tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 30/05/2024. Aprovado 21/03/2025
Avaliação: revisão duplo-anônimo

Expansão automática de léxico para Análise de Sentimentos de textos no domínio do Mercado Financeiro Brasileiro

AUTOMATIC LEXICON EXPANSION FOR SENTIMENT ANALYSIS: OF TEXTS IN THE BRAZILIAN FINANCIAL MARKET DOMAIN

AMPLIACIÓN AUTOMÁTICA DEL LÉXICO PARA EL ANÁLISIS DE SENTIMIENT O DE TEXTOS EN EL DOMÍNIO DEL MERCADO FINANCIERO BRASILEÑO

Thiago Monteles de Sousa
Universidade Federal de Goiás (UFG)
thiagomonteles@discente.ufg.br

Deborah Silva Alves Fernandes
Universidade Federal de Goiás (UFG)
deborah@inf.ufg.br

Kéthlyn Campos Silva
Universidade Federal de Goiás (UFG)
kethlyncampos@discente.ufg.br

Márcio Giovane C. Fernandes
Universidade Estadual de Goiás (UEG)
marcio.giovane@ueg.br

Resumo

Este artigo explora a geração de léxicos especializados para o Mercado Financeiro Brasileiro (MFB), adotando uma abordagem híbrida que combina a criação de um léxico em português com a análise de sentimentos em *tweets* e notícias do MFB. A metodologia consiste em uma série de etapas que expandem um léxico semente por meio de técnicas como Word2Vec, sinônimos/antônimos e Pointwise Mutual Information (PMI). Os resultados demonstram que a abordagem lexical alcançou um *F1-Score* de 71,5% na classificação de *tweets* e 68,4% em notícias, enquanto a combinação do léxico com o modelo de aprendizagem de máquina support vector machine (SVM) resultou em um *F1-Score* de 80% para *tweets*. Além disso, o estudo destaca a eficácia da lematização no pré-processamento para melhorar a precisão e cobertura do léxico como também a oportunidade da abordagem demonstrada na criação de léxicos específicos.

Palavras-chave: expansão lexical; mercado financeiro brasileiro; processamento de linguagem natural; redes sociais.

Abstract

This article investigates the generation of specialized lexicons for the Brazilian Financial Market (MFB) through a hybrid approach that integrates the construction of a Portuguese lexicon with sentiment analysis

applied to *tweets* and financial news. The proposed methodology involves a sequence of steps aimed at expanding a seed lexicon by employing techniques such as Word2Vec, synonym and antonym extraction, and Pointwise Mutual Information (PMI). Experimental results indicate that the lexical approach achieved an F1-score of 71.5% in tweet classification and 68.4% in news classification. Furthermore, when combined with the Support Vector Machine (SVM) learning model, the lexicon attained an F1-score of 80% for tweets. The study also underscores the effectiveness of lemmatization in preprocessing, both for improving the accuracy and expanding the coverage of the lexicon, as well as the potential of the proposed methodology for developing domain-specific lexical resources.

Keywords: lexical expansion; brazilian financial market; natural language processing; social networks.

Resumen

Este artículo explora la generación de léxicos especializados para el Mercado Financiero Brasileño (MFB), adoptando un enfoque híbrido que combina la creación de un léxico en Portugués con el análisis de sentimientos en *tweets* y noticias del MFB. La metodología consiste en una serie de pasos que expanden um léxico semilla mediante técnicas como Word2Vec, sinónimos/antónimos y la Información Mutua Puntual (PMI). Los resultados demuestran que el enfoque léxico alcanzó un f1-score de 71.5 % en la clasificación de *tweets* y 68.4% en noticias, mientras que la combinación del léxico con el modelo de aprendizaje automático máquina de vectores de soporte (SVM) resultó en un f1-score de 80% para *tweets*. Además, el estudio destaca la eficacia de la lematización en el preprocesamiento para mejorar la precisión y cobertura del léxico, así como la oportunidad de la aproximación demostrada en la creación de léxicos específicos.

Palabras clave: expansión léxica; mercado financiero brasileño; procesamiento de lenguaje natural; redes sociales.

Formatações gerais

Com a popularização das plataformas de redes sociais online, como o Twitter (conhecido como X a partir de 2023), Facebook, LinkedIn e outras, milhares de usuários têm interagido com postagens e mensagens que abordam uma grande variedade de tópicos. Essas plataformas tornaram-se espaços onde os usuários expressam suas opiniões cada vez mais e também as utilizam como instrumento para a tomada de decisões (Bos; Frasincar, 2022). O Twitter, em particular, é uma das redes sociais mais populares no mundo. A rede permitia que cada usuário publicasse mensagens chamadas *tweets*, com limite de 4 mil caracteres para a versão paga e 280 para a gratuita. Esses *tweets* são visualizados por outros usuários por meio do compartilhamento de publicações (*retweets*) e interações, tornando-se uma fonte relevante para acompanhar tendências e opiniões (Carosia; Coelho; Silva, 2020).

No contexto do mercado financeiro, o Twitter é uma plataforma amplamente utilizada por investidores para expressar suas opiniões em razão da simplicidade e influência midiática nas dinâmicas de preços das ações. No entanto, dada a enorme quantidade de publicações, análises manuais se tornam inviáveis. Nesse cenário, a Análise de Sentimentos (AS), uma abordagem de Processamento de Linguagem Natural (PLN), é empregada para extrair indicadores automáticos das opiniões. A AS divide as tarefas em identificação da polaridade (positiva ou negativa) e da emoção associada, como felicidade ou tristeza (Pereira, 2021). Existem duas abordagens principais: Aprendizagem de Máquina (AM), que oferece resultados promissores, mas requer grande quantidade de dados rotulados, tornando o processo trabalhoso e custoso; e a abordagem lexical, que

se baseia na Orientação Semântica das palavras nos textos, proporcionando facilidade de construção, seja de forma automática, seja por meio de textos relacionados ao mercado financeiro (Mahmood *et al.*, 2020).

