UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise de sentimentos em avaliações de disciplinas: um estudo de caso no ensino superior

Erick Vansim Previato

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

Erick Vansim Previato

Análise de sentimentos em avaliações de disciplinas: um estudo de caso no ensino superior

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Wallace Correa de Oliveira Casaca

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Previato, Erick Vansim
P944a Análise de sentimentos em avaliações de disciplinas: um estudo de caso no ensino superior / Erick Vansim Previato; orientador Wallace Correa de Oliveira Casaca. -- São Carlos, 2025.
44 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Análise de sentimentos. 2. Aprendizado de máquina. 3. Processamento de linguagem natural. 4. BERT. 5. Ensino superior. I. Casaca, Wallace Correa de Oliveira, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Erick Vansim Previato

Sentiment analysis in subject evaluations: a case study in higher education

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Wallace Correa de Oliveira Casaca

Original version

São Carlos

2025

Dedico este trabalho à minha noiva, à minha família e aos meus amigos,
que sempre estiveram ao meu lado, oferecendo apoio,
amor e incentivo em cada etapa desta jornada.
Dedico também à nossa fiel companheira de quatro patas, Serena,
que partiu durante o MBA, mas deixou lembranças de amor incondicional,
alegria e lealdade. Sua presença iluminou nossos dias
e continuará viva em nossos corações para sempre.

AGRADECIMENTOS

Agradeço, em primeiro lugar, à minha noiva, Maria Fernanda, por todo o carinho, cuidado e incentivo, e por me inspirar diariamente a ser uma pessoa melhor.

Aos meus pais, expresso minha eterna gratidão pelo amor incondicional e pelo apoio constante em todas as etapas da minha vida.

Agradeço também ao MBA em Ciências de Dados pela oportunidade e pela bolsa concedida, e ao ICMC pela disponibilização dos dados que tornaram possível a realização deste trabalho.

Por fim, registro meu sincero agradecimento ao meu orientador, professor Wallace, pela paciência, disponibilidade e generosidade em compartilhar seu conhecimento, contribuindo de forma essencial para o desenvolvimento deste trabalho.

"Em tempos de crise, os sábios constroem pontes,
enquanto os tolos constroem muros."
T'Challa - Pantera Negra

RESUMO

PREVIATO, E. V. **Análise de sentimentos em avaliações de disciplinas: um estudo de caso no ensino superior**. 2025. 44 p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

A avaliação de disciplinas é um instrumento fundamental para a gestão acadêmica e a melhoria contínua no ensino superior, gerando um vasto volume de dados textuais que torna a análise manual inviável e custosa. Este trabalho propõe a aplicação de técnicas de Processamento de Linguagem Natural (PLN), especificamente a Análise de Sentimentos, para classificar automaticamente o feedback discente. O estudo de caso utilizou uma base de dados contendo cerca de 16.500 avaliações de um instituto da Universidade de São Paulo (USP), coletadas entre 2017 e 2023. A metodologia incluiu o pré-processamento dos dados e uma rotulação híbrida assistida por Large Language Models (LLM). Foram submetidos ao processo de *fine-tuning* modelos baseados na arquitetura *Transformer*, especificamente o BERTimbau (pré-treinado em português) e o XLM-RoBERTa (multilíngue), nas versões Base e Large. Os resultados demonstraram a superioridade dos modelos pré-treinados na língua nativa, com o BERTimbau-Large alcançando o melhor desempenho, registrando uma Acurácia de 90,67% e um *F1-Score* de 90,37%. O estudo conclui que a utilização de modelos de linguagem específicos para o português é eficaz para capturar nuances do contexto acadêmico brasileiro, oferecendo uma ferramenta robusta para apoiar a tomada de decisão institucional

Palavras-chave: Análise de sentimentos. Aprendizado de máquina. Processamento de linguagem natural. BERT. Ensino superior.

ABSTRACT

PREVIATO, E. V. **Sentiment analysis in subject evaluations: a case study in higher education**. 2025. 44 p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Subject evaluation is a fundamental tool for academic management and continuous improvement in higher education, generating a vast volume of textual data that makes manual analysis unfeasible and costly. This work proposes the application of Natural Language Processing (NLP) techniques, specifically Sentiment Analysis, to automatically classify student feedback. The case study used a dataset containing approximately 16,500 evaluations from an institute of the University of São Paulo (USP), collected between 2017 and 2023. The methodology included data preprocessing and a hybrid labeling strategy assisted by Large Language Models (LLM). Transformer-based models, specifically BERTimbau (pre-trained in Portuguese) and XLM-RoBERTa (multilingual), in both Base and Large versions, were submitted to fine-tuning. The results demonstrated the superiority of models pre-trained in the native language, with BERTimbau-Large achieving the best performance, recording an Accuracy of 90.67% and an F1-Score of 90.37%. The study concludes that using language models specific to Portuguese is effective in capturing nuances of the Brazilian academic context, offering a robust tool to support institutional decision-making.

Keywords: Sentiment analysis. Natural language processing. Machine learning. BERT. Higher education.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT	Bidirectional Encoder Representations from Transformers
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
MCC	Coeficiente de Correlação de Matthews
PCA	Principal Component Analysis
PLN	Processamento de Linguagem Natural
RoBERTa	A Robustly Optimized BERT Pretraining Approach
SVM	Support Vector Machines
USP	Universidade de São Paulo

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A era digital em que vivemos tem gerado um volume de dados sem precedentes, transformando a forma como interagimos, consumimos e expressamos nossas opiniões. A internet, as redes sociais, fóruns, plataformas de ensino e reviews de produtos tornaram-se repositórios vastos de informações textuais, que refletem os pensamentos, sentimentos e percepções dos indivíduos sobre os mais diversos tópicos, produtos e serviços. No contexto educacional, a opinião do aluno, expressa através de comentários e realização de avaliações, torna-se uma opção para a melhoria contínua da qualidade do ensino e da gestão acadêmica das instituições educacionais (Bóbó *et al.*, 2019).

