**Comparação entre Modelos para Análise de Sentimentos no Contexto de Avaliações de Filmes**

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 Doutora em Desenvolvimento Econômico, UNICAMP. Orientadora MBA Data Science & Analytics. – Rua Alexandre Herculano, 120 – Vila Monteiro; CEP:13418-445 Piracicaba, São Paulo, Brasil.

\*autor correspondente: rodrigofranciozi@outlook.com

**Comparação entre Modelos para Análise de Sentimentos no Contexto de Avaliações de Filmes**

**Introdução**

Nos últimos anos a indústria audiovisual cresceu consideravelmente e junto a esse crescimento advém o interesse de agradar ao 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 se tornado plataformas cada vez mais relevantes para a expressão de sentimentos e ideias, especialmente no que diz respeito a filmes e séries. Segundo Faisal e Ubaid (2023), muitos indivíduos utilizam esses espaços virtuais para compartilhar suas avaliações sobre obras cinematográficas, influenciando assim as decisões de outros consumidores em relação ao investimento de tempo e dinheiro em determinadas produções.

Essa tendência de compartilhamento de opiniões e avaliações não apenas molda a percepção do público, mas também serve como um valioso recurso para os produtores de conteúdo. Conforme apontado por Haoran (2023), os produtores podem aproveitar essas informações disponíveis nas redes sociais para obter as reações do público, facilitando a identificação de potenciais sucessos ou fracassos, permitindo-lhes adaptar e aprimorar suas obras cinematográficas.

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 (Faisal e Ubaid, 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.

A importância do emprego de técnicas de PLN na análise de sentimentos de filmes, reflete tanto na automatização do processo de coleta e interpretação de um grande volume de dados, como também oferece uma visão mais quantitativa das percepções sobre determinados espectadores. Ao utilizar PLN, pesquisadores e profissionais podem extrair nuances que outrora seriam difíceis de se identificar manualmente, facilitando assim uma compreensão mais abrangente a respeito das reações do público (Yankang Su e Zbigniew j. Kabala, 2023).

O objetivo principal desse trabalho é 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**

**Base de dados**

A base de dados foi obtida por meio da aplicação de uma técnica de raspagem de dados (“web scrapping”) dentro da plataforma do IMDB. “Web Scrapping” consiste na extração de dados da internet de maneira programática, transformando-os em uma base de dados estruturada e permitindo uma coleta de grandes volumes de maneira automatizada minimizando potenciais erros (Mine Çetinkaya-Rundel e Mine Dogucu, 2021).

A lista de filmes foi selecionada de maneira aleatória levando em consideração os anos de lançamento que se encontram entre 2022 e 2023, para que as classificações fossem realizadas em comentários avaliados recentemente, mantendo apenas comentários em inglês devido ao site ser de origem norte américa e a natureza desses comentários estar majoritariamente na língua nativa. Após a obtenção e estruturação dos dados, técnicas de limpeza e transformação como a remoção de espaços em branco, emoticons, números e caracteres especiais, assim como a padronização de palavras para letras minúsculas e remoção de palavras vazias foram aplicadas. As palavras vazias representam uma lista de palavras que tipicamente aparecem com uma certa frequência na construção de frases, mas que não denotam uma significância relevante ao usuário final Faisal e Ubaid (2023). O fluxograma que implementa as etapas de transformação está descrito na Figura 1.

Ao final de todas as etapas obteve-se um *n* amostral de 14.480 comentários para a definição das polaridades juntamente com o treino e teste dos modelos.

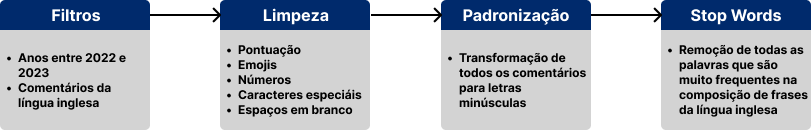


Figura 1. Fluxo de limpeza e tratamento de dados  
Fonte: Dados originais da pesquisa

**Métodologias**

Foi necessário a implementação de um processo de rotulagem de comentários, para que esses possam ser classificados de acordo com a polaridade (positivo, negativo e neutro) presente no texto. Essa se torna uma das grandes dificuldades, devido a subjetividade presente na tarefa de classificação, assim como o tempo necessário para que a rotulagem seja feita.

Foram implementadas duas técnicas de aprendizado não supervisionado a método de comparação para facilitar o processo de rotulagem, que consiste na implementação de dicionários léxicos contendo palavras que já possuem sentimentos atrelados e expressos de uma maneira quantitativa, podendo ser números entre -1 e 1 onde -1 representa uma polaridade mais negativa, 0 neutra e 1 positiva (Fabrício *et al*., 2015).

