Um *Pipeline* de Pré-Processamento de Dados Textuais em Português para Análise de Redes Sociais

Livia A. dos Santos[1], Orlando B. Coelho *(in memoriam)*[1]**, Renata Araujo[1,2], Ivan Carlos A. Oliveira[1]**

[1]Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie
São Paulo - SP - Brasil.

[2]Programa de Pós-Graduação em Sistemas de Informação - EACH/USP
São Paulo - SP - Brasil

liviaalabarse.santos@mackenzista.com.br, {renata.araujo, orlando.coelho, ivan.oliveira}@mackenzie.br

Abstract. *Preprocessing is a fundamental step in processing textual data, especially when working with text analysis, data mining or machine learning. In particular, textual data from social networks offers challenges to pre-processing, mainly due to its informal structure. This article presents a pipeline to perform 9 basic processing activities to guarantee the quality and consistency of brazilian Portuguese textual data sets extracted from social networks. Tests were conducted on datasets containing 8,000, 20,000, and 60,000 tweets, demonstrating the pipeline's performance in terms of accuracy, noise reduction, and processing time.*

Resumo. *O pré-processamento é uma etapa fundamental no processamento de dados textuais, especialmente quando se trabalha com análise de textos, mineração de dados ou aprendizagem de máquina. Em particular, os dados textuais provenientes das redes sociais oferecem desafios ao pré-processamento, principalmente devido a sua estrutura informal. Este artigo apresenta um pipeline compreendendo 9 atividades básicas de processamento para garantia da qualidade e consistência de conjuntos de dados textuais em português brasileiro extraídos de redes sociais. O pipeline foi testado em conjuntos de 8 mil, 20 mil e 60 mil tweets, demonstrando sua eficácia em termos de precisão, redução de ruído e tempo de processamento.*

1. Introdução

O pré-processamento ou preparação de dados textuais é uma das etapas mais dispendiosas e importantes no ciclo de vida de aplicações de ciência e análise de dados devido a sua natureza não-estruturada. Ela é potencializada quando se utiliza textos publicados em redes sociais. O processamento de textos de redes sociais apresenta uma variedade de desafios únicos, destacando-se a presença de gírias, erros ortográficos, emojis, abreviações e a natureza informal da linguagem. Desta forma, a etapa de pré-processamento pode conter diversas fases, tais como: remoção de URLs, exclusão de *stopwords* e símbolos especiais [Kurniawan 2020].

Muitas pesquisas envolvendo a mineração de dados textuais em redes sociais apresentam o pré-processamento nos idiomas inglês, chinês, espanhol e outras línguas [Yang e Zang 2018][Shen et. al. 2019][Kurniawan 2020][Osakwe e Cortes 2021]. No Brasil, as pesquisas na área de análise de redes sociais têm abordado esses desafios,

destacando a necessidade da limpeza e pré-processamento adequados para garantir a qualidade dos dados e o bom desempenho de análises subsequentes [Garcia et. al. 2023] [Cardozo e Freitas 2021][Nascimento et. al. 2021][Medeiros e Borges 2019][Kansaon et. al. 2018][Souza et. al. 2017]. No entanto, são poucas as pesquisas que organizam e disponibilizam para uso das comunidades científicas ou de prática, seus *pipelines* de processamento. Neste artigo, apresentamos um *pipeline* de pré-processamento de dados textuais no idioma português brasileiro para o estudo e a análise de mensagens publicadas em redes sociais e, a partir disso, permitir que aplicações que fazem uso desses dados possam extrair conhecimentos mais assertivos do seu conteúdo.

O artigo está organizado em mais três seções. A Seção 2 descreve o *pipeline* proposto, as bibliotecas e ferramentas indicadas em cada fase. A Seção 3, destaca os testes realizados com o uso do *pipeline*. A Seção 4 apresenta as conclusões, propostas de melhorias e trabalhos futuros.