Dessa forma, este trabalho tem como objetivo abordar as possibilidades de geração de vocabulários especializados, examinando uma perspectiva híbrida para criar um léxico em Português voltado ao domínio do Mercado Financeiro Brasileiro (MFB). O intuito é identificar palavras que indiquem graus de otimismo ou pessimismo em textos relacionados ao campo-lavo, contribuindo para a aplicação de PLN nesse contexto da língua portuguesa, área que ainda apresenta escassez de estudos publicados (Januário *et al.*, 2021; Pereira, 2021). Assim, este trabalho visa contribuir para o avanço da área de PLN e fornecer recursos para a criação de léxicos com domínios específicos. Nesse contexto, será elaborada uma estratégia para validar os vocabulários obtidos em tarefas de AS no âmbito do MFB.

As principais contribuições deste trabalho estão resumidas a seguir.

- Elaborar diferentes configurações para a geração de léxicos do domínio-alvo, resultando na criação de léxicos do campo especializado.
- Testar o desempenho dos léxicos por meio da análise de sentimentos em *tweets* e notícias no campo do Mercado Financeiro Brasileiro.
- Comparar o desempenho entre abordagem lexical, aprendizagem de máquina supervisionado e uma proposta que mescle as duas abordagens na tarefa de classificação de sentimentos.

Trabalhos relacionados

A abordagem lexical é um recurso presente em várias atividades de processamento de linguagem natural, como análise de sentimentos, classificação de textos, recuperação de opinião, identificação de temas, entre outras. Quando elaborados de forma adequada, os léxicos podem fornecer uma boa capacidade de classificação, além de poderem ser utilizados como recursos adicionais aos modelos de aprendizagem de máquina (Oliveira; Cortez; Areal, 2016). Detectar subjetividades em sentenças e classificá-las em uma classe é um desafio, especialmente em domínios específicos, como o mercado de ações (Das *et al.*, 2022), doenças (Jung *et al.*, 2021), documentos jurídicos (Smywinski-Pohl *et al.*, 2019) e outros que exigem corpora especializado.

A construção de um dicionário de léxicos pode seguir diferentes abordagens. Uma delas é totalmente manual, como em Loughran e McDonald (2011), que apresenta uma popular coleção de palavras rotuladas para o domínio do mercado financeiro. Para isso, foram utilizados documentos de textos extraídos do portal U.S *Securities and Exchange Commission* entre 1994 e 2008, resultando em seis grupos de palavras. Outra abordagem é de forma automática, como o realizado por Smywiński-Pohl *et al.* (2019). Neste, é proposta a construção de um dicionário polonês, que mapeia a relação entre os termos jurídicos e extrajurídicos. Para isso, os pesquisadores compilaram documentos judiciais e extrajudiciais e realizaram etapas de pré-processamento para a redução de ruídos. Posteriormente, foram elaborados dois dicionários que combinam n-gramas obtidos por meio da ferramenta SRILM toolkit e a semelhança de cosseno entre os vetores dos termos dos dois dicionários com o auxílio do modelo *Word2Vec.*

Além das abordagens de construção mencionadas, existe uma abordagem híbrida, que utiliza um conjunto de palavras como semente para um contexto específico e um processo de expansão desse vocabulário. No estudo de Bos e Frasincar (2022), foram avaliadas três abordagens para a expansão automática de léxicos relacionados ao mercado financeiro: uma baseada na probabilidade de pertencimento das palavras a conjuntos positivos ou negativos, utilizando a medida *Pointwise Mutual Information* (PMI); outra que usa uma adaptação da medida *Term Frequency-Inverse Document Frequency* (TF-IDF), considerando documentos como categorias e avaliando a frequência das palavras em várias categorias; e uma terceira que emprega o Word2Vec como embedding de palavras para definir a proximidade entre conjuntos de palavras e termos da vizinhança para classificar em categorias apropriadas.

O processo de avaliar a qualidade do léxico em tarefas de AS pode ser entendido por meio da abordagem de um analisador lexical que faz a soma das pontuações dos termos-alvo, também conhecido como *Sentiment Orientation* (SO), como é usado em Oliveira, Cortez e Areal (2016), Carosia, Coelho e Silva (2020), Shan, Jiang e Wang (2021) e Wang *et al.* (2020). Uma opção com aprendizagem de máquina supervisionada consiste em utilizar SO para incrementar essas informações como entrada para um classificador de sentimentos. Um exemplo de aplicação dessa abordagem é o estudo realizado por Bos e Frasincar (2022), que utilizou support vector machine (SVM) com Bag-Of-Words (BOW) para codificação de texto na validação de um léxico de mercado financeiro americano e obteve uma acurácia de 75,1%.

Como mencionado em Pereira (2021), poucos trabalhos focam a análise dos textos na língua portuguesa. Assim, pensando nesse panorama apresentado pela revisão bibliográfica, em que se observa um déficit de propostas que adotam léxicos especializados no contexto da língua portuguesa, neste artigo será adotada uma estratégia para a construção automática de léxicos específicos para o domínio do mercado financeiro brasileiro.

Metodologia

Este capítulo fornece uma descrição detalhada dos procedimentos adotados neste estudo. No início, foi apresentado o conjunto de dados utilizado. Em seguida, foi apresentado o protocolo de processamento de texto aplicado em todas as etapas. Por último, foi detalhada a proposta de construção do léxico-alvo, incluindo a criação do léxico semente e suas variações.

