ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

ANÁLISE COMPARATIVA ENTRE TÉCNICAS DE NORMALIZAÇÃO DE ERROS TEXTUAIS PROPOSITAIS NO TWITTER

Thainá MARINI[1]; Taffarel BRANT-RIBEIRO[2]

RESUMO

Durante a pandemia da COVID-19, a expansão tecnológica resultou no aumento do uso de redes sociais e comunicação a distância, onde foram evidenciados diversos benefícios e malefícios de seu uso contínuo. Neste contexto, foi observada a tendência da utilização de expressões escritas propositalmente de maneira errada como forma de comunicação. Foram-se categorizados os erros intencionais mais comuns encontrados na rede social Twitter, tais como: a troca de números por letras com a mesma fonética e a substituição do acento agudo pela letra "h". O objetivo deste trabalho foi analisar a eficácia da correção dessas expressões utilizando as técnicas encontradas na literatura: N-Grama e Medida de Distância de Levenshtein. Após a extração de *tweets* e implementação das técnicas, foram realizados testes alterando os parâmetros para avaliar a eficácia. Embora os testes demonstraram acurácia de 100% para ambas as técnicas na categoria 2, foi possível concluir que a Medida de Distância de Levenshtein foi a mais adequada para corrigir erros intencionais nas diversas categorias estudadas, obtendo acurácias de 100% em diferentes parâmetros.

Palavras-chave: Processamento de Linguagem Natural; Redes Sociais; Correção Textual Automatizada, N-Grama, Medida de Distância de Levenshtein.

1. INTRODUÇÃO

Com o advento da tecnologia e o isolamento social causado devido ao período pandêmico, as redes sociais passaram a impactar ainda mais a vida cotidiana (AFFUM, 2022). Devido a isso, a ampla participação de usuários nas redes sociais tem permitido a observação de um novo fenômeno comportamental da geração atual que consiste na utilização da linguagem escrita de um modo distinto do convencional no meio *offline*.

Segundo Gallardo e Kobayashi (2021), o desenvolvimento desta nova forma de escrita prejudicou a importância da norma padrão da língua portuguesa devido à variação linguística. Analisando esse novo fenômeno de escrita distinta é possível observar que, frequentemente, o erro é cometido propositalmente (LAW, 2022).

O Twitter é uma rede social com alta aglomeração de indivíduos online, possuindo cerca de 19 milhões de usuários cadastrados (KEMP, 2022). Devido ao ambiente informal de comunicação, é comum encontrar grande quantidade de conteúdo digital com erros ortográficos. Neste contexto, é interessante observar e documentar os erros intencionais cometidos pelos usuários, a fim de verificar a viabilidade de corrigi-los.

Dentre esses erros cometidos, é possível destacar aqueles que consistem em substituições de letras por números, trocas de letras com fonéticas semelhantes e acréscimo da letra "h" ao final da palavra para expressar entonação, como pode ser observado na Tabela 1.

[1]Bolsista PIBIC, IFSULDEMINAS – *Campus* Passos. E-mail: thainamnobrega@hotmail.com
[2]Orientador, IFSULDEMINAS – *Campus* Passos. E-mail: brant.ribeiro@ifsuldeminas.edu.br

Tabela 1: Categorias de erros que foram utilizadas nesta pesquisa.

Categoria	Descrição	Exemplo
1	Substituição de vogais por números visualmente parecidos.	“P0l1t1c4” - Política
2	Substituição de letras por caracteres especiais visualmente parecidos.	“√er$átil” - Versátil
3	Substituição de sílabas por números com fonéticas semelhantes.	“9dades” - Novidades
4	Substituição do acento til pelo sufixo “aum”.	“Coraçaum” - Coração
5	Substituição de letras que possuem fonéticas semelhantes.	“Xurrasco” - Churrasco
6	Acréscimo da letra “h” para expressar entonação.	“Obrigadah” - Obrigada

Fonte: Autoria Própria.

