educational domain, contrasting established lexical approaches (TextBlob and VADER), a domain-specific expert model (Pred-inter), and a Transformer-based discriminative model provided by the Hugging Face library. The experimental evaluation is conducted using the Stanford MOOCPosts dataset, which contains forum messages manually annotated with sentiment scores. To enable direct comparison, the outputs of all tools were normalized to a common scale. The analysis combines quantitative metrics, such as overall accuracy and mean absolute error, with a qualitative examination of classification divergences. The results show that the domain-specialized model significantly outperforms the other approaches, reinforcing the importance of domain adaptation in educational sentiment analysis. Lexical methods achieve moderate performance, whereas the Transformer-based model, despite its effectiveness in general-purpose sentiment tasks, performs poorly in the educational context considered. Finally, the feasibility of applying large language models is discussed, highlighting practical limitations and open challenges.

Keywords: Sentiment Analysis; Education; Natural Language Processing; Learning Analytics; Domain Adaptation.

Agradecimentos

A todos os meus parentes, pelo encorajamento e apoio.

Ao professor Victor Ströele pela orientação, amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computução
UFJF	Universidade Federal de Juiz de Fora
PLN	Processamento de Linguagem Natural
MOOC	Massive Open Online Course
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
SVM	Support Vector Machine

1 Introdução

A integração de Tecnologias da Informação e Comunicação no ambiente escolar alterou significativamente a dinâmica de ensino-aprendizagem. A educação híbrida e o uso de ambientes virtuais tornaram-se onipresentes, gerando um volume sem precedentes de dados digitais (ROMERO; VENTURA, 2010). Nesse cenário, as interações dos alunos — que antes se perdiam na oralidade da sala de aula — agora ficam registradas em fóruns, chats e sistemas de avaliação.

O campo da Mineração de Dados Educacionais surge, portanto, como uma disciplina vital para transformar esses "rastros digitais" em conhecimento pedagógico útil (ROMERO; VENTURA, 2010). Dentro deste escopo, a análise do feedback discente é fundamental. A literatura especializada aponta que o acompanhamento contínuo das dúvidas e dos comentários dos alunos é um fator determinante para o sucesso acadêmico e para a retenção (TINTO, 1975). No entanto, o desafio reside na escalabilidade: em cursos massivos ou em grandes plataformas, é inviável para os docentes processarem manualmente o volume de mensagens geradas.

Nesse cenário, surge a Análise de Sentimentos, uma subárea do Processamento de Linguagem Natural (PLN), que busca automatizar a identificação de estados emocionais em textos postados por estudantes, permitindo que gestores atuem de forma proativa. A evasão escolar e o desengajamento são problemas crônicos, especialmente na educação a distância. Segundo (TINTO, 1975), a integração social e acadêmica é fundamental para a retenção dos alunos. Ferramentas que detectam precocemente sentimentos negativos permitem intervenções pedagógicas rápidas, potencialmente reduzindo a evasão e promovendo um ambiente de aprendizagem mais acolhedor.

Embora a Análise de Sentimentos seja uma área consolidada na indústria, aplicada largamente em *reviews* de produtos e mercado financeiro (PANG; LEE, 2008), sua aplicação direta no contexto educacional enfrenta barreiras específicas. A linguagem utilizada por estudantes em ambientes virtuais é frequentemente marcada por informalidade, uso de gírias, ironias e vocabulário técnico específico da disciplina estudada.

Ferramentas generalistas de análise léxica, como o VADER (*Valence Aware Dictionary and sEntiment Reasoner*), demonstraram eficácia em redes sociais (HUTTO; GILBERT, 2014), mas sua precisão em contextos pedagógicos específicos ainda é objeto de debate. Da mesma forma, bibliotecas populares, como o TextBlob, oferecem facilidade de implementação, mas apresentam limitações por analisarem as palavras de forma isolada.

O problema central desta pesquisa reside na incerteza quanto à ferramenta mais adequada para este domínio. Assim, considerando as especificidades do discurso no ambiente educacional, a questão de pesquisa que norteou o desenvolvimento deste trabalho é: "Como o desempenho da ferramenta especializada 'Pred-inter' se compara, em termos de classificação de polaridade, com ferramentas léxicas consolidadas (VADER, TextBlob) e modelos de última geração baseados em Transformers?"

A falta de estudos comparativos que confrontem diretamente modelos treinados especificamente para a educação com grandes modelos de linguagem justifica a investigação.

Apoiando-se na evolução recente das arquiteturas de redes neurais, especificamente o advento dos *Transformers* introduzidos por (VASWANI et al., 2017), e na importância do contexto para a semântica, formula-se a seguinte hipótese:

H1: A ferramenta Pred-inter, por ser especializada no domínio educacional, apresentará desempenho superior ao das abordagens estritamente baseadas em léxicos (TextBlob e VADER). Contudo, espera-se que os modelos baseados em *Transformers* (via Hugging Face), devido ao mecanismo de atenção e ao pré-treinamento massivo contextual, alcancem resultados equivalentes ou superiores em precisão, estabelecendo-se como o novo estado da arte também na mineração de sentimentos educacionais.

Além das abordagens descritas, este trabalho também contempla uma análise exploratória sobre a viabilidade do uso de Modelos de Linguagem de Grande Porte (LLMs) na análise de sentimentos em contexto educacional. Embora esses modelos tenham apresentado avanços significativos em tarefas gerais de compreensão de linguagem, sua aplicação prática neste estudo encontrou limitações relevantes. Em razão dessas restrições, não foi possível obter resultados quantitativos suficientemente estáveis para uma comparação direta com as demais ferramentas avaliadas. Dessa forma, o uso de LLMs é

tratado neste trabalho como um estudo de viabilidade, com foco na análise crítica de seus limites e desafios, em vez de uma avaliação de desempenho propriamente dita.

O objetivo deste trabalho de conclusão de curso é realizar um estudo comparativo de eficácia e acurácia na análise de sentimentos em contexto educacional, confrontando o software especialista "Pred-inter" com duas categorias distintas de ferramentas de Processamento de Linguagem Natural: abordagens léxicas (TextBlob e VADER) e modelos discriminativos baseados em Transformers (Hugging Face/BERT).

Objetivos específicos:

1. Pré-processar o dataset: Realizar a limpeza e a normalização de um corpus pré-existente de interações estudantis, adequando-o para a entrada nos diferentes algoritmos.

2. Normalizar as escalas de avaliação: Desenvolver métricas de conversão para alinhar a escala numérica do Pred-inter (0 a 4) às escalas de polaridade das demais ferramentas.

3. Executar a classificação automatizada: Submeter as frases às quatro ferramentas propostas, coletando métricas de polaridade.

4. Comparar métricas de desempenho: Analisar os resultados utilizando indicadores estatísticos de acurácia, confrontando as saídas dos modelos generalistas com as do modelo especialista.

5. Validar a eficácia no domínio: Discutir qualitativamente os erros e acertos de cada ferramenta em relação aos aspectos específicos desta pesquisa.

Validar se modelos clássicos e leves (léxicos) ainda são competitivos frente a modelos computacionalmente custosos (Deep Learning) é essencial para a tomada de decisão na engenharia de software educacional. Conforme (LIU, 2012) destaca, a análise de sentimentos é altamente dependente do domínio; portanto, verificar a eficácia de um modelo treinado especificamente para a educação (Pred-inter) contribui para o avanço da área de *Learning Analytics*.

O presente trabalho está organizado em cinco capítulos, descritos brevemente a seguir:

- Capítulo 2 - Fundamentação Teórica: Apresenta os conceitos-chave de Processamento de Linguagem Natural, detalha as abordagens léxicas versus aprendizado de máquina e revisa trabalhos relacionados à análise de sentimentos na educação.

- Capítulo 3 - Metodologia: Descreve o conjunto de dados utilizado, o funcionamento do software Pred-inter e das bibliotecas comparadas (TextBlob, VADER, Hugging Face), bem como as métricas de conversão de escalas adotadas.

- Capítulo 4 - Análise de Resultados: Exibe os dados obtidos, comparando as classificações através de gráficos e tabelas, discutindo casos de concordância e discordância entre os algoritmos.

- Capítulo 5 - Considerações Finais: Retoma os objetivos, valida a hipótese inicial e sugere trabalhos futuros, concluindo quanto à viabilidade do uso das diferentes ferramentas no contexto educacional.

2 Fundamentação Teórica e Trabalhos Relacionados

Este capítulo fundamenta os conceitos de Processamento de Linguagem Natural (PLN), detalhando a transição de métodos baseados em regras para arquiteturas de Aprendizado Profundo (*Deep Learning*). Discute-se especificamente a mecânica algorítmica por trás das ferramentas comparadas: léxicos (TextBlob, VADER) e modelos transformadores (BERT/Hugging Face).

2.1 PLN e Representação Vetorial

O Processamento de Linguagem Natural evoluiu de sistemas simbólicos para abordagens estatísticas robustas. Segundo (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011), o PLN moderno busca converter a linguagem humana não estruturada em representações numéricas que algoritmos possam processar.

Para que o computador entenda um texto, é preciso traduzir as palavras para a linguagem matemática. Antigamente, os sistemas apenas verificavam se uma palavra existia ou não na frase, tratando sinônimos como 'feliz' e 'contente' como coisas totalmente diferentes, pois a grafia não era a mesma. A grande revolução trazida por (MIKOLOV et al., 2013), com a técnica de *Word Embeddings*, foi a criação de um tipo de 'mapa geográfico' das palavras. Nessa abordagem, cada palavra recebe uma coordenada numérica, e palavras com significados semelhantes são colocadas matematicamente muito próximas umas das outras nesse mapa. Assim, o computador consegue 'perceber' que 'Rei' está para 'Rainha' assim como 'Homem' está para 'Mulher', capturando a semântica por trás das letras.

A Análise de Sentimentos é definida por (PANG; LEE, 2008) como o tratamento computacional de opiniões e de subjetividade em textos. De forma prática, esse processo funciona como um sistema de classificação automática: o objetivo é fazer com que o

computador leia um texto (como o comentário de um aluno) e atribua a ele um 'rótulo' que identifique a emoção predominante, geralmente classificando-o em uma das três categorias básicas: Positivo, Negativo ou Neutro.