Conjunto de Dados	Otimistas	Pessimistas	Total
Conjunto de *tweets* (AUTOR/A, 2019)	2048	1180	3228
Conjunto de Notícias (Januário *et al.*, 2021)	555	273	828

Quadro 1 - Informações dos conjuntos de dados para avaliar o desempenho final dos léxicos
Fonte: Elaborado pelos(as) autores(as).

Base de dados

Um dos conjuntos de textos adotadas é composto de 1.031.419 *tweets* distintos. As mensagens foram coletadas no ano de 2019, e foi possível utilizar uma API[1] fornecida pelo Twitter para esse fim. Foram utilizados os nomes de empresas e seus *tickers*[2] como filtros para a seleção das publicações. A coleta foi realizada conforme descrito em (AUTOR/A, 2019).

No contexto da avaliação de léxicos para a classificação de sentimentos em *tweets* sobre o MFB, utilizou-se um conjunto de teste composto de 3228 *tweets* rotulados, dos quais 2048 foram categorizados como otimistas e 1180 como pessimistas. Adicionalmente, na classificação de notícias do MFB, empregou-se um conjunto de teste composto de 828 notícias rotuladas, com 555 classificadas como otimistas e 273 como pessimistas, conforme produzido por Januário *et al.* (2021).

Pré-processamento dos textos

Durante os experimentos, todos os textos foram submetidos a etapas pré-processamento, o que é crucial para assegurar a qualidade dos dados utilizados nos próximos experimentos. O processo envolve a normalização das palavras para minúsculas, a remoção de stopwords usando a lista para a língua portuguesa disponível no Natural Language Toolkit (NLTK) (Bird, 2006), a eliminação de menções a usuários e a exclusão de URLs, hashtags, números, emoticons e pontuações. Além disso, o processo inclui a geração de tokens, que consiste na separação dos textos em palavras individuais.

Construção lexical

O fluxo do método principal, denominado como a primeira configuração do léxico, é apresentado na Figura 1. O método de construção e expansão automática do léxico é composto distintas. A primeira é a criação de um conjunto de palavras que represente o domínio, e a segunda etapa é a expansão das palavras.

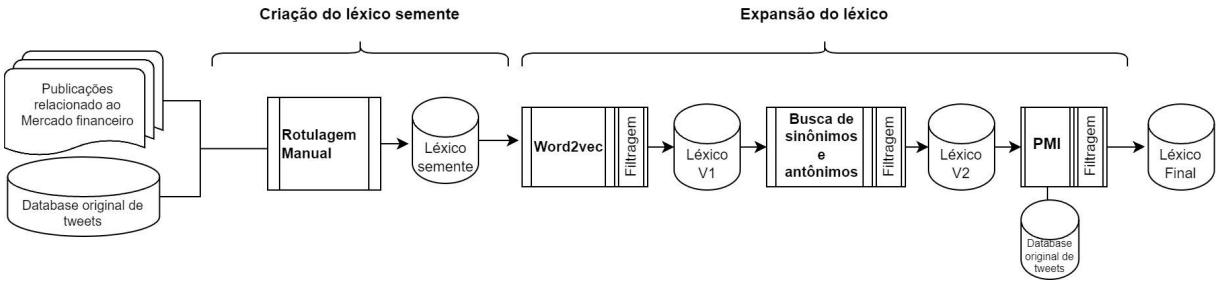

Figura 1 – Fluxo da construção lexical da principal configuração proposta. Semente (S) + Word2Vec (W2V) + Sinônimos e Antônimos (S/A) + Pointwise Mutual Information (PMI)
Fonte: Elaborado(a) pelos(as) autores(as).

1 Application Programming Interface.

2 Rótulos utilizados para identificar ações de uma empresa.

LÉXICO SEMENTE

O processo da criação do conjunto semente inicia-se com uma análise exploratória do corpora, buscando identificar as palavras que potencialmente exercem maior influência nos textos do domínio-alvo por conta da repetição observada. A primeira etapa para o domínio do MFB foi a seleção de conjuntos de textos que incluíram *tweets* e notícias coletadas no ano de 2019, submetidos ao pré-processamento dos textos, conforme descrito anteriormente, a fim de obter um conjunto de palavras sem possíveis ruídos. Dessa forma, foi possível gerar uma nuvem de palavras por meio do uso da biblioteca WordCloud, cujo resultado é apresentado na Figura 2.

Com base na WordCloud, as principais palavras identificadas são selecionadas e rotuladas manualmente. Exemplos dessas palavras incluem “lucro”, “sobe”e “alta”, as quais receberam uma classificação **otimista**. Por outro lado, palavras como “baixa”, “perda”e “queda” receberam uma classificação **pessimista**. Além da observação inicial da nuvem de palavras, também foi realizada uma leitura visual de alguns *tweets* e notícias recolhidos de sites de conteúdo relacionados ao MFB, como Informoney e Estadão Investimentos, a fim de catalogar suas palavras-chave.

EXPANSÃO LEXICAL

Com o léxico semente estabelecido, o processo de expansão lexical automático começa com o uso de *word embeddings*, que criam representações de palavras em espaços n-dimensionais, capturando relações sintáticas e semânticas. Foi utilizado um modelo Word2vec de 600 dimensões para a língua portuguesa, previamente treinado por Hartmann *et al.* (2017).

A expansão pelo Word2vec (W2V) começa com a divisão do léxico semente (S) em otimista e pessimista. Para cada léxico com n palavras, são criados lotes (batch) com três palavras em ordem de inserção do conjunto semente, por exemplo, $batch_1$(*palavra*$_1$, *palavra*$_2$, *palavra*$_3$), até que as n palavras estejam em lotes com no máximo três palavras. Posteriormente, o $batch_i$ é então entregue ao modelo pré-treinado Word2vec. Dessa forma, busca-se, dentro do espaço de representação, as palavras com os K vizinhos mais próximos, com base no cosseno de similaridade entre as palavras de entrada e seus vizinhos. Os três primeiros termos são selecionados para serem considerados como palavras candidatas, passando pela filtragem mencionada e, por fim, incorporados ao conjunto da rotulagem em expansão atualmente.

