p-ISSN 2301–4156 | e-ISSN 2460–5719

Synonym Recognition Influence in Text Similarity Detection Using Winnowing and Cosine Similarity

Santi Purwaningrum[1], Agus Susanto[2], Ari Kristiningsih[3]

[1,2] Department of Informatics and Business Engineering Politeknik Negeri Cilacap, Cilacap 53212 INDONESIA (tel.: 0282-533329; fax: 0274-4321982, email: [1]santi.purwaningrum@pnc.ac.id, [2]agussusanto@pnc.ac.id)
[3] Department of Agroindustry Product Development, Politeknik Negeri Cilacap, Cilacap 53212 INDONESIA (tel.: 0282-533329; fax: 0274-4321982, email: [3]ari.kristiningsih@pnc.ac.id)

[Received: 19 January 2023, Revised: 4 July 2023]
Corresponding Author: Santi Purwaningrum

ABSTRACT — Plagiarism is an act of imitating, quoting and even copying or acknowledging the work of others as one's own work. A final project is one of the mandatory requirements for students to complete learning at college. It must be written by the students based on their own ideas. However, there is much plagiarism because it is easy to carry out just by simply copying the text of other people's ideas and then pasting it into a worksheet and admitting that the ideas are theirs. In addition, replacing some words in other people's sentences with their own language style without properly acknowledging the original source of the quotation is also an act of plagiarism. A manual check for the final project also becomes an issue for the final project coordinator, i.e., it needs high accuracy and a relatively long time to check the plagiarism in the final project document. Therefore, implementing plagiarism detection mechanisms is necessary to mitigate the escalation of plagiarism occurrences. In response to those matters, this study aims to design a system capable of identifying textual similarities by focusing on sentences containing synonymous words. One of the used algorithms is synonym recognition, which detects words that possess synonymous meanings by comparing each term with the entries in a dictionary. The synonym recognition is combined with the winnowing method, functioning as a fingerprint-based text weighting. After the weight of each document is obtained, the similarity level between documents is calculated with the cosine similarity algorithm. The inclusion of synonym recognition in conjunction with the winnowing weighting method resulted in a notable gain of 3.11% in the average similarity scores for title and abstract detection, compared to the absence of synonym recognition. The results show that the used algorithms are accurate with accuracy testing and root mean squared error (RMSE).

KEYWORDS — Synonym Recognition, Winnowing, Cosine Similarity, Plagiarism.

I. INTRODUCTION

A final project is a compulsory prerequisite for students pursuing diploma-level courses in order to obtain an associate expert and applied bachelor's degree. It is taken subsequent to the completion of an industrial internship. Therefore, the students are expected to derive potential topics for their final projects from the case studies encountered during their internship experience. Case studies are not required to be conducted at the internship industry site. Students are allowed to find another case study site with a topic that suits the needs of the site.

A system implementation is not a big deal for vocational students. Nevertheless, the process of drafting the final project report has certain challenges for students, particularly when it comes to articulating ideas or concepts so that they are not considered plagiarism. Plagiarism may occur either deliberately or unintentionally. Due to a lack of reading or difficulty in finding references, the concepts or ideas that will be used as the topic of the final project can be the same or have been made before [1], [2].

The process of uploading titles, abstracts, proposals, journals, posters, and final project reports to the system is carried out by the final project coordinator. However, the final project coordinator only uploads without paying attention to each title, or abstract, to the student's final project report. Consequently, it results in a failure to identify numerous instances of similarity across final project reports, both across different classes and within the same class. On the other hand, the examination process by the final project coordinators requires a considerable amount of time and high concentration because they have to check the final project one by one.

Furthermore, alongside the aforementioned issue, the copy-paste phenomenon has also developed. It is conducted by only substituting words containing synonyms without citing their sources. Early detection of similarities in titles and abstracts of scientific papers is needed to reduce acts of plagiarism arising from the students' lack of creativity.

Many researchers are interested in conducting research on plagiarism cases. The methods used for plagiarism detection text weighting include the winnowing algorithm, the Boyer-Moore algorithm, and the Rabin-Karp algorithm (or also known as Karp-Rabin) [3], [4]. The winnowing algorithm has been used in a study to detect the similarity level of thesis titles submitted compared with preexisting thesis titles [5]. The findings of this study indicate that with the score of *n*-gram = 3, window = 3, and prime number = 2, a similarity score of 73.86% has been achieved. This score suggests a significant degree of similarity between the examined titles. With the input of *n*-gram = 7, window = 9, and prime number = 2, a similarity score of 19.82% has been achieved, indicating a moderate plagiarism level [5].

A study describes an overview of the definition of plagiarism, tools to detect plagiarism, comparison matrices, obfuscation methods, data sets used for comparison, and the algorithm types. It is concluded that three algorithms that are often used in plagiarism detection are running Karp-Rabin greedy string tiling (RKR-GST), the winnowing algorithm, and the implementation process in the tokenization [6].

