Timor-Leste Journal of Engineering and Science

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

Tetun Language Plagiarism Detection With Text Mining Approach Using N-gram and Jaccard Similarity Coefficient

Edio da Costa, Vasco Soares Mali
Department of Computer Science, School of Engineering and Science, Dili Institute of Technology, Timor-Leste
Email: ediocosta73@gmail.com, vascosoares96@gmail.com

ABSTRACT

The objective of this research is to develop Tetun language plagiarism detection application with the Text Mining approach that performs Tokenizing and Filtering that use to extract and select a word list from the title of the thesis that is submitted by the students. The n-grams and Jaccard Similarity Coefficient methods are used to retrieve the letter characters in the document to be matched and calculate what percentage of the similarities in the processed thesis title. The dataset used in this study was obtained from the Dili Institute of Technology (DIT) Library with a total of 1000. The word dictionary used consists of 2.560 Word Lists and 8.972 Stop Words that were obtained from the Language Centre of DIT. The result of experiment shows that the performance detection plagiarism obtained the highest precision and recall is 0.90 and 0.94

Keywords: Tetun language, plagiarism detection, text mining, n-grams, jaccard similarity coefficient.

Received July 10, 2021; Revised September 29, 2021; Accepted November 04, 2021

1. Introduction

The advancement of information technology has resulted in massive textual material that is open to appropriation. In academic writing, plagiarism detection is important for an educational institution (Oberreuter and Velásquez, 2013). Due to researchers' misconduct, a plethora of plagiarism detection systems has been developed. However, most plagiarism detection systems on the market do not support for all language (Al-thwaib and Hammo, 2020).

Several studies on plagiarism detection in the academic world have been conducted, such as plagiarism detection for student reports (Sakamoto and Tsuda, 2019), academic Arabic Corpus for detection plagiarism (Al-thwaib and Hammo, 2020), author's Style for plagiarism detection in academic environments (Vysotska, 2018), academic for self-plagiarism detection (Horbach and Hal, 2019) and plagiarism in nursing education (Carter, Hussey and Forehand, 2019).

Several researchers have developed tools for plagiarism detection on textual documents in several languages such as Turnitin, Grammarly, SmallSeoTools, Plagramme, Unichek, and Plagscan, majority of these applications are implemented for the English language (Turnitin, 2021; Grammarly, 2021; SmallSeoTools, 2021; Plagrame, 2021; Ho *et al.*, 2017). However, most of these tools and applications are used to detect plagiarism in English text documents. There are three types of categories in plagiarism such as (i) detection plagiarism based on them on the database including website page, (ii) detecting plagiarism using searching mechanism such as Google; (iii) comparison instrument tools for compared documents (Fish and Hura, 2013; Metz, 2016; Henriques, 2015).

There are several approaches and methods used for the detection of plagiarism such as TF-IDF, Vector Space Model, and Cosine Similarity. TF-IDF and Cosine Similarity are methods of treating a text as a set of words (Sakamoto and Tsuda, 2019). While n-grams is a method that uses the value of n. Where n is the desired word separator factor (Mcnamee, 2004). The characters in the text are used to determine the level of similarity of words. The advantage of the n-grams method is that it does not identify the writing errors made by the author. Based on several studies conducted it shows that the most effective value is 3 and 4 (Kosmajac and Keselj, 2017), So in this study, we opted to use the n-grams size of 4 for character 4-grams.

Tetun language is one of the two official languages of the Republic Democratic of Timor-Leste (Klinken, 2015). In the last twenty years, many references have been written in Tetun, such as in universities in Timor-Leste, many students have written their thesis in Tetun. Many writers have written in various orthographies, for example, some have written '*ne'ebe*', *'nebe'*, and some have written *'nasaun*', 'nação', '*nacao'*. Based on low decree 1/2004 the official orthography used is *Institutu Nasionál de Linguística* (INL) orthography. Currently, the Dili Institute of Technology (DIT) has also developed its own orthography, most of which follow the INL orthography. The rules for writing DIT phonology follow today's Tetun rules, which do not use accents, and have few features because according to many studies this is not necessary (Klinken, 2019).

So, this research uses DIT orthography because almost of all students studying at DIT write their thesis using DIT

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

orthography. Until now, the title of the research, thesis, has not been widely published online in Tetun, so this research proposes to develop a plagiarism detection application in Tetun language texts that focus on checking two input compared documents. The objective of this research is to build the detection plagiarism application in Tetun Language text document to assist the process of identifying the level of plagiarism in the title of the thesis that submitted by students.

The research consists of six sections: Following this section, the second describes theoretical concept and framework, the third section describes the proposed method for Tetun language detection plagiarism using text mining approaches with tokenizing and filtering using to extract the text document. Next, we use the n-grams method to take letters from several strings from a word that is continuous to end the document with 4-grams. The fourth section describes the experimental results of keywords extraction with tokenizing, filtering, hashing method with 4-grams, and performance of detection plagiarism with Jaccard Similarity Coefficient. The fifth section is discussion. Finally, this research is concluded in section six.

