UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

THIAGO MONTELES DE SOUSA

Expansão automática de léxico para Análise de Sentimentos de Twitter

Uma abordagem para o domínio do Mercado Financeiro Brasileiro

Goiânia
2023

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [ ] Tese [x] Outro*:_monografia _____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

THIAGO MONTELES DE SOUSA

3. Título do trabalho

"**Expansão automática de léxico para Análise de Sentimentos de Twitter**"

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [ ] SIM [x] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 23/08/2023, às 08:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thiago Monteles De Sousa**, **Discente**, em 23/08/2023, às 10:50, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3976819** e o código CRC **6BDC9EFA**.

THIAGO MONTELES DE SOUSA

Expansão automática de léxico para Análise de Sentimentos de Twitter

Uma abordagem para o domínio do Mercado Financeiro Brasileiro

Trabalho de Conclusão apresentado à Coordenação do Curso de Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Bacharel em Computação.

Área de concentração: Otimização.

Orientadora: Profa. Deborah Silva Alves Fernandes

Goiânia
2023

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Sousa, Thiago Monteles de
Expansão automática de léxico para Análise de Sentimentos de Twitter [manuscrito] : Uma abordagem para o domínio do Mercado Financeiro Brasileiro / Thiago Monteles de Sousa. - 2023.
50 f.: il.

Orientador: Prof. Dr. Deborah Silva Alves Fernandes.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Ciência da Computação, Goiânia, 2023.
Bibliografia.
Inclui lista de figuras, lista de tabelas.

1. Léxico. 2. Analises de Sentimentos. 3. Mercado Financeiro Brasileiro. 4. Processamento de Linguagem Natural. I. Fernandes, Deborah Silva Alves, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Ao(s) **vinte e dois dias** dia(s) do mês de **agosto** do ano de **2023** iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado "**Expansão automática de léxico para Análise de Sentimentos de Twitter**", de autoria de **THIAGO MONTELES DE SOUSA**, do curso de **Ciência da Computação**, do(a) **Instituto de Informática** da UFG. Os trabalhos foram instalados pelo(a) **profa. Dra. Deborah S. A. Fernandes (INF/UFG)** com a participação dos demais membros da Banca Examinadora: **profa. Dra. Luciana de Oliveira Berratta (INF/UFG)**. Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 10,0 , tendo sido o TCC considerado **aprovado**.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 22/08/2023, às 16:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luciana De Oliveira Berretta**, **Professora do Magistério Superior**, em 22/08/2023, às 16:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3976820** e o código CRC **B9A3EE0E**.

Thiago Monteles de Sousa

Graduando em Ciência da Computação na UFG - Universidade Federal de Goiás. Ao longo da graduação, participou de uma iniciação científica na área de Análise de Sentimentos para o Mercado Financeiro Brasileiro. Atualmente, integra o projeto promovido pela UNICAMP denominado H.IAAC (IA para Arquiteturas Móveis e Cognitivas), onde faz parte da equipe de Processamento de Linguagem Natural.

Resumo

. **Expansão automática de léxico para Análise de Sentimentos de Twitter**. Goiânia, 2023. 50p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

Este estudo investiga as oportunidades na criação de léxicos especializados, através da exploração de uma abordagem híbrida. O objetivo é construir um léxico em Português voltado para o mercado financeiro brasileiro. Para isso, foram identificadas palavras-chave que refletem níveis de otimismo ou pessimismo em textos desse domínio. A seguir, realizou-se uma série de expansões dos conjuntos de palavras, empregando métodos de busca por sinônimos e antônimos, além de técnicas probabilísticas para identificar novas palavras. Por fim, os léxicos obtidos foram validados em tarefas de análise de sentimentos específicas para o mercado financeiro brasileiro.

Palavras–chave

Léxico, Abordagem Híbrida, Análise de Sentimentos, Processamento de Linguagem Natural, Mercado Financeiro Brasileiro, Tomada de Decisão

Abstract

. **Automatic lexicon expansion for Twitter Sentiment Analysis: An approach to mastering the Brazilian Financial Market**. Goiânia, 2023. 50p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

This study investigates the opportunities in creating specialized lexicons by exploring a hybrid approach. The objective is to build a lexicon in Portuguese aimed at the Brazilian financial market. For this, keywords were identified that reflect levels of optimism or pessimism in texts of this domain. Next, a series of expansions of the sets of words was carried out, using search methods for synonyms and antonyms, in addition to probabilistic techniques to identify new words. Finally, the lexicons obtained were validated in specific sentiment analysis tasks for the Brazilian financial market.

Keywords

Lexicon, Hybrid Approach, Sentiment Analysis, Natural Language Processing, Brazilian Financial Market, Decision Making

Este trabalho é dedicado ao meu pai, o qual mesmo sem querer me mostrou o mundo da computação. Expresso também minha sincera gratidão à minha mãe e irmã, que sempre estiveram ao meu lado, assim como aos meus amigos, cujo apoio inestimável foi essencial ao longo dessa jornada.

Agradecimentos

Agradeço à minha orientadora, Professora Deborah S. A. Fernandes, por me introduzir ao pensamento acadêmico e por ter me apresentado ao campo do Processamento de Linguagem Natural.

Immanuel Kant,
Sapere aude.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 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 [3]. 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 [4].

No domínio do mercado financeiro, alguns investidores frequentemente expressam suas opiniões no *Twitter*, aproveitando a plataforma por suas mensagens diretas, limitadas pela quantidade de caracteres, e pela facilidade de uso. Tendo como uma importante característica a influência midiática na dinâmica dos preços das ações, a qual embora prever flutuações seja uma atividade complexa, o *Twitter* pode servir como uma métrica para avaliar o clima dos investidores em relação às movimentações do mercado de ações [6]. Entretanto, levando em conta o número de publicações nas mídias sociais, torna-se inviável para o ser humano realizar análises manuais de uma grande quantidade de publicações, a fim de obter uma opinião consistente ou indicação sobre a situação visualizada. Diante desse cenário, a técnica de análise de sentimentos (AS) é uma abordagem de Processamento de Linguagem Natural (PLN) visando extrair de forma automática um indicador dessas opiniões [15].

Na SA, as tarefas são divididas entre a procura pela opinião, buscando identificar a polaridade, ou seja, o grau de positividade ou negatividade, e a procura pela emoção, onde a classificação se refere a um sentimento associado, como felicidade ou tristeza.

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

As técnicas para realizar SA podem ser separadas em: abordagens baseadas em *machine learning*, onde algoritmos aprendem padrões nos dados e os utilizam para prever novos exemplos; abordagens baseadas em léxicos, que realizam a classificação com base na orientação semântica das palavras presentes nos textos alvos, computando a soma ou média dos pesos dos termos encontrados; e abordagens baseadas em conceitos, que utilizam ontologias ou redes semânticas para realizar a classificação semanticamente [16].

Entretanto, existem diferenciais com relação aos recursos necessários para a utilização dessas abordagens, visto que as técnicas de *machine learning* e lexical apresentam vantagens distintas no processo de análise de sentimentos (AS). Na abordagem com *machine learning*, embora ofereça resultados mais promissores, é exigida uma grande quantidade de dados rotulados para o treinamento e teste dos modelos supervisionados, tornando o processo trabalhoso e custoso. Por outro lado, a abordagem lexical traz como vantagem a facilidade de construção, quando realizada de forma automática ou estendida, a partir de textos do campo alvo [13].

Considerando o uso da abordagem lexical, a criação do conjunto de palavras pode ser dividida em duas formas primárias: manualmente rotulada ou automaticamente. Na primeira, a qualidade das rotulagens das palavras é um diferencial positivo, porém, a dificuldade de recursos humanos traz um peso extra, gerando empecilhos na sua criação. Por outro lado, na construção automática, são utilizadas informações presentes em textos, bem como técnicas probabilísticas, o que torna possível encontrar palavras que condizem com o domínio esperado. No entanto, esse tipo de construção pode resultar em léxicos com mais ruídos de termos que não contribuem com o domínio [14].

Este trabalho tem como objetivo abordar as possibilidades de geração de vocabulários especializados, examinando uma abordagem híbrida para a criação de um léxico em Português, voltado para o domínio do mercado financeiro brasileiro. O objetivo é buscar palavras que possam indicar graus de otimismo ou pessimismo em textos relacionados ao campo alvo, contribuindo para a aplicação de Processamento de Linguagem Natural (PLN) neste contexto para a língua portuguesa, a qual ainda apresenta escassez de estudos publicados [6, 16]. 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 mercado financeiro brasileiro.

O resumo das principais contribuições deste trabalho são as seguintes:

- Elaboração de diferentes configurações para construção de léxicos do domínio alvo.
- Teste do desempenho dos léxicos através da análise de sentimentos em *tweets* e notícias no campo mercado financeiro brasileiro.
- Comparação do desempenho da abordagem lexical com abordagens baseadas em *machine learning* na classificação de sentimentos.

