**Título do trabalho de conclusão de curso** (Deve conter no máximo 15 palavras e não deve conter expressões redundantes como: “Estudo de...”; “Influência de...”; “Elaboração de...” “Efeito de...“, “Análise de...”)

Rodrigo Franciozi Rodrigues da Silva¹\*;Adâmara Santos Gonçalvez Felício2

1 Analista de dados. Avenida Getúlio Vargas, 405 – Baeta Neves; 09751-250 São Bernardo do Campo, São Paulo, Brasil

2 Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País

\*autor correspondente: rodrigofranciozi@outlook.com

**Título do trabalho de conclusão de curso** (Deve conter no máximo 15 palavras e não deve conter expressões redundantes como: “Estudo de...”; “Influência de...”; “Elaboração de...” “Efeito de...“, “Análise de...”)

**Resumo**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares.

**Palavras-chave:** Análise de sentimentos; Aprendizado profundo; Classificação de texto; Modelos de aprendizado de máquina; PLN.

**Introdução**

Nos últimos anos a indústria audiovisual cresceu consideravelmente e junto a esse crescimento advém o interesse de agradar o público para que obras de sucesso sejam criadas. Uma maneira de se verificar tal sucesso é por meio da checagem geral de notas e pela leitura de comentários sobre críticos especializados e o público geral que já assistiram determinado conteúdo. (Yassine, 2019)

Quando observado o conjunto dos comentários produzidos sobre o assunto, identificamos plataformas online que trabalham no registro das críticas e avaliações por meio do texto escrito como o “Rotten Tomatoes”, “Internet Movie Database” [IMDb] e “Metacritic”. Esses sites contém uma biblioteca vasta de registros que funcionam como bases de dados, para que pesquisadores consigam avaliar o engajamento e potencial sucesso de produções cinematográficas. (Harsh et al., 2023)

As redes sociais têm um papel cada vez mais relevante, devido ao fato de muitos indivíduos as utilizarem para expressar seus sentimentos e ideias ao redor de um ou mais tópicos de interesse, sendo um deles as avaliações a respeito de filmes e séries da atualidade. Consumidores ao redor do mundo costumam utilizar essas avaliações como um norteador para decidir se determinada obra vale ou não o investimento de tempo e dinheiro por parte deles. Sabendo disso, produtores de filmes podem se apoiar nesses comentários para conseguir potencialmente prever quais tem altas chances de sucesso ou fracasso. (Ubaid e Faisal, 2023)

A análise de sentimentos, conhecida também como mineração de opinião, consiste em uma técnica que avalia a polaridade dos sentimentos, emoções e atitudes de determinado indivíduo de acordo com um objeto de interesse podendo ele ser um filme, produto, serviço e organização Azilawati et al. (2019). Essa tecnologia se baseia em inteligência artificial que usa o processamento de linguagem natural [PLN] para transformar textos não estruturados em dados normalizados e adequados para análises de algoritmos de aprendizado de máquina, que por sua vez auxiliam na identificação de “insights” e tomada de decisão a respeito de um determinado tópico de interesse. (Ubaid e Faisal, 2023)

De acordo com Yassine (2019), até pouco tempo essas análises de comentários eram feitas de forma bastante manual, mas com o advento de técnicas de aprendizado de máquina como a análise de sentimentos, essa realidade acabou sendo modificada.

O objetivo principal desse trabalho é investigar e comparar a eficácia de modelos conhecidos como clássicos e de aprendizado profundo, especificamente o Naïve Bayes e “Bidirectional Encoder Representations from Transformers” [BERT]. Procurou-se determinar qual modelo oferece maior acurácia, precisão, revocação e pontuação F1 ao classificar sentimentos contidos em comentários de filmes de acordo com as polaridades positiva, negativa e neutra.