No domínio educacional, a análise enfrenta o desafio da "dependência de domínio". (KASTRATI et al., 2021), em um mapeamento sistemático, alertam que modelos treinados em dados genéricos (como resenhas de filmes) falham ao interpretar o vocabulário pedagógico, em que termos como "desafiador" podem denotar engajamento positivo, o que difere do uso comum.

2.2 Abordagens Léxicas e Baseadas em Regras

As abordagens léxicas fundamentam-se na premissa da composicionalidade, segundo a qual o sentimento de uma sentença é derivado da soma ou da média das Orientações Semânticas (SO) de suas palavras constituintes. Diferentemente de modelos estocásticos que exigem treinamento, estes métodos utilizam bases de conhecimento linguístico pré-computadas. Segundo (TABOADA et al., 2011), a principal vantagem desta abordagem é a interpretabilidade: é possível rastrear exatamente qual palavra levou à classificação positiva ou negativa, uma transparência frequentemente ausente em redes neurais profundas.

2.2.1 TextBlob e Heurísticas

A biblioteca TextBlob utiliza uma abordagem baseada nas médias aritméticas da polaridade dos léxicos subjacentes. Embora computacionalmente leve, estudos comparativos, como o de (RIBEIRO et al., 2016), apontam que métodos puramente baseados em dicionários sofrem com a incapacidade de resolver polissemias (palavras com múltiplos significados) sem o auxílio de contexto profundo. Textblob opera como uma interface simplificada para a biblioteca *Pattern*, desenvolvida por (SMEDT; DAELEMANS, 2012). O cálculo de polaridade não é uma simples contagem de palavras, mas um processo de média ponderada de adjetivos conhecidos no léxico da WordNet.

O algoritmo funciona da seguinte maneira:

1. **Tagging Part-of-Speech (POS):** O texto é, primeiramente, processado para identificar a classe gramatical de cada palavra. O TextBlob foca principalmente em adjetivos, pois estes carregam a maior carga subjetiva.

2. **Averiguação de Intensidade:** Cada palavra no léxico possui tuplas de valores, com polaridade ($p \in [-1, 1]$) e subjetividade ($s \in [0, 1]$).

3. **Tratamento de Negação Naïve:** O algoritmo verifica a presença de negadores léxicos (ex.: "não", "jamais") na janela imediatamente anterior ao adjetivo e inverte a polaridade resultante.

Após a atribuição dos valores individuais modificados pelas regras acima, o TextBlob agrega as pontuações para gerar a polaridade global da frase. O cálculo final é realizado por meio da **média aritmética simples** das polaridades de todos os léxicos identificados.

Matematicamente, se uma frase contém n palavras com carga sentimental (após ajustes de negação), a polaridade final P é dada pela Equação 2.1, onde p_i é a polaridade da i-ésima palavra.

O resultado é um valor escalar flutuante no intervalo $[-1, 1]$. Esta abordagem de média simples é justamente o ponto criticado na literatura, pois dilui sentimentos intensos quando a frase contém muitas palavras neutras ou levemente opostas.

$$P = \frac{\sum_{i=1}^{n} p_i}{n} \tag{2.1}$$

2.2.2 VADER (Regras Sintáticas)

O VADER (*Valence Aware Dictionary and sentiment Reasoner*), proposto por (HUTTO; GILBERT, 2014), representa o estado da arte em métodos baseados em léxicos. Diferente de dicionários tradicionais, o léxico do VADER foi validado via *Wisdom of the Crowds* (Sabedoria das Multidões), utilizando avaliadores humanos para calibrar a intensidade de mais de 7.500 itens lexicais, incluindo *emoticons* (ex.: ":-)") e acrônimos de internet (ex.: "LOL").

A robustez do VADER reside na aplicação de cinco regras heurísticas que alteram

a valência da sentença em tempo de execução:

1. **Pontuação:** O uso de exclamações aumenta a magnitude da intensidade (ex.: "Bom!!!" > "Bom").

2. **Capitalização:** Palavras em caixa alta recebem um incremento na pontuação de valência, o que simula ênfase vocal.

3. **Modificadores de Grau:** Advérbios intensificadores (ex.: "extremamente") ou atenuadores (ex.: "levemente") multiplicam ou dividem o escore da palavra seguinte.

4. **Conjunções Adversativas:** O algoritmo detecta a conjunção "mas", reduzindo o peso da oração anterior e aumentando o da posterior, mimetizando a interpretação cognitiva humana de que a segunda parte da frase carrega a opinião dominante.

5. **Janela de Negação Tri-grama:** O VADER verifica as três palavras anteriores a um termo de sentimento para detectar negações, capturando estruturas complexas que escapam a abordagens simples.

Após a soma das valências de todas as palavras da frase (ajustadas pelas heurísticas), o VADER não utiliza uma média simples. Para garantir que o resultado final esteja sempre contido no intervalo normalizado $[-1, 1]$, utiliza-se a seguinte função de normalização não-linear (Equação 2.2). Onde:

- x é a soma bruta das valências das palavras;

- α é uma constante de normalização (geralmente 15), que "suaviza"a curva, aproximando o comportamento de uma função sigmoide.

$$Score_{norm} = \frac{x}{\sqrt{x^2 + \alpha}} \tag{2.2}$$

Essa formulação matemática permite que o VADER compare sentenças de diferentes comprimentos de forma justa.

2.3 Aprendizado Supervisionado

O Aprendizado de Máquina Supervisionado funciona de forma semelhante a um aluno que estuda com exemplos prontos. O computador recebe várias frases que já foram classificadas (o "gabarito") e aprende a identificar padrões nelas. Segundo Kotsiantis (2007), algoritmos dessa categoria, como o SVM, buscam traçar, matematicamente, uma linha divisória clara que separe o grupo de comentários positivos do grupo de negativos.

A grande limitação dos modelos antigos era ler o texto palavra por palavra, muitas vezes 'esquecendo' o começo da frase quando chegavam ao final. A arquitetura Transformer, proposta em (VASWANI et al., 2017), resolveu isso por meio do mecanismo de 'Auto-Atenção'. Em termos simples, esse mecanismo permite que o computador olhe para a frase inteira de uma só vez e decida quais palavras são mais importantes para dar sentido às demais. Por exemplo, na frase 'o banco fechou cedo', a atenção do modelo conecta a palavra 'banco' com 'fechou', entendendo que se trata de uma instituição financeira, e não de um assento de praça.

O modelo BERT, apresentado por Devlin et al. (2019), aprimorou essa ideia ao ler o texto de forma 'bidirecional'. Diferente de nós, que lemos estritamente da esquerda para a direita, o BERT analisa o contexto que vem antes e o que vem depois da palavra simultaneamente. É como um exercício de preencher lacunas: para entender o significado de uma palavra no meio de uma frase, o modelo olha 'para a esquerda' (o que já foi dito) e 'para a direita' (o que será dito) ao mesmo tempo. Para facilitar o uso dessa tecnologia complexa, utiliza-se a biblioteca Hugging Face, que, segundo Wolf et al. (2020), oferece versões otimizadas e mais leves desses modelos.

2.3.1 O Modelo PRED-INTER: Predição de Intervenções Pedagógicas

O Pred-inter (*Prediction of Pedagogical Interventions*) é uma ferramenta computacional proposta por Rossi (2022), desenvolvida para automatizar o suporte ao estudante em Ambientes Virtuais de Aprendizagem. A premissa central do sistema é solucionar o problema de escalabilidade em cursos massivos (MOOCs), em que o volume de interações

nos fóruns supera a capacidade humana de monitoramento por parte do docente.

O sistema Pred-inter pode ser classificado como um modelo de aprendizagem supervisionada, mas com uma vantagem importante: é um especialista. A eficácia desse tipo de ferramenta baseia-se no que Pan e Yang (2010) chamam de Adaptação de Domínio. Os autores explicam que, quando um algoritmo é treinado especificamente com dados da área em que atuará (neste caso, a educação), tende a ser melhor do que ferramentas genéricas. Isso acontece porque ele aprende o vocabulário específico e as gírias dos estudantes, detalhes que ferramentas treinadas com textos gerais da internet costumam perder.

Diferentemente de analisadores de sentimento genéricos, o Pred-inter não se limita a classificar a polaridade do texto. Ele utiliza técnicas de Deep Learning para compor um sistema de decisão que sugere ações pedagógicas concretas.

2.3.2 Corpus de Treinamento e Mapeamento de Classes

A arquitetura do Pred-inter baseia-se no aprendizado supervisionado, utilizando o dataset *Stanford MOOCPosts* como fonte de dados. Este conjunto de dados contém 29.550 postagens extraídas de fóruns de cursos nas áreas de Ciências Humanas, Medicina e Educação.

Originalmente, as frases deste dataset foram rotuladas manualmente por avaliadores humanos em uma escala Likert de 1 a 7. Para a construção do modelo de análise de sentimentos do Pred-inter, Rossi (2022) estabeleceram um mapeamento de discretização dessas notas originais para 5 classes fundamentais (0 a 4), conforme descrito a seguir:

- Classe 0 (Muito Negativo): Notas originais entre 1,0 e 1,5;
- Classe 1 (Negativo): Notas originais entre 2,0 e 3,0;
- Classe 2 (Neutro): Notas originais entre 3,5 e 5,0.
- Classe 3 (Positivo): Notas originais entre 5,5 e 6,0;
- Classe 4 (Muito Positivo): Notas originais entre 6,5 e 7,0.

2.3.3 Arquitetura e Processamento

O treinamento do modelo segue o protocolo padrão de Aprendizado de Máquina, onde 80% do corpus é dedicado ao processo de aprendizado, permitindo que o sistema identifique os padrões linguísticos e as regras de classificação, enquanto os 20% restantes são reservados para testes e validação.

Ao processar uma nova sentença, a saída do Pred-inter não é um valor determinístico único, mas sim uma distribuição de probabilidade sobre as cinco classes possíveis. O valor final (classe predita) é determinado pela classe com maior probabilidade no vetor de saída.

```
[caption={Exemplo de saída do modelo Pred-inter}, label={cod:
   saida_pred}]
Classe predita: 3 (Positivo)
Distribuição: [0.008, 0.010, 0.059, 0.859, 0.061]
```

Nota: Neste caso, o modelo tem 85,9% de certeza de que a frase pertence à classe 3.

