UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
SISTEMAS DE INFORMAÇÃO

LUCAS GLASNER REGIS

TÉCNICAS DE NORMALIZAÇÃO DE TEXTO PARA PLN COM DADOS DO FACEBOOK: UM MAPEAMENTO SISTEMÁTICO DA LITERATURA

Recife
2022

LUCAS GLASNER REGIS

TÉCNICAS DE NORMALIZAÇÃO DE TEXTO PARA PLN COM DADOS DO FACEBOOK: UM MAPEAMENTO SISTEMÁTICO DA LITERATURA

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fernando Maciano de Paula Neto

Recife

2022

LUCAS GLASNER REGIS

TÉCNICAS DE NORMALIZAÇÃO DE TEXTO PARA NLP COM DADOS DO FACEBOOK: UM MAPEAMENTO SISTEMÁTICO DA LITERATURA

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Recife, 12 de Maio de 2022

BANCA EXAMINADORA

_____

Prof. Fernando Maciano de Paula Neto (Orientador)

UNIVERSIDADE FEDERAL DE PERNAMBUCO

_____

Prof. Frederico Luiz Gonçalves de Freitas (2º membro da banca)

UNIVERSIDADE FEDERAL DE PERNAMBUCO

AGRADECIMENTOS

Agradeço primordialmente a meus pais Oswaldo Régis e Márcia Glasner e a todos que de forma direta ou indireta iluminam minha jornada.

Aos professores Vinicius Garcia, José Carlos Cavalcante, Renato Vimieiro e Fernando por serem verdadeiros guias.

A Patrick Gouy e RECRUT.AI, Renan Hannouche, Lucas Dantas, Raphael Alencar, Túlio Hoffimann, Lucas Santos, Jorge Lobo e todos com quem já empreendi.

A meus grandes amigos Elther Oliveira e Mateus Ponciano (o Matt).

Finalmente, gostaria de agradecer a todos do CIn Centro de Informática da UFPE.

"Todos os modelos são incorretos,
mas alguns são úteis."
(George Box)

RESUMO

A Normalização de Texto (NT) é uma etapa importante para os algoritmos de aprendiza- gem de máquina e mineração de dados no processamento de linguagem natural (PLN). Ela padroniza os dados de entrada e tem mostrado melhorar o desempenho e a precisão dos modelos de PLN. Neste trabalho, realizamos um mapeamento da literatura das técnicas de NT e outros pré-processamentos no contexto de redes sociais, em específico, do Fa- cebook, utilizando como bibliotecas digitais os repositórios da ACM, Emerald, IEEE, Science Direct, Scopus e Springer de artigos publicados em periódicos e conferências.

Palavras-chave: Processamento de linguagem natural, Pré-processamento, Normalização de texto, Aprendizagem de máquina, Facebook

ABSTRACT

Text Normalization (TN) is an important step for machine learning and data mi- ning algorithms in natural language processing (NLP). It standardizes input data and has been shown to improve the performance and accuracy of NLP models. In this work, we conduct a literature mapping of TN techniques and other pre-processings in the context of social networks, specifically Facebook, using as digital libraries the ACM, Emerald, IEEE, Science Direct, Scopus and Springer repositories of articles published in journals and conferences.

Keywords: Natural Language Processing, Preprocessing, Text Normalization, MachineLearning, Facebook

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

UFPE	Universidade Federal de Pernambuco
PLN	Processamento de linguagem natural
NT	Normalização de texto
AM	Aprendizagem de máquina
AP	Aprendizagem profunda
IA	Inteligência artifical
WE	Word embedding - Representação vetorial de palavras
GloVe	Vetores Globais para Representação de Palavras
API	Interface de Programação de Aplicação
URL	Localizador Unificado de Recursos
IP	Protocolo de Internet
JSON	Notação de Objeto JavaScript
CSV	Valores Separados Por Virgula

SUMÁRIO

1 INTRODUÇÃO

Na era da inteligência artificial, geração de valor a partir de dados textuais escritos por humanos está cada vez mais em voga. Como sub-área de IA, Processamento de linguagem natural (PLN) tem suas próprias demandas de tratamento e uso de dados. Dentre estas, Normalização de Texto (NT) que é a etapa de pré-processamento de sistemas de PLN (Rahate and Chandak, 2019) tem um impacto significativo na qualidade geral desses modelos e sistemas (Li et al., 2018).

A etapa de pré-processamento e normalização de dados textuais é a fase onde, após a coleta de dados, se normaliza e padroniza as *strings* (cadeias de caracteres) com o intuito de diminuir tokens únicos e assim melhorar o desempenho, custo computacional, facilitar análises descritivas e melhorar a acurácia de modelos preditivos.

As formas mais comuns de NT envolvem remoção e substituição simples de carac- teres, técnicas de extração do radical de palavras, separação em tokens, além de métodos avançados de substituição e etiquetagem gramatical.

Dentre as várias fontes de dados textuais disponíveis hoje em dia, as redes sociais tem se mostrado uma grande geradora de valor com diversas aplicações para a área de PLN. Os dados advindos do Facebook, por ser a maior rede social usado no mundo, foram escolhidos como foco desse estudo. Tais dados gerados por usuários comuns sofrem com a informalidade, múltiplos tipos de erros, linguagem de internet e outros desafios que tornam a etapa de NT uma peça importante em tarefas de PLN (Krchnavy and Simko, 2017).

Este trabalho faz uma pesquisa sobre pré-processamento e normalização de textos publicados no Facebook e realiza um mapeamento da literatura das técnicas de NT, utilizando como bibliotecas digitais os repositórios da ACM, Emerald, IEEE, Science Direct, Scopus e Springer de artigos publicados em periódicos e conferências. Além disso, um possível caminho a ser seguido para o desenvolvimento de uma solução de NT é apresentado.

1.1 Motivação

O uso da Internet e das redes sociais vem evoluindo com o passar dos anos e, apesar de frequentes escândalos e má reputação com conflitos de interesse ao público, o Facebook

e outras redes sociais cada vez mais têm novos usuários em suas plataformas.

Este cenário tende a aumentar, algo que pôde ser observado com a globalização da pandemia da variante do Coronavírus SARS-CoV-2, no ano de 2020 e que continua ainda a pervadir o mundo mesmo no momento de publicação deste trabalho em 2022. Com isso, muitas áreas da sociedade que não utilizavam, tiveram de rapidamente passar por um processo forçado de transformação digital e se voltar ainda mais para o uso das tecnologias da Internet e redes sociais.

A quantidade, e potencial qualidade, de dados gerados por essas redes online tem um insurgente aumento e não dá sinais de diminuir a qualquer momento. Nesse contexto, as áreas de ciência dos dados, engenharia de aprendizagem de máquina (AM) e PLN ganham um grande impulso e que deve ser explorado enquanto ganha tração para os mais variados casos de uso.

Para tal, como sub-área de IA, PLN tem uma série de características únicas no ramo pois lida com dados de natureza específica, *strings* ou cadeias de caracteres em inglês, e necessita de limpeza e tratamento de dados, além de várias formas de enriquecê-los para garantir que o melhor potencial possa ser extraído dessas fontes.

O Facebook, como maior rede social global usada no mundo, foi escolhido como plataforma de onde os dados advindos serão focados neste mapeamento.

1.2 Objetivo

Este trabalho tem como objetivo realizar um mapeamento dos métodos propostos em outros estudos para pré-processamento e normalização de dados textuais do Facebook, e com isso realizar uma análise dos resultados obtidos a partir desse mapeamento. Dentro dos estudos analisados serão procuradas respostas para as seguintes questões:

- Q1: Após a NT, qual processamento é realizado?
- Q2: Qual o idioma dos dados utilizados?
- Q3: Como os dados utilizados foram coletados?
- Q4: Quais os desafios encontrados em NT com dados de redes sociais?
- Q5: Qual o impacto da NT na performance e acurácia dos modelos?

Ademais, visando apoiar o desenvolvimento de soluções de NT, houve um enfoque nas técnicas e métodos utilizados, gerando-se uma Questão Central de Pesquisa:

- QCP: Quais os principais algoritmos de pré-processamento e normalização de texto para PLN com dados da rede social Facebook?

Ao responder essas perguntas, este estudo servirá de apoio a quem se interessar em entender e avaliar os métodos atuais de pré-processamento e NT. Além disso, as respostas presentes neste trabalho servem de guia para quem vier a implementar alguma solução de NT com dados de redes sociais, em específico do Facebook.

O capítulo a seguir mostrará a metodologia utilizada para conduzir o estudo.

2 METODOLOGIA

O trabalho foi conduzido conforme o processo descrito em (Petersen et al., 2008). De acordo com os autores, uma revisão sistemática em engenharia de software deve ser dividida em cinco etapas: (i) definição das questões de pesquisa, (ii) realização de buscas por artigos primários relevantes nas fontes de pesquisa escolhidas, (iii) triagem dos artigos,
(iv) busca de palavras-chave nos resumos, e (v) extração de dados e mapeamento.