Figura 2 – Nuvem de palavras dos termos mais frequentes nos corpora de *tweets* do domínio do Mercado Financeiro Brasileiro
Fonte: Elaborado pelos(as) autores(as).

A segunda extensão envolve a expansão por Sinônimos e Antônimos (S/A), utilizando uma técnica de *web scraping* no site de dicionário online DICIO. Para isso, foi utilizada a biblioteca Python chamada Beautiful Soup.

O processo começa com a extração das palavras do léxico a ser estendido. Para cada palavra, é acessado o endereço virtual da página correspondente no site do dicionário, e as informações sobre sinônimos e antônimos são extraídas. Cada sinônimo é rotulado com a mesma orientação da palavra original, enquanto os antônimos recebem uma orientação oposta.

O passo para a expansão (S/A) é realizado extraindo-se as palavras do léxico a ser estendido. Para cada palavra, é feito o acesso à URL correspondente no site do dicionário e, por meio da ferramenta Beautiful Soup, são extraídas as informações sobre sinônimos e antônimos. Cada termo candidato sinônimo é rotulado com a mesma orientação da palavra que está sendo estendida. Por outro lado, caso o termo candidato seja um antônimo, ele será associado a uma orientação oposta. Por fim, os candidatos são submetidos a uma filtragem para verificar se os termos já estão presentes no léxico-alvo.

A terceira extensão, conhecida como “expansão por PMI”, utiliza a medida probabilista *Pointwise Mutual Information* (PMI) para quantificar o sentimento de uma palavra com base em sua probabilidade de ocorrência em um conjunto de dados. Essa abordagem amplamente explorada em trabalhos anteriores, como Oliveira, Cortez e Areal (2016), Losada e Gamallo (2020), Bos e Frasincar (2022), avalia a força de uma palavra em ser considerada positiva ou negativa em relação ao domínio em questão.

A medida estatística PMI é definida pela Equação 1:

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad (1)$$

Onde x e y são variáveis ou conjuntos de variáveis, *P*(x, y) representa a probabilidade conjunta de x e y ocorrerem, e *P*(x) e *P*(y) representam as probabilidades marginais de ocorrência de x e y no conjunto de variáveis.

A Orientação Semântica (OS) de uma nova palavra x é calculada como a diferença entre a força associada ao conjunto de palavras positivas (*setPositivo*) e a força associada ao conjunto de palavras negativas (*setNegativo*), conforme a Equação 2.

$$OS(x) = PMI(x, setPositivo) - PMI(x, setNegativo) \quad (2)$$

Essa diferença reflete a intensidade da associação da palavra com cada conjunto, permitindo inferir seu sentimento em relação ao domínio de interesse.

Com isso, é possível realizar uma série de passos com o objetivo de estender um conjunto de palavras. As etapas do procedimento estão ilustradas na Figura 3.

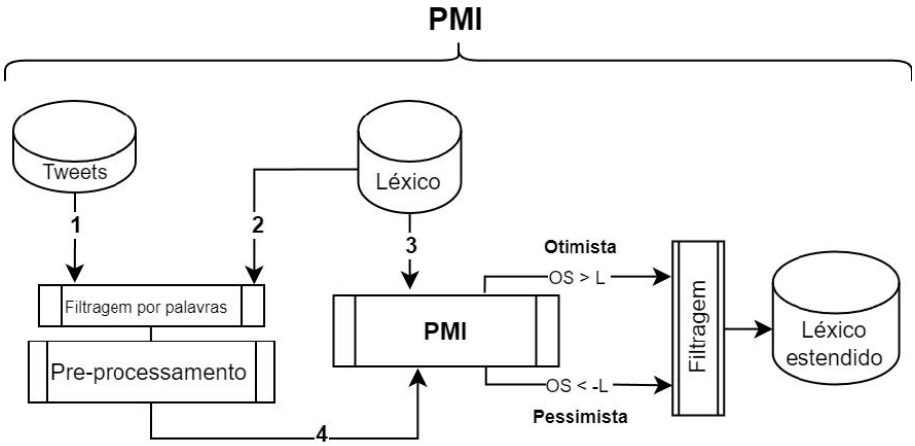

Figura 3 – Fluxo da extensão do léxico utilizando a medida Pointwise Mutual Information.
Fonte: Elaborado pelos(as) autores(as).

O processo começa filtrando *tweets* que contenham palavras do léxico a ser ampliado (etapas 1 e 2 da Figura 3), considerando que o PMI necessita relacionar as ocorrências dessas palavras-chave com outras, indicando maior probabilidade de co-ocorrência com sentimentos semelhantes. Em seguida, os *tweets* filtrados passam por pré-processamento e geram uma sequência de *tokens*. Esses *tokens*, juntamente com as palavras do léxico a ser ampliado, são usados para calcular a Orientação Semântica (OS) da nova palavra (etapas 3 e 4).

Por fim, a orientação semântica (OS) da nova palavra x será definida pela diferença entre o PMI com o léxico otimista e com o léxico pessimista. Para definir a polaridade da nova palavra, utiliza-se um limiar *L*, onde, caso o resultado de *O* seja maior que *L*, será considerado otimista, e caso seja menor que *–L*, será considerado pessimista:

$$OS(x) = \begin{cases} Otimista; & \text{se } OS \geq L; \\ Pessimista; & \text{se } OS \leq -L; \end{cases} \quad (3)$$

Dessa forma, a construção do léxico final para a configuração, que começou com o léxico semente e foi ampliada pelas etapas Word2Vec, busca por sinônimos/antônimos e, finalmente, a busca por novos termos por meio da medida PMI é concluída.