2.3.4 Lógica de Intervenção Pedagógica

A inovação do Pred-inter reside na combinação da Análise de Sentimentos com duas outras dimensões afetivas e cognitivas: Urgência e Confusão (ambas classificadas em níveis Alto, Médio e Baixo). O sistema cruza a classe de sentimento inferida com os níveis de urgência e confusão detectados para determinar a intervenção pedagógica mais adequada. Segundo Rossi (2022), as ações resultantes podem ser:

1. Ajuda do Tutor: O professor é notificado para intervir pessoalmente.
2. Ajuda Automática: O sistema sugere materiais complementares (vídeos, artigos) sem intervenção humana.
3. Ajuda da Classe: O sistema notifica outros alunos para promover a aprendizagem colaborativa.
4. Mensagem Automática: Envio de feedbacks motivacionais (ex: reforço positivo para alunos engajados).

A Figura 2.1 ilustra o fluxo de decisão do Pred-inter para a seleção da intervenção.

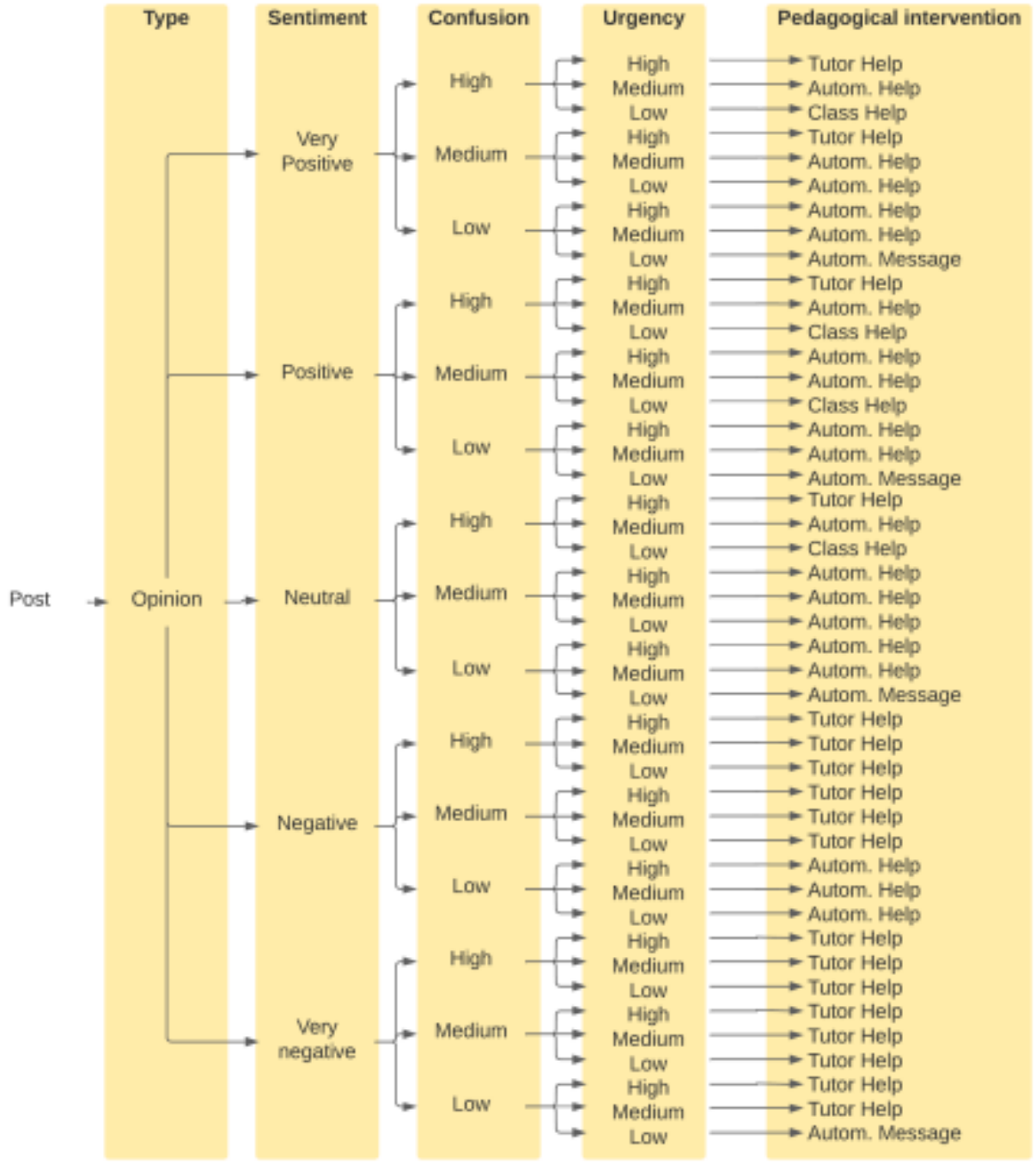

Figura 2.1: Fluxo de decisão para intervenções pedagógicas no Pred-inter.
Fonte: Adaptado de Rossi (2022).

2.4 Trabalhos Relacionados

A literatura sobre Análise de Sentimentos em contextos educacionais apresenta uma clara evolução tecnológica. Esta seção categoriza os trabalhos anteriores em três ondas distintas: *benchmarks* de métodos tradicionais, a consolidação do Deep Learning e a emergência recente dos Grandes Modelos de Linguagem (LLMs).

2.4.1 Comparativos Clássicos: Léxicos vs. Machine Learning

Nos primeiros anos da mineração de dados educacionais, o debate centrava-se entre a eficiência dos métodos léxicos e a eficácia dos classificadores estatísticos (SVM, Naive Bayes). Ribeiro et al. (2016) estabeleceram um marco com o estudo *SentiBench*, demonstrando que o VADER superava métodos de aprendizado de máquina em textos curtos e informais de redes sociais, devido às suas heurísticas de pontuação e negação.

No entanto, Ortigosa, Martín e Carro (2014) aplicaram métodos híbridos em ambientes de E-learning (Facebook) e alertaram para a "dependência de domínio". Os autores constataram que dicionários genéricos falham ao interpretar termos pedagógicos, sugerindo que a adaptação do léxico é crucial para a precisão na educação.

2.4.2 A Era do Deep Learning e BERT na Educação

A segunda onda de pesquisas concentra-se na contextualização semântica. Kastrati et al. (2021) realizaram um mapeamento sistemático da literatura, focado exclusivamente no feedback estudantil. A revisão de 67 estudos primários concluiu que arquiteturas baseadas em Deep Learning superam consistentemente as abordagens léxicas.

Mais especificamente, Chyr et al. (2020) aplicaram o modelo BERT para classificar postagens em fóruns de medicina. Os resultados mostraram que o BERT, graças ao mecanismo de atenção, alcançou acurácias superiores a 90%, estabelecendo os Transformers discriminativos como o "padrão-ouro" para a tarefa, superando as limitações de contexto apontadas anteriormente em (RIBEIRO et al., 2016).

Diante desse panorama, observa-se uma evolução contínua que parte de regras heurísticas rígidas até a sofisticação semântica dos modelos de codificação. Contudo, o surgimento recente dos Grandes Modelos de Linguagem (LLMs), como a família GPT, instaura um novo paradigma que desafia a hegemonia de modelos discriminativos, como o BERT. Enquanto o padrão-ouro estabelecido por Chyr et al. (2020) foca na representação latente para a classificação, os LLMs introduzem capacidades de raciocínio few-shot e de compreensão contextual ampliada. Este Trabalho de Conclusão de Curso propõe-se, portanto, a preencher a lacuna comparativa atual, avaliando se o desempenho superior reportado nas arquiteturas de Deep Learning clássicas ainda se sustenta diante das capacidades emergentes das abordagens baseadas em LLMs no contexto educacional.

3 Metodologia

Este capítulo descreve o desenho experimental, as ferramentas computacionais e os critérios analíticos adotados para o estudo comparativo. A pesquisa caracteriza-se como experimental e de natureza mista (quantitativa-qualitativa), visando confrontar a eficácia de um modelo especialista com ajuste fino no domínio educacional (Pred-inter) contra soluções de propósito geral consolidadas na literatura.

Para uma análise abrangente, as ferramentas de comparação foram estratificadas em dois paradigmas tecnológicos distintos:

1. **Abordagens Léxicas:** Baseadas em dicionários e regras (VADER, TextBlob);
2. **Modelos Discriminativos:** Baseados em Transformers pré-treinados (BERT/Hugging Face);

Desta forma, busca-se verificar se a especialização do treinamento (domain adaptation) supera a robustez e a capacidade de generalização dos grandes modelos modernos."

3.1 Fonte de Dados e Pré-processamento

O estudo fundamenta-se no dataset Stanford MOOCPosts, que reúne mensagens publicadas em fóruns de discussão de cursos massivos abertos online. O dataset é composto por quatro tabelas, cada uma associada a um tipo específico de anotação manual. Duas dessas tabelas são voltadas à análise de sentimento, sendo que uma utiliza três classes e a outra utiliza cinco classes, diferenciando muito negativo, negativo, neutro, positivo e muito positivo. Neste trabalho, foi utilizada exclusivamente a tabela de sentimento com cinco classes, por permitir uma análise mais fina da polaridade expressa nas mensagens. As outras duas tabelas correspondem às anotações de confusão e de urgência, ambas baseadas em três classes, e têm como objetivo identificar, respectivamente, o nível de confusão do autor e a urgência percebida na mensagem.

Para isolar o objeto deste estudo, foi desenvolvido um algoritmo de pré-processamento para extração seletiva de atributos. O procedimento consistiu na leitura do arquivo original e na geração de um novo artefato estruturado (*clean dataset*), mantendo-se apenas duas colunas fundamentais:

- **Text:** O conteúdo textual da postagem (variável independente/entrada).
- **Sentiment(1-7):** A nota atribuída manualmente por avaliadores humanos, em uma escala contínua de 1 a 7 (variável dependente/alvo).

A coluna `Sentiment(1-7)` atua, portanto, como o **Ground Truth** (Padrão-Ouro) da pesquisa. É a referência de correção à qual as ferramentas automáticas devem se aproximar para serem consideradas precisas.

A Tabela 3.1 ilustra a transformação dos dados realizada para a construção do arquivo de análise.

