Universidade Federal de Goiás
Escola de Engenharia Elétrica, Mecânica e de Computação

Heinrych Matheus Gomes Andrade

Desenvolvimento de *lexicon* no contexto de violência contra a mulher em língua portuguesa para automação de classificação de tweets.

Goiânia

2023

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 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 dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) 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 Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as):

Heinrych Matheus Gomes Andrade

Título do trabalho:

*Desenvolvimento de lexicon no contexto de violência contra a mulher em língua portuguesa para automação de classificação de tweet*s

2. 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)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. 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.

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

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 24/08/2023, às 16:38, 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 **Heinrych Matheus Gomes Andrade**, **Discente**, em 24/08/2023, às 16:40, 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 **3991697** e o código CRC **182F8142**.

Referência: Processo nº 23070.024561/2023-71 SEI nº 3991697

Heinrych Matheus Gomes Andrade

Desenvolvimento de *lexicon* no contexto de violência contra a mulher em língua portuguesa para automação de classificação de tweets.

Trabalho de conclusão de curso apresentado na Escola de Engenharia Elétrica, Mecânica e de Computação como requisito para a conclusão do curso de Engenharia de Computação e obtenção do título de Engenheiro de Computação

Universidade Federal de Goiás – UFG Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)

Orientadora: Profa. Dra. 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.

Andrade, Heinrych Matheus Gomes
Desenvolvimento de lexicon no contexto de violência contra a mulher em língua portuguesa para automação de classificação de tweets [manuscrito] / Heinrych Matheus Gomes Andrade. - 2023.
XV, 15 f.: il.

Orientador: Prof. Deborah Silva Alves Fernandes.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), , Goiânia, 2023.
Bibliografia.
Inclui gráfico, tabelas.

1. análise de sentimentos. 2. TF-IDF. 3. violencia contra a mulher.. 4. PMI. 5. lexicon. I. Fernandes, Deborah Silva Alves, orient. II. Título.

CDU 519.25

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

ATA DE AVALIAÇÃO DE PROJETO FINAL CURSO

( ) Eng Elétrica	( ) Eng Mecânica	(X) Eng Computação PFC 1 ( ) PFC 2 (X)

Título do Trabalho

Desenvolvimento de lexicon no contexto de violência contra a mulher em língua portuguesa para automação de classificação de tweets

Banca Avaliadora

Membro 1	Deborah Silva Alves Fernandes
Membro 2	Sandrerley Ramos Pires
Membro 3	Nádia Félix Felipe da Silva

Data da Defesa

24/08/2023

Discentes

Matrícula	Nome
201802726	Heinrych Matheus Gomes Andrade

NOTAS										
	Membro 1			**Membro 2**			**Membro 3**			
Matrícula	**NPT**	**NTE**	**NAA**	**NPT**	**NTE**	**NAA**	**NPT**	**NTE**	**NAA**	**Média***
201802726	10	10	10	10	10	10	10	10	10	10,0

NPT – Nota plano de trabalho;
NTE – Nota do trabalho escrito;
NAA – Nota de apresentação e arguição

Para Eng. Elétrica, Mecânica e PFC2 da Eng. Da Computação: NF = 0,1 x NPT + 0,45xNTE + 0,45 x NAA

Para PFC1 da Eng. Da Computação: NF = 0,3 x NPT + 0,7 x NAA

* A aprovação do(s) aluno(s) está condicionada à apresentação do trabalho final ao orientador com todas as correções sugeridas pela banca.

Observações:

Preencher com modificações solicitadas, caso existam. Em caso de reprovação, informar a justificativa.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 24/08/2023, às 10:09, 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 **Sandrerley Ramos Pires**, **Professor do Magistério Superior**, em 24/08/2023, às 16:13, 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 **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 24/08/2023, às 16:26, 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 **3985723** e o código CRC **E5980C69**.

Referência: Processo nº 23070.024561/2023-71 SEI nº 3985723

Desenvolvimento de lexicon no contexto de violência contra a mulher em língua portuguesa para automação de classificação de tweets

Heinrych Matheus Gomes Andrade[1], graduando em Engenharia de Computação. [1]EMC/UFG Deborah Alves Fernandes[2], Professora Associada. [2]INF/UFG E-mails: heinrych@discente.ufg.br[1], deborah.fernandes@ufg.br[2]

***Resumo*—Este artigo discute a análise de sentimentos em tweets relacionados à violência contra a mulher em língua portuguesa por meio de abordagens de expansão léxica. O processo começa com a criação de um *lexicon* semente composto por unigramas, bigramas e trigramas, originados de artigos acadêmicos e tweets sobre violência contra a mulher. Em seguida, o corpus é classificado usando esse *lexicon*. Várias expansões são realizadas sobre este conjunto inicial: por meio de PMI com unigramas, bigramas e trigramas, através de TF-IDF com as palavras mais frequentes, via dicionário de sinônimos, e finalmente por meio de uma LSTM treinada com um conjunto de dados classificados. A eficácia de cada expansão do *lexicon* é avaliada por meio da classificação do conjunto de dados utilizando a métrica PMI. Além disso, é implementado o método de agrupamento "Agglomerative Clustering"para a classificação final dos tweets, baseando-se nas pontuações PMI calculadas a partir do *lexicon* expandido. Por fim, é realizada uma comparação entre a eficácia do agrupamento com e sem o uso das pontuações.**

***Palavras-chave*—análise de sentimentos, classificação de sentimentos, tweets, *lexicon* semente, Pontuação de Informação Mútua (PMI), Agglomerative Clustering, TF-IDF, LSTM, violencia contra a mulher.**

***Abstract*—This article discusses sentiment analysis in tweets related to violence against women in Portuguese through lexical expansion approaches. The process begins with the creation of a seed lexicon composed of unigrams, bigrams, and trigrams derived from academic articles and tweets about violence against women. Subsequently, the corpus is classified using this lexicon. Several expansions are performed on this initial set: through PMI with unigrams, bigrams, and trigrams, using TF-IDF with the most frequent words, via a synonym dictionary, and finally through an LSTM trained with classified data. The effectiveness of each lexicon expansion is assessed by classifying the dataset using the PMI metric. Additionally, the "Agglomerative Clustering"method is implemented for the final classification of tweets, based on PMI scores calculated from the expanded lexicon. Finally, a comparison is conducted between the effectiveness of clustering with and without the use of PMI scores.**

***Index Terms*—sentiment analysis, sentiment classification, tweets, seed lexicon, Pointwise Mutual Information (PMI), Agglomerative Clustering, TF-IDF, LSTM, violence against women.**

1. Introdução

Este estudo está alinhado aos Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU) [1] e se concentra na análise da violência virtual contra mulheres no contexto do Twitter. Ele se insere na área de análise de sentimentos e classificação de textos no processamento de linguagem natural. O enfoque principal é compreender e quantificar a prevalência e percepção desse problema, contribuindo para a literatura acadêmica e fornecendo informações relevantes para a formulação de políticas que promovam a igualdade de gênero (ODS 5).

Para abordar essa questão, o estudo propõe uma metodologia de classificação de sentimentos expressos em tweets em língua portuguesa. Nessa abordagem, é desenvolvido um *lexicon* personalizado, explorando o conceito de PMI (Pointwise Mutual Information - Informação Mútua Pontual) com o intuito de aprimorar a precisão da classificação. O cálculo desse índice é realizado entre palavras consideradas positivas ou negativas e as palavras presentes nos textos dos tweets. Esse cálculo resulta em pontuações que refletem a polaridade dos sentimentos expressos.

O processo se inicia com a construção de um *lexicon* semente composto por unigramas, bigramas e trigramas, derivados de artigos acadêmicos e tweets sobre a violência contra a mulher. Seguidamente, realizam-se várias expansões deste via TF-IDF (*Term Frequency-Inverse Document Frequency*), PMI, um dicionário de sinônimos e uma LSTM (*Long Short-Term Memory*), cada uma proporcionando uma dimensão adicional para capturar mais nuances e variações linguísticas para uma melhor classificação dos sentimentos positivos e negativos.

A validação da abordagem proposta se dá através da comparação das previsões do modelo com classificações manuais, utilizando um conjunto de dados composto por tweets em língua portuguesa que discutem a temática da violência contra a mulher. Adicionalmente, a avaliação da eficácia da abordagem é ampliada por meio da aplicação do método de agrupamento hierárquico *Agglomerative Clustering* na classificação adicional dos tweets. Esse agrupamento é realizado com base nas pontuações PMI calculadas a partir do *lexicon* expandido, tanto considerando essas pontuações como não considerando.

Os detalhes dos métodos empregados, os resultados da classificação e as discussões serão abordados nas seções seguintes deste estudo. Inicialmente, serão apresentados os trabalhos relacionados, seguidos pela fundamentação teórica das ferramentas utilizadas. Em seguida, será descrita a metodologia empregada, por fim a conclusão do estudo.

2. Trabalhos Relacionados

Esta seção apresenta trabalhos agrupados conforme as abordagens empregadas na análise de sentimentos. Na primeira

seção, 2.1, as pesquisas investigam a aplicação de técnicas de aprendizado de maquina em diferentes cenários para compreensão de sentimentos. Na segunda seção, 2.2, o foco está na expansão e criação de *lexicons* de sentimentos em vários idiomas, oferecendo uma abordagem fundamentada em regras.

A Tabela I apresenta um resumo dos métodos, conjuntos de dados, pré-processamentos, técnicas de classificação e precisões obtidas dos estudos. Ela oferece uma visão comparativa dos diferentes métodos e técnicas utilizados para lidar com o desafio da análise de sentimentos e suas aplicações em diversas áreas, como o abuso doméstico, a bolsa de valores e a variação linguística do Twitter. A precisão dos métodos varia conforme a complexidade do problema e o tipo de dado analisado, sendo evidenciado que a análise de sentimentos é uma tarefa complexa que requer um cuidadoso pré-processamento dos dados e a escolha apropriada da técnica de classificação.

2.1 Abordagens de Análise de Sentimentos com Métodos de Aprendizado de Máquina

A aquisição de dados anotados em domínios variados pode ser um desafio. Diante disso, [2] propôs a expansão de um *lexicon* de opiniões geral com palavras específicas de cada domínio. Dois conjuntos de dados foram selecionados para o estudo: uma coleção com 56 tópicos do Microblog Track 2011 e a coleção Stanford Sentiment140. Esses conjuntos foram expandidos com base em emoticons, mesclados e filtrados para manter apenas os tweets que continham emoticons. Em seguida, eles foram agrupados em conjuntos resultando em um dataset com 22.782 tweets. Para a expansão baseada em tópicos, foi utilizada a técnica do PMI. Por fim, para a classificação dos sentimentos, foram empregados os algoritmos *Naive Bayes* Multinomial (NBM) e *SentiStrength* (SS). O SS obteve o melhor resultado, alcançando uma precisão de 88,29% para a expansão baseada em tópicos.

O estudo de [3] abordou a detecção de discurso, especialmente violência, diante do grande volume de dados disponíveis. Superando limitações de conjuntos de dados restritos, os pesquisadores expandiram o conjunto OLID (14.100 tweets em inglês) [4] com o conjunto SOLID (9 milhões de tweets), ambos para identificação de linguagem ofensiva. Aquisição via API de Streaming do Twitter e Twython, com o auxílio das palavras de paradas mais frequentes na lingua inglesa (por exemplo, *the*, *of* e, *to*) para assegurar a aleatoriedade dos tweets coletados. O pré-processamento incluiu identificação de idioma e remoção de links. Foram obtidos 12 milhões de tweets, 75% rotulados por co-treinamento. Modelos como BERT (*Bidirectional Encoder Representations from Transformers*), LSTM, PMI e FT foram testados, treinados com OLID. Resultados indicaram que o modelo BERT alcançou 92,3% de precisão no SOLID, superando 92,2% anterior.

