RBTI – Revista Brasileira em Tecnologia da Informação
Volume 06, número 01 – 2024
ISSN: 2675-1828
Editor Científico: Jaime Cazuhiro Ossada

Uma revisão sistemática da literatura sobre a utilização de algoritmos de *Machine Learning* para realização de análise de sentimentos

Stewart Evangelista Gonçalves
Instituto Federal de Alagoas (IFAL), stewartegoncalves@gmail.com

RESUMO

A análise de sentimentos é uma técnica que visa mensurar comentários e notícias por meio de inteligência artificial, é crescente a popularização da área de machine learning a fim da realização dessa análise. O objetivo deste trabalho é investigar e mapear qual é a linguagem mais utilizada no ramo de machine learning, bem como quais algoritmos de NLP são mais utilizados para aplicação da análise de sentimento, e também encontrar qual algoritmo possui maior acurácia nesta aplicação. Como metodologia, foi elaborada uma Revisão Sistemática de Literatura - RSL, onde foram analisados 22 artigos das fontes: Google Scholar e Scielo. Tais estudos foram selecionados com base em critérios de inclusão e exclusão e também via String de busca. Foi observado que a linguagem mais utilizada é Python, já os algoritmos mais utilizados são Naive Bayes, Máquinas de Vetores de Suporte (SVM), Floresta aleatória (Random Forest) e Regressão Logística. Dentre eles, o algoritmo que apresentou maior acurácia ficou sob contexto entre SVM e Random Forest.

Palavras-chave: Análise de Sentimentos; Machine Learning; NLP.

Data de Submissão: 09/10/2023

Data Aceito Publicação: 15/05/2024

A Systematic Literature Review on the Use of Machine Learning Algorithms for Sentiment Analysis

ABSTRACT

Sentiment analysis is a technique that aims to measure comments and news through artificial intelligence. The objective of this work is to investigate and map which is the most used language in the field of machine learning, as well as which NPL algorithms are most used for the application of sentiment analysis, and also to find which algorithm has the highest accuracy in this application. As a methodology, a Systematic Literature Review - RSL was elaborated, where 22 articles from the sources: Google Scholar and Scielo were analyzed. Such studies were selected based on inclusion and exclusion criteria and also via String search. It was observed that the most used language is Python, since the most used algorithms are Naive Bayes, Support Vector Machines (SVM), Random Forest and Logistic Regression. Among them, the algorithm that presented the highest accuracy was under context between SVM and Random Forest.

Keywords: Sentiment Analysis; Machine Learning; NLP.

1. Introdução

A análise de sentimentos é uma técnica que tem sido cada vez mais utilizada para extrair informações valiosas de dados não estruturados, como textos de mídias sociais, avaliações de produtos, entre outros. A aplicação de técnicas de análise de sentimentos pode permitir às empresas compreenderem melhor a opinião do público sobre seus produtos ou serviços, identificar problemas e oportunidades de melhoria, e desenvolver estratégias de marketing mais eficazes.

De acordo com Pang e Lee (2008), a análise de sentimentos é o processo de identificação e extração de informações subjetivas de fontes não estruturadas, tais como opiniões, sentimentos e emoções expressos em textos. Essas informações podem ser úteis para diversas áreas, incluindo negócios, política e saúde.

Recentemente o uso de técnicas de machine learning e processamento de linguagem natural (PLN) tornou-se popular para realizar análise de sentimentos em grandes conjuntos de dados de texto. Esta área se trata de problemas relacionados com a compressão e extração de informações a partir de documentos em textos escritos (Freitas, 2014). Algoritmos de aprendizado de máquina têm se mostrado eficazes na realização da análise de sentimentos, permitindo que as empresas possam analisar grandes volumes de dados de maneira eficiente e precisa.

Segundo Ethem Alpaydin (2010), no aprendizado de máquina utiliza-se um modelo definido com base em alguns parâmetros de dados. O aprendizado em si é a otimização dos parâmetros do modelo por meio de treinamento de dados ou de experiências passadas, podendo o modelo ser preditivo – prevendo o futuro de algum cenário –, descritivo – a fim de ganhar conhecimento com os dados –, ou ambos.

