Timor-Leste Journal of Engineering and Science

Vol.4, Issue 2, pp.1-10, 2024
Available online at: http://tljes.org/index.php/tljes/data

Classification of Tetun Language Documents Based on INL and DIT Orthography with a Text Mining Approach

Edio da Costa, Almeida Barreto
Department of Computer Science, School of Engineering and Sceince, Dili Institute of Technology, Dili, Timor-Leste
Email: ediocosta73@gmail.com,

ABSTRACT

The main problem in language classification is the complexity and intricacy of accurately tracing these relationships, such as language evolution, contact and borrowing words, which makes it difficult to classify the orthography used. In both government and non-government institutions, many individuals write documents using varying spellings. At the Dili Institute of Technology (DIT), a unique spelling system has been developed alongside adherence to the guidelines of the National Institute of Linguistics (INL). The DIT orthography, based on contemporary Tetun, does not employ accents, as numerous studies have indicated that accents are unnecessary. This research aims to develop an application that classifies documents using a text mining approach, with tokenization and filtering based on word lists from INL and DIT orthographies. This process aims to categorize submitted user documents accurately. The documents used in this research consist of INL and DIT orthographic. The word list dictionary from INL comprises 1,487 words from the Tetun-Portuguese dictionary, while the DIT word list includes 756 words collected from the DIT Language Center and additional sources. The research findings indicate that the system can classify documents based on predefined orthographic categories.

Keywords: *Orthography classification, tetun language, INL and DIT, text mining and orthography*

Received January 16, 2024; Revised March 20, 2024; Accepted May 10, 2024

1. Introduction

Tetun is the second official language in Timor-Leste, originating from Malayo-Polynesian and heavily influenced by the Portuguese language (Klinken, 2016). According to Decree-Law No. 1/2004, the official orthography in use is that of the National Institute of Linguistics (INL). The INL orthography is significantly influenced by Portuguese, resulting in borrowing some Portuguese words (Lusismu). Thus, INL Tetun prefers to write with accents to indicate the tonic syllable of words (Ofisiál & Sousa, 2014). However, the Dili Institute of Technology (DIT) has also developed its orthography based on INL guidelines. The rule for writing at DIT follows contemporary Tetun phonology and does not use accents because many studies have shown that they are unnecessary. The observations indicate that documents are written using different orthographies in government institutions, academics, electronic media, and many non-governmental organizations (Silva, 2021). Newspapers mainly use Tetun, and more than ten national online news websites actively broadcast news in Tetun every day (Jesus, 2023)

DIT Tetun serves as a guide that primarily teaches simple Tetun, known as simple orthography, which can also be found in the word finder. In DIT Tetun, when writing according to the pronunciation we hear when people speak, accents are not used to mark the tonic syllable, for example: *"nebee"*, *"nee"*, and *"koalia"*. The main differences are that INL uses acute accents: *"ha'u"*, *"di'ak"*, *"maña"*, *"falla"*, *"animál"*, *"laran-moras"*, "*ha'u-nia"*. However, in DIT Tetun, the differences are *"hau"*, *"diak"*, *"deit"*, *"manha"*, *"falha"*, without using accents (Klinken, 2017).

The complexity of INL and DIT orthographies in Tetun involves handling loanwords, diacritics, spelling conventions, and phonetic representation. INL orthography typically adapts foreign loanwords to fit Tetun's phonetic and orthographic norms, using diacritics to indicate specific sounds and stress patterns, thus enhancing phonetic accuracy and making the language more intuitive for native speakers. In contrast, DIT orthography retains the original spelling of loanwords and uses fewer diacritics, simplifying writing but potentially reducing phonetic clarity. This divergence affects the written form and the ease with which speakers adapt to new words and pronunciation rules (Sousa, 2014; Ground el al., 2020; Mart,2020). These differences impact the standardization and usage of Tetun across various contexts, including education, media, and official communications. INL orthography, promoted by official institutions, is widely accepted and used in formal settings, leading to a more uniform linguistic landscape. Conversely, DIT orthography is more prevalent in academic, technical, or regional contexts, reflecting a broader linguistic diversity and introducing variation in written forms. This duality necessitates careful consideration in text processing and classification tasks to ensure comprehensive language representation (Ofisiál & Sousa, 2014; Greg, 2022).

Recent studies in text mining approach for orthography classification have demonstrated significant advancements because of the boom of textual applications such as news article portals and social networking gates (Muaad et al., 2022). The study proposed various methods for detecting text formality, a crucial aspect of orthography classification (Dementieva et al., 2023). Another relevant research in text mining show cased the high-performance classification of different orthographies using advanced computational techniques. Research conducted by (Huang et al. 2020; Zhang and Liu 2019; Kim et al. 2021; Li and Wang 2020; Silva & Pereira 2022) have all contributed to improving the efficiency and accuracy of orthography classification systems through innovative text mining approaches.

In recent years, the classification of orthography and applying text mining techniques, mainly using Decision Tree (DT) methods, have been extensively investigated. These studies underscore the effectiveness of DT in managing complex orthographic data across multiple languages, highlighting its high performance. For example, Zhang and Liu (2019) utilized DT methods with text mining to classify English orthographic features effectively. Similarly, Li and Wang (2020) applied DT to educational datasets, revealing patterns and deviations in orthographic usage. Dementieva et al. (2023) demonstrated the utility of DT in classifying texts based on formality levels, thereby contributing to more nuanced text classification systems.

