Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11224
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Indonesian Spelling Error Detection and Type Identification using Bigram Vector and Minimum Edit Distance Based Probabilities

Emmy Erwina[1)*], Tommy[2)], Mayasari[3)]
[1)2)3)]Universitas Harapan Medan, Indonesia
[1)] emmyerwina8@gmail.com, [2)] tomshirakawa@gmail.com, [3)] mayasaribuya1989@gmail.com

Submitted : Oct 1, 2021 | **Accepted** : Oct 15, 2021 | **Published** : Nop 24, 2021

Abstract: Spelling error has become an error that is often found in this era which can be seen from the use of words that tend to follow trends or culture, especially in the younger generation. This study aims to develop and test a detection and identification model using a combination of Bigram Vector and Minimum Edit Distance Based Probabilities. Correct words from error words are obtained using candidates search and probability calculations that adopt the concept of minimum edit distance. The detection results then identified the error type into three types of errors, namely vowels, consonants and diphthongs from the error side on the tendency of the characters used as a result of phonemic rendering at the time of writing. The results of error detection and identification of error types obtained are quite good where most of the error test data can be detected and identified according to the type of error, although there are several detection errors by obtaining more than one correct word as a result of the same probability value of these words.

Keywords: bigram; minimum edit distance; probabilities; spelling; vector;

INTRODUCTION

Spelling errors in Indonesian have become a natural thing in the current era of globalization. This can be seen from the high use of non-standard languages among the community as can be seen among students (Supriadin, 2016) (Ningrum, 2019) and on public area (Sirait, 2021). The phenomenon of spelling errors that are often found in society is influenced by various factors such as culture, trends and the influence of foreign languages (Erwina, 2012). Spelling error detection can be the first step in helping to overcome the high level of spelling errors in Indonesian. Various studies on the detection and correction of errors in Indonesian have been carried out. Several models have been developed for error detection and correction in Indonesian. These models were developed with various approaches and case studies such as a spell checker for patient complaints (Ratnasari, Kusumadewi, & Rosita, 2017), word correction for Indonesian historic newspapers (Purwantoro, Akbar, & Hidayati, 2019), and Indonesian spelling error correction (Santoso, Yuliawati, Shalahuddin, & Wibawa, 2019). The error detection stage is an important stage in correcting errors contained in words. One of the most common approaches to detecting errors in words is to look for differences between words with a dictionary or lookup table. In simple terms, a word is categorized as a word error if there is a difference with the whole word in the lexicon such as probability of similarity (Samanta & Chaudhuri, 2013) (Aşliyan, Günel, & Yakhno, 2007) (Christanti & Naga, 2018). A direct comparison between error words and lexicon will of course require expensive computational costs, especially if the dictionary or lexicon used has a large size. Several alternatives can be applied to increase efficiency in error word detection, such as selecting n-candidates. Xiang Tong proposed a simple way to generate candidate words by using vector space information retrieval technique (Salton, 1989), which then uses Term Freuqeuncy (TF) scoring to form a list of candidates (Tong & Evans, 1996).

Bigram is a form of n-gram which is an ordered pair of words or characters from the observed text. Bigram can be used as a feature in calculating the probability of the existence of a string in another string that has been used for a long time as can be seen in previous studies such as the role of bigrams in words and non-words perception (Rice & Robinson, 1975) and word clustering (Martin, Liermann, & Ney, 1998). This study aims to identify the types of misspellings errors and map them into pronunciation errors such as vowel errors, diphthong errors and consonant errors. The identification of the type of error used in this study adopts an error correction model that uses the concept of a minimum edit distance as can be seen in several studies related to error correction in the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11224
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Indonesian language. Rather than using direct Minimum Edit Distance like Damrau and Levenshtein Distance (Kamayani, Reinanda, Simbolon, Soleh, & Purwarianti, 2011) (Wibawa, Yuliawati, Santoso, Shalahuddin, & Wirawan, 2020), this study uses probability values as done by Xiang Tong (Tong & Evans, 1996) and (Brill & Moore, 2000).