A avaliação de disciplinas, em particular, é um processo crucial para instituições de ensino superior, pois permite coletar *feedback* direto dos estudantes sobre a qualidade do conteúdo, a metodologia de ensino, a atuação docente e a infraestrutura disponível (Romero; Ventura, 2020). Em muitos dos casos, essas avaliações são analisadas manualmente: um processo que se torna inviável e ineficiente diante de um grande volume de dados gerado ao longo dos anos. A análise individualizada de milhares de comentários, como no caso da base de dados de mais de 8 mil avaliações de disciplinas de um instituto da Universidade de São Paulo (USP), consome tempo e recursos significativos, além de ser suscetível a vieses humanos e à perda de informações valiosas contidas no conjunto dos dados.

Posto o cenário acima, uma ferramenta proeminente é a Análise de Sentimentos (também conhecida como *Opinion Mining*): um campo da Inteligência Artificial e do Processamento de Linguagem Natural (PLN) que visa extrair, identificar e quantificar as polaridades emocionais (positiva, negativa ou neutra) expressas em textos (Liu, 2012). Por meio da aplicação de técnicas computacionais, a análise de sentimentos permite transformar dados textuais não estruturados em informações acionáveis, oferecendo uma visão geral e automatizada do sentimento predominante em um conjunto de avaliações. Neste caso, o objetivo principal é compreender a atitude, o tom e as emoções expressas em um determinado conteúdo textual (Liu, 2020).

A aplicação da análise de sentimentos no contexto das avaliações de disciplinas de instituições de ensino permite que gestores educacionais obtenham uma compreensão do cenário de forma ágil e eficiente. Em vez de examinar cada comentário individualmente, é possível identificar padrões, tendências e pontos críticos em um conjunto de avaliações de uma ou mais disciplinas em um dado semestre, por exemplo. Isso facilita a tomada de decisões estratégicas, como a revisão de planos de ensino, a capacitação docente, a melhoria de recursos didáticos, a otimização das condições de oferta das disciplinas e a

evolução de equipamentos tecnológicos, contribuindo assim diretamente para a qualidade do ensino e a satisfação dos alunos.

1.2 Objetivos

A relevância da análise de sentimentos cresceu conforme a quantidade de informações também evoluiu, tornando-se uma ferramenta para gestão de tomada de decisão.

Portanto, este trabalho busca explorar a aplicação da análise de sentimentos para automatizar e otimizar o processo de avaliação de disciplinas, proporcionando aos gestores uma ferramenta para a compreensão da percepção dos alunos e a consequente melhoria contínua da qualidade educacional.

Em termos de aplicabilidade e prova de conceito, a pesquisa visa validar a escalabilidade da solução por meio de um estudo de caso real utilizando a base de dados de um instituto da USP, demonstrando seu potencial para replicação em outras instituições. Os fundamentos teóricos que embasam esses objetivos são detalhados no Capítulo 2.

2 REVISÃO BIBLIOGRÁFICA

2.1 Análise de Sentimentos

A Análise de Sentimentos, ou *Opinion Mining*, é um subcampo do Processamento de Linguagem Natural (PLN) que se dedica à identificação e extração da polaridade (positiva, negativa, neutra) e da subjetividade de textos, classificando-os de acordo com o sentimento expresso (Pang; Lee, 2004). O objetivo principal é determinar a atitude de um autor em relação a um tópico, produto, serviço ou entidade. Essa atitude pode ser um julgamento ou avaliação, um estado emocional, como raiva ou alegria, ou a intenção comunicativa, como a compra de um produto, por exemplo (Liu, 2012).

A origem da análise de sentimentos pode ser traçada até o início dos anos 2000, impulsionada pelo crescimento exponencial da internet e o consequente aumento na produção de conteúdo textual, como avaliações de produtos, posts em blogs e fóruns de discussão (Pang; Lee, 2004). Embora o estudo da opinião em textos não seja um conceito novo, a abordagem computacional para extrair e resumir essas opiniões ganhou destaque com o advento de técnicas de Aprendizado de Máquina e o volume massivo de dados disponíveis. Os trabalhos pioneiros de Pang, Lee e Vaithyanathan (2002) e Turney (2002) são frequentemente citados como marcos iniciais nesse campo, demonstrando a viabilidade de aplicar algoritmos de aprendizado supervisionado para classificar a polaridade de textos.

2.1.1 Tipos e Níveis de Análise de Sentimentos

A análise de sentimentos pode ser categorizada em diferentes tipos e níveis de granularidade, dependendo do objetivo e da profundidade da análise desejada. A Análise de Sentimentos Baseada em Léxico utiliza léxicos (listas de palavras) pré-definidos, onde cada léxico é associado a uma pontuação de sentimento (positiva, negativa ou neutra). O sentimento de um texto é então calculado com base na soma ou média das pontuações das palavras presentes no texto (Hu; Liu, 2004). Embora simples e eficaz para muitas aplicações, sua principal limitação reside na incapacidade de lidar com ironia, sarcasmo e o contexto em que as palavras são utilizadas.

A Análise de Sentimentos Baseada em Aprendizado de Máquina utiliza algoritmos de aprendizado de máquina, supervisionado ou não supervisionado, para classificar o sentimento de um texto. Modelos são treinados em grandes conjuntos de dados rotulados, no caso supervisionado, para aprender a associar características do texto a uma polaridade de sentimento (Medhat; Hassan; Korashy, 2014).

Em relação aos níveis de granularidade, a análise de sentimentos pode ser realizada em três níveis: nível de documento, nível de sentença e nível de aspecto ou entidade.

No nível de documento, a classificação do sentimento é realizada em um documento inteiro como positivo, negativo ou neutro. É útil para obter uma visão geral do sentimento em relação a um tópico (Pang; Lee, 2004).

No nível de sentença, o sentimento de cada frase é analisado individualmente dentro de um documento. Permite identificar nuances e diferentes sentimentos expressos em diferentes partes de um texto (Liu, 2012).