Furthermore, Garcia and Martinez (2021) employed DT to classify orthographic variants in indigenous texts, showcasing the method's ability to handle the orthographic diversity inherent in these languages. Choi and Lee (2021) investigated the application of DT methods for orthography classification in technical texts demonstrating the method's capability in handling specialized vocabulary and structure. Additionally, Rahman and Ahmed (2022) examined the classification of orthographic variants and highlighted the effectiveness of DT. These studies illustrate the versatility and robustness of DT methods in orthography classification and text mining.

This research aims to classify Tetun documents according to INL and DIT orthographies using a text-mining approach. This involves performing case-folding, tokenizing, and filtering with Decision Tree (DT) methods to aid writers in maintaining consistency when composing formal documents, thereby avoiding orthographic confusion.

2. Literature Review

2.1. Text Classification

Text classification is the problem of assigning a document. Analyzing text has become an essential part of our lives because of the increasing amount of text data, making text classification a big problem (Muaad et al., 2022). A supervised learning framework is commonly used to train a text classifier (Peng and Huang, 2006). Text classification is the process of assigning one or more predefined categories to text according to its content (Samah et al., 2022), which combines classification concepts with a supervised learning approach, and the results show a significant increase in performance. Several studies have been conducted on text classification using decision trees, demonstrating their effectiveness in various applications. For instance, Zhang (2023) integrated text and table extraction for materials science publications, while Chen et al. (2019) focused on fraud detection. Additionally, Jalal (2022) applied decision trees to general text classification, and Wang et al. (2024) utilized them for classifying defect texts. These studies have consistently shown good performance across different domains.

2.2. Text Mining

Text mining plays an important role in unveiling purified information from many documents in a satisfactory time (Accuosto and Saggion, 2020). Typically, the phases conducted before text analysis using Text Mining methods include pre-processing steps such as tokenizing, filtering, removing stopwords, tagging, and stemming (Asiyah & Fithriasari, 2016). However, in this research, only two stages, tokenizing and filtering, including case-folding, are implemented, as proposed by Costa and Mali (2021).

Numerous studies employing text-mining approaches have demonstrated superior classification performance. For example, Muaad et al. (2022) focused on Arabic document classification, while Lian et al. (2024) utilized text-mining techniques to analyze sentiments and themes in Chinese documents. Sudigyo et al. (2023) applied text mining to extract supplementation intervention research from Indonesian documents, and Piriyakul et al. (2024) evaluated brand equity in English documents. These studies highlight the effectiveness of text mining in diverse linguistic contexts.

2.3. Tetun Orthography

Orthography is a part of grammar that teaches the correct way to describe words. The areas of orthography include: "alphabet, writing symbols, accentuation rules, how to express vowels and consonants, syllabic derivation, translinear syllabification, phonetic and graphic relations between words, the use of capital letters at the beginning of sentences, and punctuation (Crystal, 2003). Orthography is also a branch of linguistics that teaches the proper way to form phrases through writing, indicating that orthography is a sub-discipline of linguistics that also teaches spelling (Thoyyibah, 2019).

Based on documents written in the Tetun language, there is currently a mixture of orthographies due to the adaptation from the National Institute of Linguistics (INL) and Tetun DIT. The Tetun DIT orthography has changed to simplify its use. Previously, the Portuguese way of writing *"nh"* and *"lh"*, as seen in "*senhora*" and "*ilha*", was written with the letters *"ny"* and *"ly"* because *"nh"* also appears in some Tetun words, particularly "*bainhira*", "*bainhaat*", and "*bainhitu*". Additionally, words with *"aan"* are now written separately rather than joined; for example, "*foti aan*" is now written instead of "*foti-an*" (Klinken et al., 2017). When using the National Institute of Linguistics (INL) orthography, there are differences from Tetun DIT because some words are different.

DIT orthography adapts Portuguese letters to match Tetun pronunciation better. For example, the Portuguese *"nh"* and *"lh"* are replaced with *"ny"* and *"ly"*. This adaptation makes the written form more phonetic and makes it easier for Tetun speakers to pronounce correctly (Klinken, 2016). Tetun INL orthography retains the original Portuguese letters 'nh' and 'lh'. This system maintains a closer link to the historical and etymological roots of the words, preserving their original Portuguese form (Ofisiál & Sousa, 2014).

INL orthography uses accents to indicate the tonic syllable in words, similar to Portuguese. This helps preserve correct pronunciation and distinguish between words that might otherwise look similar. Tetun INL orthography follows more complex rules influenced by Portuguese orthography (Ofisiál & Sousa, 2014). Meanwhile, DIT orthography does not use accents. Studies have shown that accents are unnecessary for understanding or pronunciation in Tetun, so they are omitted to simplify the writing process (Klinken, 2016). Table 1 shows the different orthography of INL and DIT.