In the weighting process, another study uses the Rabin-Karp algorithm combined with the synonym recognition approach to identify words that have been modified into word

p-ISSN 2301–4156 | e-ISSN 2460–5719

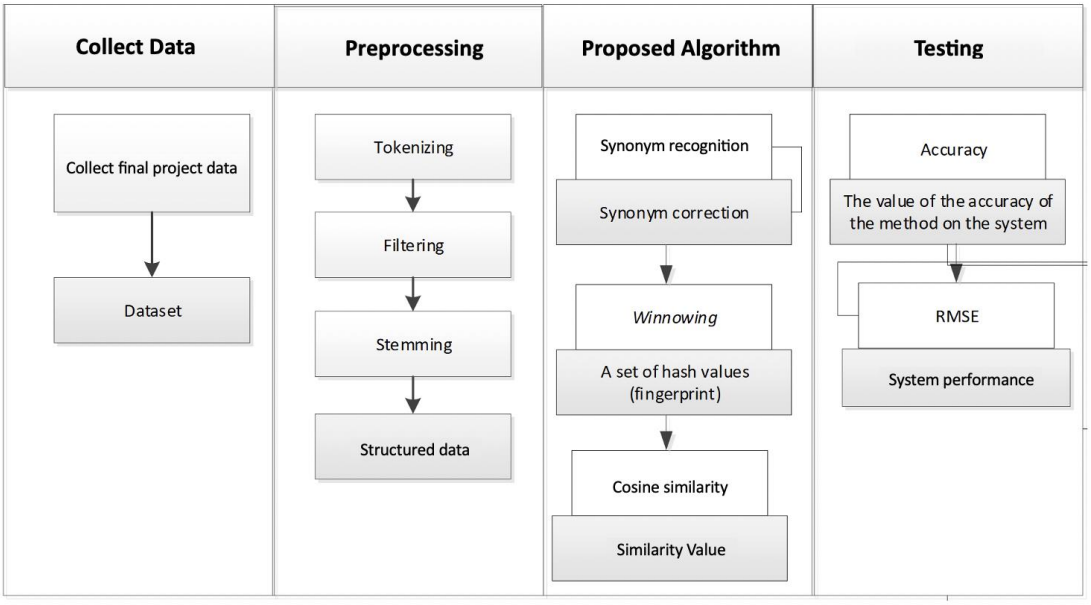

Figure 1. Research stages.

forms with similar meanings. The Rabin-Karp algorithm involves a step known as the rolling hash, which necessitates the provision of a base score for the weighting process. Not all numbers can be used in the base score in this algorithm. The rationale behind this phenomenon is that in certain cases, an incorrect base score yields an identical hash score to another hash score corresponding to a different word [7].

A study aiming at reducing bias in a description answer scoring system using the cosine similarity method with term frequency-inverse document frequency (TF-IDF) weighting and word matching by adding linear regression has been conducted [8]. Frequently, the assessment system provides outputs with higher scores than manual assessment by lecturers, indicating the presence of bias in the scoring. Therefore, linear regression is used to reduce bias to prevent the score of the description answer generated by the system from overestimating or underestimating the manual score given by the lecturer. In the linear regression process, the system score refers to the answer key that has been made, which is adjusted to the lecturer's assessment. Therefore, lecturers are expected to give an objective assessment [8].

Previous study has used the synonym recognition approach to identify terms containing synonyms and have used cosine similarity to determine similarity scores. The text weighting process of each compared document has used term frequency (TF), and root mean squared error (RMSE) has been used to test the accuracy. There are several stages in RMSE testing, namely 20, 50, 100, and 116 data. The highest RMSE score achieved in answer number 1 corresponding to a sample size of 20 students, which was 2.07, and the lowest was associated with a sample size of 50 students, which was 6.16. In answer number 2, the highest RMSE score of 8.94 was obtained in the group of 50 students, and the lowest score at 8.00 was in the group of 20 students. The accuracy testing has a relatively high error score because the words in the student's answers are included in the stopword dictionary, resulting in their loss in the filtering process. Therefore, in further research, it is expected that the similarity score between the two texts can be increased [9].

State of the art of previous research has discussed text similarity detection with many methods used for weighting a text, including Rabin-Karp, winnowing, Smith-Waterman and Manber, and TF [10]–[12]. In addition, algorithms used to find similarity scores, including Jaccard, dice similarity, neural network, and cosine similarity, are considered to have better similarity rates [13]. Reference [14] has investigated the collaborative use of the cosine similarity method with other weighting algorithms, such as artificial neural network (ANN) and support vector machine (SVM). However, the combination of winnowing method has not been explored in the study.

The main objective of this study is to detect the similarity of titles and abstracts in students' final assignments earlier by focusing on words with synonyms. The method used to calculate the similarity level between documents is cosine similarity. In the similarity method, the weighting process is usually conducted using TF-IDF to determine the weight score of each document. However, in this study, the weighting score of each document uses the winnowing algorithm, which is a fingerprint-based weighting. This algorithm is combined with a synonym recognition method used to detect words in documents compared with a dictionary of synonyms. With the combination of synonym recognition, final assignments with synonym-based paraphrases are expected to be detected. The performance of each combined method is measured using the accuracy parameter, while RMSE is used to determine the error score of the score recommended by the system against the actual score [15].

II. METHODOLOGY

In this stage, the impact of the synonym recognition algorithm in conjunction with the winnowing method and cosine similarity, is examined to detect the similarity between titles and abstracts of final projects focusing on words containing synonyms.

Figure 1 illustrates the stages in the research, from the initial process, i.e., data collection, initial data processing, algorithm implementation, i.e., synonym recognition for detection of text containing synonyms and winnowing algorithm for weighting a text in the document (fingerprint), to the final process that obtains the percentage of similarity score results from the compared documents and the accuracy score of the system using winnowing, synonym recognition, and cosine similarity