LITERATURE REVIEW

Spelling Error

Spelling error was a form of unusually written word that makes the text harder to read and process (Hládek, Staš, & Pleva, 2020). In general, spelling errors are in the form of typing errors in text which will result in a string experiencing excess, deficiency and change of one or more characters compared to text that matches the lexicon. Spelling Error Detection and Correction generally consists of three main stages, namely lexicon preparation, candidate generation, string correction based on intended word and context (Hládek, Staš, & Pleva, 2020). Knowledge and information about the causes of spelling errors are needed in developing spelling error detection and correction models. Modeling the causes of spelling errors can be seen in the research conducted by Deorowicz and Ciura which describes spelling errors into the following categories (Deorowicz & Ciura, 2005) :

1. Mistyping, is the simplest error that can be in the form of addition, subtraction, substitution and transposition of the characters contained in the text.
2. Misspellings, are errors resulting from errors resulting from rendering the phones of the word. Some misspellings can also be found in the form of mistyping, but not because of accident, but due to habit or lack of understanding.
3. Vocabulary Incompetence Errors, are errors that are usually found in the use of prefixes, suffixes and affixes.

Bigram Vector

Bigram Vector is one form of vector space information retrieval technique (Salton, 1989) recommended by Xiang Tong in his research to generate string candidates (Tong & Evans, 1996). Xiang Tong indexes each data based on a trigram letter that includes the start and end symbols in the string. As an example quoted from Xiang Tong's research, a string "selamat" will produce a vector consisting of trigram letters as follows:

"selamat" => "#selamat#" => {#se, sel, ela, lam, ama, mat, at#}

In this study, the same concept is used but the vector formed consists of bigram letters so that the string "selamat" will produce the following vectors :

"selamat" => "#selamat#" => {#s, se, el, la, am, ma, at, t#}

Minimum Edit Distance Base Probability

Measurement of the similarity between a string with another string can be done by measuring the probability of occurrence of a string character in another string. If given a string w, then the probability of a character from string s in string w can be denoted by $p(s|w)$. If the characters contained in the string w are denoted by $t_1, t_2, \dots, t_i$ and $s_1, s_2, \dots, s_j$ are characters from the string s then $p(s|w)$ is the dot product of the probability $pr(t_{1,i}|s_{1,j})$ (Tong & Evans, 1996).

$$pr(i|j) = max \begin{cases} pr(i-1|j) * pr(ins(t_i)) \\ pr(i|j-1) * pr(del(s_j)|s_j) \\ pr(i-1|j-1) * pr(t_i|s_j) \end{cases} \quad (1)$$

Where $pr(ins(y))$ is the probability of character y being inserted.
$pr(del(y)|y)$ is the probability of character y being deleted.
$pr(x|y)$ is the probability of character y being replaced by character x.

In cases where character confusion probabilities are not available, the probabilities can be estimated by (Tong & Evans, 1996) :

$$pr(y|x) = \begin{cases} \alpha, if y = x \\ \frac{1-\alpha}{N}, otherwise \end{cases} \quad (2)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11224
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$pr(del(x)|x) = pr\big(ins(x)\big) = \frac{1-\alpha}{N} \quad (3)$$

Where N = printable character count.

METHOD

In this study, the lexicon data used is the Indonesian language dictionary (Badan Pengembangan dan Pembinaan Bahasa, 2018). The error data used in this study is observational data obtained from previous research references which will be used as test data for the model built. This study conducted a test using a single word input which aims to detect and identify the word error. Each error word will be tested into the detection model and the type of error will be identified from the error words. The error detection process is carried out to find the correct word from the input word. Based on the correct word obtained, then the type of error will be identified using the following simple error criterias :

1. Vocal Error, a vowel error is identified if one of the characters from a different bigram vector in the input character is a vowel character.
2. Consonant Errors, consonant errors are identified if the character errors found are consonant characters.
3. Diphthong error, diphthong error is identified if the two characters of the different bigram vector in the input character are vowel characters.

The simplification of the criteria for determining the error type is done to reduce the complexity of how words are rendered based on their phonemes, so that special phonemes representations of words are ignored and follow the character rendering of the lexicon used. Broadly speaking, the stages of the fault type detection and identification process can be seen in Fig. 1.