As técnicas escolhidas foram o “Opinion Lexicon”, composto por uma lista de aproximadamente 13.000 palavras incluindo gírias e abreviações na língua inglesa, que foi desenvolvido a partir de textos de avaliações de sites de compra e o “Valence Aware Dictionary and Sentiment Reasoner” [VADER] que é uma composição de outros dicionários já bem estabelecidos, com o adendo de emoticons, acrônimos e gírias (Fabrício *et al*., 2015).

Ambas as saídas dos dicionários são valores que podem ser -1, 0 e 1, representando as polaridades. Esses valores são somados para cada uma das avaliações e classificados de acordo com os resultados, sendo negativos para somas menores do que 0, positivos para maiores que 0 e neutras iguais a zero.

Dois modelos de aprendizado de máquina a título de comparação foram treinados e testados, avaliando métricas como Acurácia, Sensibilidade, Precisão e Teste F1 a fim de escolher aquele que melhor se ajustou a tarefa de predição das classificações de polaridade.

Devido aos rótulos estarem desbalanceados, uma técnica de “oversampling” envolvendo repetição foi aplicada com o auxílio do método “resample”, tendo como base a classificação positivo que contém o maior número de exemplos. Essa técnica de “oversampling” é utilizada quando se tem dados insuficientes, fazendo com que haja uma expansão da base em relação as amostras com menor volume de informações, gerando mais dados ao invés de deletar os já existentes. Caso essa técnica não seja aplicada, pode existir um impacto no treinamento dos modelos justamente por ferir a capacidade de generalização buscada, tendo em vista que com menos exemplos fica mais complicado do modelo conseguir identificar nuances (Nassera *et al.*, 2023).

Para ambos os modelos foi realizado um particionamento de 80:20 entre os dados alocados para treinamento e teste respectivamente. Há 100 anos atrás, o economista Italiano Wilfred Pareto fez a observação de que 80% de toda a riqueza estava concentrada em 20% da população, o que mais tarde a partir do desenvolvimento de uma fórmula matemática acabou descrevendo esse fenômeno de distribuição desigual presente em seu país. Essa formulação ficou conhecida como distribuição de Pareto e mais tarde, em 1940 como regra de Pareto devido a generalização por parte de J.M Juran. Essa abordagem segue o princípio de Pareto onde 20% dos fatores estão representados nos 80% outros fatores (Rosie *et al.*, 2014).

O primeiro modelo treinado foi o Naïve Bayes. Esse modelo é derivado do teorema de Bayes, onde se calcula a probabilidade de um evento acontecer com base em probabilidades condicionais, podendo ser aplicado no contexto de análise de sentimentos que leva em consideração a probabilidade de um texto pertencer a determinada classificação (Murtadha *et al.*, 2022).

A técnica de vetorização denominada “term frequency-inverse document frequency” [TF-IDF] foi aplicada permitindo a conversão de dados textuais em numéricos. Essa estatística reflete a importância das palavras dentro de um documento e é definida pelo número de vezes que determinada palavra aparece, em comparação com o número de documentos que contenham a palavra. Entenda documentos como sendo cada comentário presente na base de dados (Christine *et al*., 2023).

Por se tratar de uma classificação multi-classes, ou seja, contendo mais do que duas categorias a serem classificadas, foi aplicado o Naïve Bayes multinomial que é especificamente adequado para variáveis que apresentam contagens de palavras e que normalmente são obtidas usando técnicas de bolsa de palavras ou TF-IDF. Esse modelo assume como base que as palavras pertencentes a uma determinada classe são independentes entre si de acordo com determinada classificação, seguindo uma distribuição multinomial. Dessa forma calcula-se a probabilidade de um documento pertencer a essa classe, de acordo com a frequência de contagem das palavras naquele documento (Christine *et al.*, 2023).

O segundo modelo treinado foi o BERT, que consiste em um modelo de rede neural desenvolvido pela Google em 2018 com o intuito de compreender o contexto de frases, auxiliando máquinas no processo de linguagem natural com o emprego da arquitetura de transformadores e transferência de aprendizado. BERT é o primeiro modelo a atingir o estado da arte com relação a performance para atividades em nível de sentença e “token” (Jacob *et al*. 2018).