Embora a análise de sentimentos seja amplamente aplicada a conteúdos em inglês, a disponibilidade de recursos para o idioma urdu é limitada, como salientado por [5]. Vale ressaltar que o urdu é a língua nacional oficial do Paquistão e é amplamente falado no sul da Ásia [6]. Diante desse cenário, os autores propuseram um conjunto de dados que abrange 1372 expressões no dialeto urdu. Com o objetivo de criar um conjunto de dados mais abrangente, eles empregaram a extração multimodal, utilizando recursos textuais, áudios e vídeos. Para a coleta de dados, foram utilizadas diversas ferramentas e técnicas, como OPENSMILE, 3D-CNN e camadas de BLSTM (Long Short-Term Memory), sendo cada camada uma combinação de incorporações de palavras pré-treinadas com 300 dimensões do modelo fastText. Na etapa final, foram comparados os tipos de dados individualmente e suas combinações. Os resultados mostraram uma precisão de 84,32% para os algoritmos modais e 95,35% para os algoritmos que utilizaram recursos multimodais.

Tendo em vista o crescimento significativo das pesquisas voltadas para a análise de sentimentos no Twitter, [7] apresentam um estudo detalhado de análise de sentimentos baseado em regressão ordinal, aplicando técnicas de aprendizado de máquina. O conjunto de dados foi coletado utilizando a API do Twitter, e os tweets foram anotados como positivos ou negativos. Este conjunto de dados é disponibilizado pelo recurso corpora do Natural Language Toolkit (NLTK) e consiste em 10 mil postagens. No processo de pré-processamento dos dados, foram removidos hiperlinks, stopwords e erros ortográficos. Além disso, as letras foram convertidas para minúsculas, as palavras foram reduzidas ao seu radical, os termos foram tokenizados e os emoticons foram substituídos pelo sentimento correspondente. Em seguida, foi realizada a vetorização TF-IDF para a extração de recursos. Na etapa de classificação, foram utilizados algoritmos como regressão logística multinomial (SoftMax), Support Vector Regression (SVR), Árvores de Decisão (DTs) e Random Forest (RF). As classes consideradas foram "muito negativo", "negativo", "neutro", "positivo"e "muito positivo", sendo que o método Árvore de Decisão obteve a melhor precisão, com 91,81% de acurácia.

Considerando a crescente conscientização em relação às questões públicas, como os direitos das mulheres, a análise de sentimento desempenha um papel importante no reconhecimento das atitudes das pessoas. [8] abordam a análise de sentimentos aplicada à violência contra as mulheres na língua árabe. Para este estudo, foram coletados cinco mil tweets por meio da API do Twitter, provenientes de países árabes. Após a coleta dos dados, foi realizada a limpeza, que envolveu a remoção de links, hashtags, emojis e símbolos ruidosos. Como resultado desse processo, o conjunto de dados foi reduzido para três mil e setecentos tweets. Em seguida, ocorreu a rotulação manual dos dados, seguida pelo pré-processamento. Durante o pré-processamento, os dados foram tokenizados, as stopwords foram removidas e os N-gramas ((Sequências contíguas de N elementos) foram gerados. Para a classificação dos dados, foram empregadas várias ferramentas, como SVM, *K-Nearest Neighbors* (KNN), Decision Trees e *Naive Bayes*. Ao comparar a precisão dos diferentes algoritmos, observou-se que o SVM obteve o melhor desempenho, com uma precisão de 78,25%.

Nesta pesquisa, [9] tem como objetivo desenvolver um sistema de detecção de ações negativas cometidas por seres humanos. Os dados foram coletados de um sistema de relatórios online chamado "maps.safecity", que abrange informações relacionadas ao COVID-19, violência doméstica, assédio online, perseguição, fotografias indesejadas, vaias/assobios, agressão se-

xual, entre outros. Durante a etapa de limpeza dos dados, foram aplicadas as seguintes transformações: conversão para letras minúsculas, remoção de pontuações, stopwords, palavras muito específicas, tags, emojis e emoticons. No pré-processamento, o texto foi tokenizado e foi realizada a lematização e o *stemming* das palavras. Para a classificação dos dados, foi inicialmente empregado o método TF-IDF para identificar os termos mais relevantes nas sentenças. Em seguida, foram aplicados diversos modelos de classificação, sendo que o classificador SGD obteve os melhores resultados com uma precisão de 81%, em segundo lugar, os classificadores Linear SVC e SVC alcançaram uma precisão de 79%.

Por intermédio da aplicação de técnicas de mineração de dados, [10] empregou técnicas de mineração de dados para identificar camadas ocultas e estruturas temáticas em tweets relacionados à violência contra a mulher. A coleta de dados foi realizada por meio da API do Twitter, usando a palavra-chave "violência doméstica", resultando em um total de 322.863 tweets coletados. No pré-processamento, os dados passaram por uma etapa de limpeza, onde símbolos, referências a usuários, URLs e caracteres não ingleses foram removidos. As mensagens foram convertidas em bigramas (sequências de duas palavras) usando a função CountVectorizer3. Em seguida, foi definido o número de tópicos e aplicado o algoritmo LDA para a redução da dimensionalidade, resultando em uma matriz documento-termo. A partir dessa matriz, foram identificados os termos mais frequentes nos tweets relacionados à violência contra a mulher, tais como "greg hardy", "hardy girlfriend", "dallas cowboys", entre outros. Esses resultados demonstram a viabilidade da mineração de dados no contexto da agressão direcionada ao gênero.

O estudo de [11] abordou a categorização automatizada e análise de assédio sexual no fórum SafeCity, em contexto de crescimento da hashtag MeToo. Foram coletadas 9.892 histórias com detalhes de incidentes, localização e tipos de assédio. A partir desses dados, três tags principais foram consideradas: toque/olhar/comentar. Foram utilizados dois métodos de avaliação dos dados: modelos de rótulo único, utilizando a métrica de precisão, e modelos de multi-rótulo, utilizando a métrica de escore de Hamming. Três formas de classificação foram aplicadas, utilizando os métodos *redes neurais convolucionais* (CNN), LSTM e uma combinação de CNN e LSTM. A composição obteve o melhor resultado com multi-rótulo de 86,5%. Além disso, foram empregadas ferramentas de visualização, como análise LIMI, incluindo mapa de calor, clusterização e visualização t-SNE.

Diante do aumento de ataques de assédio online, principalmente direcionados a mulheres devido ao anonimato proporcionado pelas redes sociais, [12] realizaram um experimento visando detectar mensagens misóginas no Twitter por meio de análise de sentimento. O estudo utilizou o corpus MisoCorpus-2020, que engloba subconjuntos como Violência Contra Mulheres Relevantes (VARW), Espanhol Europeu vs. Latino-americano (SELA) e Dominância, Assédio Sexual e Tipo Estereótipo (DDSS). Após remover tweets fora do domínio, o pré-processamento incluiu a eliminação de duplicações de tabulações, correção de erros ortográficos, aplicação de incorporações de sentenças e identificação de tags de palavras. O conjunto final continha 7.682 tweets, avaliados com os algoritmos de classificação RF, Sequential Minimal Optimization (SMO) e *Linear Support Vector Machine* (LSVM). O algoritmo SMO se destacou com a maior precisão no conjunto MisoCorpus-2020, alcançando 85,175% de acurácia.

Explorando o tópico da violência doméstica e seus custos para estatísticas, [13] realizou pesquisa qualitativa sobre abuso doméstico. Utilizando dados do reddit.com, abrangendo termos ligados à violência, raiva e ansiedade, buscava analisar seu impacto nas vítimas. A coleta envolveu 21.133 submissões e 349.277 comentários, passando por etapas de pré-processamento, como remoção de subreddits sem comentários. Com N-gramas separados em unigramas (palavras individuais no texto), bigramas (pares de palavras consecutivas) e trigramas (conjuntos de três palavras consecutivas), aplicou-se revisão sistemática para rótulos preditivos, os dados foram colocados em caixa baixa, normalizados, reduzidos aos radical e as stopwords foram removidas. O modelo SVM Linear alcançou 90% de acurácia na classificação dos dados iniciais. Ampliando o conjunto para 1.334 submissões mais comentários, uso de representações vetoriais no spaCy resultou em 94% de precisão após classificação

Segundo a Hipótese do Mercado Eficiente, eventos e notícias impactam o mercado financeiro de forma significativa. [14] exploraram a análise de sentimento nesse contexto, focando na avaliação de tweets sobre uma empresa automobilística em português. Especialistas classificaram um conjunto de 2.132 notícias e 11.027 tweets. Os dados passaram por limpeza, removendo símbolos, links, emoticons e stopwords com NLTK. Com *stemming*, palavras foram reduzidas ao radical e convertidas em vetores por bag of words. O Sentiment Classifier, incluindo *Naive Bayes*, *Maximum Entropy* (ME), SVM e *Multilayer perceptron* (MLP), classificou sentimentos, destacando *Naive Bayes* com 80,4% de precisão. Para o segundo módulo, usando MLP, 16.352 tweets não rotulados foram analisados diariamente, vinculando sentimentos a movimentos no mercado. A perspectiva de retweets teve a melhor acurácia, alcançando 60,0%.

O trabalho de [15] introduz o modelo *Aspect-Dependent Convolutional Neural Network* (ADeCNN) para análise de sentimento a nível de aspecto, superando limitações da abordagem tradicional de polaridade positiva/negativa. ADeCNN utiliza Redes Neurais Convolucionais Deformáveis (DeCNN) com Bi-LSTM (*Bidirectional Long Short-Term Memory*) e atenção ao nível de frase, lidando com distâncias entre palavras. O modelo GMemN2N (*Gated Memory Networks for Neural Machine Translation*) é empregado para gerar pesos de atenção distintos com base no alvo, útil para extrair recursos de sentimentos de forma contextual. Testado no conjunto SemEval com dados do Twitter, restaurantes e laptops (18.547, 4.480 e 2.958 exemplos), ADeCNN supera modelos como Bi-LSTM, TD-LSTM (*Target-Dependent Long Short-Term Memory*) e IAN (Classificação em Nível de Aspecto), destacando-se no conjunto de restaurantes com 84,03% de precisão.

Considerando a negligência frequente da correlação entre o contexto global e a polaridade do sentimento em muitos estudos, [16] propõem o modelo de aprendizado multilíngue *Local and Global Context Fusion* (LGCF), que aborda de forma interativa

os contextos local e global. Usando seis conjuntos de dados do SemEval de diferentes domínios (laptop14, restaurante14, Twitter, camiseta e televisão), o LGCF engloba camadas de incorporação de entrada, enfoque nos contextos local e global, além de uma camada de aprendizado interativo de recursos. No pré-processamento, empregaram o modelo pré-treinado BERT para mapear palavras. O foco no contexto local compreende autoatenção de várias cabeças (MHSA), CDM (máscara dinâmica de recursos de contexto) e peso dinâmico de recursos de contexto (CDW). O contexto global é construído com unidades recorrentes fechadas bidirecionais (BGRU), CNN e normalização de camadas (LN). Dentre os modelos comparados, o LGCF-CDM-CDW apresentou os resultados mais promissores, com a maior acurácia de 93,86% obtida no conjunto de dados "camiseta".

