International Seminar of Research Month 2021
Volume 2022
http://dx.doi.org/10.11594/nstp.2022.2433

NST PROCEEDINGS

OPEN ACCESS

Conference Paper

A Rule-based Spelling Checker for Correcting Punctuation Errors in Indonesia Text using KEBI 1.0 Checker

Tresna Maulana Fahrudin*, Ilmatus Sa'diyah, Latipah, Ibnu Zahy' Atha Illah, Cagiva Chaedar Bey Lirna, Burhan Syarif Acarya

Department of Data Science, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya 60294, Indonesia

**Corresponding author:*
E-mail:
tresna.maulana.ds@upnjatim.ac.id

ABSTRACT

Punctuation errors in Indonesian often occur in Indonesian scientific writing. This error was caused intentionally and unintentionally by the author. The factor of intentional is because the author does not have time to make independent improvements after writing, while unintentional occurs because of the author's ignorance of the appropriate form of punctuation. Therefore, the punctuation error detection system in Indonesian writing was developed to overcome them. The name of the application is KEBI 1.0 Checker which was a web-based application and provided three features, namely punctuation detection, typo detection, and standard words detection. The research proposed to evaluate the performance of application based on punctuation errors using a rule-based spelling checker that can be detected were in the form of comma punctuation, hyphens in repeat words, and writing of pre-, post-, and inter- particles, as well as writing prefixes with additional English verbs. After the application was tested on ten scientific papers, KEBI 1.0 Checker showed accuracy for correcting punctuation errors reached an accuracy of 76.67%, precision of 100%, and recall of 76.67%.

Keywords: Rule-based, spelling checker, correcting punctuation errors, Indonesian Text, KEBI 1.0 Checker

Introduction

Writing is one of the language skills that need to be learned by researchers, students, and people who want to express their ideas, opinions, and thoughts so that others can read them. Writing an article is part of a linguistic competency that demands more attention. Punctuation errors are errors that often occur in writing. Several researchers search for punctuation errors using the Boyer-Moore stream search algorithm by finding punctuation marks in a collection of words, then checking the conditions for using punctuation (Anggraini et al., 2016). The correct use of punctuation marks can prevent the reader from misunderstanding the meaning conveyed by the author (Yakhontova, 2020). For example, some authors do not use a comma in the last item to separate it from other items, "I like apples, bananas and grapes", the correct one is "I like apples, bananas, and grapes". Sometimes punctuation errors often go unnoticed and are not handled by the researchers themselves, peer reviewers, journal editors, and language instructors. Punctuation marks show the relationships between textual passages and help to produce easy-to-understand and readable sentences that are the "building blocks" of any consistent research text (Yakhontova, 2020). A study conducted by Laili (2018), showed that the highest type of error was negligence. From the results of the essay test of 11 punctuation marks containing 10 items, the most dominant types of errors made by students were exclamation marks, dots, and commas. Many errors in the

How to cite:
Fahrudin et al. (2022). A rule-based spelling checker for correcting punctuation errors in indonesia text using KEBI 1.0 Checker . *International Seminar of Research Month 2021*. NST Proceedings. pages 207-214. doi: 10.11594/nstp.2022.2433

use of punctuation also occur in high school students. Research conducted by Husada et al. (2018) shows that from several punctuation errors that occur, the use of "comma" punctuation errors is the largest use punctuation errors, which is 36%.

According to Apriliana (2020), the correct use of language rules following the General Guidelines for Indonesian Spelling (PUEBI) is one of the important factors in the writing of academic and scientific papers. The selection of words in writing must be based on writing rules such as meaning rules, syntax rules, constitutional rules, and social relations rules that support writing to be more valuable, structured, and easy to understand by the community. The use of the spelling Indonesian in public writing is still often found, for example, in papers written by students. The improper use of Indonesian spelling often happens because some people in the writing process are not guided by the rules of the correct language. In addition, the spelling error in Indonesian scientific papers can also be caused because the author being less accustomed to using spelling, lacking understanding of spelling, and the author's inattention to writing.