Já no nível de aspecto, o mais granular, foca na identificação do sentimento em relação a aspectos ou entidades específicas dentro de um texto (Liu, 2012). Por exemplo, em uma avaliação de disciplina, pode-se analisar o sentimento em relação ao professor, material didático ou metodologia de avaliação. Este nível é particularmente relevante para o presente trabalho, pois permite identificar sentimentos específicos para as categorias e critérios das avaliações.

2.1.2 Aprendizado Supervisionado e Não Supervisionado na Análise de Sentimentos

A escolha entre abordagens de Aprendizado Supervisionado e Não Supervisionado é fundamental na construção de um modelo de análise de sentimentos robusto e efetivo. A frente de Aprendizado Supervisionado requer um conjunto de dados rotulados, onde cada exemplo de texto já possui uma polaridade de sentimento predefinida, seja ela positiva, negativa ou neutra. O algoritmo aprende a mapear características do texto a partir desses rótulos (Patel; Sachin, 2024). A vantagem é que geralmente resulta em modelos com maior precisão quando há dados rotulados de alta qualidade e em quantidade suficiente. Já a desvantagem é com relação à rotulagem de dados, pois é um processo demorado e custoso, especialmente para grandes volumes de texto (Medhat; Hassan; Korashy, 2014). Algoritmos populares na vertente de Aprendizado Supervisionado são: Máquinas de Vetores de Suporte (SVM), *Naive Bayes*, Regressão Logística, Redes Neurais e Árvores de Decisão (Shaukat *et al.*, 2020).

O Aprendizado Não Supervisionado não requer um conjunto de dados rotulados. Algoritmos desta família buscam padrões e estruturas inerentes aos dados para inferir o sentimento. Métodos baseados em léxico são exemplos de abordagens não supervisionadas, onde a polaridade é inferida a partir de dicionários de sentimentos (Hu; Liu, 2004). Outras técnicas, como modelagem de tópicos, podem ser usadas para identificar temas e associar sentimentos a eles. Uma vantagem é que não exige o esforço de rotulagem manual, tornando-o mais escalável para grandes volumes de dados. Já a desvantagem é que pode apresentar menor precisão em comparação com abordagens supervisionadas e tem dificuldade em capturar nuances de linguagem, como sarcasmo e negação (Liu, 2012). Algoritmos clássicos na frente de Aprendizado Não Supervisionado são: Análise de Componentes Principais (PCA), *K-Means* (para agrupamento de textos por sentimento implícito) e Modelagem de Tópicos como o *Latent Dirichlet Allocation* (LDA) (Islam, 2018).

Para o contexto das avaliações de disciplinas, onde a rotulagem manual de 8 mil comentários seria um processo exaustivo, a consideração de abordagens não supervisionadas ou a combinação com técnicas semi-supervisionadas, onde uma pequena parte dos dados é rotulada manualmente e usada para treinar um modelo inicial, pode ser uma estratégia viável. Porém, caso haja possibilidade de utilizar a abordagem supervisionada, ou seja, caso não tenha um custo muito elevado para realizar a rotulagem dos dados, pode-se tornar uma estratégia mais assertiva.

2.1.3 Fatores Importantes na Análise de Sentimentos

A aplicação da análise de sentimentos não é trivial e envolve a consideração de diversos fatores que podem impactar a precisão e a utilidade dos resultados:

- Pré-processamento de Texto: Etapa utilizada para limpar e normalizar o texto antes da análise. Inclui a remoção de *stopwords* (palavras comuns como "e", "o", "a"), pontuações, caracteres especiais, lematização (redução de palavras à sua forma base) e *stemming* (redução de palavras ao seu radical) (Ganesan, 2022).
- Tratamento de Negação: A presença de negações como "não gostei" e "nunca farei", pode inverter o sentido de uma frase e deve ser tratada adequadamente para evitar classificações errôneas (Liu, 2012).
- Linguagem Específica e Gírias: Termos específicos de um domínio, como jargões acadêmicos e gírias, podem não estar presentes em léxicos genéricos e requerem adaptação ou construção de léxicos especializados (Pang; Lee, 2004).
- Ironia e Sarcasmo: A detecção de ironia e sarcasmo é um desafio significativo na análise de sentimentos, pois a polaridade literal das palavras é oposta à intenção do autor (Liu, 2012).
- Emojis e Emoticons: Em textos informais, emojis e emoticons são expressivamente utilizados para transmitir sentimentos e devem ser considerados na análise (Salgado, 2024).
- Modelo de Idioma: A análise de sentimentos para a língua portuguesa apresenta desafios específicos devido à sua complexidade gramatical, polissemia e riqueza de sinônimos, exigindo modelos e recursos linguísticos adaptados (Souza; Nogueira; Lotufo, 2020).

2.1.4 Ferramentas e Modelos

Nos últimos anos, a área de análise de sentimentos tem sido significativamente impactada pelo avanço dos modelos de *Deep Learning* e *Transformers*. Esses modelos, pré-treinados em grandes volumes de texto, demonstraram capacidades impressionantes

de compreensão de linguagem e geração de texto, superando abordagens tradicionais em muitas tarefas de PLN, incluindo a análise de sentimentos (Devlin *et al.*, 2019).

É importante ressaltar que a escolha da ferramenta ou modelo depende de diversos fatores, como o volume e a natureza dos dados, a disponibilidade de recursos computacionais, a necessidade de precisão e a granularidade da análise desejada. Para o contexto das avaliações de disciplinas em português, o uso de modelos *Transformers* pré-treinados em português, como o BERTimbau ou RoBERTa com *fine-tuning*, ou ajuste fino, para o domínio específico, pode oferecer os melhores resultados em termos de precisão e compreensão contextual.

2.1.5 Trabalhos Relacionados

A análise de sentimentos em avaliações de disciplinas no contexto educacional tem se mostrado uma ferramenta valiosa para compreender a percepção dos estudantes e subsidiar a gestão acadêmica. Alguns estudos recentes exploram essa aplicação, utilizando diferentes técnicas e abordagens para extrair informações de dados textuais.