Uma versão pré-treinada do BERT foi utilizada, mas com uma adaptação referente ao volume de classes preditas que passou de duas para três. Devido a essa mudança nas classes apenas a estrutura definida pela rede neural foi aproveitada, havendo a necessidade da realização de um retreino com os dados presentes no estudo. A técnica de “tokenização” foi aplicada para configurar corretamente os dados da camada de entrada da rede neural, habilitando assim a possibilidade de treinamento do modelo.

A função de perda escolhida foi a entropia cruzada que é universalmente utilizada em trabalhos que tangem processos de classificação em redes neurais. Além disso, essa função se torna mais favorável em termos de otimização quando se trabalha com classificações de mais de duas classes Like e Mikhail (2021). O otimizador mais comum e o escolhido para atualização dos pesos das redes foi o “Adaptative Moment Estimation with Weight Decay” [AdamW] que consiste em uma versão modificada da otimização estocástica de primeira ordem Adam. Ambos de diferem devido ao ADAMw omitir uma das etapas de correção de viés auxiliando no processo de evitar o sobre ajuste na etapa de treinamento (Tianyi *et al*., 2021).

Devido a limitações técnicas presentes no processo de treinamento, foi considerado um tamanho de lote de 32 e um total de 3 épocas. Como a função de perda continuou apresentando uma redução ao invés de estagnação, isso dá uma margem para a continuação do treinamento considerando um volume maior de épocas, devido ao fato dos algoritmos de aprendizado profundo terem como objetivo principal a minimização das funções de perda entre as predições dadas pelo modelo e os valores reais pertencentes a base de dados de treinamento (Alexander *et al.*, 2023).

Todos as etapas foram aplicadas por meio da linguagem Python, em particular com o uso das bibliotecas “selenium”, “requests” e “random” para extração de dados da internet e aplicação de aleatoriedades, “nltk”, “pandas” e “re” para limpeza, tratamento e pré-processamento dos dados, “sklearn” e “transformers” para implementação de modelos de aprendizado de máquina e “mathplotlib” e “seaborn” para a criação de gráficos e visualizações.

**Resultados Preliminares**

A Figura 2 mostra 4 classificações realizadas para ambos os dicionários léxicos, sendo consideradas padrão aquelas contendo apenas a implementação da técnica, e customizada que envolveram uma remoção de palavras extras que aparecem com muita frequência em ambas as avaliações, sem trazer um significado relevante.

Figura 2. Classificações das polaridades por meio dos dicionários léxicos  
Fonte: Resultados originais da pesquisa



O dicionário “Opinion Lexicon” customizado foi escolhido como rótulo oficial das avaliações, devido a apresentar uma sensibilidade maior em suas classificações, principalmente levando em consideração os comentários de natureza neutra.

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 3.

