Analise de regressores na tarefa de estimar o sentimento em um conjunto de dados de letras musicais

Rafael A. S. Silva[1] **Carlos H. V. Moraes**[2] **Melise M. V. Paula**[3]

[1]Instituto de Matemática e Computação (IMC) – Universidade Federal de Itajubá (UNIFEI)
Caixa Postal 50 – 37500 903 – Itajubá – MG – Brasil

Abstract. *The Sentiment Analysis Technique is increasingly being used for analysis of textual content present mainly in social networks, this technique is derived from artificial intelligence and usually estimates a positive, negative or neutral value for the analyzed sentences through an automatic processing of natural language. In this article, the study target for sentiment analysis is musical lyrics. The letters used belong to the data set Music Dataset: Lyrics and Metadata from 1950 to 2019, they were introduced to the library Vader in order to obtain the feeling value and then, MLP, Decision Tree, Random Forest and KNN were presented to the regressors so that their performance to perform the task of obtaining the feeling was evaluated. The results lead to the conclusion that regressors are able to assess sentiment, attesting to its applicability and contribution to the advancement of research in this domain*

Resumo. *A Técnica de analise de sentimento cada vez mais está sendo utilizada para análises de conteúdos textuais presentes principalmente em redes sociais, essa técnica é derivada da inteligência artificial e costuma estimar um valor positivo, negativo ou neutro para as sentenças analisadas através de um processamento automático de linguagem natural. Neste artigo o alvo de estudo para analise de sentimento são letras musicais. As letras utilizadas pertencem ao conjunto de dados Music Dataset: Lyrics and Metadata from 1950 to 2019, elas foram apresentadas a biblioteca Vader para que se obtesse o valor do sentimento e em seguida, foram apresentadas aos regressores MLP, Árvore de Decisão, Floresta Aleatória e KNN para que fosse avaliado o seu desempenho para a realização da tarefa de obtenção do sentimento. Os resultados levam a concluir que regressores são capazes de avaliar o sentimento, atestando sua aplicabilidade e contribuição para o avanço de pesquisas neste domínio.*

1. Introdução

O Processamento de Linguagem Natural (PLN) consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em alguma líingua natural (e.g. tradução e interpretação de textos, busca de informações em documentos e interface homem-máquina) [do Lago Pereira 2019]

A relevância deste trabalho se da pelo fato de que, do ponto de vista linguístico, observar os aspectos morfológicos, sintáticos e semânticos em letras musicais nos ajudam a identificar visões de mundo carregadas de emoção e beleza. Percebe-se, portanto, que o sentimento é um elemento fundamental para os estudos sobre o funcionamento linguístico das letras de música [Souza and Café 2018]

Deste modo, o objetivo desta pesquisa é analisar a capacidade dos regressores Multilayer perceptron(MLP), Árvore de decisão, K-Nearest Neighbors(KNN) e Floresta Aleatória de estimarem os valores das polaridades dos sentimentos vinculados a letras musicais presentes na base de dados 'Music Dataset: Lyrics and Metadata from 1950 to 2019', fazer um comparativo e por fim indicar qual regressor é o mais recomendado.

2. Revisão da literatura

Na literatura existem diversos estudos com o propósito de analisar sentimentos e emoções através de dados textuais. No trabalho proposto por [Kumar and Benitta 2022] ele busca analisar o sentimento extraidos de comentários sobre filmes presentes na base de dados IMDB, utilizando técnicas de Machine Learning e Deep Learning. A parte principal do trabalho é uma combinação de memória longa e curta (LSTM) e rede neural recorrente (RNN) usada para classificar os sentimentos com alto grau de precisão em um curto período.

O estudo de [Xiang et al. 2021] apresenta uma análise usando substituição léxica focada em parte de fala (POS) para aumento de dados (PLSDA), com o objetivo de melhorar o desempenho de algoritmos de aprendizado de máquina na análise de sentimentos.

Em outro trabalho que também utiliza de avaliações da base de dados IMDB realizado por [Chatterjee et al. 2021], é anotado as avaliações em uma das três classes: positivas, negativas e neutras em um conjunto de dados chamado de JUMRv1. Para a avaliação do JUMRv1, é adotado uma abordagem exaustiva testando várias combinações de incorporação de palavras, métodos de seleção de recursos e classificadores. Também é analisado as tendências de desempenho, se houver, eles tentam explicá-las.

