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

Thiago Monteles de Sousa[1], **Deborah S. A. Fernandes**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)

thiagomonteles@discente.ufg.br ,deborah.fernandes@ufg.br

Abstract. *This article investigates the opportunities in creating specialized lexicons with a focus on building a glossary in Portuguese aimed at the Brazilian Financial Market (MFB). The methodology used involves the design of a sequence of steps aimed at enriching a set of seed words, which is subsequently used in the task of analyzing sentiments in tweets and news related to the MFB domain. As results, a f1-score of 71.5% was achieved in the classification of tweets and a f1-score of 67.9% in news, both in the lexical approach. Furthermore, a mixed approach, combining the lexicon with the machine learning support vector machine model, achieved a f1-score of 77.4% in classifying tweets .*

Resumo. *Este artigo investiga as oportunidades na criação de léxicos especializados com foco na construção de um glossário em Português voltado para o Mercado Financeiro Brasileiro (MFB). A metodologia empregada envolve a concepção de uma sequência de etapas visando enriquecer um conjunto de palavras semente, que é posteriormente utilizado na tarefa de análise de sentimentos em tweets e notícias relacionadas ao domínio do MFB. Como resultados, foram alcançados um f1-score de 71,5% na classificação de tweets e um f1-score de 67,9% em notícias, ambos na abordagem lexical. Além disso, uma abordagem mista, combinando o léxico com o modelo de aprendizagem de máquina support vector machine, atingiu um f1-score de 77,4% na classificação de tweets.*

1. Introdução

Com a crescente popularização das plataformas de redes sociais *online*, como o *Twitter*[1], Facebook[2], LinkedIn[3] e outras, milhares de usuários têm interagido com postagens e mensagens que abordam uma grande variedade de tópicos. Essas plataformas se tornaram espaços onde os usuários expressam suas opiniões cada vez mais e também as utilizam como instrumento para a tomada de decisões [Bos and Frasincar 2022]. O *Twitter*, em particular, é uma das redes sociais mais populares no mundo, permitindo que cada usuário publique 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 através do compartilhamento das publicações (*retweets*) e interações, tornando-se uma fonte relevante para acompanhar tendências e opiniões [Carosia et al. 2020].

[1]www.twitter.com
[2]www.facebook.com
[3]www.linkedin.com

No contexto do mercado financeiro, o *Twitter* é uma plataforma amplamente utilizada por investidores para expressar suas opiniões devido à sua 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 ou a partir 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 a criação de um léxico em Português, voltado para o domínio do Mercado Financeiro Brasileiro (MFB). O objetivo é buscar palavras que possam indicar graus de otimismo ou pessimismo em textos relacionados ao campo alvo, contribuindo para a aplicação PLN neste contexto para a língua portuguesa, a qual ainda apresenta escassez de estudos publicados [Januário et al. 2022, Pereira 2021]. Visando 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.

O resumo das principais contribuições deste trabalho são as seguintes:

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

2. 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 e 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 et al. 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 [Smywiński-Pohl et al. 2019] e outros que exigem corpora especializado.

Sua construção pode ser dividida em totalmente manual, como em [Loughran and Mcdonald 2011], que apresenta uma popular coleção de palavras rotuladas para o domínio do mercado financeiro. Para isso, foi utilizado 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 redução de ruídos. Posteriormente, foram elaborados dois dicionários que combinam n-gramas obtidos através 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 and 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 através 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 et al. 2016, Carosia et al. 2020, Shan et al. 2021, Wang et al. 2020](SO) Uma opção com aprendizagem de máquina supervisionado 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 and 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 na análise dos textos na língua portuguesa, e pensando nesse panorama apresentado pela revisão bibliográfica, o qual 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.

3. Metodologia

Este capítulo fornece uma descrição detalhada dos procedimentos adotados neste estudo. No início da seção 3.1, foi apresentado o conjunto de dados utilizado. Em seguida, na seção 3.2, foi apresentado o protocolo de processamento de texto aplicado em todas as etapas. Por último, na seção 3.3, foi detalhada a proposta de construção do léxico alvo, incluindo a criação do léxico semente e suas variações.

3.1. Base de dados

Uma dos conjuntos de textos adotadas é composta por 1.031.419 *tweets* distintos. As mensagens foram coletadas no ano de 2019, utilizando a 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 [Fernandes et al. 2019].