2.1 Pesquisa

Como fontes de artigos publicados em periódicos e conferências, a busca realizada incluiu seis bases de dados: ACM, Emerald, IEEE, Science Direct, Scopus e Springer. Todas essas bibliotecas digitais possuem páginas Web onde foi possível realizar as buscas pelas palavras-chave de interesse.

A pesquisa nessas fontes foi feita inicialmente com base nas palavras-chave: *NLP, Text Normalization e Facebook*. Após a análise dos resultados foram acrescentados: *Natu- ral Language Processing, Text Preprocessing, Machine Learning e Artificial Intelligence*, o que resultou na seguinte consulta:

("NLP" OR "Natural Language Processing") AND
("Text Normalization" OR "Text Preprocessing") AND
("Machine Learning" OR "Artificial Intelligence") AND ("Facebook")

A abreviação *NLP* foi estendida para ampliar resultados e mostrou-se necessário o acréscimo dos termos *Machine Learning e Artificial Intelligence* para garantir que, dentro do contexto de PLN, seriam retornados resultados que envolvessem treinamento de modelos, pois frequentemente o termo PLN é usado genericamente para qualquer operação com texto.

2.2 Triagem dos artigos

A primeira busca retornou um total de 8612 artigos, levando em consideração todas as fontes. Para selecionar os artigos mais relevantes para o estudo, foram realizadas três etapas de filtragem, descritas a seguir:

2.2.1 Etapa 1

Considerando que a primeira busca não levou em consideração nenhum filtro dis- ponível nas fontes pesquisadas, foi necessário realizar uma triagem por data, linguagem, tipo de documento e disponibilidade. Com isso os seguintes filtros foram aplicados:

- Apenas publicações realizadas nos últimos cinco anos
- Apenas artigos em inglês
- Apenas artigos e trabalhos de conferências
- Apenas artigos acess´ıveis (acesso livre ou para membros da UFPE)

A Tabela 1 mostra os resultados após a aplicacão desses filtros por fonte:

Springer	83
IEEE	327
ACM	4
Emerald	2
Scopus	32
Science Direct	86

Tabela 1: Triagem Etapa 1

Ainda nesta etapa, a base IEEE, em sua página Web, demandou uma etapa a mais que as outras envolvendo o desenvolvimento de uma *string* de busca com sintaxe diferente da oficial para garantir a associação das palavras-chave como nas demais bases. Com esta nova consulta, os resultados entraram na mesma ordem de grandeza que as outras e a seguir foram aplicados os mesmos filtros.

2.2.2 Etapa 2

O primeiro filtro não levou em consideração o conteúdo dos artigos, por isso foi necessário a avaliação dos títulos e resumos de cada artigo para verificar se estavam coe- rentes com o tema proposto. Nessa etapa foram retirados todos que não utilizavam dados advindos do Facebook, como por exemplo artigos que usavam exclusivamente conteúdo textual do Twitter, Reddit e outras redes sociais.

Nessa etapa foram selecionados apenas artigos que utilizavam dados do Facebook. A Tabela 2 mostra o resultado obtido:

Springer	6
IEEE	12
ACM	1
Emerald	0
Scopus	5
Science Direct	13

Tabela 2: Triagem Etapa 2

2.2.3 Etapa 3

As etapas anteriores não levaram em consideração a qualidade dos artigos selecionados, por isso foi realizada uma análise na introdução, conclusão e partes de cada artigo para verificar se as perguntas propostas nesse estudo seriam respondidas. Com este intento, foi utilizado mecanismos de busca textual por palavras-chave associadas as questões. Além disso, foram excluídos estudos secundários.

Nessa etapa, apenas artigos que responderam bem ao menos três perguntas foram selecionados. Os resultados dessa etapa estão apresentados na Tabela 3:

Springer	3
IEEE	5
ACM	0
Emerald	0
Scopus	4
Science Direct	8

Tabela 3: Triagem Etapa 3

Essa tabela mostra que a fonte com a maior quantidade de artigos selecionados foi a Science Direct com 8 artigos, enquanto que a Emerald e ACM apresentaram a menor quantidade não tendo nenhum artigo selecionado.

2.2.4 Resultado

O objetivo desse processo foi selecionar os artigos mais relevantes para responder as questoes propostas sobre métodos de pre-processamento e NT com dados do Facebook. Os resultados apresentados na próxima seção são referentes a análise apenas dos artigos selecionados.

A Figura 1 mostra um resumo do processo de triagem dos artigos por fonte.

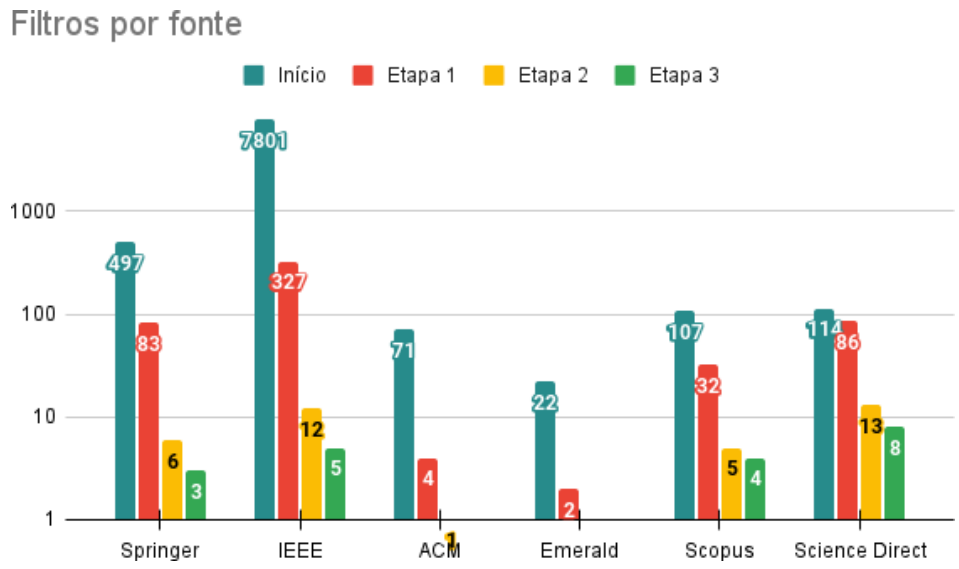

Figura 1: Resumo Triagem

A Tabela 4 mostra um identificador, o t´ıtulo, o ano e a fonte de cada artigoselecionado:

ID	T´ıtulo	Ano	Fonte
PS1	Abusive Comments Detection in Bangla-English Code-mixed and Transliterated Text Jahan et al. (2019)	2019	IEEE
PS2	Generate a list of Stop Words in Moroccan Dialect from Social Network Data Using Word Embedding NASSR et al. (2021)	2021	IEEE

PS3	Machine Learning-Based Automated Tool to De-tect Sinhala Hate Speech in Images Silva et al. (2021)	2021	IEEE
PS4	Sentiment Analysis of Social Network Posts in Slovak Language Krchnavy and Simko (2017)	2017	IEEE
PS5	Social Networking Sites Data Analysis using NLP and ML to Predict Depression Hossain et al. (2021)	2021	IEEE
PS6	Multi-level embeddings for processing Arabic so-cial media contents Moudjari et al. (2021)	2021	Science Direct
PS7	Multi-label Arabic text classification in Online Social Networks Omar et al. (2021)	2021	Science Direct
PS8	Geographic Disaggregation of Textual Social Me-dia Data: A Machine Learning-based ApproachZahir (2022)	2021	Science Direct
PS9	Covid-19 fake news sentiment analysis Iwendi et al. (2022)	2022	Science Direct
PS10	Context-sensitive normalization of social mediatext in bahasa Indonesia based on neural word embeddings Kusumawardani et al. (2018)	2018	Science Direct
PS11	An integrated framework of learning and eviden-tial reasoning for user profiling using short texts Vo et al. (2021)	2020	Science Direct
PS12	An integrated multi-node Hadoop framework to predict high-risk factors of Diabetes Mellitus using a Multilevel MapReduce based Fuzzy Clas- sifier (MMR-FC) and Modified DBSCAN algo-rithm Ramsingh and Bhuvaneswari (2021)	2021	Science Direct
PS13	A textual-based featuring approach for depres-sion detection using machine learning classifiersand social media texts Chiong et al. (2021)	2021	Science Direct

PS14	An effective Decision Support System for social media listening based on cross-source sentiment analysis models Ducange et al. (2019)	2018	Scopus
PS15	Machine Learning-Based Analysis of the Association between Online Texts and Stock Price Movements František Dařena and Žižka (2018)	2018	Scopus
PS16	Social Media Cross-Source and Cross-Domain Sentiment Classification Zola et al. (2019)	2019	Scopus
PS17	Suspicious Activity Detection of Twitter and Facebook using Sentimental AnalysisAl Mansooriet al. (2020)	2020	Scopus
PS18	Not All Swear Words Are Used Equal: Attention over Word n-grams for Abusive Language Identification Jarquín-Vásquez et al. (2020)	2020	Springer
PS19	SentiVerb system: classification of social mediatext using sentiment analysis Singh and Sachan (2019)	2019	Springer
PS20	Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Me-dia Comments in Spanish Tessore et al. (2022)	2020	Springer

Tabela 4: Triagem Resultado

3 RESULTADOS DA PESQUISA

Nesse capítulo serão mostrados os resultados das perguntas definidas no objetivo desse estudo, e também uma analise quantitativa e qualitativa desse resultado. Além disso, uma análise geral da pesquisa e do conteúdo dos artigos também será apresentada.