2. Related Work

Text mining is used to identify new information or terms from large amounts of unstructured text documents (Kumar and Tripathi, 2015), Text mining has implemented in several fields such as information trend analysis (Valsamidis et al., 2013) and extraction of text information from different websites sources (Openminted, 2016). In the academic environment text also implemented such as detection plagiarism thesis of student in Indonesian language (Parwita, Indradewi and Wijaya, 2019), detecting plagiarism in source codes of students learning programming languages (Sharma and Sharma, 2015) and plagiarism detection assignments in the format of text documents (Hasan, Wicaksana and Hansun, 2018). Text mining approach using n-grams has been carried out by (Loseu, Ghasemzadeh and Jafari, 2012) to help course coordinators efficiently select course topics that cover course specifications without manual work. Text mining is divided into seven groups namely; (i) Document classification, (ii) Information extraction, (iii) Web mining, (iv) Document classification, (v) Information search and acquisition, (vi) Natural language processing, and (vii) Concept extraction (Aninditya, Hasibuan, and Sutoyo, 2019).

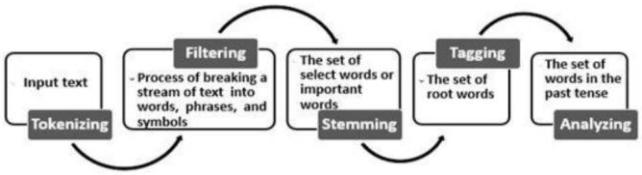

Figure 1. The stages of text mining (Da Costa, Tjandrasa and Djanali, 2018)

Figure 1. showed the five stages of Text Mining that are used to extract the keyword from the unstructured document (Mooney R.J., 2006). However, in this study, we used two processes, namely: Tokening and Filtering. Tokenization is an important process used to break the text into parts of a word (Putra, Gunawan, and Suryatno, 2018).

N-gram is one of the most widely used methods in text mining (word processing) and language processing. The n-grams method is used to generating of words or characters from a word that is continuous read from the source text to the end of the document (Parwita, Indradewi and Wijaya, 2019). In detecting plagiarism the n-gram method greatly affects the level of accuracy or similarity of a document being compared (Yudhana *et al.*, 2018). There are tree popular of n-grams that consists of bi-gram (2 words), tri-gram (3 words), and four-gram (4 words) (Cavnar and Trenkle, 2001). Based on the value of n, the extraction of documents using n-grams has been carried out by (Tanantong, Kreangkriwanich, and Laosen, 2020), the results of the study showed the highest precision value is 70%. The same research was also carried out by (Setiawan et al., 2018), the results of study showed the value of the precision obtained was 80%. Research conduct by (Suzuki et al., 2008) also uses n-grams as feature terms improving to improve accuracy. Because in multi-language identification, n-grams does not depend on grammar but depends on the number of characters being compared.

Plagiarism is a crime that is familiar in the academic world. In the English dictionary, The Oxford Advanced Learner's Dictionary defines plagiarism as ‘to copy somebody else’s idea or words and use it like if they were one’s own (Yudhana *et al.*, 2018). Plagiarism is divided into 2 types, namely Literal Plagiarism, and Intelligent Plagiarism (Alzahrani, Salim and Abraham, 2012). Literal plagiarism (paraphrasing) is the act of copying and pasting writings sourced from the internet without mentioning the original document reference source. Intelligent plagiarism (summarization) is an act where the author does not include sources from other people's writing and acknowledges it as his own writing. Plagiarism is a problem that can be treated from two perspectives, prevention and detection (Oberreuter and Velásquez, 2013). The methods of copy plagiarism detection can be concluded are easier to implement, and can solve the problem in different levels, from simple manual comparison to complex automatically algorithms (Potthast and Holfeld, 2011),

3. Research Methods

The Block diagram in Figure 2 describes the all process of plagiarism detection in the Tetun Language. The process consists of pre-processing and testing.

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

3.1. Characteristics of Tetun Language

The Tetun language has twenty-six alphabet letters (Klinken, Ribeiro, and Tilman, 2016). In terms of the structure of the Tetun language, especially orthography, the use of vocabulary and grammar is varied (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 write in Tetun *"nasaun"*. Furthermore, writing Vocabulary for engineering in Tetun, many people have problems because of the lack of technical terms, for example, the words *"hosting"* we do not find 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 consists of 2.560 Word Lists and 8.972 Stop Words.

