Age Group Based Document Classification in Bahasa Indonesia

by Budi Irmawati

Submission date: 11-Apr-2023 08:11AM (UTC-0500)
Submission ID: 2061525044
File name: icadeisPaper14.pdf (222.07K)
Word count: 4461
Character count: 23315

2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)

978-1-7281-8272-8

Age Group Based Document Classification in Bahasa Indonesia

M. Iqbal D. Putra
Department of Informatics Engineering
University of Mataram
Mataram, Indonesia 83125
iqbaldwinfor@gmail.com

*Budi Irmawati
Department of Informatics Engineering
University of Mataram
Mataram, Indonesia 83125
budi-i@unram.ac.id

Wirarama Wedashwara
Department of Informatics Engineering
University of Mataram
Mataram, Indonesia 83125
wirarama@unram.ac.id

Dita Pramesti
Department of Information System
Telkom University
Bandung, Indonesia
ditapramesti@telkomuniversity.ac.id

Siti Oryza Khairunnisa
Faculty of System Design
Tokyo Metropolitan University
Tokyo, Japan
siti-oryza-khairunnisa@ed.tmu.co.jp

***Abstract*— Internet provides articles that may be categorized to various target readers based on genders, ages, hobbies, etc. To make sure that readers consume a proper article based on their age group, methods and training data were proposed and collected to classify the articles. This paper reported a document classification based on age groups using a binary classification method for Indonesian documents. The document classification used the term frequency and inverse document frequency (TF-IDF) features run on the *Multinomial Naïve Bayes Classifier*. The dataset was crowdsourced from three different sites: bobo.grid.id, hai.grid.id, and www.detik.com for three age group readers such as elementary school children, teenagers, and adults. The experimental results obtained 0.9406, 0.9341, and 0.9374 of precision, recall, and F-score respectively. This experiment also reported that for the datasets that were not stemmed performed better than those that were stemmed. It shows that the stemming process, which usually be done in the document classification, throws some information in the Indonesian texts. However, because this behavior was not happen on nouns, our future work is to elaborate further on the role of affixations in the lower age group documents.**

Keywords – Age groups, Document classification, Multinomial Naïve Bayes, TF-IDF

I. Introduction

Internet provides various articles that can be accessed easily. The articles are also vary in category and may also be categorized to different target readers. Among the category, the articles are intended to be read by a specific age group such as children, teenagers, or adults.

Age-based document classification aims to identify the age of the authors or to measure whether a document is appropriate to a specific age readers. The former is to find out author profiling of a document in the internet, from its linguistic characteristics such as word choices and writing style [1]. Text-based age prediction is important to avoid false information, such as detecting a criminal activities of an elder people that pretend to be younger people to get closer to children [2], [3]. Age, gender, and demographics are personal data that are not always available in an author profile [4]. The later is to make sure that a document does not contain harmful content if it is read by children. We focused on the later case by evaluating documents written for children and teenagers. The main goal is to verify whether a document is appropriate for children. As the comparison to the two ages, we also evaluated adults dataset to see if the vocabulary used in the three ages were different.

Even though the main goal is to verify whether a document is appropriate for children, other aspect to be evaluated later was whether the vocabulary in the lowest age group may be easier to be learned by language learners. To simplify the problem of language resources, the datasets were collected from online resources of those three ages.

In simplification research, a model trained on dataset of simple sentences was used to built an application to help learners understand structures and word choices [5]. This result showed that a simple vocabulary was useful for language learners to build sentences in their early learning steps. Therefore, we assume that learners in the beginner level are able to develop sentences using simple vocabulary such as ones used by children. We will work further to find simplification features that may be useful to build a tool for language learners.

II. Previous Works

In a cybersecurity monitoring tool to detect sexual predators in social media, a 380,000 Dutch chat posts collected from social media was classified by [3]. They used binary classifier to detect the real age and gender of users where those information are not available in their profile. For the classification of documents written to specific age group, a simple linear regression on shallow text features proposed by [6]. It was evaluated on genres of blogs, telephone conversation, and online forum posts and obtained correlation up to 0.74.

