Uma Aplicação Baseada em Árvores de Vocabulário para Mineração de Opinião

Guilherme Neri Bustamante Sá[1], **Daniel Oliveira de Freitas**[1], **Fábio Paulo Basso**[1]

[1]Universidade Federal do Pampa (UNIPAMPA) – Alegrete, RS, Brasil

{guilhermesa.aluno,danielodf2.aluno, fabiobasso}@unipampa.edu.br

Abstract. *This work proposes a hybrid algorithm for sentiment analysis, combining lexicon-based techniques with classical machine learning methods. The approach is based on a vocabulary tree structure. The model was evaluated using the Corpus Collection dataset, achieving an accuracy of 61%, a result close to literature benchmarks, despite the absence of parameter optimization. Initial results indicate potential for future improvements, such as exploring new lexicons, structural adjustments to the tree, and aspect-level sentiment analysis.*

Resumo. *Este trabalho propõe um algoritmo híbrido para análise de sentimentos, combinando técnicas baseadas em léxicos com métodos clássicos de aprendizado de máquina. A abordagem é fundamentada em uma estrutura de árvore de vocabulário. O modelo foi avaliado utilizando o conjunto de dados Corpus Collection, obtendo uma acurácia de 61%, valor próximo aos referenciais da literatura, ainda que sem otimizações de parâmetros. Os resultados iniciais indicam potencial para aprimoramentos futuros, como a exploração de novos léxicos, ajustes estruturais na árvore e análise em nível de aspecto.*

1. Introdução

Entre as diversas métricas disponíveis para avaliar a presença de uma marca, as *menções de marca* têm papel central. Elas representam ocasiões em que uma marca é citada em redes sociais, avaliações de clientes, notícias ou outros meios. Monitorar essas menções é essencial para entender a percepção pública e interagir de forma eficaz com o público, respondendo rapidamente a críticas e aproveitando elogios para fortalecer a imagem da marca.

Diante disso, este trabalho propõe um algoritmo de análise de opiniões que seja ao mesmo tempo adaptável e eficiente, mesmo com conjuntos de dados reduzidos. Para tal, utiliza-se uma abordagem baseada em árvores de vocabulário, com extração e indexação de descritores a partir de léxicos.

2. Sistema Proposto

Neste trabalho, propomos um sistema para análise de sentimentos baseado em *árvores de vocabulário* em conjunto com léxicos. Essa abordagem combina a interpretabilidade de técnicas lexicais com a escalabilidade e robustez de modelos de aprendizado de máquina.

O algoritmo proposto é composto por duas etapas principais: **ajuste** e **predição**. Na etapa de ajuste, os descritores são extraídos de dicionários léxicos e indexados em

nossa árvore de vocabulários. Já a etapa de predição concentra-se na identificação de padrões ou sentimentos semelhantes, propagando os descritores obtidos através do texto de entrada pela estrutura, calculando a pontuação de similaridade e ranqueando as opiniões ou textos de acordo com os resultados obtidos.

A etapa inicial de ambas as fases recebe um vetor de palavras como entrada. Como nossa abordagem se fundamenta em algoritmos de agrupamento, como o k-means [MacQueen et al. 1967], que requerem representações numéricas, é necessário não apenas tokenizar as palavras, mas também convertê-las em vetores de embeddings. Para essa tarefa, utilizamos a biblioteca *SentenceTransformers*[1], uma ferramenta desenvolvida especificamente para facilitar o acesso, uso e treinamento de embeddings semânticos aplicados a textos.

As palavras da etapa de ajuste são obtidas a partir do léxico AFINN [Årup Nielsen 2011], um dicionário amplamente utilizado em tarefas de análise de sentimentos por atribuir valores de polaridade às palavras com base em sua carga emocional. O arquivo do AFINN é estruturado de forma que a primeira coluna contém os termos e a segunda apresenta valores numéricos inteiros que indicam a polaridade (negativa ou positiva) de cada palavra. Essa representação quantitativa é utilizada na etapa posterior para o cálculo das pontuações na árvore de vocabulário.

A construção da árvore de vocabulário começa com a propagação do vetor a partir do nó raiz. Em cada nível, os descritores (no nosso caso, vetores de embeddings) são agrupados em k clusters obtidos pelo algoritmo k-means, formando k subárvores . Esse processo recursivo continua até atingir uma profundidade máxima l ou até que o número de elementos em um nó seja menor ou igual a k. Assim, o parâmetro k define o número de filhos por nó, enquanto l define a profundidade da árvore.

Após a construção recursiva da árvore, realizamos o mapeamento dos *embeddings* contidos nas folhas em relação às palavras do léxico, a fim de capturar os respectivos valores de polaridade. Com isso, a pontuação total de cada folha é calculado como a soma das pontuações dos *embeddings* nela contidos.

Para a etapa de predição, um texto qualquer é utilizado como entrada. Esse texto é segmentado em um vetor de palavras, e em seguida, as posições que contêm termos irrelevantes são removidas por meio de um filtro de *stop words*. A partir desse ponto, o processamento segue de forma semelhante à fase de ajuste, até alcançar a estrutura da árvore de vocabulário. Nela, cada palavra transformada em um vetor de embeddings é passada individualmente para o nodo raiz da árvore. A partir do nodo raiz, utilizamos o modelo k-means treinado para predizer qual subárvore a palavra deve percorrer até chegar em uma folha.