Tabela 1. Informações dos conjuntos de dados para avaliar o desempenho final dos léxicos

Conjunto de Dados	Otimistas	Pessimistas	Total
Conjunto de Tweets [Fernandes et al. 2019]	2048	1180	3228
Conjunto de Notícias [Januário et al. 2022]	555	273	828

No contexto da avaliação de léxicos para a classificação de sentimentos em *tweets* sobre o MFB, utilizou-se um conjunto de teste composto por 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 por 828 notícias rotuladas, com 555 classificadas como otimistas e 273 como pessimistas, conforme produzido por [Januário et al. 2022].

3.2. Pre-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, URLs, hashtags, números, emoticons e pontuações. Além disso, o processo inclui a geração de *tokens*, que consiste na separação dos textos em palavras individuais.

3.3. Construção lexical

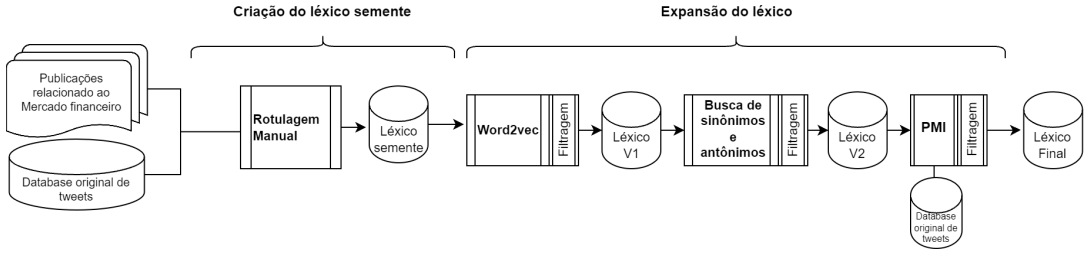

Figura 1. Fluxo da construção lexical da principal configuração proposta. Semente (S) + Word2Vec (W2V) + Sinônimos e Antônimos (S/A) + *Pointwise Mutual Information* (PMI).

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 por duas etapas distintas, sendo a primeira delas a criação de um conjunto de palavras que represente o domínio e a segunda etapa a expansão das palavras.

[1]Application Programming Interface.

[2]Rótulos utilizados para identificar ações de uma empresa.

3.3.1. Léxico semente

O processo da criação do conjunto semente se inicia com uma análise exploratória do corpora, buscando identificar as palavras que potencialmente exercem maior influência nos textos do domínio-alvo por conta da sua 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, que foram 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* [3], cujo resultado é apresentado na Figura 2.

Figura 2. Nuvem de palavras dos termos mais frequentes nos corpora de *tweets* do domínio do Mercado Financeiro Brasileiro.

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 recolhida de sites de conteúdo relacionados ao MFB como o Informoney[4] e Estadão investimentos[5], a fim de catalogar suas palavras-chave.

3.3.2. 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]. O léxico semente (S) é dividido em otimista e pessimista, sendo que 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-treinando 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

[3]https://amueller.github.io/word_cloud

[4]https://www.infomoney.com.br/

[5]https://einvestidor.estadao.com.br

filtragem mencionada e, por fim, incorporados ao conjunto da rotulagem em expansão atualmente.

A segunda etapa de extensão envolve a expansão por Sinônimos e Antônimos (S/A), utilizando uma técnica de *web scraping* no site de dicionário online DICIO[6]. 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 acessada 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.

A terceira etapa, conhecida como "extensão 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 et al. 2016, Losada and Gamallo 2020, Bos and 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 fórmula $PMI(x, y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$, 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 $OS(x) = PMI(x, setPositivo) - PMI(x, setNegativo)$. 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 objetivo de estender um conjunto de palavras, as etapas do procedimento são ilustradas na Figura 3.

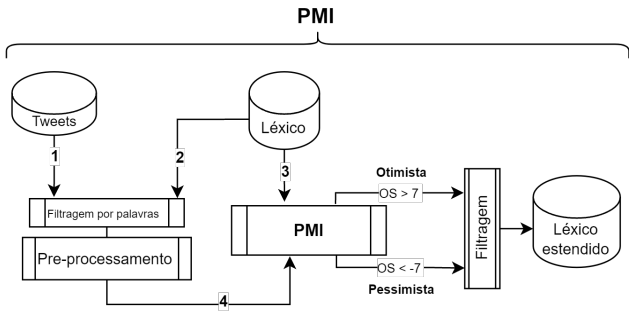

Figura 3. Fluxo da extensão do léxico utilizando a medida *Pointwise Mutual Information*.