* Corresponding author

Classification is also a popular technique in language learning research. Learners need to be in at least medium level to write sentences as good as ones written by native speakers. Error correction research in Bahasa Indonesia are still in the beginning stage due to lacks of language resources. Recent works utilized machine translation technique to correct errors in sentences written by learners [7]. Therefore, a large-scaled artificial sentences was developed from sentences written by native speakers using classification method. The correct words in the original sentences were replaced by the erroneous words adopted from sentences written by language learners [8]. The same technique for preposition error correction is done in [9], [10]. The data were used to build model to correct preposition mistakes made by the language learners.

People also assumed that texts written by or for children in their first language (L1) are good reference for beginner learners to develop simple sentences [11]. Previous research [5] used a corpus of Japanese children taken from Tatoeba Corpus[1] and Hiragana Times Corpus[2] to develop the training data while the testing data were taken from the 2011 data of Lang-8 website. Basically, it provides a verb-noun collocation suggestion for learners of Japanese. This system provides examples related to the noun or to the verb of a verb-noun collocation that want to be written by a learner. This research proofed that documents written for young age group is useful to build a training model for an assisted tool.

III. Methods

Figure 1 explains the pipeline of the classification experiments such as text preprocessing and feature extraction. The preprocessing was separated to tokenizing, case folding, stopwords filtering, and stemming. In tokenization, one token represents single meaning in the vocabulary. As Indonesian has complex morphological features, it used morphology analyzer [12] to make sure that the tokens were cleaned from any clitics. In the case folding, the tokens were transferred to low case characters to avoid redundancy. Then, we filtered the stopwords. Stopwords are words that do not relate to any class or category so their existence does not contribute any information. In stemming, the tokens were returned to their basic word, without affixes. Stemming is a common step used in many classification documents. The purpose of stemming is to get smaller vocabulary and to avoid tokens with the same root word are considered as different tokens. However, we also conducted experiments that used non-stemmed dataset as comparison. This experiment used stemmer algorithm proposed by Nazief and Adriani in Sastrawi library [12]. This stemmer had also been used by [13] to compare the performance of Nazief and Adriani and Porter stemmer on Winnowing algorithm to measure plagiarism score.

A. Datasets

The datasets were crowdsourced from bobo.grid.id, hai.grid.id, and detik.com. The first two are magazines for children and teenagers, while the latest is a newspaper. The number of vocabulary of each dataset is listed in Table I. From each category, we collected 100 articles so we have totally 300 articles. Table II shows examples from each category. This table shows that sentence in adults class is more formal than the other two classes. Adults age magazines like Femina and Tempo also fit with adults dataset but we assumed that the vocabulary of these two magazines are almost similar to the newspaper. However, we need to confirm our assumption with an experiment.

[1]http://tatoeba.org/eng/
[2]http://www.hiraganatimes.com

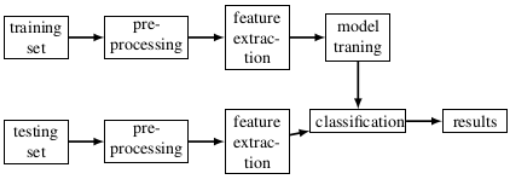

Fig. 1. The clasification pipeline.

TABLE I
The Statistics of the Datasets

Datasets	Vocabulary Size
Children	4,994
Teenagers	5,713
Adults	5,665

TABLE II
Sentence examples from each class

Classes	Sentences
Children	Yuk teman-teman kita belajar membuat mainan. Mainan kali ini berasal dari barang yang sudah tidak terpakai. Begini lho caranya!
Teenagers	Ini dia beberapa alasan yang buat cowok-cowok nggak bisa berhenti main game. Jangan sampai lo ...tan juga guys.
Adults	Kasus pembunuhan terjadi pada sepasang kekasih di Jakarta Utara. "Korban dibunuh dengan cara dipukuli pada bagian kepala dan leher" kata seorang petugas kepolisian yang diwawancarai.