p-ISSN 2301–4156 | e-ISSN 2460–5719

TABLE I
EXAMPLES OF INPUT DATA

Title	Abstract
Aplikasi Media Pembelajaran Pengenalan Aksara Jawa Menggunakan Augmented Reality di Smartphone Android (Studi Kasus: SDN 07 Adipala)	Teknologi mengalami kemajuan yang pesat dan mempengaruhi dunia pendidikan juga berbagai inovasi telah dilakukan untuk menunjang kegiatan belajar mengajar. Proses belajar mengajar bahasa Jawa yang dilakukan guru saat ini masih menggunakan cara ceramah juga membosankan inovasi teknologi yang dapat digunakan dan dapat membantu siswa untuk belajar bahasa Jawa salah satunya adalah Augmented Reality (AR). Penelitian ini memiliki rumusan masalah yaitu bagaimana merancang dan membuat aplikasi Augmented Reality (AR). Pembelajaran pengenalan aksara Jawa di smartphone Android Penulis menggunakan metode literature dan observasi lapangan sebagai metode pengumpulan data dan sebagai metode pengembangan sistem penulis menggunakan MDLC (Multimedia Development Life Cycle). Kata kunci: Augmented Reality, MDLC, Smartphone Android, Aksara Jawa, Teknologi, Pendidikan, Belajar Mengajar
Aplikasi Pembelajaran Pengenalan Rangka Manusia Menggunakan Augmented Reality (AR). Berbasis Smartphone Android (Studi Kasus: SD Negeri Jepara Kulon 01, Binangun)	Pemahaman belajar siswa dapat meningkat dengan tersedianya media belajar yang menarik. Media belajar yang menarik dapat memudahkan pengajar dalam menyampaikan materi. Penelitian ini diambil dari studi lapangan dengan guru dan juga siswa sekolah dasar di SDN Jepara Kulon 01, Binangun. Kurangnya alat peraga rangka manusia dan metode pembelajaran di SDN Jepara Kulon 01, Kecamatan Binangun yang kurang menarik karena hanya tersedianya gambar rangka manusia 2D membuat siswa merasa bosan, sehingga dapat menjadi salah satu penyebab kurangnya pemahaman siswa terhadap materi yang diajarkan. Penelitian ini dilakukan dengan menerapkan teknologi Augmented Reality (AR) yang diimplementasikan dalam bentuk aplikasi berbasis android. Teknologi Augmented Reality (AR) merupakan perpaduan antara 2D, 3D, dan dunia nyata yang digabung dalam satu objek dengan satu teknologi yang dapat digunakan sebagai media pembelajaran di bidang multimedia. Aplikasi ini dikembangkan menggunakan metode Multimedia Development Life Cycle (MDLC). Hasil kuisioner menunjukan aplikasi ini dapat membantu guru dan siswa dalam mengenalkan serta memahami bagian - bagian rangka manusia dan juga menarik minat belajar dibanding menggunakan buku dengan persentase yang didapat 28% menyatakan setuju dan 72% menyatakan sangat setuju dengan aplikasi ini. Kata Kunci: android, Augmented Reality (AR), media pembelajaran, rangka manusia

methods. Accuracy and RMSE tests were conducted to assess the influence of synonym recognition on the system with a combination of winnowing and cosine similarity weighting.

A. DATA COLLECTION

The initial step of the research process involves determining the methodology and location for data acquisition. The data acquisition involves collecting primary data, i.e., data directly obtained from the case study site, i.e., the Department of Informatics Engineering at Politeknik Negeri Cilacap.

B. DATA PROCESSING

The dataset is in the form of titles and abstracts of Indonesian final assignments of the students of Politeknik Negeri Cilacap majoring in Informatics Engineering class of 2021 and 2022. There were 182 final project reports which were then categorized into three sections based on the respective theme of the final project. Specifically, there were 42 reports related to augmented reality (AR), 63 reports related to decision support systems, and 77 reports related to e-commerce-based information systems. The average title length is between 12 to 15 words per title, while the average abstract length is 350 to 500 words per abstract.

In this process the data that had to be uploaded were prepared, i.e., the title and abstract text documents of the final project, so that they could be processed in the text similarity measurement method. The data were obtained from the final project information system of the Department of Informatics Engineering of Politeknik Negeri Cilacap. The collected data were in the form of a complete final project report, which was subsequently organized into separate documents based on their titles and abstracts. Examples of input data were in the form of final project report data organized into titles and abstracts and saved in the form of PDF files.

Table I presents examples of each document with titles and abstracts used as datasets for model testing. Documents from each title and abstract were compared and searched for words containing synonyms, and then a text weighting generating a fingerprint was made. Subsequently, the similarity score of each document's fingerprint was calculated.

C. PREPROCESSING

Preprocessing is the initial stage of the text mining process. It serves to transform the unstructured text data into structured text data [16], [17]. In preprocessing, a series of steps were taken to remove parts of text that were not needed in a document because they would become noise in the subsequent process. Preprocessing is divided into three distinct stages: tokenizing, filtering, and stemming [18], [19].

Tokenizing is the process of separating individual words in a document, commonly known as tokens [20]. This process also converted all uppercase letters into lowercase letters. In addition, tokenizing also removed all punctuation marks, numbers, and symbols, as they have no unique score because they were not related to the string to be processed.

The filtering process serves to remove meaningless words. The meaningless words are commonly referred to as stopwords. Some examples of stopwords are "juga," "dan," "untuk," and "adalah" [21]. It is necessary to delete these stopwords because if conjunctions frequently appear in a sentence, the text similarity percentage is very high, and it interferes with the accuracy of the text similarity method [22].

The stemming process serves to remove affixes on a word in a text document, so the word taken is the root word. It was conducted to facilitate the subsequent process. Some examples of affixes are "mem-," "-kan," "ber-," "-pun," and "me-an" [23]. The obtained root words that were used as tokens within each text content in order to enhance efficiency and accuracy in syntactic matching. For instance, document 1 contained the word "belajar" and document 2 contained the word "mengajar." After the stemming process, the words "belajar dan mengajar"

p-ISSN 2301–4156 | e-ISSN 2460–5719

changed to “ajar” because “ajar” is the root word of “belajar” and “mengajar.”