Kaur, Sehra S e Sehra K (2017) realizaram uma revisão sistemática de literatura sobre técnicas de análise de sentimentos em mídias sociais e encontraram que a abordagem de aprendizado de máquina é a mais utilizada na literatura, com destaque para algoritmos como Support Vector Machines (SVM), Redes Neurais Artificiais (ANN) e Árvores de Decisão.

Além disso, a análise de sentimentos tem sido amplamente aplicada em pesquisas de software, permitindo a identificação de problemas e tendências em diferentes fases do ciclo de vida do software. Mäntylä (2018) realizou um estudo empírico de análise de sentimentos em avaliações de usuários, bugs e pedidos de melhoria em projetos de software livre, identificando diferenças significativas nos sentimentos expressos em cada tipo de feedback.

Em suma, a análise de sentimentos tem se mostrado uma técnica valiosa para empresas e pesquisadores que desejam compreender melhor a opinião do público sobre produtos, serviços e outros assuntos. Com o uso de técnicas avançadas de aprendizado de máquina, é possível automatizar o processo de análise de sentimentos, tornando-o mais rápido e eficiente.

2. Metodologia

Para realização desta revisão sistemática de literatura, foram selecionados artigos que tivessem como foco algoritmos e técnicas utilizadas na análise de sentimentos em

relação a frases e textos. Esta pesquisa utiliza-se de caráter exploratório, pois busca-se um aprofundamento em trabalhos acadêmicos já publicados. Foram excluídos os artigos que não estavam em português ou inglês, que não estavam disponíveis na íntegra, ou que tinham pouca relevância para o tema da revisão.

A presente RSL foi dividida em três fases: Planejamento, condução e documentação, seguindo o protocolo proposto por Kitchenham e Charters (2007).

No período de planejamento, foi realizado o preenchimento do protocolo de RSL, assim, definindo alguns pontos como: objetivo, string de busca, fontes de dados, critérios de inclusão e exclusão. Foi elaborada também as questões de pesquisas, além de definir as fontes de dados que seriam utilizadas na pesquisa.

Na condução iniciou-se a pesquisa bibliográfica de fato, utilizando a string definida no passo acima e exercendo alguns filtros de melhorias, foi obtido fontes literárias preciosas para a relevância do trabalho em questão. Todo esse passo foi elaborado com o objetivo de responder às questões de pesquisas definidas.

Na fase de documentação, foram respondidas as questões pesquisadas, assim resultando no relatório final.

2.1. Planejamento

2.1.1. Questões de pesquisa

Seguindo o objetivo da RSL, temos a necessidade de encontrar estudos acadêmicos que possuam algoritmos de machine learning que realizem a análise de sentimentos em comentários ou notícias, a fim de criar uma base teórica para desenvolvimento de um estudo posterior. Visando o objetivo, foram definidas três questões de pesquisa:

Questão 1 (Q1) – Quais são as linguagens de programação mais frequentemente utilizadas na implementação de análise de sentimentos?

Questão 2 (Q2) – Quais são os algoritmos de Aprendizado de Máquina mais comumente utilizados na análise de sentimentos?

Questão 3 (Q3) – Dentre os algoritmos de machine learning mais utilizados no contexto de análise de sentimentos, quais apresentam melhores dados de acurácia em comparações entre si?

2.1.2. Fontes de dados

A fonte de dados escolhida primeiramente foi o *Google Scholar*, pela facilidade de utilização e por conter conteúdos de outras plataformas, após isso, a fim de realizar uma busca mais elaborada, foram utilizadas outras fontes, como no quadro a seguir:

Quadro 1. Fontes de dados selecionadas

Fonte de dados	Link
Google Scholar	https://scholar.google.com
Scielo	https://www.*Scielo*.org

2.1.3. Critérios de inclusão e exclusão

De acordo com Vom Brocke (2009), o processo de inclusão e exclusão deve ser transparente para que a revisão possua credibilidade, assim, temos no Quadro 2, uma série de critérios utilizados no refinamento da fonte, visando a facilidade de uma possível replicação por alguém que venha a se basear neste trabalho.