Um estudo recente de Koufakou (2024) propõe uma arquitetura para mineração de sentimentos a partir de avaliações online de cursos. Para isso, utilizou-se uma abordagem baseada em aprendizado de máquina e modelos como o BERT e RoBERTa, destacando a importância da análise de sentimentos para melhorar a qualidade do ensino e a experiência do aluno. A pesquisa mostra como a identificação de padrões de sentimentos pode fornecer *feedback* construtivo para educadores e administradores de cursos, auxiliando na tomada de decisões estratégicas para uma melhoria contínua.

Na mesma direção, Shaik *et al.* (2023) investigaram a aplicação de técnicas de análise de sentimentos para avaliar o *feedback* de estudantes. O trabalho ressalta a capacidade da análise de sentimentos na identificação de pontos fortes e fracos dos programas, permitindo uma resposta mais ágil e direcionada às necessidades dos estudantes.

Já no estudo de Grljević, Bošnjak e Kovačević (2022), explorou-se o uso de *big data* e análise de sentimentos para entender a percepção dos alunos sobre o ambiente de aprendizado em instituições de ensino, utilizando informações de um *website* com mais de 8000 avaliações. Embora o foco seja mais amplo, o trabalho aborda a relevância da análise de sentimentos para aprimorar a experiência educacional, fornecendo informações sobre o engajamento e a satisfação dos estudantes.

Esses trabalhos demonstram a relevância e o potencial da análise de sentimentos para aprimorar a gestão educacional, oferecendo uma visão estratégica a partir de grandes volumes de avaliações, o que corrobora com o objetivo proposto para este trabalho.

No próximo capítulo, será detalhada a metodologia de desenvolvimento utilizada para o processo de análise de sentimentos em avaliação de disciplinas.

3 DESENVOLVIMENTO

Este capítulo detalha a metodologia de pesquisa empregada, abrangendo a descrição da base de dados utilizada, o processo de pré-processamento, a rotulação dos dados, o *fine-tuning* dos Modelos de Linguagem de Grande Escala (LLMs) e, por fim, as métricas de avaliação de desempenho dos modelos classificadores de sentimento desenvolvidos.

3.1 Descrição e Tratamento da Base de Dados

A base de dados utilizada neste trabalho é composta por cerca de 16.500 registros de avaliações anônimas de estudantes, coletadas no período de 2017 a 2023. Cada registro corresponde a uma avaliação individual, contendo informações referentes ao curso, disciplina, turma, data e hora da resposta, bem como as notas atribuídas e os comentários textuais dos alunos.

3.1.1 Estrutura dos Dados

O *dataset* apresenta a seguinte estrutura de colunas:

- Identificadores e Contexto: ID (identificador único), Curso, Disciplina, Turma e Data/Hora do preenchimento.
- Notas Quantitativas (C1 a C14): 14 colunas contendo notas inteiras de 1 a 5, distribuídas em três categorias de avaliação.
- Comentários Qualitativos (C15, C16, C17): 3 colunas de texto livre, que são o foco da análise de sentimentos.

As três categorias de avaliação são apresentadas na Tabela 1.

Tabela 1 – Categorias de avaliação e seus respectivos critérios

Categoria	**Notas**	**Comentário**	**Contexto**
I	C1, C2, C3	C15	Características e condições de oferta da disciplina
II	C4 a C9	C16	Práticas e atividades do docente na disciplina
III	C10 a C14	C17	Autoavaliação na disciplina

Fonte: Elaborada pelo autor (2025).

As colunas C15, C16 e C17 são o foco principal deste estudo, pois contêm os textos utilizados na análise de sentimentos. As demais colunas são utilizadas como atributos

de apoio, podendo contribuir para a contextualização dos sentimentos expressos pelos estudantes.

As Figuras 1 e 2 apresentam amostras do *dataset* para exemplificar a estrutura dos dados.

Figura 1 – Amostra do *dataset* - Parte 1

	ID	Cursos ID	Disciplina	Turma	Data Hora	C1	C2	C3	C4	C5	...
3022	2922	Bacharelado em Estatística	Tópicos de Matemática	2018101	2018-08-06 11:53:12	5	5	5	5	5	...
4280	4146	Bacharelado em Matemática Aplicada e Computaçã...	Cálculo I	2019101	2019-07-29 15:33:41	1	1	1	1	1	...
1290	1284	Engenharia de Computação	Modelagem Orientada a Objetos	2017201	2017-12-04 08:38:21	2	4	5	2	2	...
2691	2605	Bacharelado em Sistemas de Informação	Introdução à Teoria da Computação	2018101	2018-07-17 08:07:27	5	5	1	4	5	...
1268	1262	Bacharelado em Ciências de Computação	Laboratório de Introdução à Ciência da Computa...	2017201	2017-12-03 17:06:33	5	5	4	5	5	...

Fonte: Elaborada pelo autor (2025).

Figura 2 – Amostra do *dataset* - Parte 2

...	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
...	5	5	4	5	4	4	4	NaN	NaN	NaN
...	1	1	3	5	5	4	1	Aula confusa, provar teoremas sem uma explicaç...	ele não se preparava para as aulas, tentando s...	Ele não ligava muito pra sala, não deu trabalh...
...	4	5	3	4	3	2	3	NaN	NaN	NaN
...	5	5	4	3	3	4	4	NaN	NaN	NaN
...	5	5	5	4	5	3	5	NaN	NaN	NaN

Fonte: Elaborada pelo autor (2025).

3.1.2 Tratamento dos Dados

O processo de tratamento dos dados foi conduzido com o objetivo de garantir a qualidade e a consistência das informações textuais antes da aplicação das técnicas de Processamento de Linguagem Natural (PLN). As etapas de tratamento seguiram os seguintes passos.