D. SYNONYM RECOGNITION

Synonym recognition is one of the methods of detecting text plagiarism instances with the synonym approach [7]. Wordnet is a lexical database intended for use under program control [24]. The database contains various types of vocabulary organized according to their synonyms according to a lexical concept. A word with two or more meanings of the same word is called a synonym. In the synonym recognition process, a list of words that have synonym meanings is required. This word list was obtained from the Great Dictionary of Indonesian Language (Kamus Besar Bahasa Indonesia, KBBI) online website and kamusbesar.com [15].

Figure 2 explains the stages of the synonym recognition method, from input to output. The preprocessed text undergoes the synonym recognition stage in order to detect plagiarism activity in a text document. This stage compares one document with another by detecting words that contain synonyms. The synonym recognition process compares the word contained in the document with the thesaurus contained in the database.

The winnowing algorithm alone cannot detect plagiarism of a document if the words in the document have experienced word modification, yet the word’s meaning remains the same. Therefore, a synonym recognition algorithm is required to overcome the word modification problem. The synonym recognition process changes all words diagnosed as synonyms of a word contained in the synonym dictionary which is considered as the main word [25].

E. WINNOWING

In the winnowing process, data are processed from string to numeric data. In the process, the user must enter parameter scores, namely *k*-gram, hash, and window. To determine the *k*-gram, hash, and window scores, parameter testing must be conducted with the sample data to obtain a good similarity score.

The input of the winnowing algorithm is the root word text from the data preprocessing results. Then, the output is a collection of hash scores. A hash score is a numerical score formed from the American Standard Code for Information Interchange (ASCII) table calculation of each character. The hash score can also be referred to as a fingerprint, which is used as an indicator to compare the similarity between text documents [26].

Before starting the winnowing process, a synonym recognition process was performed, which was the processing of preprocessed root words compared with the synonym dictionary. Next, the data entered the *k*-gram process, which was used to obtain a new set of strings from the old set of strings, with a *k*-gram score determined by the user. The set of strings was grouped into a new set of strings, which was a concatenation of the initial strings, with the length of the concatenated strings being *k*.

Then, a rolling hash process was carried out, which served to generate the hash score of each gram that had been formed. The conversion of a string of characters into a score or code that then becomes the sign of the string of characters is called hashing and the resulting score can be referred to as a hash score. The obtained hash score is already numerical data. This process could be conducted using the following formula.

$$H(c_1..c_l) = c_1.b^{(l-1)} + c_2.b^{(l-2)} + .. + c_{(l-1)}.b + c_1 \quad (1)$$

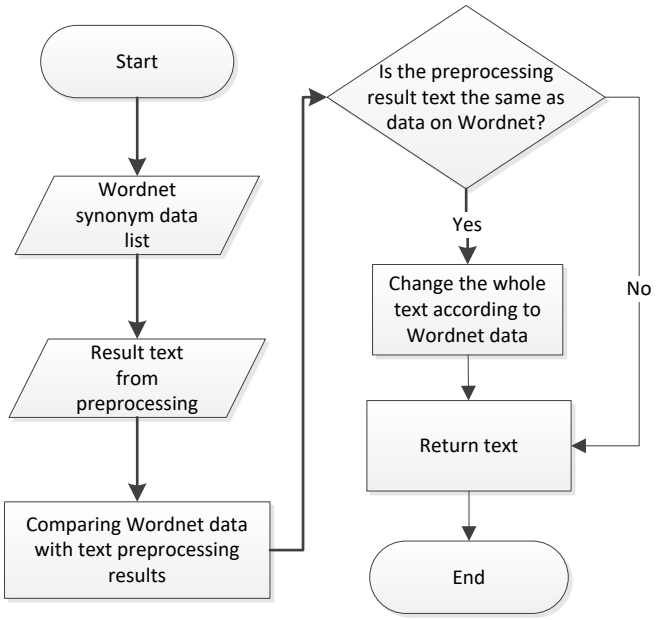

Figure 2. Synonym recognition process.

where *c* is the ASCII score of each character, *l* is the length of the string, and *b* is the user-defined hash base score. From this process, the hash score of each gram will be obtained.

The next step was to form a window on the hash score results of each previous gram. The window size was also determined by the user. The window process was carried out by determining the hash score of each window to be used as a document fingerprint and if there were the same hash score, the rightmost hash score would be selected. Then, the final process of the winnowing algorithm was to determine the smallest hash score from each window to be used as a fingerprint on a document.

F. COSINE SIMILARITY

Cosine similarity is one method that can be used to calculate the text similarity level in a sentence or document. Cosine similarity has a high accuracy score to determine the similarity level because it does not affect the length of a word or sentence in a compared document [27], [28]. It was calculated with the following formula.

$$similarity(X,Y) = \frac{(X \cap Y)}{|X|^{0,5}.|Y|^{0,5}} \quad (2)$$

where $X \cap Y$ is the number of words contained in document *X* and contained in document *Y*, $|X|$ is the number of words contained in document *X*, and $|Y|$ is the number of words contained in document *Y*.

G. ACCURACY

Accuracy is used to provide an assessment of predictive results which corresponds to the actual data. The higher the accuracy score, the more accurate or better the performance of the used method. The accuracy function is to determine the level of accuracy of recommendations for methods used in a system. The accuracy formula is written as follows [29]–[31].

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

where *TN* is the number of negative data detected correctly (true negative), *FP* is negative data detected as positive data (false positive), *TP* is positive data detected correctly (true positive), and *FN* is positive data detected as negative data (false negative).