**Material e Métodos**

Os TCCs dos cursos de especialização USP/ESALQ devem apresentar um caráter aplicado. Pesquisas bibliográficas ou documentais, ou seja, ‘formas’ de pesquisa baseadas exclusivamente em dados não experimentais, serão consideradas e aceitas somente se elaboradas utilizando métodos de análise de metadados ou grupos de dados. Entre os trabalhos com esse tipo de abordagem, também conhecidos como estudos secundários, são aceitos o Mapeamento Sistemático [MS] e a Revisão Sistemática [RS]. Mais informações sobre os estudos secundários podem ser consultadas no arquivo de Instruções para elaboração de Trabalhos de Conclusão de Curso. Todos os trabalhos deverão ser executados de forma que os(as) alunos(as) utilizem as ferramentas e os conhecimentos adquiridos ao longo do curso. Situações excepcionais serão avaliadas pela coordenação do programa. Recomenda-se que a escolha do tema seja realizada em função da identificação com a área de conhecimento e/ou demanda profissional do aluno. É fundamental que antes de iniciar a redação do seu TCC, o aluno considere se terá acesso irrestrito aos dados necessários, fontes de consulta e, principalmente, se a metodologia da pesquisa está ao alcance do conhecimento e domínio do assunto, tanto do(a) aluno(a) quanto do(a) orientador(a).

Considerando que as pesquisas envolvendo a participação, direta ou indireta, de seres humanos, a experimentação com animais e aquelas pesquisas que representem risco potencial ao meio ambiente devem obedecer a um conjunto de diretrizes específicas, é responsabilidade do aluno e orientador conhecerem as regras pertinentes à sua área de pesquisa. Para mais informações, consulte o manual de “Normas para Elaboração do Trabalho de Conclusão de Curso”.

O título da seção Material e Métodos deve ser grafado em negrito, alinhado à esquerda, as primeiras letras das palavras em letra maiúscula. O texto da seção deve ser redigido no pretérito perfeito do indicativo (passado), na forma impessoal e pode ser dividido em subtópicos com formatação de acordo com a descrição no item 1.1 Formato e margens.

Neste tópico deve ser apresentada a descrição detalhada de todo o material utilizado e todos os métodos utilizados na obtenção de dados e análise dos resultados da pesquisa, de forma a permitir fácil compreensão e interpretação dos resultados, bem como a reprodução do estudo ou a utilização do método por outros profissionais. No entanto, apenas novos procedimentos devem ser descritos detalhadamente; procedimentos previamente publicados (descritos na literatura) devem ser citados e referenciados, mencionando de modo breve quaisquer modificações significativas efetuadas.

Na descrição do local da pesquisa não deve ser mencionado o nome da propriedade ou instituição ou empresa onde foi realizado o estudo, ou onde foram coletados os dados, bem como nome do(s) proprietário(s). O local deve ser identificado apenas pela localização geográfica (coordenadas) ou nome da cidade e estado. A população estudada, técnicas e métodos de amostragem, espécie animal ou vegetal utilizada, com a pertinente classificação zoológica ou botânica, técnicas agronômicas ou zootécnicas e outras, além da descrição dos procedimentos analíticos, devem ser detalhados na sequência lógica em que o trabalho foi conduzido. A opção pela discriminação ou identificação da propriedade, instituição ou empresa, bem como do nome do eventual proprietário, imputa ao autor a inteira responsabilidade pela obtenção da autorização, junto aos mesmos, para a citação e consequente total isenção do MBA USP/ESALQ por esta opção.

**Resultados Preliminares**

Utilizou-se uma matriz de confusão multiclasses para quantificar a precisão dos modelos de classificação propostos, um método que proporciona uma visão clara de seus comportamentos em todas as classes. Nessas matrizes as colunas representam as previsões do modelo e as linhas os valores reais. Cada célula mostra o número de amostras real que foram previstas na categoria correspondente e está representada na Figura 2.