Table 1. Differences Between DIT and INL Orthography

Aspect	Tetun DIT Orthography	Tetun INL Orthography
Alphabet Usage	It uses letters similar to Portuguese but adapted for Tetun pronunciation. *For example, it uses* "*NY*" and *“ly”* instead of Portuguese *“nh”* and *“lh*."	Uses more Portuguese letters. Example: *“nh”* and *“lh”* are retained.
Accentuation	Does not use accents, as studies show they are unnecessary.	It uses accents to indicate the tonic syllable in words.
Word Separation	Words are written separately without connecting dashes. Example: *“foti aan”* instead of *“foti-an”*.	Words may be connected with hyphens or written together based on Portuguese influence.
Consistency	Aims for more straightforward and more consistent spelling based on pronunciation.	Follows more complex rules influenced by Portuguese orthography.
Example Words	- *senhora* (written as *“senyora”*) - *ilha* (written as *“ilya”*)	- *senhora* (written as *“senhora”*) - *ilha* (written as *“ilha”*)

3. Research Methodology

Figure 1 describes all processes of classification based on DIT and INL orthography. Tetun comprises twenty-six alphabet letters (Klinken, Ribeiro, and Tilman, 2016). Regarding its structure, particularly orthography, the use of vocabulary and grammar varies (Klinken, 2019). Many different orthographic writings such as "*ne'ebé*" some write "*ne'ebé*" and some write "*nebe*". So those words adopted from Portuguese must be written using the rules of the Tetun language, for example, the word "*nação*" *(Portugis)* and written in Tetun "*nasaun*". Additionally, when writing technical vocabulary in Tetun, many encounter difficulties due to the lack of technical terms, such as the word "hosting," which is not found in the Tetun dictionary.

The stop word and word list of the Tetun Language in this study were obtained from the Language Centre of Dili Institute of Technology, which used 2.560 Word Lists and 8.972 Stop Words. The word list from the INL consists of 1.487 words collected from the Tetun-Portuguese dictionary.

3.1. Pre-processing

Table 2 shows the sample of dataset orthography Tetun DIT and INL document with corresponding words, which are used for case studies to classify Tetun orthography. The purpose is to obtain and categorise results according to their orthographic standards. The data set referenced in the following table is sourced from Ofisiál & Sousa (2014) and Klinken (2017). The pre-processing process was used to extract the data collected. In the extraction process in this research, the authors adopt the Tokenizing and Filtering proposed by (Da Costa, Tjandrasa and Djanali, 2018).

The pre-processing stage is the subsequent step to minimize non-essential attributes in the classification process. At this stage, the input data is in raw form, and this process aims to produce high-quality documents that are expected to streamline the classification process. The first step involves cleaning, which encompasses several procedures such as removing numbers and performing case folding to reduce variability in the text, ensuring that words like "*Tetun*" and "*tetun*" are treated as the same token and remove the punctuation such as ‘(‘,’)’,’.’,’-‘,'/','?'. This normalization is crucial for the accurate matching of words and helps reduce the dataset's dimensionality. Table 3 shows examples of documents and pre-processing results using the clean number and case folding algorithm proposed by (Bird, S., Klein, E., & Loper, 2009; Manning, Raghavan, & Schütze, 2008; Arief and Deris, 2021). Next, break down the text into individual words using a tokenizing algorithm. The text was successfully segmented into tokens, the basic units for further analysis. This step facilitated the identification and analysis of key terms within the documents. The last process is filtering, removing stop words and other irrelevant information and filtering other common words. This ensured that the remaining tokens were more meaningful for the classification process. For instance, after filtering, the token list might be reduced to "*ami*", "*sira*", "*no*", focusing on the most relevant keywords. (Costa and Mali, 2021) This process is done by comparing the text documents with a word in the stop word dictionary of the two orthographies; the results of pre-processing are shown in Tables 3 and 4.

Vol.4, Issue 2, pp.1-10, 2024
Available online at: http://tljes.org/index.php/tljes/data