- Verificação de dados ausentes: realizou-se uma inspeção para identificar e tratar valores nulos (`NaN`) ou vazios, especialmente nas colunas de texto (C15, C16 e C17). Caso os três comentários fossem ausentes, o registro era descartado;

- Uniformização de tipos de dados: garantiu-se que as colunas de datas (Data Hora) estivessem no formato de data/hora adequado e que as notas (C1 a C14) fossem tratadas como variáveis numéricas (inteiros);

- Limpeza de texto: embora o pré-processamento mais detalhado ocorra na Seção 3.2, uma limpeza inicial no texto livre incluiu a remoção de caracteres especiais não informativos ou pontuações excessivas (ex.: "!!!!!"), conversão para letras minúsculas e tratamento de *emoticons* ou *emojis*, que não foram observados na base.

Esses procedimentos visaram reduzir a dimensionalidade do texto e eliminar ruídos que poderiam comprometer o desempenho dos modelos de aprendizado de máquina.

3.1.3 Separação de Dados para Rotulação Manual

Para o treinamento dos modelos de classificação, é imprescindível um conjunto de dados com sentimento pré-rotulado. Uma amostra de aproximadamente 500 registros foi extraída da base de dados principal, buscando representar a diversidade temporal (2017–2023), de cursos e disciplinas, a fim de evitar viés em subconjuntos específicos.

Essa amostra foi exportada para um novo arquivo `.csv`, contendo as colunas com os comentários, C15, C16 e C17. Este arquivo serviu como base para o pré-processamento descrito na Seção 3.2, no qual foram adicionadas novas colunas, Sentimento_C15, Sentimento_C16 e Sentimento_C17 para armazenar os rótulos de sentimento (negativo, neutro e positivo) em cada uma delas.

3.2 Pré-processamento e Rotulação dos Dados

Esta seção detalha as etapas de pré-processamento e a estratégia de rotulação empregada para a criação do *dataset* de treinamento.

3.2.1 Pré-processamento de Texto

O pré-processamento do texto livre é crucial para otimizar o desempenho dos modelos de *Machine Learning* e LLMs. Os seguintes passos foram aplicados às colunas C15, C16 e C17:

- Remoção de *stop words*: eliminação de palavras comuns da língua portuguesa que não agregam valor ao sentimento (ex.: de, a, o, um, uma);

- Tokenização: divisão do texto em unidades menores (palavras ou subpalavras) que os modelos conseguem processar;

- Lematização/*Stemming*: redução das palavras às suas raízes (lematização) ou radicais (*stemming*) para tratar variações morfológicas (ex.: “estudando”, “estudou”, “estuda” para “estudar”);

3.2.2 Rotulação Auxiliada por LLM

A rotulação manual de 500 registros é um processo custoso em tempo. Para otimizar e agilizar o processo, utilizou-se uma estratégia de pré-classificação com LLM, seguida por validação humana.

Essa pré-classificação foi realizada com o auxílio do modelo Gemini, que analisou os comentários das colunas C15, C16 e C17 e atribuiu rótulos de sentimento positivo, negativo ou neutro. Isso gerou três novas colunas, Sentimento_C15, Sentimento_C16 e Sentimento_C17, contendo as classificações preliminares.

A validação humana foi então conduzida sobre os 500 registros, ajustando classificações ambíguas ou incorretas e garantindo que os rótulos refletissem fielmente o contexto acadêmico e a intenção do aluno.

Ao final desta etapa, obteve-se um *dataset* rotulado, pronto para ser utilizado no treinamento e *fine-tuning* dos modelos.

3.2.3 Preparação para o Treinamento do Modelo

Com os dados rotulados, a base foi dividida em dois subconjuntos: 80% para treinamento e 20% para validação. Essa divisão visa garantir um balanceamento adequado entre a capacidade de aprendizado e a generalização dos modelos.

3.2.4 Modelagem e Fine-Tuning dos Modelos de Linguagem

Após o pré-processamento e rotulação, aplicou-se a técnica de *fine-tuning* para adaptar modelos pré-treinados de arquitetura *Transformer* ao domínio específico dos comentários acadêmicos em português. Foram empregados dois modelos de referência, ambos pré-treinados em grandes corpora de texto em língua portuguesa: o BERTimbau e o RoBERTa (variantes em português).

O objetivo do *fine-tuning* foi gerar dois modelos classificadores otimizados, capazes de identificar sentimentos nos comentários das avaliações realizadas pelos alunos e classificá-los em positivo, negativo ou neutro.

3.3 Avaliação dos Modelos Classificadores

A performance dos modelos BERTimbau e RoBERTa, após *fine-tuning*, foi avaliada utilizando métricas padrão de classificação, calculadas a partir da Matriz de Confusão.

3.3.1 Matriz de Confusão e Métricas de Desempenho

A Matriz de Confusão é uma tabela que permite visualizar o desempenho de um algoritmo de classificação, indicando o número de classificações corretas e incorretas por classe. Os termos-chave utilizados são:

- Verdadeiros Positivos (VP): instâncias positivas classificadas corretamente como positivas;
- Verdadeiros Negativos (VN): instâncias negativas classificadas corretamente como negativas;
- Falsos Positivos (FP): instâncias negativas classificadas incorretamente como positivas (Erro Tipo I);
- Falsos Negativos (FN): instâncias positivas classificadas incorretamente como negativas (Erro Tipo II).

Com base na Matriz de Confusão, as métricas que foram utilizadas para comparar a eficácia dos classificadores são descritas a seguir.

Acurácia (*Accuracy*). Mede a proporção de predições corretas em relação ao total de instâncias. É uma métrica simples, mas pode ser enganosa em casos de desequilíbrio de classes.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

Precisão (*Precision*). Mede a proporção de instâncias classificadas como Positivas que são realmente Positivas. Responde à pergunta: "Das que o modelo classificou como Positivas, quantas estavam corretas?".

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Revocação (*Recall* ou Sensibilidade). Mede a proporção de instâncias Positivas que foram corretamente identificadas. Responde à pergunta: "Das que são realmente Positivas, quantas o modelo identificou?".

$$\text{Revocação} = \frac{VP}{VP + FN}$$

Métrica F1 (*F1-Score*). É a média harmônica da Precisão e da Revocação, sendo uma métrica que busca um equilíbrio entre ambas. É particularmente útil em casos de desequilíbrio de classes, onde se deseja que o modelo tenha tanto alta precisão quanto alta revocação.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Coeficiente de Correlação de Matthews (MCC). Mede a qualidade das classificações binárias, levando em conta verdadeiros e falsos positivos e negativos. É uma métrica robusta, especialmente em casos de classes desbalanceadas.

$$\text{MCC} = \frac{(VP \cdot VN) - (FP \cdot FN)}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Kappa de Cohen. Mede o grau de concordância entre as predições do modelo e os rótulos reais, ajustando para a concordância que poderia ocorrer por acaso.