Quadro 2. Critérios de inclusão e exclusão

Critérios de inclusão	Critérios de exclusão
Tccs, monografias, artigos.	Não deixa explícito o algoritmo
Publicado entre 2018 e 2023	Não aborda análise de sentimentos
Informar algoritmo	Análise de sentimentos facial, não textual
Explicitar base de dados utilizada	Livros
Base pública	Indisponíveis e repetidos

2.2. Condução

2.2.1. Busca dos trabalhos

Para conseguir encontrar artigos relevantes para o trabalho proposto, fez-se necessária a criação de uma *String* de busca, onde a mesma passou por um processo de refinamento, procurando torná-la mais específica e condizente com o que queremos, e assim ficou:

(“Machine Learning”or “Deep Learning” or NLP or “aprendizagem de máquina") and ("análise de sentimento”) and (algorithms or algoritmo")

2.2.2. Estratégia de seleção

A pesquisa inicial resultou em 60 trabalhos provenientes de ambas as fontes utilizadas. Ao aplicarmos o primeiro critério de seleção de trabalhos publicados após 2018, o número de resultados reduziu para 42. Dessa forma, prosseguimos com o processo de filtragem, utilizando critérios de inclusão (CI) e critérios de exclusão (CE).

Após a aplicação dos CI e CE sobraram 22 estudos que se adequaram ao objetivo final desta RSL. A Tabela 1 mostra a quantidade de estudos selecionados por fonte de dados.

Tabela 1. Quantidade de trabalhos selecionados por fonte

Fonte de dados	Qtd de resultados	Qtd após seleção
Google Scholar	42	22
Scielo	0	0

A planilha detalhada com título dos trabalhos, autores, ano, base de dados, algoritmos utilizados e uma breve análise da conclusão pode ser acessada no seguinte link: https://tinyurl.com/ybtfab3y

3. Resultados

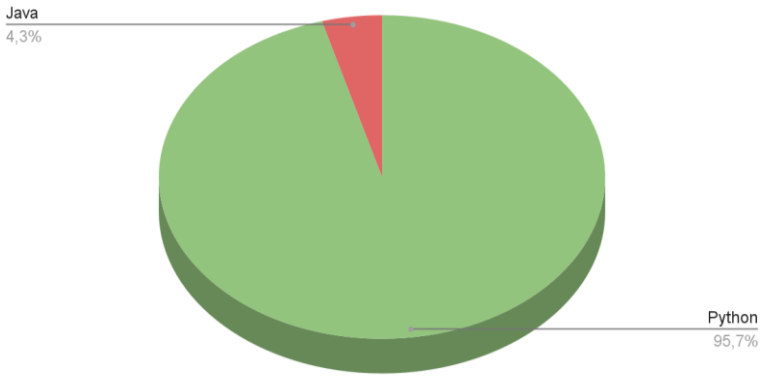
Com base na investigação e análise dos estudos realizados, foi possível responder às perguntas de pesquisa propostas e alcançar o objetivo da revisão sistemática da literatura. A seguir, cada questão será abordada e respondida com base nos dados obtidos a partir dos estudos selecionados.

Q1 – Quais são as linguagens de programação mais frequentemente utilizadas na implementação de análise de sentimentos?

No total, foram analisados 22 trabalhos com aplicações práticas de análise de sentimentos, um dos pré-requisitos da seleção, foi que dentro dos artigos tivessem exemplificações e trechos de códigos utilizados nos algoritmos, a fim de avaliar de fato a linguagem adotada, sem suposições.

O foco da questão é identificar qual linguagem está sendo mais utilizada no momento, para replicação em um trabalho futuro, por isso é tão importante que os artigos sejam novos e atualizados.