Alguns trabalhos como Carosia, Coelho e Silva (2020), Das *et al.* (2022) e Shan, Jiang, Wang (2021) rotulam o peso das palavras do léxico com +1 para positivas e -1 para negativas. Entretanto, essa abordagem considera que todos os termos possuem o mesmo grau de importância para a definição do sentimento associado ao texto-alvo. Uma abordagem contrastante é o uso personalizado de pesos associados a cada termo, como é feito em trabalhos como Oliveira, Cortez e Areal (2016), Bos e Frasincar (2022), Wang *et al.* (2020) e Losada e Gamallo (2020), em que o processo de rotulagem é utilizado para definir um peso para as palavras. Neste trabalho, será utilizada a pontuação PMI para determinar os pesos associados aos termos dos léxicos, conforme as etapas a seguir.

- Peso para palavras das etapas Semente, W2V e S/A: os termos obtidos nessas etapas, por tratar-se de palavras obtidas por relação semântica direta da rotulagem manual feita no conjunto semente, receberão pontuação máxima (+1 para otimistas e -1 para pessimistas).

- Peso para palavras da etapa PMI: será utilizada a função Min-Max Scaling, disponível na biblioteca *scikit-lear*[3], para normalizar entre +1 e -1 a pontuação PMI obtida para cada palavra estendida nessa etapa. A palavra com a maior pontuação PMI otimista será normalizada para +1, e a palavra com a menor pontuação PMI pessimista será normalizada para -1.

O conjunto final com os pesos personalizados segue o modelo apresentado no Quadro 2.

Dessa forma, a construção do léxico final, que iniciou-se com o léxico semente e foi ampliada pelas etapas Word2Vec, busca por sinônimos/antônimos e, finalmente, a busca por novos termos por meio da medida PMI (S+W2V+S/A+PMI) foi finalizada. Com o objetivo de verificar a melhor configuração e o impacto das etapas de expansão no léxico final, foram implementadas variações de configurações do léxico para serem avaliadas em experimentos na classificação de *tweets* e Notícias. Todas as configurações são apresentadas no Quadro 3.

Palavra	Etapa de ingresso	Peso	Rótulo
positiva	Semente(S)	+1	Otimista
recuar	Word2Vec	-1	Pessimista
perder	S/A	-1	Pessimista
conseguindo	PMI	+0.814	Otimista
greve	PMI	-1	Pessimista

Quadro 2 - Exemplo de pesos para palavras vindas de diferentes etapas
Fonte: Elaborado pelos(as) autores(as).

Cosntrução	Etapas
1	S
2	S+PMI
3	S+S/A+PMI
4	S+W2V+S/A+PMI

Quadro 3 - Configurações dos léxicos finais
Fonte: Elaborado pelos(as) autores(as).

Configurações dos experimentos de classificação

Os experimentos realizados neste estudo têm como objetivo testar diferentes configurações de léxicos gerados pelo processo descrito anteriormente. Inicialmente, aplicou-se uma abordagem que utiliza a técnica de soma das pontuações dos termos do léxico para a classificação de textos do MFB. Essa abordagem consiste em calcular uma pontuação total para cada texto, somando as pontuações individuais dos termos presentes no léxico, e utilizando essa pontuação para determinar a classificação do texto.

Em um segundo experimento, foram implementadas duas técnicas de aprendizado supervisionado: *Naive Bayes* (NB) e *Support Vector Machine* (SVM). Essas técnicas foram escolhidas devido à sua eficácia em tarefas de classificação de texto e foram implementadas com a biblioteca *scikit-learn* em Python. Para a representação dos textos, utilizou-se a abordagem *bag-of-words* (BOW), que transforma os textos em

3 https://scikit-learn.org

vetores de frequência de palavras, permitindo que os algoritmos de aprendizado de máquina processem os dados de forma eficiente.

No terceiro experimento, procurou-se enriquecer a representação dos textos utilizando informações adicionais fornecidas pelo analisador lexical. Essas informações foram integradas à representação matricial do texto, com o intuito de melhorar o desempenho dos modelos de classificação. A finalidade foi explorar se a inclusão de características léxicas adicionais poderia proporcionar um aumento na precisão dos modelos.

O treinamento em todas as tarefas de aprendizado supervisionado foi conduzido utilizando a técnica de validação cruzada *K-Fold*. Essa técnica envolve a divisão dos dados em K subconjuntos (ou "*folds*"), realizando múltiplos treinamentos e validações cruzadas.

Otimização do limiar para a etapa PMI

Para definir o limiar adequado para a rotulagem das palavras na etapa de extensão por PMI, será empregada a otimização bayesiana visando maximizar a métrica F-score na classificação de sentimentos de *tweets* e notícias, ao mesmo tempo que se minimiza a porcentagem de textos não classificados para ambos os corpus. Para isso, foi proposta uma métrica que consolida tais informações. A Equação 4 apresenta a função S(L), que pode ser entendida como uma combinação linear das métricas:

$$S(L) = \alpha \cdot (f1\_score\_T + f1\_score\_N) - \beta \cdot (Unclassified\_T - Unclassified\_N) \quad (4)$$

Onde $f1_{score}$ representa as pontuações F1 para os conjuntos de dados de *tweets* (T) e de notícias (N), e *Unclassified* representa a porcentagem de classificações não identificadas nesses conjuntos de dados. Os valores α e β são pesos atribuídos a cada métrica, refletindo sua importância relativa.

De acordo com o trabalho de Gardner *et al.* (2014), a Equação 5 apresenta a função fundamental da otimização bayesiana, representada da seguinte forma:

$$\min_{x \in X} f(x) \quad (5)$$

Onde $f(x)$ é a função objetivo que se deseja minimizar, e X é o espaço de busca. A otimização bayesiana envolve dois componentes principais: o Processo Gaussiano (GP) e a Função de Aquisição.