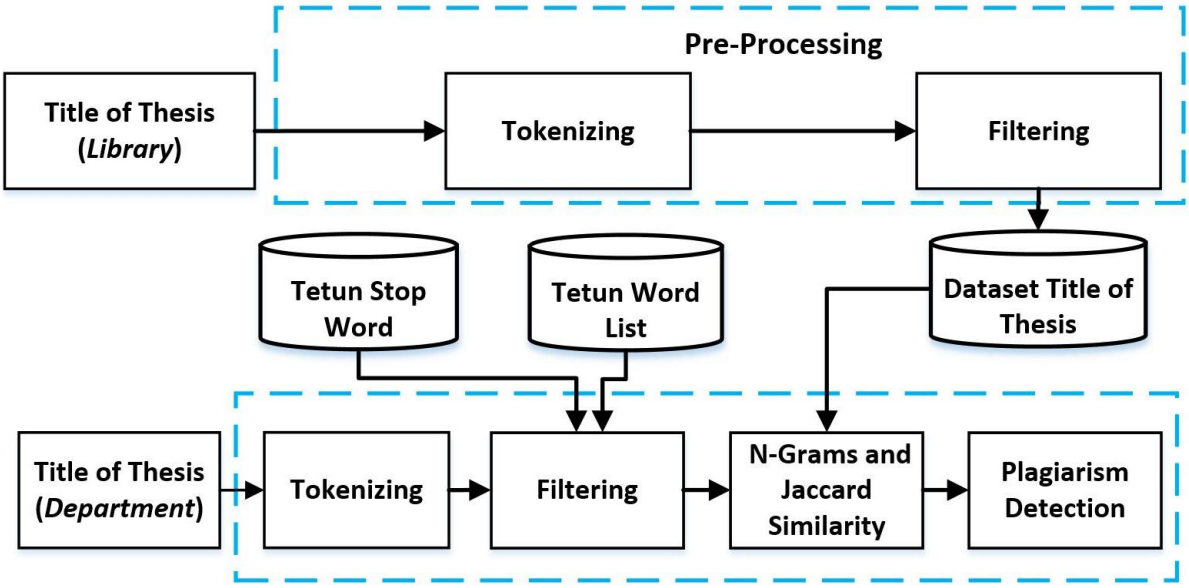

Figure 2. Block Diagram of the Research

3.2. Pre-processing

The dataset title of the thesis in this study was obtained from the library of the Dili Institute of Technology (DIT). The dataset consists of 1000 titles of articles, obtained from 2017 to 2020. The pre-processing process was used to extract the data are collected from the library. To extract the title of the thesis, the authors adopt the Tokenizing and Filtering are proposed by (Da Costa, Tjandrasa and Djanali, 2018).

Table 1. Example Title of Thesis

Code	Title of Text Input
DS1	*Sistema Monitorizasaun Dadus Monografia iha Dili Institute of Technology (DIT) Bazeadu Web*
DS2	*Sistema Deteksaun Plagiarismu Titulu monografia iha Dili Institute of Technology (DIT) Bazeadu ba Web*

To perform detection plagiarism of the title of thesis, we propose preprocessing text with case folding and tokenizing. Table 1. Shows the sample of the thesis obtained from the library. The list of the keyword that obtained in the process of the Tokenizing is used as a model to identify the similarity of the text that input by the department. Table 2. shows the result of Tokenizing for document DS1 is twelve tokens and document DS2 is thirteen tokens.

Table 2. The Result of Tokenizing and Filtering

Result of Tokenizing		Result of Filtering	
DS1	**DS2**	**DS1**	**DS2**
sistema monitorizasaun dadus monografia iha dili institute of technology dit bazeadu web	sistema deteksaun plagiarismu titulu monografia iha dili institute of technology dit bazeadu web	sistema monitorizasaun dadus monografia dili institute technology web	sistema deteksaun plagiarismu titulu monografia dili institute technology web

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

Then, perform filtering to remove stop words, we first eliminate the punctuation such as '(',')','.','-','/','?' Next, use the filtering to eliminate the word is unimportant such as "iha", "of", "dit", "bazeadu", This approach is done by comparing the text documents with a word in the stop word dictionary. Then the important word list that obtained in document DS1 and DS2 are 8 and 9.

Next, using n-grams for text categorization, then we split the sentence into the n categorization. An n-grams categorization is a substring of n consecutive words (Al-thwaib and Hammo, 2020). In this study we use the n-grams method take letters from several string from a word is continuously to end the of the document (Parwita, Indradewi and Wijaya, 2019). There is tree popular of n-grams that consists of bi-gram (2 words), tri-gram (3 words), and four-gram (4 words). The value n that use in the research is n=4. The objective of this step is to generate the text based on the number of n (Badawy et al., 2018). Then the results of n-grams that obtained in the document DS1 and DS1 are shown in Table 3.