Tabela 3.1: Fluxo de Processamento do Dataset

Dataset Original (Bruto)	→	**Dataset de Análise (Processado)**
Colunas: Unnamed:0, Text, Opinion(1/0), Question(1/0), Answer(1/0), **Sentiment(1-7)**, Confusion(1-7), Urgency(1-7), CourseType, forum_post_id...	→	**Col. 1:** Text (Entrada) **Col. 2:** Sentiment(1-7) (Gabarito) **Col. 3+:** [Resultados das Ferramentas]

Fonte: O autor (2025).

Este novo arquivo processado serve como insumo para o *pipeline* de inferência. À medida que o *script* de avaliação executa as ferramentas (VADER, TextBlob, Pred-inter, Huggingface), as notas preditas são concatenadas em novas colunas, permitindo uma comparação linha a linha entre as notas geradas pela máquina e o *Ground Truth* humano na etapa de análise estatística.

O script de automação foi desenvolvido para garantir a reprodutibilidade, iterando sobre o dataset e armazenando os resultados de cada ferramenta de forma estruturada.

3.2 Configuração das Ferramentas de Análise

O experimento foi concebido para confrontar quatro paradigmas distintos de inferência. Todas as ferramentas foram configuradas para produzir, ao final do processamento, um valor numérico contínuo normalizado no intervalo $[-1, 1]$, permitindo a comparação direta de intensidade.

3.2.1 Abordagens Léxicas e Estatísticas

Estas ferramentas operam nativamente ou foram adaptadas para a escala $[-1, 1]$:

- **VADER:** Executado via biblioteca `vaderSentiment`. Utilizou-se o método `polarity_scores` para extrair a métrica `compound`, que fornece a polaridade normalizada por heurísticas gramaticais;
- **TextBlob:** Executado via propriedade `sentiment.polarity`, retornando valores flutuantes derivados da média dos léxicos;
- **Hugging Face (Transformers):** Utilizou-se o modelo `distilbert-base-uncased-finetuned-sst-2-english`. Como este modelo retorna originalmente um escore de confiança $[0, 1]$ associado a um rótulo categórico ("POSITIVE" ou "NEGATIVE"), aplicou-se um pós-processamento de inversão de sinal aos rótulos negativos, mapeando o resultado final para a escala contínua $[-1, 1]$.

3.2.2 Modelo Especialista (Pred-inter)

O Pred-inter, baseado na arquitetura BERT e treinado com dados educacionais, classifica as sentenças em 5 classes discretas ($c \in \{0, 1, 2, 3, 4\}$). Para viabilizar a comparação com as demais ferramentas de escala contínua, implementou-se o método do **Valor Esperado Normalizado**:

1. Extração da distribuição de probabilidade (P) das classes via função *Softmax*;
2. Cálculo da média ponderada (Valor Esperado): $Score_{raw} = \sum_{i=0}^{4}(P_i \times i)$;
3. Transformação linear para o intervalo-alvo:

$$Score_{norm} = \frac{Score_{raw} - 2}{2} \tag{3.1}$$

Desta forma, uma predição de classe 0 (Muito Negativo) resulta em -1.0 e de classe 4 (Muito Positivo) em $+1.0$.

3.3 Critérios de Categorização e Alinhamento de Escalas

Um desafio metodológico central deste estudo é a heterogeneidade entre a escala de anotação humana (*Ground Truth*) e a escala das ferramentas automáticas. Para o cálculo da

acurácia, adotou-se um protocolo de discretização em 5 classes.

Para a construção do gabarito de referência, este trabalho adota, estritamente, o protocolo de discretização estabelecido por (ROSSI, 2022), previamente detalhado na Fundamentação Teórica (Seção 2.6.1).

A aplicação deste critério sobre as notas originais manuais ($N_{humano} \in [1.0, 7.0]$) resulta na seguinte segmentação de classes, utilizada aqui para o cálculo da Acurácia:

- **Classe 0 (Muito Negativo):** $1.0 \leq N \leq 1.5$
- **Classe 1 (Negativo):** $2.0 \leq N \leq 3.0$
- **Classe 2 (Neutro):** $3.5 \leq N \leq 5.0$
- **Classe 3 (Positivo):** $5.5 \leq N \leq 6.0$
- **Classe 4 (Muito Positivo):** $6.5 \leq N \leq 7.0$

Para as ferramentas automáticas ($S_{tool} \in [-1.0, 1.0]$), adotou-se uma discretização em intervalos equidistantes de amplitude 0,4. Esta abordagem assume uma distribuição linear da intensidade do sentimento na ausência de regras de domínio específicas:

- **Classe 0:** $S < -0.60$
- **Classe 1:** $-0.60 \leq S < -0.20$
- **Classe 2:** $-0.20 \leq S < +0.20$
- **Classe 3:** $+0.20 \leq S < +0.60$
- **Classe 4:** $S \geq +0.60$

Esta metodologia garante que, embora as escalas de origem sejam distintas, a comparação final ocorra sobre categorias de intensidade semanticamente equivalentes.

3.4 Procedimentos de Análise dos Resultados

A avaliação da eficácia das ferramentas foi conduzida em duas etapas complementares: uma análise estatística do desempenho (Quantitativa) e uma investigação semântica dos erros (Qualitativa).

3.4.1 Análise Quantitativa e Métricas

Para a avaliação numérica, as saídas contínuas de todas as ferramentas foram convertidas em 5 classes discretas, seguindo os critérios de categorização definidos na seção anterior. Com os dados padronizados, aplicam-se as seguintes métricas:

1. **Acurácia Global (Accuracy):** Representa o percentual de correspondência exata entre a classe predita e a classe real. É a métrica primária para ranqueamento das ferramentas.

2. **Erro Médio Absoluto (MAE):** Calculado com base em valores contínuos para avaliar a precisão da intensidade. Como o *Ground Truth* original (y_{orig}) utiliza a escala $[1, 7]$ e as ferramentas utilizam a escala $[-1, 1]$, foi aplicada uma transformação linear na nota humana para viabilizar a subtração direta:

$$y_{\text{norm}} = \frac{y_{\text{orig}} - 4}{3} \tag{3.2}$$

   Desta forma, uma nota humana 1.0 torna-se -1.0, uma nota 4.0 torna-se 0.0 e uma nota 7.0 torna-se 1.0. O MAE é então calculado entre este y_{norm} e a predição da ferramenta ($\hat{y}$):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{norm,i} - \hat{y}_i| \tag{3.3}$$

3. **Matriz de Confusão (5×5):** A matriz de confusão é uma tabela cruzada na qual as linhas correspondem às classes definidas como *Ground Truth* e as colunas, às classes preditas pelas ferramentas. No contexto de uma escala ordinal de intensidade de sentimento (0 a 4), essa visualização permite observar como as predições se distribuem em relação aos valores reais, evidenciando as classes mais frequentemente atribuídas por cada ferramenta, bem como os padrões de acerto e de erro ao longo da escala. A partir dessa representação, a matriz auxilia na compreensão do comportamento geral das ferramentas ao longo da escala de sentimento.

3.4.2 Análise Qualitativa

Esta etapa tem como objetivo compreender, de forma qualitativa, as causas das classificações incorretas observadas nos experimentos, investigando por que diferentes ferramentas

divergem na interpretação de uma mesma sentença. Diferentemente das análises quantitativas apresentadas anteriormente, esta seção adota uma abordagem exploratória e interpretativa, baseada na inspeção manual dos erros.

A análise foi conduzida em dois eixos principais. O primeiro consistiu na identificação de erros compartilhados entre as ferramentas, buscando padrões de falha recorrentes. O segundo eixo envolveu a categorização manual das frases incorretamente classificadas, com o objetivo de associar cada erro a um possível fator linguístico ou contextual que tenha influenciado a decisão do modelo.

No que se refere à intersecção de erros, foi realizada uma análise simples dos conjuntos de frases classificadas incorretamente, permitindo a identificação de dois cenários de interesse. O primeiro corresponde aos casos de erro universal, nos quais todas as ferramentas avaliadas falharam simultaneamente. O segundo cenário refere-se à eficácia do modelo especialista, representado pelas frases em que apenas o Pred-inter apresentou classificação correta, sugerindo benefícios decorrentes de sua adaptação ao domínio educacional.

Paralelamente, foi realizada uma categorização manual das frases classificadas incorretamente. Diferentemente de uma simples contagem de erros, essa etapa buscou identificar o motivo provável de cada falha, associando a frase a uma categoria explicativa. Esse processo resultou no conjunto de categorias apresentado na Seção de Categorização de Amostras, como erros de intensidade, desconsideração de contexto, viés semântico em domínio especializado, classificação indevida de textos não emotivos e superestimação de positividade em contextos acadêmicos ou profissionais.

4 Resultados

O experimento foi conduzido em um conjunto de 29.550 postagens educacionais extraídas de fóruns de cursos massivos (MOOCs), conforme descrito no Capítulo 3. Cada instância contém:

- Texto original da postagem;
- Nota de sentimento atribuída manualmente por avaliadores humanos (escala 1–7);
- Classes discretizadas de sentimento (0 a 4), consideradas como Ground Truth;
- Predições contínuas e discretizadas produzidas pelas quatro ferramentas analisadas: VADER, TextBlob, Hugging Face, Pred-inter.

Essa estrutura permitiu uma comparação linha a linha, garantindo consistência estatística e reprodutibilidade dos resultados.

Na Tabela 4.1, a coluna *Sentiment (1–7)* representa a nota atribuída manualmente por avaliadores humanos, enquanto *GT* indica a respectiva classe discretizada (*Ground Truth*). As colunas VADER, TextBlob e HF apresentam escores contínuos normalizados no intervalo [-1, 1], e as colunas marcadas com "C" indicam as classes discretizadas correspondentes utilizadas no cálculo das métricas quantitativas. As predições do modelo Pred-inter não são exibidas nesta tabela por questões de legibilidade e formatação; contudo, seus escores contínuos e classes discretizadas foram armazenados no conjunto de dados final, seguindo o mesmo procedimento adotado para as demais ferramentas, e foram integralmente considerados nas análises apresentadas nas seções subsequentes.