Um Processo Gaussiano é usado para prever o valor de uma função com base em observações anteriores. Ele fornece:

- **Média (μ):** Representa a estimativa média da função no ponto x.
- **Desvio padrão (σ):** Representa a incerteza dessa estimativa.

Como explicado por Snoek, Larochelle e Adams (2012), esses dois elementos modelam a função objetivo que queremos otimizar. Com base no GP, a função de aquisição decide onde olhar em seguida para encontrar o valor mínimo da função. Ela

equilibra entre explorar novas áreas (onde a incerteza é alta) e explorar áreas promissoras (onde os valores conhecidos são bons). A Equação 6 representa a função de aquisição ***Expected Improvement*** **(EI)**:

$$EI(x) = \sigma(x)[z\Phi(z) + \phi(z)] \tag{6}$$

- $\sigma(x)$ é a incerteza no ponto x.
- z é uma medida de quão bom x parece ser, calculada como $z = \frac{\mu(x) - f_{\text{melhor}}}{\sigma(x)}$, onde f_{melhor} é o melhor valor encontrado até agora.
- Φ e ϕ são funções da distribuição normal.

O processo termina ao atingir uma convergência quando a função de aquisição deixa de sugerir pontos significativamente diferentes ou quando a quantidade determinada de repetições é alcançada. Neste experimento, foram utilizadas 100 iterações, com o intervalo de valores *L* variando entre 0 e 25.

Construção	Otimistas	Pessimistas	Total
S	75	75	150
S+PMI	507	1153	1660
S+S/A+PMI	1492	1518	3010
S+W2V+S/A+PMI	1685	1946	3631

Quadro 4 - Quantidade de palavras dos dicionários
Fonte: Elaborado pelos(as) autores(as).

Resultados e discussões

Neste capítulo, serão expostos os resultados dos experimentos realizados neste trabalho. Primeiramente, serão apresentados os resultados referentes à quantidade de termos decorrentes da expansão lexical proposta neste estudo, assim como os resultados da otimização de limiar para definição da polaridade dos termos. Em seguida, serão discutidos os desempenhos dos léxicos em tarefas de classificação de sentimentos em *tweets* e notícias, ambas sobre o domínio do mercado financeiro brasileiro. Por fim, será feita uma comparação entre o desempenho do método lexical e o método supervisionado.

Expansão lexical

Os resultados da expansão lexical apresentados no Quadro 4 é derivado da proposta de construção apresentada anteriormente.

A expansão foi realizada inicialmente por meio da criação de uma semente (S). A coleta da semente resultou em um conjunto de 75 palavras consideradas otimistas para o contexto proposto, assim como outras 75 palavras consideradas pessimistas.

Em seguida, foi realizada a primeira variação no pipeline de expansão (S+PMI), na qual se observou o impacto da expansão buscando apenas as palavras por meio da medida probabilística PMI após o léxico semente. Isso resultou em 507 palavras otimistas e 1153 palavras pessimistas.

Uma segunda variação foi realizada, na qual foram buscados sinônimos e antônimos (S/A) do conjunto semente, seguidos pelo uso do PMI. Essa abordagem (S+S/A+PMI) resultou em 1492 palavras otimistas e 1518 palavras pessimistas.

Por fim, uma última variante da expansão lexical foi realizada, buscando a similaridade entre as palavras por meio da word embedding Word2Vec (W2V). Em seguida, foi feita uma busca por sinônimos e antônimos, finalizando com a busca por termos mais específicos em relação aos já expandidos utilizando o PMI. Essa abordagem (S+W2V+S/A+PMI) resultou em um conjunto otimista de 1685 palavras e 1946 palavras pessimistas.

O limiar selecionado para o processo de expansão do léxico foi determinado com base nos resultados da otimização bayesiana, conforme ilustrado na Figura 1. Na Figura 4, podemos observar o comportamento das métricas ao longo de diferentes limiares. Para facilitar a visualização das métricas, foi utilizada a função de desempenho $S(L)$, que combina o *F1-Score* e a taxa de classificação incorreta. Essa função foi normalizada para variar entre 0 e 1, sendo 1 o ponto máximo de desempenho e 0 o mínimo.

Com isso, o valor máximo de 1 acontece quando $L = 3,67$. Nesse ponto, as métricas individuais registram os seguintes valores: *F1-Score* para *tweets* (*F1_tweets*) é 0,7154, a proporção de *tweets* não classificados (Não classificado(T)) é 0,02, *F1-Score* para notícias (*F1_Noticias*) é 0,684, e a proporção de notícias não classificadas é zero.

Figura 4 – Variação das métricas normalizadas em função do limiar *T*. A linha sólida representa *S*(*T*), enquanto as linhas tracejadas representam as métricas *F*1_*tweets*, *F*1_*Noticia*s, Não classificado(*T*) e Não classificado(*N*)
Fonte: Elaborado pelos(as) autores(as).

Desse modo, a Figura 4 destaca a importância de selecionar um limiar adequado para balancear a maximização das métricas de desempenho e a minimização das classificações não identificadas. Escolher um limiar apropriado é crucial para garantir a eficácia da abordagem de expansão lexical apresentada neste trabalho.

Desempenho do léxico na classificação de *tweets* e notícias

Para avaliar o desempenho nas tarefas de classificação de sentimentos em *tweets* e notícias, foram consideradas as métricas de precisão, revocação (*recall*), *F1-Score* e acurácia geral. Além disso, utilizou-se a métrica denominada não classificados, que consiste na verificação da porcentagem de textos sem inferência da classe (Bos; Frasincar, 2022).

Na avaliação do desempenho das diferentes configurações de construção lexical, foram analisados os sentimentos de um conjunto de 3228 *tweets* categorizados manualmente. Esses resultados podem ser visualizados na Tabela 1.

