Comparative Analysis of Intentional Grammatical Error Correction Techniques on Twitter/X

Thainá Marini **Taffarel Brant-Ribeiro**

Federal Institute of Education, Science and Technology of the South of Minas Gerais
IFSULDEMINAS, Campus Passos – Passos, Minas Gerais, Brazil

thainamnobrega@hotmail.com brant.ribeiro@ifsuldeminas.edu.br

Abstract

During the COVID-19 pandemic, rapid proliferation of technologies led to an increased dependence on social media and remote communication. This shift highlighted a noteworthy trend: the deliberate use of inaccurately written expressions as a unique mode of communication. These expressions often take form of intentional misspellings, such as substituting letters with similar phonetic sounding numbers or replacing acute accents with letter "h". The main goal of this study was to evaluate the effectiveness of correcting these intentionally incorrect expressions using techniques documented in existing literature, specifically the N-Gram, Levenshtein Distance Measure, and Soundex phonetic algorithm. After assembling a dataset of posts and applying these correction techniques, series of tests were conducted, incorporating various parameter configurations to determine their effectiveness. Results revealed a 100% accuracy rate for Levenshtein Distance and N-Gram techniques for one of the error categories we analysed. Also, excluding the initial letter from the Soundex code improved its accuracy, although it ranged from 22% to 96%. Nevertheless, the Levenshtein Distance Measure approach emerged as the most significant option for correcting intentional errors in various examined categories, achieving 100% accuracy rate across a range of parameter permutations.

1 Introduction

With advent of technology and social isolation caused by the pandemic period, social networks have gained an even greater influence on everyday life (Affum, 2022). Consequently, widespread engagement of users on social media has allowed the observation of a new behavioral phenomenon in the current generation. This phenomenon involves the use of written language in a distinct manner from conventional offline mediums.

According to Gallardo and Kobayashi (2021), the development of this new form of writing has diminished the importance of standard norms of Portuguese language due to linguistic variation. While analyzing this novel phenomenon of distinct writing, it is often possible to observe that errors are committed intentionally (Law, 2022).

Twitter/X is a social network with a substantial congregation of online individuals, having approximately 19 million users (Kemp, 2022). Due to its informal communication environment, a significant amount of digital content with spelling errors can commonly be encountered. In this context, we considered intriguing to observe and document intentional errors committed by users in order to assess feasibility of correcting them automatically.

Thus, our motivation aims to impact Natural Language Processing (NLP) tools by finding effective techniques to correct these intentional errors. Among these intentional errors, notable instances include substitution of letters with numbers, exchange of letters with phonetically similar counterparts, and addition of letter "h" at end of words to convey intonation, as shown in Table 1.

One of greatest challenges in interpreting these orthographically incorrect data is the impact that a minor writing error can have on the functioning of a sophisticated NLP tool (Hu et al., 2020). By developing techniques for automatically correcting these errors, it is possible to enhance quality and reliability of analyses of large volumes of textual data. Thus, in this work, we aimed to analyze techniques that could identify and correct these intentional grammatical errors efficiently.

2 Related Work

The complexity of analyzing user-provided data extends beyond the Portuguese language, as illustrated in Demir and Topcu (2022). In their work, a graph-based tool for text normalization in turk-

Category	Description	Example
1	Replacement of vowels with visually similar numbers.	“P0l1t1c4” - Política
2	Replacement of letters with visually similar symbols.	“V€rs@til” - Versátil
3	Replacement of syllables with phonetically similar numbers.	“9dades” - Novidades
4	Replacement of tilde accent with the suffix “aum”.	“Coraçaum” - Coração
5	Replacement of letters with similar phonetics.	“Xurrasco” - Churrasco
6	Addition of the letter “h” to express intonation.	“Obrigadah” - Obrigada

Table 1: Error categories utilized in this research.

ish language was developed, effectively mitigating noise interference in user-generated texts.

Application of the Levenshtein Distance Measure for spelling correction and text standardization has also been a prevalent approach. Ortega et al. (2022) formulated a comprehensive approach to address the challenges of enhancing quality of galician text data for Natural Language Processing applications. The authors integrated the Levenshtein Distance into a set of heuristics to improve coherence and correctness of galician corpus.