Tabela 4.1: Exemplo do arquivo final de saída utilizado na análise dos resultados

Text	**Sent. (1–7)**	**GT**	**VADER**	**VADER C**	**TextBlob**	**TB C**	**HF**	**HF C**
This lecture was very clear and helpful...	6.0	3	0.624	3	0.500	3	0.981	4
I am completely lost in this topic...	2.0	1	-0.636	1	-0.400	1	0.912	4
The assignment is challenging but fair...	5.5	3	0.318	3	0.250	3	0.874	4 5
I did not understand the explanation...	2.5	1	-0.421	1	-0.300	1	0.901	4
Great course, learned a lot so far...	6.5	4	0.812	4	0.700	4	0.993	4

Fonte: O autor (2026).

4.1 Análise Quantitativa de Desempenho

4.1.1 Acurácia Global

A Tabela 4.2 apresenta a acurácia global das ferramentas, considerando a correspondência exata entre a classe predita e o *Ground Truth* humano (escala de 5 classes).

Tabela 4.2: Acurácia global das ferramentas de análise de sentimentos

Ferramenta	**Acurácia (%)**
VADER	23.11%
TextBlob	59.79%
Hugging Face	1,51%
Pred-inter	**88.07%**

Fonte: O autor (2026).

Os resultados evidenciam que o Pred-inter apresentou desempenho amplamente superior, confirmando a hipótese H1 de que modelos especializados no domínio educacional são mais eficazes. Já o TextBlob apresentou desempenho intermediário, beneficiando-se da simplicidade do léxico, mas ainda assim limitado pela ausência de contexto profundo. Com relação ao VADER, este apresentou baixa acurácia, sugerindo que suas heurísticas, embora eficazes em redes sociais, não capturam adequadamente o vocabulário pedagógico. Por fim, o modelo Hugging Face, apesar de ser o estado da arte em contextos gerais, apresentou desempenho extremamente baixo, indicando uma forte incompatibilidade entre o domínio de treinamento (reviews) e o domínio educacional.

4.1.2 Erro Médio Absoluto

Além da classificação discreta, avaliou-se a precisão na intensidade do sentimento, por meio do Erro Médio Absoluto, calculado sobre valores contínuos normalizados no intervalo [1, 1].

Tabela 4.3: Erro Médio Absoluto das ferramentas de análise de sentimentos

Ferramenta	**MAE**
VADER	0,4481
TextBlob	0,1925
Hugging Face (DistilBERT)	0,8651
Pred-inter	**0,1071**

Fonte: O autor (2026).

Observa-se que o modelo Pred-inter não apenas apresenta o maior número de acertos,

como também estima de forma mais adequada a intensidade emocional das postagens, o que reflete o menor erro médio absoluto entre as ferramentas avaliadas. O TextBlob, apesar de suas limitações, apresentou desempenho razoável na estimativa da intensidade média do sentimento. Em contraste, o modelo Hugging Face apresentou o maior valor de MAE, reforçando a hipótese de um desalinhamento semântico entre o corpus educacional analisado e o modelo pré-treinado utilizado.

4.2 Matrizes de Confusão e Tipologia dos Erros

Nesta seção, são apresentadas as matrizes de confusão das ferramentas analisadas, bem como uma discussão qualitativa sobre os tipos de erro observados, distinguindo erros leves e erros críticos conforme a distância entre a predição e o *Ground Truth*.

4.2.1 Matriz de Confusão do Modelo Pred-Inter

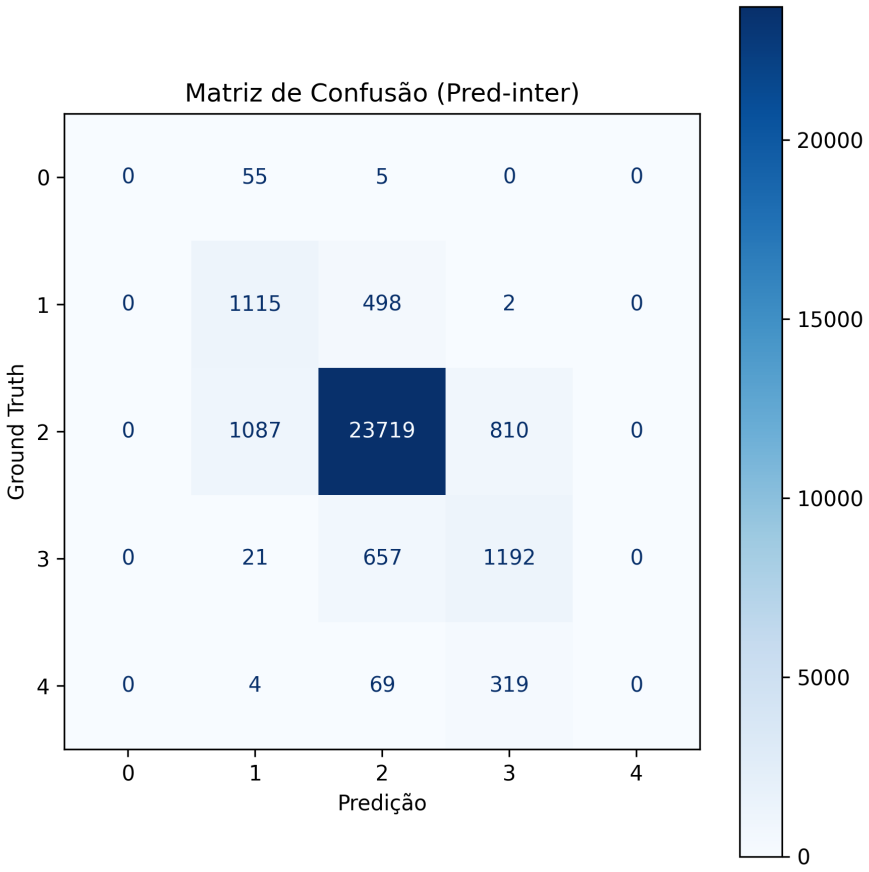

Figura 4.1: Matriz de confusão do modelo Pred-inter

A matriz de confusão do modelo Pred-inter (Figura 4.1) evidencia um padrão de acerto

fortemente concentrado na classe intermediária de sentimento (classe 2), na qual foram corretamente classificadas 23.719 instâncias, o que corresponde à ampla maioria dos acertos do modelo. Observa-se ainda um número significativamente menor de acertos nas classes adjacentes, com 1.115 instâncias corretamente classificadas como classe 1 e 1.192 como classe 3. Não foram observados acertos nas classes extremas (0 e 4), indicando que o modelo tende a evitar classificações de polaridade muito intensa, mesmo quando essas categorias estão presentes no *Ground Truth*.

Em relação aos erros, a maior concentração ocorreu na predição da classe 1, totalizando 1.087 instâncias incorretas, o que sugere uma tendência do modelo a subestimar a intensidade negativa em casos limítrofes. Esse comportamento resulta predominantemente em erros leves, caracterizados por deslocamentos de uma classe em relação ao valor real, enquanto erros críticos — em que a predição diverge em mais de duas classes — são pouco frequentes. Tal padrão indica que, embora conservador na atribuição de extremos, o Pred-inter mantém boa aderência ordinal à avaliação humana, errando majoritariamente quanto à intensidade, e não à polaridade, do sentimento.

4.2.2 Matriz de Confusão do Modelo VADER

A Figura 4.2 apresenta a matriz de confusão da ferramenta VADER, revelando um comportamento significativamente distinto do observado no modelo especialista. Embora o maior número de acertos concentre-se na classe intermediária de sentimento (classe 2), com 5.861 instâncias corretamente classificadas, observa-se uma elevada dispersão de erros nessa mesma classe. Em particular, frases rotuladas como classe 2 foram frequentemente classificadas incorretamente, sendo 1.583 preditas como classe 0, 2.613 como classe 1, 6.342 como classe 3 e 9.217 como classe 4.

Esse padrão indica uma incapacidade do VADER em estimar adequadamente a intensidade do sentimento em contextos educacionais, resultando não apenas em erros leves, mas também em uma quantidade expressiva de erros críticos, em que a predição se desloca em duas ou mais classes em relação ao valor real. A forte tendência à superestimação da intensidade positiva, evidenciada pelo grande número de predições na classe 4 para instâncias neutras ou moderadas, sugere que heurísticas léxicas eficazes em domínios como redes sociais não se transferem adequadamente para o discurso acadêmico, no qual termos como “desafiador” ou “complexo” podem assumir conotação positiva. Como consequência, o VADER apresenta baixa aderência ordinal ao *Ground Truth*, comprometendo sua confiabilidade nas análises de sentimento em ambientes

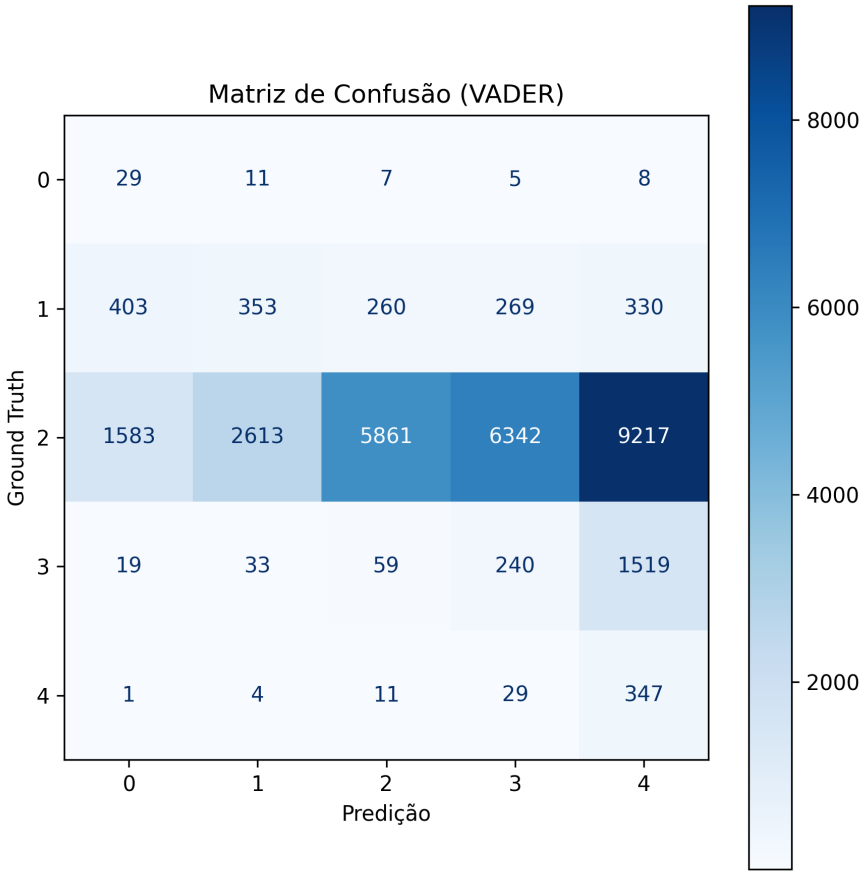

Figura 4.2: Matriz de confusão do modelo VADER

educacionais.