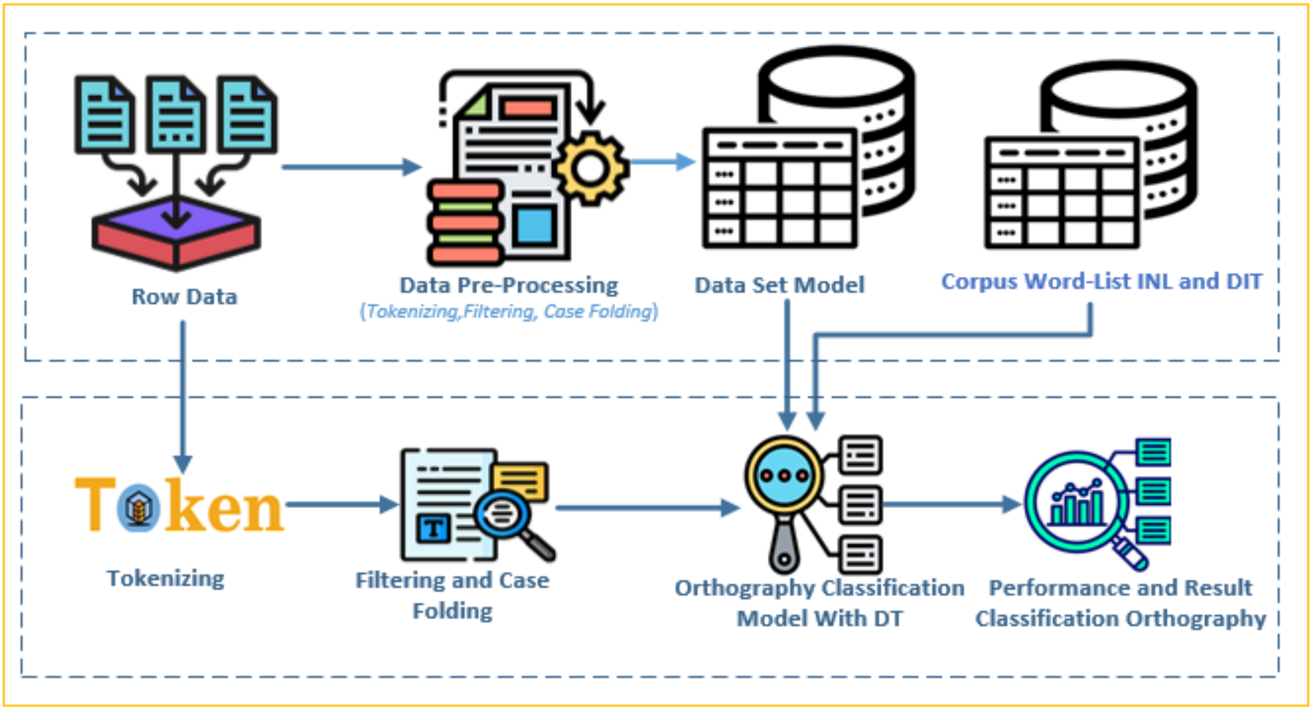

Row Data
Data Pre-Processing
(Tokenizing,Filtering, Case Folding)
Data Set Model
Corpus Word-List INL and DIT
Token
Tokenizing
Filtering and Case Folding
Orthography Classification Model With DT
Performance and Result Classification Orthography

Figure 1. System Proposed

Table 2. Sample of Dataset

Document	Classification
Iha mementu solene ida nee, permite hau nudar reitor atu hatoo liafuan dezenvolvimentu. Nebee DIT hahuu halao tiha ona. Hanesan estudante, importante tebes ba ita atu kompriende instrusaun nebee ema foo, no halo tuir. Se ita la halo tuir instrusaun nebee mestri foo iha ezame laran, ita la liu ezame nee. Se ita nia patraun haruka ita halo buat ruma, maibee ita la kompriende ka la halo tuir, patraun bele hirus. Se ita lee sala instrusaun kona ba oinsaa atu hemu aimoruk ruma karik, aimoruk nee bele halo ita lanu, bele mos halo ita mate. Klasifikasaun hanesan materia ida nebee geralmente iha data mining, klasifikasaun nee rasik sai hanesan teknika nebee mak utilija hodi hetan valor husi data sampel hodi sai grupu ida. Metodu klasifikasaun nee mos peskizador sira utiliza ona mak hanesan Support Vector Machine (SVM), Cosine Similarity hodi halo kategorizasaun ba jornal ka artigu, atu nunee fasil atu hatene kategoria jornal ka artigu sira…………………………………………………… ………………………………………………………… ………………………..	DIT
Vokabuláriu lian tetun iha disionáriu ne’e iha liafuan oioin ne’ebé deskreve lisan tradisionál. Iha Timór Lorosa’e, no balada no ai indíjena sira, maibé iha móstermu modernu barak hosi mundu siénsia, polítika, teknolójia no relijiaun nian. Ne’e mós atu ajuda ema iha Timór Lorosa’e sé seidauk hatene liafuan ortografia no vokabuláriu ofisial. Ne’ebé Institutu Nasionál Linguístika nian promove hela, no ema hosi raiseluk sé serbisu iha Timór Lorosa’e sé presiza tradús loos iha leet lia-portugés no lia-indonéziu. Semántíku nu’udar parte ida hosi gramátika ne’ebé estuda signifikadu liafuan nian ho modifikasaun sentidu nian ne’ebé sira sofre liuhosi tempu no espasu. Área ne’ebé envolve iha semántika mak hanesan. Pontuasaun ne’e importante tebtebes. Dala barak ita kompreende laloos saida mak ita rejista liuhosi prosesu hakerek tanba falta koloka pontuasaun para fasilita leitura ba ema seluk. Atu ita kompreende di’ak liu tan kona-ba pontusaun sei hare iha oin. Atu ita koñese no komprende kle’an liu kona-ba parte tolu ortografia nian ne’e, mai ita halehat hamutuk asuntu idaidak kona-ba ortografia padronizada..	INL

Table 3. Sample of Text Input and Result of Clean Number and Case Folding

	Sample Text
Text Input	Vokabuláriu lian tetun iha disionáriu ne’e iha liafuan oioin ne’ebé deskreve lisan tradisionál. Iha Timór Lorosa’e, no balada no ai indíjena sira, maibé iha móstermu modernu barak hosi mundu siénsia, polítika, teknolójia no relijiaun nian. Ne’e mós atu ajuda ema iha Timór Lorosa’e sé seidauk hatene liafuan ortografia no vokabuláriu ofisial. Ne’ebé Institutu Nasionál Linguístika nian promove hela, no ema hosi raiseluk sé serbisu iha Timór Lorosa’e sé presiza tradús loos iha leet lia-portugés no lia-indonéziu. Semántíku nu’udar parte ida hosi gramátika ne’ebé estuda signifikadu liafuan nian ho modifikasaun sentidu nian ne’ebé sira sofre liuhosi tempu no espasu. Área ne’ebé envolve iha semántika mak hanesan. Pontuasaun ne’e importante tebtebes. Dala barak ita kompreende laloos saida mak ita rejista liuhosi prosesu hakerek tanba falta koloka pontuasaun para fasilita leitura ba ema seluk. Atu ita kompreende di’ak liu tan kona-ba pontusaun sei hare iha oin. Atu ita koñese no komprende kle’an liu kona-ba parte tolu ortografia nian ne’e, mai ita halehat hamutuk asuntu idaidak kona-ba ortografia padronizada.
Result of Clean Number and Case Folding	vokabuláriu lian tetun iha disionáriu ne’e iha liafuan oioin ne’ebé deskreve lisan tradisionál iha timór lorosa’e no balada no ai indíjena sira maibé iha móstermu modernu barak hosi mundu siénsia polítika teknolójia no relijiaun nian ne’e mós atu ajuda ema iha timór lorosa’e sé seidauk hatene liafuan ortografia no vokabuláriu ofisial ne’ebé institutu nasionál linguístika nian promove hela no ema hosi raiseluk sé serbisu iha timór lorosa’e sé presiza tradús loos iha leet lia-portugés no lia indonéziu semántíku nu’udar parte ida hosi gramátika ne’ebé estuda signifikadu liafuan nian ho modifikasaun sentidu nian ne’ebé sira sofre liuhosi tempu no espasu área ne’ebé envolve iha semántika mak hanesan pontuasaun ne’e importante tebtebes dala barak ita kompreende laloos saida mak ita rejista liuhosi prosesu hakerek tanba falta koloka pontuasaun para fasilita leitura ba ema seluk atu ita kompreende di’ak liu tan kona-ba pontusaun sei hare iha oin atu ita koñese no komprende kle’an liu kona-ba parte tolu ortografia nian ne’e mai ita halehat hamutuk asu ntu idaidak kona-ba ortografia padronizada

