SOCIOECONOMIC ANALYTICS

Predicting sentiments in Spotify comments: A comparative analysis of Machine Learning Models

Filipe Augusto Felix de Queiroz [1] *
Igor Barbosa Negreiros [2]
Giovana de Souza [3]
Débora Cordeiro de Sousa [4]
Sílvio Fernando Alves Xavier Júnior [5]

[1 2 3 4 5] Departamento de Estatística, State University of Paraíba, Campina Grande, Brazil
Emails: [1] filipe.queiroz@aluno.uepb.edu.br; [2] igor.negreiros@aluno.uepb.edu.br; [3] giovana.souza@aluno.uepb.edu.br; [4] debora.cordeiro@servidor.uepb.edu.br; [5] silvio@servidor.uepb.edu.br.
*Corresponding Author

How to cite this paper: de Queiroz, F. A., Negreiros, I. B., de Souza, G., de Sousa, D. C., Xavier Júnior, S. F. A. (2024). Predicting sentiments in Spotify comments: A comparative analysis of Machine Learning Models. *Socioeconomic Analytics,* 2(1), 107-113. https://doi.org/10.51359/2965-4661.2024.265070

RESEARCH ARTICLE

Socioeconomic Analytics
https://periodicos.ufpe.br/revistas/SECAN/
ISSN Online: 2965-4661

Submitted on: 11.11.2024.
Accepted on: 15.12.2024.
Published on: 27.12.2024.

Abstract

Using data from user sentences on Spotify, this work explores through Natural Language Processing positive and negative sentiments in each comment. We compare different statistical modeling and Machine Learning techniques, identifying the ones with the greatest accuracy in predicting sentiments. As a result, the assessment supports most of the sentences presented with negative connotations. As for modeling, the Logistic Regression and Random Forest models resulted in better accuracy.

Keywords

Sentiment Analysis, Natural Language Processing, Machine Learning, Logistic Regression, Random Forest, Spotify

1. Introdução

No cenário atual, as opiniões acerca de serviços ou produtos são essenciais ao desempenho do mercado. Diariamente as pessoas relatam suas experiências com determinadas marcas ou empresas, sejam tais experiências boas ou desagradáveis, afinal quem antes de comprar um produto não pesquisa antes os comentários a respeito. Com o alto uso dos meios sociais, ocorreu também a naturalização dos ditos feedbacks, entretanto, os feedbacks retratam não só a satisfação das pessoas, mas também seus sentimentos em relação ao produto adquirido ou ao serviço prestado. Nesse sentido, é necessário entender qual o sentimento demonstrado pelos consumidores ao disseminaremsuas avaliações.

Diante disso, técnicas de Cambria & White (2014) abordam, em sua pesquisa, diversos avanços e desafios no campo do Processamento de Linguagem Natural, enfatizando as principais tendências e inovações que têm impulsionado o progresso dessa área. Eles destacam a relevância de combinar abordagens distintas, como as baseadas em regras e as estatísticas, para superar limitações passadas e alcançar avanços significativos na compreensão e processamento da linguagem natural. Serão aplicadas a fim de compreender a emoção expressa pelos usuários do Spotify em suas avaliações do aplicativo. Além disso, os modelos Regressão Logística, *Random Forest, Suport Vector Machine* (SVM) (Lorena & de Carvalho (2007)) e *Gradient Boosting* serão aplicados, para a identificação daquele que melhor prevê o sentimento dos usuários do aplicativo.

2. Revisão literária

O aprendizado de máquina, ou do inglês *Machine Learning* (ML) pode ser descrito como conjunto de técnicas estatísticas juntamente com computação, na qual por meio de algoritmos consegue identificar padrões para tomar decisões ou fazer previsões. O termo Machine Learning foi utilizado a primeira vez em 1969 por Arthur Samuel, um pioneirono campo de inteligência artificial. Ele usou o termo ao desenvolver um software que jogava damas e melhorava suas jogadas com o tempo, aprendendo a partir da experiência. Samuel definiu Machine Learning como o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados (Samuel (1959)).