4.2.3 Matriz de Confusão do TextBlob

A matriz de confusão da Figura 4.3 mostra que a ferramenta TextBlob apresenta um comportamento mais consistente do que o VADER, porém ainda aquém do observado no modelo especialista. O maior número de acertos concentra-se na classe intermediária de sentimentos (classe 2), com 16.254 instâncias corretamente classificadas, o que indica maior estabilidade na identificação de sentimentos moderados.

No entanto, observa-se uma tendência recorrente à superestimação da intensidade positiva, evidenciada pelo elevado número de erros em que instâncias rotuladas no *Ground Truth* como classe 2 foram classificadas como classe 3, totalizando 7.568 ocorrências. Esse padrão sugere que o TextBlob, embora mais equilibrado, apresenta dificuldade para distinguir nuances sutis entre sentimentos neutros e levemente positivos em contextos educacionais.

A classe 3 representa ainda a segunda classe com maior número de acertos, totalizando 1.180 instâncias corretamente classificadas, o que reforça a inclinação do modelo a deslocar as

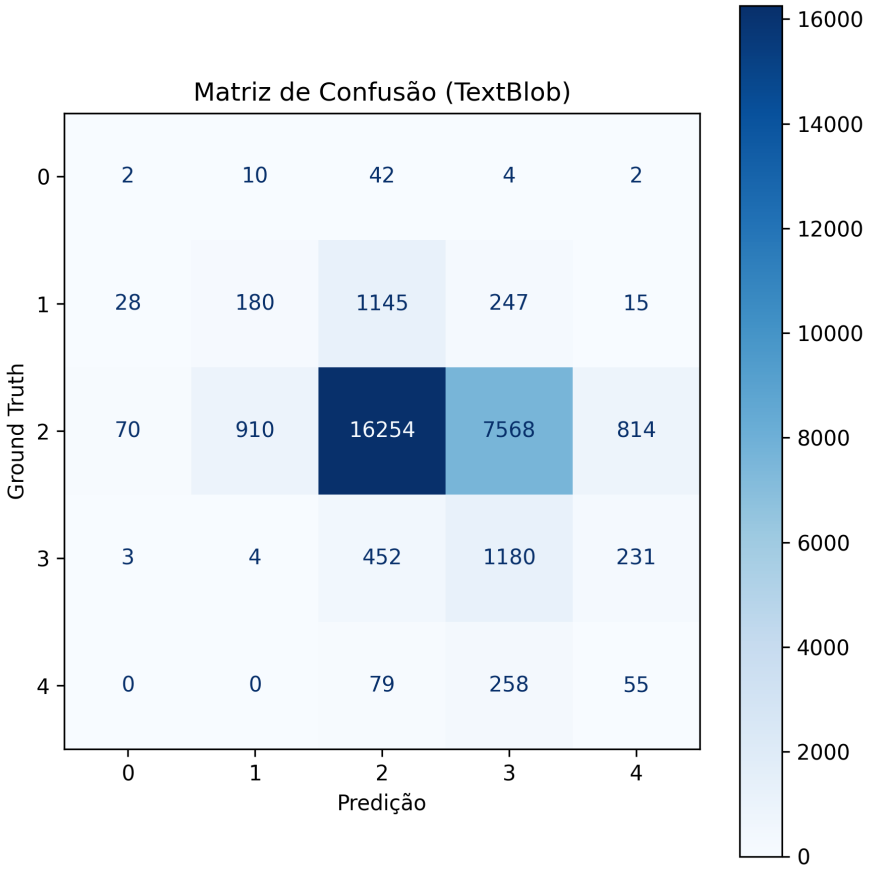

Figura 4.3: Matriz de confusão do modelo TextBlob

predições para níveis superiores de positividade. Em termos de tipologia de erro, predominam erros leves, caracterizados por deslocamentos de uma única classe, com baixa incidência de erros críticos. Ainda assim, essa tendência à superestimação compromete a fidelidade da estimativa de intensidade emocional, limitando a aplicabilidade do TextBlob em análises educacionais que demandam maior sensibilidade semântica.

4.2.4 Matriz de Confusão do Hugging Face

A matriz de confusão do modelo Hugging Face (Figura 4.4) revela um comportamento altamente instável no contexto educacional analisado, caracterizado por um baixo índice geral de acertos e por erros de grande magnitude. O maior número de acertos foi observado na classe 4, com apenas 375 instâncias corretamente classificadas, o que evidencia baixa aderência global ao *Ground Truth*.

Em contraste, os erros concentram-se majoritariamente em deslocamentos extremos, especialmente em instâncias rotuladas como classe 2, que foram frequentemente classificadas como classe 0 (14.616 ocorrências) ou classe 4 (10.543 ocorrências). Padrões semelhantes são

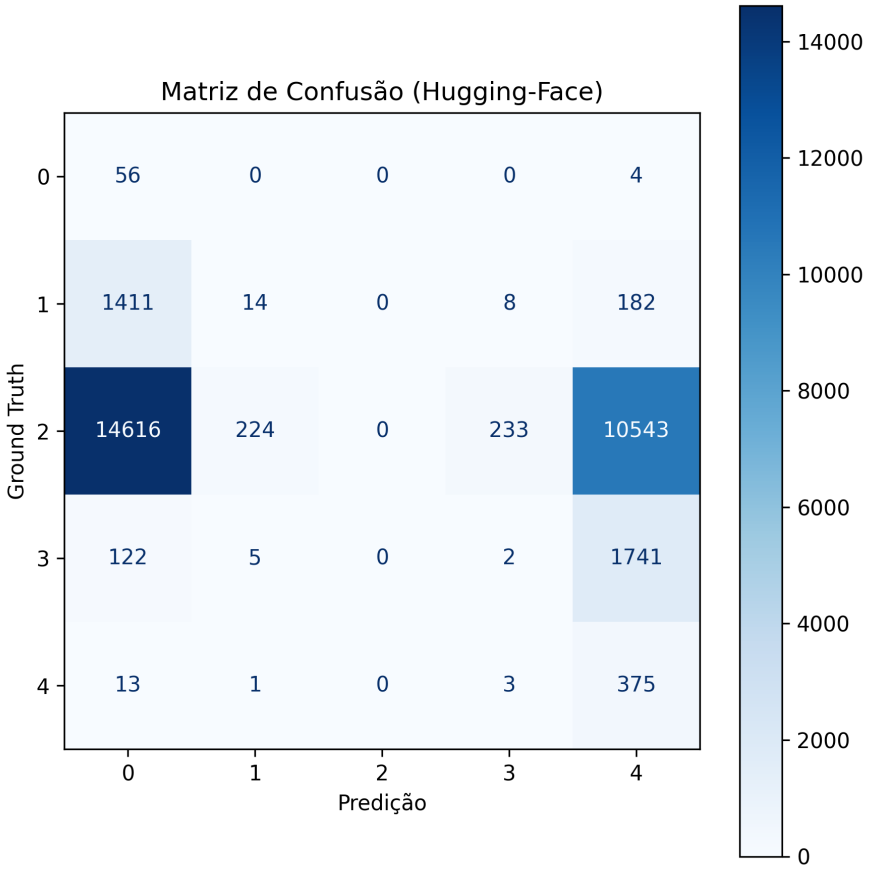

Figura 4.4: Matriz de confusão do modelo Hugging Face

observados para a classe 3, com 1.741 instâncias preditas incorretamente como classe 4, e para a classe 1, com 1.411 instâncias classificadas incorretamente como classe 0.

Esse comportamento indica uma forte polarização das predições, com tendência a colapsar para os extremos da escala, produzindo predominantemente erros críticos, em que a diferença entre a predição e o valor real é igual ou superior a duas classes. Tal padrão sugere que o modelo, embora baseado em arquiteturas modernas de Deep Learning, não foi capaz de capturar adequadamente as nuances semânticas do discurso educacional, reforçando a hipótese de que modelos generalistas, quando não adaptados ao domínio de aplicação, podem apresentar desempenho inferior ao de abordagens mais simples.

4.3 Análise Qualitativa das Divergências

Nesta seção, realiza-se uma análise qualitativa das divergências entre as ferramentas de análise de sentimentos, com o intuito de compreender por que algumas sentenças geraram interpretações distintas. Para isso, são examinados exemplos representativos, considerando aspectos

linguísticos e contextuais, como a forma do texto, o vocabulário empregado e a intensidade do sentimento expresso.

4.3.1 Erro Universal

A análise qualitativa inicia-se pela investigação de um caso extremo, no qual todas as ferramentas avaliadas produziram erros críticos, com diferença de duas ou mais classes em relação ao *Ground Truth*. A sentença apresentada a seguir exemplifica uma combinação rara de fatores linguísticos e estruturais que dificultam a interpretação automática do sentimento, mesmo para modelos baseados em Deep Learning.

> This is an incredible course, and Jo is a revolutionary. This online class should be offered in every school as an inservice course for teachers._x0007__x0007_My only problem is that I happened to listen to Harry Shearer's *Le Show* a week ago and he had a piece on the gratuitous use of the word *"so"*._x0007__x0007_Alas, I am now having trouble getting past Jo's frequent "so"s._x0007__x0007_So, I will just try to ignore them. :¿)

A sentença apresentada foi rotulada manualmente por avaliadores humanos como *Muito Positiva* (classe 4), o que reflete uma avaliação global amplamente favorável do curso e da instrutora, apesar da presença de uma crítica pontual. No entanto, observou-se uma divergência acentuada entre as ferramentas automáticas: o modelo Pred-inter classificou a frase como classe 1, enquanto VADER e Hugging Face atribuíram a classe 0, indicando um sentimento fortemente negativo. O TextBlob, por sua vez, apresentou uma classificação intermediária (classe 2), mais próxima da avaliação humana, embora ainda distante da intensidade originalmente atribuída. Essa discrepância evidencia a dificuldade das ferramentas em hierarquizar informações discursivas, distinguindo críticas secundárias de avaliações globais, especialmente em textos longos, com sentimento misto e presença de ruídos de codificação.