Table 3. The Result of 4-grams

DS1	sist iste stem tema emam mamo amon moni onit nito itor tori oriz riza izas zasa asau saun aund unda ndad dadu adus dusm usmo smon mono onog nogr ogra graf rafi afia fiad iadi adil dili ilii liin iins nsti stit titu itut tute utet tete etec tech echn chno hnol nolo olog logy ogyw gywe yweb
DS2	sist iste stem tema emam emad made adet dete etek teks eksa ksau saun aunp unpl npla plag lagi agia giar iari aris rism ismu smut muti utit titu itul tulu ulum lumo umon mono onog nogr ogra graf rafi afia fiad iadi adil dili ilii liin iins inst nsti stit titu itut tute utet tete etec tech echn chno hnol nolo olog logy ogyw gywe yweb

Next, we use the hashing method to convert each character into numbers based on the value of ASCII html. The result hashing of string "sist" is as follows:

Ascii s = 115
Ascii i = 105
Ascii s = 115
Ascii t = 116

$\mathbf{H_{(sist)}} = asci_{(s)} \times 2^4 + asci_{(i)} \times 2^3 + asci_{(s)} \times 2^2 + asci_{(t)} \times 2^1$
$= 115 * 16 + 105 * 8 + 115 * 4 + 116 * 2$
$= 1840 + 840 + 460 + 232$
$= 3372$

So that the hashing result of document DS1 and DS2 are shown in the Figure 2, total number of n-grams of the documents DS1=58 and DS2=67.

Next, we use the Jaccard similarity coefficient for calculating similarity on two samples or documents being compared using the following formula

$$Jaccard(D,Q) = \frac{|D \cap Q|}{|D \cup Q|} = \frac{|D \cap Q|}{|D| + |Q| - |D \cap Q|} x100\%$$

Where D is fingerprint generate from the data in the document 1 and Q is fingerprint generate from the data in the document 2. Then the result hashing in document DS1 = 58 and DS2 = 67, so that the value of intersection = 47, while of union = 57, then Jaccard Similarity Coefficient = 47/57 = 0.82*100% = 82%. Based on the result concluded that the similarity of the DS1 and DS2 is 82%.

3.3. Performance Evaluation

Evaluation of the performance of the plagiarism detection system in this study using the confusion matrix. The matrix evaluation is opted from (Oberreuter and Velásquez, 2013) and are common information that is implemented in the case of detecting plagiarism. True Negative (TN) means that the title of a thesis was categorized as plagiarized, which is incorrect. F-measure means the mean between the value of precision and recall. True Positive (TP) means that the title of thesis found to be plagiarized, the detection which is the correct plagiarism. False Positive (FP) means that the title of the thesis that should have been categorized as plagiarized was not. The accuracy is evaluated based on the correct percentage of detection plagiarism.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$F - Measure = \frac{2\ x\ Recall\ x\ Precision}{Recall + Precision} \quad (4)$$

Table 4. shows some sample of dataset that collected from the library of DIT. Total datasets in this research is 1000, start 2017 to 2020. Categorized in to four categories such as Computer Science, Civil Engineering, Mechanical Engineering, and Petroleum Engineering.

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

3372	3308	3286	3280	3410	3352	3346	3274	3382	3094
3304	3128	3346	3148	3010	3050	3214	3450	3376	3294
3306	3288	3250	3152	3218	2982	3060	3066	2988	3082
3174	3208	3190	3252	3354	3420	3394	3308	3458	3436
3330	3146	3268	3044	3078	3204	3302	3290	3270	3280
3202	3338	3006	3070	3256	3026	3058	3214	3424.	

a

3010	3050	3214	3450	3376	3294	3306	3288	3250	3152
3218	2982	2982	3060	3066	2988	3082	3174	3208	3190
3252	3354	3420	3374	3308	2842	3436	3330	3146	3268
3044	3078	4254	3302	3290	3270	3280	3210	3318	2964
2826	3058	3214	3424.						

b

Figure 2. The Result of Hashing Documents DS1 (a) DS2 (b)

Table 4. Sample of Dataset

Code	Thesis Title
DS1	Aplikasaun E-learning alfabetu iha escola pre-primaria Dominicana Hera Utiliza Macromedia Flash.
DS2	Sistema informasaun E Commerce faan tais iha Tais Market. Estudu kazu halao iha Tais Market.
DS3	Dezenvolve sistema monitorizasaun dadus bolseirus iha Ministeriu Saude bazeadu ba web
DS4	Sistema informasaun dadus partisipasaun feto Timor-Leste iha vida politika bazeadu ba website Estudu Kazu Fundasaun Caucus Feto Iha Politika
DS5	Sistema monitorizasaun dadus monografia iha Dili Institute Of Technology (DIT) bazeadu web.
DS6	Dezenu casing ba operasaun perfurasaun uza metodu Maximum Load iha posu "X" kampu "Y".
DS7	Evaluasaun no optimasaun ba bomba Sucker Rod utiliza metodu Tryal and Error HODI hasae laju produsaun ba posu"X" kampu "Y"
DS8	Analiza kondisaun estrada a'at uza metodu Bina Marga no PCI, estudu kazu Perumnas Ai-Lok Laran.
DS9	Analiza misturasaun Betaun FC 17,5 MPA ho utilizasaun material raihenek husi Mota We-Lala tuir padraun SNI.
. . . .	. . . .
DS1000	Dezenhu Meja teste Alternador kareta ho bateria 12V

Table 5. Distribution of Document

Category	Number of Title	Training	Testing
Computer Science	280	196	84
Civil Engineering	280	196	84
Mechanical Engineering	150	105	45
Petroleum Engineering	290	203	87

4. Results and Analysis

The dataset in this study consists of 1000 titles of the thesis with 4 categories, such as Computer Science, Civil Engineering, Mechanical Engineering, and Petroleum Engineering. The dataset was collected from the library of DIT in the year of 2017 to 2020. The corpus consists of 2.560 word-list and 8.972 stop words collected from the Centre of Language Studies DIT. Also, the proposed algorithm was implemented on the Php platform. We also use the DIT orthography because almost all students studying at DIT write their thesis using DIT orthography. The objective of the testing is to identify system performance with the Text mining approach using n-grams and Jaccard Similarity to identify the range of plagiarism based on the title of the thesis submitted by the students.