Os autores [17] abordaram os desafios relacionados ao tempo necessário para a classificação de sentimentos em tweets, o que representa um obstáculo para empresas que buscam automatizar esses processos. Para resolver o problema de desempenho, propuseram um novo método que utilizava uma arquitetura de aprendizado profundo chamada ULMFiT (Universal Language Model Fine-tuning) em conjunto com a técnica de *Support Vector Machine* (SVM). Os pesquisadores adotaram como base de dados o Wikitext-103, que contém 103 milhões de palavras. Inicialmente, aplicaram filtros, particionamento e remoção de stopwords nos dados não estruturados. Em seguida, realizaram a expansão dos dados, removendo espaços e tabulações, e tokenizando-os com a biblioteca *spaCy*. As palavras foram enumeradas, e aquelas com baixa frequência foram removidas. O próximo passo foi o pré-processamento dos dados utilizando o modelo de linguagem AWD-LSTM, que ajudou a reduzir a dimensionalidade da base de dados. Em seguida, foi aplicado o ajuste fino ULMFiT em conjunto com SVM. Os testes foram realizados em três conjuntos de dados: Twitter US Airlines, IMDB e Debate GOP, contendo, respectivamente, 14.640, 50.000 e 13.871 tweets. Os resultados obtidos foram uma precisão de 98,78%, 99,71% e 95,78% para cada conjunto de dados, respectivamente.

2.2 Expansão e Aplicações de Lexicons *de Opinião em Análise de Sentimentos*

O estudo de [18] apresentou uma abordagem supervisionada para expandir um *lexicon* de opinião, usando dois conjuntos de dados: Edinburgh corpus (ED) e Stanford Sentiment corpus (STS). Após filtrar tweets em inglês, o ED continha 2.138.622 tweets e o STS tinha 1.600.000. Os tweets anotados foram pré-processados, incluindo conversão para minúsculas, tokenização e marcação de part-of-speech (POS) com TweetNLP. Em dois experimentos, uma série temporal foi calculada para 10 mil palavras identificadas com tags POS nos conjuntos de dados. A série *Semantic Orientation* (SO) e *Stochastic Gradient Descent* (SGD) foram usadas para aprender a relação entre palavras e sentimento, com SO obtendo melhor precisão no STS, chegando a 77% de acurácia. No segundo experimento, o *lexicon* foi expandido com classificadores treinados para polaridade de cada tweet, usando SentiWordNet e regressão logística. A métrica *Area Under the Curve* (AUC) mostrou que a melhor precisão foi alcançada com o *lexicon* expandido do STS, atingindo 84% de classificação.

A pesquisa de [19] propôs um método de análise de sentimentos em língua tâmil, usando regras para classificar sentimentos como positivos e negativos. O corpus *UJ_Tamil*, proveniente de várias fontes como noolaham, Wikipédia, Twitter, Facebook e sites de resenhas de filmes, passou por pré-processamento para remover elementos indesejados (etiquetas, HTML, palavras em inglês, símbolos, caracteres repetidos e emoticons), resultando em 1.377.412 sentenças. O processo incluiu o uso de Word2Vec e FastText para expandir o *lexicon* de sentimentos inicial, derivado do SentiWordNet e gírias do Twitter. O *lexicon* final continha 10.537 palavras positivas e 12.664 negativas. O método de análise de sentimentos baseado em regras foi aplicado para prever sentimentos nos textos, resultando em uma precisão de 88% para o *lexicon* final.

O artigo de [20] apresenta a criação de um *lexicon* de sentimentos em indonésio por meio de um processo pipeline automático. O método envolveu várias etapas. Inicialmente, foram geradas palavras-sementes a partir da extração de termos do WordNet Bahasa. Essas palavras-sementes foram então associadas ao SentiWordNet, selecionando apenas aquelas com pontuações de sentimento acima de 0,5. Esse processo resultou em um conjunto de sinônimos com 1.957 palavras, que serviu como base do *lexicon*. Foram também adicionadas palavras formais usando o recurso Kateglo, que forneceu sinônimos e antônimos, além de gírias e emoticons populares. Para enriquecimento adicional, um corpus chamado TrustedCompany, contendo dados dos 15 principais mercados online, contribuiu com 2.625 palavras. Ao final, o *lexicon* resultante incluiu um total de 11.804 palavras. Nos testes realizados, o *lexicon* obteve um F1-Score de 88,8% ao ser avaliado com dados de mercado online usando o modelo CBOW (*Continuous Bag-of-Words*), e uma precisão de 95,7% ao ser avaliado com dados do Twitter usando o modelo Skip-gram.

Com o aumento expressivo da presença nas redes sociais e a imensa produção de dados, o reconhecimento e a identificação de emoções emergiram como um campo de pesquisa significativo. [21] abordaram o desafio de desenvolver um *lexicon* emocional versátil, aplicável a diversos domínios. Eles propuseram a criação do *lexicon* CB-Lex, utilizando um corpus não rotulado (JIRA, com 700 mil relatórios e 2 milhões de comentários) e *lexicons*-semente (Emolex e EmoSenticNet). A construção do *lexicon* envolveu técnicas como POS-tagging para identificar palavras, gerando três conjuntos de dados: o primeiro com adjetivos e advérbios, o segundo apenas com adjetivos, e o terceiro apenas com advérbios. Foram eliminadas as stopwords com base em dicionários. Em seguida, as proximidades das palavras candidatas foram calculadas usando o algoritmo PMI, a partir do *lexicon*-semente. A versão expandida do Emolex, que incluiu adjetivos e advérbios, apresentou o melhor desempenho, com uma medida F-Measure de 76,69%. Adicionalmente, as variantes do Emolex demonstraram maior precisão nos resultados.

No artigo de [22], é apresentada uma técnica para a expansão automática de um *lexicon* de sentimentos na língua árabe, considerando as limitações de recursos *lexicons* e conjuntos de dados disponíveis. O *lexicon* utilizado como ponto de partida

foi o NileULex, que é uma base construída manualmente e contém 6.287 palavras com polaridades negativas, positivas e neutras, usando o padrão do árabe moderno padrão (MSA) e o árabe egípcio. Inicialmente, para a expansão lexical, foi utilizada a técnica de incorporação de palavras. Para gerar o corpus de entrada, foi empregada uma base pré-treinada chamada Twitter CBOW AraVec, que consiste em mais de 70 milhões de tweets em árabe. A análise de proximidade entre os contextos em que as palavras aparecem resultou em um corpus expandido contendo 36.775 termos. Os resultados obtidos após a aplicação de algoritmos de classificação foram uma precisão de a 58%.

Considerando a ampla variação linguística presente no Twitter, [23] conduziram um estudo de análise de sentimento em tweets na língua portuguesa. Para a seleção dos dados de teste, utilizaram a API do Twitter para coletar 1.700 tweets com as hashtags "win"e "fall", buscando tweets com polaridades positivas e negativas. Dentre os dados coletados, 540 correspondiam a negações de termos. Para o *lexicon* semente, os pesquisadores utilizaram o OpLexicon, composto por 15.000 palavras com polaridades, adicionado aos adjetivos anotados no Sentilex, o que proporcionou uma vantagem para a análise em nível sentencial. Dois escopos de negação foram empregados: o primeiro baseado em janela (NegWind) e o segundo considerando a negação em toda a frase. Além disso, também foi considerado um escopo sem negação (NoNeg). No que diz respeito à classificação, foram testados dois conjuntos de dados: o OptionLexicon e o SentiLex. Conforme esperado, o primeiro conjunto obteve uma precisão melhor, atingindo 74%.

Os autores do trabalho [24] desenvolveram um *lexicon* baseado em microblogs sobre bolsa de valores, iniciando com a remoção de termos pouco frequentes e substituição de cashtags, menções e URLs por tags. A limpeza também incluiu a exclusão de mensagens compostas apenas por esses termos. O processo envolveu marcação de fala com POS, tokenização e lematização usando o Stanford CoreNLP. A criação do *lexicon* compreendeu três medidas estatísticas TF-IDF, Ganho de Informação (Tradução de IG) e Informação Mútua Pontual (Tradução de PMI), e medidas complementares como Pdays para avaliar a porcentagem de dias com mesma polaridade, e Massoc para calcular correlação. Foram criados 12 *lexicons* no total. Os conjuntos de treinamento incluíram 250.000 posts e a análise considerou unigramas e bigramas. Duas abordagens de avaliação foram aplicadas: treinamento com janela de 75% e 25%, e treinamento contínuo de 20, onde 2/3 dos dados foram usados para treinamento e o restante para teste. Os resultados indicam que a melhor performance foi alcançada pelo *lexicon* gerado com a abordagem PMI, combinado com Pdays e Massoc, resultando em um *F1-Score* de 85,5%.

3. Referencial teórico

Nesta seção, será apresentada uma visão geral das técnicas empregadas para analisar a eficácia dos procedimentos experimentais realizados. As técnicas que serão discutidas são: PMI, LSTM, TF-IDF e *AgglomerativeClustering*. Cada subseção a seguir definirá o que é cada técnica, explicar como ela funciona, suas propriedades e aplicações, e finalmente, como ela será utilizada neste experimento.

3.1 Point Mutual Information (PMI)

A Informação Mútua Ponto a Ponto é uma medida estatística que quantifica a relação entre dois eventos em um contexto específico. Ao contrário da Informação Mútua (MI), que se concentra na sobreposição de informações entre duas variáveis aleatórias, o PMI avalia a associação entre eventos individuais [25]. PMI é frequentemente utilizado em processamento de linguagem natural e recuperação de informação para medir a relação semântica entre palavras. Ele é positivo quando os eventos são dependentes e negativo quando eles são independentes.

O PMI entre dois eventos A e B é calculado a partir das probabilidades conjuntas $p(A, B)$, $p(A)$ e $p(B)$. A fórmula do PMI é definida da seguinte forma [25]:

$$PMI(A, B) = \log_2 \left(\frac{p(A, B)}{p(A) \cdot p(B)} \right)$$

O PMI mede a proporção entre a probabilidade conjunta dos eventos A e B e o produto de suas probabilidades marginais. Valores positivos de PMI indicam uma associação maior do que a esperada, enquanto valores negativos indicam uma associação menor do que a esperada.

- **Simetria**: O PMI é simétrico, ou seja, $PMI(A, B) = PMI(B, A)$.
- **Independência**: Se os eventos A e B forem independentes, então $PMI(A, B) = 0$.
- **Limitação**: O PMI não está limitado superiormente, o que pode dificultar a comparação entre diferentes pares de eventos. Portanto, é comum normalizá-lo para ter um valor máximo de 1 em caso de associação perfeita.

O PMI é amplamente utilizado em várias áreas, incluindo:

a) **Linguística**: Na análise de linguagem, o PMI é utilizado para medir a associação entre palavras em um corpus de texto. É frequentemente aplicado em tarefas como detecção de colocações, extração de termos-chave e classificação de documentos.
b) **Mineração de dados**: O PMI é útil na descoberta de padrões e associações entre itens em conjuntos de dados. Ele é aplicado em problemas de recomendação, análise de cestas de compra e detecção de associações em redes complexas.
c) **Bioinformática**: Na análise de sequências genômicas, o PMI é usado para identificar relações entre elementos como genes, regiões promotoras e proteínas. Ele ajuda a entender a função e a evolução dos organismos.
d) **Recuperação de Informação**: O PMI é aplicado na recuperação de informações em sistemas de busca, ajudando a melhorar a precisão das consultas e a relevância dos resultados retornados.