4.3.2 Eficácia do Especialista

A análise dos casos exclusivos do modelo Pred-inter permite avaliar, de forma isolada, suas limitações e vantagens em relação às demais ferramentas, conforme apresentado na Tabela 4.4. Observam-se dois cenários distintos: instâncias em que apenas o Pred-inter produziu uma classificação incorreta e outras em que apenas esse modelo obteve a classificação correta.

Os erros exclusivos do Pred-inter correspondem a uma fração residual do conjunto de dados e concentram-se predominantemente na subestimação da intensidade do sentimento,

sem inversão de polaridade nem ocorrência recorrente de erros extremos. Em contraste, os acertos exclusivos representam um volume expressivo, indicando maior capacidade do modelo para identificar avaliações neutras ou moderadamente positivas em contextos educacionais. Esse padrão reforça a hipótese de que a adaptação ao domínio permite ao Pred-inter lidar melhor com as nuances do discurso educacional, apresentando maior proximidade com as avaliações humanas, ainda que tenda a evitar a atribuição de valores extremos de sentimento.

Tabela 4.4: Casos exclusivos do modelo Pred-inter

Situação	**Classe Predita**	**Classe Correta**	**Quantidade**
Erros exclusivos do Pred-inter			
Pred-inter errou sozinho	3	4	42
Pred-inter errou sozinho	2	4	6
Pred-inter errou sozinho	1	0	1
Acertos exclusivos do Pred-inter			
Pred-inter acertou sozinho	2	2	7.231
Pred-inter acertou sozinho	Outras classes	Correspondentes	1.104
Total de erros exclusivos	–	–	**52**
Total de acertos exclusivos	–	–	**8.335**

Fonte: O autor (2026).

4.3.3 Categorização de Amostras

Com o objetivo de compreender, de forma qualitativa, os tipos de erros cometidos pelas ferramentas avaliadas, foi realizada uma análise manual de amostras classificadas incorretamente. Para cada ferramenta, foram selecionados aleatoriamente cinquenta casos de erro, que foram posteriormente analisados e agrupados em categorias conceituais. Essa categorização não tem como objetivo estabelecer uma taxonomia definitiva de erros na análise de sentimentos, mas sim oferecer uma interpretação organizada dos padrões observados ao longo dos experimentos.

É importante destacar o caráter inerentemente interpretativo dessa etapa. A classificação dos erros envolve julgamento humano e, portanto, reflete tanto a percepção do autor deste trabalho quanto as decisões dos anotadores originais do dataset. Dessa forma, os resultados apresentados nesta seção não devem ser interpretados como verdades absolutas, mas sim como uma leitura possível dos dados, construída a partir da experiência empírica e das limitações inerentes ao processo de anotação manual de sentimentos.

As amostras analisadas foram organizadas em cinco categorias principais, descritas a seguir.

Erro de intensidade

Esta categoria engloba casos em que a ferramenta identificou corretamente a polaridade geral do sentimento, porém errou na intensidade atribuída. Exemplos típicos incluem classificações entre positivo e muito positivo ou entre negativo e muito negativo, sem que haja confusão entre classes de polaridade distintas, como negativo, neutro ou positivo.

Exemplo de frase representativa desta categoria: *"Awesome idea! Do all the groups improve?"*

A frase acima foi classificada como muito positiva pela ferramenta e como positiva por humanos, indicando subjetividade na avaliação.

Ignorou contexto em detrimento de sinais lexicais locais

Nesta categoria estão os erros em que o modelo atribuiu o sentimento com base em palavras ou expressões isoladas, desconsiderando o contexto global da frase. Nesses casos, termos com carga emocional aparente acabam por sobrepor o significado discursivo completo, levando a classificações equivocadas.

Exemplo de frase representativa desta categoria: *"The same problem without solution for three days."*

Na frase acima, fica evidente a frustração do autor. O gabarito para esta frase seria 1, negativo, mas a ferramenta avaliou como neutro, talvez pela falta de elementos lexicos suficientes para forçar a nota para baixo. A palavra "problem", sendo a única com carga negativa, não foi suficiente para que o modelo percebesse que o sentimento era negativo.

Erro por viés semântico em domínio especializado

Esta categoria reúne erros associados ao uso de linguagem técnica ou especializada, comum em contextos acadêmicos e educacionais. Observou-se que, em diversas situações, os modelos apresentaram dificuldade em interpretar corretamente termos com significado específico no domínio, resultando em classificações de sentimento inadequadas.

Exemplo de frase representativa desta categoria: *"Asystole and PEA are non shockable.....the best we can do is to continue CPR whenever you get one. The medications and stuff can be considered after."*

Um exemplo representativo de erro por viés semântico em domínio especializado pode ser observado na frase acima. O texto faz uso intensivo de jargão médico e terminologia técnica

associada a protocolos clínicos de emergência. Embora trate de situações críticas, a frase é predominantemente descritiva e instrucional, não expressando avaliação emocional por parte do autor. Modelos de análise de sentimentos que não incorporam conhecimento de domínio tendem a interpretar termos como *asystole*, *non shockable* e *CPR* como indicadores de polaridade negativa, ou a classificá-los como neutros de forma acrítica, demonstrando dificuldade em diferenciar linguagem técnica de manifestação subjetiva de sentimento.

Erro por classificação de texto não emotivo

Incluem-se nesta categoria os casos em que o texto analisado apresenta caráter predominantemente informativo ou descritivo, sem manifestação clara de emoção, mas, ainda assim, é classificado pela ferramenta como positivo ou negativo. Esse tipo de erro foi recorrente em mensagens neutras, especialmente em solicitações, relatos factuais ou descrições de procedimentos.

Exemplo de frase representativa desta categoria: *"Ah, sorry, misunderstood the system! Of course, for the median no rounding is needed."*

Para a frase acima, determinada ferramenta classificou-a como muito negativa, demonstrando um erro grosseiro em uma frase neutra.

Erro de intensidade positiva no contexto acadêmico ou profissional

Esta categoria refere-se a um padrão específico observado, sobretudo, em textos relacionados a ambientes acadêmicos ou profissionais. Verificou-se uma tendência de superestimação da positividade em frases que empregam linguagem cordial, avaliativa ou protocolar, mesmo quando o sentimento expresso é neutro ou apenas levemente positivo.

Exemplo de frase representativa desta categoria: *"and remember, the purpose of the course isn't to produce a document for actual publication but to improve your writing skills!"*

A frase acima foi superestimada assim como outras por estar atrelada ao contexto pedagógico/acadêmico, ela foi rotulada como neutra por humanos e recebeu muito positivo por determinada ferramenta. Este padrão se repetiu em uma proporção significativa nas amostras coletadas, justificando a criação desta categoria.

4.3.4 Distribuição dos erros por categoria

Após a categorização das amostras, os cinquenta erros selecionados para cada ferramenta foram quantificados conforme as categorias definidas. Para facilitar a comparação entre os métodos avaliados, os resultados são apresentados em tabelas que contêm tanto a frequência absoluta quanto a proporção percentual de cada categoria de erro.

Nesta análise, as categorias de erro são interpretadas da seguinte forma:

- CATEGORIA 1: Erro de intensidade;
- CATEGORIA 2: Ignorou contexto em detrimento de sinais lexicais locais;
- CATEGORIA 3: Erro por viés semântico em domínio especializado;
- CATEGORIA 4: Erro por classificação de texto não emotivo;
- CATEGORIA 5: Erro de intensidade positiva no contexto acadêmico ou profissional.

VADER

Tabela 4.5: Distribuição das categorias de erro para o VADER

Categoria	Quantidade	Percentual (%)
CATEGORIA 1	21	42
CATEGORIA 2	21	42
CATEGORIA 3	3	6
CATEGORIA 4	1	2
CATEGORIA 5	4	8
Total	50	100

Text Blob

Tabela 4.6: Distribuição das categorias de erro para o Text Blob

Categoria	Quantidade	Percentual (%)
CATEGORIA 1	22	44
CATEGORIA 2	21	42
CATEGORIA 3	1	2
CATEGORIA 4	4	8
CATEGORIA 5	2	4
Total	50	100

Hugging-Face

Tabela 4.7: Distribuição das categorias de erro para o Hugging-Face

Categoria	Quantidade	Percentual (%)
CATEGORIA 1	17	34
CATEGORIA 2	9	18
CATEGORIA 3	3	6
CATEGORIA 4	15	30
CATEGORIA 5	6	12
Total	50	100

Pred-Inter

Tabela 4.8: Distribuição das categorias de erro para o Pred-Inter

Categoria	Quantidade	Percentual (%)
CATEGORIA 1	17	34
CATEGORIA 2	13	26
CATEGORIA 3	14	28
CATEGORIA 4	6	12
Total	50	100

4.4 A Viabilidade de LLMs

Com o objetivo de avaliar a viabilidade prática do uso de modelos de linguagem de grande porte na análise de sentimentos, foram realizadas tentativas de processamento do dataset por meio de três sistemas distintos: ChatGPT, Gemini e DeepSeek, este último acessado via API local utilizando o framework Ollama. Em todos os casos, buscou-se aplicar o mesmo comando, solicitando a classificação das frases da primeira coluna do arquivo em uma escala discreta de cinco níveis, variando de muito negativo a muito positivo.

No caso do ChatGPT, a principal limitação observada foi a inviabilidade de enviar o dataset completo para análise. O arquivo utilizado neste trabalho possui aproximadamente 16 MB e contém um grande volume de textos, o que excede as restrições práticas de entrada impostas pelo modelo em ambientes interativos.