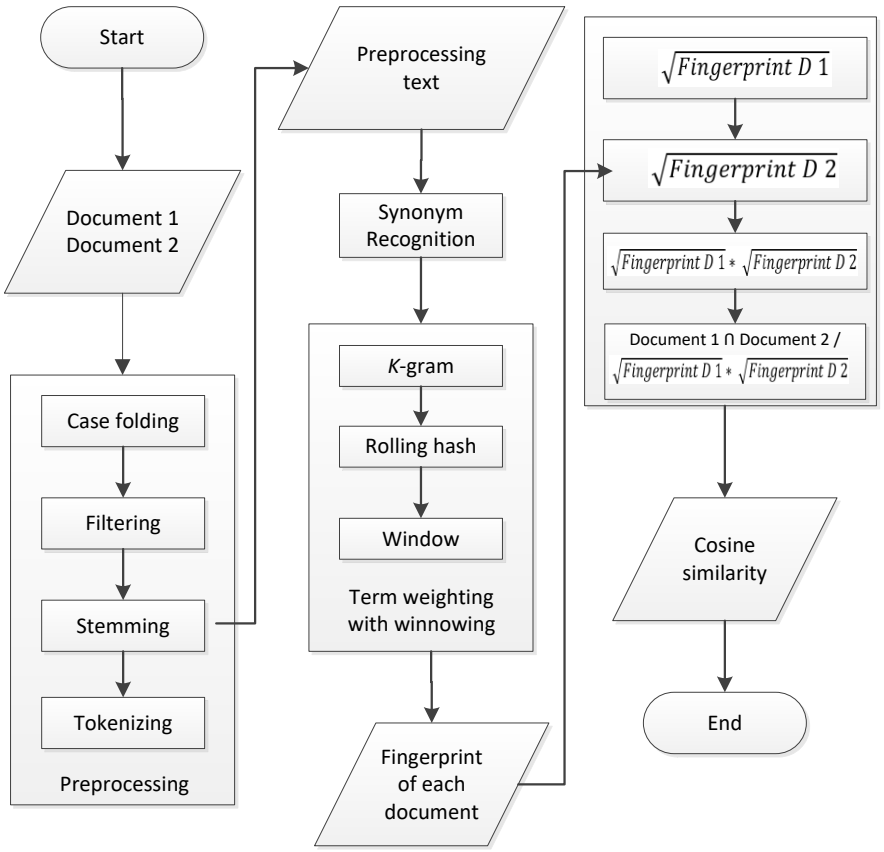

Figure 3. Flowchart of title and abstract similarity detection system.

H. ROOT MEAN SQUARE ERROR (RMSE)

RMSE is used to determine the performance of the system created. RMSE is the square root of the average squared difference between predictions and actual observations [32]. The greater the similarity score between the system and manual calculation, the more suitable the method used in the system.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}. \quad (4)$$

III. RESULTS AND DISCUSSION

Figure 3 describes the process of algorithms used in text similarity detection in the title and abstract of a final project that focuses on text containing synonyms. It is necessary to test the scores of *k*-gram, rolling hash, and window parameters to obtain maximum results in similarity detection in titles and abstracts in the weighting process using the winnowing method. The parameter testing was carried out using the example of two titles and abstracts of the decision support system with parameter scores as shown in Table II.

Table II presents the results of the winnowing parameter test. Each parameter was tested with different scores using a dataset of two titles and a final project abstract on the topic of decision support systems. The best parameter results in text weighting were obtained with the score of *k*-gram = 2, hash = 5, and window = 7, with an average title similarity score of 55.84% and abstract similarity of 59.42%. Therefore, these parameters were used in this study to analyze the influence of the synonym recognition algorithm on the title and abstract of the final project in the text weighting of the winnowing algorithm. Those parameters were combined with the cosine similarity method to determine the similarity between documents.

TABLE II
PARAMETER TESTING

K-gram	Hash	Window	Title Similarity (%)	Abstract Similarity (%)
5	7	2	24.50	26.21
7	2	5	18.81	19.37
2	5	7	55.84	59.42
2	7	5	51.74	57.33
7	5	2	19.44	20.50
5	2	7	25.55	27,67

The initial process was carried out by comparing two text documents, each of which was preprocessed. The example of preprocessing results in title 1 is “Aplikasi media belajar kenal aksara Jawa Augmented Reality smartphone android studi kasus SDN Adipala,” and the results of preprocessing title 2 is “Aplikasi belajar kenal rangka manusia Augmented Reality basis smartphone android studi kasus SD Negeri Jepara Kulon Binangun.”

After preprocessing was complete, the next step was the process of detecting words that contain synonyms. This synonym recognition method can add, delete, and change synonym words contained in the synonym dictionary database. Strings in the title or abstract were compared or matched with the synonym dictionary. A synonym dictionary is a collection of synonym-based words in Indonesian. The synonym dictionary focuses on three major topics, namely on AR, e-commerce-based information systems, and decision support systems. If there were title and abstract words that were the same as the synonym dictionary, the string contained in the title or abstract was replaced with the string as in the synonym dictionary. For instance, after going through preprocessing and

p-ISSN 2301–4156 | e-ISSN 2460–5719

synonym recognition, title 1 becomes “Aplikasi media belajar kenal aksara Jawa Augmented Reality smartphone android studi kasus SDN Adipala” and title 2 becomes “aplikasi tatar kenal rangka manusia Augmented Reality basis smartphone android studi kasus SD Negeri Jepara Kulon Binangun.” There is one string of document 1 and 2 titles identified as having words containing synonyms found in the synonym database dictionary, namely “belajar” to “tatar.”

After synonym recognition, the next process was text weighting of each compared document. The weighting process using the winnowing method was carried out by finding the fingerprint score of each document. The initial weighting process was to enter the parameter scores of *k*-gram, rolling hash, and window.