Figura 1. Linguagens de programação mais utilizadas



Presente em 95,7% dos trabalhos analisados como a linguagem mais utilizada para implementação da análise de sentimentos, o Python demonstra predominância para atividades envolvendo *Data science*, já que possui um uso intuitivo para computação quantitativa e analítica, além de contar com um grande banco de dados de bibliotecas voltadas para inteligência artificial e *machine learning.*

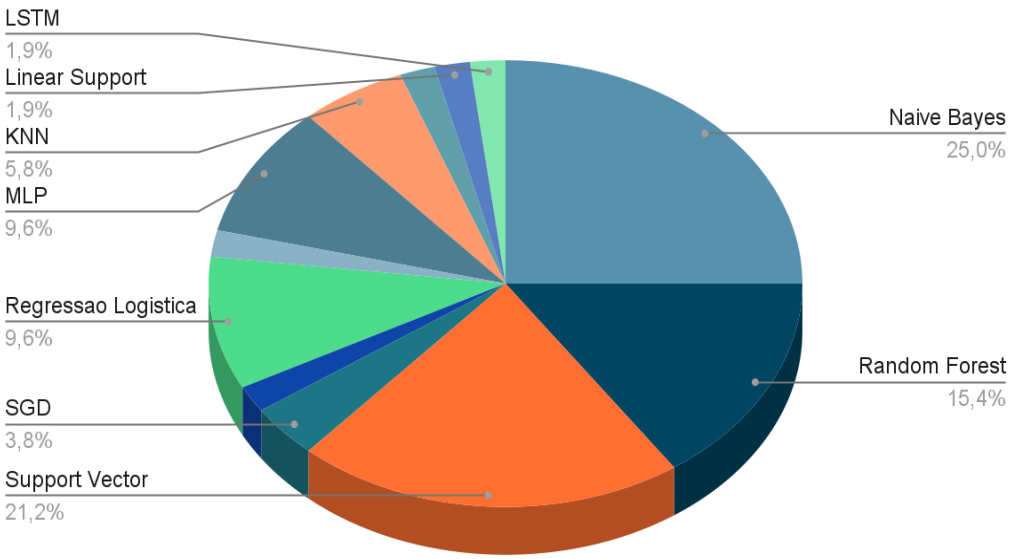
Quadro 3. Linguagens mais utilizadas na implementação de análise de sentimentos

Linguagem	Refs dos artigos
Python	[A01], [A02], [A03], [A04], [A05], [A06], [A07], [A08], [A10], [A11], [A12], [A13], [A14], [A15], [A16], [A17], [A18], [A19], [A20], [A21], [A22].
Java	[A09]

Q2 – Quais são os algoritmos de Aprendizado de Máquina mais comumente utilizados na análise de sentimentos?

Dos 22 artigos analisados, cada um apresentou sua particularidade e diversidade em técnicas, fomentando a pluralidade de respostas. Foram mapeados todos os algoritmos de classificação utilizados em todos os trabalhos, considerando que alguns trabalhos possuem mais de um algoritmo em seu corpo, em caso de existir no trabalho estudado algum tipo de comparação entre os modelos.

Figura 2. Algoritmos tradicionais mais utilizados



Conforme o gráfico acima, percebe-se que três algoritmos possuem maior predominância em utilizações, que são: *Naive Bayes, Support Vector Machine* (SVM) e *Random Forest*, respectivamente.

O *Naive Bayes* ganha pontos por possuir diversas aplicações e por sua simplicidade aliada aos bons resultados que gera. Já o SVM é eficaz na classificação de dados complexos e separáveis por hiperplanos, o que é muito útil quando há um grande número de atributos e a separação das classes é clara. Por sua vez, o *Random Forest* é baseado em conjunto de árvores de decisão. Ele cria várias árvores independentes e, em seguida, combina suas previsões para obter um resultado final.

Observa-se que cada técnica possui suas vantagens e desvantagens, e a escolha de qual técnica utilizar depende do objetivo da análise e dos dados disponíveis.

Quadro 4. Algoritmos tradicionais mais utilizados

Algoritmos	Refs dos artigos	Número de citações
Naive Bayes	[A01], [A02], [A04], [A05], [A06], [A07], [A08], [A09], [A10], [A15], [A18], [A20], [A21]	13
Random Forest	[A01], [A06], [A08], [A09], [A10], [A14], [A15], [A20]	8
Support Vector Machines (SVM)	[A02], [A03], [A06], [A08], [A09], [A10], [A12], [A15], [A18], [A19], [A20]	11
SGD	[A02], [A08]	2
XGBoost	[A02]	1
Regressão Logística	[A03], [A05], [A06], [A08], [A15]	5
LGBM	[A05]	1
MLP	[A05], [A08] [A11], [A16], [A20]	5
KNN	[A06], [A08], [A12]	3
Decision Tree	[A15]	1