3.1 Importância do tema

Ao analisar os artigos selecionados foi confirmado que o tema de pré-processamento e NT, estudado nesse trabalho, tem ganhado cada vez mais relevância nos últimos anos. Ao olhar para o ano de publicação dos artigos selecionados, foi observado que a maioria foi publicada no ano de 2021. O ano de 2022 naturalmente apresenta um valor abaixoque 2021, pois a pesquisa foi realizada no primeiro semestre de 2022.

A Figura 2 mostra a quantidade de artigos por ano:

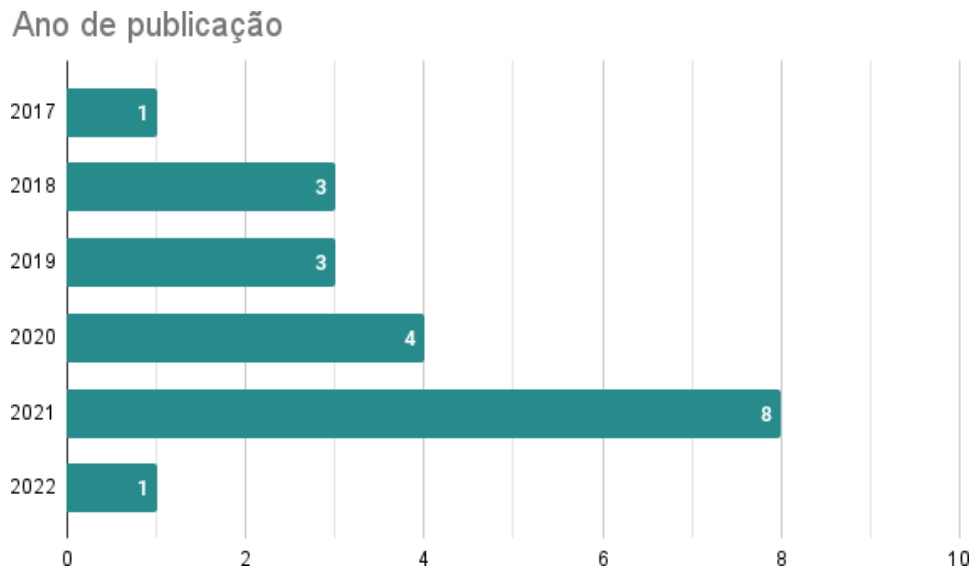

Figura 2: Ano de publicação

Esse aumento na quantidade de estudos que referenciam o tema confirma o que foi apresentado pelos autores Omar et al., que afirmam, em seu estudo publicado no ano de 2021, que a redução do uso de memória e requisitos de processamento, através do pré-processamento de dados, é de alta importância mesmo com os avanços de hardware
dos últimos anos (Omar et al., 2021). Os autores Ramsingh e Bhuvaneswari também

afirmam que a limpeza de dados ajuda a tornar os dados mais adequados para analytics, plataformas que ajudam a tomar decisões, cada dia mais requisitadas na era da ciência de dados (Ramsingh and Bhuvaneswari, 2021).

A próxima seção corresponde a análise realizada a partir das respostas coletadas da questão de pesquisa Q1: Após a NT, qual processamento é realizado?

3.2 Processamentos realizados após NT

A primeira questão, Q1, proposta nos objetivos desse trabalho visa entender que processamento é realizado nos estudos analisados para entender que técnicas têm sido usadas atualmente na NT com dados do Facebook. De forma geral, foi verificado que os artigos têm como objetivo principal realizar classificações com o uso de técnicas de inteligência artificial e aprendizagem de máquina para diversos casos de uso.

Essa clara tendência em utilizar aprendizagem de máquina para gerar valor com dados textuais de redes sociais é justificada pelo fato de que esses sistemas superam os métodos tradicionais, baseados no trabalho humano.

A maioria dos trabalhos analisados se usou de técnicas de classificação com AM como foi o caso de Chiong et al. que visava realizar detecção de depressão através de textos curtos (Chiong et al., 2021). Singh e Sachan criaram um sistema para classificação de textos de mídias sociais com análise de sentimento, no artigo *SentiVerb system: clas- sification of social media text using sentiment analysis* (Singh and Sachan, 2019).

Os autores Jahan et al., assim como Mansoori et al., buscavam realizar detecção de comentários abusivos e comportamento suspeito de outras atividades ilegais. O primeiro utilizou-se de classificação de AM, já o segundo, uma abordagem de análise de sentimento baseada em léxico (vocabulário).

Alguns estudos, como o dos autores Moudjari et al. e Iwendi et al., propuseram uma abordagem de aprendizagem profunda, com o intuito de realizar análise de sentimento.
Um dos artigos chegou a realizar análise baseada em AM da associação entre textos online e movimentos de preços de ações (František Dařena and Žižka, 2018).

Um pequeno conjunto de estudos focaram em realizar anotação de dados (*data annotation*, tarefa de etiquetar manualmente um conjunto de dados para treinamento). Um exemplo é o estudo dos autores Tessore et al., que desenvolveram um dataset em espanhol etiquetado com emoções atreladas à palavras e frases (Tessore et al., 2022).

Por fim, um único artigo dos autores Kusumawardani et al. se usou de técnicas de aprendizagem profunda para criação de Word Embeddings (representações vetoriais de palavras) (Kusumawardani et al., 2018).

A figura 3 abaixo mostra a quantidade de artigos que se encaixam em: Classificação de AM, Classificação de AP, Data Annotation, Criação de Word Embedding e uso de análise de sentimento baseada em léxico:

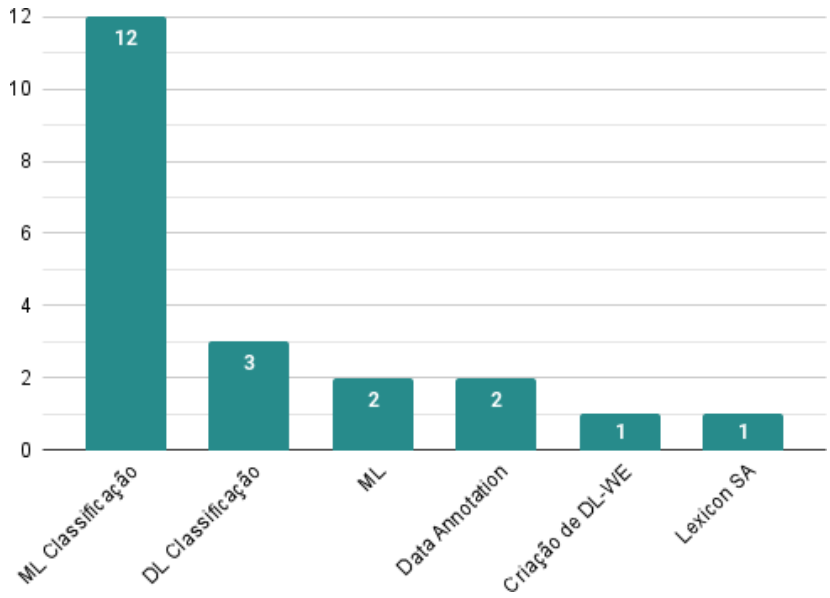

Figura 3: Processamentos realizados após NT

A seção a seguir mostra a análise realizada a partir das respostas para a segunda questão de pesquisa Q2: Qual o idioma dos dados utilizados?

3.3 Idiomas utilizados

Diferentes idiomas têm diferentes necessidades de pré-processamento e NT (Omar et al., 2021). Durante a realização das questões de pesquisa deste trabalho foi inferido que a depender do idioma, diferentes técnicas seriam utilizadas para se adequar às necessidades linguísticas do idioma em que os dados se encontravam. Apesar de não subdividir as técnicas que estavam sendo utilizadas à que linguagens, foi procurado entender quais línguas tem maior presença nas pesquisas científicas com dados do Facebook nos últimos anos.

A questão de pesquisa Q2 procura responder, de forma sucinta e direta, quais idiomas foram os mais encontrados e fazer uma análise geral de comentários sobre as especificidades das l´ınguas.

A maioria dos estudos analisados trabalhavam com dados em l´ıngua inglesa, se- guido imediatamente pelo árabe. Foi encontrado um número significativo de artigos que não explicitavam de forma clara com que idioma estavam trabalhando. Apesar da possibi- lidade de inferir que seriam em inglês ou na língua nativa do autor principal, foi preferível mantê-los como classe Não Responde para uma maior conformidade.

A Figura 4 mostra o resultado dos idiomas das bases de dados utilizadas nos artigos. Alguns artigos que que usaram dados com mais de uma l´ıngua foram contados 2 vezes em cada classe de idioma utilizado.