This study continues from previous research by Fahrudin et al. (2021) that discusses the ability of the error detection and correction of Indonesian non-standard words and typos in an application called KEBI 1.0 Checker. Therefore, this study proposes adding a feature for a spelling checker for correcting punctuation errors in Indonesian text using a rule-based. The addition of this feature will enhance the module of KEBI 1.0 Checker in detecting and correcting spelling errors.

Material and Methods

Spelling errors often occur when writing Indonesian text. These errors can be in the form of a lack of knowledge of the author of the Indonesian language, the author's unintentional negligence, and several other things that allow errors in writing good and correct Indonesian according to PUEBI and KBBI (Badan Pengembangan dan Pembinaan Bahasa, 2016) (Tim Penyusun Kamus Pusat Bahasa, 2008). This study will focus on discussing the Indonesian spelling error detection feature based on punctuation marks that were added to KEBI 1.0 Checker modules. The detection of punctuation spelling errors is related to a) contradictory conjunctions, b) connecting expressions between sentences, c) complete repetition of words, d) bound forms that become the object of discussion in the form of particles, and e) the combination of Indonesian prefixes elements with foreign language verb aspects. Figure 1 shows the punctuation error detection process using a rule-based spelling checker.

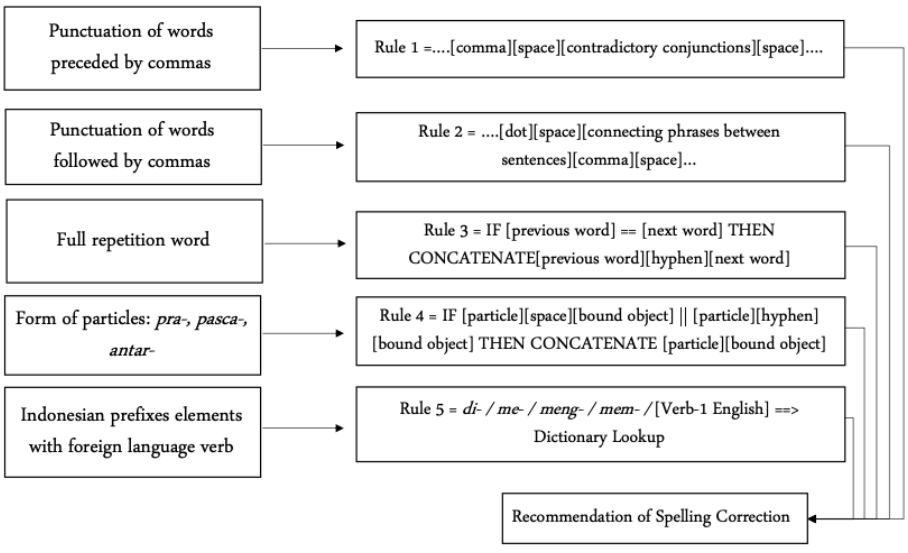

Figure 1. Punctuation error detection process using the rule-based spelling checker

Contradictory conjunctions

This relates to use of the words *"walaupun", "tetapi", "melainkan", "sedangkan",* and etc. KEBI 1.0 Checker will detect punctuation spelling errors using rule-based that have the arrangement "...[comma][space][contradictory conjunction][space]...". This means that whenever a punctuation error is found in the contradictory conjunction, the sentence does not follow the rules and a correction suggestion will appear. On other hand, if there are no punctuation errors in the contradictory conjunction, the sentence has complied with the rules. For example the following sentences, "*Raina memiliki kemampuan berhitung yang hebat* **tetapi** *ia tidak pandai menggunakan bahasa asing"*. If we follow the PUEBI rules, we need to put a comma (,) before the word ***tetapi***. The complete sentence correctly is "*Raina memiliki kemampuan berhitung yang hebat,* ***tetapi*** *ia tidak pandai menggunakan bahasa asing".*