Construção	Acurácia	Precisão	Recall	F1	Não classificadas
S+PMI	61,9%	65,1%	61,9%	63,4%	10,5%
S+PMI (lematizado)	66,5%	66,9%	66,5%	65,9%	4%
S+S/A+PMI	65,7%	68,6%	65,7%	67%	8,6%
S+S/A+PMI (lematizado)	71,7%	72,2%	71,7%	71,5%	2,8%
S+W2V+S/A+PMI	65%	66,7%	65%	65,7%	6,4%
S+W2V+S/A+PMI (lematizado)	71,1%	71,9%	71,1%	70,5%	2%

Tabela 1 - Avaliação dos léxicos de sentimento financeiro na classificação de *tweets* no conjunto de dados relacionados ao Mercado Financeiro Brasileiro (em %, melhores valores em negrito)
Fonte: Elaborado pelos(as) autores(as).

Todas as variações do léxico foram comparadas com uma abordagem de pré-processamento dos termos, em que as palavras, tanto no dicionário proposto quanto nos *tweets* a serem classificados, foram lematizadas, reduzindo-as ao seu lema raiz (Jung *et al.*, 2021). O melhor resultado foi obtido na proposta S+S/A+PMI (lematizado), com *F1-Score* de 71,5%, e uma precisão de 71,7%. Em contraste, sua versão não lematizada apresentou uma diferença negativa de até 6% na acurácia e 4,5% no F1. Isso se deve à facilidade de identificação dos termos uma vez normalizado, o que reduz a dimensionalidade dos termos, simplificando a comparação e o reconhecimento das palavras nos textos. No entanto, em termos de porcentagem de *tweets* que não foram classificados em razão da soma dos termos zerados ou à falta de cobertura das palavras do léxico nos tweets-alvo, a proposta mais adequada foi a S+W2V+S/A+PMI (lematizado), com um resultado de 2%, sendo assim considerando o léxico com a maior cobertura das palavras no conjunto de teste. Isso se deve principalmente as 621 palavras a mais nessa construção em comparação com a melhor abordagem geral.

O uso do léxico não se limita a textos no nível de sentença, mas também pode ser aplicado em documentos com textos mais extensos. Para testar o dicionário que obteve o melhor resultado previamente apresentado na Tabela 1, foi realizada uma avaliação do desempenho na classificação de um conjunto de notícias sobre o MFB, produzido por Januário *et al.* (2021). Essas 828 notícias possuem rótulos que indicam se o sentimento é otimista (555 notícias), refletindo uma alta expectativa de um determinado investidor em relação a uma ação, ou negativo, considerando um contexto pessimista (273 notícias).

	Acurácia	*F1-Score*
(1) Baseline (original) (JANUÁ- RIO *et al.*, 2021)	57,1%	57,4%
(2) Baseline (com stemming) (JANUÁRIO *et al.*, 2021)	58,2%	58,8%
(3) S+S/A+PMI	63,1%	63,4%
(4) S+S/A+PMI (com stemming)	58,5%	59,2%
(5) S+S/A+PMI (com lematização)	68,3%	68,4%

Tabela 2 - Desempenho médio das acurácias e *F1-Score* de notícias rotuladas usando o léxico de melhor pontuação (S+S/A+PMI) em comparação com o baseline (em %, melhores valores em negrito)
Fonte: Elaborado pelos(as) autores(as).

A Tabela 2 compara diferentes métodos de classificação, incluindo a linha de base original e variações do léxico com várias técnicas de processamento de texto. A linha de base original alcançou 57,1% de acurácia e um valor de *F1-Score* de 57,4%. Com a aplicação de *stemming* no pré-processamento, houve uma leve melhoria para 58,2% de acurácia e 58,8% de valor de *F1-Score*. No entanto, o uso do léxico proposto neste trabalho teve um impacto ainda mais significativo. A abordagem S+S/A+PMI alcançou 63,1% de acurácia e 63,4% de *F1-Score*. Com lematização, por sua vez, resultou em melhorias adicionais, elevando a acurácia para 68,3% e o valor de *F1-Score* para 68,4%.

Comparação com método supervisionado

Uma comparação foi realizada entre o desempenho do léxico de melhor resultado demonstrado anteriormente com os métodos SVM e NB, além de uma abordagem mista com analisador lexical. Os experimentos utilizaram um subconjunto de 2000 *tweets* do conjunto original, criado por meio da técnica de *Random Undersampling*.

Métrica	Léxico	SVM	NB	SVM+Léxico	NB+Léxico
F1-Score	67,8%	78,9% ± 0,7	76,4% ± 0,7	80% ± 0,6	77,7% ± 0,8

Tabela 3 - Comparação com método supervisionado treinando usando validação cruzada *K-Fold* K=50 em um sub-conjunto de 2000 *tweets*
Fonte: Elaborado pelos(as) autores(as).

Conforme ilustrado na Tabela 3, observou-se que o método de aprendizado de máquina SVM alcançou um valor de *F1-Score* de 78,9% com um desvio padrão de 0,7%, enquanto o NB atingiu 76,4% com um desvio padrão de 0,7%. Por outro lado, o léxico proposto alcançou um valor de *F1-Score* de 67,8%. Já quando é utilizado o léxico com as abordagens de AM, os melhores resultados foram alcançados, tendo como destaque SVM + Léxico com 80% de *F1-Score*. Essa comparação indica que os métodos supervisionados tiveram um desempenho superior em comparação ao uso único do vocabulário utilizado na classificação de *tweets* relacionados ao MFB, como também visto em Januário *et al.* (2021) e Das *et al.* (2022). Além disso, ao incluir informações adicionais do vocabulário como parte do treinamento, a abordagem supervisionada resulta em ganhos de desempenho.