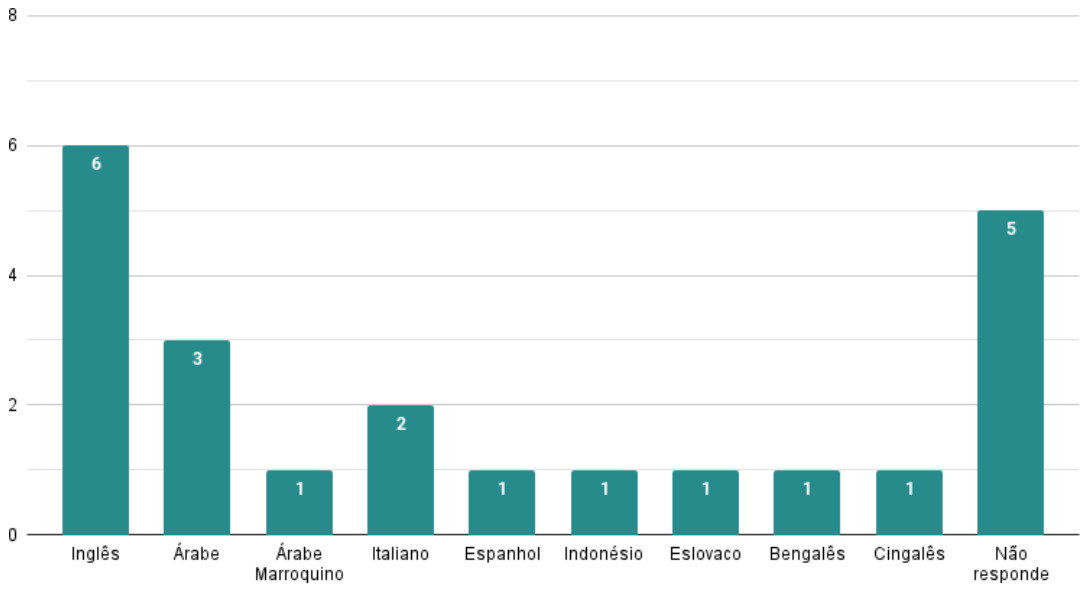

Figura 4: Idiomas utilizados

Observa-se uma alta variedade de idiomas mesmo com apenas 20 artigos seleci- onados, 10 classes foram contadas entre elas: 1 para Não Responde, 7 ramos de língua diferentes, 1 dialeto árabe e 2 indo-arianos.

O idioma inglês, confirmando ser o idioma mais utilizado globalmente para pes- quisas, teve 6 artigos seguido do árabe com também 6 artigos, 1 deles contudo, artigo dos autores NASSR et al., se especializava no dialeto marroquino, também conhecido com Darija, sendo contado separadamente o que expandiu ainda mais a alta variabilidade de

idiomas encontrados.

Dentre os autores de língua árabe, Omar et al. em seu artigo sobre multi-classificação de texto, pontuam que a língua árabe é altamente flexional e derivacional com várias for- mas de palavras e diacríticos (acentuações) que transformam radicalmente o sentido das palavras (Omar et al., 2021), dando forte ênfase a etapa de NT como ponto crucial para PLN em árabe.

Dentre os 20 artigos, 2 deles realizaram análises e experimentos com 2 idiomas. Os autores Jahan et al. em *Abusive Comments Detection in Bangla-English Code-mixed and Transliterated Text* realizaram detecção de comentários abusivos tanto em bengalês (língua oficial de Bangladesh) como em inglês, além de experimentar técnicas de tradução e mistura dos 2 idiomas (Jahan et al., 2019).

Ademais, os autores Zola et al. tiveram resultados de pesquisa bastante promissores com classificação de sentimento tanto em inglês quanto em italiano (Zola et al., 2019).

A próxima seção apresenta o resultado obtido das respostas da questão de pesquisa Q3: Como os dados utilizados foram coletados?

3.4 Meios de coleta de dados

A questão de pesquisa Q3 visa entender quais os meios empreendidos para coletar os dados da rede social Facebook. Esta é uma questão considerada delicada no meios científicos e de ciência de dados pois as APIs (interfaces de comunicação via programação) do Facebook são bastante restritivas, o que torna a coleta de dados muito limitada se comparadas a outras redes como o Reddit e Twitter que são bem mais permissivas.

5 artigos afirmaram utilizar a Facebook API. E, ratificando o ponto supracitado, os autores Ramsingh e Bhuvaneswari descreveram as limitações da API como 600 dados por IP e no máximo 50 operações por lote. Para contornar estas limitações os autores de- senvolveram um modelo linear para extrair o máximo de dados com a menor redundância possível. De forma automática, em diferentes pontos no tempo, requisições foram feitas e os dados eram extra´ıdos em JSON em tempo real (Ramsingh and Bhuvaneswari, 2021). Os pacotes de software utilizados no artigo foram descritos.

2 artigos afirmaram utilizar a técnica de Web crawling (rastreador Web) onde um programa testa diferentes URLs de forma a conseguir acessar endereços Web válidos. Os autores Vo et al. desenvolveram uma ferramenta para rastrear automaticamente dados

de usuários selecionados aleatoriamente no Twitter e Facebook (Vo et al., 2021).

Segundo Omar et al., a coleta de dados em mídias sociais online é um processo demorado e representa uma parte essencial da classificação de texto (Omar et al., 2021), por conseguinte, os autores programaram um rastreador Web (Web Crawler) que rola automaticamente para baixo páginas selecionadas do Facebook e Twitter mostrando todas as postagens e tweets e, em seguida, reúne todas as postagens, comentários e tweets e armazena-os em um arquivo de texto. Os autores coletaram 60.000 postagens em duas semanas.

Alguns poucos artigos explicitaram o uso de técnicas de *scraping* (varredura e extração de conteúdo de páginas Web) como meio de coleta, a exemplo Jahan et al. e Mansoori et al. É importante salientar que, apesar de ser possível inferir que os artigos que se referem ao meio de coleta via Web crawling também utilizaram Web scraping, foi preferido reportar isoladamente as classes na Figura 5.

Junto ao uso da Facebook API, também com 5 artigos, o uso de datasets prontos foi uma das fontes de dados coletados mais reportada. Pode-se inferir que alguns desses datasets utilizaram-se da API do Facebook para sua construção mas foi preferível, para efeito deste mapeamento, reportar os dados como a classe Datasets prontos.

Finalmente, 2 artigos reportaram coleta manual de dados e 4 não responderam. A Figura 5 mostra a quantidade de artigos para cada meio de coleta de dados.

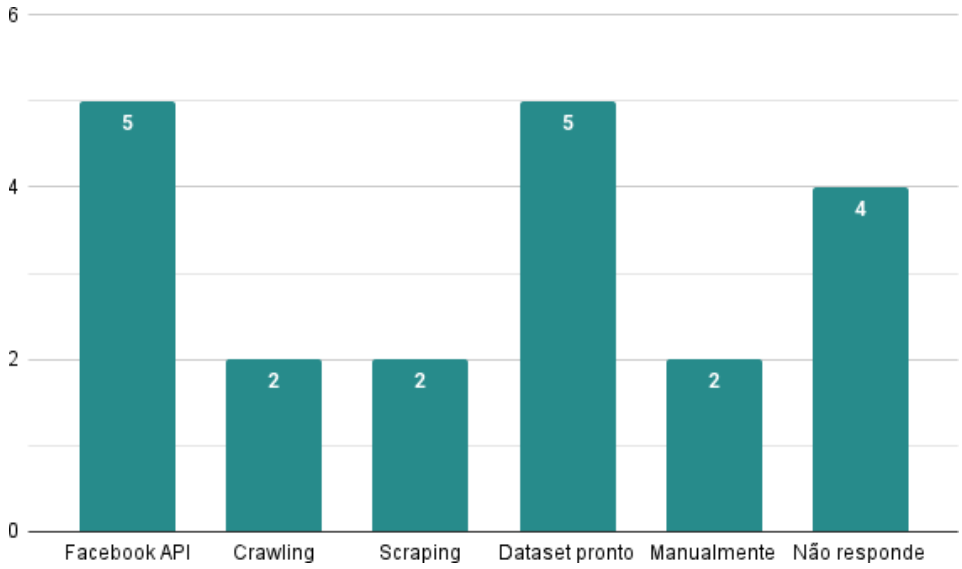

Figura 5: Meios de coleta de dados

A seção a seguir mostra a análise realizada das respostas para a questão de pesquisa Q4: Quais os desafios encontrados em NT com dados de redes sociais?

3.5 Desafios encontrados

A questão de pesquisa Q4 visa entender quais os principais desafios encontrados ao se tentar pré-processar e normalizar dados textuais advindos de redes sociais. Das questões propostas nesse estudo, essa foi a que teve a menor quantidade de respostas. Dos 20 artigos selecionados, 15 deles não apresentam os desafios encontrados no desenvolvimento de soluçoes para NT.