Connecting phrases between sentences

This relates to use of the words "*Oleh karena itu*", *"Jadi", "Meskipun demikian", "Akan tetapi", "Jika demikian", "Namun", "Selain itu", "Sehubungan dengan itu", "Walaupun demikian",* and some examples of connecting phrases between other sentences. KEBI 1.0 Checker will detect spelling punctuation errors using rule-based that have the structure "...[dot][space][connective expression between sentences][comma][space]...". This means that whenever a punctuation error is found in the connecting phrases between sentences, the sentence does not follow the rules and a correction suggestion will appear. On other hand, if there are no punctuation errors, the sentence has complied with the rules. For example the following sentences, "*Toni telah menempuh program magister di salah satu Perguruan Tinggi****, namun*** *ia belum mendapatkan pekerjaan"*. If we follow PUEBI rules, the word ***namun*** should be preceded by a dot (.), not a comma (,). In addition, the first letter of the word ***namun*** is changed to a capital letter. The complete sentence correctly is "*Toni telah menempuh program magister di salah satu Perguruan Tinggi****. Namun,*** *ia belum mendapatkan pekerjaan".*

Full repetition word

This relates to use of the words *"anak-anak", "sama-sama", "jalan-jalan"*, and some other examples of full repetition words. KEBI 1.0 Checker will detect spelling errors of punctuation using rule-based which have the structure "IF [previous word] == [next word] THEN CONCATENATE [previous word][hyphen][next word]". If there are similarities between the previous word and next word, between the two words will be given a "hyphen" in the middle. On other hand, if no similarities are found, it is not a full repetition word. For example the following sentences, "*Ayah mengajak adi dan dita jalan jalan ke pantai"*. If we follow PUEBI rules, the word ***jalan jalan*** should be affixed with a hyphen (-) in the middle. The complete sentence correctly is "*Ayah mengajak adi dan dita* ***jalan-jalan*** *ke pantai".*

The bound form that becomes the object of discussion is in the form of particles

This relates to the use of particle *"pra-", "pasca-",* dan *"antar-"*. KEBI 1.0 Checker will detect spelling errors of punctuation using rule-based which have the structure "IF [particle][space][bound object] || [particle][hyphen][bound object] THEN CONCATENATE [particle][bound object]. If there is a word consisting of particles containing "space" with the bound object or other conditions, the particle containing "hyphen" with the bound object, it will be concatenated without spaces or hyphen. For example the following sentences, "*Dia telah menyelesaikan studi di tingkat* ***pasca sarjana*** *dalam bidang ilmu komputer"*. If we follow the PUEBI rules, the word ***pasca sarjana*** must be written without spaces. The complete sentence correctly is "*Dia telah menyelesaikan studi di tingkat* ***pascasarjana*** *dalam bidang ilmu komputer".*

The combination of Indonesian prefixes elements with foreign language verb aspects

This relates to common words and non-standard words that store data on Indonesian prefixes and English verbs such as "back up", "accept", "follow", and etc. KEBI 1.0 Checker will detect punctuation spelling errors using rule-based which have the order "*di- / me- / meng- / +* [V1 English verb]" which using dictionary lookup. It is combined with checking the availability of the word in the dictionary first. For example the following sentences, "*Dia telah* ***membackup*** *semua file ke dalam komputer miliknya*". If we follow PUEBI rules, the word ***membackup*** should be written "***mem-backup***". The complete sentence correctly is "*Dia telah* ***mem-backup*** *semua file ke dalam komputer miliknya*".

Results and Discussion

In the process of writing a scientific paper in Indonesian text, punctuation is an important part so that the writing can be read and the meaning of the sentences is conveyed. Punctuation marks in Indonesian are divided into several types, namely commas, dots, semicolons, colons, double quotes, single quotes, brackets, square brackets, hyphens, dashes, and apostrophes. However, in KEBI application, the punctuation marks that can be detected were comma punctuation in the use of conjunctions between sentences, intra-sentence conjunctions, and conjunctions between paragraphs as well as hyphens in repetition words, bound words, such as *pra-* and *pasca-,* and writing combined Indonesian prefixes with verbs in a foreign language.