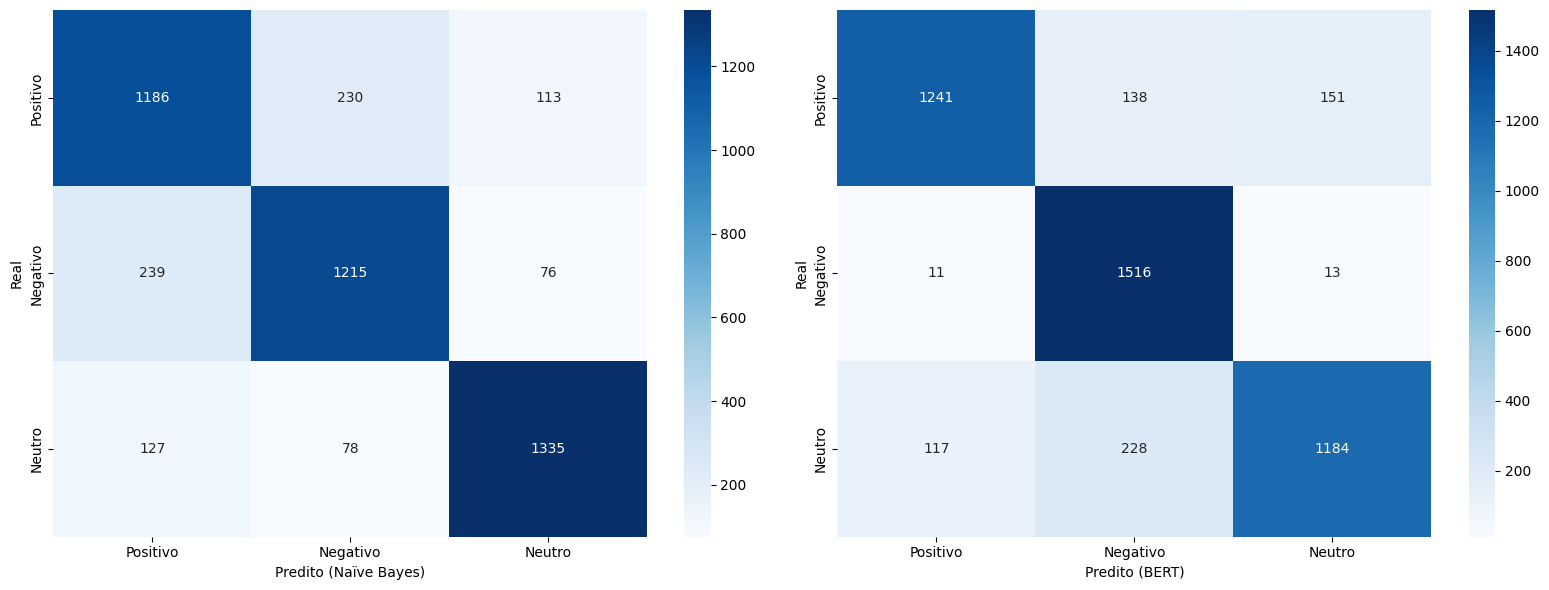


Figura 2. Comparação entre as matrizes de confusão dos modelos Naïve Bayes e BERT  
Fonte: Resultados originais da pesquisa

A partir das matrizes consegue-se chegar em mais algumas métricas importantes como a acurácia que representa o percentual de sentenças que foram corretamente classificadas, tendo como resultado os valores de 0,81% e 0,88% para os modelos Naïve Bayes e BERT respectivamente.

Além da acurácia outras três métricas bem comuns como processo de avaliação podem ser calculadas quando se trata de modelos de classificação. Entre elas temos a precisão que consiste nas previsões corretas de uma classe com relação a todas as previsões feitas para a mesma classe, a revocação que representa as previsões corretas de uma classe com relação ao total de casos que realmente pertencem a essa classe e a pontuação f1 que é uma média harmônica entre a precisão e revocação, fornecendo um valor que equilibra ambas as métricas. A Figura 3 demonstra a comparação entre as três métricas citadas. (Fabrício et al., 2015)