Cleaning numbers involves removing non-alphabetic characters, such as numbers and symbols, to reduce noise. However, only specific symbols (such as commas) are removed in this process. This distinction is made because symbols vary between the INL and DIT orthographies. The following process uses tokenization to separate each word in a separate line and then accumulate all results into one text file (Salah et al., 2022), then uses filtering to eliminate the unimportant words so the result of both processes shown in Table 4.

Table 4. Result of Tokenizing and Filtering

Tokenizing and Filtering
vokabuláriu
ne’ebé
maibé
modernu
sé
institutu
nu’udar
pontuasaun
di’ak
komprende
kle’an

Next, classification is performed using Decision Trees (DT) by calculating the Gain and Entropy values. The decision tree model learns from the training data, which already has classes such as INL and DIT orthographies. The result of entropy and gain shows INL orthography as a root word. Therefore, the DT classification results indicate that the text belongs to the Orthography INL classification (Figure 2).

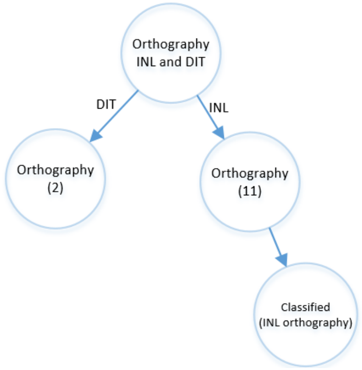


Figure 2. Result of Classification

3.2. Performance Evaluation

Performance evaluation of the plagiarism orthography classification in this study using multiclass classification models often involves using a confusion matrix, which helps in understanding the model's accuracy and error patterns across multiple classes (Saxena, 2023; Kuzu, 2023). The confusion matrix is an N×N table, where N is the number of classes. Each row of the matrix represents the instances of an actual class, while each column represents the instances of a predicted class. True Positive (TP) means that the orthography found to be classified, the detection which is the correct classification. False

Positive (FP) instances are incorrectly predicted as a particular class. False Negatives (FN) instances that belong to a class but were predicted as another class. True Negatives (TN) means the orthographies of a language Tetun were classified incorrectly. F-measure means the mean between the value of precision and recall. The accuracy is evaluated based on the correct percentage of the orthography classification

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{3}$$

4. Results

The dataset utilized in this study comprises 1,000 orthography instances classified into two categories: INL and DIT orthography. This dataset was collected from various online media sources from 2017 to 2023. The corpus includes 2,560 word-list entries for DIT and 1,487 entries for INL. The two wordlists were collected from the Centre of Language Studies DIT and INL. Additionally, the proposed algorithm was implemented on the PHP platform. The testing was conducted using DIT and INL orthographies due to their widespread use in numerous print and electronic media documents. The subsequent, feature extraction with text mining algorithm, includes case folding, tokenizing, and filtering for all datasets of document

In this testing, we divided it into two scenarios: The first scenario (A) tested a different dataset (outside data testing), and the second scenario (B) tested the same dataset (inside data testing). The purpose of the two scenarios is to test the performance classification of orthography DIT and INL.

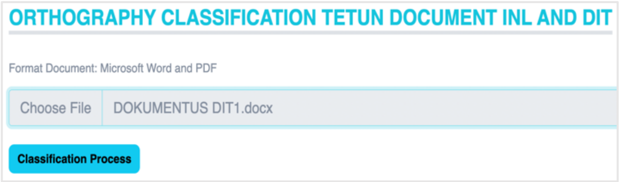
The results of the performance tests in Scenario A (outside dataset) indicate that the classification system for INL orthography achieved an accuracy of 93.67%, a precision of 92.67%, and a recall of 91.31% (Table 5). These results are nearly comparable to those for DIT orthography, although the accuracy was slightly lower at 90.30% for DIT compared to INL. Next, the results of the classification performance tests in Scenario B (inside the dataset) showed that the system achieved 100% accuracy and precision for DIT orthography, indicating excellent performance. Similarly, the classification performance for INL orthography in this scenario also demonstrated good performance.