Situação semelhante foi observada ao tentar utilizar o modelo Gemini. Assim como no ChatGPT, o sistema não permitiu o processamento direto de um arquivo desse porte. Além disso, durante o processo de raciocínio apresentado pelo próprio modelo, foi indicado o uso de ferramentas léxicas externas, como o TextBlob, para a realização da análise de sentimentos. Essa abordagem torna o uso do modelo conceitualmente redundante neste trabalho, uma vez que o

TextBlob já integra explicitamente o conjunto de ferramentas avaliadas. Dessa forma, o emprego do Gemini não acrescentaria um método independente de análise, mas apenas reutilizaria uma abordagem já contemplada no estudo.

Em relação ao modelo DeepSeek, optou-se por sua execução local, por meio da API do Ollama, a fim de contornar as limitações impostas pelas plataformas proprietárias no envio de dados. No entanto, essa alternativa apresentou restrições significativas de ordem computacional. Devido às limitações de hardware disponíveis, o tempo estimado para a conclusão da análise completa do dataset estava na ordem de dias de processamento contínuo. Além disso, os resultados parciais obtidos até o momento da interrupção do experimento indicaram um comportamento inconsistente do modelo, que atribuía exclusivamente o valor de 0.0 às frases analisadas, independentemente de seu conteúdo semântico. Tal comportamento sugere fragilidade no processo de avaliação quando submetido a tarefas extensas e repetitivas em ambiente local.

Diante desses resultados, concluiu-se que o uso de modelos de linguagem de grande porte mostrou-se inviável para os objetivos deste trabalho, ao menos no contexto de estudos acadêmicos com datasets extensos e recursos computacionais limitados. A adoção efetiva desses modelos demandaria investimentos consideráveis em infraestrutura de hardware, além da contratação de planos premium em plataformas comerciais, o que compromete sua aplicabilidade prática quando comparados a abordagens mais leves, reproduzíveis e economicamente acessíveis.

5 Considerações Finais e Trabalhos Futuros

Este trabalho teve como objetivo realizar um estudo comparativo entre diferentes técnicas de Análise de Sentimentos aplicadas ao contexto educacional, confrontando abordagens léxicas consolidadas, modelos generalistas baseados em Transformers e um modelo especialista treinado especificamente para o domínio educacional. A motivação central residiu na necessidade de compreender quais ferramentas são mais adequadas para interpretar corretamente o feedback discente em ambientes virtuais de aprendizagem, em que a linguagem apresenta características próprias e dependência do contexto.

Os experimentos foram conduzidos sobre um conjunto expressivo de postagens extraídas de fóruns de cursos massivos (MOOCs), utilizando como referência um *Ground Truth* humano anotado manualmente. Para garantir a comparabilidade entre ferramentas heterogêneas, foi adotado um protocolo de normalização de escalas e de discretização em cinco classes de intensidade emocional, permitindo a avaliação tanto por métricas categóricas quanto por contínuas.

Os resultados obtidos evidenciam de forma consistente que a especialização no domínio educacional é um fator determinante para o desempenho em tarefas de análise de sentimentos nesse contexto. O modelo Pred-inter apresentou desempenho amplamente superior às demais ferramentas avaliadas, alcançando a maior acurácia global e o menor erro médio absoluto. Esses resultados indicam que o modelo não apenas classifica corretamente a polaridade do sentimento, mas também estima com maior precisão sua intensidade, aproximando-se da avaliação humana de forma mais fiel.

Em contraste, as abordagens léxicas, representadas pelo VADER e pelo TextBlob, apresentaram desempenho limitado. Embora o TextBlob tenha obtido resultados intermediários, sua incapacidade de lidar adequadamente com ambiguidades semânticas e com vocabulário pedagógico específico compromete sua eficácia em cenários educacionais mais complexos. O VADER, por sua vez, demonstrou baixa acurácia, sugerindo que heurísticas eficazes nas redes sociais não se transferem diretamente para o discurso acadêmico.

De forma particularmente relevante, os resultados do modelo Hugging Face baseado em DistilBERT evidenciaram que modelos generalistas de última geração não garantem desempenho satisfatório quando aplicados fora de seu domínio de treinamento original. Apesar de

sua robustez em tarefas de análise de sentimentos em reviews e redes sociais, o modelo apresentou elevado erro médio e baixa correspondência com o Ground Truth humano no contexto educacional, reforçando a importância da adaptação de domínio.

Os resultados obtidos permitem afirmar que a hipótese proposta foi confirmada apenas parcialmente. Conforme esperado, o modelo Pred-inter, por ser especializado no domínio educacional, apresentou desempenho superior às abordagens estritamente léxicas, como VADER e TextBlob, tanto em termos de acurácia quanto na estimativa da intensidade do sentimento. No entanto, diferentemente do que se supunha inicialmente, os modelos baseados em Transformers avaliados por meio da biblioteca Hugging Face não alcançaram desempenho equivalente ou superior, apresentando resultados inferiores, inclusive aos modelos léxicos. Esse comportamento sugere que o uso de arquiteturas avançadas e pré-treinamento massivo, por si só, não garante bom desempenho em tarefas de análise de sentimentos educacionais, reforçando a importância da adaptação ao domínio e da adequação do modelo ao tipo de linguagem analisada.

As principais contribuições desta pesquisa incluem a realização de um estudo comparativo sistemático entre ferramentas léxicas, modelos generalistas de Deep Learning e um modelo especialista no contexto educacional, bem como a proposição de um protocolo de alinhamento de escalas que possibilita a comparação entre saídas contínuas e discretas de diferentes algoritmos. Além disso, os resultados obtidos permitem a validação empírica da dependência de domínio na análise de sentimentos educacionais e demonstram a viabilidade do uso do Pred-inter como ferramenta de suporte pedagógico automatizado, indo além da mera classificação de polaridade.

Apesar dos resultados positivos, este trabalho apresenta algumas limitações. Primeiramente, o estudo concentrou-se em um único dataset, ainda que amplamente utilizado na literatura. A generalização dos resultados para outros contextos educacionais, idiomas ou níveis de ensino requer investigações adicionais. Além disso, os modelos baseados em Transformers utilizados não passaram por fine-tuning específico com dados educacionais, o que pode ter prejudicado seu desempenho.

Outra limitação diz respeito à natureza subjetiva da anotação humana, que, embora seja tratada como Ground Truth, pode conter ambiguidades inerentes à interpretação emocional de textos curtos.

Como trabalhos futuros, sugere-se a adaptação de modelos baseados em Transformers, como BERT ou RoBERTa, por meio de treinamento adicional com textos provenientes de contextos educacionais, como fóruns de cursos e ambientes virtuais de aprendizagem. Essa adaptação

permitiria avaliar se modelos originalmente treinados em textos genéricos passam a apresentar desempenho mais próximo ao de modelos especialistas. Também é indicada a expansão do estudo para outros idiomas, em especial o português, considerando o crescimento de plataformas educacionais nacionais e a escassez de estudos comparativos nesse contexto.

Conclui-se que a análise de sentimentos no contexto educacional exige abordagens que considerem as especificidades linguísticas e semânticas desse domínio. Modelos especializados, como o Pred-inter, demonstram ser mais adequados para essa tarefa, oferecendo não apenas maior precisão, mas também maior potencial de aplicação prática em sistemas de apoio à decisão pedagógica. Assim, este trabalho contribui para o avanço da área de Learning Analytics e reforça a importância da adaptação de domínio no desenvolvimento de soluções baseadas em Processamento de Linguagem Natural para a educação.

Referências

CHYR, C. et al. Classifying medical student forum posts using deep learning and bert. In: *Proceedings of the 13th International Conference on Educational Data Mining (EDM)*. [S.l.: s.n.], 2020.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). *NAACL-HLT (1)*. Association for Computational Linguistics, 2019. p. 4171–4186. ISBN 978-1-950737-13-0. Disponível em: ⟨http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19⟩.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 8, n. 1, p. 216–225, May 2014. Disponível em: ⟨https://ojs.aaai.org/index.php/ICWSM/article/view/14550⟩.

KASTRATI, Z. et al. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. *Applied Sciences*, MDPI, v. 11, n. 9, p. 3986, 2021.

KOTSIANTIS, S. B. Supervised machine learning: A review of classification techniques. In: *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. NLD: IOS Press, 2007. p. 3–24. ISBN 9781586037802.

LIU, B. Sentiment analysis and opinion mining. In: . [S.l.: s.n.], 2012. v. 5. ISBN 978-3-642-19459-7.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: . Red Hook, NY, USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119.

NADKARNI, P.; OHNO-MACHADO, L.; CHAPMAN, W. Natural language processing: An introduction. *Journal of the American Medical Informatics Association : JAMIA*, v. 18, p. 544–51, 09 2011.

ORTIGOSA, A.; MARTÍN, J. M.; CARRO, R. M. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, Elsevier, v. 31, p. 527–541, 2014.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, 2010.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1–2, p. 1–135, jan. 2008. ISSN 1554-0669. Disponível em: ⟨https://doi.org/10.1561/1500000011⟩.

RIBEIRO, F. N. et al. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, SpringerOpen, v. 5, n. 1, p. 1–29, 2016.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. *Trans. Sys. Man Cyber Part C*, IEEE Press, v. 40, n. 6, p. 601–618, nov. 2010. ISSN 1094-6977. Disponível em: ⟨https://doi.org/10.1109/TSMCC.2010.2053532⟩.

ROSSI, D. *PRED-INTER: automatic prediction of pedagogical interventions*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Juiz de Fora (UFJF), Juiz de Fora, 2022. Disponível em: ⟨https://repositorio.ufjf.br/jspui/handle/ufjf/14652⟩.

SMEDT, T. D.; DAELEMANS, W. Pattern for python. *The Journal of Machine Learning Research*, JMLR. org, v. 13, n. 1, p. 2063–2067, 2012.

TABOADA, M. et al. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 37, n. 2, p. 267–307, jun. 2011. ISSN 0891-2017. Disponível em: ⟨https://doi.org/10.1162/COLI_a_00049⟩.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, [Sage Publications, Inc., American Educational Research Association], v. 45, n. 1, p. 89–125, 1975. ISSN 00346543, 19351046. Disponível em: ⟨http://www.jstor.org/stable/1170024⟩.

VASWANI, A. et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: LIU, Q.; SCHLANGEN, D. (Ed.). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020. p. 38–45. Disponível em: ⟨https://aclanthology.org/2020.emnlp-demos.6/⟩.