O restante deste trabalho está organizado da seguinte forma: o Capítulo 2 apresentará alguns trabalhos relacionados ao tema proposto, seguido pelo Capítulo 3, o qual abordará os conceitos teóricos necessários para o desenvolvimento deste estudo. O Capítulo 4 detalhará os métodos e materiais utilizados para realizar a pesquisa. Em seguida, o Capítulo 5 descreverá e analisará os resultados obtidos. Por fim, o Capítulo 6 apresentará as conclusões do trabalho.

CAPÍTULO 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 elaborado de forma adequada, os léxicos podem fornecer uma boa capacidade de classificação, além de poderem ser usados como recursos adicionais aos modelos de *machine learning* [15]. Ao longo dos anos, as abordagens de análise de sentimentos têm se tornado cada vez mais exploradas. Detectar subjetividades em sentenças e classificá-las em uma classe é um desafio, especialmente em domínios específicos, como o mercado de ações [5], doenças [10], documentos jurídicos [19] e outros que exigem um corpus especializado.

Sua construção pode ser dividida em totalmente manual, como em [12], 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 em [14], que propõe, via um conjunto de palavras considerado semente para o domínio proposto, um *pipeline* (série de passos) de expansão composto por três etapas: a primeira utiliza o *word embedding SkipGram* para identificar palavras candidatas para expansão. Na segunda, as palavras candidatas são filtradas através da verificação de sua presença em uma base de dicionário enciclopédico previamente selecionada. Na terceira etapa, são buscados diversos sinônimos dos termos selecionados. Concluindo o *pipeline* através da análise para determinar o grau de relevância de cada variação dos sinônimos encontrados, utilizando como medida as intersecções existentes entre as palavras filtradas e os conjuntos de sinônimos candidatos.

Além das abordagens de construção citadas, existe uma híbrida, usando dicionários de palavras já rotulados, adaptando para um perfil especializado, como em [18], que apresenta um método automatizado a partir da extração de termos de um texto base previamente categorizado com um conjunto semente de palavras. O método utiliza a fusão de um conjunto de termos semente com léxicos generalistas, realizando uma expansão lexical com *Word2Vec* e uma filtragem que mede a probabilidade de ocorrência das novas palavras no contexto do domínio. Para a categorização das novas palavras, é utilizada a

medida probabilística *Pointwise Mutual Information (PMI)* que estabelece a relação com conjuntos de textos do domínio específico. A construção totalmente manual, embora geralmente apresente uma boa acurácia nas tarefas, é bastante custosa tanto do ponto de vista de requisitar esforço humano especializado como do tempo dedicado à seleção e rotulagem das palavras. Esse desafio também se aplica à abordagem híbrida, que requer um conjunto de palavras ou textos previamente classificados.

Uma etapa crucial na extração de termos é a utilização de um corpus de texto com fontes provenientes do domínio desejado. Esses documentos fornecem informações úteis sobre as palavras mais frequentemente utilizadas no domínio proposto e podem incluir fontes como jornais, artigos e publicações de blogs, entre outros. Na literatura existe uma gama de trabalhos que utilizam essa estratégia como o realizado por [19]. 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 que abrangem domínios políticos e jurídicos. As etapas de pré-processamento foram realizadas para separar o texto em sentenças, gerar tokens e realizar a lematização e agregação de cada termo com sua classificação gramatical. Em seguida, foram criados dois dicionários combinando n-gramas encontradas através da ferramenta *SRILM toolkit*. Posteriormente, a semelhança de cosseno entre os vetores dos termos dos dois dicionários foi calculada com auxílio do algoritmo *Word2Vec*, permitindo estabelecer a relação entre termos jurídicos e extrajurídicos.

Em [10] é apresentada a criação de um léxico especializado para inscrições de triagem de voluntários em pesquisas de câncer de mama. Para tal, foi utilizado um banco de dados inicial de inscrições e suas avaliações finais, das quais foram extraídas as palavras mais relevantes através da combinação dos algoritmos de *frequency inverse document frequency (TF-IDF)* e *bag of words*. Uma equipe de especialistas validou os termos encontrados pelo algoritmo e, posteriormente, o dicionário foi ampliado pela fusão com glossários hospitalares do domínio do câncer. O resultado foi um léxico especializado com 4918 termos, sendo 2678 aprovados e 2240 negados. Para avaliar a eficiência do léxico, foi realizado um experimento utilizando TF-IDF para medir a combinação do dicionário com a base de dados SNOMED CT, que compila terminologias médicas amplamente utilizadas. O resultado foi uma cobertura de 41% nos termos alvos provido pela base de dados de teste em relação ao léxico construído.

Quando a construção do léxico é realizada por um conjunto de palavras sementes para o contexto proposto, a etapa de expansão é fundamental para aumentar a quantidade de termos relacionados aos inicialmente selecionados. Outros trabalhos, como os de [14, 18], usam *Word Embeddings* para expansão, assim como também [5] que propôs a geração de léxico do mercado financeiro. Para isso, são inseridas pares de palavras com sua conotação previamente identificada, sendo elas sinônimos ou antônimos. Caso

as palavras sejam sinônimos, são geradas duas listas de palavras e em seguida é realizada uma busca por interseções entre os termos iniciais. Já, se forem antônimos, é realizada a divisão das listas. Para isso foi utilizado o modelo de *embedding FastText*, buscando os K termos mais similares através da similaridade cosseno no espaço vetorial do vocabulário treinado. Depois da formação dos grupos, foi avaliado o rendimento dos léxicos criados em tarefas de classificação de sentimentos, usando como codificador de texto o algoritmo TF-IDF com o *framework AutoGloun*. O datatest *Financial Phrase Bank (FPB)* foi usado para comparar notícias financeiras com os sentimentos rotulados. Dessa forma, foi possível atingir uma acurácia de 70% ao usar o vocabulário construído.

Os autores de [3] avaliaram diferentes abordagens para a criação automática de léxicos relacionados ao mercado financeiro. São utilizadas três propostas: baseadas na probabilidade, na recuperação de informações e na *embedding* de palavras consciente aos sentimentos. A primeira baseia-se na probabilidade de uma palavra pertencer ao conjunto positivo ou negativo, adaptando o algoritmo *Pointwise Mutual Information (PMI)* para atribuir o valor máximo a uma palavra que apenas aparece em determinado sentimento. A segunda proposta utiliza uma adaptação da medida *Term Frequency-Inverse Document Frequency (TF-IDF)*, considerando documentos como categorias. Quanto mais uma palavra aparecer em várias categorias, menor é seu peso decisório para classificação de um texto. A terceira abordagem usa o *Word2Vec* como *embedding* de palavras para definir a proximidade entre conjuntos de palavras e classificar em uma categoria apropriada.

Outra estratégia adicional quando já se tem exemplos de termos alvos é procurar novos candidatos a expansão lexical usando medidas de estatística que busca a frequência e a co-ocorrencia de termos entre coleções de documentos. Como em [20] que visa construir um léxico específico de domínio sem ambiguidade, utilizando informações prévias de léxicos existentes e corpus de domínios específicos. Para tal, é proposta a integração dos valores de sentimento calculados pelo algoritmo modificado *term frequency–inverse document frequency (TF-IDF)*, que utiliza rótulos de polaridade previamente classificados e características de *part-of-speech (POS)* para obter valores sentimentais mais refinados. Os cálculos de refinamento de polaridade dos termos são baseados em quatro modificações do TF-IDF: a primeira visa medir a intensidade do sentimento dentro de um POS específico, a segunda busca a importância da palavra em uma certa categoria de polaridade, a terceira força a singularidade do sentimento associado ao POS e ao corpus de domínio, e a quarta soma a polaridade resultante das três medidas anteriores em um determinado corpus. Finalmente, o valor sentimental da palavra será a diferença do sentimento associado a cada polaridade e corpus base.

Em [15] é apresentado um método para criar léxicos baseados em microblogs do mercado financeiro. O estudo utiliza três algoritmos de aprendizado de máquina *Term frequency-inverse document frequency (TF-IDF), Information Gain (IG)* e *Pointwise*

Mutual Information (PMI) para gerar três léxicos diferentes usando postagens rotuladas da comunidade do site *StockTwits*. Para aprimorar cada léxico, foram criadas três versões adicionais: a primeira incluiu a medida Pdays, que indicava o sentimento mais associado a uma palavra ao longo do período das postagens coletadas; a segunda versão incluiu a medida Massoc, que considerava a frequência da palavra nas postagens; a terceira versão combinou ambas as medidas. O estudo demonstrou que a melhor abordagem foi a utilização do algoritmo PMI com as medidas de Pdays e Massoc, que resultaram em uma acurácia de 79,3% na soma dos indicadores de sentimento associados ao léxico.

Em [1] é proposto um modelo híbrido para extração automática de expressões n-gramas em inglês, denominadas *multiwords*, baseado em busca por padrões gramaticais, associações e similaridade de contexto. O método consiste em extrair os n-gramas de um corpus chave e filtrá-los com base em regras gramaticais e medidas estatísticas de associação e similaridade. São aplicados dois métodos de filtragem, *Association measure based filtering*, visando filtrar as palavras com base na frequência de aparição conjunta dos termos. Isso é feito usando os algoritmos *Dice's Coefficient (DC)*, *Point Wise Mutual Information (PMI)* e *Context similarity based filtering*, que buscam identificar a similaridade dos n-gramas em relação a um corpus de controle. Isso é feito utilizando a medida *Term Frequency - Inverse Document Frequency (TF-IDF)* para selecionar apenas conjuntos de palavras que possuem alta similaridade com documentos escolhidos.