Por fim no estudo onde o autor [Shaukat et al. 2020] também utiliza a base de dados IMDB, a tarefa de mineração de opinião a partir de resenhas de filmes foi alcançada com o uso de redes neurais treinadas no "Movie Review Database" emitido por Stanford, em conjunto com duas grandes listas de palavras positivas e negativas. A rede treinada conseguiu atingir uma precisão final de 91%.

3. Metodologia

A metodologia a ser utilizada neste trabalho é baseada na metodologia de outros dois trabalhos. O primeiro é o Metodologias para Análise de Sentimentos de Tweets sobre o Mercado Financeiro [Medeiros 2019]. A metodologia apresentada neste TCC é descrita da seguinte forma: primeiramente, os tweets são carregados de um arquivo em disco, e são pré-processados, gerando-se um espaço vetorial. Em seguida, ocorre a extração de tópicos e a redução de dimensionalidade em paralelo. O resultado da redução de dimensionalidade é utilizado para classificar os sentimentos, para agrupar os tweets e para análise visual, a fim de tornar a investigação de aspectos em comum nos dados mais objetiva.

o segundo é o Estudo e avaliação de métodos de análise de sentimentos baseada em aspectos para textos opinativos em português [Machado 2018]. A metodologia utilizada nesta tese é dividida em duas etapas, na primeira é identificada as caracteristicas do produto avaliado mencionados no texto, na segunda o foco é identificar a polaridade (positiva ou negativa) relacionado a cada aspecto do produto.

A princípio, neste trabalho é coletada a base de dados, realizado o pré-processamento da mesma, realizada a mineração do sentimento utilizando a biblioteca Vader. Esta biblioteca gera os valores dos sentimentos positivo, negativo ou neutro para cada musica no conjunto de dados, e no fim é utilizado regressores que consigam atingir valores próximos aos valores gerados pela biblioteca Vader.

3.1. Base de Dados

A base de dados utilizada foi a *Music Dataset: Lyrics and Metadata from 1950 to 2019*[1], criada por Moura et al. (2020). Este conjunto de dados fornece uma lista de letras músicais do ano de 1950 a 2019 descrevendo metadados de música como tristeza, dança, volume, acústica, etc. Os autores também fornecem algumas informações como letras que podem ser usadas para processamento de linguagem natural.

Para este estudo a principal informação do conjunto são as letras de cada música, a partir delas que a biblioteca vader gera os sentimentos positivo, negativo e neutro, e os valores dos sentimentos gerados são os tutores dos regressores que estão sendo utilizados.

3.2. Pré-processamento

Nesta etapa do trabalho foi removida da base de dados original 27 colunas que não agregavam ao trabalho, sendo elas: *release_date, len, dating, violence, world/life, night/time, shake the audience, family/gospel, romantic, communication, obscene, music, movement/places, light/visual perceptions, family/spiritual, like/girls, sadness, feelings, danceability, loudness, acousticness, instrumentalness, valence, energy, topic, age e artist_name.*

Após a remoção dessas informações, seria aplicado a remoção das *stop-words* presentes nas letras musicais, porém os autores ao construirem o conjunto de dados já haviam realizado essa etapa, portanto ela não foi necessária.

3.3. Vader

A biblioteca *VADER*[2] (*Valence Aware Dictionary and sEntiment Reasoner*) é uma ferramenta de análise de sentimentos de código aberto baseada em regras léxicas que estão especificamente vinculadas com os sentimentos expressos nas mídias sociais.

Neste trabalho esta biblioteca foi utilizada para analisar as letras das músicas, e à partir delas é gerado os valores de sentimento positivo, negativo ou neutro. A medida que os sentimentos eram calculados os valores referentes a cada música eram adicionados ao conjunto de dados.

3.4. Regressores

Neste estudo foram utilizados os Regressores *MLP(Multilayer Perceptron)*, Árvore de Decisão, Floresta Aleatória e *KNN(K-Nearest Neighbors)*, empregando as implementações disponibilizadas pela biblioteca de Machine Larning do Python, chamada *scikit-learn*[3].