KEBI 1.0 Checker provided several features that can detect and correct non-standard Indonesian words, typos, and punctuation in a scientific paper. However, this study only focuses on detecting spelling errors based on punctuation. To detect punctuation errors, users can access the page following https://kebi.id. Through this page, users can improve their Indonesian scientific papers before they are published or read by the general public. The documents used in this experiment were obtained from paper assignments written by ten students in Indonesian text. Figure 2 shows the punctuation detection page of KEBI 1.0 Checker web-based application.

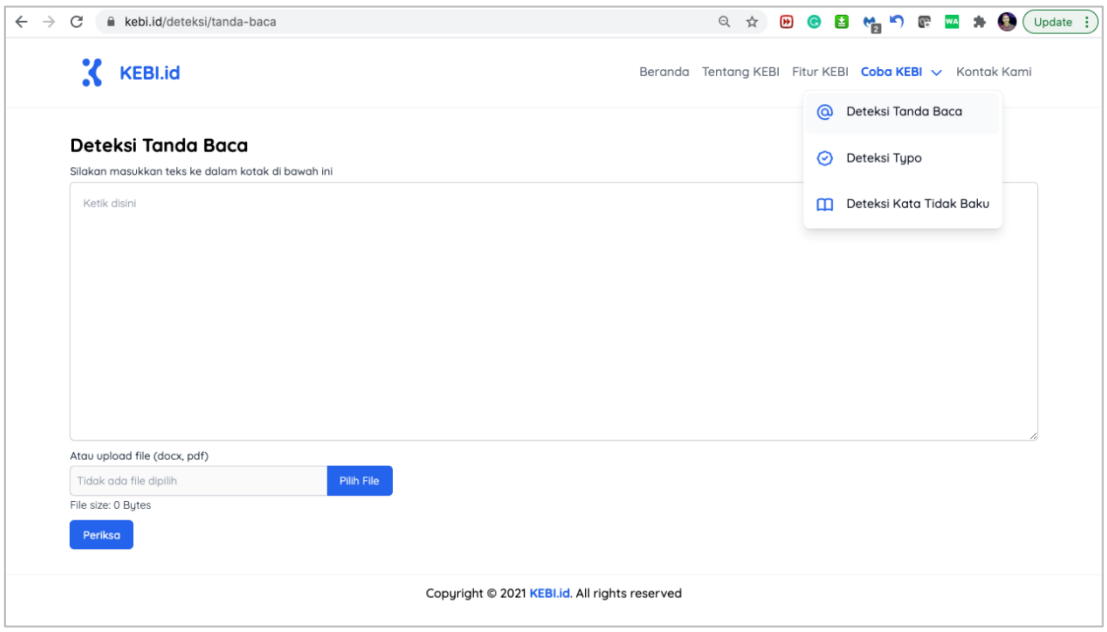

Figure 2. Punctuation detection page of KEBI 1.0 Checker web-based application

To determine the performance of KEBI application, both in terms of accuracy, a test scenario was needed through the accuracy of detecting and correcting Indonesian spelling errors in scientific article documents. After submitting ten manuscripts of scientific papers in the Indonesian

language, KEBI has given pretty good detection performance. Table 1 shows the confusion matrix of spelling correction based on punctuation errors in KEBI.

Table 1. Confusion matrix of spelling correction based on punctuation errors in KEBI application