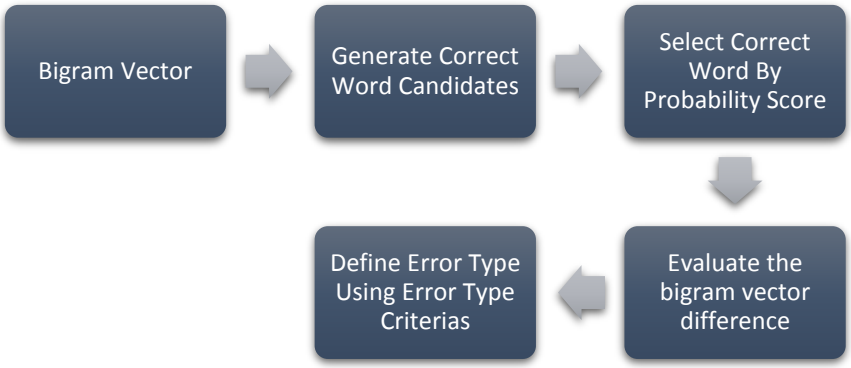

Fig. 1 Model Workflow

The error detection process begins by forming a Bigram Vector from the input word. For example, the error word used as the input word is "sampek" so that the Bigram Vector obtained is:

"sampek" => {#s, sa, am, mp, pe, ek, k#}

Based on the Bigram Vector obtained, words that have one or more of the bigram vector input words will be filtered to simplify computing as can be seen in fig. 2. Correct word candidates are then generated by calculating the Term Frequency score and selecting n-candidates based on the scores obtained as can be seen in table 1.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11224
e-ISSN : 2541-2019
p-ISSN : 2541-044X

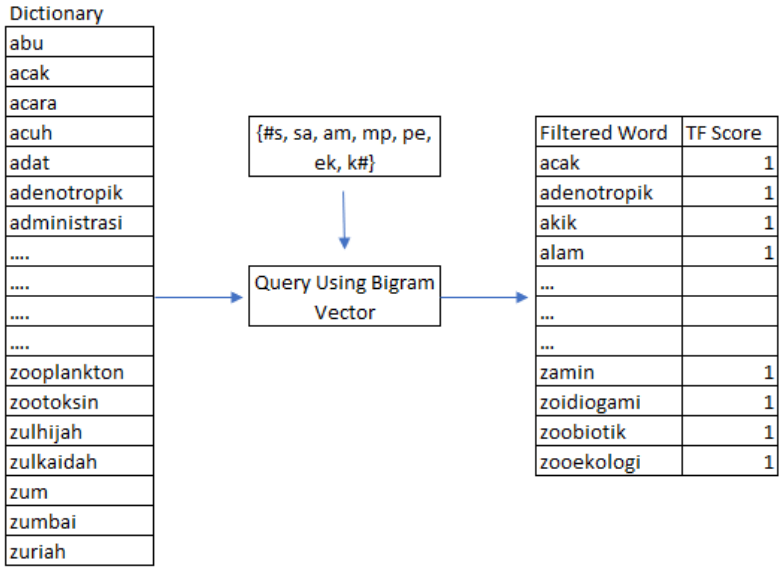

Fig. 2 Bigram Vector Filtering

Table 1. Candidate Generation, TF Score <= 4

Filtered Word	Bigram Vector	TF Score	Candidate
acak	{#a, ac, ca, ak, k#}	1	No
…	…	…	…
sampah	{#s, sa, am, mp, pa, ah, h#}	4	Yes
sampai	{#s, sa, am, mp, pa, ai, i#}	4	Yes
sampang	{#s, sa, am, mp, pa, an, ng, g#}	4	Yes
…	…	…	…
zooekologi	{#z, zo, oo, oe, ek, ko, ol, lo, og, go, i#}	1	..