2. Descrição do *Pipeline*

O *pipeline* desenvolvido neste trabalho[1] tem o objetivo de garantir a qualidade e consistência de conjuntos de dados textuais em português brasileiro advindos de quaisquer redes sociais, mas sua aplicabilidade vai além e pode ser empregado em diferentes bases de dados textuais. Ele combina bibliotecas/ferramentas para lidar com desafios de sua manipulação:

- **NLTK (Natural language Toolkit) v: 3.8.1 (**https://www.nltk.org/**):** Utilizada para processamento de linguagem natural. Oferece uma variedade de ferramentas e recursos, incluindo tokenização e *stopwords*, que são essenciais para o pré-processamento de texto.
- **Demoji v: 1.1.0 (**https://pypi.org/project/demoji/**):** Empregada para lidar com emojis presentes nos dados textuais. Fornece métodos para mapear emojis para rótulos específicos.
- **Enelvo v: 0.15 (**https://pypi.org/project/enelvo/**):** Ferramenta desenvolvida para normalização de textos em português, com problemas como erros ortográficos, gírias da internet e siglas.
- **Cryptography v: 42.0.5 (**https://pypi.org/project/cryptography/**):** Responsável por proteger o conteúdo do dicionário de usuários "user_dict", garantindo que os dados sensíveis não possam ser facilmente acessados ou lidos por terceiros não autorizados.

O *pipeline* de pré-processamento é composto por etapas que transformam o texto bruto em uma representação para análise subsequente. Etapas ilustradas a seguir.

1. **Substituição de Vírgulas:** todas as vírgulas são temporariamente substituídas por um rótulo ("*chavevirg*") para evitar conflitos durante a normalização de texto com a ferramenta Enelvo relacionadas a números com vírgulas. A substituição temporária facilita a manutenção da integridade dos dados numéricos e é revertida posteriormente. Ex. **Entrada:** ["quanto foi? 5,70?", "que triste, queria ter ido"] **Saída:** ["quanto foi? 5chavevirg70", "que tristechavevirg queria ter ido"]

[1] https://github.com/ciberdem/ProjetoHEIWA-FAPESP/tree/main/CuradoriaExtracaoDados

2. **Normalização com *Enelvo*:** Utiliza a ferramenta *Enelvo* para normalizar erros ortográficos, gírias, siglas e outros aspectos do texto. Ex. **Entrada:** [“ururguau”, “disculpa qq coisa!”, "Vc eh muitooooo legal", "Oii, To trabahlando hj"] **Saída:** [“uruguai”, “desculpa qualquer coisa”, “você é muito legal”, “oi to trabalhando hoje”]
3. **Substituição de *Emojis*:** Substitui *emojis* encontrados no texto por rótulos específicos para uniformizar sua representação. Ex. **Entrada:** ['😀', '😋', ':)', ':('] **Saída:** ['grinningface', 'facesavoringfood', 'emojipositivo', 'emojinegativo"]
4. **Substituição de Usuários:** Anonimiza usuários mencionados no texto (@usuário), substituindo-os por rótulos específicos, com um dicionário de usuários já mapeados em um arquivo criptografado chamado '*user_dict.txt*' para recuperar o rótulo correspondente. Caso contrário, cria um novo rótulo e armazena no dicionário. Ex. **Entrada:** ['oi @maria', 'gostei mt de vcs @pedro @maria', ‘vamos pra praia @pedro @julia @maria?’] **Saída:** [‘oi @user1', ‘gostei mt de vcs @user2 @user1’, ‘vamos pra praia @user2 @user3 @user1?’]
5. **Remoção de *URLs*:** Remove *URLs* do texto. Ex. **Entrada:** [‘amei essa música! https://www.youtube.com/watch?v=dQw4w9WgXcQ’] **Saída:** [‘amei essa música!’]
6. **Reversão da substituição de Vírgulas:** Restaura as vírgulas substituídas no início do *pipeline*. Ex. **Entrada:** [“quanto foi? 5chavevirg70”, “que tristechavevirg queria ter ido”] **Saída:** [“quanto foi? 5,70?”, “que triste, queria ter ido”]
7. **Remoção de pontuação e Caracteres Especiais:** Remove pontuação e caracteres especiais, exceto quando são partes de *hashtags*, datas ou números com vírgula. Ex. **Entrada:** [“vai ser dia 20/05?”, “que divertido!!!! #praia”, “ quanto foi? 5,70?”] **Saída:** [“vai ser dia 20/05”, “que divertido #praia”, “quanto foi 5,70”]
8. **Remoção de *Stopwords*:** (Opcional) Remove *Stopwords*, palavras que não contribuem significativamente para o significado do texto. Ex. **Entrada:** [“vou para praia hoje”, “vou parar de fazer isso”] **Saída:** [“vou praia hoje”, “vou parar fazer”]
9. ***Tokenização***: (Opcional) Divide o texto em palavras, *hashtags*, datas e números com vírgula. Cria uma lista de itens separados. Ex. **Entrada:** [“vou para praia hoje”, “vou parar de fazer isso”] **Saída:** [“vou, para, praia, hoje”, “vou, parar, de, fazer, isso”]