Um dos maiores desafios para a interpretação destes dados ortograficamente incorretos é a influência que um pequeno erro de escrita pode ter para o funcionamento de uma ferramenta sofisticada de processamento de linguagem natural (HU et al., 2021). Assim, se fez necessária a análise de uma metodologia que identifique e corrija de forma eficiente os erros aqui elencados.

Entre as técnicas encontradas na literatura é possível destacar o N-Grama, que consiste em uma ordem de N palavras ou letras, tal como um bigrama, por exemplo, que é formado pela sequência de duas palavras ou letras. Essa técnica é utilizada para a comparação de candidatos que possuem a maior quantidade de n-gramas em comum para a correção da palavra incorreta (JURAFSKY; MARTIN, 2023).

Também pode-se citar a técnica da Medida de Distância de Levenshtein, que é a métrica mais conhecida para medir a distância entre duas palavras: A e B. Essa medida é definida como o número mínimo de operações necessárias para transformar uma palavra em outra, considerando adições, remoções ou trocas de letras (PATRIARCA; HEINSALU; LÉONARD, 2020).

O objetivo deste trabalho consistiu em comparar e avaliar as técnicas de Medida de Distância de Levenshtein e N-Grama. O propósito foi determinar qual delas demonstrou acurácia superior na correção dos erros intencionais descritos na Tabela 1, encontrados na rede social Twitter.

2. MATERIAL E MÉTODOS

Inicialmente, foram coletados *tweets* publicados entre os meses de janeiro e abril de 2023 e classificados como português, utilizando a biblioteca em Python snscrape. Em seguida, os dados foram pré-processados e analisados individualmente. Nessa análise foram criadas funções para cada categoria de erro, por exemplo: para a categoria 1, foi verificado se a palavra possuía o padrão de conter números e letras e, para a categoria 6, foi verificado se a palavra apresentava o padrão de vogal seguido da letra “h”.

Neste passo, com todos os arquivos de texto contendo possíveis palavras válidas para a pesquisa, foi realizada uma análise manual com o objetivo de assegurar que cada termo seria alocado em sua categoria correspondente e, em seguida, foi inserida a correção referente a cada termo. Com a base léxica desenvolvida, contendo cerca de 900 termos incorretos, empregou-se a linguagem de programação Python 3 e as bibliotecas Levenshtein e NLTK para realizar a implementação das técnicas de medida de distância de Levenshtein e N-Grama, respectivamente.

Com as técnicas implementadas, foram feitos testes com cada categoria, comparando cada erro com todas as palavras corretas. Para otimizar as técnicas utilizadas, foram realizados testes empíricos ajustando os parâmetros e analisando seus comportamentos. Para a técnica N-Grama, variou-se a quantidade de sequências a serem separadas e testou-se a inclusão de um símbolo chamado "*pad symbol*". Este símbolo teve como objetivo melhorar a comparação de palavras que possuem a mesma letra inicial e final, ao separá-las em uma sequência distinta do restante do termo. Ao utilizá-lo, foi observada uma melhora nos resultados, sendo assim mantida essa decisão em todos os testes. Quanto à Medida de Distância de Levenshtein, durante cada teste ajustou-se individualmente o valor de apenas uma das 3 operações. Dessa forma, foi obtido, devido aos resultados consistentes, o conjunto médio de parâmetros com os valores {1,1,1} (referentes às operações de inserção, exclusão e substituição, respectivamente).

Após a finalização da testagem, foi calculada a métrica de acurácia para cada um dos testes. Tal métrica foi computada como a razão entre as sugestões corretas apresentadas por cada técnica pelo número total de termos de cada categoria de erros.

3. RESULTADOS E DISCUSSÕES

Após a análise das diferentes variações aplicadas às técnicas N-Grama e Medida de Distância de Levenshtein, foram obtidos os valores de acurácia apresentados na Tabela 2. A tabela apresenta os resultados para cada categoria (1 a 6) e para cada combinação de parâmetros utilizados.