A regressão com o *Multi Layer Perceptron* foi implementada através da classe *MLPRegressor*. O algoritmo de Árvore de decisão através da classe *DecisionTreeRegressor*.

[1]https://www.kaggle.com/datasets/saurabhshahane/music-dataset-1950-to-2019
[2]https://github.com/cjhutto/vaderSentiment
[3]https://scikit-learn.org/stable/

Para o algoritmo de Floresta Aleatória, foi utilizada a classe *RandomForestRegressor*. Por fim, o algoritmo de *K-Nearest Neighbors* através da classe *KNeighborsRegressor*.

As regressões foram realizadas em duas etapas, na primeira foi utilizado os parâmetros padrões das bibliotecas de cada regressor, na segunda foi utilizado variação de parâmetros, e com isso é observado uma melhora nos resultados das métricas.

3.5. Métricas de validação

Para implementação das métricas e da validação cruzada, a biblioteca *scikit-learn* também foi empregada. As métricas de validação MSE, MAE, e RMSE foram implementadas respectivamente através dos métodos *mean_squared_error, mean_absolute_error* e *mean_squared_error* passando o parâmetro *squared* falso, esses métodos são fornecidos através do módulo *sklearn.metrics*. Também foi utilizado a métrica Score(R^2) que é fornecida pelas própias classes dos regressores.

Para as validações cruzadas foi utilizado o método *cross_validate*, do módulo *sklearn.model_selection* utilizando uma validação de 5 pastas.

4. Resultados e análises

Este Capítulo apresenta os resultados obtidos, para cada regressor são analisadas as suas métricas de avaliação. Na Seção 4.1 são descritos e analisados os resultados dos regressores sem variar nenhum parâmetro e nem utilizar validação cruzada. A Seção 4.2 apresenta os resultados dos regressores com os parâmetros padrões e utilizando uma validação cruzada de 5 pastas. Por fim na seção 4.3 é analisado os regressores utilizando uma variação de parâmetros e validação cruzada também de 5 pastas.

4.1. Regressores sem variação de parâmetro e validação cruzada

Na tabela 1 são apresentadas os valores das métricas dos regressores, se observa que em todas as métricas o regressor MLP se destaca em relação aos outros regressores utilizados.

Tabela 1. Métricas dos regressores sem variação de parâmetro e validação cruzada

	MAE	MSE	RMSE	Score(R^2)
MLP	**0.05185**	**0.00513**	**0.07135**	**0.794**
Árvore de Decisão	0.10030	0.01935	0.13778	0.226
Floresta Aleatória	0.07237	0.00984	0.09851	0.613
KNN	0.10478	0.01819	0.13452	0.277

A métrica *'score'* é de 0.794, o valor desta métrica quão mais próximo de 1 melhor, pois ela calcula qual a porcentagem da variança que pôde ser prevista pelo modelo de regressão e, portanto, revela o quão "próximo" as medidas reais estão do nosso modelo.

A métrica MSE possui um valor de 0.00513, o valor desta métrica quão mais próximo de 0 melhor, pois ela pega a diferença entre o valor predito pelo modelo e o valor real, eleva o resultado ao quadrado e faz a mesma coisa com todos os outros pontos, somando e dividindo pelo número de elementos preditos, por elevar o erro ao quadrado ela penaliza predições muito distantes das reais.

A métrica RMSE possui um valor de 0.07135, o valor desta métrica quão mais próximo de 0 melhor, o RMSE entra como uma forma de melhorar a interpretabilidade da métrica MSE, fazendo a raiz quadrada do resultado do MSE.

A métrica MAE possui um valor de 0.05185, o valor desta métrica quão mais próximo de 0 melhor, o Erro Absoluto Médio consiste na média das distâncias entre valores preditos e reais. Diferentemente do MSE e do RMSE, essa métrica não "pune" tão severamente os *outliers* do modelo.

4.2. Regressores sem variação de parâmetro e com validação cruzada

Na tabela 2 são apresentadas os valores das métricas dos regressores após a validação cruzada de 5 pastas, se observa uma pequena piora nos resultados em todos os regressores, mas o MLP continua apresentando os melhores valores em suas métricas.