Na abordagem lexical, a variação dos resultados está ligada à formulação do léxico utilizado e seu domínio. Exemplos disso são o estudo de Jung *et al.* (2021), que cobriu 41% dos termos de vocabulário conhecidos em triagens de câncer de mama. Já Wang *et al.* (2020) obteve 69,6% de acurácia na análise de sentimentos de comentários de filmes. Resultados semelhantes ocorrem no contexto financeiro, como acurácia

de 70% em publicações sobre o sistema financeiro americano Das *et al.* (2022) e a pontuação *F1-Score* de 58,2% Januário *et al.* (2021).

Conclusão

Este artigo comparou distintas abordagens para a criação e expansão automática de léxicos em língua portuguesa, focando a aplicação ao cenário do Mercado Financeiro Brasileiro, que apresenta poucos estudos relacionando tanto a língua portuguesa quanto o uso desses conjuntos de palavras especializados em tarefas de suporte na tomada de decisão por meio da análise de mensagens (Pereira, 2021). Os resultados alcançados destacaram um desempenho promissor na avaliação de sentimentos presentes em *tweets* e notícias relacionadas ao mercado, o que potencialmente poderia oferecer informações valiosos para a orientação de decisões e a análise do panorama desse contexto.

Foram apresentadas três abordagens de construção lexical com variações de pré-processamento, resultando em seis configurações finais para léxicos no contexto do Mercado Financeiro Brasileiro. Os experimentos abrangeram análise de sentimentos em mensagens curtas, como *tweets* relacionados ao mercado brasileiro, e em textos maiores, como notícias do mesmo domínio. A configuração S+S/A+PMI (com lematização) obteve o melhor desempenho, alcançando um *F1-Score* de 71,5% para a classificação de *tweets* e 68,4% para notícias, superando o baseline para notícias (JANUÁRIO *et al.*, 2021). Além disso, a abordagem lexical, combinada com o modelo *Support Vector Machine*, alcançou um *F1-Score* de 80%.

Dessa forma, o método proposto permite a criação de léxicos personalizados que podem ser ajustados de acordo com o contexto temporal dos dados, se adaptando às variações nas nuances da linguagem ao longo do tempo. Essa flexibilidade é particularmente relevante, pois permite a criação de léxicos específicos para diferentes áreas ou períodos, refletindo melhor as variações na linguagem utilizada. No contexto do mercado financeiro, por exemplo, isso possibilita uma análise de sentimentos mais precisa e dinâmica, levando em consideração as atualizações do mercado e os eventos que possam afetar as decisões dos investidores. Esse tipo de abordagem também pode ser aplicado em outras áreas, como a análise de tendências em redes sociais ou o monitoramento de crises, em que a atualização constante do vocabulário é crucial para uma compreensão eficaz das mensagens e sentimentos em evolução.

Referências

BIRD, S. *NLTK*: The Natural Language Toolkit. Barcelona: Association for Computational Linguistics, 2006.

BOS, T.; FRASINCAR, F. Automatically building financial sentiment lexicons while accounting for negation. *Cognitive Computation*, [*s. l.*], v. 14, p. 442-460, 2022.

CAROSIA, A. E.; COELHO, G. P.; SILVA, A. E. Analyzing the brazilian financial market through portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, London, v. 34, p. 1-19, 2020.

DAS, S. R.; DONINI, M.; ZAFAR, M. B.; HE, J.; KENTHAPADI, K. Finlex: An effective use of word embeddings for financial lexicon generation. *The Journal of Finance and Data Science*, Elsevier, v. 8, p. 1-11, 2022.

GARDNER, J. R.; KUSNER, M. J.; XU, Z. E.; WEINBERGER, K. Q.; CUNNINGHAM, J. P. Bayesian optimization with inequality constraints. *ICML*, [*s. l.*], p. 937-945, 2014.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *Proceedings of Symposium in Information and Human Language Technology*, Uberlândia, p. 122-131, oct. 2017.

JANUÁRIO, B. A.; CAROSIA, A. E. d. O.; SILVA, A. E. A. da; COELHO, G. P. Sentiment analysis applied to news from the brazilian stock market. *IEEE Latin America Transactions*, [*s. l.*], v. 20, n. 3, p. 512-518, 2021.

JUNG, E.; JAIN, H.; SINHA, A. P.; GAUDIOSO, C. Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis. *Health Informatics Journal*, [*s. l.*], v. 27, 2021.

LOSADA, D. E.; GAMALLO, P. Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, [*s. l.*], v. 54, p. 1-24, 2020.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, [*s. l.*], v. 66, p. 35-65, 2011.

MAHMOOD, A. T.; KAMARUDDIN, S. S.; NASER, R. K.; NADZIR, M. M. A combination of lexicon and machine learning approaches for sentiment analysis on facebook. *Journal of System and Management Sciences*, Pequim, v. 10, p. 140-150, 2020.

OLIVEIRA, N.; CORTEZ, P.; AREAL, N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, Elsevier, v. 85, p. 62-73, 2016.

PEREIRA, D. A. A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, [*s. l.*], v. 54, p. 1087-1115, 2021.

SHAN, R.; JIANG, T.; WANG, Y. Research on the construction of domain sentiment lexicon based on label propagation algorithm. *ACM International Conference Proceeding Series*, [*s. l.*], p. 1024-1029, 2021.

SMYWIŃSKI-POHL, A. *et al*. Automatic construction of a polish legal dictionary with mappings to extra-legal terms established via word embeddings. *ICAIL*, [*s. l.*], p. 234-238, 2019.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, [*s. l.*], v. 25, p. 2951-2959, 2012.

WANG, Y. *et al*. Automatic construction of domain sentiment lexicon for semantic disambiguation. *Multimedia Tools and Applications*, [*s. l.*], v. 79, n. 31-32, p. 22355-22373, 2020.