Os desafios relatados se resumem em: complexidade morfológica e linguagem de internet, que refere-se a informalidade gramatical de textos dos usuários na Web.

Os autores Silva et al. afirmam em seu artigo sobre detecção de discurso de ódio em cingalês (língua do Sri Lanka) que os textos dos usuários não respeitam regras de gramática, linguagem ou ortografia (Silva et al., 2021), dificultando a criação de regras de normalização e aumentando o número de *features* (classes) para termos iguais.

Na mesma linha, Jahan et al. afirmam que há desrespeito pela ortografia, gramática e pontuação corretas quando se trata de escrever comentários na internet (Jahan et al., 2019). Kusumawardani et al. acrescentam a este contexto o uso de gírias e abreviaçoes de internet (Kusumawardani et al., 2018), que complicam ainda mais o processo de NT já que frequentemente são gerados novos neologismos e expressões de acordo com as tendências vigentes.

Krchnavy e Simko, em seu artigo sobre análise de sentimento em eslovaco, são os que melhor expõem os desafios presentes na tarefa, ao afirmar que é preciso superar a complexidade morfológica das postagens, a qualidade da escrita e afirmam que a linguagem de internautas comuns muitas vezes sofre de informalidade, multiplicidade de estilos, neologismos, além de vários tipos de erros (Krchnavy and Simko, 2017).

A Figura 7 mostra a quantidade de respostas coletadas para cada desafio. Os artigos que mencionaram mais de um desafio foram contados uma vez para cada classe.

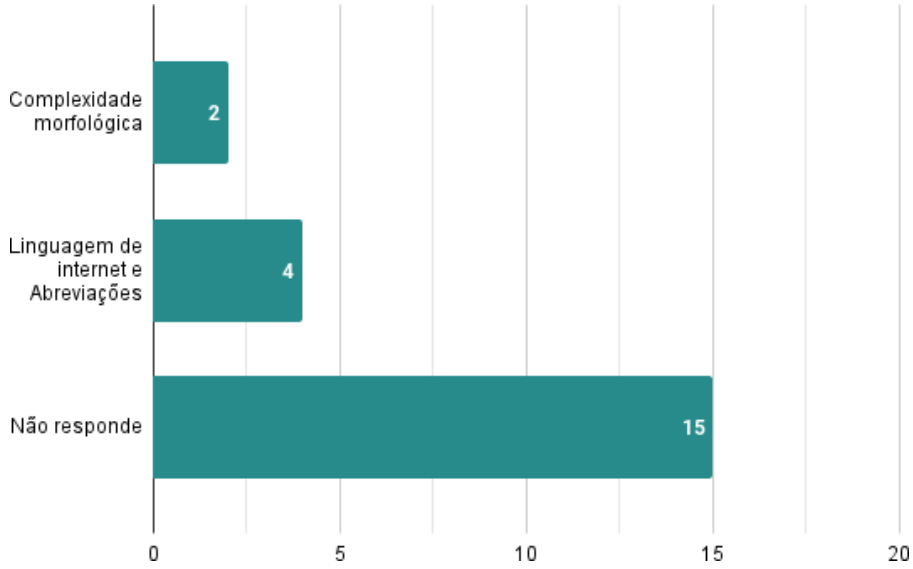

Figura 6: Desafios encontrados

A próxima seção apresenta os resultados encontrados a respeito da questão de pesquisa Q5: Qual o impacto da NT na performance e acurácia dos modelos?

3.6 Impacto na performance e acurácia

A questão Q5 tem como objetivo mapear o impacto que a etapa de NT tem sobre o desempenho (performance) e acurácia de modelos treinados (precisão associada à tarefa onde a normalização é aplicada), com e sem, esta etapa de tratamento e limpeza de dados.

Infelizmente, apesar de ser este o grande objetivo da etapa de pré-processamento de dados para PLN, melhorar a qualidade, performance e acurácia dos modelos, quase metade dos artigos selecionados não responde a questão, 10 deles afirmam melhorias tácitas sem medi-las e apenas 1 trabalho realiza medição do impacto na acurácia.

Os artigos de NASSR et al. e Hossain et al. afirmam categoricamente que a NT permite obter dados de boa qualidade e garante um melhor desempenho em análises (NASSR et al., 2021), assim como ajuda a aumentar a precisão do classificador. Apesar da asserção não há medições nem uma explicação do porque dessas conclusões.

Ainda na classe de artigos que afirmam melhorias sem medir, contudo expondo alguns pontos, Ramsingh e Bhuvaneswari trazem que a limpeza ajuda a melhorar a qua- lidade dos dados, ao convertê-los de modo adequado para análise, e afirmam melhorar

a precisão do resultado (Ramsingh and Bhuvaneswari, 2021). Trazendo um ponto rele- vante para entender a afirmação de melhoria tácita, Chiong et al. reportam que a NT tem um bom desempenho em trabalhos anteriores e é um passo essencial para reduzir a diversidade de palavras e facilitar o reconhecimento (Chiong et al., 2021). De fato, ao se diminuir a diversidade de palavras, em especial as que são iguais mas grafadas de maneira diferente ou equívoca, infere-se um menor número de classes para a mesma quantidade de instâncias melhorando desempenho e acurácia por conseguinte.

Zola et al. trazem o mesmo ponto ao descrever que a NT pode ajudar na redução do número de features no texto, aumentando o desempenho da classificação (Zola et al., 2019). Na mesma linha, Mansoori et al. urgem que há potencial impacto no desempenho final e afirmam que a etapa é de grande necessidade, principalmente se os dados forem textos retirados do Facebook e/ou Twitter, caracterizados como dados não estruturados (Al Mansoori et al., 2020).

Por fim, ainda nesta classe, contudo dando explanações mais contundentes, Omar et al. discorrem que o pré-processamento é uma etapa importante no processo de clas- sificação de texto. A principal importância de aplicar a NT é reduzir o número total de features no conjunto de dados, bem como melhorar o desempenho do classificador em termos de requisitos de recursos e precisão de classificação. Esta etapa [...] resulta na redução do número de termos com os quais o classificador precisa trabalhar. Consequente- mente, reduz os requisitos de memória e processamento do sistema de classificação (Omar et al., 2021).

A Figura 7 mostra a quantidade de artigos para cada classe de resposta sobre o impacto da NT na performance. Como pode-se ver vários artigos afirmam melhoria sem medir.

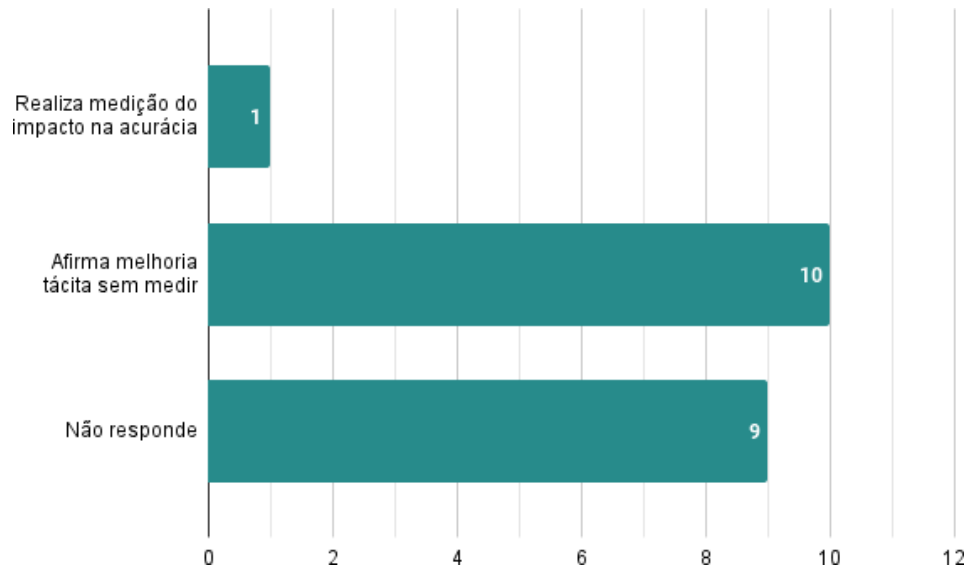

Figura 7: Impacto na performance

O único artigo que realiza avaliação e medição do impacto da NT na performance de classificadores é o grande *Sentiment Analysis of Social Network Posts in Slovak Language* que realizou o experimento com e sem cada etapa de pré-processamento. A influência das etapas de NT na acurácia foi confirmada como positiva para todos os classificadores usados por eles (Na¨ıve Bayes, Maximum Entropy e Support Vector Machine) em uma grandeza de até 12% no melhor caso.

Krchnavy e Simko confirmam pelos resultados que todas as etapas de pré-processamento influenciam no desempenho final da classificação. Em alguns casos o desempenho melhora um pouco, em outros a melhora é mais distinta. Curiosamente, para todas as abordagens baseadas em aprendizado de máquina (NB, ME, SVM) o desempenho foi alcançado com a mesma configuração utilizando as técnicas de normalização de emojis, substituição de diacríticos (acentuações) e extração de radical (Krchnavy and Simko, 2017). (Todas essas técnicas e métodos serão minunciosamente detalhadas na próxima sessão).