3.2 Long Short-Term Memory

A técnica LSTM, conforme descrito no [26], é um tipo de rede neural recorrente que permite a dependência entre

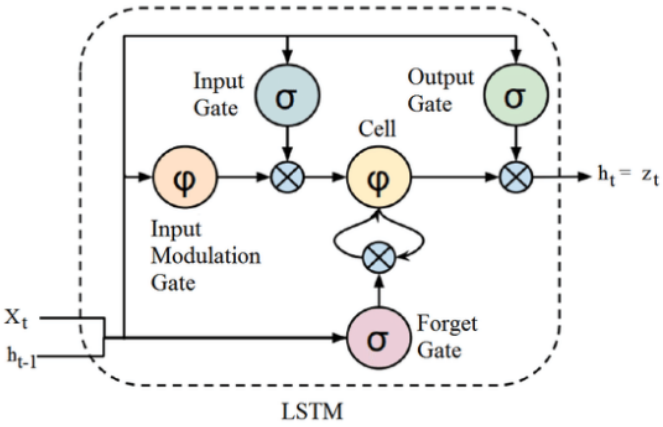
Tabela I. Trabalhos relacionados

Ref.	Tipo de dado	Tamanho da base	Técnicas de pré-processamento	Técnicas para classificação	Resultado (%)
[2]	Twiiter	22.782 tweets	Limpeza	SentStraight e Naiva Bays Multinomial	Acurácia - 88,29%
[3]	Twitter	12 milhões	—	LSTM, PMI, BERT e FT	F1-Sore - 81,6%
[5]	Youtube	1.375 expressões	OPENSMILE, 3D-CNN e BLSTM	BERT, LSTM	Acurácia - 95,35%
[7]	Twitter	10.000 twittes	Remover hiperlinks, erros ortográficos, Tokenização, caixa-baixa, reduzir palavras ao radical, substituir emoticoins pelo sentimento, TF-IDF	SoftMax, SVR, DTs e RF	Acurácia - 91,81%
[8]	Twitter	3.700 tweets	Remoção de stopwords, tokenização e geração de N-gramas	SVM, KNN, DT e NB	Precisão - 78,25%
[9]	*maps.safecity*"	—-	Conversões lowercase, remoção de: pontuações, stopwords, palavras muito específicas, tags, emoji e emoticons, e TF-IDF	RF, NB multinominal, SVC linear, SGD, Bernoulli NB, TD, K-NN	F1-Score 80%
[10]	Twitter	322.863 tweets	removendo símbolos, referências a usuários, URLs e caracteres não ingleses	—	—
[11]	site maps.safecity	9.892 histórias	Remoção de pontuações, stopwords, emojis, emoticons, tags HTML, tokenização, correção ortográfica e redução ao radical	CNN-LSTM, random forest, Gaussian, NB, SVM	Pontuação de Hamming - 86,5%
[12]	Twitter	7.682 tweets	Foram removidas tabulações duplas e símbolos duplicados, caixa-baixa, remoção de linhas em branco, tags HTML, menções e símbolos repetidos continuamente, os erros ortográficos foram corrigidos,	RF, Support Vector machines, LSVM e SMO	Acurácia - 85,16%.
[13]	Reddit	21.133 submissões, e 349.277 comentários	Revisão sistemática, rótulos preditivos, identificação de tags, processamento de dados, caixa baixa, normalização, redução ao radical, remoção de stopwords, links externos, URLs, substituição por "url"	Linear SVM, Nayve Bayes, Random Forest, Regressão Logística e Perceptron	Acurácia - 94%
[14]	Twitter e Noticias	2.132 notícias e 11.027 tweets	Limpeza dos dados, remoção de stopwords por meio do NLKT, tecnicas de streamming para reduzir ao radical e tecninca do bag do words	Naive Bayes, ME, SVM e MLP	F1-Score - 72,7%
[15]	Twiiter, Restaruant, Laptop	—-	–	Bi-LSTM, rede de memória de longo prazo dependentes do alvo (TD-LSTM) e classificação ao nível de aspecto (IAN), ADeCNN	Precisão - 84,03%
[16]	Twiiter	Total de 26.312 comentários	–	BERT	Acurácia - 98,26%
[17]	Twiiter	14.640, 50.000, 13871 tweets	Separação do texto em seções, tokenização, remoção de stopwords, remoção de ruídos, separação do conjunto por palavras chave	ULMFiT-SVM	Acurácia - 99,78%
[18]	Twitter	3.738.622 tweets	Limpeza dos dados, tag POS, pré-treino Linear Regressor	Sentimento baseado em *lexicons*	AUC - 84%
[19]	noolaham, wikipédia, twitter, facebook e sites de resenha de filmes	22664 palavras	Limpeza dos dados, expansão por meio do sentiwordnet	Sentimento baseado em *lexicons*	Acurácia - 88%
[20]	Twiiter e mercados online	11.804 palavras	Limpeza dos dados, redução ao valor base, marcação de fala POS, definição de valor das palavras, PMI	Sentimento baseado em *lexicons*	F1-Score - 88,8%
[21]	JIRA	18.705 palavras	Limpeza, POS-tagging, PMI, remoção de stopwords	Sentimento baseado em *lexicons*	F-Measure - 76,69%
[22]	Twitter	36.775 mil tweets		Sentimento baseado em *lexicons*	Precisão - 58%
[23]	Twitter	15.000 palavras	Heurísticas de normalização léxica, tratamento de variações lexicais e abreviações em textos de microblogs	Análise de Sentimento baseado em *lexicons*	F1-Score - 74%
[24]	Twitter	250.000 tweets	Remoção de termos de baixa frequência, exclusão de mensagens apenas com cashtags, links url, mensões ou avaliações. A marcação de parte da fala (POS), tokenização e lematização foi feita por meio do Stanford CoreNLP	PMI, TF-IDF, IG, Pdays, Massoc	F1-Score - 85,5 %

sequências de dados longas. Frequentemente utilizada para resolver problemas de processamento de linguagem natural, como classificação, tradução e geração de texto, sua arquitetura é composta por unidades de memória capazes de reter informações por períodos prolongados. Isso se mostra especialmente útil em tarefas de processamento de linguagem que demandam o manejo de grandes sequências, permitindo ao modelo aprender padrões de dependência complexos. A Figura 1, fornece uma representação visual da estrutura de uma célula LSTM, ilustrando sua capacidade de armazenamento e processamento de informações sequenciais extensas.

As Redes Neurais Recorrentes (RNNs) são amplamente utilizadas para modelar sequências de dados, mas enfrentam o desafio de capturar dependências de longo prazo. A arquitetura LSTM foi desenvolvida para lidar com esse problema, permitindo que a rede aprenda a lembrar informações relevantes por

Fig. 1. Exemplo de estrutura LSTM [26]



longos períodos.

A unidade básica de uma LSTM é composta por células de memória, portões de entrada, portões de esquecimento e portões de saída. Esses componentes trabalham em conjunto para controlar o fluxo de informações e decidir quais informações são importantes de serem lembradas ou descartadas.

A fórmula que descreve o funcionamento de uma célula LSTM é a seguinte:

$$
\begin{aligned}
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
g_t &= \tanh(W_g \cdot [h_{t-1}, x_t] + b_g) \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t \\
h_t &= o_t \cdot \tanh(c_t)
\end{aligned}
$$

Onde: - x_t é a entrada na posição de tempo atual t - h_t é a saída na posição de tempo atual t - c_t é o estado da célula na posição de tempo atual t - i_t, f_t e o_t são os portões de entrada, esquecimento e saída, respectivamente - g_t é o vetor de atualização da célula - W_i, W_f, W_o, W_g são as matrizes de pesos para os diferentes portões - b_i, b_f, b_o, b_g são os vieses para os diferentes portões - σ é a função sigmoide e $\tanh$ é a função tangente hiperbólica

A arquitetura LSTM supera as limitações das RNNs tradicionais, permitindo a captura de dependências de longo prazo em sequências de dados. Isso é especialmente útil em problemas em que a memória de longo prazo é importante, como tradução automática, geração de texto e análise de séries temporais.

Uma das principais vantagens da LSTM é capacidade de lidar com o desvanecimento do gradiente (*vashing gradient problem*), comum nas redes neurais tradicionais [27]. O desvanecimento ocorre quando os gradientes utilizados para a atualização dos pesos se toram pequenos o suficiente para dificultar o aprendizado de dependências de longo prazo. As unidades de memória do LSTM foram projetadas para mitigar tal problema, permitindo a preservação das informações por um período maior.

3.3 TF-IDF

A técnica TF-IDF (*Term Frequency-Inverse Document Frequency*) é uma medida estatística amplamente utilizada para avaliar a importância de um termo em relação a um documento em uma coleção de documentos. Ela é comumente empregada em tarefas de processamento de texto, como recuperação de informações, classificação de documentos e mineração de texto [28, 29, 30]

O cálculo do TF-IDF envolve duas etapas principais: o cálculo da frequência do termo (TF) e o cálculo do fator de ponderação inverso do documento (IDF). A combinação desses dois valores resulta no TF-IDF para cada termo em um documento.

A frequência do termo refere-se à contagem do número de ocorrências de um termo específico em um documento. Geralmente, a frequência do termo é normalizada para evitar que documentos mais longos tenham um peso maior. Uma das formas mais comuns de calcular o TF é utilizando a fórmula do TF ponderado [31]:

$$TF_{t,d} = \frac{f_{t,d}}{\max\{f_{w,d} : w \in d\}}$$

Onde: - $f_{t,d}$ é a frequência do termo t no documento d. - $\max\{f_{w,d} : w \in d\}$ é a frequência máxima de qualquer termo no documento d.

O fator de ponderação inverso do documento é calculado para avaliar a raridade de um termo em toda a coleção de documentos. Termos raros geralmente têm maior importância. O IDF é calculado utilizando a fórmula [31]:

$$IDF_t = \log\left(\frac{N}{df_t}\right)$$

Onde: - N é o número total de documentos na coleção. - df_t é o número de documentos que contêm o termo t.

O TF-IDF é obtido multiplicando a frequência do termo (TF) pelo fator de ponderação inverso do documento (IDF). Isso ajuda a identificar termos frequentes em um documento específico, mas raros na coleção de documentos na totalidade. A fórmula para o cálculo do TF-IDF é:

$$TF\text{-}IDF_{t,d} = TF_{t,d} \times IDF_t$$

A técnica TF-IDF é amplamente aplicada em várias tarefas de processamento de texto, incluindo:

- Recuperação de informações e sistemas de busca.
- Classificação de documentos e categorização de textos.
- Extração de palavras-chave e resumos automáticos.
- Análise de sentimento e mineração de opinião.

3.4 AglomerativeClustering

O AgglomerativeClustering, ou agrupamento aglomerativo, é uma técnica de aprendizado não supervisionado que agrupa dados semelhantes em clusters [32, 33]. Ele é utilizado em uma ampla variedade de aplicações, incluindo análise de redes sociais, agrupamento de documentos, segmentação de mercado, entre outras. Ao contrário do K-means que particiona os dados em clusters, o agrupamento aglomerativo segue uma

abordagem hierárquica, começando com cada dado como um cluster individual e fundindo-os até atingir um único cluster ou um número especificado de clusters [32, 33, 34, 35].