After the correct word candidates are obtained, the best correct word will be selected using a probability calculation from the bigram vector between the error word and the correct word candidates. Each candidate correct error will have its probability value calculated first which can be described as follows:

$$pr(sampah|sampah) = pr(\#s|\#s) * pr(sa|sa) * pr(am|am) * pr(mp|mp) * pr(pa|pa) * pr(ah|ah) * pr(h\#|h\#)$$

By using alpha = 0.9, it is obtained:

$$pr(sampah|sampah) = 0.9 * 0.9 * 0.9 * 0.9 * 0.9 * 0.9 * 0.9 = 0.4782969$$

After the probability for the correct word is calculated, then the probability error word for the correct word will be calculated which can be described as follows:

$$pr(sampek|sampah) = pr(\#s|\#s) * pr(sa|sa) * pr(am|am) * pr(mp|mp) * pr(pe|pa) * pr(k\#|ah) * pr(del(h\#)|h\#)$$

$$pr(sampek|sampah) = 0.9 * 0.9 * 0.9 * 0.9 * 0.014 * 0.014 * 0.014 = 1.80034E-06$$

So that the final probability is obtained as follows:

$$pr(sampek|sampah) = \frac{pr(sampek|sampah)}{pr(sampah|sampah)} = \frac{1.80034E-06}{0.4782969} = 3.76406E-06$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11224
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Probability Score

Candidates	Probability	Correct Word
sampah	3.76406E-06	Yes
sampai	3.76406E-06	Yes
sampak	5.8552E-08	No

From the calculation of probability, obtained two correct words with the greatest probability value. The next step is to identify the type of error using simple criteria based on the bigram vector difference between the correct words obtained and the error word input. In the correct word “sampah”, the difference in the bigram vector obtained is “pa”, “ah”, “h#” from the bigram vector, the character differences with the bigram vector from the error word will be searched which can be seen in table 3.

Table 3. Bigram Vector Differences and Error Types

	Bigram Vector		
“sampek”	pe	ek	k#
“sampah”	pa	ah	p#
differences	a	ah	P
Error Type	vocal	consonant	consonant

Based on the differences in bigram vectors as can be seen in table 4, the first difference between the two bigram vectors is "pe" and "pa" where the difference in the character of the two bigram vectors is "a" which based on the simple criteria used can be identified the errors found are vowel errors. In the next bigram vector, the error types are consonants and consonants, so for the correct word "sampah" from the error word "sampek" two errors are obtained, namely vowels and consonants. Meanwhile, for the correct word character “sampai”, there are vowel errors and diphthong errors.

RESULT

The spelling error detection and identification model is built to be able to detect errors contained in words and identify the types of spelling errors of the word based on the difference with the original word. The model used in this study is specifically used to identify errors in a word without paying attention to the context of the sentence. Model testing is done by detecting and identifying errors in the dataset obtained from observations of respondents which can be seen in table 4.

Table 4. Dataset

	id_data	id_person	id_jenis_kata	id_jenis_salah	id_profesi	id_tingkat_profesi	jk	usia	label_profesi	label_tingkat_profesi	label_jenis	kelas_salah	kat
0	1	1	3	1	1	1	Laki-Laki	46	Guru	SD	Kata Keterangan	Vokal	
1	2	1	1	2	1	1	Laki-Laki	46	Guru	SD	Kata Kerja	Diftong	sa
2	3	1	1	1	1	1	Laki-Laki	46	Guru	SD	Kata Kerja	Vokal	
3	4	1	2	1	1	1	Laki-Laki	46	Guru	SD	Kata Sifat	Vokal	
4	5	1	9	3	1	1	Laki-Laki	46	Guru	SD	Kata Hubung	Konsonan	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
102	126	12	9	2	2	4	Laki-Laki	35	Dosen	S1	Kata Hubung	Diftong	
103	127	12	9	3	2	4	Laki-Laki	35	Dosen	S1	Kata Hubung	Konsonan	
104	128	12	1	3	2	4	Laki-Laki	35	Dosen	S1	Kata Kerja	Konsonan	
105	129	12	9	2	2	4	Laki-Laki	35	Dosen	S1	Kata Hubung	Diftong	
106	130	12	12	2	2	4	Laki-Laki	35	Dosen	S1	Kata Tanya	Diftong	ba

107 rows × 13 columns

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, October 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11224
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The attributes used in this study are only error words contained in the dataset by ignoring other attributes. The examples of incorrect words contained in the dataset can be seen in table 5. The type of error for each error word in the dataset has been known to obtain the accuracy of the model used.