No	Punctuation Error	Correction by KEBI	Correction Result		TP	FP	FN	TN
			Validation	System				
1	*satunya di bandara**, akan tetapi** bahasa daerah juga ...*	*...satunya di bandara**. Akan tetapi,** bahasa daerah jug...*	√	√	1			
2	*...g hanya menggunakan **bahasa bahasa** daerah atau bahasa ...*	*Di Indonesia penggunaan bahasa pada ruang publi...*	√	√	1			
3	*...i eksplorasi adalah **penggunaan penggunaan** bahasa untuk menjel...*	*Bahasa Indonesia memiliki kedudukan sanga...*	√	√	1			
4	*...Data yang diperoleh **selanjutnya** dianalisis secara t...*	*...Data yang diperoleh**. Selanjutnya,** dianalisis secara...*	×	√			1	
5	*...n Bahasa Indonesia. **Namun** petugas Bandar Udar...*	*...n Bahasa Indonesia. **Namun,** petugas Bandar Ud...*	√	√	1			
6	*...ua bahasa tersebut. **Padahal** penggunaan bahasa d...*	*...ua bahasa tersebut**, padahal** penggunaan bahasa d...*	√	√	1			
7	*...i maksud yang sama**, sebenarnya** kedua pernyataan te...*	*...i maksud yang sama**. Sebenarnya,** kedua pernyataan ...*	√	√	1			
8	*...Bandara Kualanamu. **Oleh sebab itu** sepatutnya diganti ...*	*...Bandara Kualanamu. **. Oleh sebab itu,** sepatutnya digant...*	√	√	1			
9	*...ik kumpul bantuan”**, setelah itu** baru ditulis kalima...*	*...ik kumpul bantuan”**. Setelah itu,** baru ditulis kali...*	√	√	1			
10	*...onesia dengan tepat **namun** untuk skala interna...*	*...onesia dengan tepat**. Namun,** untuk skala inter...*	√	√	1			
11	*...amanan penerbangan. **Selain itu** dapat pula disertai...*	*...amanan penerbangan. **Selain itu,** dapat pula disert...*	√	√	1			
12	*...andas, bongkar muat **barang barang** dan naik turun penu...*	*Bandara merupakan kawasan atau l...*	√	√	1			
13	*...t komunikasi lancar **antar satu** dengan yang lainnya...*	*...t komunikasi lancar **antar-satu** dengan yang lainnya...*	√	√	1			
14	*...dan bahasa Inggris. **sedangkan** sisanya menggunakan...*	*...dan bahasa Inggris**, sedangkan** sisanya menggunakan...*	√	√	1			
15	*...an aman dan nyaman. **Untuk itu** maka dilakukan sebu...*	*...an aman dan nyaman. **Untuk itu,** maka dilakukan se...*	√	√	1			
	To be continued…							
16	*...amatan penerbangan. **Pada dasarnya** barang berbahaya da...*	*...amatan penerbangan. **Pada dasarnya,** barang berbahaya ...*	√	√	1			
	To be continued..							
17	*...ang sudah saya baca **sebelumnya**. makalah ini menggu...*	*...ang sudah saya baca**. Sebelumnya,** makalah ini meng...*	x	√			1	

18	*...terjadinya tabrakan **antar pesawat** saat di udara, dan ...*	*...terjadinya tabrakan **antarpesawat** saat di udara, dan ...*	√	√	1			
19	*...isien Ditinjau Dari Undang Undang Nomor 1 Tahun 2009 ...*	*Martam, N. K. (2018). Tanggung Ja...*	x	√			1	
20	*...atawan mancanegara. **Oleh karena itu** bahasa asing diperl...*	*...atawan mancanegara. **Oleh karena itu,** bahasa asing dipe...*	√	√	1			
21	*... wisatawan domestik **melainkan** juga wisatawan manc...*	*... wisatawan domestik**, melainkan** juga wisatawan man...*	√	√	1			
22	*... yaitu perkantoran**, jadi** bandar udara termas...*	*... yaitu perkantoran**. Jadi,** bandar udara term...*	√	√	1			
23	*...an Bahasa Indonesia **setelah itu** dilanjut dengan pen...*	*...an Bahasa Indonesia**. Setelah itu,** dilanjut dengan p...*	√	√	1			
24	*...pang dengan banyak. **Sedangkan** bis merupakan kata ...*	*...pang dengan banyak**, sedangkan** bis merupakan kata ...*	√	√	1			
25	*...budayaan Indonesia**, oleh karena itu** secara tidak langsu...*	*...budayaan Indonesia**. Oleh karena itu,** secara tidak lang...*	√	√	1			
26	*...ga negara Indonesia **sebaiknya** menggunakan bahasa ...*	*...ga negara Indonesia**. Sebaiknya,** menggunakan bahas...*	x	√			1	
27	*...da abad ke-18. Hari **jadi** kota ini ditetapkan...*	*...da abad ke-18. Hari**. Jadi,** kota ini ditetapk...*	x	√			1	
28	*... Kota Pekanbaru dan **sebelumnya** bernama Bandara Sim...*	*... Kota Pekanbaru dan**. Sebelumnya,** bernama Bandara S...*	x	√			1	
29	*...gedukasi kita bahwa **sebenarnya** penggunaan bahasa y...*	*...alam bahasa melayu**. Sebenarnya,** akrab disebut kap...*	x	√			1	
30	*...an bahasa Indonesia **tetapi** tentang fungsi dari...*	*...an bahasa Indonesia**, tetapi** tentang fungsi dar...*	√	√	1			
		Total			23	0	7	0