Also, utilization of N-Grams, a common approach for textual analysis, had a pivotal role in Alcoforado et al. (2022). They proposed a novel hybrid model that combined the Transformer architecture with unsupervised learning, referred as ZeroBERTo. Their model achieved proficiency at classifying texts without requirements for labeled training data. This approach employed a statistical model that leverages the N-gram technique for topic modeling in unlabeled documents.

3 Background

In this section, we cover the theoretical concepts of our work. Specifically, Levenshtein Distance is defined in subsection 3.1, N-Gram is explained in subsection 3.2, and Soundex Phonetic Algorithm is presented in subsection 3.3.

3.1 Levenshtein Distance

The Levenshtein Distance is the best known metric for measuring distance/difference between two words. This measure is defined as the minimum number of operations required to transform one word into another, considering additions, deletions, or substitutions of letters (Patriarca et al., 2020).

E.g., to calculate the minimum distance between three words, namely: "mais" (1), "mas" (2), and "más" (3), the following analyses are performed:

"Mais" – "Mas": Deletion of letter "i" (1 edit).

"Mais" – "Más": Deletion of letter "i" and substitution of "a" with "á" (2 edits).

"Mas" – "Más": Substitution of "a" with "á" (1 edit).

The Levenshtein distance can be organized into a matrix $L = L_{ij}$. Therefore, the aforementioned example is represented in a 3x3 matrix, with the main diagonal set to zero since no words are identical, as shown in the following matrix:

$$L_{ij} = \begin{bmatrix} L11 & L12 & L13 \\ L21 & L22 & L23 \\ L31 & L32 & L33 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

3.2 N-Gram

The N-Gram approach involves an order of N words or letters, e.g. a bi-gram, which is formed by a sequence of two words or letters. This technique is employed to compare candidates that share the highest number of common n-grams to rectify incorrect words (Jurafsky and Martin, 2023). As exemplified in Figure 1, which depicts the word “artigo” with N values of 1, 2, and 3.

N = 1	A - R - T - I - G - O
N = 2	AR - RT - TI - IG - GO
N = 3	ART - RTI - TIG - IGO

Figure 1: Example of word “artigo” (article) using different n-gram values.

3.3 Soundex

Soundex is a phonetic algorithm that encodes homophones with the same indexing code, searching for words that have a similar phonetic representation (Araujo et al., 2021). Each Soundex code consists of four digits: the first digit is the word’s first letter, and the next three digits are numbers obtained from the remaining letters, according to

Value	Letter(s)
0	A, E, I, O, U, H, W, Y
1	P, B, M
2	F, V
3	T, D, N
4	L, R
5	S, Z
6	J, DI, GI, TI, CH, LH, NH
7	K, C, G, Q
8	X

Table 2: Letter encoding of Soundex algorithm adapted for Brazilian Portuguese (Ruberto and Antoniazzi, 2017).

Table 2. This table utilizes encoding values adapted for Brazilian Portuguese, as presented in Ruberto and Antoniazzi (2017). For example, the word "artigo" (article, in Portuguese) is encoded in Soundex as "A437" based on the rules provided in Table 2 and shown Figure 2.

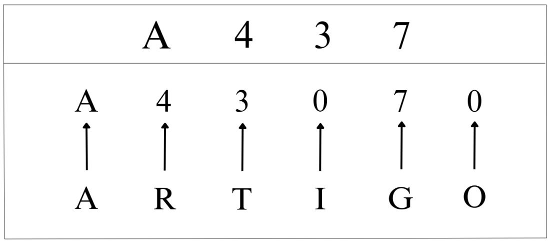

Figure 2: Example of word "artigo" in Soundex code.

4 Method

Initially, posts published between January and April of 2023 were collected and classified as Portuguese using the Python library snscrape. Subsequently, data underwent preprocessing and individual analysis. During this phase, functions were created for each error category. E.g., for Category 1, we checked which character strings had a pattern of containing both numbers and letters, and, for Category 6, we examined which words had a pattern of a vowel followed by letter "h".