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

[6]https://www.dicio.com.br/

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).

Se o resultado de OS for maior que 7, a palavra é considerada otimista; se for menor que 7, é considerada pessimista. Palavras com OS entre esses valores (menor que 7 e maior que -7) são descartadas, pois possuem alta co-ocorrência em ambos os conjuntos ou baixa frequência geral. A escolha desse limiar foi realizada de maneira empírica, observando que sete é um bom equilíbrio entre não deixar passar palavras com baixa Orientação Semântica (neutras) e deixar passar poucas palavras.

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 através da medida PMI é concluída. Visando verificar a melhor composição 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. Totalizando 3 configurações de construção lexical (S+PMI, S+S/A+PMI, S+W2V+S/A+PMI).

Os experimentos neste estudo visam testar diferentes configurações de léxicos gerados pelo processo descrito anteriormente. Inicialmente, foi aplicada uma abordagem que utiliza a técnica de soma das pontuações dos termos do léxico para classificar textos do MFB. Em um segundo experimento, foram usadas duas técnicas de aprendizado supervisionado: *Naive Bayes* (NB) e *Support Vector Machine* (SVM), implementadas com a biblioteca *scikit-learn*[7] em Python e representação matricial *bag-of-words* (BOW). Em um terceiro experimento, as informações do analisador lexical foram usadas como entrada adicional para a representação matricial do texto. O treinamento em todas as tarefas de aprendizado supervisionado foi conduzido com validação cruzada *K-Fold*.

4. Resultados e Discussões

Os resultados da expansão lexical, detalhados na Tabela 2, foram obtidos por meio de diferentes abordagens de expansão a partir de um léxico semente inicial composto por 75 palavras otimistas e 75 palavras pessimistas.

Tabela 2. Quantidade de palavras dos dicionários.

Construção	Otimista	Pessimista	Total
S	75	75	150
S+PMI	253	651	904
S+S/A+PMI	1087	1210	2297
S+W2V+S/A+PMI	1512	1385	2897

4.1. Desempenho do léxico na classificação de *tweets* e notícias

Para avaliar o desempenho nas tarefas de classificação de sentimentos associados ao conjunto de *tweets* e notícias, foram utilizados a média ponderada na precisão, revocação *(recall)*, *F1-Score* e acurácia geral, além disso, a métrica denominada não classificados

[7]https://scikit-learn.org

que consiste na verificação da porcentagem de textos que não tiverem a inferência da classe [Bos and 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 3 à seguir:

Tabela 3. Avaliação dos léxicos de sentimento financeiro na classificação de *tweets* no conjunto de dados relacionados ao Mercado Financeiro Brasileiro (em %, melhores valores em negrito).

Construção	Acurácia	Precisão	*Recall*	F1	Não classificadas
S+PMI	61,7%	66,9%	61,7%	64,2%	15,9%
S+PMI (lematizado)	67,3%	68,5%	67,3%	67,8%	6,2%
S+S/A+PMI	65,4%	68,3%	65,4%	66,8%	9,9%
S+S/A+PMI (lematizado)	**71,3%**	**71,8%**	**71,3%**	**71,5%**	2,8%
S+W2V+S/A+PMI	63,4%	66,1%	63,4%	64,7%	7,3%
S+W2V+S/A+PMI (lematizado)	69%	69,2%	69%	69,1%	**2,4%**

Todas as variações do léxico foram comparadas com uma abordagem de pré-processamento dos termos, onde 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 71,8%. Em contraste, sua versão não lematizada apresentou uma diferença negativa de até 5,9% na acurácia e 4,7% no F1. Isso se deve à facilidade de identificação dos termos uma vez normalizado reduzindo a dimensionalidade dos termos, simplificando a comparação e reconhecimento das palavras nos textos. No entanto, em termos de porcentagem de *tweets* que não foram classificados devido à 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 2,4%, sendo assim considerando o léxico com a maior cobertura das palavras no conjunto de teste. Isso se deve principalmente às 600 palavras a mais nessa construção em comparação à 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 3, foi realizada uma avaliação do desempenho na classificação de um conjunto de notícias sobre o MFB, produzido por [Januário et al. 2022]. 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).

A Tabela 4 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 64,9% de acurácia e 64,8% de *F1-Score*. Com lematização, por sua vez, resultou em melhorias adicionais, elevando a acurácia para 68,1% e o valor de *F1-Score* para 67,9%.

Tabela 4. 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)).