A acurácia mais significativa foi alcançada na categoria 2, obtendo-se 100% em ambas as técnicas. Isso indica a eficácia dessas abordagens nessa categoria, sendo capazes de encontrar a correspondência correta para todos os termos, mesmo com diferentes combinações de parâmetros.

Além disso, foi notado que o uso de N-Gramas com valor N menor que 2 resultou numa diminuição significativa na acurácia. Portanto, a utilização de sequências unitárias não é viável para o contexto dos erros intencionais. Da mesma forma, aumentar o valor da operação de substituição não se mostrou efetivo, sendo recomendável manter todos os valores de operação equivalentes.

Por fim, é possível observar que a técnica Medida de Distância de Levenshtein apresentou desempenho mais consistente em comparação ao método N-Grama. Isso aconteceu devido a apenas na categoria 3 o N-Grama ter alcançado acurácias melhores que a Medida de Distância de Levenshtein e, em todas as outras categorias, o método de Levenshtein foi superior ao N-Grama.

Tabela 2: Valores de acurácia obtidos após realização dos testes.

Acurácia								
	Levenshtein				**N-Grama**			
	{1,1,1}	**{2,1,1}**	**{1,2,1}**	**{1,1,2}**	**1**	**2**	**3**	**4**
Cat. 1	**1,0**	**1,0**	**1,0**	0,949	0,268	0,978	0,978	0,986
Cat. 2	**1,0**	**1,0**	0,947	0,947	0,421	**1,0**	**1,0**	**1,0**
Cat. 3	0,732	0,616	0,755	0,651	0,348	0,953	0,976	0,976
Cat. 4	0,777	0,777	0,925	0,481	0,111	0,814	0,814	0,814
Cat. 5	0,954	0,857	0,926	0,920	0,164	0,914	0,937	0,937
Cat. 6	0,958	0,958	0,872	0,931	0,191	0,926	0,926	0,913

Fonte: Autoria Própria.

4. CONSIDERAÇÕES FINAIS

Com base nos resultados obtidos neste trabalho, é possível concluir que a técnica mais adequada para corrigir erros intencionais nas categorias estudadas é a Medida de Distância de Levenshtein, demonstrando uma maior acurácia em comparação à técnica N-Grama. Notavelmente, os resultados também foram consistentes com a técnica N-Grama, viabilizando a utilização dessa abordagem na tarefa de corrigir erros intencionais. Por fim, com o intuito de se obter um desempenho ainda mais aprimorado, podem ser realizados novos testes com outras abordagens.

AGRADECIMENTOS

Os autores deste trabalho agradecem ao IFSULDEMINAS pelo auxílio financeiro concedido por meio do Edital 31/2022 do Programa PIBIC / PIBIC Jr.

REFERÊNCIAS BIBLIOGRÁFICAS

AFFUM, M. Q. The effect of the internet on student's studies: a review. **Library Philosophy and Practice (e-journal)**, 2022.

GALLARDO, B. C.; KOBAYASHI, E. Internetês versus escrita formal: a nova escrita e seus desdobramentos. **Web Revista SOCIODIALETO**, v. 11, n. 33, p. 1-18, 2021.

HU, Y. et al. Misspelling Correction with Pre-trained Contextual Language Model. In: INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS & COGNITIVE COMPUTING, Banff, Canada, p. 144-149, 2021.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**. 2023. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/> . Acesso em jul. 2023.

KEMP, S. Digital 2022: Brazil. DataReportal. 2022. Disponível em: <https://datareportal.com/reports/digital-2022-brazil?rq=brazil> . Acesso em set. 2022.

LAW, J. Reflections of the French nasal vowel shift in orthography on Twitter. **Journal of French Language Studies**, v. 32, n. 2, p. 197-215, 2022.