Cada vez que uma palavra alcança uma folha, considera-se que essa folha foi ativada uma vez. Assim, o valor total associado ao texto é calculado como uma soma ponderada das pontuações das folhas, de acordo com a frequência de ativação de cada uma. A equação que representa essa soma é dada por:

[1] `https://sbert.net/`

$$\text{Score}_{\text{texto}} = \sum_{i=1}^{n} s_i \cdot x_i \tag{1}$$

onde:

- s_i é a pontuação associada à folha i,
- x_i é o número de vezes que a folha i foi ativada,
- n é o número total de folhas da árvore.

A pontuação resultante é usada para classificar a entrada em relação às opiniões ou textos indexados anteriormente. Textos com sentimentos semelhantes terão pontuações mais próximas, facilitando o agrupamento ou a identificação de padrões emocionais em grandes conjuntos de dados. Essa abordagem não apenas identifica a polaridade emocional do texto de entrada, mas também permite a análise de similaridade com outras entradas.

Por causa da grande dependência do algoritmo utilizado para agrupar os *embeddings* e do dicionário léxico utilizado como entrada, a precisão dos resultados depende diretamente da qualidade da árvore construída durante a fase de ajuste e da representatividade dos léxicos utilizados.

3. Resultados

Para uma validação inicial da viabilidade do algoritmo proposto, buscamos comparar nossa abordagem com técnicas que disponibilizassem metodologias detalhadas e conjuntos de dados acessíveis publicamente. Essa busca se deu através de um estudo sistemático por trabalhos focados na mineração de opinião sobre produtos ou empresas. Optamos por utilizar o conjunto de dados *Corpus Collection*[Taboada et al. 2006], que foi utilizado também nos trabalhos [Taboada et al. 2011] e [Vilares et al. 2017].

Essa escolha se deu pela transparência em relação a metodologia e aos resultados dos trabalhos e pela configuração simples e robusta do dataset. Nele, as avaliações são rotuladas como ”recomendado”ou ”não recomendado”, correspondendo, respectivamente, a sentimentos positivos e negativos. O conjunto inclui oito categorias de produtos distintas, cada uma composta por 25 avaliações positivas e 25 negativas, totalizando 400 entradas.

A métrica adotada para avaliação do desempenho na análise de sentimentos foi a acurácia, por ser utilizada nos outros dois estudos, permitindo a comparação dos resultados. O estudo de [Vilares et al. 2017] relatou uma precisão de 65% usando o conjunto de dados Corpus Collection, um resultado bastante alinhado com o desempenho de [Taboada et al. 2011], que alcançou 65,5%. Esses benchmarks fornecem uma linha de base confiável para comparação com nosso algoritmo proposto.

Na avaliação realizada, o algoritmo proposto obteve uma acurácia de 61%, conforme apresentado na Tabela 1. Embora esse resultado esteja abaixo dos valores de referência, destaca-se o potencial de aprimoramento da abordagem. Não foram realizados testes para identificar o número ideal de clusters no léxico nem houve otimização sistemática da profundidade da árvore de vocabulário, fatores que podem impactar diretamente o desempenho.

Tabela 1. Accuracy reported by the studies when using the dataset provided by [Taboada et al. 2006]

Studys	Accuracy
Our study	61%
S08 [Vilares et al. 2017]	65%
SO-CAL [Taboada et al. 2011]	65.5%

Essas limitações indicam a necessidade de refinamento, mas também evidenciam o potencial da proposta como base para métodos mais avançados de análise de sentimentos e extração de opiniões.

4. Conclusão

Observamos que o uso de técnicas híbridas, que combinam léxicos com algoritmos clássicos de aprendizado de máquina, demonstra elevado potencial na área de análise de sentimentos. O algoritmo proposto ainda apresenta amplo espaço para aprimoramentos, especialmente devido à sua forte dependência do léxico adotado, do algoritmo de agrupamento e da função de agregação utilizada, neste caso, uma soma simples.

Trata-se de um trabalho em desenvolvimento, com diversas possibilidades de extensões futuras. Entre as direções promissoras, destacam-se: a otimização dos parâmetros estruturais da árvore (como número de nós e profundidade), a experimentação com novos léxicos, a adoção de funções de agregação mais robustas e a transição da análise para uma granularidade em nível de aspecto. Essas possibilidades serão exploradas em experimentos futuros por meio de testes de hipóteses aplicados ao modelo proposto.

Referências

[MacQueen et al. 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Taboada et al. 2006] Taboada, M., Anthony, C., and Voll, K. D. (2006). Methods for creating semantic orientation dictionaries. In *LREC*, pages 427–432.

[Taboada et al. 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

[Vilares et al. 2017] Vilares, D., Gómez-Rodríguez, C., and Alonso, M. A. (2017). Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

[Årup Nielsen 2011] Årup Nielsen, F. (2011). Afinn: A new word list for sentiment analysis. `https://github.com/fnielsen/afinn`. Accessed: August 2025.