O processo para avaliar a qualidade do léxico em tarefas de AS pode ser medido através uma abordagem de um analisador lexical que realiza a técnica de soma das pontuações dos termos alvo, comumente denominado como *Sentiment Orientation (SO)* como utilizado em [15, 4, 18, 20]. O uso de SO foi também demonstrado em [11] que apresenta metodologias para avaliar a construção de léxicos na triagem de indivíduos com sintomas depressivos, utilizando dois léxicos previamente rotulados e um corpus de textos classificados com conteúdo depressivo. Foram definidas três estratégias de avaliação: *Lemma-PoS, Word Embedding-based* e contexto explícito baseado em frequência e co-ocorrências de termos. O trabalho concluiu que a melhor estratégia foi a utilização do Lemma-PoS com a variação lexical de uni-gramas. Esta conseguiu classificar e acompanhar 78% dos indivíduos depressivos, sendo suas classificações obtidas pela medida SO que é feito através contagem do número de ocorrências de termos depressivos nos léxicos em documentos não classificados, fornecendo uma pontuação SO que indica o grau do sentimento associado ao contexto alvo.

Uma alternativa supervisionada para a soma dos termos é utilizar características provenientes do léxico para criar uma pré-classificação dos textos de treino. Isto pode ser feito adicionando informações, como a soma das pontuações das palavras apenas positivas e das negativas presentes no texto alvo, o número de mensagens positivas ou negativas no texto e as palavras com maior força no texto. Dessa forma, essas informações são

utilizadas como entrada para um classificador de sentimentos supervisionado, treinando assim algum modelo que realize essa soma do valor predominante no texto por meio de uma representação vetorial do problema. Um exemplo de aplicação dessa abordagem é o estudo realizado por [3], que utilizou *support vector machine* com BOW para codificação de texto na validação de um léxico de mercado financeiro e obteve uma acurácia de 75,1%.

Como mencionado em [16], 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 trabalho de projeto final de curso será adotada uma estratégia para a construção automática de léxicos específicos para o domínio do mercado financeiro brasileiro.

CAPÍTULO 3

Referencial teórico

A técnica de Análise de Sentimentos (AS) desempenha um papel fundamental na identificação da forma como os sentimentos são expressos em textos, permitindo a detecção de opiniões positivas ou negativas sobre um assunto específico. Essa análise pode ser realizada por meio de abordagens baseadas em Aprendizado de Máquina (AM) ou por meio de abordagens lexicais [16]. Este capítulo tem como objetivo apresentar as técnicas e conceitos fundamentais para o desenvolvimento deste trabalho. Iniciando no subcapítulo 3.1 com uma apresentação sobre os conceitos da abordagem lexical. Em seguida, no 3.2, serão abordadas as técnicas de preparação do texto. Por fim, no subcapítulo 3.3, serão apresentadas as abordagens que se baseiam em AM que serão usadas como referência para comparação com a abordagem lexical.

3.1 Abordagem lexical

A abordagem lexical em AS é tipicamente realizada através de um dicionário de palavras que apresenta uma pontuação ou uma rotulagem simples dos termos (positivo ou negativo), de acordo com o contexto proposto [6]. Diferentes métodos para a construção de um dicionário apresentam semelhança na utilização de um processo que parte de um conjunto inicial de palavras (semente) [14, 18]. Esses processos de ampliação consideram alguns critérios de semelhança entre as palavras através de dicionários ou representações vetoriais dos termos, ou com teorias estatísticos que levam em conta a frequência das palavras, além de poderem se valer de outros vocabulários criados, ajustando-os à área de atuação em questão.

3.1.1 Expansão baseada em similaridade

Uma abordagem rápida é a utilização de dicionários compilando palavras com seus respectivos significados, além de outros termos que representa uma relação de sinônimo ou antônimo ao termo alvo. Essa abordagem, como é usada em [14], tem uma vantagem pela facilidade na busca do termo alvo X, retornando um conjunto A e B, onde

A representa os sinônimos que podem ser entendidos como tendo a mesma orientação do termo chave X e antônimos B atribuindo uma orientação oposta à palavra inicial.

Outro método, dessa vez utilizado em [5, 3, 19, 11, 14] é a busca da similaridade de um termo através do uso de modelos pré-treinados de *word embeddings*. Trata-se de um método de representação vetorial de palavras em um espaço N-dimensional, no qual cada palavra é mapeada para um vetor denso de números reais. Dado um vocabulário V de palavras únicas, uma word embedding pode ser representada por uma matriz W de dimensão V x N. O treinamento de uma word embedding pode ser formalizado como um problema de otimização, no qual o objetivo é encontrar a matriz W que maximize a probabilidade de prever uma palavra em um contexto dado. Para buscar as K palavras que contenham maior similaridade com o termo X, é realizada uma busca no espaço vetorial W(X) através da similaridade de cosseno.

3.1.2 Expansão baseada em probabilidade

A abordagem de expansão lexical baseada em probabilidade tem como objetivo quantificar o sentimento de uma palavra com base em sua frequência de ocorrência em um conjunto de dados. A medida resultante dessa abordagem indica a força da palavra em ser considerada positiva ou negativa em relação ao domínio em questão. Um método probabilístico comumente utilizado nesse contexto é a *Pointwise Mutual Information (PMI)*, que avalia a associação entre duas palavras ou conjuntos de palavras. A PMI é capaz de fornecer uma indicação da polaridade de um termo com base em um conjunto de corpus utilizado como referência. Trabalhos anteriores, como [15, 11, 3], têm explorado essa abordagem para a expansão de léxicos.

A medida estatística PMI é definida como:

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3-1}$$

Nesta equação, x e y são variáveis ou conjuntos de variáveis. $p(x,y)$ representa a probabilidade conjunta de x e y ocorrerem, enquanto $p(x)$ e $p(y)$ representam as probabilidades marginais de ocorrência de x e y no conjunto de variáveis.

No modelo PMI, a orientação semântica O de uma determinada palavra é calculada pela diferença entre a força associada ao conjunto de palavras positivas e a força associada ao conjunto de palavras negativas. Essa diferença reflete a intensidade da associação da palavra com cada um desses conjuntos, permitindo inferir seu sentimento em relação ao domínio de interesse.

A orientação semântica O da nova palavra é definida como:

$$O(x) = PMI(x, setpositivo) - PMI(x, setNegativo) \tag{3-2}$$

Na expansão lexical, o termo x representa o item lexical candidato a ser inserido no dicionário já rotulados. O conjunto de palavras ou frases previamente consideradas positivas é referido como $setPositivo$, enquanto $setNegativo$ se refere ao conjunto já considerado negativo.

3.2 Processamento de texto

O uso de técnicas de processamento de texto desempenham um papel fundamental na preparação dos dados para a análise de sentimentos, ajudando a reduzir o ruído e a complexidade dos textos, ao mesmo tempo em que mantêm as informações relevantes para a tarefa em questão [17]. É nesse objetivo que o trabalho presente utilizará de conceitos comumente utilizados nesse processo, incluindo *stopwords*, *stemming*, lematização e *Bag-of-Words* (BoW).

Tabela 3.1: *Exemplo prático de remoção de stopwords*

Texto Original	Texto sem Stowords
"O valor da ação hoje está em uma crescida."	"valor ação hoje crescida."

stopwords, como utilizadas em [15, 3, 10, 14, 17], referem-se a um conjunto de palavras ou léxico que representa as palavras mais comuns na língua e que geralmente não possuem um impacto significativo no processamento final do texto. Exemplos de *stopwords* incluem artigos, preposições e pronomes, como 'o', 'a', 'este', 'ele', entre outros. A remoção dessas palavras durante o pré-processamento do texto oferece uma normalização, resultando em um texto contendo apenas palavras mais relevantes, como demonstrado na tabela 3.1.

Tabela 3.2: *Exemplo prático de stemming e lematização*

Texto Original	"Cresceram a quantidade de compras"
Texto com *stemming*	"Cresc a quantidad de compr"
Texto com Lematização	"Cresce a quantidade de compra"

A técnica de *stemming*, utilizada em [5, 6, 17], tem como objetivo reduzir palavras à sua forma raiz ou radical, removendo sufixos e prefixos, conforme ilustrado na tabela 3.2. Isso permite agrupar palavras semelhantes que compartilham a mesma raiz, reduzindo a dimensionalidade do vocabulário e tratando palavras relacionadas como iguais. Por outro lado, a lematização é uma técnica semelhante ao *stemming*, mas leva em consideração a estrutura morfológica das palavras. Diferentemente do *stemming*, que apenas corta sufixos e prefixos, a lematização analisa a palavra considerando sua classe gramatical e a converte para sua forma base ou lema. Essa abordagem resulta em palavras mais significativas e bem formadas, e foi utilizada em [14, 15, 19, 11, 10].