Tabela 2. Métricas dos regressores sem variação de parâmetro e com validação cruzada

	MAE	MSE	RMSE
MLP	**0.05214**	**0.00520**	**0.07165**
Árvore de Decisão	0.10169	0.01968	0.13878
Floresta Aleatória	-	-	-
KNN	0.10588	0.01882	0.13647

O regressor floresta aleatória não possui resultados pois não conseguiu finalizar os cálculos em menos de 5 horas, em seu funcionamento ele faz uma combinação de diversas árvores de decisão, é um regressor um pouco mais robusto porém exige um custo computacional maior, desta forma não se obteve o valor de suas métricas.

4.3. Regressores com variação de parâmetro e validação cruzada

Na tabela 2 são apresentadas os valores das métricas dos regressores após a validação cruzada de 5 pastas e com variação de parâmetros, e neste caso se observa um resultado melhor em relação aos analisados anteriormente.

Tabela 3. Métricas dos regressores com variação de parâmetro e validação cruzada

	MAE	MSE	RMSE
MLP	**0.04667**	**0.00441**	**0.06609**
Árvore de Decisão	0.10171	0.01809	0.13366
Floresta Aleatória	-	-	-
KNN	0.09280	0.01414	0.11860

O MLP continua apresentando os melhores valores em suas métricas, para esse regressor foram variados os seguintes parâmetros: *solver* entre *lbfgs, adam* e *sgd*, sendo *lbfgs* o melhor. Sua função de ativação foi variada entre *logistic* e *relu* sendo a *relu* a melhor, por fim sua taxa de aprendizado foi variada entre *constant* e *adaptative* onde a *constant* foi a melhor.

Para a árvore de decisão o único parâmetro variado foi a sua profundidade máxima, que variou entre *none*, 2, 3, 5, 8, 10, 20 e 30, sendo o melhor a de profundidade

20. O regressor de floresta aleatória também não apresentou resultados pelos mesmos motivos citados na seção 4.2

Para o regressor knn o único parâmetro variado foi o número de vizinhos, que variou entre 5, 10, 15, 20, 25, 30, 35 e 40, sendo o melhor o de 40.

5. Conclusão

Através deste estudo foi possível concluir que regressores podem ser utilizados na tarefa de análise de sentimento. entre os regressores analisados o mlp foi o que apresentou os melhores resultados em suas métricas, portanto é o mais adequado para analisar a base de dados utilizada neste estudo.

A variação de parâmetros dos regressores apresentou uma melhora nos resultados das métricas de maneira geral, porém o custo computacional para realizar essa variação também aumenta. Ao realizar uma análise sob essa perspectiva isso deve ser levado em consideração.

Como trabalhos futuros é relevante fazer uma análise do sentimento presente em músicas levando em consideração não apenas a sua letra, mas também sua melodia. Outro estudo a ser realizado é uma analise para verificar se o gênero musical também influencia no valor dos sentimentos das músicas, tentando encontrar um padrão.

Referências

Chatterjee, S., Chakrabarti, K., Garain, A., Schwenker, F., and Sarkar, R. (2021). Jumrv1: A sentiment analysis dataset for movie recommendation. *Applied Sciences*, 11(20):9381.

do Lago Pereira, S. (2019). Processamento de linguagem natural.

Kumar, R. and Benitta, A. (2022). Imdb movie reviews sentiment classification using deep learning. Technical report, EasyChair.

Machado, M. T. (2018). *Estudo e avaliação de métodos de análise de sentimentos baseada em aspectos para textos opinativos em português*. PhD thesis, Universidade de São Paulo.

Medeiros, M. C. (2019). Metodologias para análise de sentimentos de tweets sobre o mercado financeiro.

Shaukat, Z., Zulfiqar, A. A., Xiao, C., Azeem, M., and Mahmood, T. (2020). Sentiment analysis on imdb using lexicon and neural networks. *SN Applied Sciences*, 2(2):1–10.

Souza, R. R. and Café, L. M. A. (2018). Análise de sentimento aplicada ao estudo de letras de música. *Informação & Sociedade*, 28(3).

Xiang, R., Chersoni, E., Lu, Q., Huang, C.-R., Li, W., and Long, Y. (2021). Lexical data augmentation for sentiment analysis. *Journal of the Association for Information Science and Technology*, 72(11):1432–1447.