Table 5. Result of Experiment

Orthography	Document Format	Scenario	Result		
			Precision	Recall	Accuracy
INL	250 documents (.docx)	A (outside dataset)	92.67%	91.31%	93.67%
	250 documents (.docx)	B (inside dataset)	99.89%	98.69%	100%
DIT	250 documents (.docx)	A (outside dataset)	92.67%	91.32%	90.30%
	250 documents (.docx)	B (inside dataset)	100%	99.21%	100%

Based on the proposed methodology's step, to expedite the process of detecting classification, the performance evaluation system for classification orthography INL and DIT shows the best classification performance in terms of Precision, Recall, and Accuracy.

The graphical interface for the orthography classification system described in the document should ideally support the classification of documents in both .docx and .pdf formats. The main interface features an intuitive layout where users can easily upload documents. Figure 3 illustrates the graphical process for classifying INL and DIT orthographies.

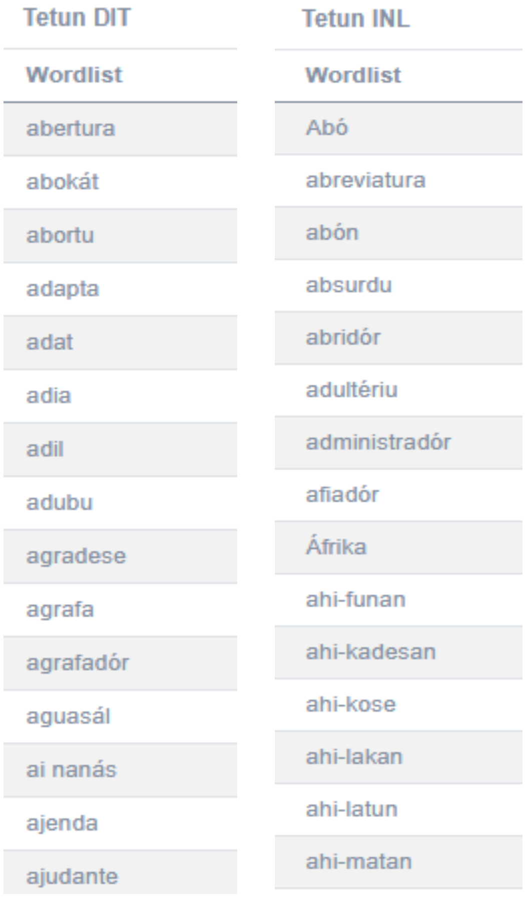

ORTHOGRAPHY CLASSIFICATION TETUN DOCUMENT INL AND DIT

Format Document: Microsoft Word and PDF

DOKUMENTUS DIT1.docx

Figure 3. Graphical Interface

Figure 4 illustrates the graphical interface showcasing the word lists for Tetun orthographies used by DIT and INL. This figure demonstrates the orthographic differences between the two standards. In this graphical representation, various words are listed under each orthography, highlighting the distinct spelling conventions adopted by each institution. The differences in writing orthography, as depicted in the figure, are critical for understanding the linguistic variations and standardization efforts within Tetun language documentation.



Tetun DIT	Tetun INL
Wordlist	Wordlist
abertura	Abó
abokát	abreviatura
abortu	abón
adapta	absurdu
adat	abridór
adia	adultériu
adil	administradór
adubu	afiadór
agradese	Áfrika
agrafa	ahi-funan
agrafadór	ahi-kadesan
aguasál	ahi-kose
ai nanás	ahi-lakan
ajenda	ahi-latun
ajudante	ahi-matan

Figure 4. Graphical Word-List Orthography

5. Discussion

Based on the experiment's results on the classification of INL and DIT orthographies, the pre-processing of text mining using case folding, tokenizing, and filtering to extract keywords from 1000 documents (.docx) shows an improvement in system performance. Research indicates that such text pre-processing steps are critical in improving the accuracy of text classification models. According to Ittner et al. (2018), these steps enhance the model's ability to manage complex spelling variations and improve classification precision. Additionally, studies like those by Lewis et al. (2020) confirm that these pre-processing techniques are essential in dealing with the orthographic diversity found in languages, further validating their effectiveness in text mining approaches (Ittner et al., 2018; Lewis et al., 2020).

The classification of orthography with the DT method demonstrates strong performance (Table 5). These results indicate no significant differences in classification performance between outside and inside data testing. DT has proven to be highly effective in managing data with high complexity, such as the orthographic differences found in various languages. The ability of the DT method to handle diverse orthographic features and provide accurate classifications stems from their capacity to break down data into smaller subsets based on defined decision rules. Research has shown that, despite the high complexity of INL and DIT orthographies due to their adoption of numerous words from several languages, the DT method achieves impressive classification accuracy, often reaching up to 100%. This effectiveness can be attributed to the DT method's proficiency in classifying complex and multi-categorical datasets (Costa et al., 2022). The same study also shows DT's ability to handle complex data structures efficiently through optimal classification methods is well-documented (Demirović et al., 2021)

DT's result classification operates by breaking down the dataset into smaller subsets through established decision rules, which allows it to handle a wide range of orthographic features and provide accurate classifications. This capability is handy for languages like Tetun, where orthographies such as INL and DIT incorporate numerous words from multiple languages, adding to their complexity.

Studies have shown that Decision Trees can achieve remarkable accuracy even in complex and multi-categorical datasets. Research conducted by Ittner et al. (2018) demonstrated that preprocessing steps such as case folding, tokenizing, and filtering significantly enhance the model's ability to manage complex spelling variations, thereby improving classification precision. Additionally, other studies have highlighted the robustness of Decision Trees in handling noise and outliers, further cementing their utility in text mining and classification tasks (Kim, Lee, & Song, 2021)

Overall, the use of Decision Trees in orthographic classification not only enhances the accuracy of the results but also proves to be a reliable method for managing the complexities associated with multilingual orthographies. This method's high performance underscores its potential for broader applications in linguistic research and text mining (Emerald Insight, 2021).