Tabela 3.3: *Exemplo prático de Bag-of-Words (BoW)*

Frase	Comprei	essa	ação	é	ruim
Comprei essa ação	1	1	1	0	0
Essa ação é ruim	0	1	1	1	1

A representação *Bag-of-Words* (BoW) ultilizada em [6, 17] e demonstrada na tabela 3.3 é uma abordagem comum para transformar o texto em uma forma numérica adequada para análise. Nesse método, um vocabulário é criado a partir de todas as palavras únicas nos documentos. Cada documento é então representado por um vetor, onde cada posição do vetor corresponde a uma palavra do vocabulário e o valor indica a frequência ou presença dessa palavra no documento. Essa representação permite que os algoritmos de aprendizado de máquina lidem com dados textuais.

3.3 Abordagem com aprendizagem supervisionada

A aprendizagem de máquina supervisionada é amplamente utilizada em técnicas de análise de sentimentos, com destaque para os classificadores tradicionais Naive Bayes (NB) e Support Vector Machines (SVM) [4]. O Naive Bayes é um classificador estatístico que utiliza o teorema de Bayes para prever a probabilidade de um conjunto de características pertencer a uma classe específica, sendo considerado "ingênuo"devido à suposição de independência condicional entre as características. Por outro lado, o SVM busca criar um hiperplano no espaço vetorial que seja capaz de separar as diferentes classes [6].

Essas técnicas de aprendizagem de máquina supervisionada têm sido aplicadas na previsão da orientação semântica de textos sobre o mercado financeiro, como demonstrado em trabalhos anteriores [3, 4, 6]. Essas abordagens têm demonstrado eficácia na classificação e análise de sentimentos quando já há uma grande quantidade de exemplos rotulados, sendo, portanto, o maior desafio para a realização de modelos de AS ao se utilizar o método supervisionado.

CAPÍTULO 4

Materiais e métodos

Este capítulo apresenta uma descrição detalhada dos procedimentos adotados neste trabalho. Inicialmente, na Seção 4.1, é abordado o conjunto de dados utilizado em todo o estudo. Em seguida, na Seção 4.2, são apresentadas o protocolo de processamento de texto realizadas em todas as etapas. A primeira etapa, na Seção 4.3, apresenta a proposta de construção do léxico alvo, explorando os dados para um melhor entendimento do domínio do léxico proposto, criando o léxico semente e as variações de construção lexical. Será detalhado o processo de seleção de termos, a associação de polaridades e a criação de um conjunto de palavras-chave relacionadas ao domínio financeiro. Na segunda etapa, Seção 4.4, são descritos os experimentos a serem realizados para avaliar os léxicos gerados e também as configurações dos classificadores utilizados, as métricas de avaliação e os procedimentos para treinamento e teste dos modelos. Por fim, na Seção 4.5, são apresentadas as métricas utilizadas para avaliar os resultados obtidos.

A figura 4.1 ilustra o fluxo dos assuntos que serão detalhados neste capítulo.

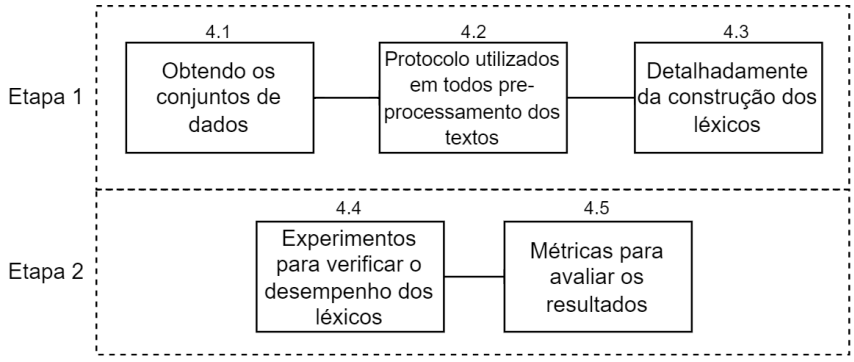

Figura 4.1: *Fluxo dos assuntos que serão aboradados no Capítulo 4.*

4.1 Base de dados

Uma das bases de dados 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 [8].

Tabela 4.1: *Informações dos conjuntos de dados para avaliar o desempenho final dos léxicos*

Conjunto de Dados	Total de Exemplos	Otimistas	Pessimistas
Conjunto de Tweets	3228	2048	1180
Conjunto de Notícias	828	555	273

Para o conjunto de teste, utilizado na avaliação dos léxicos na classificação de sentimentos de *tweets* relacionados ao mercado financeiro brasileiro, foram rotulados manualmente 800 *tweets* provenientes do banco de dados mencionado anteriormente. Além disso, foram utilizados mais 2428 *tweets* rotulados manualmente fornecidos por [8]. O conjunto de teste final consiste em 3228 *tweets* rotulados, sendo 2048 considerados otimistas e 1180 pessimistas. Devido a esse desequilíbrio, foi criada uma variação balanceada utilizando a técnica de *Random Oversampling*. Essa técnica consiste em selecionar aleatoriamente exemplos do conjunto minoritário e criar cópias adicionais incrementadas no conjunto final, resultando em um conjunto balanceado.

Para a classificação de notícias relacionadas ao domínio financeiro brasileiro, foi utilizado um conjunto de teste composto por 828 notícias rotuladas, produzido por [6]. Dentre esse total, 555 notícias possuem rótulos que indicam um sentimento otimista, refletindo uma alta expectativa de um determinado investidor em relação a uma ação, enquanto 273 notícias têm um contexto pessimista. Uma visão geral dos conjuntos que serão usados nesse estudo é apresentada na tabela 4.1.

4.2 Pre-processamento dos textos

Durante todos os experimentos, tanto as mensagens coletadas do *Twitter* quanto as notícias coletadas passaram por um processo de pré-processamento. Essa etapa teve como objetivo remover informações desnecessárias para o processamento nas etapas seguintes. Como já exemplificadas na Subseção 3.2, foram adotadas as seguintes etapas:

1. Normalização das palavras: todas as palavras são convertidas para minúsculas.

[1] Application Programming Interface.
[2] Rótulos utilizados para identificar ações de uma empresa.

2. Remoção de *stopwords*: uma lista de palavras amplamente utilizadas na linguagem, como pronomes, numerais e outras com pouca relevância para o modelo, para isso foi utilizado a lista para língua portuguesa disponível em *Natural Language Toolkit* (NLTK)[3] [2].
3. Remoção de menções a usuários, URLs, hashtags, números, emoticons e pontuações.
4. Tokenização: o texto é separado em palavras individuais, a qual são chamados de *tokens*.

Além disso, nos *tweets*, foram empregados os seguintes procedimentos adicionais:

5. Remoção de *tweets* que, após a aplicação das etapas anteriores, resultaram em menos de 3 *tokens*.
6. Remoção de *tokens* que ocorreram menos de 10 vezes no conjunto de palavras resultante da etapa anterior.

Essas ações visam simplificar e limpar o texto, removendo elementos que não são relevantes para a análise de sentimentos. O pré-processamento dos textos é uma etapa fundamental para garantir a qualidade dos dados utilizados nos experimentos subsequentes.

4.3 Construção lexical

O método de construção e expansão automática do léxico consiste em duas etapas distintas. Na primeira etapa, é realizada a seleção manual de palavras para a criação de um conjunto semente. Em seguida, na etapa de expansão do léxico, são empregados três passos principais: utilização de uma *Word Embedding* chamada *Word2Vec* (W2V) previamente treinada em português, busca por sinônimos e antônimos (S/A) relevantes e identificação de novos termos com base em abordagem probabilísticas dos corpora selecionados (PMI). Cada passo da etapa de expansão do léxico é submetido a um processo de filtragem para assegurar a ausência de interseções entre as classes positivas e negativas no contexto-alvo. O fluxo do método principal denominado como a primeira configuração do léxico é apresentado na Figura 4.2, e todas as etapas serão detalhadamente descritas a seguir.

[3] https://www.nltk.org

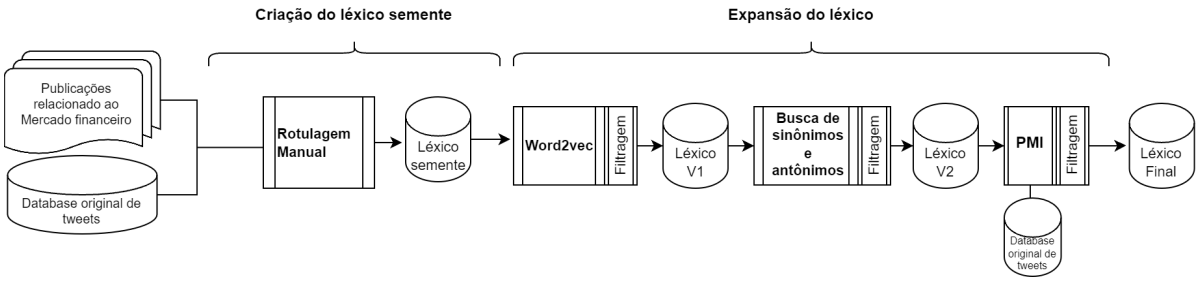

Figura 4.2: *Fluxo da construção lexical da principal configuração proposta. Semente(S) + Word2Vec(W2V) + Sinônimos e Antônimos (S/A) + Pointwise Mutual Information (PMI)*

4.3.1 Léxico semente

O léxico semente é composto por uma coleção de palavras com a polaridade do sentimento previamente classificado. Essa rotulagem provê-se de informações sobre algumas palavras comumente usadas no contexto alvo. Dessa forma, as etapas seguintes utilizaram o conhecimento intrínseco nos conjuntos semente para, então, buscar novas palavras para serem inseridas no dicionário final.