	Acurácia	*F1-Score*
(1) Baseline (original) [Januário et al. 2022]	57.1%	57.4%
(2) Baseline (com *stemming*) [Januário et al. 2022]	58.2%	58.8%
(3) S+S/A+PMI	64,9%	64,8%
(4) S+S/A+PMI (com steminização)	64,4%	63,6%
(5) S+S/A+PMI (com lematização)	**68,1%**	**67,9%**

4.2. 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 utilizando a técnica de *Random Undersampling*.

Tabela 5. Comparação com método supervisionado treinando usando validação cruzada *K-Fold* K=5 em um sub-conjunto de 2000 *tweets*.

	Léxico	SVM	NB	SVM+Léxico	NB+Léxico
F1-Score	67,1%	73,1 $\pm$ 0,2	73,7 $\pm$ 0,2	**77,4** $\pm$ 0,3	75,3 $\pm$ 0,3

Conforme ilustrado na Tabela 5, observou-se que tanto método de aprendizado de máquina SVM, como NB alcançaram um valor de *F1-Score* próximo de 73% com um desvio padrão de 0,2%. Por outro lado, o léxico proposto alcançou um valor de F1 de 67,1%. Já quando é utilizado o léxico com as abordagens de AM, os melhores resultados foram alcançados, tendo como destaque SVM + Léxico com 77,4% de *F1-Score*. Esta 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. 2022, Das et al. 2022]. Além disso, ao incluir informações adicionais do vocabulário como parte do treinamento, 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 por [Das et al. 2022] e a pontuação F1 de 58,2% obtido por [Januário et al. 2022].

5. Conclusão

Este artigo comparou distintas abordagens para a criação e expansão automática de léxicos em língua portuguesa, focando na 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 através 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 67,9% para notícias, superando o *baseline* em [Januário et al. 2022] para notícias. Além disso, a abordagem lexical, combinada com o modelo *Support Vector Machine*, alcançou um *F1-Score* de 77,4%.

Para trabalhos posteriores, poderá ser incluso estratégias visando a identificação de *n-gramas*, que consistem em conjuntos de palavras que aparecem frequentemente juntas em textos do domínio. Adicionalmente, a avaliação do uso de léxicos com abordagens de modelos sequencias em redes neurais.

Referências

Bird, S. (2006). Nltk: The natural language toolkit.

Bos, T. and Frasincar, F. (2022). Automatically building financial sentiment lexicons while accounting for negation. *Cognitive Computation*, 14:442–460.

Carosia, A. E., Coelho, G. P., and Silva, A. E. (2020). Analyzing the brazilian financial market through portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, 34:1–19.

Das, S. R., Donini, M., Zafar, M. B., He, J., and Kenthapadi, K. (2022). Finlex: An effective use of word embeddings for financial lexicon generation. *Journal of Finance and Data Science*, 8:1–11.

Fernandes, D. S. A., Fernandes, M. G. C., Borges, G. A., and Soares, F. A. (2019). Decision-making simulator for buying and selling stock market shares based on twitter indicators and technical analysis. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2626–2632.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks.

Januário, B. A., Carosia, A. E. d. O., Silva, A. E. A. d., and Coelho, G. P. (2022). Sentiment analysis applied to news from the brazilian stock market. *IEEE Latin America Transactions*, 20:512–518.

Jung, E., Jain, H., Sinha, A. P., and Gaudioso, C. (2021). Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis. *Health Informatics Journal*, 27.

Losada, D. E. and Gamallo, P. (2020). Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, 54:1–24.

Loughran, T. and Mcdonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66:35–65.

Mahmood, A. T., Kamaruddin, S. S., Naser, R. K., and Nadzir, M. M. (2020). A combination of lexicon and machine learning approaches for sentiment analysis on facebook. *Journal of System and Management Sciences*, 10:140–150.

Oliveira, N., Cortez, P., and Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54:1087–1115.

Shan, R., Jiang, T., and Wang, Y. (2021). Research on the construction of domain sentiment lexicon based on label propagation algorithm. *ACM International Conference Proceeding Series*, pages 1024–1029.

Smywiński-Pohl, A., Lasocki, K., Wróbel, K., and Strzała, M. (2019). Automatic construction of a polish legal dictionary with mappings to extra-legal terms established via word embeddings. *Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019*, pages 234–238.

Wang, Y., Yin, F., Liu, J., and Tosato, M. (2020). Automatic construction of domain sentiment lexicon for semantic disambiguation. *Multimedia Tools and Applications*, 79:22355–22373.