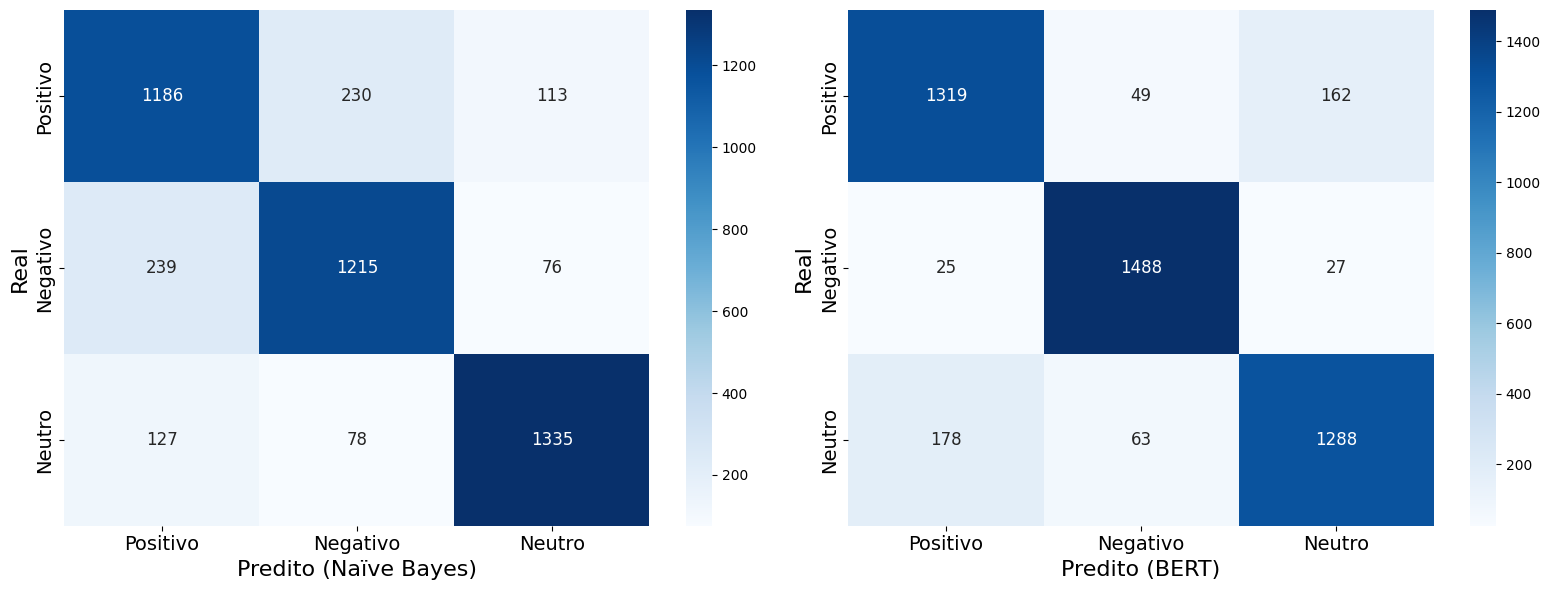


Figura 3. Comparação entre as matrizes de confusão dos modelos Naïve Bayes e BERT  
Fonte: Resultados originais da pesquisa

O modelo Naïve Bayes obteve um desempenho comparativamente melhor na classificação dos sentimentos de categoria neutra, o que não se aplicou para as duas outras categorias, sugerindo uma maior dificuldade em diferenciar as classes positivo a negativo. Por outro lado, o BERT apresentou um desempenho superior, com quantidades significativamente maiores de predições corretas e menos confusões entre as duas categorias, como evidenciado pelos valores elevados na diagonal principal.

No geral as taxas de acurácia total para ambos os modelos foi de 0,81% para o Naïve Bayes e 0,89% para o BERT, o que denota uma eficácia maior nas classificações por parte do modelo de rede neural sendo consistente com a natureza do algoritmo que, como um modelo de linguagem textual profundo, é capaz de compreender com maior capacidade a complexidade e os matizes do texto.

Além da acurácia, outras três métricas bem comuns como processo de avaliação foram 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 4 demonstra a comparação entre as três métricas citadas.