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}$$

onde P_o é a proporção de observações concordantes e P_e é a proporção esperada de concordância por acaso.

Essas métricas permitem avaliar a capacidade dos modelos em classificar corretamente os sentimentos, considerando o equilíbrio entre falsos positivos e falsos negativos.

3.4 Conclusão

Este capítulo apresentou a metodologia adotada para o desenvolvimento dos modelos classificadores de sentimento, desde a descrição e tratamento da base de dados, passando pelo pré-processamento e rotulação, até o *fine-tuning* dos modelos e a avaliação de seu desempenho. As etapas descritas garantem a robustez do processo e a confiabilidade dos resultados obtidos, que serão discutidos no próximo capítulo.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta a avaliação experimental dos modelos de aprendizado profundo baseados em *Transformers*, aplicados à tarefa de análise de sentimentos de comentários de alunos sobre cursos de graduação.

4.1 Configuração experimental e métricas

Para garantir a confiabilidade dos resultados, o conjunto de dados processado foi dividido na proporção de 80% para treinamento e 20% para validação, utilizando estratificação para manter a distribuição original das classes, conforme descrito na Seção 3.2.

A avaliação dos modelos foi conduzida utilizando quatro métricas principais, escolhidas para oferecer uma visão global do desempenho dos classificadores, especialmente considerando o possível desbalanceamento das classes. As métricas utilizadas foram: Acurácia, *F1-Score*, Coeficiente de Correlação de Matthews (MCC) e Kappa de Cohen.

4.2 Comparativo de desempenho dos modelos

Foram treinados e avaliados dois modelos pré-treinados baseados em *Transformers*, sendo eles BERTimbau e RoBERTa. Dentre esses modelos, foram utilizados duas variações de cada um: BERTimbau-Base, BERTimbau-Large, XLM-RoBERTa-Base e XLM-RoBERTa-Large. A tabela 2 e o gráfico 3 apresentam um resumo dos resultados obtidos após o *fine-tuning* para cada modelo nas métricas selecionadas.

Tabela 2 – Comparação de métricas entre os modelos avaliados.

Modelo	**Acurácia**	***F1-Score***	**MCC**	**Kappa**
BERTimbau-Base	0.880000	0.875245	0.800280	0.798613
BERTimbau-Large	0.906667	0.903733	0.845077	0.843959
XLM-RoBERTa-Base	0.833333	0.769192	0.733353	0.701813
XLM-RoBERTa-Large	0.813333	0.749260	0.693264	0.668874

Fonte: Elaborada pelo autor (2025).

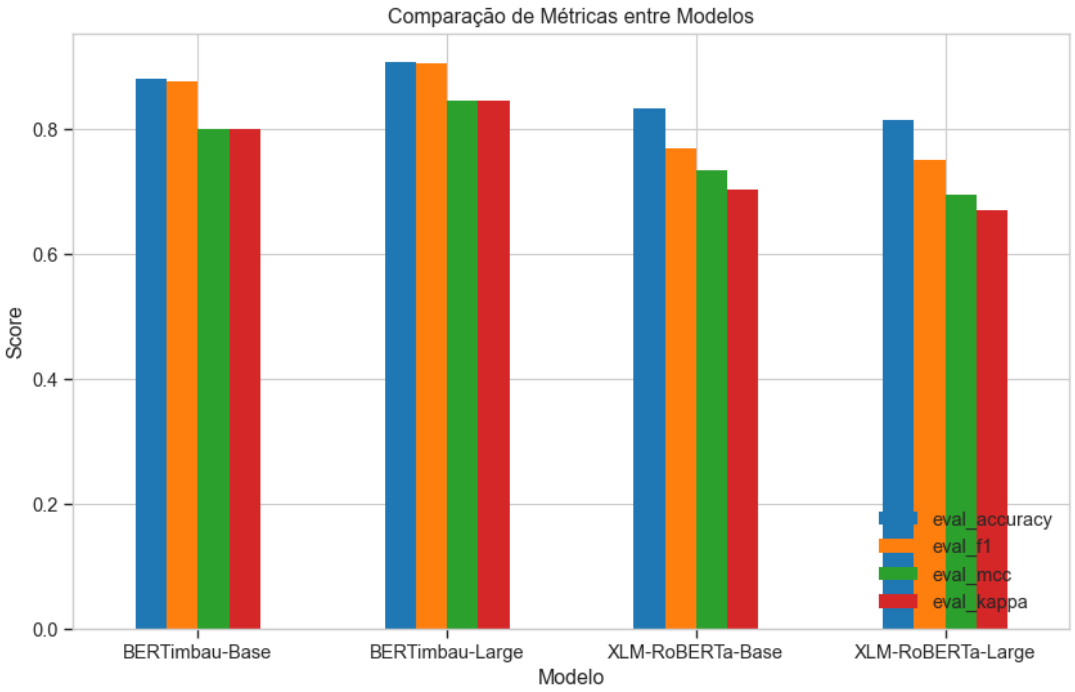

Figura 3 – Desempenho dos modelos nas métricas avaliadas.

Fonte: Elaborada pelo autor (2025).

Observa-se que o modelo BERTimbau-Large apresentou o melhor desempenho em todas as métricas avaliadas, alcançando uma acurácia de 90,67% e um *F1-Score* de 90,37%. Este resultado corrobora a hipótese de que modelos pré-treinados especificamente em língua portuguesa, que é o caso do BERTimbau, tendem a capturar melhor as nuances do idioma e, portanto, superar modelos multilíngues como o XLM-RoBERTa para tarefas específicas em português.