Every character string containing a potential valid word for our study was also checked manually to ensure that each term was assigned to its corresponding category. Subsequently, we inserted the appropriate correction of each term. After finishing these steps, our lexical base had approximately 900 terms. In each category, the corresponding numbers were as follows: (1) 380, (2) 20, (3) 90, (4) 30, (5) 180, and (6) 220. This distribution highlights the higher frequency of usage by users in categories 1 and 5. Next, we employed Python 3 programming language and the "Levenshtein" and "NLTK" libraries to implement the Levenshtein Distance and N-Gram measurement techniques, respectively.

Regarding the Soundex Phonetic Algorithm, due to absence of pre-existing implementations for Brazilian Portuguese in Python, a manual implementation was developed. This implementation incorporated encoding values adapted for Brazilian Portuguese, as detailed in Table 2 and presented in Ruberto and Antoniazzi (2017).

Similarly, in response to the observed trend among Twitter/X users of substituting syllables with numbers that sound alike (Error Category 3), we have created an additional encoding for Soundex (Table 3). This table groups values from Table 2 and was developed to speed up the representation of words with similar pronunciations, aligning with the current communication standards in the context of Twitter/X.

Thereafter, tests were conducted for each category, comparing each error with all correct words. To optimize the techniques, empirical tests were performed by adjusting parameters and analyzing their behaviors. For the N-Gram, the number of separated sequences varied, and the inclusion of a symbol called "pad symbol" was tested. This symbol aimed to enhance comparison of words that had the same initial and final letters by separating them into a distinct sequence from the rest of the term.

Regarding the Levenshtein Distance Measure, during each test the values of only one of three operations were adjusted individually. Thus, due to consistent results, an average parameter set with values 1,1,1 (referring to insertion, deletion, and

Value	Pronounce
1	"um"/"hum"
3+5	"dois"/"dos"
3+4+5	"três"/"tris"
5+5 || 7+5	"seis"/"ceis"
5+3	"sete"/"set"
3	"oito"/"oi to"
3+2	"nove"/"novi"
3+5	"dez"/"des"
8+3+5	"quinze"
2+1+3	"vinte"/"vim te"

Table 3: Encoding of number pronunciation adapted for Brazilian Portuguese.

	Levenshtein				N-Gram				Soundex	
	{1,1,1}	{2,1,1}	{1,2,1}	{1,1,2}	1	2	3	4	W/ 1°L.	W/O 1 L.°
Cat. 1	1.0	1.0	1.0	0.94	0.26	0.97	0.97	0.98	0.88	0.96
Cat. 2	1.0	1.0	0.94	0.94	0.42	1.0	1.0	1.0	0.94	0.94
Cat. 3	0.73	0.61	0.75	0.65	0.34	0.95	0.97	0.97	0.02	0.81
Cat. 4	0.77	0.77	0.92	0.48	0.11	0.81	0.81	0.81	0.33	0.22
Cat. 5	0.95	0.85	0.92	0.92	0.16	0.91	0.93	0.93	0.32	0.39
Cat. 6	0.95	0.95	0.87	0.93	0.19	0.92	0.92	0.91	0.94	0.93

Table 4: Accuracy values obtained after our tests.

substitution operations, respectively) was obtained.

During the testing of Soundex, two modifications were also made: firstly, the length of resulting encoded word was adjusted, and it was observed that increasing the code length did not yield significant improvements in accuracy. Therefore, we decided to maintain a code length of 4 characters in all tests. Secondly, we observed that omitting the first letter of each word significantly improved accuracy. Consequently, this decision was maintained throughout all tests.

In order to observe the technique's success rates, we calculated the obtained accuracy in each test. This metric was computed as a ratio between correct suggestions provided by each technique and the total number of terms in each error category.

5 Results and Discussion

After testing the N-Gram, Levenshtein Distance Measure, and Soundex techniques, we obtained the accuracy values presented in Table 4. The Table illustrates results for each error category (1 to 6) and for each combination of parameters used. The highest accuracy was achieved in Category 2, with 100% accuracy rate for both N-Gram and Levenshtein techniques. This highlights effectiveness of these approaches in this category, as they were capable of identifying correct matches for all terms, even with different parameter combinations.

Additionally, we observed that using N-Grams with an N value lower than 2 resulted in a significant decrease of its accuracy. Therefore, the use of unitary sequences does not appear to be viable in the context of intentional errors. Similarly, increasing the value of operations did not prove to be more effective, and it is recommended to keep all operation values equivalent.