B. Feature Extraction

After the preprocessing steps, we extracted features using the sophisticated TF-IDF (*Term Frequency-Inverse Document Frequency*) score to weight a word related to a category. This method combines TF, the number of a word in a document, and IDF. IDF assigns a higher weight to words that only occurs in a few categories [12], [14], [15]. Equation (1) represents the TF-IDF formulas.

$$\text{tf-idf}=tf_{i,j} \times \text{idf}_j \quad (1)$$
$$\text{idf}_i=\log(\frac{N}{k})$$

where term frequency $tf_{i,j}$ is the frequency of `term` i in the `document` j while N is the number of documents and k is the number of documents containing `term` i.

C. Classification

Feature extraction resulted TF-IDF scores calculated from the tokens in the dataset as the training model. The classifications were done in 10-fold cross-validation [16] using the *Multinomial Naïve Bayes* classifier and were repeated 30 times. On each iteration, the dataset was repositioned randomly to avoid the same combination of the training and the testing data. The *Multinomial Naïve Bayes* classifier was also used in [1] and performed well.

We did five scenarios in this experiment. The main experiment used all tokens and was named as (1) All Tokens (AT) scenario. Other four scenarios were (2) Only Verbs (OV), (3) Only Nouns (ON), (4) Verbs and Nouns (VN), and (5) Verbs, Nouns, and Adjectives (VNA). For the five scenarios, we differentiated the datasets for STEMMED and NON-STEMMED ones. It means that we have ten different scenarios.

Then, to verify the results, we also prepared nine articles out of ones used in the test set of the five scenarios. From these nine articles, each class consists of three articles. In this verification, we applied the best model and compared the class predicted by the system with the class assigned by a linguist of Indonesian.

D. Evaluation Metric

The classification results were evaluated in precision, recall, and F-score. **Precision** is the percentage of prediction given by the system that are correct, **recall** is the percentage of item in a class that are correctly identified by the system, and **F-score** combines these two metric into a single metric. Equations (2) to (4) shows the formulas of the precision, recall, and F-score.

$$\text{precision (P)} = \frac{\text{\# correct items predicted by the system}}{\text{\# items predicted by the system}} \quad (2)$$

$$\text{recall (R)} = \frac{\text{\# correct items predicted by the system}}{\text{\# items}} \quad (3)$$

$$\text{F-score} = \frac{2PR}{P+R} \quad (4)$$

IV. Result and Discussion

As explained in Subsection III-C, we did experiment with five scenarios (AT, OV, ON, VN, and VNA). Other than the classification, we also evaluated the vocabulary based on the part-of-speech[3], frequent words, influence of the stemming, and thresholding. We also reported the comparison between the prediction given by the system and the class assigned by human.

[3]**Part-of-speech** is a word category based on its grammatical property in a sentence [14].

TABLE III
Classification Results of All Scenarios.

Datasets	Non-Stemmed			Stemmed		
	P	*R*	*F*	*P*	*R*	*F*
AT	0.9406	0.9341	**0.9374**	0.9360	0.9298	0.9229
OV	0.8629	0.8253	**0.8437**	0.8220	0.7866	0.8039
ON	0.9132	**0.9080**	0.9106	**0.9194**	0.9061	**0.9127**
VN	0.9167	0.9038	**0.9102**	0.9127	0.8961	0.9043
VNA	0.9143	0.9005	**0.9073**	0.9088	0.8924	0.9005

A. Classification Trained on All Scenarios

Table III lists the classification results of the system trained on NON-STEMMED and STEMMED datasets for all scenarios. The table shows that the NON-STEMMED dataset performed better for almost all scenarios (**written in bold**), except for ON scenario. The table also shows that the classification results trained on AT dataset performs the best. We predict that the success of the AT scenario is because the dataset vocabularies in other scenarios are much smaller than the vocabulary used in AT scenario.

Table III also shows that ON has quite different behaviour (written in red), because its precision for the system trained on STEMMED dataset is better than that trained on the NON-STEMMED dataset. However, its recall follows the general phenomena. This experiment proofed that the affixes did not help nouns much even though the NON-STEMMED datasets work well in other scenarios.