O processo da criação do conjunto semente se inicia com uma análise exploratória dos corpora, buscando identificar as palavras que potencialmente exercem maior influência nos textos do domínio-alvo por conta da sua repetição observada. Conforme ilustrado na Figura 4.3.

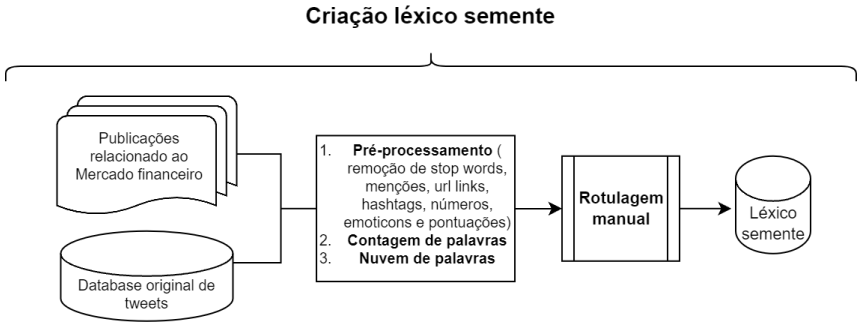

Figura 4.3: *Fluxo da construção do léxico semente.*

A primeira etapa para o domínio do mercado financeiro brasileiro foi a seleção de conjuntos de textos mencionados anteriormente no subcapítulo 4.1. Esses conjuntos incluíram *tweets* e notícias coletadas no ano de 2019, que foram submetidos ao pré-processamento dos textos, conforme descrito em 4.2, a fim de obter um corpus sem possíveis ruídos, como números, hashtags e outros elementos que não contribuem para a análise final. Dessa forma, foi possível gerar uma nuvem de palavras por meio do uso

da biblioteca *WordCloud* [4], cujo resultado é apresentado na Figura 4.4 a seguir:

Figura 4.4: *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", "subiu", "batendo"e "gostei", as quais receberam uma classificação **otimista**. Por outro lado, palavras como "baixa", "perda"e "suspende"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 mercado financeiro brasileiro como o Informoney[5] e Estadão investimentos[6], a fim de catalogar suas palavras-chave. Como resultado, o conjunto lexical semente consiste em 75 palavras consideradas otimistas e 75 palavras consideradas pessimistas.

4.3.2 Expansão lexical

Com o léxico semente criado, o processo de expansão lexical se inicia para configuração presente na figura 4.2 com o objetivo de estender os termos já catalogados, utilizando o conhecimento de textos da língua portuguesa ou dicionários.

Cada etapa de extensão (W2V, S/A e PMI) envolve uma filtragem entre os termos candidatos, verificando a existência de interseção entre as novas palavras e o conjunto

[4] https://amueller.github.io/word_cloud
[5] https://www.infomoney.com.br/
[6] https://einvestidor.estadao.com.br

oposto ao rótulo das palavras de entrada. Por exemplo, na expansão do léxico otimista, verifica-se se as palavras candidatas já estão inseridas ao léxico pessimista.

Em anos recentes, o uso de *word embeddings* tem se mostrado eficiente quanto à criação de representações de palavras em espaços n-dimensionais, podendo captar através de grandes conjuntos de corpus as relações sintáticas e semânticas dos termos. Por conta dessas características, eles têm sido utilizados em diversas tarefas de processamento de linguagem natural, como expansões de léxicos, como mostrado nos capítulos 2 e 3.

Com isso, foi utilizado um modelo para língua portuguesa Word2vec[7] de 600 dimensões, treinado por [9].

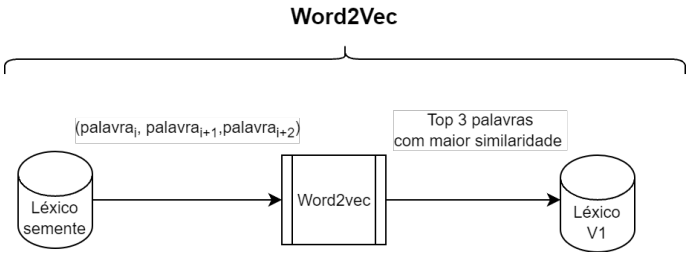

Figura 4.5: *Fluxo da extensão do léxico utilizando Word2Vec (W2V).*

Para expandir a lista semente (S) como demonstrado na Figura 4.5, inicialmente S é dividido entre um léxico 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 $batch_1(palavra_1, palavra_2, palavra_3)$, esse processo irá acontecer até que as n palavras estejam em lotes com no máximo três palavras. Cada $batch_i$ 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 filtragem mencionada e, por fim, incorporados ao conjunto da rotulagem em expansão atualmente.

A segunda etapa de extensão é classificada como expansão por sinônimos e antônimos, na qual se realiza uma busca no site de dicionário online de Português DICIO[8]. Para essa finalidade, foi empregada uma técnica de *web scraping*, que consiste na extração de dados de sites da web, convertendo-os em informações estruturadas. Para isso, utilizou-se a biblioteca em Python chamada *Beautiful Soup*[9]. O procedimento é ilustrado nas etapas apresentadas na figura 4.6 a seguir:

[7] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc
[8] https://www.dicio.com.br/
[9] https://www.crummy.com/software/BeautifulSoup

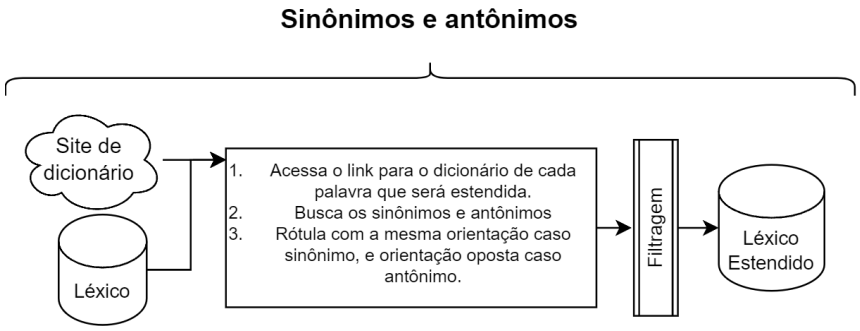

Figura 4.6: *Fluxo da extensão do léxico utilizando sinônimos e antônimos*

O processo para a expansão de Sinônimos e Antônimos (S/A) é realizado inicialmente 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. Finalizando o processo, os candidatos são submetidos a uma filtragem para verificar se já estão presentes no léxico alvo.

A terceira etapa, classificada como "extensão por PMI", refere-se ao uso da medida probabilista *Pointwise Mutual Information*, com o objetivo de buscar novas palavras que não possuam uma relação semântica tão inerente como as encontradas na etapa de extensão por em S/A e W2V. As etapas do procedimento são ilustradas na figura 4.7 a seguir:

O processo se inicia filtrando os *tweets* que contenham palavras do léxico a ser estendido (etapas 1 e 2 da figura 4.7). Isso se deve ao fato de que o PMI é uma medida probabilista que relaciona as ocorrências da palavra-chave (do léxico) com outras palavras (ou conjunto de palavras), sendo que quanto maior essa co-ocorrência, maior a probabilidade de estarem na mesma orientação de sentimento. Em seguida, os *tweets* filtrados são pré-processados, conforme enumerado em 4.2. Os *tweets* processados geram uma sequência de *tokens*, que são passados ao PMI, juntamente com as palavras do léxico que será estendido (etapas 3 e 4). Conforme foi descrito em 3.1.2, as palavras candidatas terão sua pontuação PMI calculada tanto com o conjunto pessimista do léxico como também com o conjunto otimista.