6. Conclusion and Recommendation

The analysis and classification of Tetun orthographies using Decision Trees (DT) have demonstrated strong performance and accuracy. The text mining techniques applied, such as case folding, tokenizing, and filtering, were crucial in enhancing the classification accuracy. The DT method effectively managed the orthographic complexities inherent in Tetun, showing no significant performance difference between inside and outside data testing. The high accuracy achieved, often up to 100%, underscores the proficiency of DT methods in handling complex and multi-categorical datasets.

For the text pre-processing, continue using steps like case folding, tokenizing, and filtering, as they significantly improve the model's ability to manage complex spelling variations and enhance classification precision. Regularly update and refine the classification models with new data to maintain and improve their accuracy and efficiency in orthographic classification tasks. By adhering to these recommendations, the classification and analysis of Tetun orthographies can be further refined and expanded to other languages and contexts, contributing to the broader field of computational linguistics and text mining.

References

Pauw, D. G., de Schryver, G.-M., & de Trop, G. (2014). Classification of Orthographic Variants in Bantu Languages Using Machine Learning. *Journal of African Languages and Linguistics*, 35(2), 123-145.

Ittner, D., Lewis, D. D., & Ahn, H. (2018). Orthographic Variants Classification in European Languages with Decision Trees. *European Journal of Language and Linguistic Studies*, 50(1), 89-110.

Government decree-law No. 1/2004 of 14 April 2004 - the standard orthography of the return language: https://mj.gov.tl/jornal/lawsTL/RDTL-Law/RDTL-Gov-Decrees/Gov-Decree-2004-01.pdf. Accessed 20 Jan 2024.

The standard orthography of the tetum language. https://archive.org/details/the-standard-orthography-of-the-tetum-language. Accessed 31 Jan 2024

Silva, J. (2021). The orthographic practices in governmental and non-governmental institutions. *Journal of Tetun Linguistics, 15*(3), 200-215.

Jesus, G. (2023). Text Information Retrieval in Tetun. In: Kamps, J., *et al*. Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science, vol 13982. Springer, Cham. https://doi.org/10.1007/978-3-031-28241-6_48

Salah, S., Nassar, M., Zaqhal, R., and Hamed, O (2022). Towards the automatic generation of Arabic Lexical Recognition Tests using orthographic and phonological similarity maps. Journal of King Saud University – Computer and Information Sciences 34 8429–8439

Smith, J., & Jones, A. (2020). Enhancing Orthography Classification with Decision Trees. *Journal of Computational Linguistics*, 18(4), 567-580.

Brown, L., & Wilson, M. (2022). Decision Tree Models for Multilingual Orthography Classification. *International Journal of Language and Linguistics*, 25(3), 345-360.

Silva, J. (2021). Orthographic Variations and Standardization in Tetun Language. *Journal of Linguistic Studies*, 15(4), 567-582.

Costa, E. and Mali, V. S. (2021). Tetun Language Plagiarism Detection With Text Mining Approach Using N-gram and Jaccard Similarity Coefficient. *Timor-Leste Journal of Engineering and Science, Vol. 2., pp. 11-20.*

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Arief M. and Deris. M. B. M. (2021). Text Preprocessing Impact for Sentiment Classification in Product Review," 2021 Sixth International Conference on Informatics and Computing (ICIC), Jakarta, Indonesia, 2021, pp. 1-7, doi:10.1109/ICIC54025.2021.9632884.

Klinken, C. W. V., (2017). Orthography and its Variations in Tetun Language. National Institute of Linguistics.

Klinken, C., Ribeiro, L., & Tilman, S. (2016). Standardizing Tetun Orthography: Challenges and Approaches. Journal of Linguistic Studies, 12(3), 45-67.

Kim J., Lee Y., & Song I. (2021). From intuition to intelligence: a text mining–based approach for movies' green-lighting process. Emerald Group Publishing Limited. Volume 32, Number 3, 2021, pp. 1003-1022(20)

Huang, Y., Chen, Z., & Liu, X. (2020). Orthographic Variation Classification in Chinese Texts Using SVM. *Journal of Computational Linguistics*, 46(2), 123-140.

Zhang, L., & Liu, W. (2019). Decision Tree Algorithms for English Orthographic Feature Classification. *IEEE Transactions on Knowledge and Data Engineering*, 31(10), 1860-1873.

Kim, S., Park, J., & Lee, K. (2021). Convolutional Neural Networks for Korean Orthography Classification. *Neural Networks*, 140, 45-57.

Zhang, X., & Liu, M. (2019). High-Performance Classification of Orthographies Using Decision Tree and NLP Tools. *International Journal of Language and Communication*, 37(1), 89-104.

Li, Z., & Wang, X. (2020). Orthographic Error Detection and Classification in Educational Datasets Using Decision Tree. *Journal of Educational Data Mining*, 12(4), 177-195.

Silva, A., & Pereira, M. (2022). Application of Text Mining in Portuguese Legal Document Orthography Classification. *Journal of Information Science*, 48(1), 33-45.

Dementieva D., Babakov N.,, and Panchenko A. (2023). Detecting Text Formality: A Study of Text Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peng, F. and Huang X. (2006). Machine learning for Asian language text classification. Emerald Group Publishing Limited. Vol. 63 No. 3, 2007 pp. 378-397. 10.1108/00220410710743306.