The results of VN scenario is also interesting because it outperforms the results of OV and VNA scenarios. Therefore, it would be interesting to explore further what are the more informative features in verbs and nouns and whether the features are related to the affixes.

B. Evaluation Based on POS

From the datasets, we listed dominant tokens from each class. Table IV lists the first five highest frequency tokens from all POSs. The table shows that children and teenager datasets shared three similar frequent tokens (*ada*, *juga*, and *lo*). We cannot put these words into the stop words because they are rarely appear in adults dataset. The tokens *ada* and *juga* also have higher TF-IDF score compare to other tokens. Even though *tahun* is common in all age group texts but we did not consider it as a stop words because its frequency in teenagers dataset is lower. Therefore, we conclude that children and teenagers datasets have quite similar vocabulary than the adults dataset.

We also listed dominant tokens based on part-of-speeches. Tables V, VI, and VII list the first five highest frequency tokens based on verbs, nouns, and adjectives. From Table V, we found that in the teenagers class, there are some verbs that were in their root word (**written in bold**). In Indonesian, verbs formed in root word are usually used in casual style or in imperative sentences. Therefore, it can be inferred that the teenagers documents may contain more casual words.

TABLE IV
THE FIRST FIVE HIGHEST FREQUENCY WORDS IN EACH CLASS.

Classes	Tokens	Frequency	TF-IDF
children	teman	826	-5.080
	ada	409	-9.005
	juga	350	-7.283
	lo	276	-7.651
	tahun	258	-6.763
teenagers	nggak	240	-5.860
	bisa	218	-6.789
	juga	181	-7.747
	ada	150	-9.721
	lo	127	-8.223
adults	pada	190	-7.768
	tidak	143	-7.063
	tahun	138	-7.247
	dia	128	-7.720
	mereka	116	-7.077

TABLE V
THE FIRST FIVE VERBS WITH HIGHEST FREQUENCY FROM EACH CLASS.

Classes	Tokens	Frequency	TF-IDF
children	menjadi	126	-5.734
	tersebut	112	-5.945
	membuat	107	-5.483
	memiliki	99	-5.639
	merupakan	87	-5.728
teenagers	tersebut	91	-5.890
	buat	79	-5.976
	punya	44	-5.804
	bikin	42	-5.306
	membuat	42	-6.145
adults	mengatakan	103	-5.313
	tersebut	93	-6.021
	menjadi	65	-6.279
	menyebut	56	-5.560
	menyatakan	55	-5.453

Different from the verb POS, we cannot draw any conclusion from noun and adjective POS listed in Tables VI and VII because the words whose higher TF-IDF score even appear in all classes. We may need more filters to identified nouns and adjectives in each class. Or, we may need to combine words as n-gram or with the head of each word.

C. Evaluation Based on the Frequent Words

Based on our evaluation on Subsection IV-B, we drew a hypothesis that the children and teenagers datasets have almost similar vocabulary. To justify the hypothesis, we listed words that frequently appear in one or two datasets. Table VIII shows the first five words from the list. From this table, we may conclude that some words are dominant in a specific class (the tokens are written in red) even though they also appear in other class, but less frequent. For example, the word *kamu* ('you') will rarely appear in adults dataset because people will use *anda* ('you') in formal text or conversation.

On the other hand, Table IX lists the first five words that appear in all classes. Even though the frequency of some words

TABLE VI
THE FIRST FIVE NOUNS WITH HIGHEST FREQUENCY FROM EACH CLASS.

Classes	Tokens	Frequency	TF-IDF
children	teman	818	-4.569
	tahun	254	-6.290
	bulan	214	-5.506
	orang	212	-6.168
	nama	125	-6.261
teenagers	lagu	86	-5.526
	bisa	79	-6.174
	orang	68	-6.991
	film	66	-5.749
	waktu	63	-6.787
adults	*tahun*	129	-6.866
	pesawat	112	-5.526
	orang	93	-6.889
	kata	82	-6.654
	anak	79	-6.313

TABLE VII
THE FIRST FIVE ADJECTIVES WITH HIGHEST FREQUENCY FROM EACH CLASS.