O algoritmo do agrupamento aglomerativo pode ser implementado nas seguintes etapas:

a) Inicialização: Inicialmente, cada ponto de dado é considerado como um cluster individual. Portanto, se houver N pontos de dados, haverá N clusters no início do algoritmo.
b) Cálculo da Matriz de Distância: Uma matriz de distância é calculada, onde cada entrada na matriz representa a distância entre dois clusters. A distância pode ser calculada de várias maneiras, sendo as mais comuns a distância euclidiana, a distância manhattan, entre outras.
c) União de Clusters: O par de clusters com a menor distância é identificado e unido em um único cluster. Isso reduz o número total de clusters em um.
d) Atualização da Matriz de Distância: A matriz de distância é atualizada para refletir a distância entre o novo cluster formado e os clusters existentes. Isso pode ser feito de várias maneiras, dependendo do critério de ligação utilizado. O critério de ligação determina a distância entre os clusters. Os critérios mais comuns incluem: - Ligação Única: a distância entre dois clusters é definida como a menor distância entre quaisquer dois pontos nos clusters [32, 33, 34, 35]. - Ligação Completa: a distância entre dois clusters é definida como a maior distância entre quaisquer dois pontos nos clusters [32, 33, 34, 35]. - Ligação Média: a distância entre dois clusters é definida como a média das distâncias entre todos os pares de pontos nos clusters [32, 33, 34, 35]. - Ligação de Ward: a distância entre dois clusters é definida com base no aumento da soma dos quadrados dentro do cluster resultante após a fusão [33, 35].
e) Repetição: As etapas de união de clusters e atualização da matriz de distância são repetidas até que todos os dados sejam agrupados em um único cluster ou até que um número especificado de clusters seja atingido [32, 33, 34, 35].
f) Considerações Finais: O agrupamento aglomerativo é um método poderoso de agrupamento que pode revelar hierarquias naturais nos dados. No entanto, ele tem algumas limitações. Por exemplo, uma vez que dois clusters são unidos, eles não podem ser divididos em etapas posteriores [32, 33]. Além disso, o algoritmo é sensível à escolha do critério de ligação e da métrica de distância. Por fim, o agrupamento aglomerativo pode ser computacionalmente intensivo para conjuntos de dados muito grandes, devido à necessidade de calcular e atualizar a matriz de distância [32, 33, 34].

4. METODOLOGIA

Nesta seção, será apresentada a metodologia empregada para realizar a expansão léxica relacionada ao contexto de violência contra a mulher em língua portuguesa. A metodologia é estruturada em quatro fases distintas: "ETAPA EXTRAÇÃO E TRATAMENTO", "ETAPA EXPANSÃO", "ETAPA CLASSIFICAÇÃO", "ETAPA VALIDAÇÃO", conforme Figura 2.

4.1 Coleta de Dados

Para obter os dados apresentados na Figura 2, item B, foi desenvolvida uma aplicação *crawler* que se conectou à API do Twitter para capturar e armazenar as informações relevantes. O filtro de palavras foi configurado para capturar apenas os tweets que contivessem termos relacionados à violência contra a mulher. Essa abordagem permitiu direcionar a coleta para mensagens específicas que pudessem fornecer percepções sobre o tema em questão. A coleta foi realizada como parte de um projeto de PFC da UFG [36], e os exemplos estão indicados na Tabela II.

Tabela II. Exemplos de tweets do corpus antes e pós-processamento.

tweets	tweets_processados
RT @menezes_anaaa: desculpa per te chamar de idiota cabeâ^šÃŸâ^šÂ£o, eu tava tentando flertar	rt desculpa per te chamar de idiota cabecao eu tava tentando flertar
mano sâ^šÃ©rio eu acordei pra assistir a aula e ele me vem com MEIA HORA de â€šÃ„Ãºaulaâ€šÃ„Ã¹ e â€šÃ„Ãºexplicaâ^šÃŸâ^šÂ£oâ€šÃ„Ã¹ PUTA QUE PARIU	mano serio eu acordei pra assistir a aula e ele me vem com meia hora de auaulaau e auexplicacaoau puta que pariu
Kkkkkkkkkkkkk ok imagina ser burra desse jeito nâ^šÂ£o ironicamente	kk ok imagina ser burra desse jeito nao ironicamente
RT @atbcons: Oi DIREITA, dia de fortalecer nossa uniâ^šÂ£o!	rt atbcons oi direita dia de fortalecer nossa uniao
RT @umbeargja: Eita que sâ^šâ%¢¥ falta uma putinha pra mamar aqui na Madruga https://t.co/9DRa2UdJDI	rt umbeargja eita que so falta uma putinha pra mamar aqui na madruga http

Para a seleção dos textos, utilizaram-se palavras-chave relacionadas ao contexto abordado, tais como: "violência doméstica", "feminicídio", "assédio sexual"e outras expressões similares. Foram estabelecidos critérios de inclusão para garantir a relevância e diversidade dos textos selecionados, levando em consideração a representatividade regional, fontes confiáveis e variedade de contextos discursivos.

4.2 Pré-processamento do Texto

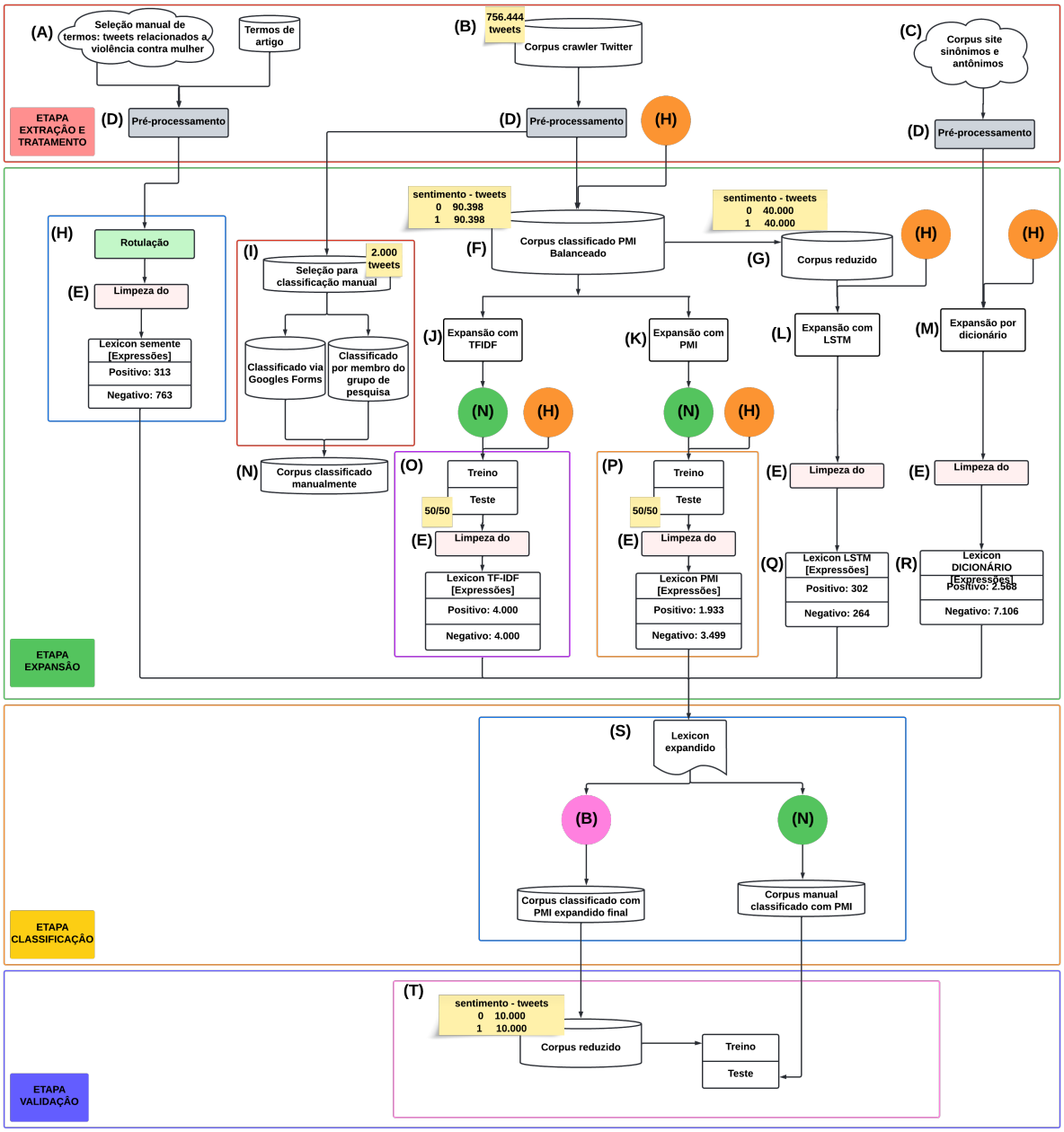
A etapa inicial envolveu o pré-processamento direcionado à expansão léxica Figura 2, item D. Nesse estágio, técnicas de processamento de linguagem natural e análise de texto foram empregadas, utilizando bibliotecas específicas da linguagem Python, conforme apresentado na Tabela III. O propósito desta fase é purificar e normalizar o texto, visando aprimorar a qualidade dos dados e simplificar a futura aplicação de algoritmos de aprendizado de máquina.

Tabela III. Atividades de pré-processamento realizadas utilizando diferentes bibliotecas em Python.

Biblioteca	**Atividades Realizadas**
Biblioteca Padrão de Python	Conversão para minúsculas, Expansão de abreviações, Correção de codificação, Remoção de caracteres sequenciais repetidos
NLTK (Natural Language Toolkit)	Tokenização, Lematização
Biblioteca re (expressões regulares do Python)	Remoção de tokens específicos, Remoção de acentos, Remoção de caracteres especiais, Remoção de tokens não alfabéticos
Biblioteca pandas	Reconstrução do texto, Remoção de duplicatas, Remoção de valores nulos

É relevante ressaltar que, no estágio de pré-processamento textual, optou-se por não realizar a exclusão das *stopwords*

Fig. 2. Desenho do experimento.



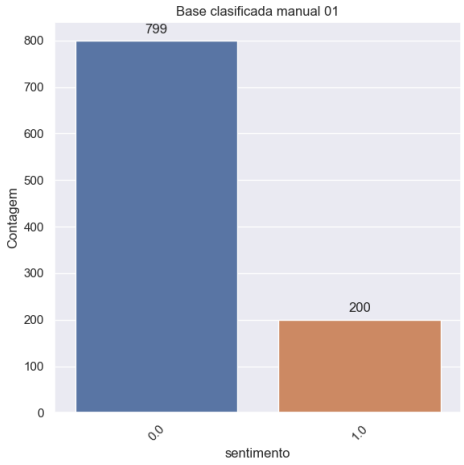
(termos de uso comum que não contribuem de maneira significativa para a análise de sentimento, como artigos, preposições e pronomes), visto [23]. Isso se deve ao fato de que, no contexto do reconhecimento da violência contra a mulher, as stopwords podem desempenhar um papel importante. Certos termos e palavras-chave específicas são fundamentais para identificar e contextualizar situações de violência. Por exemplo, pronomes e artigos como "ela"ou "a"podem indicar o gênero feminino em expressões como "ela é uma vagabunda"ou "a puta". Portanto, a presença das stopwords não foi removida, a fim de preservar a relevância das palavras no processo de análise.

4.3 Classificação manual

Após a etapa de preparação da base de dados, o primeiro passo foi selecionar 2.000 tweets para a classificação manual, como pode ser visto na Figura 2, item I. A primeira parte desses dados foi categorizada usando o *Google Forms*, onde foram disponibilizados questionários que permitiam aos respondentes indicar se o tweet continha ou não violência contra a mulher. A classificação final foi determinada pela mediana das notas atribuídas por cada avaliador. Os avaliadores foram três mulheres, que não têm afiliação acadêmica. A distribuição das classificações dos tweets é visualizada na Figura 3.