Ademais, os autores expandem a discussão trazendo que a normalização de emo- ticons introduziu melhorias de 0,64 a 3,65% para diferentes classificadores. A maior influência foi notada no caso do classificador Naïve Bayes. A melhora causada pela re- construção diacrítica variou de 1,33 a 9,80%. Esta etapa de pré-processamento mostrou-se particularmente útil para o classificador de Maximum Entropy. O mesmo se aplica à ex-

tração de radical. Quando desligado, o desempenho foi inferior, variando de 3,86% para SVM a 11,56% para classificador ME. Todas essas etapas de pré-processamento provaram ser importantes para a análise de sentimentos em eslovaco (Krchnavy and Simko, 2017).

A próxima seção apresenta os resultados encontrados a respeito da questão central de pesquisa, QCP: Quais os principais algoritmos de pré-processamento e normalização de texto para PLN com dados da rede social Facebook?

3.7 Questão central de pesquisa

Essa seção é dividida em 3 partes, na primeira é apresentada uma análise das classes de técnicas utilizadas para o pré-processamento e NT. Na segunda parte, uma análise da classe de Remoções e por fim, na terceira, um aprofundamento nos métodos específicos de Substituição encontrados nos artigos.

O objetivo de esmiunçar cada categoria de técnicas visa facilitar um possível ca- minho para o desenvolvimento de uma solução de NT.

A Figura 8 mostra a quantidade total de funções executadas em cada categoria de técnicas. 109 operações de NT foram executadas ao todo nos 20 artigos selecionados.

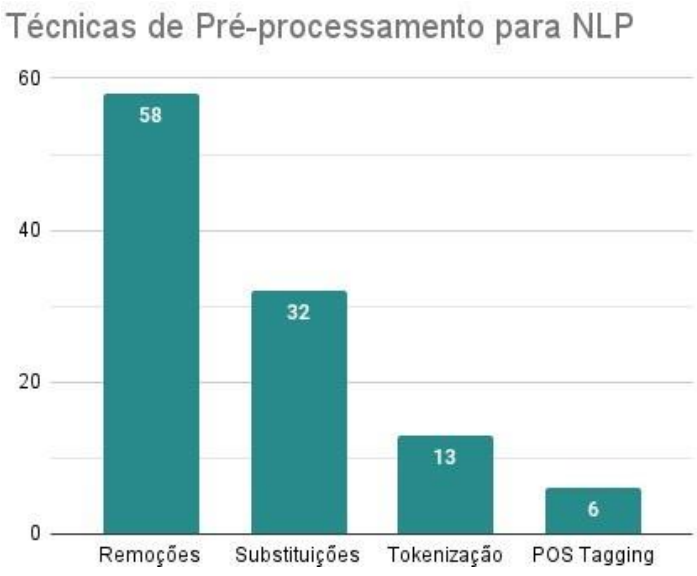

Figura 8: Técnicas de pré-processamento

A primeira categoria, ou classe, de técnicas de NT são chamadas de Remoção (por vezes chamada, limpeza) de texto (Iwendi et al., 2022), que se define como uma etapa

de remoção simples de caracteres (e.g. números, pontuações, emojis) ou remoçao da palavra inteira que contenha esses caracteres (e.g. remoção de hashtags, #love; menção de usuário, @username).

Da mesma forma, a segunda categoria de técnicas são chamadas de reconstrução ou Substituição, onde ao invés de se remover um caractere, ou palavra (conjunto de caracteres), substitui-se o mesmo por outro como é o caso de substituições simples por caracteres minúsculos e reconstrução de diacríticos (acentuações) e outras mais complexas como correções ortográficas, normalização de emojis por sentimentos, traduções e até extração do radical de palavras.

Essas transformações podem ser mais simples e diretas ou mais rebuscadas como remoção de HTML Tags, diferentes tipos de espaço em branco, todo e qualquer caractere não alfa numérico e URLs por exemplo.

As formas mais comuns presentes na literatura de se realizar essas operações são através de funções simples que recebem uma string e varrem toda a cadeia de caracteres para removê-los ou substituí-los. Frequentemente, o uso de expressões regulares é empre- endido direta ou indiretamente, pois muitas das funções prontas em algumas linguagens se utilizam do uso de regexes (expressões regulares) para executar tais operações.

Em termos de ferramental, as formas mais comuns presentes na literatura são realizadas através da biblioteca padrão da linguagem de programação Python, que contém seu módulo String. A linguagem C também é utilizada, apesar de pervadir menos o universo dos desenvolvedores, e em alguns casos vê-se também o uso da linguagem e ecossistema Java (Zahir, 2022).

Além dessas 2 classes, é importante evidenciar a presença de 2 outras categorias de NT que foram bastante vistas nos artigos selecionados. A Tokenização (neologismo, do inglês Tokenization) e a Etiquetagem gramatical (POS Tagging).

A Tokenização é um processo muito simples contudo bastante útil desde a geração de Analytics descritivos (plataformas de análise de dados e tomada de decisão que não envolvem modelos de IA) até mesmo modelos preditivos. A etapa de Tokenização realiza a separação de cada palavra isoladamente através de um separador (normalmente o espaço) e acrescenta aspas simples ao começo e fim da mesma, constituindo-se assim um token (Ducange et al., 2019). Essa etapa auxilia várias tarefas de PLN e tem-se provado uma operação presente em vários casos de uso. A nível de implementação, normalmente, a

Tokenização é feita se usando bibliotecas que tem a função já implementada. Dentro do ecossistema Python, a NLTK (Natural Language ToolKit) foi a ferramenta mais reportada para esta funcionalidade (Iwendi et al., 2022), entre outras como reportado nos artigos de Iwendi et al. e Hossain et al.

Por fim, a categoria com menos menção foi a Etiquetagem gramatical (POS Tag- ging) que pega partes-do-discurso humano e atrela automaticamente a ela classes gramaticais e outras informações como uma etiqueta que descreve aquela palavra. É uma técnica que tende a cair mais em desuso com o tempo, devido a métodos mais avançados de PLN (como Word Embeddings - Representações vetoriais de palavras) conseguirem inferir essas características através de vastas quantidades de dados textuais. Ainda assim, também é uma técnica que é utilizada através de ferramentas terceiras, a exemplo, a própria NLTK supracitada (Silva et al., 2021).

3.7.1 Remoções

As técnicas mais utilizadas de NT foram métodos de remoção e limpeza de texto. Dentre estas, podemos ver Remoções simples de caracteres ou palavra inteira como é o caso de números, pontuações, emojis, espaços em branco, e @usernames e hashtags, respectivamente.

Remoções mais rebuscadas com o uso de regexes mais complexas podem ser vistas no caso de remoção de URLs, HTML Tags e caracteres não alfanuméricos. Neste último, a nível de programação de sistemas de NT, é valido exemplificar que os autores Krchnavy e Simko fizeram uso da função *s.isalpha()* da biblioteca padrão de Python (Krchnavy and Simko, 2017), o que facilita bastante o uso de métodos de NT.

Por fim, um método diferenciado mas de alta relevância e presença nos artigos analisados é a Remoção de Stop Words (palavras de parada – tradução livre) que são palavras que podem ser consideradas irrelevantes, normalmente conjunções e preposições, que não agregam muita semântica ao texto (Krchnavy and Simko, 2017).

Apesar da presença nos artigos, o autor deste mapeamento aponta que a depender da arquitetura de PLN escolhida, é vital *manter* essas palavras como é o caso de estruturas com WE. A nível de implementação, a abordagem utilizada é remoção via presença em um dicionário pré-pronto. Mais uma vez se vê o uso da ferramenta NLTK, técnica de construção de dicionário manual, ou até mesmo fusão de termos presentes nas ferramentas

padrão acrescidas de termos customizados manualmente (Hossain et al., 2021).

A Figura 9 mostra a quantidade de Remoções de cada tipo executadas em todos os artigos.

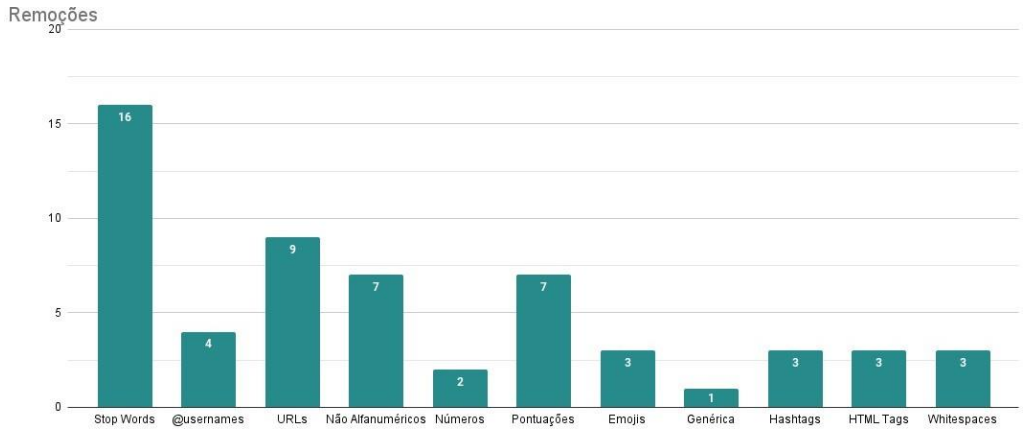

Figura 9: Remoções

3.7.2 Substituições

Por fim, as técnicas de substituição podem ser entendidas como Substituições sim- ples (e.g. substituição por caracteres minúsculos, reconstrução de diacríticos), Substi- tuições complexas como normalização de emojis para sentimentos, correção ortográfica, tradução e por último as técnicas de extração de radical (Stemming e Lemmatization).