Embora o *pipeline* proposto tenha se mostrado eficaz em vários aspectos do pré-processamento, a ferramenta *Enelvo* apresentou algumas limitações durante a normalização de textos. Em particular, identificamos que a ferramenta não lida bem com certos tipos de gírias e abreviações frequentes em redes sociais. Isso sugere a necessidade de explorar alternativas para complementar a *Enelvo*, como o *Hunspell (*https://hunspell.github.io/*)*, para atender às peculiaridades do português brasileiro usado em redes sociais.

3. Uso do *pipeline*

Para avaliar a eficácia do *pipeline* de pré-processamento de textos em redes sociais, três métricas principais foram utilizadas, sendo elas:

- **Precisão da *Enelvo***, mede o percentual de palavras que foram corretamente normalizadas. Nesta avaliação, foi utilizado um conjunto de dados contendo 100

frases com erros ortográficos e gírias, comparando os resultados obtidos pela ferramenta com os resultados esperados. A ferramenta atingiu uma precisão de 85%, indicando boa capacidade de correção, embora haja espaço para melhorias, especialmente em casos de gírias ou abreviações com mais de um significado.

- **Tempo de Processamento**, avaliou o seu desempenho em termos de tempo de processamento. Ao lidar com conjuntos contendo: a) 8 mil *tweets*, o *pipeline* foi capaz de concluir o processo em 11 minutos; b) 20 mil *tweets*, concluiu o processo em cerca de 30 minutos; e c) 60 mil *tweets*, o tempo de processamento aumentou para aproximadamente 1 hora e 42 minutos. Todos esses experimentos foram conduzidos utilizando o *Google Colab*[2], em sua versão gratuita, como ambiente de desenvolvimento e execução.
- **Redução de Ruído**, avaliou a sua eficiência em remover elementos indesejados, como URLs e caracteres especiais. Para isso, foi utilizado um outro conjunto de dados com 100 frases contendo esses elementos, os seus resultados foram comparados com o conjunto de resultados esperados após a remoção de ruídos. O *pipeline* apresentou um desempenho notável, alcançando uma taxa de acerto de 97% dos caracteres especiais e 100% das *URLs*.

Conduzimos testes em um *dataset* composto por cerca de 20 mil postagens em português brasileiro obtidas do *Twitter (X)*. O *pipeline* foi aplicado sequencialmente a cada postagem, seguindo a ordem previamente descrita. Durante o processo, ele conseguiu lidar de forma eficaz com os desafios comuns encontrados em dados textuais em português de redes sociais. Correções ortográficas foram aplicadas, emojis foram substituídos por etiquetas específicas, menções de usuários foram anonimizadas, URLs foram removidas e a pontuação e caracteres especiais foram tratados.

4. Conclusão

Neste trabalho, foi apresentado um *pipeline* para o pré-processamento de dados textuais de redes sociais em português composto de 9 fases sequenciais, sendo as duas últimas opcionais, oferecendo flexibilidade e personalização ao processo.