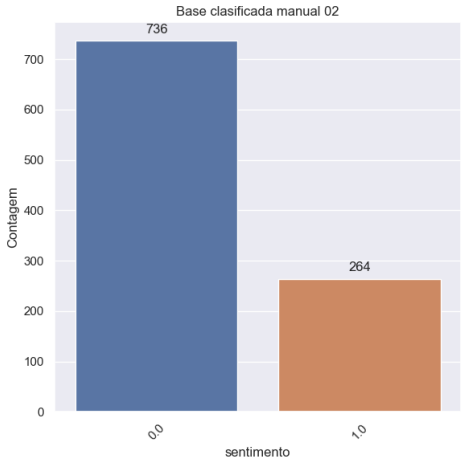
Enquanto a segunda parte, destinada à validação da classificação pelo *lexicon*, foi realizada por outro membro do grupo de pesquisa. A distribuição das classificações realizadas está

Fig. 3. Dados classificados Google Forms



ilustrada na Figura 4.

Fig. 4. Dados classificados pela equipe



As diretrizes fornecidas para a classificação e rotulagem dos dados incluíram a seguinte abordagem: primeiramente, avaliar a presença de palavras e expressões relacionadas à violência contra a mulher; em segundo lugar, analisar o contexto dos tweets e as conotações das palavras, considerando suas diferentes utilizações; por fim, levar em consideração a maneira como as palavras foram empregadas para expressar emoções e atitudes. Exemplos das classificações realizadas de acordo com essas diretrizes estão apresentados na Tabela IV.

Tabela IV. Exemplo de tweets classificados manualmente

tweets	Avaliação
vagabundau	1.0
yasfmff puta	1.0
rt amantedoscasais levou a magrinha safada pro meio do mato e comeu ela gostoso http tcohevkdkrnzv	1.0
rt unstvabler selenators amar passar vergonha como a fracassada da fav deles poder da carreira pra alguem se ela nao tem ela mesmo assua	0.0
euamandafurtado kk puta que pariu e uma duvida muito pertinente	0.0
poxa jao ja era	0.0

4.4 Extração dos Termos para o lexicon *Semente*

Para a elaboração do *lexicon* semente e para ampliar a abrangência dos termos relacionados ao discurso de ódio direcionado às mulheres, foram efetuadas coletas de palavras e expressões em duas etapas como indicado na (Figura 2 item A). Na primeira, foram utilizadas expressões extraídas do trabalho de monografia de [36]. A segunda fase envolveu uma busca por unigramas, bigramas e trigramas no Twitter. Para tal, realizou-se pesquisas com hashtags como "#sexismo", "#violênciacontramulheres", "#discursoodioso"e "#Violência-ContraMulher". O propósito era de coletar informações nos ambientes onde a violência ocorre, proporcionando assim uma cobertura mais ampla dos termos utilizados nesses locais, as expressões coletadas foram adicionadas ao *lexicon* semente presente no Figura 2, item H.

4.5 Limpeza do lexicon

A limpeza do *lexicon* desempenha um papel importante na análise de sentimentos, assegurando que as palavras utilizadas para classificar frases sejam pertinentes e coerentes com o contexto. Esse processo é indicado na Figura 2, item E. Nesta subseção, serão apresentadas as técnicas empregadas, com ênfase nas distinções entre o *lexicon* positivo e o negativo. O *lexicon* positivo incorpora termos que indicam a presença de violência contra a mulher, enquanto o *lexicon* negativo abrange casos nos quais não há indícios dessa violência.

Foram aplicadas duas etapas distintas na limpeza dos *lexicons*: uma destinada a ambos os *lexicons* positivo e negativo, e outra especificamente voltada ao *lexicon* positivo.

Na etapa de limpeza comum a ambos os *lexicons*, é executada uma operação de remoção de stopwords. Onde, a expressão é descartada do *lexicon* apenas se todas as palavras forem stopwords.

Adicionalmente, o *lexicon* positivo passa por duas etapas suplementares para excluir termos que possam restringir a análise à violência contra a mulher. Na primeira, uma função é empregada para identificar e remover adjetivos no gênero masculino, utilizando a biblioteca NLTK e o WordNet. Se uma palavra for reconhecida como um adjetivo masculino, ela é excluída do *lexicon* positivo. Na segunda, outra função é utilizada para determinar se uma frase contém pronomes de primeira ou terceira pessoa do singular no gênero masculino. Caso sejam identificados, a frase pode estar se referindo a uma pessoa específica e, portanto, não é adequada para o *lexicon* positivo.

Em resumo, o objetivo da implementação dessas etapas é evitar expressões no *lexicon* positivo nas quais o locutor promova características para si ou a agressão não seja direcionada ao público do gênero feminino. O resultado é um *lexicon* positivo focado em aspectos relevantes à violência contra a mulher e livre de referências masculinas que poderiam restringir a análise.

4.6 Classificação da base com o lexicon *semente*

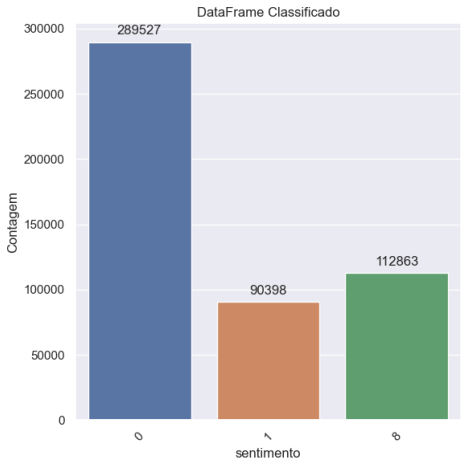
Para realizar a expansão léxica com termos frequentes e palavras relacionadas com maior precisão, a primeira etapa

de expansão léxica envolveu a classificação da base de dados por meio do *lexicon* semente, utilizando o algoritmo PMI projetado para categorizar frases [3]. Cada sentença foi inicialmente tokenizada usando a biblioteca NLTK para garantir a consistência dos dados, e, em seguida, procedeu-se à identificação de bigramas e trigramas na mesma.

Posteriormente, o cálculo das medidas estatísticas de PMI foi realizado para os n-gramas identificados entre os *lexicons* semente, positivo e negativo, e cada tweet individual, conforme [24]. Esses cálculos foram conduzidos em três componentes principais: unigramas, bigramas e trigramas. Por fim, as pontuações resultantes de cada avaliação foram somadas.

A fase seguinte envolve a classificação das frases com base nas pontuações computadas. Um tweet é classificado como positiva se o PMI em relação ao *lexicon* positivo for maior do que o relacionado ao *lexicon* negativo. Se os valores forem iguais, ele é classificado como neutro. Essa etapa é aplicada exclusivamente nesta fase inicial, servindo para minimizar possíveis erros nas fases subsequentes. Nesse contexto, são excluídos os tweets classificados como neutros. A distribuição dos resultados é ilustrada na Figura 5.

Fig. 5. Distribuição das classificações com *lexicon* semente

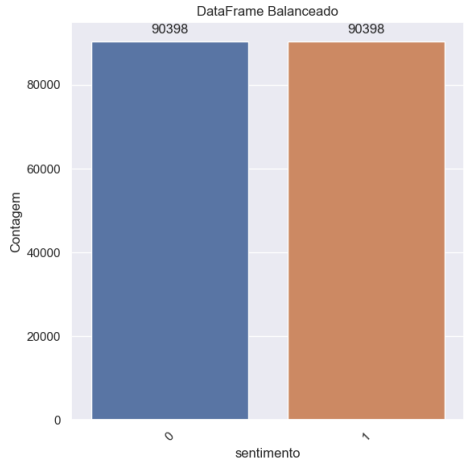


Para aprimorar a distribuição da classificação, foi aplicada a técnica de reamostragem *downsample* da biblioteca sklearn.utils do Python. Essa técnica alinhou o número de sentenças positivas e negativas no corpus como visto na Figura 6, buscando minimizar distorções na análise. Esses resultados da classificação inicial foram registrados, incluindo as sentenças classificadas, valores de PMI positivo e negativo, além dos rótulos de sentimento correspondentes, resultando na base apresentada na Figura 2, item F.

4.7 Expansão do lexicon *com TF-IDF*

A extração de características desempenha um papel fundamental na análise de texto, com o método TF-IDF sendo uma abordagem amplamente adotada nesse domínio, como apresentado por [24]. Nesta seção, será discutida a aplicação específica do TF-IDF para a extração de características no contexto da análise de sentimentos, conforme ilustrado na Figura 2, item J e O.

Fig. 6. Distribuição das classificações depois do balanceamento



O procedimento começa com a vetorização dos dados de texto usando a classe *TfidfVectorizer()* do módulo *sklearn.feature_extraction.text*. Essa classe transforma os textos em uma representação numérica adequada para algoritmos de aprendizado de máquina. O vetorizador é configurado considerando critérios como a frequência mínima de documentos (*min_df*), o número máximo de características (*max_features*) e o intervalo de n-gramas.

Após a adaptação do vetorizador aos textos pré-processados, obtém-se a matriz TF-IDF, uma representação quantitativa em que cada linha corresponde a um documento e cada coluna a um termo. Os valores em cada célula refletem a importância do termo no respectivo documento com base no cálculo do TF-IDF. Com a matriz TF-IDF em mãos, prossegue-se para a seleção das características. Inicialmente, são extraídos os termos incorporados pelo vetorizador, os quais são ordenados com base nos valores de TF-IDF.

Para otimização mais eficaz, emprega-se *GridSearch* na base de dados classificada manualmente destinada para treinamento. Essa técnica de otimização auxilia na seleção refinada dos parâmetros. Por fim, com base nos parâmetros *min_df* e *max_features*, são selecionadas as características com os maiores valores de TF-IDF. Esse processo garante a escolha das palavras mais representativas e significativas como características para o modelo de aprendizado de máquina, melhorando sua eficiência e precisão na análise de sentimentos. Esse procedimento é aplicado tanto aos dados classificados como positivos quanto aos negativos, resultando na expansão dos *lexicons* respectivos.

Em conclusão, a extração de características por meio do TF-IDF é um processo em várias etapas, que envolve o pré-processamento de dados, a transformação de dados textuais em uma representação numérica e a seleção de termos significativos para serem usados como características. Esta metodologia efetivamente converte texto não estruturado em dados estruturados e informativos que podem ser utilizados por algoritmos de aprendizado de máquina para realizar tarefas como a análise de sentimentos.

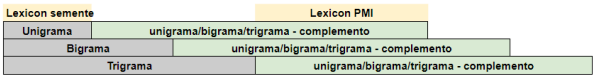
4.8 Expansão do lexicon *com PMI*

A expansão do *lexicon* é um procedimento importante em processamento de linguagem natural que ajuda a enriquecer o conjunto de palavras-chave ou frases utilizadas para análises subsequentes. Nesta etapa é realizada a expansão do *lexicon* por meio do índice de Pontuação de Informação Mútua [24], o processo pode ser visto na Figura 2, item K e P.

Inicialmente, os dados da Seção 4.6 são divididos em dois subconjuntos, um para sentimentos positivos e outro para negativos, a fim de processá-los separadamente para a expansão do *lexicon*. Para cada subconjunto, os tweets são combinados em um único documento e tokenizado em palavras individuais usando a classe *RegexpTokenizer*, baseada em expressões regulares.

Depois, para cada tipo de n-grama, bigrama, tri-grama e quadrigrama (quatro palavras consecutivas), são identificadas as combinações mais frequentes de palavras e calculadas as respectivas pontuações de PMI, a forma das combinações é apresentada na Figura 7.