As técnicas de extração de radical, Stemming e Lemmatization, foram as mais utilizadas. Stemming advém do inglês *stem*, que significa caule, e em essência visa extrair apenas o radical da palavra e não preocupar-se com suas inflexões e variedade gramatical de tempo e modo. O problema dessa abordagem é que com frequência ela substitui por uma palavra sem sentido intrínseco o que pode afetar análises de diversos tipos (Silva et al., 2021). Já a técnica de Lemmatization busca o lema da palavra mantendo sempre um significado junto ao radical. A nível de implementação essas técnicas frequentemente são aplicadas utilizando-se bibliotecas terceiras de software.

Foi observada uma técnica incomum de NT, uma substituição de hashtags em palavras separadas. Os autores Jarquín-Vásquez et al. descrevem que hashtags foram segmentadas em palavras para enriquecer o vocabulário (Jarquín-Vásquez et al., 2020). Apenas uma instância desta técnica foi observada.

Correções ortográficas ficaram em segundo lugar com 7 artigos mencionando uso e 1 artigo realizou experimentos com tradução de texto (Jahan et al., 2019).

A Figura 10 mostra a quantidade de Substituições de cada tipo executadas em todos os artigos.

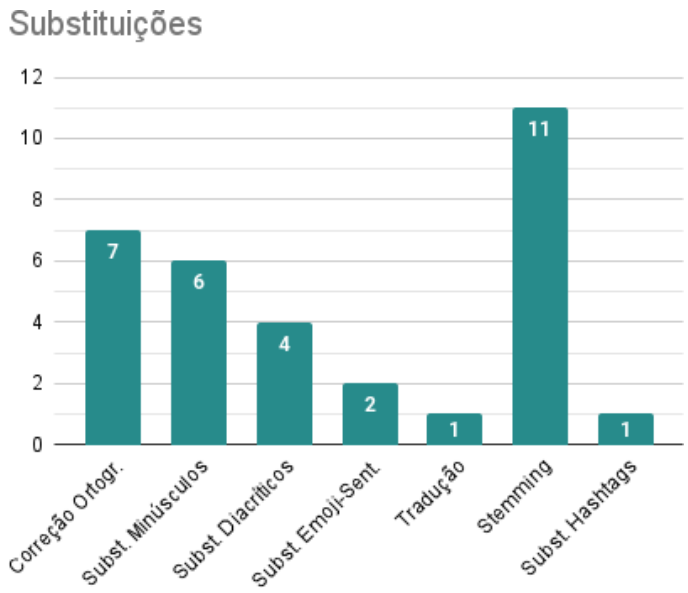

Figura 10: Substituições

4 ANÁLISE

Esse capítulo apresenta uma análise do autor deste trabalho a respeito do que foi encontrado nos resultados da pesquisa, e nas respostas coletadas no cap´ıtulo anterior. As próximas seções mostram uma análise sobre o estado da arte encontrado e apresentam uma possível forma de se criar uma solução para o pré-processamento e normalização de texto com dados do Facebook.

4.1 Estado da Arte

A pesquisa realizada sobre o pré-processamento e NT com dados de rede sociais, e a análise das respostas obtidas para as perguntas propostas nesse trabalho, permitiram um entendimento do atual estado da arte a respeito desse tema. Foi possível entender como a NT tem sido desenvolvida, as técnicas utilizadas nesta etapa e os desafios que podem ser encontrados.

4.1.1 Desempenho dos modelos após NT

Em termos de acurácia, as soluções propostas apontam um bom desempenho na diminuição de tokens únicos, o que se direciona para melhoria de performance e acurácia, além de 1 artigo que prova através de avaliações uma melhora de até 11,56% em tarefas de classificação de AM. Vale ressaltar, porém, que o impacto na acurácia não foi medido na vasta maioria dos artigos e, uma vez que, custo computacional e acurácia são, de forma geral, as métricas que norteiam a melhoria provida pelas etapas de pré-processamento, estas devem ser melhor avaliadas em outros estudos que realizem tal medição. O enfoque no uso de dados da rede social Facebook é visto como grande limitação deste estudo para análises mais profundas de desempenho e acurácia, pois a mesma é uma das que tem acesso menos permissível para coleta de dados (Ramsingh and Bhuvaneswari, 2021).

A acurácia e performance foram escolhidas como medida para avaliar o desempenho dos modelos nesse trabalho, porque é a métrica mais utilizada, e ao menos um estudo pode comprovar a eficácia da etapa de NT para melhorá-las, além de outros que referenciaram de forma tangencial a medida.

De toda forma, o desempenho das etapas de NT analisadas se mostra encorajador,

e é possível afirmar que as técnicas apresentadas são efetivas até certa extensão (baseando- se na diminuição de tokens/features únicos) ao melhoramento de análises, treinamento e predições com modelos de PLN.

Apesar dos resultados observados, em sua maioria, as soluções propostas pelos estudos analisados não realizaram medições e avaliações comparativas. Essa questão foi apontada como uma falta nos estudos atuais, e um ponto de melhoria para estudos futuros. Além disso, esses estudos não apresentaram o uso dos modelos ao longo do tempo, o que iria, provavelmente, apontar uma necessidade de retreinamento para os classificadores, visto que a linguagem de internet e redes sociais tende a transformar-se rapidamente.

4.1.2 Técnicas utilizadas

Conforme afirmado em 3.2, as soluções atuais de NT têm sido voltadas para a otimização de modelos de aprendizagem de máquina, especialmente tarefas de classificação de AM. Na seção 3.7 foram apresentadas as principais categorias de técnicas e métodos para NT com dados de redes sociais nos estudos analisados. Ficando assim, claro, queas técnicas de remoção e substituição utilizadas no pré-processamento e NT são muito importantes para melhorar o desempenho final de modelos de PLN.

A análise realizada permite afirmar que, com apenas transformações básicas nas cadeias de caracteres, é possível desenvolver um modelo que atinja uma acurácia satis- fatória. Entretanto, a combinação de técnicas de remoção, substituição simples, substi- tuição avançada, além de Tokenização, e, a depender da arquitetura, etiquetagem grama- tical, apresentam melhor resultado.

No que tange as técnicas de coleta de dados do Facebook, apesar da limitaçao de acesso imposta por sua API, a mesma parece ser uma via básica inicial de se angariar dados. E tendo em vista essas restrições, uma combinação de técnicas de acesso HTTP automatizado, Web crawling e Web scraping se mostrou bastante promissora como forma de atingir a maior quantidade possível de dados obtendo menor redundância possível. Ademais, diferentes fontes de dados pré-prontas (datasets) que aportam conteúdos da rede se mostra como via alternativa ou até complementar aos métodos expostos.

Na área linguística idiomática foi encontrada uma alta variedade de diferentes idiomas realizando as mesmas transformações fundamentais de NT. O que nos permite ver o interesse nas mais diversas áreas do mundo por dados do Facebook para geração

de valor e pesquisas. Ademais, pode-se inferir que soluções de NT a serem desenvolvidas podem sim atender até as mais remotas línguas com as devidas configurações setadas.

4.2 Construção de uma solução de NT

A análise realizada, a partir dos resultados da pesquisa, permitiu o entendimento necessário para o processo de desenvolvimento de uma solução de pré-processamento e NT com dados do Facebook. Para a construção de uma pipeline de NT, será preciso, inicialmente, coletar os dados da plataforma. Na seção 3.4, foram identificadas meios de coleta das quais é possível extrair dados para compor a base. Buscas na internet por datasets prontos devem sempre ser uma das primeiras opções escolhidas pois contornam a necessidade de coleta direto da rede social. Ademais, uma combinação de técnicas automatizadas de acesso a Facebook API mais Web crawling e Web scraping em paralelo irão conduzir o desenvolvedor a um bom número de entradas para compor sua base de dados.

A seguir, deve-se realizar uma análise exploratória dos dados e visualizar a variabilidade das cadeias de caracteres. Esta etapa de ciência de dados em PLN é cabal para se otimizar que técnicas e métodos específicos serão escolhidos e utilizados.

A partir da base de dados rica em material da rede social e uma visualização da distribuição de diferentes tipos de caracteres, é necessário implementar, em código, scripts com funções que recebam a string e a retornem com a transformação desejada. Um método por funcionalidade a ser executada, e operá-la proceduralmente da primeira função até a última iteradamente até a conclusão. A linguagem Python é sugerida para este intento.