2.1 Machine Learning

Novas tecnologias são normalmente projetadas para tornar o processo mais gerenciável, preciso, mais rápido ou menos dispendioso. Elas também nos permitem realizar ou desenvolver trabalhos anteriormente inatingíveis em determinadas circunstâncias. Um dos procedimentos científicos em rápida expansão para uso prático nos últimos anos tem sido a Inteligência Artificial (IA). ML é uma aplicação de IA na qual algoritmos treinados em dados produzem IA. A inteligência artificial e o aprendizado de máquina cresceram em popularidade na última década graças aos avanços significativos na tecnologia da computação. Esta popularidade resultou em avanços dramáticos na capacidade de reunir e analisar enormes quantidades de dados (Alkhatib et al. (2023))

No Aprendizado Supervisionado as máquinas são ensinadas a resolver problemas com a ajuda de humanos, que reúnem e identificam dados e depois os “alimentam” aos sistemas. Um computador recebe características de dados específicas para examinar para detectar padrões, classificar itens e avaliar se sua previsão está correta ou incorreta. No Aprendizado

Não-supervisionado os métodos estão principalmente preocupados com o agrupamento de dados não categorizados. Nesta categoria de aprendizagem, as máquinas aprendem a identificar padrões e tendências em dados não rotulados sem que ninguém seja supervisionado por humanos. Por fim, no Aprendizado por Reforço em um ambiente confinado estranho a eles, os modelos devem resolver um problema por meio de uma série de tentativas e erros. As máquinas são punidas por erros e recompensadas por boas tentativas, semelhante a um cenário em muitos jogos (Arulkumaran et al. (2017))

2.2 Processamento Natural de Linguagem (NLP)

O Processamento de Linguagem Natural (NLP) é uma área da ciência da computaçãoe da inteligência artificial que se dedica ao desenvolvimento de sistemas capazes de compreender, interpretar e gerar a linguagem humana. Essa área combina conhecimentos de linguística, ciência da computação, estatística e aprendizado de máquina para facilitar a interação entre humanos e máquinas (Silva & Souza (2014)).

NLP é usado em várias atividades relacionadas a linguagem, perguntas, questionamentos, classificação de várias maneiras e conversa entre pessoas. Abaixo na figura 1 tem-se exemplos práticos de aplicações de NLP.

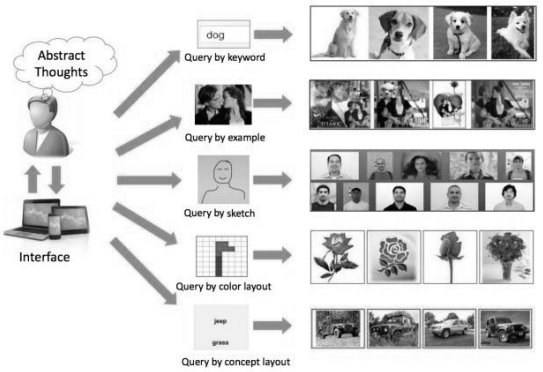


Figura 1 – Aplicações NLP. (adaptado de Piplani & Bamman (2018))

Modelos de NLP funcionam encontrando relações entre as partes constituintes da linguagem - por exemplo, as letras, palavras e frases encontradas em um conjunto de dados de texto. As arquiteturas de NLP usam vários métodos para pré-processamento de dados, extração de recursos e modelagem.

O primeiro passo para compreender o NLP é focar no significado do diálogo e do discurso em um contexto específico. O objetivo principal é facilitar conversas significativas entre um voicebot e um humano. Por exemplo, ao dar comandos a chatbots, como “mostre-me as melhores receitas'” ou “toque música para festa”', estamos dentro do escopo dessa etapa. Isso envolve entender e responder às solicitações do usuário dentro do contexto da conversa em andamento.