Ao comparar as versões Base e Large dos modelos, nota-se que o aumento na capacidade do modelo, mais camadas e parâmetros, contribui significativamente para a melhoria do desempenho, especialmente no caso do BERTimbau. Porém, como a versão Base já apresenta um desempenho robusto, ela pode ser considerada uma alternativa viável em cenários com restrições computacionais.

A análise do valor de MCC, acima de 0.84 para o BERTimbau-Large, indica uma forte correlação entre as previsões do modelo e as classes reais, validando a capacidade do modelo em lidar com a tarefa de classificação de sentimentos de forma eficaz.

4.3 Análise de erros e matriz de confusão

Para compreender as limitações dos modelos, analisou-se a matriz de confusão do melhor modelo, o BERTimbau-Large, conforme ilustrado na Figura 4.

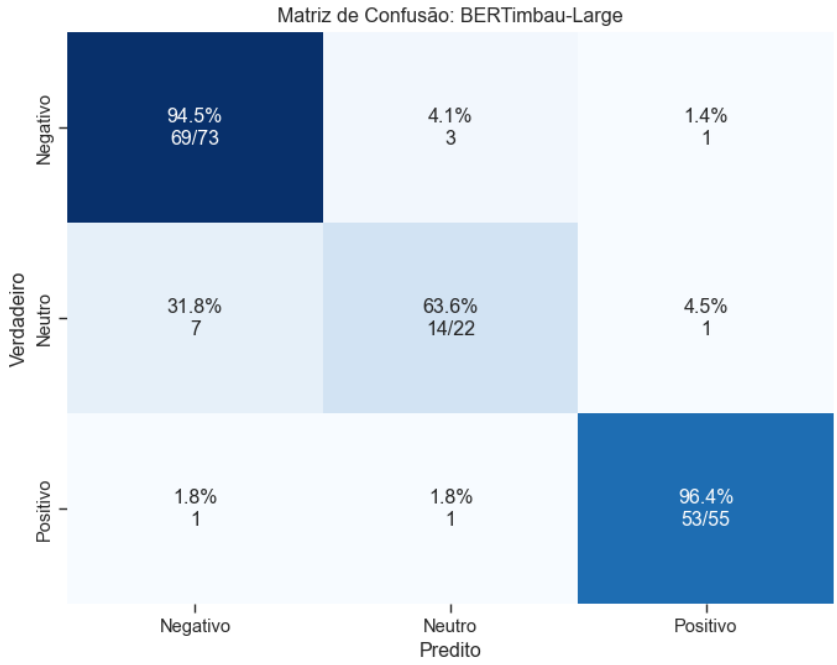

Figura 4 – Matriz de confusão do modelo BERTimbau-Large.

Fonte: Elaborada pelo autor (2025).

A análise da matriz de confusão revela que a maior fonte de confusão ocorre entre as classes Neutro e as classes polares, Positivo e Negativo. Isso é um fenômeno esperado em tarefas de análise de sentimentos, dado que comentários neutros frequentemente contêm elementos ambíguos ou ironias que modelos de linguagem, mesmo avançados, podem ter dificuldade em interpretar corretamente sem um contexto mais amplo. Além disso, a subjetividade na rotulação manual dos dados pode introduzir ruído, especialmente em casos onde o sentimento expresso não é claramente definido.

Por outro lado, a distinção entre as classes Positivo e Negativo foi realizada com alta precisão, com poucos casos de inversão de polaridades, falsos positivos ou falsos negativos, indicando que o modelo é eficaz em capturar os sinais mais evidentes de sentimento, o que é crucial para a aplicação prática do classificador em ambientes de produção.

4.4 Validação qualitativa com dados reais

Para testar a aplicabilidade do modelo em um cenário real, foi realizada uma validação qualitativa utilizando algumas frases inéditas, não presentes no conjunto de dados original. As frases foram selecionadas para representar diferentes nuances de sentimento. O modelo com o melhor desempenho classificou corretamente comentários como “Péssima didática, não aprendi nada e as provas não faziam sentido.” como Negativo, e “A ementa da disciplina é fantástica, muito atual e aplicada ao mercado.” como Positivo, ambos com alta confiança, maior que 95%.

Nessa validação, as seguintes frases foram testadas:

1. O professor explica muito bem, mas a sala é muito barulhenta.
2. Péssima didática, não aprendi nada e as provas não faziam sentido.
3. A ementa da disciplina é fantástica, muito atual e aplicada ao mercado.
4. Aula normal, nada demais.
5. O curso superou minhas expectativas, recomendo a todos!
6. Não há o que opinar sobre o docente.

A Tabela 3 apresenta os resultados da classificação dessas frases pelos quatro modelos treinados, com o número de confiança entre parênteses.

Tabela 3 – Resultados da validação qualitativa com frases inéditas.

Frase	**BERTimbau-Large**	**BERTimbau-Base**	**RoBERTa-Large**	**RoBERTa-Base**
1	Negativo (52%)	Negativo (71%)	Negativo (67%)	Negativo (73%)
2	Negativo (96%)	Negativo (91%)	Negativo (90%)	Negativo (85%)
3	Positivo (98%)	Positivo (97%)	Positivo (94%)	Positivo (95%)
4	Positivo (93%)	Negativo (56%)	Negativo (49%)	Positivo (91%)
5	Positivo (98%)	Positivo (97%)	Positivo (99%)	Positivo (95%)
6	Neutro (57%)	Neutro (63%)	Negativo (56%)	Negativo (41%)

Fonte: Elaborada pelo autor (2025).

Os resultados da Tabela 3 revelam que, embora todos os modelos apresentem alta performance em sentenças de polaridade explícita (Frases 2, 3 e 5), há uma divergência significativa em sentenças neutras ou ambíguas. Nota-se que os modelos da família BERTimbau foram os únicos capazes de identificar a neutralidade da Frase 6, enquanto os modelos RoBERTa tenderam a classificar falsos negativos. Esse fenômeno sugere que o pré-treinamento do BERTimbau em larga escala com dados do português confere uma vantagem semântica para interpretar expressões de indiferença ou objetividade típicas do ambiente acadêmico brasileiro.