Classes	Tokens	Frequency	TF-IDF
children	*ada*	385	-4.228
	lain	90	-4.808
	salah	80	-4.820
	satu	67	-4.748
	besar	60	-4.631
teenagers	*ada*	115	-5.132
	jadi	63	-4.480
	baru	52	-4.824
	sama	50	-4.612
	lain	46	-5.180
adults	*ada*	67	-5.627
	seorang	53	-4.559
	setempat	51	-4.064
	lain	50	-5.076
	baru	40	-5.053

TABLE VIII
LIST OF THE FIRST FIVE WORDS AND THEIR FREQUENCY, THAT APPEAR IN ONE OR TWO DATASETS.

Tokens	Frequency		
	children	***teenagers***	***adults***
teman	951	15	6
lo	311	141	0
nggak	1	269	0
kamu	58	83	0
bakal	0	82	6

TABLE IX
LIST OF THE FIRST FIVE WORDS AND THEIR FREQUENCY, THAT APPEAR IN ALL CLASSES.

Tokens	Frequency		
	children	*teenagers*	*adults*
menjadi	**137**	35	57
tersebut	**106**	79	88
ada	**385**	115	67
lain	**90**	46	50
baru	50	**52**	40

TABLE X
DERIVATION OF ***salah: Adjective*** BY CHANGING THE AFFIXES.

Words+affixes	Affixed Words	POS	Meaning
ber-*salah*	bersalah	verb	guilty
meN-*salah*-kan	menyalahkan	verb	blame
ke-*salah*-an	kesalahan	noun	mistake

is significantly higher (**written in bold**) from theirs in other classes but the value is not enough to to classify them to a specific class.

D. Evaluation on the NON-STEMMED *Dataset*

Based on Table II, we concluded that the NON-STEMMED dataset performs better than the STEMMED dataset. Therefore, we need to explore further the features provided by the NON-STEMMED dataset that affect the results. We predict that in the NON-STEMMED dataset, the complexity of the affixes influences the results.

In Indonesian, adding affixes conduces an inflection or derivation. Derivation will change the words POS, while inflection will not. The affixes are mostly derivational [17]. Table X shows three examples of adding different affixes to the word *salah* (adjective: 'wrong'). These examples show that the word POS is changed to other POS as well as changed the meaning.

We have found that the NON-STEMMED dataset performed better than the STEMMED one. Therefore, we listed the NON-STEMMED words as in Table XI. As an example, the word *buruk* ('bad') got the prefix *meN* to be *meN-buruk* ('to be bad') and may also be added by two prefixes *meN-* and *per-* to be *meN-per-buruk* ('worsen'). The *meN-per-* affix is rarely used in children vocabulary and may be consider as a complex affix. The performance of NON-STEMMED datasets were better because the stemming process will leave the tokens in their root word that will also shorten the vocabulary. It means that the words with different affixes will be treated as the same words. Even though reduce vocabulary work well in other languages like English, French, or Japanese, as Indonesian is rich in morphology [17], we need to explore further to find more common words used in children writings related to the simpler affixes. The further exploration will useful to build words choice recommendation for beginner learners. We assume that the beginner learners will use simpler word form.

TABLE XI
SAMPLE OF AFFIXED WORDS IN EACH CLASS.

Tokens	children	teenagers	adults
antara	antaranya	antaranya	antara, antaranya
buruk	**buruk**	**memburuk**	**memperburuk**
budaya	kebudayaan, kebudayaannya	berbudaya	budayanya
jangan	jangan	jangankan, janganlah	jangan
negara	negara	senegaranya	kenegaraan

TABLE XII
THRESHOLDING ON THE TF-IDF SCORE.

Thresholds	P	R	F
$>$ -11.8	0.9406	0.9341	**0.9373**
$>$ -11.6	0.9392	0.9325	0.9358
$>$ -11.3	0.9391	0.9325	0.9358
$>$ -10.1	0.9302	0.9203	0.9252

TABLE XIII
THRESHOLDING ON THE FREQUENCY.