The next, feature extraction with n-grams for all titles of the thesis was performed. This process is the value of n-grams obtained from the title of the thesis are saved into the memory of hashing table. In this study, we used the 4-grams features of all titles of the thesis were extracted and testing procedures were used separately for each feature.

Then we divided system detection plagiarism using 2 scenarios: The first scenario, testing for different thesis title dataset (outside data testing), the second testing for the same thesis title dataset (inside data testing). The purpose of the two scenarios is to test the performance of plagiarism detection performance of student thesis titles.

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

Table 6. Result of Experiment

Dataset /year	Computer Science		Civil Engineering		Mechanical Engineering		Petroleum Engineering	
	Scenario A	Scenario B	Scenario A	Scenario B	Scenario A	Scenario B	Scenario A	Scenario B
2017	75.30%	100%	82.30%	100%	89.67%	100%	82.22%	100%
2018	73.67%	100%	80.67%	99.99%	76.10%	100%	81.34%	100%
2019	72.33%	99.21%	79.33%	100%	76.56%	100%	78.20%	100%
2020	68.34%	100 %	75.34%	100%	75.25%	99.92%	78.60%	98.99%

The first scenario (A) uses testing data from different years, namely the title of the thesis in 2017 as testing data and 2018 as comparison data for the Computer Science category. The same experiment was carried out for all categories. The results of the experiment in the Computer Science category with the thesis title in a different year (outside data testing) showed the highest percentage in 2017 and 2018 is 75.30% and 73.33, while in 2019 and 2020 the percentage level of similarity decreased to 72.33% and 68.34% (Table 6). Furthermore, for the Civil Engineering category, the highest percentage of similarity was in 2017 at 82.30%, followed by 2018 and 2019. For the Mechanical Engineering category, the experimental results showed the highest percentage of similarity levels in 2017 and 2019 is 89.67% and 76.56%, while in 2020 the percentage level of similarity decreased to 75.25% (Table 6). And finally, for the Petroleum Engineering category, the highest percentage of similarity in 2017 and 2018 was 82.22% and 81.34% (Table 6).

Next, for scenario B using testing data and comparison data in the same year with the same thesis title. The objective of this scenario is to test the system whether can detect similarities with 100% the same or not. For the Computer Science category, the result of experimental in 2017, 2018, and 2020 obtained the same level of similarity, namely 100% (Table 6). Furthermore, the result of the experiment for the Civil Engineering category shows that 2017, 2019, and 2020 obtained the same level of similarity, namely 100% (Table 6). For the Mechanical Engineering and Petroleum Engineering categories, in 2017, 2018, and 2019 they obtained the same percentage level of similarity, namely 100% (Table 6).

Based on the step of the proposed methodology, in order to expedite the process of detecting plagiarism the performance evaluation system for all categories (Table 7) showed the category of Computer Science and Civil Engineering obtained the highest recall of 0.94 and 0.92 compared to the other categories. While for precision, the Civil Engineering and Computer Science category still obtains the highest performance is 0.90 and 0.89 (Table 7). Furthermore, the f-measure for the Civil Engineering category obtained the highest performance compared to Computer Science is 0.89 and 0.87, while the lowest category was Mechanical Engineering is 0.70 (Table 7).

5. Discussion

The resulting experiment in this study is based on the dataset of four proposed categories, they are Computer Science, Civil Engineering, Mechanical Engineering, and Petroleum Engineering. This study uses 2 steps. First, using a text mining approach with tokenizing (remove spaces and punctuation), case-folding (changes capital letters to lowercase) and filtering to extract keywords from the thesis title using Tetun DIT orthography. After filtering with stop words for each thesis title, the result shows that the number of tokens decreased by 35% with respect to the first calculation. This approach is very effective for performing feature extraction of the text document (Setiawan *et al.*, 2018) regard to the detection of plagiarism cases (Kumar and Tripathi, 2015). Once the processes with tokenization and filtering have been carried out, we analyzed the words that appear most frequently. Figure 3 (a), shows the most frequently Words List that often appear words are directly related to the Computer Science categories, such as: sistema, informasaun, maneza, php, mysql, bazeadu, maneza, website, etc. Figure 3 (b), shows the result of 3-grams of words "sistema", "informasaun" and so on.