Table 5. Samples of Dataset Error Words

No.	Error Words	Error Type
1	pake	Diftong
2	teros	Vokal
3	kenderaan	Vokal
4	naek	Vokal
5	kalo	Diftong
6	seneng	Vokal
7	akherat	Vokal
8	liat	Konsonan
9	tutop	Vokal
10	kalo	Diftong
11	karna	Vokal
12	belom	Vokal
13	liat	Konsonan
14	trus	Vokal

Every error word contained in the dataset has been pre-processed to eliminate duplication. The results of the detection and identification carried out in this study can be seen in table 6. Accuracy will be measured by comparing the correct word generated and the type of error identified.

Table 6. Test Results

Error Types	Error Words	Count of Correct Error Words	Count of False Correct Words	Count of Correct Error Type Identified	Count of False Error Type Identified
Vokal	53	49	12	48	14
Konsonan	22	20	7	20	10
Diftong	31	29	10	29	5

The test results as can be seen in table 6 show the number of error words for each type of error, the true words detected column shows the number of correct words that were correctly detected from the given error words. False words detected shows an incorrect number of correct words. The accumulation of true error words detected and false words detected does not match the number of error words used for each type of error due to the possibility that the model produces more than one correct words, so any excess correct words that do not match will be counted as false words. This also applies to the true error type and false error type columns.

DISCUSSIONS

Detection and identification of spelling errors in Indonesian has its own challenges, considering the many variants of spelling errors found. The results obtained from this study indicate that almost all spelling errors contained in the dataset can be detected properly, where for the type of vowel error, 49 of 53 error words can be detected, which is around 92%, but because the model uses probability assessment, there is a possibility produces more than one correct word so that the excess will be considered as a false correct word which in the vowel error type is obtained by 22% of the total test. In the identification of error types, the number of correct error types identified is very dependent on the detected correct words, so the number of correct error types identified will not differ significantly from the number of detected correct error words.

CONCLUSION

The Bigram Vector model is used in this study to simplify the detection process by searching for correct word candidates from the lexicon or a large word dictionary. The final correct word is obtained by calculating the probability that adheres to the minimum edit distance concept. The error type is then obtained from the bigram

*name of corresponding author

vector differences using simple criteria. The results of the tests carried out show that for the type of vocal error, the accuracy of correct word detection is 92% with a false correct word size of 22%, correct error type identification is 90.56% and false error type identification is 26.41%. For other errors such as diphthongs and consonants, the percentage obtained is not so different from the data for vowel errors, so it can be concluded that the model used has the same relative performance against the three types of errors.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Indonesian Ministry Of Education, Culture, Research and Technology for funding this work through "Penelitian Dasar Unggulan Perguruan Tinggi (PDUPT)" grant of 2021.

REFERENCES

Aşliyan, R., Günel, K., & Yakhno, T. (2007). Detecting misspelled words in Turkish text using syllable n-gram frequencies. International Conference on Pattern Recognition and Machine Intelligence, (pp. 553-559). Springer, Berlin, Heidelberg. Retrieved from https://link.springer.com/chapter/10.1007/978-3-540-77046-6_68

Badan Pengembangan dan Pembinaan Bahasa, K. (2018). Kamus Besar Bahasa Indonesia Edisi 5. [Big Indonesian Dictionary 5th Edition].