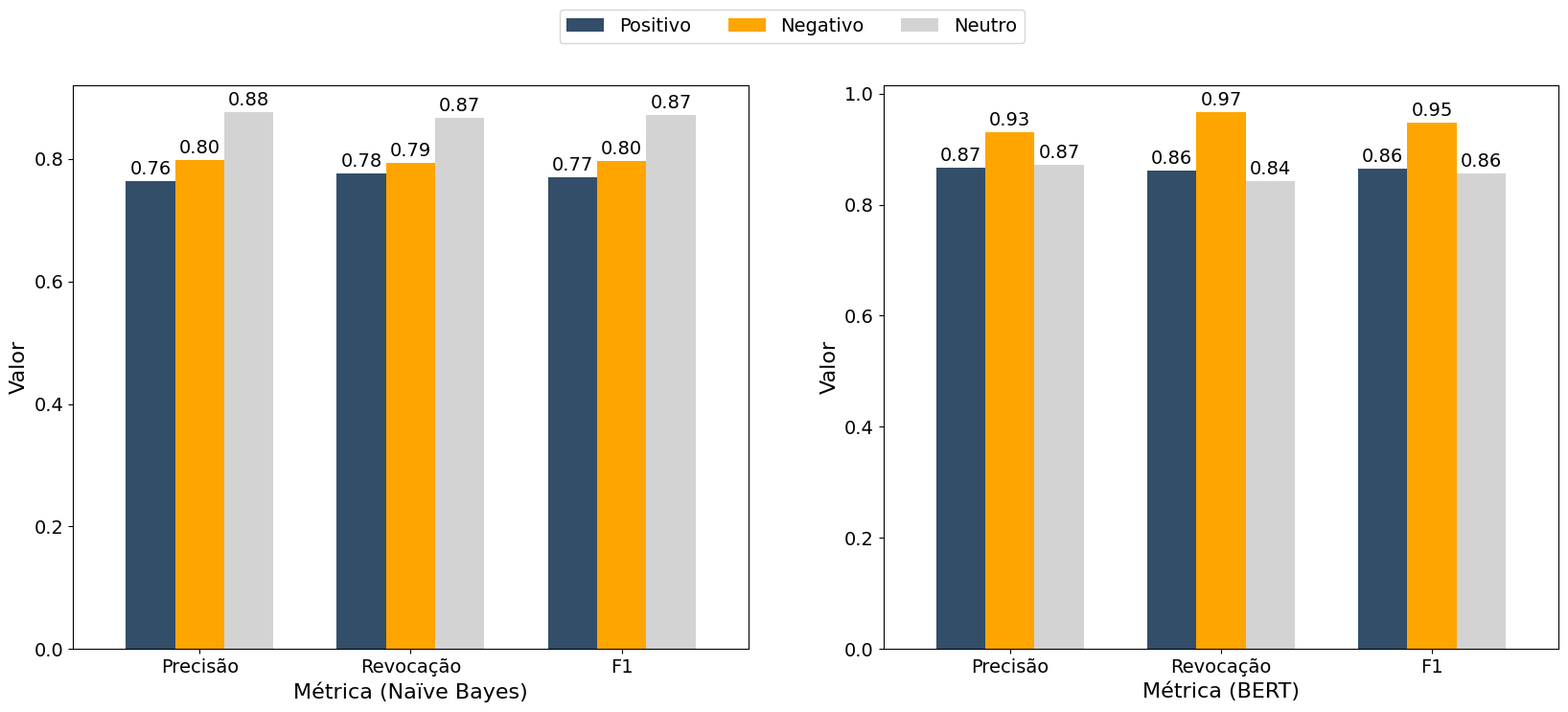


Figura 4. Comparativo de métricas de avaliação para modelos de classificação Naïve Bayes e BERT  
Fonte: Resultados originais da pesquisa

No gráfico correspondente ao Naïve Bayes, o modelo apresentou um equilíbrio entre todas as métricas para as diferentes classes de sentimentos, indicando uma consistência em seu desempenho. Essa consistência é importante, pois sugere que o modelo não favoreceu indevidamente nenhuma classe em particular durante o processo de classificação.

Embora no cenário geral o modelo BERT tenha superado o Naïve Bayes, o modelo mais simples exibiu uma capacidade notável nas classificações dos sentimentos neutros, o que é relevante em contextos onde essa distinção é crucial.

Por outro lado, o modelo de rede neural evidenciou uma performance superior, com uma predominância particularmente expressiva para as métricas revocação e pontuação F1, levando em consideração os sentimentos das classes positiva e negativa. O BERT com sua arquitetura avançada, demonstrou uma consistência que contribui para uma taxa menor de falsos negativos e uma e uma melhor identificação global dos sentimentos.

**Referências**

Alexander Demidovskij, Artyom Turgaryov, Aleksei Trutnev, Marina Kazyulina, igor Salnikov, Stanislav Pavlov. Lightweight and Elegant Data Reduction Strategies for Training Acceleration of Convolutional Neural Networks. MDPI Journal, 2023.

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

Abayomi Bello, Man-Fai Leung, Sin-Chun Ng. A BERT Framework to Sentiment Analysis of Tweets. Sensors, 2023.

Christine Dewi, Rung-Ching Chen, Henochi Juli Cristiano, Francesco Cauteruccio. Multinomial Naïve Bayes Classifier for Sentiment Analysis of Internet Movie Database. Vietnam Journal of Computer Science, 2023.

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

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

Haoran Li. Sentiment Analysis on Internet Movie Database (IMDb) Movie Review Dataset: Hyperparameters Tuning for Naïve Bayes Model. Department of Material Sciences and Engineering, 2023.

Harsh Sharma, Prakash Rokade, Reena Gunjan, Satyajit Pangaonkar. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language, 2018.

Like Hui, Mikhail Belkin. Evaluation Of Neural Architectures Trained with Square Loss vs Cross-Entrophy in Classification Taks. International Conference on Learning Representations, 2021.

Mine Çetinkaya-Rundel, Mine Dogucu. 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. Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets. Department of Computer science, University of Technology, Baghdad, Iraq, 2022.

Nassera Habbat, Hicham Nouri, Houda Anoun, Larbi Hassouni. Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning. Research Laboratory on New Economy and Development, 2023.

Rosie Dunfornd, Quanrong Su, Ekraj Tamang, Abigail Wintour. The Pareto Principle. The Plymouth Student Scientist, 2024.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, Yoav Artzi. Revisiting Few-Sample BERT Fine-Tuning.International Conference on Learning Representations, 2021.

Yankang Su, Zbigniew j. Kabala. Public perception of ChatGPT and Transformer Learning for Tweets Sentiment Analysis Using Wolfram Mathematica. MDPI Journal, 2023.

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