Words	Frequency
sistema	1462
informasaun	1292
maneja	340
dili	323
php	306
mysql	272
bazeadu	221
maneza	187
website	153
estudante	153
hospital	136
technology	136

a

3-grams	Hashing Result
Sis	S : 115 , i : 105 , s : 115 = 1570
ist	i : 105 , s : 115 , t : 116 = 1532
ste	s : 115 , t : 116 , e : 101 = 1586
tem	t : 116 , e : 101 , m : 109 = 1550
ema	e : 101 , m : 109 , a : 97 = 1438
mad	m : 109 , a : 97 , d : 100 = 1460
adi	a : 97 , d : 100 , i : 105 = 1386
dis	d : 100 , i : 105 , s : 115 = 1450
ist	i : 105 , s : 115 , t : 116 = 1532
str	s : 115 , t : 116 , r : 114 = 1612
tri	t : 116 , r : 114 , i : 105 = 1594
rib	r : 114 , i : 105 , b : 98 = 1528
ibu	i : 105 , b : 98 , u : 117 = 1466
bui	b : 98 , u : 117 , i : 105 = 1462

b

Figure 3. Frequency of Word List Computer Science Category (a) and Result of 3-grams (b)

More over the system forms n-grams along n, and performs a hashing technique. The result of feature

Vol.2, Issue.1, pp.11-20, 2021
Available online at: http://tljes.org/index.php/tljes/data

extraction with 4-grams showed the highest accuracy, the research conducted by (Baygin, 2019) show that the calculation time increases with the increase in the value of n-grams. The other research also showed the n-grams method does not depend on the extracted language but depends on the n-grams string. So that the results of trials conducted by (Suzuki *et al.*, 2008) on the multilingual text extraction process obtained almost the same accuracy. The experiment with the corpus database, a linguist explained that the n-gram approach was used to inquire words and sentences in the database (Hammo et.al., 2016). The result of experiments shows the greater the value of n-gram, the level of similarity of two documents have a high degree of similarity, whereas the smaller the value of n-gram, the lower level of similarity in the two documents.

The experimental results show that the percentage of similarity in Scenario A shows a significant level of plagiarism, which is above 65% and is included in the category of plagiarism (Yudhana, 2019). Based on the result of experiment in scenario A found that a major portion of thesis title was plagiarized in the highest percentage. It is concluded that there is a major portion of thesis titles were plagiarized in the highest percentage for all categories between 2017 and 2018, 2018 and 2019, and finally 2019 and 2020.

Next, the experiment results of scenario B (inside data testing) shows the performance of the plagiarism detection system for all categories with the similarity percentage above 98.99%. So the conclusion of the proposed plagiarism detection system with a text mining approach using 4-grams and Jaccard similarity coefficient shows the best performance.

No.	Word List
1	3g
2	abilidade
3	abitasaun
4	abominasaun
5	abominavel
6	abordajen
7	aboresida
8	aboresidu
9	abreviasaun
10	abreviatura

Showing 1 to 10 of 2,560 entries

a

No.	Stop Words
1	aas
2	aat
3	abituadu
4	abokate
5	abolisaun
6	abortus
7	abranja
8	abranje
9	abranjente
10	abrelata

Showing 1 to 10 of 8,973 entries

b

Figure 4. Graphical of Word Lists (a), Stop Words (b)

Title of Thesis	Detection	Percentage
Sistema Informasaun Rekrutamentu Funsionariu Foun iha Komisaun Justisaem Paz Diosece Dili	15	15%
Dezenvolve sistema monitorizasaun dadus bolseirus iha Ministeriu Saude bazeadu ba web	15	15%
Sistema informasaun distribuisaun avizu ba komunidade suku Lourba utiliza sms gateway	15	15%
Sistema informasaun E-Commerce fan tais iha Tais Market Estudu kazu halao iha	15	15%
Sistema distribuisaun informasaun orario ba Liga Futebol Amadora utiliza SMS Gateway	15	15%
Sistema informasaun distribuisaun avizu ba komunidade suku Lourba utiliza sms gateway	15	15%
Sistema distribuisaun informasaun orariu ba Liga Futebol Amadora utiliza SMS Gateway	15	15%
Sistema Informasaun Geofrafiku Ba Area Turistiku Iha Munisipiu Liquisa Bazeadu Iha Website	15	15%
Sistema informasaun faan ai funan fresku iha kompanha Delta Flores bazeadu website	14.286	14.286%
Sistema informasaun maneza dadus transporte publico iha Direcao Nacional Transporte Terestre bazeado web	19.048	19.048%
Sistema informasaun maneja dadus atividades arte cultura iha Secretario Estado Arte no Cultura	19.048	19.048%
Sistema informasaun maneja dadus monografia iha Dili Institute of Technology DIT bazeadu web	19.048	19.048%
Sistema informasaun maneza dadus kazu insidenti iha Comando Geral Policia Nacional de Timor-Leste	19.048	19.048%
Sistema informasaun maneza dadus transporte publico iha Direcao Nacional Transporte Terestre bazeado web	19.048	19.048%
Sistema informasaun maneja dadus trabalhadores estranjeirus iha Timor Leste Estudu kazu halao iha SEPFOPE	19.048	19.048%
Sistema informasaun maneza dadus produtu kafe iha kompanha Nasional Cooperative Busines Assosiation NCBA	19.048	19.048%
Sistema informasaun aluga salaun iha Timor Top	18.75	18.75%
Sistema informasaun faan telefone iha Compania Fonehaus	18.75	18.75%