Antes que um modelo processe o texto para uma tarefa específica, é comum que o texto precise ser pré-processado para melhorar o desempenho do modelo ou para converter palavras e caracteres em um formato que o modelo possa entender. Diversas técnicas podem ser utilizadas nesse pré-processamento de dados.

O processo de dividir o texto em unidades menores chamadas tokens, que podem ser palavras, subpalavras ou caracteres. Sendo elas tokenização por palavras: consiste em dividir o texto em palavras individuais; tokenização por sentença: envolve a divisão do texto em sentenças individuais; tokenização por subpalavras: refere-se à divisão de palavras em unidades menores, como prefixos, sufixos ou até mesmo caracteres individuais.

O Bag-of-Words é uma técnica amplamente utilizada no processamento de linguagem natural para representar textos de forma numérica. Nesse modelo, um texto é representado pela frequência de ocorrência de suas palavras, desconsiderando a ordem e a estrutura gramatical. Conforme destacado por \cite{harris1954}, essa abordagem baseia-se na distribuição das palavras, onde o significado é inferido a partir dos contextos em que as palavras aparecem.

O Term Frequency-Inverse Document Frequency (TF-IDF) é uma medida estatística que avalia a importância de uma palavra em um documento específico, considerando sua frequência no documento e sua ocorrência em um conjunto de documentos (corpus). Harris (1954) explica que o TF-IDF combina duas métricas: Term Frequency (TF): mede a frequência de uma palavra em um documento. Quanto mais vezes a palavra aparece, maior é o TF; e Inverse Document Frequency (IDF): avalia a raridade de uma palavra no corpus. Palavras comuns em muitos documentos recebem um IDF baixo, enquanto palavras raras recebem um IDF alto. A multiplicação de TF e IDF resulta no TF-IDF, que destaca palavras frequentes em um documento, mas raras no corpus, indicando termos potencialmente mais informativos para a caracterização do conteúdo do documento.

Por fim, as etiquetas Part of Speach (POS) são rótulos atribuídos a cada palavra em um texto, indicando sua função gramatical, como substantivo, verbo, adjetivo, entre outros. A identificação dessas etiquetas é fundamental para a compreensão da estrutura sintática e semântica de uma sentença, auxiliando em tarefas como análise sintática, tradução automática e extração de informações. Salton & Buckley (1988) propôs um método de etiquetagem baseado em regras que demonstrou eficácia na atribuição de POS tags, contribuindo significativamente para o avanço das técnicas de processamento de linguagem natural.

3. Metodologia

Os dados utilizados referem-se a comentários de usuários do aplicativo Spotify, são frases e relatos do desempenho do aplicativo e da satisfação dos clientes. Nesse sentido, serão utilizados métodos estatísticos para o cálculo de informações básicas que incluem o número de sentenças negativas e positivas, também serão aplicadas técnicas de tratamento das sentenças, remoção das palavras de parada e limpeza das informações. Posteriormente serão criadas duas nuvens de palavra, uma para cada tipo de sentimento, e por fim será realizada a comparação de modelos. As modelagens utilizadas serão a Regressão Logística (Dayton (1992)), *Random Forest* (Biau & Scornet (2016)), SVM e *Gradient Boosting* (Natekin & Knoll (2013)). Todos esses métodos têm alta capacidade de previsão, e o critério de escolha utilizado será a acurácia.