Brill, E., & Moore, R. (2000). An Improved Error Model for Noisy Channel Spelling Correction. Proceedings of the (pp. 286–293). Hong Kong: Association for Computational Linguistics. doi:https://doi.org/10.3115/1075218.1075255

Christanti, V., & Naga, D. (2018). Fast and accurate spelling correction using trie and Damerau-levenshtein distance bigram. Telkomnika, 16(2), 827-833. doi:10.12928/TELKOMNIKA.v16i2.6890

Deorowicz, S., & Ciura, M. (2005). Correcting spelling errors by modelling their causes. International journal of applied mathematics and computer science, 15(2), 275-285. Retrieved from http://zbc.uz.zgora.pl/Content/330/HTML/Vol15No2-113.pdf

Elghannam, F. (2021). Text representation and classification based on bi-gram alphabet. Journal of King Saud University - Computer and Information Sciences, 33(2), 235-242. doi:https://doi.org/10.1016/j.jksuci.2019.01.005

Erwina, E. (2012). Kajian sebutan baku bahasa Indonesia. Singapore International Press.

Hládek, D., Staš, J., & Pleva, M. (2020). Survey of Automatic Spelling Correction. Electronics, 9(10), 1670. doi:https://doi.org/10.3390/electronics9101670

Kamayani, M., Reinanda, R., Simbolon, S., Soleh, M., & Purwarianti, A. (2011). Application of document spelling checker for Bahasa Indonesia. 2011 International Conference on Advanced Computer Science and Information Systems (pp. 249-252). IEEE. Retrieved from https://www.researchgate.net/profile/Mia-Kamayani/publication/254048228_Application_of_document_spelling_checker_for_Bahasa_Indonesia/links/590e7617a6fdccad7b10dff1/Application-of-document-spelling-checker-for-Bahasa-Indonesia.pdf

Martin, S., Liermann, J., & Ney, H. (1998). Algorithms for bigram and trigram word clustering. Speech communication, 24(1), 19-37. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.2354&rep=rep1&type=pdf

Ningrum, V. (2019). Penggunaan Kata Baku Dan Tidak Baku Di Kalangan Mahasiswa Universitas Pembangunan Nasional "VETERAN" YOGYAKARTA. Jurnal Skripta : Jurnal Pembelajaran Bahasa Dan Sastra Indonesia, 5(2), 22-27.

Purwantoro, D., Akbar, H., & Hidayati, A. (2019). OCR correction for Indonesian historic newspapers using word repetition, stemmer and n-gram. In Journal of Physics: Conference Series, 1193(1), 012032.

Ratnasari, C., Kusumadewi, S., & Rosita, L. (2017). A Non-Word Error Spell Checker for Patient Complaints in Bahasa Indonesia. Int. J. Inf. Technol. Comput. Sci. Open Source, 1(1), 18-21.

Rice, G., & Robinson, D. (1975). The role of bigram frequency in the perception of words and nonwords. Memory & Cognition, 3(5), 513-518. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03197523.pdf

Salton, G. (1989). Automatic Text Processing. Addison-Wesley Publishing Company.

Samanta, P., & Chaudhuri, B. (2013). A simple real-word error detection and correction using local word bigram and trigram. Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING).

Santoso, P., Yuliawati, P., Shalahuddin, R., & Wibawa, A. (2019). Damerau levenshtein distance for indonesian spelling correction. JURNAL INFORMATIKA, 13(2), 11-15. doi:10.26555/jifo.v13i2.a15698

Sirait, Z. (2021). Penggunaan Bahasa Indonesia Di Ruang Publik Yang Tidak Memenuhi Bahasa Baku. Linguistik : Jurnal Bahasa dan Sastra, 6(1), 1-9.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Supriadin. (2016). Identifikasi Penggunaan Kosakata Baku Dalam Wacana Bahasa Indonesia Pada Siswa Kelas Vii Di Smp Negeri 1 Wera Kabupaten Bima Tahun Pelajaran 2013/2014. JIME : Jurnal Ilmiah Mandala Education, 2(2), 150-161.

Tong, X., & Evans, D. (1996). A statistical approach to automatic OCR error correction in context. In Fourth workshop on very large corpora. Retrieved from https://aclanthology.org/W96-0108.pdf

Wibawa, A., Yuliawati, P., Santoso, P., Shalahuddin, R., & Wirawan, I. (2020). Damerau Levenshtain Distance dengan Metode Empiris untuk Koreksi Ejaan Bahasa Indonesia. ILKOM Jurnal Ilmiah, 12(3), 176-182.

*name of corresponding author