Figure 5. Result of Plagiarism Detection with Jaccard Similarity Coefficient

Table 7. Result of Experiment

	Computer Science	Civil Engineering	Mechanical Engineering	Petroleum Engineering
Precision	0.89	0.90	0.84	0.85
Recall	0.94	0.92	0.89	0.82
F-measure	0.87	0.89	0.70	0.81

Table 6 shows the similarity difference of thesis titles for the Computer Sciences category between 2017 and 2018 is 1.67%, 2018 and 2019 is 1.34%, and 2019 and 2020 is 3.99%. Furthermore, for the Civil Engineering category, the difference in similarity between 2017 and 2018 is 1.67%, 2018 and 2019 is 1.34%, and 2019 with 2020 is 3.99%. For the Mechanical Engineering category, the difference between 2017 and 2018 is 12.9%, and 2019 and 2020 is 1.31%. As for the Petroleum Engineering category, the difference in similarity between 2017 and 2018 is 0.88%, and 2018 and 2019 is 3.14%. The experimental results from the four categories indicate that the difference in the degree of similarity of the thesis titles submitted by students each year is relatively high.

6. Conclusion and Future Research

The system developed with a text mining approach using Tokenizing and Filtering can effectively extract keywords based on the Tetun DIT orthography. The 3-grams method and the Jaccard similarity coefficient can be used to process text to detect indications of plagiarism based on the word similarity. Based on the result of experiment in scenario A found that a major portion of thesis title was plagiarized in the highest percentage. The results of the plagiarism detection with the Jaccard similarity coefficient method show that the highest accuracy is obtained by the Civil Engineering and Computer Science categories.

The built system still cannot be accessed online, so that in further research, it can be developed to be accessible online and to be able to detect the similarity of words not only in documents but also from website URL links.

References

Al-thwaib, E. and Hammo, B. H. (2020) 'An academic Arabic corpus for plagiarism detection : design , construction and experimentation'. International Journal of Educational Technology in Higher Education, pp. 1–26.

Alzahrani, S. M., Salim, N. and Abraham, A. (2012) 'Understanding plagiarism linguistic patterns, textual features, and detection methods', *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(2), pp. 133–149. doi: 10.1109/TSMCC.2011.2134847.

Baygin, M. (2019) 'Classification of Text Documents based on Naive Bayes using N-Gram Features', *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*. IEEE, pp. 1–5. doi: 10.1109/IDAP.2018.8620853.

Carter, H., Hussey, J. and Forehand, W. (2019) 'Plagiarism in nursing education and the ethical implications in practice', (March). doi: 10.1016/j.heliyon.2019.e01350.

Cavnar, W. B. and Trenkle, J. M. (2001) 'N-Gram-Based Text Categorization N-Gram-Based Text Categorization', *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, (December 2012), pp. 1–14.

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.pp1671-1683.

Fish, R. and Hura, G. (2013) 'Students ' perceptions of plagiarism', 13(5), pp. 33–45.

Grammarly (2021) *Great Writing, Simplified*. Available at: https://www.grammarly.com/ (Accessed: 20 July 2021).

Hasan, E. G., Wicaksana, A. and Hansun, S. (2018) 'The Implementation of Winnowing Algorithm for Plagiarism Detection in Moodle-based E-learning', *Proceedings - 17th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2018*. IEEE, pp. 321–325. doi: 10.1109/ICIS.2018.8466429.

Henriques, P. R. (2015) 'An AST-based Tool , Spector , for Plagiarism Detection : The Approach , Functionality ', pp. 153–159. doi: 10.1007/978-3-319-27653-3.

Ho, P. H. *et al.* (2017) 'Data Warehouse Designing for Vietnamese Textual Document-based Plagiarism Detection System'.

Horbach, S. P. J. M. S. and Hal, W. W. (2019) 'The extent and causes of academic text recycling or " self-plagiarism"', 48(September 2017), pp. 492–502. doi:10.1016 /j. respol.2017 09. 004.

Klinken, C. W. (2015) *Word-Finder*. Edisaun 2. Dili Institute of Technology.

Klinken, C. W. (2019) 'Dezenvolvimentu Lia-Tetun Tuir Dalan Informál', in *Timor-Leste Studies Association*.

Kosmajac, D. and Keselj, V. (2017) 'Language identification in multilingual, short and noisy texts using common N-grams', *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2018-Janua, pp. 2752–2759. doi: 10.1109/BigData.2017.8258240.