Thresholds	P	R	F
$>$ 1	0.9396	0.9333	**0.9364**
$>$ 4	0.9308	0.9219	0.9263
$>$ 6	0.9334	0.9235	0.9284
$>$ 10	0.9396	0.9333	0.9364

E. Thresholding

The experiments got precision and recall of 0.9406 and 0.9341 respectively in the dataset for AT scenario. As the TF-IDF is a good feature in document classification, we added an experiment to remove tokens whose TF-IDF score are very low. Tables XII and XIII shows the thresholds and results of AT scenario both on the TF-IDF score and the term frequency. We tried to use lower threshold but the performance was worse. The tables show that the cut threshold performed lower. Therefore, it is better not to remove words whose frequency are low.

F. Human Evaluation

To deeply evaluated our model, we compared the classifications predicted by the system and ones assigned by a linguist. The evaluation was performed on the test set of nine articles outside ones used in the training data and the testing data of the five scenarios. Table XIV presents the results. It shows that only one article predicted by the system that missed the linguist assignation. The linguist assigned **Article-4** to *teenagers* and *adults*, while the system predicted it to *teenagers*. However, the prediction given by the system is still in the linguist's choices.

TABLE XIV
EVALUATION BY THE LINGUIST.

Test sets	Linguist Choice	System Prediction
Article-1	children	children
Article-2	adults	teenagers
Article-3	children	children
Article-4	teenagers, adults	teenagers
Article-5	teenagers	teenagers
Article-6	teenagers	teenagers
Article-7	adults	adults
Article-8	children	children
Article-9	teenagers	teenagers
Article-10	children	children

V. CONCLUSION

From the discussion, we concluded that the stemming process is inefficient for the document classification in Indonesian. The system performs the best on the model trained using all tokens in the dataset. On the other hand, verbs contribute significantly because the verbs used in non-adult class, especially children are highly disambiguation.

For future work, we want to justify how the complexity of affixes features benefits to the age group document classification. The aim is to explore whether simpler affixes will develop easier words for learners. Aside of further exploration words used in lower age groups, the simpler words may be used as recommendation to develop text for children or language learners in the beginner level.

REFERENCES

[1] A. Pentel, "Effect of Different Feature Types on Age Based Classification of Short Texts," in *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2015, pp. 1–7.

[2] J. Hong, C. Mattmann, and P. Ramirez, "Ensemble Maximum Entropy Classification and Linear Regression for Author Age Prediction," in *Information Reuse and Integration (IRI), 2017 IEEE 18th International Conference on*. IEEE, 2017.

[3] J. van de Loo, G. D. Pauw, and W. Daelemans, "Text-Based Age and Gender Prediction for Online Safety Monitoring," *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, vol. 5, no. 1, pp. 46–60, 2016.

[4] R. Guimarães, R. Rosa, D. De Gaetano, D. Z. Rodríguez, and G. Bressan, "Age Groups Classification in Social Network Using Deep Learning," *IEEE Access*, pp. 1–1, 05 2017.

[5] L. Pereira, E. Manguilimotan, and Y. Matsumoto, "Leveraging a Large Learner Corpus for Automatic Suggestion of Collocations for Learners of Japanese as a Second Language," *CALICO Journal*, vol. 33, no. 3, pp. 311–332, 2016.

[6] Dong Nguyen and Noah A. Smith and Carolyn P. Rosé, "Author Age Prediction from Text using Linear Regression," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, June 2011, pp. 115–123.

[7] X. Liu, K. Cheng, Y. Luo, K. Duh, and Y. Matsumoto, "A hybrid chinese spelling correction using language model and statistical machine translation with reranking," in *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, October 2013, pp. 54–58. [Online]. Available: http://www.aclweb.org/anthology/W13-4409

[8] A. Rozovskaya and D. Roth, "Training Paradigms for Correcting Errors in Grammar and Usage," in *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. Los Angeles, California, USA: Association for Computational Linguistics, 2010, pp. 154–162. [Online]. Available: http://www.aclweb.org/anthology/N10-1018