No próximo capítulo, serão apresentadas as conclusões do trabalho, bem como sugestões para pesquisas futuras nesta área.

5 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho atingiu seu objetivo principal ao desenvolver e avaliar classificadores automáticos de sentimento para avaliações de disciplinas de uma Instituição de Ensino Superior, utilizando técnicas de *Transfer Learning* com modelos baseados em *Transformers*.

Os resultados obtidos demonstraram que a utilização de LLMs pré-treinados em português, especificamente o BERTimbau, supera abordagens genéricas ou multilíngues para este domínio específico. O modelo final alcançou uma acurácia de 90,67% e um *F1-Score* de 90,37%, métricas consideradas satisfatórias para implementação em ambiente de produção.

A análise de erros evidenciou que o principal desafio reside na classificação de comentários neutros ou ambíguos, sugerindo que a subjetividade humana na rotulação dos dados de treino é um fator limitante que o modelo tende a reproduzir.

5.1 Contribuições

As principais contribuições deste estudo incluem a estruturação de um pipeline completo de PLN para dados educacionais, desde a limpeza até o *fine-tuning*, a validação empírica da superioridade de modelos pré-treinados em português sobre multilíngues para análise de feedback acadêmico, e a entrega de um classificador capaz de processar grandes volumes de avaliações institucionais, permitindo aos gestores identificar gargalos pedagógicos com agilidade.

5.2 Trabalhos futuros

Para trabalhos futuros, sugere-se a exploração de técnicas de *data augmentation* para aumentar a diversidade do conjunto de treinamento, especialmente para a classe neutra, que apresentou maior dificuldade de classificação. Além disso, a incorporação de modelos híbridos que combinem *Transformers* com abordagens baseadas em regras ou dicionários sentimentais pode ser investigada para melhorar a robustez do classificador. Outra linha promissora é a adaptação do modelo para outras línguas e contextos educacionais, avaliando sua generalização e eficácia em diferentes ambientes acadêmicos.

Além da análise de sentimentos, futuros estudos podem explorar a extração de tópicos, facilitando a identificação automática de temas recorrentes nas avaliações, o que pode fornecer percepções adicionais para a melhoria contínua dos cursos oferecidos pela instituição.

REFERÊNCIAS

BóBó, M. *et al.* Análise de sentimento na educação: Um mapeamento sistemático da literatura. *In*: **Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019)**. [*S.l.: s.n.*]: Brazilian Computer Society (Sociedade Brasileira de Computação - SBC), 2019. (SBIE 2019), p. 249.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: https://arxiv.org/abs/1810.04805.

GANESAN, K. **The Business Case for AI: A Leader's Guide to AI Strategies, Best Practices & Real-World Applications**. Opinosis Analytics Publishing, 2022. ISBN 9781544528717. Disponível em: https://books.google.com.br/books?id=Vn34zgEACAAJ.

GRLJEVIć, O.; BOšNJAK, Z.; KOVAčEVIć, A. Opinion mining in higher education: a corpus-based approach. **Enterprise Information Systems**, Taylor & Francis, v. 16, n. 5, p. 1773542, 2022. Disponível em: https://doi.org/10.1080/17517575.2020.1773542.

HU, M.; LIU, B. Mining and summarizing customer reviews. *In*: **Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2004. (KDD '04), p. 168–177. ISBN 1581138881. Disponível em: https://doi.org/10.1145/1014052.1014073.

ISLAM, N. S. M. M. Comparative study on machine learning algorithms for sentiment classification. **International Journal of Computer Applications**, Foundation of Computer Science (FCS), NY, USA, New York, USA, v. 182, n. 21, p. 1–7, Oct 2018. ISSN 0975-8887. Disponível em: https://ijcaonline.org/archives/volume182/number21/30054-2018917961/.

KOUFAKOU, A. Deep learning for opinion mining and topic classification of course reviews. **Education and Information Technologies**, v. 29, n. 3, p. 2973–2997, 2024. ISSN 1573-7608. Disponível em: https://doi.org/10.1007/s10639-023-11736-2.

LIU, B. **Sentiment Analysis and Opinion Mining**. [*S.l.: s.n.*]: Morgan & Claypool Publishers, 2012. ISBN 1608458849.

LIU, B. Introduction. *In*: _____. **Sentiment Analysis: Mining Opinions, Sentiments, and Emotions**. [*S.l.: s.n.*]: Cambridge University Press, 2020. (Studies in Natural Language Processing), p. 1–17.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, v. 5, n. 4, p. 1093–1113, 2014. ISSN 2090-4479. Disponível em: https://www.sciencedirect.com/science/article/pii/S2090447914000550.

PANG, B.; LEE, L. **A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts**. 2004. Disponível em: https://arxiv.org/abs/cs/0409058.

PATEL, S.; SACHIN, D. "a comparative study of supervised and unsupervised classification techniques for sentiment analysis in imdb". *In*: . [*S.l.: s.n.*], 2024.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **WIREs Data Mining and Knowledge Discovery**, Wiley, v. 10, n. 3, jan. 2020. ISSN 1942-4795. Disponível em: http://dx.doi.org/10.1002/widm.1355.

SALGADO, H. Análise de sentimentos de tweets em língua portuguesa. *In*: . [*S.l.: s.n.*], 2024.

SHAIK, T. *et al.* Sentiment analysis and opinion mining on educational data: A survey. **Natural Language Processing Journal**, Elsevier BV, v. 2, p. 100003, mar. 2023. ISSN 2949-7191. Disponível em: http://dx.doi.org/10.1016/j.nlp.2022.100003.

SHAUKAT, Z. *et al.* Sentiment analysis on imdb using lexicon and neural networks. **SN Applied Sciences**, 2020.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. *In*: **Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I**. Berlin, Heidelberg: Springer-Verlag, 2020. p. 403–417. ISBN 978-3-030-61376-1. Disponível em: https://doi.org/10.1007/978-3-030-61377-8_28.