Kumar, R. and Tripathi, R. C. (2015) 'Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset', *Procedia - Procedia Computer Science*. Elsevier Masson SAS, 65(Icc), pp. 911–919. doi: 10.1016/j.procs.2015.09.062.

Loseu, B. V., Ghasemzadeh, H. and Jafari, R. (2012) 'A Mining Technique Using n-Grams andMotion Transcripts for Body Sensor Network Data Repository'.

Mcnamee, P. (2004) 'Character N -Gram Tokenization for European', pp. 73–97.

Metz, C. (2016) *Forget Apple vs. the FBI: WhatsApp Just Switched on Encryption for a Billion People*, *Wired*. Available at: http://www.wired.com/2016/04/forget-apple-vs-fbi-whatsapp-just-switched-encryption-billion-people/ (Accessed: 30 June 2018).

Oberreuter, G. and Velásquez, J. D. (2013) 'Expert Systems with Applications Text mining applied to plagiarism detection : The use of words for detecting deviations in the writing style', *Expert Systems With Applications*, 40(9), pp. 3756–3763. doi: 10.1016/j.eswa.2012.12.082.

Parwita, W. G. S., Indradewi, I. G. A. A. D. and Wijaya, I. N. S. W. (2019) 'String matching based plagiarism detection for document in Bahasa Indonesia', *Proceedings of 2019 5th International Conference on New Media Studies, CONMEDIA 2019*, pp. 54–58. doi: 10.1109/CONMEDIA46929.2019.8981821.

Plagrame (2021) *Plagiarism and originality detector*. Available at: https://www.plagramme.com/?gclid=Cj0KCQjw6NmHBhD2ARIsAI3hrM0XQjIsixEJuWVzYvXDyjVhoaL_BllHh00A39gSUw2ccO7axiho7AQaAoFBEALw_wcB (Accessed: 20 July 2021).

Potthast, M. and Holfeld, T. (2011) 'Overview of the 2nd international competition on Wikipedia vandalism detection', *CEUR Workshop Proceedings*, 1177(January 2010).

Mooney RJ, "Machine Learning Text Categorization", University of Texas Austin, 2006

Tanantong, T. K. S. and Laosen, N. (2020) 'Extraction of Trend Keywords from Thai Twitters using N-Gram Word Combination'. 17th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology. pp.320-323.

Sakamoto, D. and Tsuda, K. (2019) 'ScienceDirect ScienceDirect A Detection Method for Plagiarism Reports of Students A Detection Method for Plagiarism Reports of Students', *Procedia Computer Science*. Elsevier B.V., 159, pp. 1329–1338. doi: 10.1016/j.procs.2019.09.303.

Setiawan, E. I. *et al.* (2018) 'N-Gram Keyword Retrieval on Association Rule Mining for Predicting Teenager Deviant Behavior from School Regulation', *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2018 - Proceeding*. IEEE, pp. 325–328. doi: 10.1109/CENIM.2018.8710892.

Sharma, S. and Sharma, C. S. (2015) 'Plagiarism Detection Tool " Parikshak "'.

SmallSeoTools (2021) *Plagiarism Checker*. Available at: https://smallseo.tools/plagiarism-checker (Accessed: 20 July 2021).

Suzuki, M. *et al.* (2008) 'Multilingual text categorization using character N-gram', *SMCia/08 - Proceedings of the 2008 IEEE Conference on Soft Computing on Industrial Applications*, 2003, pp. 49–54. doi: 10.1109/SMCIA.2008.5045934.

Suzuki, M. *et al.* (2010) 'English and taiwanese text categorization using N-gram based on Vector Space Model', *ISITA/ISSSTA 2010 - 2010 International Symposium on Information Theory and Its Applications*, pp. 106–111. doi: 10.1109/ISITA.2010.5649453.

Turnitin (2021) *The new standard in academic integrity*. Available at: https://www.turnitin.com/products/originality?utm_source=Google&utm_medium=CPC&utm_campaign=APAC_ALL_AD_ID_Integrity_2021&utm_content=Originality&utm_country=ID (Accessed: 20 July 2021).

Vysotska, V. (2018) 'Defining Author ' s Style for Plagiarism Detection in Academic Environment', *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, pp. 128–133.

Yudhana, A. *et al.* (2018) 'Implementasi Deteksi Plagiarisme Menggunakan Metode n-gram dan Jaccard Similarity Terhadap Algoritma Winnowing', (3), pp. 2–7.

Yudhana, B. A. (2019) 'Implementation of Pattern Matching Algorithm for Portable Document Format'.

Klinken, C. W. Ribeiro, Leoneto da S. Tilman, C. M. (2016). Tetun ba Eskola ho Servisu 1. Pp.16-17.

Badawy, M. Mahmood, M. El-aziz, A. Hefny, H. A. A. (2018). Text Mining Approach for Automatic Selection of Academic Course Topics based on Course Specifications. 2018 14th International Computer Engineering Conference (ICENCO). pp. 162-167