Os testes fizeram uso da plataforma *Google Colab* na sua versão gratuita, sem o uso de recursos de processamento paralelo, em *datasets* com postagens do *Twitter* com quantidades de 8 mil, 20 mil e 60 mil *tweets*, com resultados satisfatórios em relação à qualidade dos dados observados, consistência e tempo de execução, apontando o potencial do *pipeline* para tarefas de análise de texto em português brasileiro. Como trabalho futuro, é possível pensar em considerar as particularidades dos textos escritos por usuários de redes sociais [Di Felippo et. al. 2021][Sanguinetti et. al. 2020]

Na construção do *pipeline*, a ferramenta Enelvo apresentou algumas falhas na normalização do texto, mostrando que uma investigação de outra tecnologia ou alteração do seu código interno pode ser adequada. A realização de mais testes envolvendo *datasets* de redes sociais com diferentes quantidades de *posts* e qualidade de conteúdo, fazendo uso de métricas (como, percentual de acertos/falhas por fase), fornecerão subsídios para avaliar com maior critério a qualidade. Testes com processamento paralelo permitirão avaliar se há melhora no seu tempo de execução.

[2] https://colab.research.google.com/

Agradecimentos

Os autores agradecem à FAPESP pelo financiamento desta pesquisa (#2021/14772-1). Renata Araujo é bolsista de produtividade em desenvolvimento tecnológico e extensão inovadora do CNPq (#305645/2022-6). Vitor dos Santos é bolsista TT1 pela FAPESP (#2023/04752-9). Livia Alabarse dos Santos é bolsista TT1 pela FAPESP (2023/04042-1).

Referências

Cardozo, L. S. e Freitas, L. A. (2021) "Análise de Sentimentos: Avaliando o Desempenho de Pré-Processamento e de Algoritmos de Aprendizagem de Máquina sobre o Dataset TweetSentBR", Em: *Brazilian Workshop on Social Network Analysis and Mining*. Evento Online. Sociedade Brasileira de Computação. p. 169-174.

Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., Silva, E. H., Roman,N. T., Pardo, T. A. S. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In: Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC.

França, T. C. e Oliveira, J. (2014) "Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013". Em: *Brazilian Workshop on Social Network Analysis and Mining*. Brasília. Sociedade Brasileira de Computação. p. 128-139.

Garcia, L. Q., Chinellato, M. H., Caseli, H. M., Oliveira, L. H. M. (2023) "Pipeline para identificação de erros lexicais e geração de sugestões de correção". Em: *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. Belo Horizonte/MG. Sociedade Brasileira de Computação. p. 357-361.

Kansaon, D. P., Brandão, M. A. e Pinto, S. A. P. (2018). "Análise de Sentimentos em Tweets em Português Brasileiro" Em: *Brazilian Workshop on Social Network Analysis and Mining*. Natal. Sociedade Brasileira de Computação, 2018.

Kurniawan, S., Gata, W., Puspitawati, D.A., Parthama, I.K.S, Setiawan, H. e Hartini, S. (2020) "Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis" Em: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing. p. 012057.

Medeiros, M. C. e Borges, V. R. P.(2019) "Tweet Sentiment Analysis Regarding the Brazilian Stock Market" En: *Brazilian Workshop on Social Network Analysis and Mining.* Belém. Sociedade Brasileira de Computação. p. 71-82.

Nascimento, R. S., Santos, G., Carvalho, F e Guedes, G. (2021) "Avaliando contribuições na substituição de termos informais em classificação de texto de redes sociais com NetSpeak-BR". Em: *Brazilian Workshop on Social Network Analysis and Mining*. Evento Online. Sociedade Brasileira de Computação. p. 181-186.

Osakwe, Z. T. e Cortés, Y. I. (2021) "Impact of COVID-19: a text mining analysis of Twitter data in Spanish language" Em: *Hispanic Health Care International*, v. 19, n. 4, p. 239-245.

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D. e Zeldes, A. (2020). "Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations". Available in: https://arxiv.org/abs/2011.02063.

Shen, C., Chen, M. e Wang, C. (2019) "Analyzing the trend of O2O commerce by bilingual text mining on social media" Em: *Computers in Human Behavior*, v. 101, p. 474-483.

Souza, B. Á., Almeida, T. G., Menezes, A. A., Figueired, C. M. S., Nakamura, F. G. e Nakamura, E. F. (2017) "Uma Abordagem para Detecção de Tópicos Relevantes em Redes Sociais Online" En: *Brazilian Workshop on Social Network Analysis and Mining.* São Paulo. Sociedade Brasileira de Computação.

Yang, Sidi, Zhang, Haiyi. (2018) "Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis" Em: *International Journal of Computer and Information Engineering*, v. 12, n. 7, p. 525-529.