Por fim, a orientação semântica (O) da nova palavra X será definida pela diferença entre o PMI com o léxico otimista e com o léxico pessimista. Caso o resultado de O seja maior que 7, será considerado otimista, e caso seja menor que 7, será considerado pessimista:

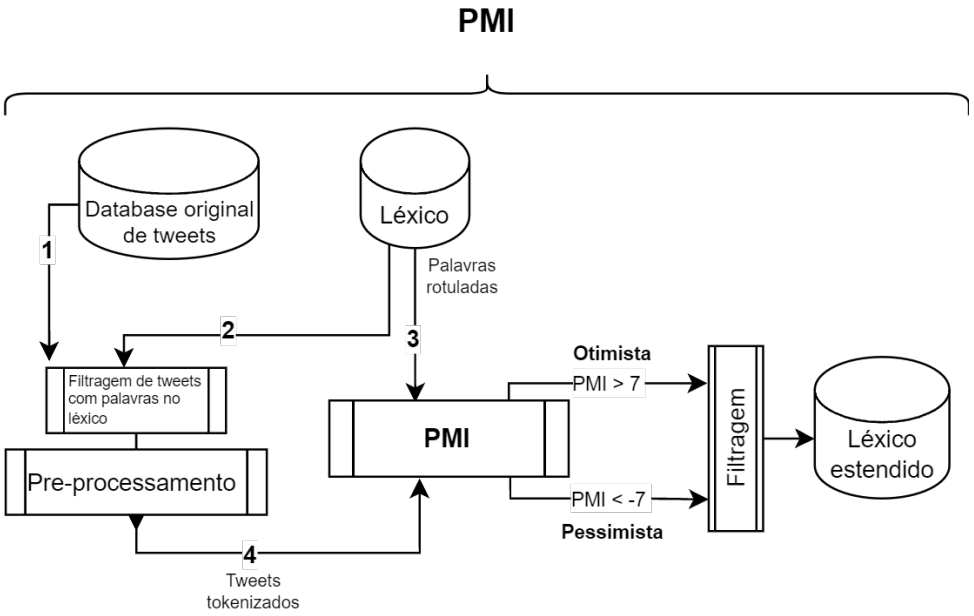

Figura 4.7: *Fluxo da extensão do léxico utilizando a medida Pointwise Mutual Information*

$$O(x) = \begin{cases} Otimista; & \text{se } O \geq 7; \\ Pessimista; & \text{se } O \leq -7; \end{cases}$$

As palavras que ficarem entre esse limiar (menor que 7 e maior que -7) serão descartadas, pois apresentaram uma alta co-ocorrência tanto no conjunto otimista quanto no pessimista ou o número de ocorrências no conjunto geral é baixa. A escolha desse limiar foi realizada de maneira empírica, observando que sete é um bom *trade-off* entre deixar passar palavras com baixa orientação semântica (neutras) e deixar passar poucas palavras. Por fim, as palavras candidatas são submetidas à filtragem para verificar se já existem no léxico que está sendo estendido.

Alguns trabalhos como [4, 5, 18] 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 [15, 3, 20, 11], onde 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, seguindo as seguintes etapas:

1. **Peso para palavras das etapas Semente, W2V e S/A**: Os termos obtidos nessas etapas, por tratarem-se de palavras obtidas por relações semântica direta da rotu-

lagem manual feito no conjunto semente, receberão pontuação máxima (+1 para otimistas e -1 para pessimistas).

2. **Peso para palavras da etapa PMI**: Será utilizada a função *Min-Max Scaling* disponível na biblioteca *scikit-learn*[10] para normalizar entre +1 e -1 a pontuação PMI obtida para cada palavra estendida nessa etapa. Sendo que 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 na Tabela 4.2 a seguir:

Tabela 4.2: *Exemplo de pesos para palavras vindas de diferentes etapas.*

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

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 (S+W2V+S/A+PMI), é concluída.

Tabela 4.3: *Configurações dos léxicos finais*

Construções	S	S+PMI	S+S/A+PMI	S+W2V+S/A+PMI

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 na Tabela 4.3.

4.4 Experimentos

Os experimentos conduzidos neste trabalho têm como objetivo testar as diferentes configurações de léxicos gerados pelo processo descrito no subcapítulo anterior 4.3. Para esse fim, foi utilizada uma abordagem de analisador que utiliza a técnica de soma das pontuações dos termos presentes no conjunto lexical para realizar a classificação de textos do domínio do mercado financeiro brasileiro.

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

4.4.1 Desempenho no conjunto de tweets

O primeiro experimento com o conjunto de *tweets* teve início com as etapas de pré-processamento do texto, cujo objetivo era reduzir possíveis ruídos que poderiam prejudicar a abordagem lexical. As etapas de pré-processamento foram enumeradas na seção 4.2. Em seguida, foram realizados experimentos aplicando a técnica de lematização dos termos, resultando em duas versões finais de avaliação: uma versão original (não lematizada) e uma versão lematizada. Esse procedimento foi realizado para cada variação dos léxicos finais apresentados na tabela 4.3.

Posteriormente, cada exemplo processado teve seu conjunto de textos separados em uma lista de palavras. Foi aplicada a abordagem lexical, que consiste na soma dos termos presentes no dicionário construído, e em seguida, foi recolhido o peso de cada palavra encontrada. Dessa forma, caso a soma dos pesos resultasse em um valor positivo, a amostra seria considerada otimista. Por outro lado, caso o resultado fosse negativo, o exemplo seria classificado como pessimista.

Na tabela 4.4, é apresentado um exemplo de um *tweet* original, o texto após o pré-processamento, a pontuação para cada palavra encontrada pelo léxico e a soma final do sentimento.

Tabela 4.4: *Exemplo da classificação de um tweet*

Texto original: @mobilon de qualquer forma, quem investiu naquela ocasião em #PETR4 ainda está ganhando mais do que os que ficaram no FGTS.
Texto processado: qualquer forma quem investir aquele ocasiao petr ainda estar ganhar mais ficar fgts
Peso por palavra: qualquer = 0.014 , investir = 0.271 , ocasiao = 1.0 , petr = 0.036 , ainda = 1.0 , ganhar = 1.0 , mais = 1.0 , ficar = 0.375
Pontuação: 4.696 (Otimista)

Um segundo experimento foi conduzido neste conjunto de *tweets*, desta vez utilizando duas técnicas de aprendizagem supervisionada: a primeira com um algoritmo de *Naive Bayes*(NB) e a segunda com *Support Vector Machine* (SVM). Ambos os algoritmos foram implementados utilizando a biblioteca *scikit-learn* em Python[11]. Em ambas as técnicas, o algoritmo recebe os exemplos que foram processados pelas etapas anteriores, além de usar como codificador de textos a representação matricial *bag-of-words* (BOW), indicando a frequência das palavras presentes em cada exemplo. O processo de treinamento foi realizado utilizando o método de validação cruzada *K-Fold*, particionando o conjunto de dados em $k=5$ subconjuntos mutuamente exclusivos de mesmo tamanho. O processo foi repetido k vezes, calculando-se a métrica de desempenho em cada subconjunto e obtendo a média dos valores como a métrica final de desempenho.

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

4.4.2 Desempenho no conjunto de notícias

Um processo semelhante ao realizado na subseção 4.4.2 foi executado no conjunto de notícias. Os exemplos passaram por uma etapa de pré-processamento para garantir a qualidade dos dados e facilitar a classificação. Posteriormente, foi criada mais uma versão, além das já descritas como versão original e versão lematizada, agora uma versão com stemização, conforme exemplificado na tabela 3.2. O processo de classificação utilizou uma abordagem de analisador lexical, com o objetivo de verificar a pontuação geral do texto alvo com base nos termos presentes também nos léxicos testados.

Todos os experimentos realizados tiverem como métodos de avaliação a utilização das métricas apresentas a seguir na seção 4.5.

4.5 Métricas de avaliação

Para avaliar o desempenho nas tarefas de classificação de sentimentos associados ao conjunto de *tweets* e notícias, serão utilizadas as seguintes métricas: acurácia (Equação 4-1), precisão (Equação 4-2), recall (sensibilidade) (Equação 4-3) e F1-score (Equação 4-4), conforme descritas em [3].

$$\text{Acurácia} = \frac{\text{Número de predições corretas}}{\text{Número total de amostras}} \tag{4-1}$$

$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos + Falsos positivos}} \tag{4-2}$$

$$\text{Recall} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos + Falsos negativos}} \tag{4-3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão + Recall}} \tag{4-4}$$

Os conceitos de verdadeiro positivo, falso positivo, falso negativo e falso negativo são compostas na avaliação das métricas de desempenho. Verdadeiro positivo (VP) representa a quantidade de amostras corretamente classificadas como positivas, enquanto falso positivo (FP) são amostras incorretamente classificadas como positivas. Verdadeiro negativo (VN) é o número de amostras corretamente classificadas como negativas, e falso negativo (FN) são amostras erroneamente classificadas como negativas.

Além dessas métricas, será utilizada mais uma medida para verificar a porcentagem de textos não classificados, conforme apresentada em [3] e demonstrado na Equação

4-5 a seguir.

$$\text{Não classificados} = \frac{\text{Quantidade não classificada}}{\text{Quantidade não classificada + Quantidade classificada}} \tag{4-5}$$

A quantidade não classificada refere-se ao número de mensagens não classificadas por não conterem palavras presentes no léxico ou devido à soma dos pesos dos termos presentes no texto alvo ter resultado zero. A quantidade classificada diz respeito ao número de mensagens que o vocabulário alcançou com as palavras presentes, a fim de realizar uma classificação.

Neste trabalho, serão empregadas as métricas previamente apresentadas, tanto em uma média simples (por exemplo, F1-macro), quanto em uma média ponderada pela quantidade de exemplos nos conjuntos de teste. Essa abordagem é adotada devido à configuração desbalanceada dos conjuntos de teste em notícias e *tweets*. Além disso, no próximo capítulo, serão realizados processos de balanceamento do conjunto com o intuito de identificar o desempenho dos léxicos em um ambiente de conjuntos de teste equilibrados.