4. Análise e discussões

Objetivou-se identificar os sentimentos transmitidos através dos comentários dos

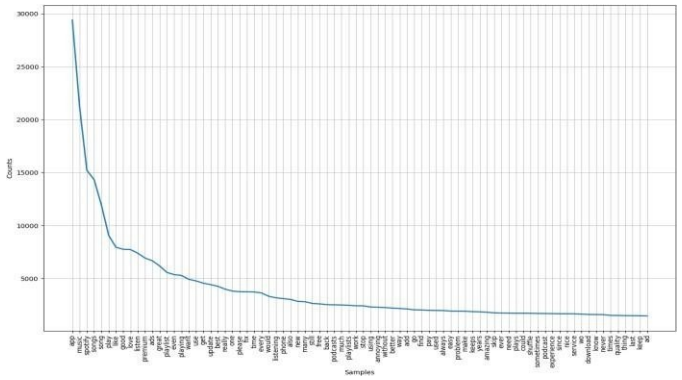
usuários do Spotify, dividindo-os entre sentimento positivo e negativo. Além disso,também foi de interesse verificar a modelagem com maior capacidade de previsão dos reviews. A tabela a seguir apresenta os principais resultados inerentes a análise.

Tabela 1 - Número de sentenças por sentimento.

Positivo	54046
Negativo	62469

Por meio da Tabela 1, é possível notar que a maioria das frases acerca do Spotify demonstram sentimento negativo, cerca de 54%. Enquanto 46% dos comentários possuem sentimento positivo.

Figura 2 - Frequência de palavras.



Através da Imagem 1, é perceptível que as palavras utilizadas com mais frequência pelos clientes do Spotify são positivas. No entanto, as palavras isoladamente não correspondem ao sentimento que toda a sentença apresenta.

Figura 3 – Nuvem de palavras para o sentimento positivo e negativo.

A Figura 2 exibe as principais palavras contidas nas sentenças positivas e negativas. Perceba que, na nuvem a esquerda estão contidas apenas palavras otimistas, corroborando com a noção de positividade das frases. Enquanto na nuvem a direita estão contidas palavras relacionadas ao sentimento negativo.

Tabela 2 - Acurácia dos modelos.

	Acurácia
Regressão Logística	0,8826
Random Forest	0,8633
SVM	0,8798
Gradient Boosting	0,8277

Por meio da Tabela 2, é possível identificar que todos os modelos aplicados obtiveram uma boa acurácia. Dentre eles, a Regressão Logística e o modelo SCV se destacam, apresentando os maiores valores de acurácia, sendo os modelos mais bem ajustados. Enquanto o Gradient Boosting resultou no menor valor de acurácia.

5. Conclusões

O método de Processamento Natural de Linguagem mostrou boa capacidade de identificação dos sentimentos das frases, por meio das nuvens de palavra foi possível verificar que as frases identificadas como positivas ou negativas eram compostas em sua maioria por palavras que refletiam tais emoções.

A utilização dos modelos de Machine Learning para previsão de sentimentos de reviews do Spotify demonstrou ótimo desempenho, juntamente com as técnicas de Processamento Natural de Linguagem. Todos os modelos apresentaram boa acurácia, com destaque para a Regressão Logística e o modelo SVM.

Referências

1. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine, 9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227
2. Lorena, A. C., & De Carvalho, A. C. P. L. F. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada, 14*(2), 43–67.
3. Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.33.0210
4. Alkhatib, R., Sahwan, W., Alkhatieb, A., & Schütt, B. (2023). A brief review of machine learning algorithms in forest fires science. *Applied Sciences, 13*(8275). https://doi.org/10.3390/app13148275
5. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240
6. Silva, E. M. da, & Souza, R. R. (2014). Fundamentos em processamento de linguagem natural: Uma proposta para extração de bigramas. *Encontros Bibli:Revista Eletrônica de Biblioteconomia e Ciência da Informação, 19*(40), 1–31.
7. Piplani, T., & Bamman, D. (2018). DeepSeek: Content-based image search & retrieval. *arXiv*. Disponível em https://arxiv.org/abs/1801.03406

8. Harris, Z. S. (1954). Distributional structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520
9. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
10. Dayton, C. M. (1992). Logistic regression analysis. *Stat, 474*, 574. Washington, DC.
11. Biau, Gérard, & Scornet, Erwan. (2016). A random forest guided tour. Test, 25, 197–227. Springer.
12. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, 21. Frontiers Media SA. https://doi.org/10.3389/fnbot.2013.00021