The example of the winnowing process results by entering *k*-gram score of 2 in title 1 is “ap, pl, li, ik, ka, as, si, it, ta, at, ta, ar, rk, ke, en, na, al, lr, ra, an, ng, gk, ka, am, ma, an, nu, us, si, ia, aa, au, ug, gm, me, en, nt, te, ed, dr, re, ea, al, li, it, ty, yb, ba, as, si, is, ss, sm, ma, ar, rt, tp, ph , ho, on, ne, ea, an, nd, dr, ro, oi, id, ds, st, tu, ud, di, ik, ka, as, su, us, ss, sd, dn, ne, eg, ge, er, rj, je, ep, pa, ar, ra, ak, ku , ul, lo, on, nb, bi, in, na, an, ng, gu, un”. The second process in winnowing is rolling hash. For example, the entered hash score was 5, then the result of the title 1 in the first string “ap”, the string “a” in the ASCII table has a score of 97, and “p” has a score of 112. Therefore, the calculation in the rolling hash formula is (97×5(2-1))+ (112×5(1-1)) = 485+112 = 597. Therefore, the overall rolling hash result of title 1 is [ap : 597, pl : 668, li : 645, ik : 632, ka : 632, as : 600, si : 680, it : 641, ta : 677, at : 601, ta : 677, ar : 599, rk : 677, ke : 636, en : 615, na : 647, al : 593, lr : 654, ra : 667, an : 595, ng : 653, gk : 622, ka : 632, am : 594, ma : 642, an : 595, nu : 667, us : 700, si : 680, ia : 622, aa : 582, au : 602, ug : 688, gm : 624, me : 646, en : 615, nt : 666, te : 681, ed : 605, dr : 614, re : 671, ea : 602, al : 593, li : 645, it : 641, ty : 701, yb : 703, ba : 587, as : 600, si : 680, is : 640, ss : 690, sm : 684, ma : 642, ar : 599, rt : 686, tp : 692, ph : 664, ho : 631, on : 665, ne : 651, ea : 602, an : 595, nd : 650, dr : 614, ro : 681, oi : 660, id : 625, ds : 615, st : 691, tu : 697, ud : 685, di : 605, ik : 632, ka : 632, as : 600, su : 692, us : 700, ss : 690, sd : 675, dn : 610, ne : 651, eg : 608, ge : 616, er : 619, rj : 676, je : 631, ep : 617, pa : 657, ar : 599, ra : 667, ak : 592, ku : 652, ul : 693, lo : 651, on : 665, nb : 648, bi : 595, in : 635, na : 647, an : 595, ng : 653, gu : 632, un : 695].

The third step in winnowing is the window. The window process is the same as *k*-gram. However, the processed data are already in the form of numbers from the previous rolling hash process. For example, the results of title 1 with the window input score of 5 were {597, 668, 645, 632, 632, 600, 680}, {668, 645, 632, 632, 600, 680, 641}, {645, 632, 632, 600, 680, 641, 677}, {632, 632, 600, 680, 641, 677, 601}, {632, 600, 680, 641, 677, 601, 677}, {600, 680, 641, 677, 601, 677, 599}, {680, 641, 677, 601, 677, 599, 677}, {641, 677, 601, 677, 599, 677, 636}, {677, 601, 677, 599, 677, 636, 615} … {595, 635, 647, 595, 653, 632, 695}.

The last step of the winnowing process is to find the minimum score of each window. If there is the same minimum score from each window, only one score is taken, which is the rightmost score. It is done to avoid redundancy, which affects accuracy and reduces processing time. The minimum score of each window is referred to as the fingerprint. The results of the fingerprint score on title 1 after the winnowing process was complete were [597, 600, 599, 593, 594, 582, 602, 605, 587, 595, 614, 615, 608, 592].

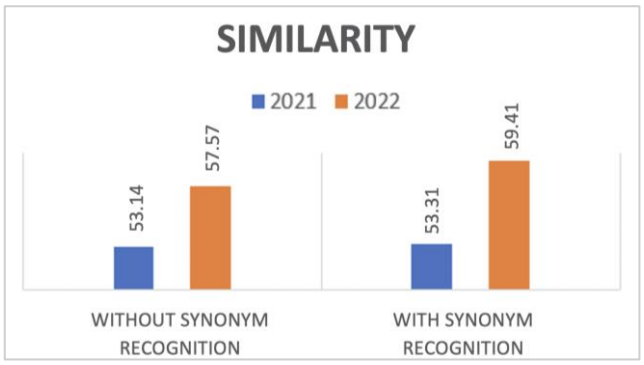

Figure 4. Influence of synonym recognition on the similarity score.

After the fingerprint score of each document was obtained, the next process was to find the similarity score of the compared text documents using the cosine similarity method. This process was carried out by dividing the same number of fingerprints from the compared documents by the sum of the square root of the fingerprints of each compared document.

The process of using synonym recognition and without synonym recognition combined with the weighting method using winnowing and the similarity search method using cosine similarity yielded the following results. In the decision support system topic, the average score of the title similarity difference in 2021 and 2022 is 6.8%, in the e-commerce-based information system topic by 1.11%, and in the AR topic by 6.07%. In addition, the average scores of abstract similarity differences on the topics of decision support systems, e-commerce-based information systems, and AR topics are 1.41%, 2.75%, and 0.52%, respectively. The average similarity scores of all topics in titles and abstracts compared using synonym recognition combined with weighting method using winnowing and similarity search method using cosine similarity are 59.32% and 55.19%, while the average scores of all topics in titles and abstracts without synonym recognition combined with weighting method using winnowing and similarity search method using cosine similarity are 54.67% and 53.63%.