Fig. 7. Funcionamento do processo de expansão com PMI



Quando uma expressão do *lexicon* semente é encontrada em um n-grama, as palavras restantes desse n-grama são adicionadas ao *lexicon* expandido, acompanhadas de suas pontuações de PMI. Isso correr para unigramas, bigramas e trigramas, seguindo o esquema apresentado na Figura 7. As entradas do *lexicon* expandido são ordenadas com base na pontuação PMI em ordem decrescente, e as primeiras *n_top* entradas são selecionadas.

Finalmente, as entradas do *lexicon* expandido são ordenadas em ordem decrescente de pontuação PMI, e as *n_top* primeiras entradas são retornadas.

Considerando o processo acima, a otimização da expansão do *lexicon* envolveu um processo de ajuste de parâmetros por meio de *GridSearch*. Nesse procedimento, por meio da base de treino, o parâmetro de *n_top* foi otimizado, com uma frequência minima de documento para cada grama de treze. Utilizando esses parâmetros, a expansão do *lexicon* foi conduzida de maneira iterativa, calculando as pontuações de PMI e selecionando as entradas principais para o *lexicon* expandido. Durante cada iteração, o desempenho das entradas expandidas foi avaliada, utilizando métricas como acurácia e F1-score. Esse processo de otimização permitiu a adaptação dos parâmetros do PMI conforme as características do conjunto de dados, melhorando a capacidade do *lexicon* expandido em capturar nuances e variações nas expressões de sentimentos presentes nos textos.

De uma perspectiva teórica, esta abordagem permite que o *lexicon* seja ampliado com base na associação estatística de palavras no corpus de texto. Isso significa que as novas palavras ou frases que são adicionadas ao *lexicon* são aquelas que ocorrem frequentemente em conjunto com as frases existentes no *lexicon*, e, portanto, são provavelmente relevantes para a análise subsequente.

4.9 Expansão do lexicon *com dicionário*

A expansão do *lexicon* utilizando um dicionário de sinônimos e antônimos representa um passo crucial na melhoria da análise de sentimentos, como feito por [20]. Isso possibilita que uma gama mais ampla de expressões seja reconhecida. Nesse caso, a expansão ocorre pela extração de sinônimos e antônimos de uma palavra pertencente ao *lexicon* semente, vide Figura 2, item M.

A expansão do *lexicon* foi realizada mediante uma abordagem programática que toma o *lexicon* semente como entrada e retorna listas de sinônimos e antônimos para cada palavra no *lexicon*. Esta abordagem utiliza consultas online a sites que fornecem dicionários de sinônimos e antônimos e extrai essas informações por meio de técnicas de raspagem de dados. Os sinônimos são obtidos a partir de consultas no site referenciado em [37], enquanto os antônimos foram obtidos no site [38]

Os sinônimos originados do *lexicon* semente positivo são incluídos na base de *lexicon* expandido positivo, enquanto a base de *lexicon* expandido negativo recebe os antônimos do *lexicon* semente positivo e os sinônimos do *lexicon* semente negativo. A ausência de adição dos antônimos da base negativa à expansão positiva acontece em virtude de muitos desses antônimos serem n-gramas positivos. Um exemplo seria o trigrama negativo "eu to feia", que gera antônimos positivos como "linda, bela, maravilhosa". Essa expansão com sinônimos e antônimos enriquece e diversifica os termos utilizados na classificação, permitindo que o modelo identifique variações e nuances linguísticas relacionadas à violência contra a mulher.

A etapa de expansão do *lexicon* com o dicionário de sinônimos é complementar a outras estratégias de expansão, como o emprego de técnicas baseadas em LSTM e a seleção de palavras frequentes por meio do TF-IDF. Essas abordagens combinadas auxiliam na criação de um *lexicon* mais robusto e eficiente para a classificação dos tweets associados à violência contra a mulher.

4.10 Expansão do lexicon *Utilizando LSTM*

Por fim, foi realizada uma expansão do *lexicon* utilizando uma abordagem baseada em LSTM [5], indicado Figura 2, item L. Essa abordagem possibilita gerar novos termos a partir do *lexicon* semente com o intuito de aumentar a abrangência e a diversidade do *lexicon* utilizado na classificação dos tweets.

Para essa expansão, foi elaborado um código que treina um modelo LSTM para prever sequências de palavras, tendo como entrada o *lexicon* semente. As frases da base de dados são tokenizadas e convertidas em sequências de palavras. Em seguida, as sequências são divididas em *input* e *output*, com o *output* convertido para representação binária (método *onehotencoding*). A construção do modelo LSTM é realizada com a ajuda da biblioteca Keras, incluindo uma camada LSTM de 128 unidades e uma camada densa de saída com ativação *softmax*. O modelo é compilado e treinado utilizando a função de perda *categorical_crossentropy* e o otimizador *adam*.

Após o treinamento, o modelo LSTM e o tokenizer são salvos. A partir do *lexicon* semente, novos termos são gerados utilizando a abordagem de predição do próximo termo com base no modelo LSTM. Os novos termos gerados são adicionados ao *lexicon* expandido positivo e negativo, respectivamente.

4.10.1 Aplicação: Considerando o tamanho da base inicial classificada através do *lexicon* semente, uma redução foi necessária devido a limitações técnicas. Utilizando o método *sample()* da biblioteca pandas, foi selecionada uma amostra de 40.000 tweets classificados como positivos, bem como 40.000 negativos para essa expansão, como visto na Figura 2, item G.

Na formulação da rede neural, o *tokenizer* e o modelo LSTM previamente treinados foram carregados. Em seguida, novos termos foram gerados utilizando a técnica baseada em LSTM. Para cada termo semente, foram gerados 50 novos termos para o *lexicon* positivo e 30 para o *lexicon* negativo.

4.11 Classificação Final Utilizando Medida de Informação Mútua Pontual

Conforme descrito na Seção 4.11, as pontuações PMI são determinadas com base nos *lexicons* de entrada. A classificação dos tweets é efetuada utilizando essas pontuações. No processo de classificação, se a pontuação positiva exceder a pontuação negativa, a frase é classificada como positiva (1). Caso contrário, é classificada como negativa (0). Esse procedimento resulta na criação da coluna 'sentimento' no DataFrame, contendo a classificação final para cada tweet. Essa etapa está ilustrada na Figura 2, item S, e também é aplicada nos estágios de treinamento e teste, representados pelos itens O e P, respectivamente.

4.12 Classificação Utilizando Agglomerative Clustering

O processo de classificação por meio do algoritmo de agrupamento foi empregado para categorizar os dados com base nos atributos extraídos dos tweets e nas pontuações de Polaridade de Informação Mútua, dataset fornecido pela Seção 4.11, e indicado na Figura 2, item T.

Por conseguinte foram selecionadas duas amostras aleatórias com 10.000 tweets, utilizando o método *sample()* da biblioteca pandas, representando sentimentos positivos e negativos. Essas amostras foram então mescladas em um novo *dataframe*, que se tornou o ponto de partida para a análise subsequente.

Posteriormente, os dados passaram por uma etapa de processamento utilizando o *TfidfVectorizer*, o qual converteu as frases em uma matriz de características TF-IDF. Adicionalmente, as pontuações PMI, tanto positivas quanto negativas, foram incorporadas à matriz de características. O resultado obtido foi uma matriz densa, combinando as características com as pontuações, a qual foi empregada como entrada para o algoritmo de agrupamento.

Para realizar o agrupamento, um objeto *AgglomerativeClustering* foi criado com dois clusters, usando o critério de ligação 'ward'. Esse objeto foi então ajustado aos dados, resultando em rótulos de cluster para cada ponto de dados.

O processo ocorreu duas vezes para o mesmo conjunto de dados, um considerando as notas dadas pelo PMI e outro usando apenas os tweets. Após a transformação desses textos pelo vetorizador, previamente treinado, as pontuações PMI correspondentes foram adicionadas. Em seguida, o algoritmo de agrupamento foi aplicado novamente, gerando rótulos de cluster para os novos dados.

Para avaliar a eficácia do algoritmo de agrupamento com e sem a inclusão das pontuações, o processo foi repetido sem as pontuações na matriz de características. O agrupamento foi realizado e os rótulos de cluster foram atribuídos aos novos dados.

5. Resultados e discussões

Nesta etapa, realizou-se a análise dos resultados provenientes da expansão léxica. Durante esse processo, exploraram-se as relações entre os termos expandidos, descritos na Tabela V, buscando identificar similaridades, contradições e complementaridades entre eles.

Tabela V. Dados quantitativos de distribuição n-gramas para cada *lexicon*

Lexicos/n-gramas	unigramas	bigramas	trigramas	total
Semente positivo	61	156	96	313
Semente negativo	311	321	131	763
TF-IDF positivo	0	3156	844	4000
TF-IDF negativo	0	3117	883	4000
PMI positivo	492	1004	437	1933
PMI negativo	428	1959	1112	3499
Dicionário positivo	2340	167	61	2568
Dicionário negativo	6784	178	144	7106
LSTM positivo	302	0	0	302
LSTM negativo	264	0	0	264
Positivo final	3195	4483	1438	9116
Negativo final	7787	5575	2270	15632

No contexto de cada expansão realizada, os conjuntos de dados base - Figuras 4 e 3 - foram submetidos a um processo de classificação, conforme detalhado na seção 4.11. Os resultados dessa classificação estão apresentados na Tabela VI.

Pela Tabela VI pode-se observar a eficácia da expansão ao complementar o *lexicon* semente. A estratégia adotada que combinou sinônimos, antônimos, PMI e um modelo LSTM, enriqueceu significativamente o repertório de termos, capturando nuances linguísticas e contextos de violência contra a mulher. A inclusão do TF-IDF também desempenhou um papel fundamental, ampliando ainda mais o *lexicon*, cujo F1-Score foi o maior. O resultado é um *lexicon* mais abrangente do que o original, capaz de mapear com precisão as complexidades da linguagem e expressões ligadas as técnicas aplicadas.

Ademais, após a execução do *Agglomerative Clustering*, ao comparar os clusters obtidos utilizando a classificação dada pelo *lexicon* final na Figura 8 e 9, pode-se constatar que a inclusão das pontuações PMI resultou em uma segmentação mais clara e representativa dos sentimentos presentes nos tweets. Isso sugere que as expansões contribuíram para a identificação de termos-chave relevantes, levando a uma melhor discriminação entre diferentes aspectos do discurso de violência contra a mulher. Os clusters formados mostraram uma distribuição mais coerente e distintiva quando as pontuações PMI foram consideradas.