No que tange as melhores funções para NT com dados do Facebook, cabe ao cientista analisar o objetivo final da tarefa de PLN a ser realizada. Aqui, o autor deste mapeamento sugere os métodos mais comuns a se aplicar, supondo um cenário de treinamento de modelo de classificação em língua inglesa: inicialmente, substituições simples de substituição por todos os caracteres para minúsculos, substituição de diacríticos (pois a língua não tem variedade gramatical dada por acentos), não é recomendado substituir ou remover emojis pois estes podem ser features do modelo, caso haja possibilidade, substituição de termos escritos errados via correção ortográfica, expansão de acrônimos via dicionário manual dos termos de internet mais comuns vigentes.

Em seguida, remoções simples de espaços em branco excessivos, HTML Tags, URLs e números. Neste ponto os dados já estão bastante enriquecidos com a limpeza e NT, ademais, a depender do caso de uso, remoção de menções @username e hashtags do tipo #love, *ou*, remoção apenas dos caracteres puros @ e #.

Ao final, em caso de arquitetura que não envolva Word Embeddings, é sugerida remoção avançada de Stop Words com auxílio da biblioteca de código aberto NLTK do ecossistema Python e, potencialmente, a remoção de todos os caracteres não alfa- numéricos. Salvar o dataset em arquivo texto tipo CSV.

Por fim, também usando a ferramenta NLTK, executar a etapa de Tokenização imediatamente antes da entrada no treinamento do modelo de classificação.

5 CONCLUSÃO E TRABALHOS FUTUROS

A geração de valor com dados textuais gerados por usuários na internet, em especial Facebook e redes sociais, tem se mostrado uma promissora fonte tanto na academia quanto no mercado. O uso da Internet e das redes sociais tende a se tornar cada vez mais amplo,o que pôde ser observado com a ocorrência da pandemia de Covid-19, no ano de 2020 até meados de 2022, na qual muitos setores tiveram de se voltar ainda mais para o uso dessa tecnologia. Com isso, o tema de PLN e IA com dados das redes ganha uma importância ainda maior do que era atribuída anteriormente, em especial para a área de ciência de dados, engenharia de AM e PLN que têm em mãos um grande recurso para gerar valor aos mais variados casos de uso.

Esse trabalho explorou as técnicas conhecidas para a pré-processamento e norma- lização de texto com dados de redes sociais, e apresentou uma análise a respeito do estado atual das soluções existentes com o objetivo de mapear tais técnicas, e apresentar uma referência para futuros estudos que tenham como objetivo desenvolver uma soluçao de NT, em especial para modelos de aprendizagem de máquina. Além disso, foi constatado um aumento de estudos relacionados ao tema, o que mostra que o pré-processamento e normalização de texto são muito importantes para garantir a qualidade de modelos de PLN.

A análise das soluções e técnicas encontradas possibilitou a conclusão de que as pipelines de NT existentes conseguem realizar melhorias no desempenho e acurácia dos modelos de PLN de forma efetiva, apresentando uma boa evolucão na acurácia dos mode- los. Além disso, foi possível concluir que os métodos para NT realizados antes do treina- mento dos classificadores são de grande importância no resultado obtido, embora tenha sido observado uma grande falta de avaliacões comparativas com e sem a normalização.

Os passos para a criação de uma solução para o tema apresentado podem ser resumidos em: 1. Coleta de dados via; buscas na Internet por datasets prontos, coletasde dados via Facebook API, Web crawling e scraping, 2. análise exploratória dos dados textuais e estratégia de normalização, 3. programação procedural de métodos de remoção e substituição de caracteres, 4. Tokenização, 5. treinamento do modelo e, por fim, avaliação dos resultados com e sem a NT é uma abordagem sugerida por esse trabalho para o desenvolvimento de uma solução.

5.1 Trabalhos Futuros

Algumas sugestões de possíveis trabalhos futuros são listadas a seguir:

- Trabalho para pré-processamento e NT em ambiente realtime, alimentado com dados vigentes de redes sociais como Twitter e Facebook.
- Estudo com avaliações comparativas com e sem cada método de pré-processamento e normalização de texto.
- Estudo para otimização da NT com uso de AP não-supervisionado para traçar as features ideais de normalização.
- Estudo para verificação da necessidade de retreinamento de modelos treinados com estratégia de NT traçada a mais de 1 ano.

REFERÊNCIAS

Saeed Al Mansoori, Afrah Almansoori, Mohammed Alshamsi, Said A Salloum, and Kha- led Shaalan. Suspicious activity detection of twitter and facebook using sentimental analysis. *TEM Journal*, 9(4):1313, 2020.

Raymond Chiong, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. A textual- based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135:104499, 2021.

Pietro Ducange, Michela Fazzolari, Marinella Petrocchi, and Massimo Vecchio. An effec- tive decision support system for social media listening based on cross-source sentiment analysis models. *Engineering Applications of Artificial Intelligence*, 78:71–85, 2019.

Jak Přichystal František Dařena, Jonáš Petrovský and Jan Žižka. Machine learning-based analysis of the association between online texts and stock price movements. *Inteligencia Artificial*, 21(61):95–110, 2018.

Md Tazmim Hossain, Md Arafat Rahman Talukder, and Nusrat Jahan. Social networking sites data analysis using nlp and ml to predict depression. pages 1–5, 2021.

Celestine Iwendi, Senthilkumar Mohan, Ebuka Ibeke, Ali Ahmadian, Tiziana Ciano, et al. Covid-19 fake news sentiment analysis. *Computers and electrical engineering*, 101: 107967, 2022.

Maliha Jahan, Istiak Ahamed, Md Rayanuzzaman Bishwas, and Swakkhar Shatabda. Abusive comments detection in bangla-english code-mixed and transliterated text. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–6. IEEE, 2019.

Horacio Jesús Jarquín-Vásquez, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. Not all swear words are used equal: Attention over word n-grams for abusive language identification. pages 282–292, 2020.

Rastislav Krchnavy and Marian Simko. Sentiment analysis of social network posts in slovak language. pages 20–25, 2017.

Renny Pradina Kusumawardani, Stezar Priansya, and Faizal Johan Atletiko. Context- sensitive normalization of social media text in bahasa indonesia based on neural word embeddings. *Procedia computer science*, 144:105–117, 2018.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-endtask-completion neural dialogue systems. 2018. URL http://arxiv.org/abs/1703. 01008.

Leila Moudjari, Farah Benamara, and Karima Akli-Astouati. Multi-level embeddings for processing arabic social media contents. *Computer Speech & Language*, 70:101240, 2021.

Zineb NASSR, SAEL Nawal, and Faouzia BENABBOU. Generate a list of stop words in moroccan dialect from social network data using word embedding. In *2021 Internatio- nal Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*, pages 66–73. IEEE, 2021.

Ahmed Omar, Tarek M Mahmoud, Tarek Abd-El-Hafeez, and Ahmed Mahfouz. Multi- label arabic text classification in online social networks. *Information Systems*, 100: 101785, 2021.

Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, 17, 06 2008.

Pooja Rahate and M. Chandak. Text normalization and its role in speech synthe- sis. 2019. doi: 10.35940/ijeat.E1029.0785S319. URL https://www.ijeat.org/ wp-content/uploads/papers/v8i5S3/E10290785S319.pdf.

J Ramsingh and V Bhuvaneswari. An integrated multi-node hadoop framework to predict high-risk factors of diabetes mellitus using a multilevel mapreduce based fuzzy classifier (mmr-fc) and modified dbscan algorithm. *Applied Soft Computing*, 108:107423, 2021.

Evangeli Silva, Maheshi Nandathilaka, Sandupa Dalugoda, Thanu Amarasinghe, Supun- mali Ahangama, and G Thilini Weerasuriya. Machine learning-based automated tool to detect sinhala hate speech in images. pages 1–7, 2021.

Shailendra Kumar Singh and Manoj Kumar Sachan. Sentiverb system: classification of social media text using sentiment analysis. *Multimedia Tools and Applications*, 78(22): 32109–32136, 2019.

Juan Pablo Tessore, Leonardo Mart´ın Esnaola, Laura Lanzarini, and Sandra Baldassarri. Distant supervised construction and evaluation of a novel dataset of emotion-tagged social media comments in spanish. *Cognitive Computation*, 14(1):407–424, 2022.

Duc-Vinh Vo, Jessada Karnjana, and Van-Nam Huynh. An integrated framework of lear- ning and evidential reasoning for user profiling using short texts. *Information Fusion*, 70:27–42, 2021.

Jihad Zahir. Geographic disaggregation of textual social media data: A machine learning- based approach. *Procedia Computer Science*, 198:367–372, 2022.

Paola Zola, Paulo Cortez, Costantino Ragno, and Eugenio Brentari. Social media cross- source and cross-domain sentiment classification. *International Journal of Information Technology & Decision Making*, 18(05):1469–1499, 2019.