Figure 4 highlights the average score of title and abstract similarity in 2021 and 2022 as a consequence of using and not using synonym recognition. The data presented in the figure shows that in 2021 and 2022 the average similarity score is increased by 2.73%.

The title and abstract similarity detection system displays the results of the similarity percentage of each title and abstract compared. The accuracy and RMSE scores are sought to determine the success rate of the algorithms in the system used. The influence of using synonym recognition combined with a weighting method using winnowing and a similarity search method using cosine similarity on title and abstract similarity detection is compared with manual calculations, yielding an average accuracy rate of 81.05%. The accuracy score was obtained by dividing the TP score of 147.57% by the amount of data. In addition, the average RMSE score is 4.38%. This score is obtained from the actual data of 63.7% minus the forecasting result score of 2.01%, followed by calculating the the square root of the difference and dividing it by the amount of data.

IV. CONCLUSION

Based on the results of the conducted tests, it can be inferred that the smaller *k*-gram parameter score is sufficient to affect

the similarity level scores between documents. Conversely, the input scores for hash and window do not significantly affect the similarity level score between documents. The highest parameter score input in the text weighting similarity score is *k*-gram = 2, hash = 5, and window = 7. The influence of synonym recognition on text weighting using the winnowing algorithm and calculating the similarity score using cosine similarity demonstrates an average increase in the similarity score of titles and abstracts by 3.11% compared to synonym recognition without the winnowing algorithm. The rise in the similarity score of titles and abstracts is derived from the average score of titles and abstracts using winnowing weighting combined with synonym recognition and calculating the similarity score using cosine similarity, amounting to 57.26%, minus the average score without synonym recognition, amounting to 54.15%. The addition of the synonym recognition method combined with winnowing weighting and calculation of similarity score using cosine similarity has the potential to increase the similarity score, especially in synonym-based text. Testing the influence of the synonym recognition method combined with the winnowing algorithm and cosine similarity yielded an accuracy score of 80.97% and RMSE of 26.14%. These results are obtained from the dataset of the final project conducted by students of Politeknik Negeri Cilacap. This dataset was selected for modeling purposes, as using other datasets may lead to the possibility of getting different results.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest in relation to the research and development of this paper.

AUTHOR CONTRIBUTION

Conceptualization, Santi Purwaningrum and Agus Susanto; methodology, Santi Purwaningrum and Agus Susanto; software, Santi Purwaningrum; validation, Santi Purwaningrum and Agus Susanto; formal analysis, Santi Purwaningrum and Agus Susanto; formal analysis, Santi Purwaningrum; investigation, Santi Purwaningrum; formal analysis, Santi Purwaningrum; resources, Santi Purwaningrum; data curation, Santi Purwaningrum; writing-original drafting, Santi Purwaningrum; writing-reviewing and editing, Ari Kristiningsih; visualization, Santi Purwaningrum; supervision, Santi Purwaningrum.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Politeknik Negeri Cilacap for their support in the form of funding grants, which made this study possible. Additionally, the authors would like to extend their appreciation to all individuals who contributed to the successful execution of this study.

REFERENCES

[1] M.H.P. Swari, C.A. Putra, and I.P.S. Handika, "Plagiarsm Checker pada Sistem Manajemen Data Tugas Akhir," *J. Sains, Inform.*, Vol. 7, No. 2, pp. 192–201, Nov. 2021, doi: 10.34128/jsi.v7i2.338.

[2] M.H.P. Swari and C.A. Putra, "Sistem Manajemen Data Skripsi (Studi Kasus: Perpustakaan Fakultas Ilmu Komputer UPN "Veteran" Jawa Timur)," *J. Pendidik. Teknol., Kejuru.*, Vol. 17, No. 2, pp. 198–209, Jul. 2020, doi: 10.23887/jptk-undiksha.v17i2.25436.

[3] F.E. Kurniawati and W.M. Pradnya, "Implementasi Algoritma Winnowing pada Sistem Penilaian Otomatis Jawaban Esai pada Ujian Online Berbasis Web," *J. Tek. Komput. AMIK BSI*, Vol. 6, No. 2, pp. 169–175, Jul. 2020, doi: 10.31294/jtk.v6i2.7838.

[4] I. Ahmad, R.I. Borman, G.G. Caksana, and J. Fakhrurozi, "Implementasi String Matching dengan Algoritma Boyer-Moore untuk Menentukan Tingkat Kemiripan pada Pengajuan Judul Skripsi/TA Mahasiswa (Studi Kasus: Universitas XYZ)," *SINTECH (Sci., Inf. Technol. J.)*, Vol. 4, No. 1, pp. 53–58, Apr. 2021, doi: 10.31598/sintechjournal.v4i1.699.

[5] N. Alamsyah and M. Rasyidan, "Deteksi Plagiarisme Tingkat Kemiripan Judul Skripsi pada Fakultas Teknologi Informasi Menggunakan Algoritma Winnowing," *Technologia*, Vol. 10, No. 4, pp. 197–201, Oct.-Dec. 2019, doi: 10.31602/tji.v10i4.2361.

[6] M. Novak, M. Joy, and D. Kermek, "Source-Code Similarity Detection and Detection Tools Used in Academia: A Systematic Review," *ACM Trans. Comput. Educ.*, Vol. 19, No. 3, pp. 1-37, May 2019, doi: 10.1145/3313290.

[7] N.P. Putra and Sularno, "Penerapan Algoritma Rabin-Karp dengan Pendekatan Synonym Recognition Sebagai Antisipasi Plagiarisme pada Penulisan Skripsi," *J. Teknol., Sist. Inf. Bisnis*, Vol. 1, No. 2, pp. 130–140, Jul. 2019, doi: 10.47233/jteksis.v1i2.52.