Por fim, considerando que as expansões realizadas com PMI e TF-IDF foram baseadas na base classificada utilizando o *lexicon* semente, um ponto de melhoria seria realizar a expansão deste utilizando uma base classificada manualmente. Outra opção seria a realização de uma coleta mais expressiva para as expansões via dicionário e LSTM. Vale ressaltar que uma das vantagens fundamentais das expansões léxicas é sua capacidade

Fig. 8. Resultado *Agglomerative Clustering* sem notas PMI

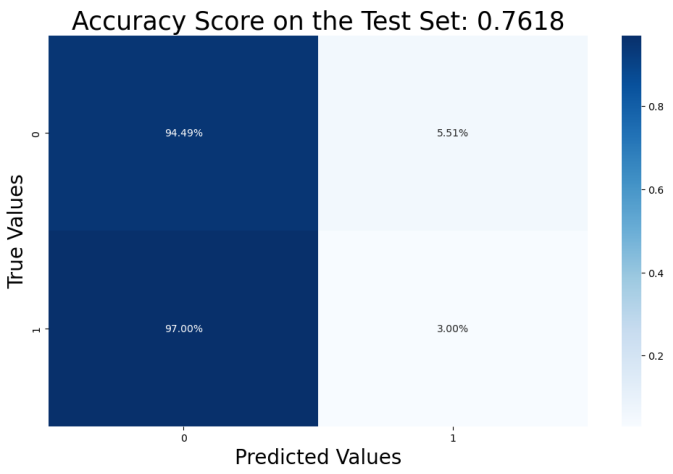
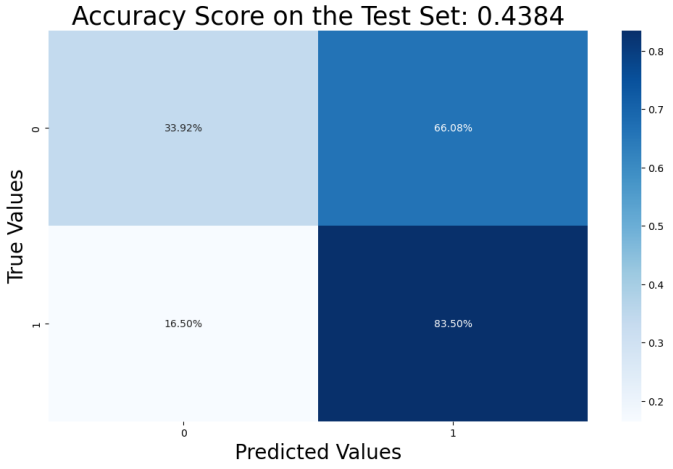


Fig. 9. Resultado *Agglomerative Clustering* com notas PMI



de abranger o domínio da linguagem, permitindo a identificação de nuances e contextos que podem escapar a modelos de aprendizado de máquina que não possuam essas informações em sua base de treinamento.

Tabela VI. Resultados com a distribuição de acurácia para cada combinação dos *lexicons* - dados de treino e validação

Treino - Tweets	799 tweets	200 tweets	-
	Acurácia neg	Acurácia pos	F1-SCORE
lexicon semente	90,74%	89,00%	90,78%
lexicon semente + TF-IDF	84,11%	89,00%	86,11%
lexicon semente + PMI	85,98%	72,50%	84,02%
lexicon semente + dicionário	91,49%	65,50%	86,27%
lexicon semente + LSTM	91,61%	70,50%	87,48%
lexicon semente + TF-IDF + PMI	82,23%	79,50%	82,91%
lexicon semente + TF-IDF + PMI + dicionário	82,85%	77,50%	82,23%
lexicon semente + TF-IDF + PMI + dicionário + LSTM	83,23%	74,50%	82,57%
Teste - Tweets	736 tweets	264 tweets	-
	Acurácia neg	Acurácia pos	F1-SCORE
lexicon semente	85,33%	79,17%	84,14%
lexicon semente + TF-IDF	79,76%	83,71%	81,66%
lexicon semente + PMI	80,43%	71,97%	78,93%
lexicon semente + dicionário	92,39%	57,58%	82,51%
lexicon semente + LSTM	89,67%	66,29%	83,36%
lexicon semente + TF-IDF + PMI	77,99%	81,44%	79,88%
lexicon semente + TF-IDF + PMI + dicionário	80,43%	78,79%	80,79%
lexicon semente + TF-IDF + PMI + dicionário + LSTM	80,43%	77,65%	80,49%

6. Conclusão

Com base nos resultados apresentados na Tabela VI, foi possível constatar que a abordagem expansão por dicionário resultou em uma acurácia mais alta para as palavras negativas. Entretanto, os resultados apresentados com TF-IDF e PMI apresentaram uma maior acurácia para as palavras positivas. É notório que o método que combinou várias estratégias, ou seja, "*lexicon* inicial + TF-IDF + PMI + dicionário + LSTM", não resultou necessariamente nos melhores resultados, no entanto, é a abordagem com maior abrangência linguística, possibilitando uma identificação melhor para a língua portuguesa.

Em termos de F1-Score, uma métrica que considera tanto a precisão quanto o *recall*, a abordagem que combinou o "*lexicon* inicial + TF-IDF"alcançou a pontuação mais alta nos dados de validação. Em resumo, os resultados sugerem que diferentes estratégias de expansão do *lexicon* e classificação de sentimentos podem apresentar variações de desempenho, dependendo da natureza dos dados e das palavras que estão sendo classificadas. Ademais, ao comparar com os resultados obtidos por [3], que aplicou uma metodologia similar usando PMI e atingiu um resultado de 68,1%, pode-se considerar que o F1-Score de 82,57% obtido neste trabalho está dentro dos padrões aceitáveis.

Agradecimentos

Dedico este projeto a todos que fizeram parte da minha jornada acadêmica até o momento. Agradeço profundamente à minha mãe, que me apoiou em cada escolha e superou desafios ao longo dos anos de estudo. Gostaria de lembrar e honrar a memória da minha querida avó, cuja influência é eterna.

Aos meus pais, Lopes e Edson, expresso minha gratidão por seu incansável apoio e dedicação em me auxiliar no meu percurso educacional.

À minha orientadora, professora Deborah, sou imensamente grato(a) pela oportunidade de adentrar no ambiente acadêmico dos docentes. Sua orientação foi inestimável na escolha do tema para meu TCC e em esclarecer minhas dúvidas sempre que necessário.

Não posso deixar de mencionar minhas amigas Deborá e Isabella, que foram minhas parceiras e proporcionaram grande auxílio ao longo do desenvolvimento deste projeto. Também sou grato aos meus amigos, Artur, Luan, Ronan e Vinicius, que têm sido meus companheiros desde o primeiro semestre e compartilharam comigo as experiências em todas as disciplinas.

Este trabalho é dedicado a todos vocês, que tiveram um impacto significativo em minha jornada acadêmica. Sua presença e apoio fizeram toda a diferença.

Referências

[1] W. Leal Filho, "Encyclopedia of the un sustainable development goals," *(No Title)*, 2021.

[2] Z. Zhou, X. Zhang, and M. Sanderson, "Sentiment analysis on twitter through topic-based lexicon expansion," 2014, pp. 98–109.

[3] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "Solid: A large-scale semi-supervised dataset for offensive language identification," 4 2020. [Online]. Available: http://arxiv.org/abs/2004.14454

[4] J. Burstein, C. Doran, and T. Solorio, "Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
[5] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu sentiment analysis via multimodal data mining based on deep learning algorithms," *IEEE Access*, vol. 9, pp. 153072–153082, 2021.
[6] T. Britannica, "Editors of encyclopaedia," *Argon. Encyclopedia Britannica*, 2020.
[7] S. E. Saad and J. Yang, "Twitter sentiment analysis based on ordinal regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019.
[8] M. Zyout and N. Hassan, "Sentiment analysis of arabic tweets about violence against women using machine learning," 1 2021.
[9] E. Alawneh, M. Al-Fawa'reh, M. T. Jafar, and M. A. Fayoumi, "Sentiment analysis-based sexual harassment detection using machine learning techniques," 2021, pp. 1–6.
[10] J. Xue, J. Chen, and R. Gelles, "Using data mining techniques to examine domestic violence topics on twitter," *Violence and Gender*, vol. 6, pp. 105–114, 6 2019.
[11] S. Karlekar and M. Bansal, "Safecity: Understanding diverse forms of sexual harassment personal stories," *arXiv preprint arXiv:1809.04739*, 2018.
[12] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García, "Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings," *Future Generation Computer Systems*, vol. 114, pp. 506–518, 2021.
[13] N. Schrading, C. O. Alm, R. Ptucha, and C. Homan, "An analysis of domestic abuse discourse on reddit," 2015, pp. 2577–2583.
[14] A. E. O. Carosia, G. P. Coelho, and A. E. A. Silva, "Analyzing the brazilian financial market through portuguese sentiment analysis in social media," *Applied Artificial Intelligence*, vol. 34, pp. 1–19, 2020.
[15] J. Zhou, S. Jin, and X. Huang, "Adecnn: An improved model for aspect-level sentiment analysis based on deformable cnn and attention," *IEEE Access*, vol. 8, pp. 132970–132979, 2020.
[16] J. He, A. Wumaier, Z. Kadeer, W. Sun, X. Xin, and L. Zheng, "A local and global context focus multilingual learning model for aspect-based sentiment analysis," *IEEE Access*, vol. 10, pp. 84135–84146, 2022.
[17] B. AlBadani, R. Shi, and J. Dong, "A novel machine learning approach for sentiment analysis on twitter incorporating the universal language model fine-tuning and svm," *Applied System Innovation*, vol. 5, 2 2022.
[18] F. Bravo-Marquez, E. Frank, and B. Pfahringer, "Positive, negative, or neutral: Learning an expanded opinion lexicon from emoticon-annotated tweets," 2015.
[19] S. Thavareesan and S. Mahesan, "Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts," 2020, pp. 272–276.
[20] R. Wijayanti and A. Arisal, "Automatic indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, 4 2021.
[21] H. S. Khawaja, M. O. Beg, and S. Qamar, "Domain specific emotion lexicon expansion," 2018, pp. 1–5.
[22] M. Youssef and S. R. El-Beltagy, "Moarlex: An arabic sentiment lexicon built through automatic lexicon expansion," vol. 142. Elsevier B.V., 2018, pp. 94–103.
[23] M. Souza and R. Vieira, "Sentiment analysis on twitter data for portuguese language," 2012.
[24] N. Oliveira, P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures," *Decision Support Systems*, vol. 85, pp. 62–73, 2016.
[25] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.
[26] D. L. Book. (Sem data) Arquitetura de redes neurais long short-term memory. Acessado em: 10/05/2023. [Online]. Available: https://www.deeplearningbook.com.br/arquitetura-de-redes-neurais-long-short-term-memory
[27] R. DiPietro and G. D. Hager, "Deep learning: Rnns and lstm," in *Handbook of medical image computing and computer assisted intervention*. Elsevier, 2020, pp. 503–519.
[28] G. Salton, "Cs a vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 6, pp. 391–398, 1975.
[29] K. Sparck-Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 5, pp. 111–121, 1972.
[30] H. Schutze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008.
[31] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
[32] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988.
[33] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed. Wiley, 2011.
[34] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Addison-Wesley Longman Publishing Co., Inc., 2006.
[35] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated, 2013.
[36] P. H. A. BATISTA, "Detecção de traços de exposição à violência contra mulher no twitter," 2021.
[37] Sinônimos.com.br. [Online]. Available: https://www.sinonimos.com.br
[38] Antônimos.com.br. [Online]. Available: https://https://www.antonimos.com.br

Heinrych Matheus Gomes, um estudante de Engenharia de Computação na Universidade Federal de Goiás, tem atuado no setor de Análise de Dados [Telemetria] na empresa Poli. Utiliza principalmente ferramentas como *Python*, *PowerBi* e *SQL* em seu trabalho. Possui forte interesse em Análise de Dados e desenvolvimento de produtos.

Deborah S. A. Fernandes trabalha no Instituto de Informática da Universidade Federal de Goiás – Campus Samambaia, Goiânia, Goiás. Além disso, é Cientista da Computação (Pontifícia Universidade Católica de Goiás), mestre em Engenharia Elétrica com ênfase em Visão Computacional (Escola de Engenharia – Universidade de Brasília/DF) e doutora em Engenharia de Sistemas Eletrônicos e Automação (Escola de Engenharia -Universidade de Brasília/DF) com ênfase em análise de dados de redes sociais e apoio à tomada de decisão.