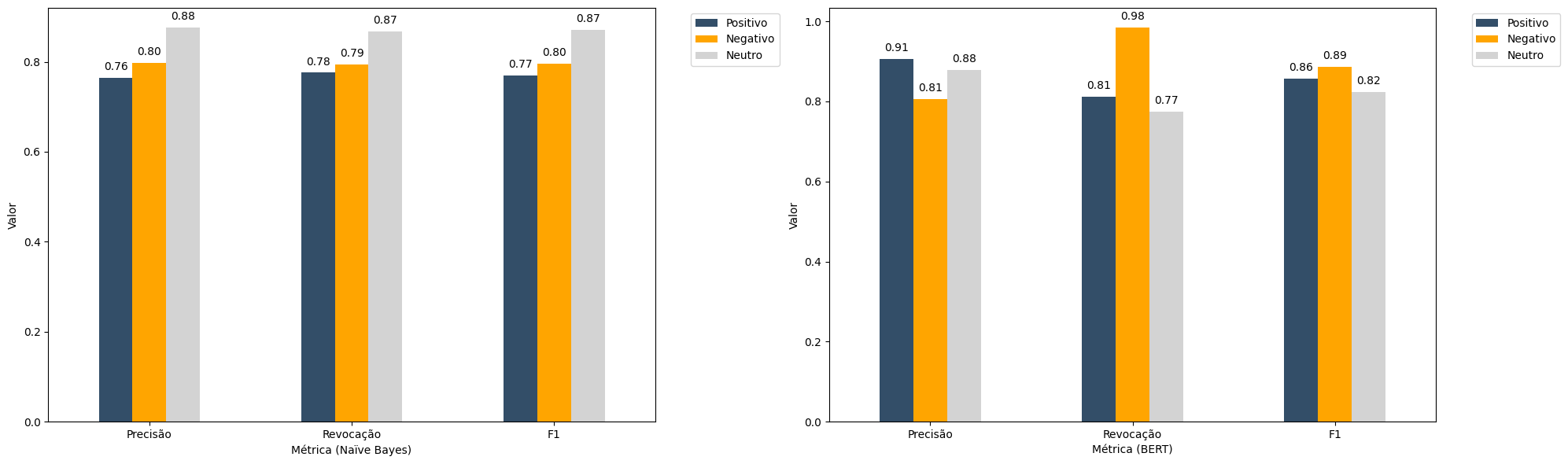


Figura 3. Comparativo de métricas de avaliação para modelos de classificação Naïve Bayes e BERT  
Fonte: Resultados originais da pesquisa

**Referências**

Azilawati Azizan, Nurul Najwa SK Abdul Jamal, Mohammad Nasir Abdullah, Masurah Mohamad, and Nurkhairizan Khairudin. Lexicon-based sentiment analysis for movie review tweets. 1st International Conference on Artificial Intelligence and Data Sciences, 2019.

Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. Sensors, 2023.

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

Fabrício Bevenuto, Filipe Ribeiro, Matheus Araujo. Métodos para Análises de Sentimentos em Mídias Sociais. Curso rápido em uma conferência de webmedia, 2015.

Harsh Sharma, Satyajit Pangaonkar, Reena Gunjan, Prakash Rokade. Sentimental Analysis of Movie Reviews Using Machine Learning. ITM Web of Conferences, 2023.

Hota HS, Sharma DK, Verma N. Lexicon-based sentiment analysis using Twitter data: a case of COVID-19 outbreak in India and abroad. Data Science for COVID-19, 2021.

Mine Dogucu & Mine Çetinkaya-Rundel. Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities, Journal of Statistics and Data Science Education, 2021.

Murtadha B. Ressan, Rehab F. Hassan. Department of Computer science, University of Technology, Baghdad, Iraq, 2022.

Rodani, Yassine. Movie Sentiment Analysis: A Multinomial Naıve Bayes-Based Approach for Assessing User and Critic Opinions, 2019.

Ubaid Mohamed Dahir, Faisal Kevin Alkindy. Utilizing Machine Learning for Sentiment Analysis of IMDB Movie Review Data. International Journal of Engineering Trends and Technology, 2023.

**Apêndice ou Anexo** (opcional)

Apêndices são textos e/ou documentos que foram elaborados pelo autor e que são importantes para complementar a argumentação do trabalho. Anexos são textos ou documentos que ilustram, mas que não foram elaborados pelos autores. Apêndices deverão seguir as mesmas normas de formatação do restante do texto, inclusive para figuras e tabelas.

O TCC deverá conter no máximo 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s).