Information
√: A spelling error was detected and corrected
× : Error detected a spelling error and does not need to be corrected

Table 1 shows KEBI application detected punctuation errors for each line of paragraphs, errors that were detected such as the use of commas in conjunctions, such *akan tetapi, namun, jadi, sebenarnya, padahal, selanjutnya, setelah itu, oleh sebab itu, selain itu, sedangkan, untuk itu, pada dasarnya, sebelumnya, oleh karena itu, melainkan, jadi,* and *tetapi*. In addition to detecting errors, KEBI also provided corrections for these errors. For example, the sentence number 14 in Table 1, "*...dan bahasa Inggris**. sedangkan** sisanya menggunakan...*", KEBI succeeded to detect errors a dot before the word *sedangkan*. Furthermore, KEBI provided corrections in the form of "*...dan bahasa Inggris**, sedangkan** sisanya menggunakan....*" The period was converted into a comma by KEBI. This sentences describe that KEBI showed good performance.

The detection results were calculated using confusion matrix. The confusion matrix determine the accuracy of a classification, including True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) (Prasetyo, 2012). For example, the sentence number 1 in Table 1 "*satunya di bandara**, akan tetapi** bahasa daerah juga ...".* According to the validator, the

sentence contained an error. KEBI application also detected that the sentence also contained an error so that the conclusion of the detection result for the sentence was true positive. It means that both the detection of spelling errors by the validator and by the system showed that there was indeed a punctuation error in the sentence. KEBI application provided suggestions for correcting the misspelling of the sentence to be *Akan tetapi,.*

However, KEBI still makes errors in detection of some conjunctions. The majority are conjunctions *sebelumnya, selanjutnya,* and *jadi*. For example, the sentence number 4 in Table 1 "*...Data yang* ***diperoleh selanjutnya*** *dianalisis secara...*," the validator does not detected any punctuation errors, but KEBI detected it as an error. KEBI provided corrections to the sentence "*...Data yang* ***diperoleh. Selanjutnya,*** *dianalisis secara....*". The corrections made by KEBI was the wrong structure sentence. According to the validator, the error in the sentence lied in writing the conjunction *selanjutnya* which should not be used.

This also happens with the conjunction *jadi*. In the sentence number 27 in Table 1 "*...da abad ke-18.* ***Hari jadi*** *kota ini ditetapkan....*", the validator also found no errors. KEBI system detected and corrected it to the sentence "*...da abad ke-18.* ***Hari. Jadi****, kota ini ditetapkan...".* The detection and correction made by KEBI was not correct, it is even changing the correct sentence into the wrong one. This finding was also an evaluation for the KEBI correction system to be able to detect comma errors in conjunctions more accurately.

When correcting hyphens, KEBI has provided corrections to spelling errors that have been made. In the sentence number 3 in Table 1 "*...i eksplorasi adalah* ***penggunaan penggunaan*** *bahasa untuk menjel...,".* KEBI was able to detect punctuation errors in words and correct the error by adding a hyphen between the two words ***penggunaan-penggunaan***.

On another aspect, KEBI actually detects errors in writing bound forms without hyphens, namely *antar pesawat* in the sentences number 18 in Table 1 "*...terjadinya tabrakan* ***antar pesawat*** *saat di udara, dan ..."* and correcting them to the sentence "*...terjadinya tabrakan* ***antarpesawat*** *saat di udara, dan ....*". KEBI has proven the capability can detecting and detecting other types of writing errors.