CAPÍTULO 5

Resultados

Nesse capítulo, serão expostos os resultados dos experimentos realizados neste trabalho. Primeiramente, na seção 5.1, serão apresentados os resultados referentes à quantidade de termos decorrentes da expansão lexical proposta neste estudo. Em seguida, serão discutidos os desempenhos dos léxicos em tarefas de classificação de sentimentos em *tweets* na seção 5.2 e em notícias na seção 5.3, ambas sobre o domínio do mercado financeiro brasileiro. Por fim, será feita uma comparação entre o desempenho do método lexical e o método supervisionado na seção 5.4.

5.1 Expansão lexical

Os resultados da expansão lexical apresentados na Tabela 5.1 é derivado da proposta de construção apresentada anteriormente no capítulo 4. 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.

Tabela 5.1: *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

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 através da medida probabilística PMI após o léxico semente. Isso resultou em 253 palavras otimistas e 651 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 1087 palavras otimistas e 1210 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 1512 palavras e 1352 palavras pessimistas.

5.2 Desempenho do léxico na classificação de *tweets*

Para avaliar o desempenho das diferentes configurações de construção lexical, foram analisados os sentimentos de um conjunto de 3228 *tweets* categorizados manualmente, dos quais 2048 foram considerados otimistas para o mercado financeiro brasileiro e 1180 pessimistas. Devido à falta de equilíbrio no conjunto de testes, foi empregada a média ponderada na precisão, revocação *(recall)* e pontuação F1, bem como a acurácia geral e a porcentagem de *tweets* não classificados. Esses resultados podem ser visualizados na tabela 5.2 à seguir:

Tabela 5.2: *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	62%	66%	63%	65%	15.9%
S+PMI (lematizado)	67%	69%	67%	67%	6.2%
S+S/A+PMI	65%	68%	65%	67%	9.9%
S+S/A+PMI (lematizado)	**71%**	**72%**	**71%**	**71%**	2.8%
S+W2V+S/A+PMI	63%	66%	63%	65%	7.3%
S+W2V+S/A+PMI (lematizado)	69%	69%	69%	69%	**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. O objetivo era identificar o impacto dessa etapa no desempenho do analisador lexical em relação à soma dos termos de sentimento presentes no texto. Conforme observado na tabela 5.2, o uso da lematização influenciou diretamente no desempenho final.

O melhor resultado foi obtido na proposta S+S/A+PMI (lematizado), com uma acurácia, Recall e F1 de 71%, e uma precisão de 72%. Em contraste, sua versão não lematizada apresentou uma diferença negativa de até 6% na acurácia e 4% no F1. Isso se deve à facilidade de identificação dos termos uma vez normalizado e redução da 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 introduzidas nessa construção em comparação à melhor abordagem geral, que apresenta uma diferença de apenas 0,4%.

Na utilização de um analisador lexical para somar os pesos dos termos, a pontuação atribuída a cada palavra em relação ao domínio alvo tem um impacto direto no resultado. Na tabela 5.3, é realizada uma comparação entre o uso padrão de pesos e o uso de pesos personalizados, que atribui uma pontuação de +1 para os resultados positivos e -1 para os negativos como utilizados em [6, 18, 5], em comparação com o uso de pesos personalizados como feitos em [3, 15, 11, 20], que utiliza do processo de construção para gerar uma pontuação. O procedimento para construção da pontuação foi demonstrada no subcapítulo (4.01 (editar)).

Tabela 5.3: *Resultados com e sem pontuação dos termos do léxico (em %, melhores valores em negrito)*

Com score lexical					
Construção	Acurácia	Precisão	Recall	F1	Não classificadas
S+S/A+PMI	65%	68%	65%	67%	9.9%
S+S/A+PMI (lematizado)	**71%**	**72%**	**71%**	**71%**	**2.8%**
Sem score lexical					
S+S/A+PMI	55%	66%	55%	59%	24%
S+S/A+PMI (lematizado)	61%	71%	61%	62%	16%

Os experimentos realizados evidenciaram que o uso de termos sem um peso variável com base no domínio afeta diretamente os resultados. Foi possível observar, para a mesma construção do léxico, uma diferença de até 10% nas métricas ao comparar o melhor desempenho com o score lexical personalizado e sem essa personalização. Essa observação sugere que, em tarefas de classificação de sentimentos, algumas palavras possuem maior influência na decisão do que outras. Isso fica ainda mais evidente ao observar a taxa de textos não classificados, onde a melhor construção entre as abordagens com e sem pontuação personalizado resultou em uma não cobertura de 16% dos *tweets*, em contraste com apenas 2,8% do melhor resultado.

5.2.1 Desempenho no conjunto de teste balanceado

Considerando a atual desproporção de exemplos entre a classe otimista, com 2048 exemplos, e a classe pessimista, com 1180 exemplos, foi utilizado a técnica *Random Over Sample* para equilibrar as classes, conforme descrito no subcapítulo 4.1.

A avaliação do desempenho no conjunto de teste balanceado é apresentada na Tabela:

Tabela 5.4: *Classificação dos tweets após o uso da técnica Random OverSample, (em %, melhores valores em negrito)*

Construção	Acurácia	Precisão	Recall	F1	Não classificados
S+PMI	59%	65%	59%	61%	15.3%
S+PMI (lematizado)	62%	67%	62%	61%	6.6%
S+S/A+PMI	61%	67%	61%	62%	9.8%
S+S/A+PMI (lematizado)	**67%**	**71%**	**67%**	**67%**	3.0%
S+W2V+S/A+PMI	60%	64%	60%	61%	7.5%
S+W2V+S/A+PMI (lematizado)	65%	68%	65%	64%	**2.6%**

De forma semelhante ao que foi apresentado na Tabela 5.2, a configuração mais eficaz continua sendo S+S/A+PMI (lematizado), considerando a média simples de precisão, *Recall* e pontuação F1. Além disso, o léxico S+W2V+S/A+PMI (lematizado) obteve a maior cobertura nos *tweets*. Nesta avaliação, os resultados foram ponderados igualmente para ambas as categorias, em contraste com a primeira avaliação apresentada nesse capítulo, na qual foi dada uma maior importância aos acertos das classificações otimistas devido à média ponderada pela quantidade de exemplos.

A comparação entre as matrizes de confusão da abordagem com balanceamento de dados e sem balanceamento é apresentada nas Tabelas da 5.5:

Tabela 5.5: *Matrizes de Confusão da configuração S+W2V+S/A+PMI (lematizado).*

Conjunto desbalanceado			
	Otimista	Pessimista	Total
Otimista	1729	287	2048
Pessimista	987	1017	2048
Conjunto balanceado			
	Otimista	Pessimista	Total
Otimista	1729	287	2048
Pessimista	581	575	1180

A proporção de falsos otimistas e verdadeiros pessimistas permanece semelhante, com a Tabela 5.5 (conjunto balanceado) apresentando uma sensibilidade de 50,7%, em contraste com a Tabela (conjunto desbalanceado) que apresenta uma sensibilidade de 49,7%. Sendo essa diferença de 1% provavelmente devido à forma como a técnica *Random OverSample* estende o conjunto minoritário, selecionando aleatoriamente exemplos existentes e adicionando-os ao conjunto final até que a classe pessimista tenha o mesmo tamanho que a classe otimista. A classificação dos *tweets* otimistas permanece a mesma para ambos os experimentos, obtendo uma sensibilidade 85,7%. É importante ressaltar que, além das classificações otimistas e pessimistas, há também a possibilidade de um

exemplo não ser classificado devido à falta de cobertura do léxico em relação às palavras do texto alvo ou à soma das pontuações dos sentimentos das palavras resultar em zero. Essa diferença também pode ser observada na coluna "Não classificados"na Tabela 5.4.

5.3 Desempenho do léxico na classificação de notícias

O uso do léxico não se limita apenas 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 5.2, foi realizada uma avaliação do desempenho na classificação de um conjunto de notícias sobre o Mercado Financeiro Brasileiro, produzido por [6]. 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).

Tabela 5.6: *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) [6]	57.1%	57.4%
(2) Baseline (com *stemming*) [6]	58.2%	58.8%
(3) S+S/A+PMI	65%	65%
(4) S+S/A+PMI (com steminização)	64%	64%
(5) S+S/A+PMI (com lematização)	**68%**	**68%**

A tabela 5.6 apresenta uma comparação entre os diferentes métodos, incluindo a linha de base original *(baseline)* e variações do léxico com diferentes técnicas de processamento de texto. A linha de base original obteve uma acurácia de 57,1% e um valor de F1 de 57,4%. Ao aplicar a técnica de *stemming* no pré-processamento dos termos, a acurácia e o valor de F1 melhoraram ligeiramente para 58,2% e 58,8%, respectivamente. No entanto, o uso do léxico proposto mostrou resultados ainda mais significativos. O método S+S/A+PMI obteve como média ponderada pela proporção do conjunto de teste uma acurácia e um valor de F1 de 65%. Quando combinado com a técnica de *stemming*, os resultados se mantiveram consistentes, com uma acurácia e um valor de F1 de 64%. No entanto, a utilização da técnica de lematização resultou em uma melhoria adicional, com uma acurácia e um valor de F1 de 68%.