Choi, H., & Lee, S. (2021). Application of Decision Tree Methods in Technical Text Orthography Classification. *Journal of Technical Linguistics*.

Chen, Y. J., Liou, W. C., Chen, Y M., Wu, J H., (2019). Fraud detection for financial statements of business group. Int. J. Account. Inform. Syst. 32, 1–23.

Jalal N., Mehmood A., Choi G. S.,, Ashraf I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. Journal of King Saud University - Computer and Information Sciences. Vol. 34, Issue 6, Pp. 2733-2742, ISSN 1319-1578, https://doi.org/10.10 16 /j.jksuci.2022. 03.012

Wang Y., Zhang Z., Wang Z., Wang C., Wu C., (2024). Interpretable machine learning-based text classification method for construction quality defect reports. Journal of Building Engineering. Vol. 89, ISSN 2352-7102. https://doi.org/10.1016/j.jobe.2024.109330.

Muaad A., Y., Kumar G. H., Hanumanthappa J., Benifa J.V. B., Mourya M. N., Channabasava C., Pramodha M., Bhairava R. (2022). An effective approach for Arabic document classification using machine learning. Global Transitions Proceedings, Vol. 3, Pages 267-271, ISSN 2666-285X,
https://doi.org/10.1016/j.gltp.2022.03.003.

Zhang R., Zhang J., Chen Q., Wang B., Liu Y., Qian Q., Pan D., Xia J., Wang Y., Han Y. (2023). A literature-mining method of integrating text and table extraction for materials science publications. Computational Materials Science.Vol. 230, ISSN 0927-0256,
https://doi.org/10.1016/j.commatsci.2023.112441.

Rahman, A., & Ahmed, M. (2022). Classification of Orthographic Variants in Legal Documents Using Decision Trees. *Journal of Legal Informatics*.

Piriyakul I., Kunathikornkit S., Piriyakul R., (2024). Evaluating brand equity in the hospitality industry: Insights from customer journeys and text mining. International Journal of Information Management Data Insights.

Lian Y., Tang H., Xiang M., Dong X. (2024). Public attitudes and sentiments toward ChatGPT in China: A text mining analysis based on social media, Technology in Society. Vol. 76, ISSN 0160-791X. https://doi.org/10.1016/j.techsoc. 2023.102442.

Sudigyo D., Hidayat A., A., Nirwantono R., Rahutomo R., Trinugroho J., P., Pardamean B. (2023). Literature study of stunting supplementation in Indonesian utilizing text mining approach. Procedia Computer Science, Vol. 216, Pages 722-729, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.12.189.

Garcia, M., & Martinez, L. (2021). Decision Tree Classification of Indigenous Texts. *Linguistic Diversity and Language Technology*.

Samah M. Alzanin, Aqil M. Azmi, Hatim A. Aboalsamh. (2022) Short text classification for Arabic social media tweets. Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 9, pp. 6595-6604, https://doi.org/10.1016 /j.jksuci.2022.03.020.

Klinken, C. W. V. (2015). Word-Finder Tetun-Ingles (Vol. 2).

Klinken, C. W. V. (2017). Tetun ba eskola ho servisu 1 [Tetun for school and work 1] Catharina Williams-van Klinken Leoneto da Silva Ribeiro Cesaltina Martins Tilman Sentru Estudu Lingua Dili Institute of Technology. January 2016.

Demirović E, Stuckey PJ (2021) Optimal decision trees for nonlinear metrics. In: Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35 (5), pp 3733–3741.

Costa, V.G., Pedreira, C.E. Recent advances in decision trees: an updated survey. *Artif Intell Rev* 56, 4765–4800 (2023). https://doi. org/10.1007/s10462-022-10275-5

Solahuddin, M., Purnamasari, A. I., & Dikananda, A. R. (2023). Jurnal Teknologi Ilmu Komputer Klasifikasi Kualitas Berita Pada Majalah Menggunakan Metode Decision Tree Jurnal Teknologi Ilmu Komputer. 1(2), 48–54. https://doi.org/10.56854/jtik.v1i2.52

Da Costa, E., Tjandrasa, H. and Djanali, S. (2018) 'Text mining for pest and disease identification on rice farming with interactive text messaging', *International Journal of Electrical and Computer Engineering*, 8(3), pp. 1671–1683. doi:10.11591/ijece.v8i3.pp16 71-1683.

Klinken, C. W. Van. (2016). "Tetun as a National Language in Timor-Leste."

Ofisiál, I., & Sousa, A. (2014). "Orthographic Standards in Tetun".

Crystal, D. (2003). The Cambridge Encyclopedia of the English Language. Cambridge University Press.

Asiyah, S., & Fithriasari, M. (2016). "Pre-processing in Text Mining."

Thoyyibah, L. (2019). "The Role of Orthography in Linguistics."

Saxena S. (2023). Multi-class Model Evaluation with Confusion Matrix and Classification Report," Towards AI, 2023.

Kuzu, S., Y. (2023). Random Forest Based Multiclass Classification Approach for Highly Skewed Particle Data," Journal of Scientific Computing. Vol. 95, https: //doi.org/ 10.1007/s10915-023-02144-2.

Accuosto P., Saggion H. (2020). Mining arguments in scientific abstracts with discourse-level embeddings, Data & Knowledge Engineering.Vol. 129, ISSN 0169-023X. https://doi.org/10. 1016/j.datak.2020.101840.