Nonetheless, when considering the Soundex technique, its performance can be summarized as follows: although it reached moderate accuracies in several categories, it did not consistently outperform the N-Gram and Levenshtein methods. Specifically, the Soundex performance varied, with an accuracy of just 22% in Category 4, in contrast to a high accuracy of 96% in Category 1. Additionally, it is worth pointing out that the newly proposed encoding depicted in Table 3 yielded some promising results, achieving an accuracy of 81%.

Lastly, we observed that Levenshtein Distance Measure exhibited more consistent results compared to the N-Gram and Soundex methods. This disparity arose because only in Category 3 the N-Gram achieved better accuracies than the Levenshtein Distance Measure, whereas in every other category the Levenshtein method outperformed N-Gram. Therefore, in this work the Levenshtein Distance Measure was considered the most consistent technique amongst all.

6 Conclusion

Based on experimental results we obtained in this study, we conclude that the most suitable technique for correcting intentional errors in the six error categories we analysed is the Levenshtein Distance Measure, which achieved a higher accuracy compared to the N-Gram and Soundex techniques.

Notably, results were also consistent with the N-Gram technique, enabling the use of this approach in tasks of correcting intentional errors. While the Soundex technique showed promise, it still requires further refinement to consistently compete with the other approaches, as discussed in section 5.

Furthermore, it is noteworthy that omitting the first letter of the Soundex code proved to enhance its accuracy, and further exploration of this approach could lead to improved results in future studies. Finally, to achieve an even enhanced performance in the task of correcting intentional grammatical errors, new tests can be conducted with alternative approaches and techniques.

References

Mark Affum. 2022. The effect of internet on students studies: A review. *Library Philosophy and Practice (e-journal)*.

Alexandre Alcoforado, Thomas P. Ferraz, Rodrigo Gerber, Enzo Bustos, André S. Oliveira, Bruno Miguel Veloso, Fábio L. Siqueira, and Anna Helena R. Costa. 2022. Zeroberto: Leveraging zero-shot text classification by topic modeling. In *Computational Processing of the Portuguese Language*, pages 125–136, Cham. Springer International Publishing.

Leonardo Araujo, Aline Benevides, and João Sansão. 2021. Desenvolvimento de um corretor ortográfico. *Texto Livre: Linguagem e Tecnologia*, 14(1):1–19.

Seniz Demir and Berkay Topcu. 2022. Graph-based turkish text normalization and its impact on noisy text processing. *Engineering Science and Technology, an International Journal*, 35:101192.

Barbara Gallardo and Eliana Kobayashi. 2021. Internetês versus escrita formal: A nova escrita e seus desdobramentos. *Web Revista SOCIODIALETO*, 11(33):1–18.

Yifei Hu, Xiaonan Jing, Youlim Ko, and Julia T. Rayz. 2020. Misspelling correction with pre-trained contextual language model. In *2020 IEEE 19th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 144–149.

Dan Jurafsky and James H. Martin. 2023. Speech and language processing (3rd ed. draft). Online; accessed in October 2023.

Simon Kemp. 2022. Digital 2022: Brazil. Online; accessed in October 2023.

James Law. 2022. Reflections of the french nasal vowel shift in orthography on twitter. *Journal of French Language Studies*, 32(2):197–215.

John E. Ortega, Iria de Dios-Flores, José R. Pichel, and Pablo Gamallo. 2022. Revisiting ccnet for quality measurements in galician. In *Computational Processing of the Portuguese Language*, pages 407–412, Cham. Springer International Publishing.

Marco Patriarca, Els Heinsalu, and Jean L. Leonard. 2020. *Languages in Space and Time: Models and Methods from Complex Systems Theory*. Physics of Society: Econophysics and Sociophysics. Cambridge University Press.

Diogo L. V. G. Ruberto and Rodrigo L. Antoniazzi. 2017. Análise e comparação de algoritmos de similaridade e distância entre strings adaptados ao português brasileiro. In *Anais da XIII Escola Regional de Banco de Dados*, page 27–36, Porto Alegre, RS, Brasil. SBC.