[9] B. Irmawati, H. Shindo, and Y. Matsumoto, "Exploiting Syntactic Similarities for Preposition Error Corrections on Indonesian Sentences Written by Second Language Learner," in *SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced Languages*, ser. Procedia Computer Science, S. Sakti, M. Adriani, A. Purwarianti, L. Besacier, E. Castelli, and P. Nocera, Eds., vol. 81. Yogyakarta, Indonesia: Elsevier, 2016, pp. 214–220. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2016.04.052

[10] ——, "Generating Artificial Error Data for Indonesian Preposition Error Correction," *International Journal of Technology (IJTech)*, vol. 8, no. 3, pp. 549–558, Apr. 2017. [Online]. Available: http://ijtech.eng.ui.ac.id/article/view/215

[11] W. Coster and D. Kauchak, "Simple English Wikipedia: A new text simplification task," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 665–669. [Online]. Available: https://www.aclweb.org/anthology/P11-2117

[12] J. Asian, "Effective Techniques for Indonesian Text Retrieval," Ph.D. dissertation, School of Computer Science and Information Technology, RMIT University Australia, 3 2007.

[13] A. Rahmatulloh, N. I. Kurniati, I. Darmawan, A. Z. Asyikin, and D. W. J, "Comparison between the Stemmer Porter Effect and Nazief-Adriani on the Performance of Winnowing Algorithms for Measuring Plagiarism," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 4, pp. 1124–1128, 2019. [Online]. Available: http://ijaseit.insightsociety.org/index.php?option=com_content&view=article&id=9&Itemid=1&article_id=8844

[14] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000.

[15] A. A. Adebiyi, O. Ogunleye, O. M. Adebiyi, and J. O. Okesola, "A Comparative Analysis of TF-IDF, LSI and LDA in Semantic Information Retrieval Approach for Paper-Reviewer Assignment," *Journal of Engineering and Applied Sciences*, vol. 14, no. 10, pp. 3378–3382, 2019.

[16] L. John, *The Cross Validation Problem*. Springer, Berlin, Heidelberg, 2005.

[17] J. N. Sneddon, A. Adelaar, D. N. Djenar, and M. C. Ewing, *Indonesian: A Comprehensive Grammar*. Routledge, Australia, 2010.

Age Group Based Document Classification in Bahasa Indonesia

ORIGINALITY REPORT

9%	8%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.scinapse.io Internet Source	2%
2	www.coursehero.com Internet Source	1%
3	acl-bg.org Internet Source	1%
4	Misbahuddin, Muhamad Syamsu Iqbal, Giri Wahyu Wiriasto. "Multi-hop Uplink for Low Power Wide Area Networks Using LoRa Technology", 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), 2019 Publication	1%
5	Submitted to University of Birmingham Student Paper	1%
6	Wirarama Wedashwara, I Wayan Agus Arimbawa, Andy Hidayat Jatmika, Ariyan Zubaidi, Tatang Mulyana. "IoT based Smart Small Scale Solar Energy Planning using	1%

Evolutionary Fuzzy Association Rule Mining", 2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS), 2020
Publication

7	Budi Dwi Satoto, Devie Rosa Anamisa, Mohammad Yusuf, Mohammad Kautsar Sophan, Siti Oryza Khairunnisa, Budi Irmawati. "Rice seed classification using machine learning and deep learning", 2022 Seventh International Conference on Informatics and Computing (ICIC), 2022 Publication	<1 %
8	aclweb.org Internet Source	<1 %
9	Lecture Notes in Computer Science, 2014. Publication	<1 %
10	Mohammad Aufar, Rachmadita Andreswari, Dita Pramesti. "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study", 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020 Publication	<1 %
11	repository.ittelkom-pwt.ac.id Internet Source	<1 %
12	patents.google.com Internet Source	

		<1%
13	Adil Ahnaf, Hossain Mohammad Mahmudul Hasan, Nabila Sabrin Sworna, Nahid Hossain. "An improved extrinsic monolingual plagiarism detection approach of the Bengali text", International Journal of Electrical and Computer Engineering (IJECE), 2023 Publication	<1%
14	hdl.handle.net Internet Source	<1%
15	www.researchgate.net Internet Source	<1%
16	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2018 Publication	<1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		