Based on evaluation using the confusion matrix, KEBI 1.0 Checker has provided good performance for detecting punctuation errors. The result was the number of TP = 23, FP = 0, FN = 7, and TN = 0. Table 2 showed the percentage accuracy of correcting punctuation errors in the KEBI application with an accuracy of 76.67%, precision of 100%, and recall of 76.67%.

Table 2. Percentage of correction accuracy of punctuation errors in KEBI Application

No.	Performance
1.	$Precision = \frac{TP}{TP+FP} = \frac{23}{23+0} = 100\ \%$
2.	$Recall = \frac{TP}{TP+FN} = \frac{23}{23+7} = 76.67\%$
3.	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{23+0}{23+0+0+7} = \frac{23}{30} = 76.67\%$

These results indicated that KEBI has good performance to be widely used as an Indonesian language error detection and correction in the punctuation aspect.

Conclusion

A punctuation mark is an important thing that must be considered in making an Indonesian scientific paper. This will affect word morphology and meaning in the sentences. KEBI 1.0 Checker as an application for correcting Indonesian spelling errors provides features for punctuation marks including contradictory conjunctions, connecting expressions between sentences,

complete repetition of words, bound forms that become the object of discussion in the form of particles, and the combination of Indonesian prefixes elements with foreign language verb aspects. The evaluation performance shows KEBI has good performance reaching an accuracy of 76.67%, precision of 100%, and recall of 76.67%.

Acknowledgment

The author would like to thank the Ministry of Education, Culture, Research, and Technology, Universitas Pembangunan Nasional "Veteran" Jawa Timur which has funded this research based on the Assignment Agreement for the Implementation of the Batch I Internal Research Program for the Basic Research Scheme (RISDA) in 2021, Number: SPP / 12 /UN.63.8/LT/IV/2021.

References

Anggraini, R. N. E., Zinni, M. A., & Rochimah, S. (2016). Kakas bantu pendeteksi kesalahan tanda baca pada karya tulis ilmiah. *Jurnal Ilmiah Teknologi Informasi, 14*(1), 117-125.

Apriliana, R. F., Firdaus, A., & Suparman, F. (2020). Kesalahan penulisan kata dan tanda baca pada online news. *Jurnal Pendidikan Bahasa dan Sastra Indonesia, 5*(1), 13-19. Doi: https://doi.org/10.30743/bahastra.v5i1.2996

Badan Pengembangan dan Pembinaan Bahasa. (2016). *Pedoman umum ejaan Bahasa Indonesia*. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Fahrudin, T. M., Sa'diyah, I., Latipah, L., Atha Illah, I. Z., Bey Lirna, C. C., & Acarya, B. S. (2021). KEBI 1.0: Indonesian spelling error detection system for scientific papers using dictionary lookup and peter norvig spelling corrector. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 12*(2), 78-90. https://doi.org/10.24843/LKJITI.2021.v12.i02.p02

Husada, S., Hidayati, H., & Humaira, H. (2018). An error analysis of using punctuation made by students in descriptive text at the second years students of SMPN 3 Wera in academic year 2017/2018. Jurnal Pendidikan Berkarakter, 1(1), 23-26. Doi:10.31764/pendekar.v1i1.253

Laili, F. N. (2018). *An errors analysis on the use of punctuation marks in students' writing (A study at second semester students of English Department of Universitas Muhammadiyah Surakarta).* Surakarta.

Prasetyo, E. (2012). *Data mining: konsep dan aplikasi menggunakan matlab*. Yogyakarta: Andi Offset.

Tim Penyusun Kamus Pusat Bahasa. (2008). Kamus Bahasa Indonesia. Jakarta: Pusat Bahasa Departemen Pendidikan Nasional.

Yakhontova, T. (2020). Punctuation mistakes in the english writing of non-anglophone researchers. Journal of Korean Medical Sains, 35(37), 1-5. Doi:10.3346/jkms.2020.35.e299