Linear Support Vector Classification (LSVC)	[A21]	1
LSTM	[A22]	1

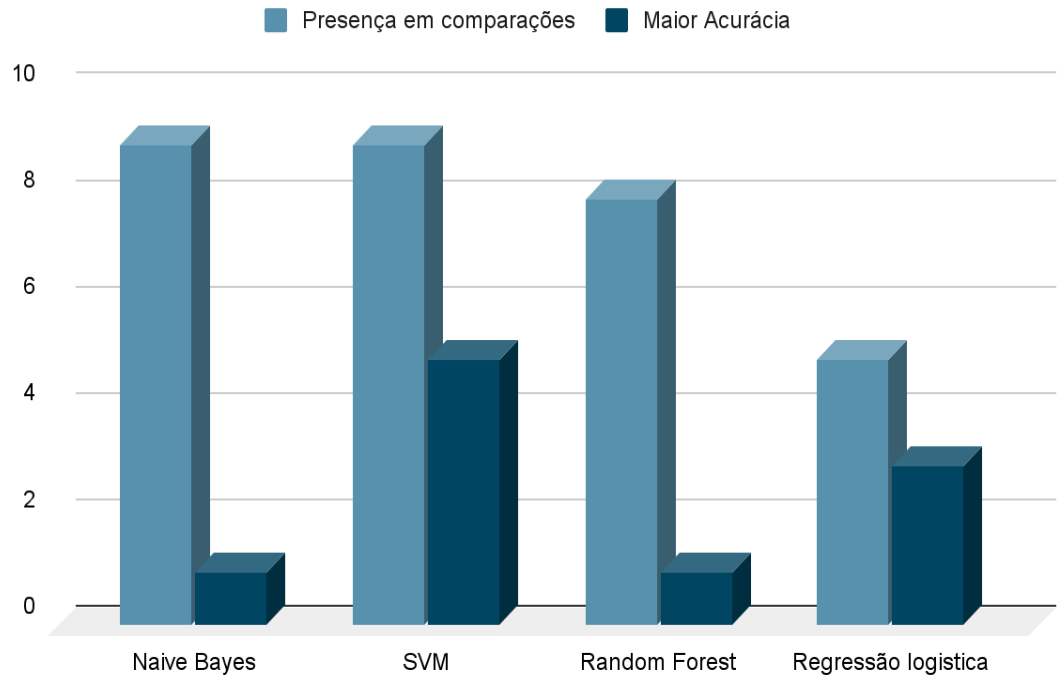
Q3 – Dentre os algoritmos de *machine learning* mais utilizados no contexto de análise de sentimentos, quais apresentam melhores dados de acurácia em comparações diretas?

Apesar de termos inúmeros algoritmos de *machine learning*, separamos os mais utilizados com base na questão anterior, e partimos para uma análise em busca da melhor acurácia.

Para essa questão foi adicionado mais um critério de inclusão: apenas trabalhos que exercem comparação de algoritmos. Com isso, dos 22 trabalhos utilizados em questões anteriores, restaram apenas 11.

Para encontrar essa resposta, utilizamos uma métrica de valor agregado sob a presença, ou seja, anota-se quantas vezes o algoritmo aparece em algum tipo de comparação, após isso, é necessário encontrar em quantos desses trabalhos o algoritmo foi tido como o de maior acurácia pelo pesquisador do estudo.

Figura 3. Algoritmos NLP com maior presença e sua acurácia



Nota-se no gráfico que o *Naive Bayes* aparece em 9 dos 11 artigos, mas apesar de sua enorme participação, ele foi considerado o de melhor acurácia em apenas 1 dos artigos em que esteve presente. Realizando uma divisão de **presença/maior acurácia**, ficamos com uma porcentagem de 11,11%.

Já o SVM, presente também em 9 artigos, consegue ser muito mais efetivo que o anterior, temos maior acurácia em 5, calculando sua porcentagem obtemos que em 55% das vezes em que ele está sendo comparado, o mesmo consegue ter maior acurácia que os outros métodos.