O uso do léxico em conjunto com a lematização também se revelou uma abordagem eficaz na classificação de notícias do mercado financeiro brasileiro, resultando em melhorias significativas na precisão e cobertura dos sentimentos expressos. A inclusão de

uma pontuação personalizada com base na frequência dos termos e o emprego de técnicas de processamento de texto, contribuíram para a precisão e relevância na classificação dos sentimentos predominantes nos exemplos. Apesar dos desafios inerentes à natureza complexa da classificação de notícias, os resultados obtidos confirmam a viabilidade e utilidade do léxico proposto como uma abordagem eficiente e rápida, podendo fornecer percepções valiosos para o domínio alvo que nesse contexto se refere a investidores, analistas e demais interessados no campo financeiro brasileiro.

5.4 Comparação com método supervisionado

Uma comparação foi realizada entre o desempenho do léxico de melhor resultado demonstrado anteriormente e um método supervisionado utilizando validação cruzada *K-Fold*, com K igual a 5, em um subconjunto de 2000 *tweets* do conjunto de teste realizado no subcapítulo 5.2, selecionado aleatoriamente 1000 *tweets* considerados otimistas e 1000 pessimistas. A Tabela 5.7 apresenta os resultados dessa comparação.

Tabela 5.7: *Comparação com método supervisionado treinando usando validação cruzada K-Fold K=5 em um subconjunto de 2000 tweets.*

Método	F1 macro
SVM	$0,73 \pm 0,02$
NB	$0,73 \pm 0,02$
Léxico	$0,67$

Observou-se que o método de aprendizado de máquina SVM (Support Vector Machine) alcançou um valor médio de F1 de $0,73$ com um desvio padrão de $0,02$, enquanto o método de Naive Bayes (NB) também obteve um valor médio de F1 de $0,73$ com um desvio padrão de $\pm 0,02$. Por outro lado, o léxico proposto alcançou um valor de F1 de $0,67$. Essa comparação indica que os métodos supervisionados tiveram um desempenho ligeiramente superior em relação ao léxico proposto na classificação de *tweets* relacionados ao mercado financeiro brasileiro. Esse fato também é demonstrado em outros trabalhos, como [3, 6, 5]. Além de métodos tradicionais de AM, técnicas de aprendizagem profundo como redes neurais apresentam vantagens no desempenho ainda mais significativas como demonstrado em [7]. Essas arquiteturas têm apresentado uma boa capacidade de generalizar a partir de pequenos conjuntos de dados, desde que sejam representativos e com uma boa qualidade de rotulagem, assim podendo por extrapolação extrair recursos de texto mais ricos[16].

Na abordagem lexical, a variação dos resultados está ligada à formulação do léxico utilizado e seu domínio. Exemplos disso são o estudo de [10], que cobriu 41% dos termos de vocabulário conhecidos em triagens de câncer de mama, e [18], que alcançou

91% de precisão na classificação de textos do cotidiano sem domínio específico. Já [20] obteve 69,6% de acurácia na análise de sentimentos de comentários de filmes. Resultados semelhantes ocorrem no contexto financeiro, como a acurácia de 70% em publicações sobre o sistema financeiro americano por [5] e a pontuação F1 de 58,2% obtido por [6].

CAPÍTULO 6

Conclusão

Este trabalho apresentou o desenvolvimento de estratégias para a criação de léxicos em domínios específicos visando identificar possíveis configurações para a construção desses vocabulários. Para isso, utilizou-se como exemplo o campo do Mercado Financeiro Brasileiro (MFB), 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.

Foram apresentadas três formas de construção lexical, cada uma delas criada uma versão a qual utiliza uma estratégia de pré-processamento das palavras chamado lematização. Isso resultou em um total de seis configurações finais para a construção de léxicos no contexto MFB. Foram conduzidos experimentos tanto para a tarefa Análise de Sentimentos (AS) em mensagens curtas, como os *tweets* relacionados ao mercado brasileiro, quanto para a AS em grandes textos, como notícias sobre o mesmo domínio. Os melhores resultados em ambas as tarefas foram obtidos utilizando a proposta de construção do léxico através da configuração que emprega o pre-processamento de lematização das palavras, em conjunto com o uso do conjunto semente, que foi estendido pela utilização da busca por sinônimos e antônimos das palavras, e finalizado pela aplicação da medida probabilística *Pointwise Mutual Information* (PMI), **S+S/A+PMI (com lematização)**. Essa configuração obteve um desempenho ponderado F1-Score de 67% na classificação dos *tweets* e 68% na classificação das notícias, superando o *baseline* apresentado em [6] para as notícias.

Em adição, foi comparado o uso da abordagem lexical para a AS de *tweets* com o uso de uma abordagem de *Machine Learning* supervisionado. Os resultados obtidos foram um F1-macro de 67% para a abordagem lexical e 73% para o modelo de *Machine Learning* supervisionado. Embora a abordagem lexical não tenha superado o uso de ML, é importante ressaltar que a aplicação de ML requer um conjunto representativo de dados previamente catalogados para o treinamento desse modelo supervisionado, o que demanda dedicação na obtenção desses exemplos. Em contraste, a abordagem lexical, com uma pequena lista de 75 palavras consideradas otimistas e pessimistas para o mercado financeiro, e o processo de extensão descrito neste trabalho, obteve um

desempenho com uma diferença de apenas 6%.

Dessa forma, este estudo em questão introduziu e comparou distintas abordagens para a expansão automática de léxicos em língua portuguesa, focando particularmente na aplicação ao cenário do mercado financeiro brasileiro. 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. Dessa forma, torna possível uma análise preliminar ágil por meio de uma rápida extração de um conjunto inicial de palavras-chave específicas para o domínio em questão. Mesmo que, apesar de o uso de aprendizado supervisionado proporcionar resultados mais consistentes, ainda requer a aquisição antecipada de quantidades significativas de dados já rotulados.

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 concepção de um algoritmo capaz de estabelecer um limiar numérico na métrica PMI para a atribuição de rótulos positivos ou negativos, buscando otimizar as métricas de avaliação.

Referências

[1] AGRAWAL, S.; SANYAL, R.; SANYAL, S. **Hybrid method for automatic extraction of multiword expressions**. *International Journal of Engineering and Technology(UAE)*, 7:33–38, 2018.

[2] BIRD, S. **Nltk: The natural language toolkit**, 2006.

[3] BOS, T.; FRASINCAR, F. **Automatically building financial sentiment lexicons while accounting for negation**. *Cognitive Computation*, 14:442–460, 2022.

[4] CAROSIA, A. E.; COELHO, G. P.; SILVA, A. E. **Analyzing the brazilian financial market through portuguese sentiment analysis in social media**. *Applied Artificial Intelligence*, 34:1–19, 1 2020.

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

[6] DE O. CAROSIA, A. E. **Sentiment analysis applied to news from the brazilian stock market**. *IEEE Latin America Transactions*, 20:512–518, 2022.

[7] DE OLIVEIRA CAROSIA, A. E.; COELHO, G. P.; DA SILVA, A. E. A. **Investment strategies applied to the brazilian stock market: A methodology based on sentiment analysis with deep learning**. *Expert Systems with Applications*, 184:115470, 2021.

[8] FERNANDES, D. S. A.; FERNANDES, M. G. C.; BORGES, G. A.; SOARES, F. A. **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)*, p. 2626–2632, Oct 2019.

[9] HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. **Portuguese word embeddings: Evaluating on word analogies and natural language tasks**. 8 2017.

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

[11] LOSADA, D. E.; GAMALLO, P. **Evaluating and improving lexical resources for detecting signs of depression in text**. *Language Resources and Evaluation*, 54:1–24, 2020.

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

[13] 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*, 10:140–150, 2020.

[14] MPOULI, S.; BEIGBEDER, M.; LARGERON, C. **Lexifield: a system for the automatic building of lexicons by semantic expansion of short word lists**. *Knowledge and Information Systems*, 62:3181–3201, 2020.

[15] OLIVEIRA, N.; CORTEZ, P.; AREAL, N. **Stock market sentiment lexicon acquisition using microblogging data and statistical measures**. *Decision Support Systems*, 85:62–73, 2016.

[16] PEREIRA, D. A. **A survey of sentiment analysis in the portuguese language**. *Artificial Intelligence Review*, 54:1087–1115, 2 2021.

[17] ROYYAN, A. R.; SETIAWAN, E. B. **Feature expansion word2vec for sentiment analysis of public policy in twitter**. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6:78–84, 2 2022.

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

[19] SMYWIŃSKI-POHL, A.; LASOCKI, K.; WRÓBEL, K.; STRZAŁA, M. **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*, p. 234–238, 2019.

[20] WANG, Y.; YIN, F.; LIU, J.; TOSATO, M. **Automatic construction of domain sentiment lexicon for semantic disambiguation**. *Multimedia Tools and Applications*, 79:22355–22373, 2020.