O *Random Forest* vem um pouco mais atrás, pois está presente em 8 artigos e é considerado o mais efetivo em apenas 1, obtendo a taxa de 12,5% após ter a presença em comparações de algoritmos dividida pela maior acurácia.

O algoritmo de Regressão Logística aparece como uma grande surpresa, apesar de ser pouco utilizado se comparado aos outros, ele apresenta um cálculo de 3 vitórias em 5 participações, ou seja, em 60% das vezes em que está envolvida em comparações, ela lidera com a melhor acurácia.

Quadro 5. Linguagens mais utilizadas na implementação de análise de sentimentos

Linguagem	Presente em	Número de citações	Maior acurácia
SVM	[A02], [A03], [A06], [A08], [A09], [A10], [A14], [A15], [A20]	9	[A02], [A06], [A09], [A10], [A14]
Regressão logística	[A03], [A05], [A06],[A08], [A15]	5	[A03], [A05], [A08]
Naive Bayes	[A01], [A02], [A05][A06], [A08], [A09][A10], [A15], [A20]	9	[A01]
Random Forest	[A01], [A06], [A08], [A09], [A10], [A14], [A15], [A20]	8	[A20]

4. Conclusão

Os resultados dos estudos analisados mostraram que o python é predominantemente utilizado quando se trata de linguagem de programação voltada para uso em machine learning, é justificável, visto que suas bibliotecas facilitam o trabalho dos desenvolvedores da área.

Com relação aos algoritmos de classificação mais utilizados para aplicação da análise de sentimentos, apesar da existência de inúmeros métodos, alguns acabam se destacando pela sua facilidade de implementação, outros por sua capacidade produtiva e outras qualidades, mas nota-se que há uma certa preferência por alguns como: Naive Bayes, SVM, Random Forest e Regressão logística.

Também foi visto que apesar da popularidade de alguns modelos, quando levamos em conta principalmente sua acurácia, alguns modelos podem não ser tão bons quanto parecem, já outros aparecem como uma grande surpresa, com números de acurácia muito bons.

Desta forma, essa revisão sistemática de literatura mostrou o que tem de mais atual no meio acadêmico na área de machine learning para elaboração de análise de sentimento utilizando algoritmos de machine learning.

Referências:

ALPAYDIN, E. **Introduction to Machine Learning**. MIT Press, 2010. Disponível em: <http://nuir.nkumbauniversity.ac.ug/xmlui/bitstream/handle/20.500.12383/1421/Introduction%20to%20Machine%20Learning,%20Second%20Edition%20(Adaptive%20Computation%20and%20Machine%20Learning)%20(%20PDFDrive%20).pdf?sequence=1>. Acesso em: 19/05/2023.

FREITAS, Cláudia. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, Rio de Janeiro, v. 13, n. 4, p. 1031-1059, dez. 2013. Disponível em: <https://doi.org/10.1590/S1984-63982013005000024>. Acesso em: 19/05/2023.

KAUR, J.; SEHRA, S. S.; SEHRA, S. K. A systematic literature review of sentiment analysis techniques. **International Journal of Computer Sciences and Engineering**, v. 5, n. 4, p. 22-28, 2017.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. In: **Technical report**, Ver. 2.3 EBSE TechnicalReport EBSE. [S.l.]: Epidemiol. Serv. Saúde, 2007.

MÄNTYLÄ, Mika; GRAZIOTIN, Daniel; KUUTILA, Miikka. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. **Computer Science Review**, v. 27, p. 16-32. fev. 2018. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1574013717300606>. Acesso em: 21/05/2023.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, v. 2, n. 1-2, p. 1-135, jul. 2008. Disponível em: <https://www.nowpublishers.com/article/Details/INR-011>. Acesso em: 19/05/2023.

VOM BROCKE, J. *et al*. Reconstructing the giant: On the importance of rigour in documenting the literature search process. *In*: Proceedings of the 17th European Conference on Information Systems, 2009, Verona. ***Anais eletrônicos*** [...] Verona, 2009, p. 1-13. Disponível em: <https://www.researchgate.net/publication/259440652_Reconstructing_the_Giant_On_the_Importance_of_Rigour_in_Documenting_the_Literature_Search_Process>. Acesso em: 25/05/2023.