[8] S. Fauziati *et al*., "Regresi Linear untuk Mengurangi Bias Sistem Penilaian Uraian Singkat," *J. Nas. Tek. Elekt., Teknol. Inf.*, Vol. 10, No. 3, pp. 221–228, Aug. 2021, doi: 10.22146/jnteti.v10i3.1983.

[9] I. Mufiid, S. Lestanti, and N. Kholila, "Aplikasi Penilaian Jawaban Esai Otomatis Menggunakan Metode Synonym Recognition dan Cosine Similarity Berbasis Web," *J. Mnemonic: J. Tek. Inform*., Vol. 4, No. 2, pp. 31–37, Sep. 2021, doi: 10.36040/mnemonic.v4i2.4067.

[10] B. Sari and Y. Sibaroni, "Deteksi Kemiripan Dokumen Bahasa Indonesia Menggunakan Algoritma Smith-Waterman dan Algoritma Nazief & Andriani," *Ind. J. Comput.*, Vol. 4, No. 3, pp. 87–98, Dec. 2019, doi: 10.21108/indojc.2019.4.3.365.

[11] M.R. Parvez, W. Hu, and T. Chen, "Comparison of the Smith-Waterman and Needleman-Wunsch Algorithms for Online Similarity Analysis of Industrial Alarm Floods," *2020 IEEE Elect. Power, Energy Conf. (EPEC)*, 2020, pp. 1–6, doi: 10.1109/EPEC48502.2020.9320080.

[12] V. Kumar, C. Bhatt, and V. Namdeo, "A Framework for Document Plagiarism Detection Using Rabin Karp Method," *Int. J. Innov. Res. Technol., Manage.*, Vol. 5, No. 4, pp. 17–30, Aug. 2021.

[13] T. Wahyuningsih, Henderi, and Winarno, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," *J. Appl. Data Sci.*, Vol. 2, No. 2, pp. 45–54, May 2021, doi: 10.47738/jads.v2i2.31.

[14] L. Meilina, I.N.S. Kumara, and N. Setiawan, "Literature Review Klasifikasi Data Menggunakan Metode Cosine Similarity dan Artificial Neural Network," *Maj. Ilm. Teknol. Elekt.*, Vol. 20, No. 2, pp. 307–314, Jul.–Dec. 2021, doi: 10.24843/mite.2021.v20i02.p15.

[15] M.N. Cholis, E. Yudaningtyas, and M. Aswin, "Pengaruh Penggunaan Synonym Recognition dan Spelling Correction pada Hasil Aplikasi Penilaian Esai dengan Metode Longest Common Subsequence dan Cosine Similarity," *InfoTekJar (J. Nas. Inform., Teknol. Jar.)*, Vol. 3, No. 2, pp. 242–246, Sep. 2019, doi: 10.30743/infotekjar.v3i2.1061.

[16] Sunardi, A. Yudhana, and I.A. Mukaromah, "Indonesia Words Detection Using Fingerprint Winnowing Algorithm," *J. Inform.*, Vol. 13, No. 1, pp. 7–15, Jan. 2019, doi: 10.26555/jifo.v13i1.a8452.

[17] M.R. Faisal, D. Kartini, A.R. Arrahimi, and T.H. Saragih, *Belajar Data Science: Text Mining Untuk Pemula I*. Banjarbaru, Indonesia: Scripta Cendekia, 2023.

[18] H.A. Rouf, A. Wijayanto, and A. Aziz, "Deteksi Plagiarisme Skripsi Mahasiswa dengan Metode Single-link Clustering dan Jaro-Winkler Distance," *J. Pilar Teknol.*, Vol. 5, No. 1, pp. 26–31, Mar. 2020, doi: 10.33319/piltek.v5i1.50.

[19] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge, England: Cambridge University Press, 2008.

[20] S.P. Gunawan, L. Dwika, and A.R. Chrismanto, "Analisis Fitur Stilometri dan Strategi Segmentasi pada Sistem Deteksi Plagiasi Intrinsik Teks*," J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 4, No. 5, pp. 988–997, Oct. 2020, doi: 10.29207/resti.v4i5.2486.

[21] N.C. Haryanto, L.D. Krisnawati, and A.R. Chrismanto, "Temu Kembali Dokumen Sumber Rujukan dalam Sistem Daur Ulang Teks," *J. Teknol., Sist. Komput.*, Vol. 8, No. 2. pp. 140–149, Apr. 2020. doi: 10.14710/jtsiskom.8.2.2020.140-149.

[22] I.M.S. Putra, P. Jhonarendra, and N.K.D. Rusjayanthi, "Deteksi Kesamaan Teks Jawaban pada Sistem Test Essay Online dengan Pendekatan Neural Network," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 5, No. 6, pp. 1070–1082, Dec. 2021, doi: 10.29207/resti.v5i6.3544.

[23] N.L.W.S.R. Ginantra and N.W. Wardani, "Implementasi Metoda Naïve Bayes dan Vector Space Model dalam Deteksi Kesamaan Artikel Jurnal Berbahasa Indonesia," *J. Infomedia*, Vol. 4, No. 2, pp. 94–100, Dec. 2019, doi: 10.30811/jim.v4i2.1530.

[24] R.P. Nuristiqomah and Y. Anistyasari, "Pengembangan Kamus Istilah Basis Data Berbasis Website Menggunakan Algoritma Cosine Similarity untuk Meningkatkan Hasil Belajar Siswa," *J. IT-EDU*, Vol. 5, No. 2, pp. 621–630, 2021.

[25] R. Nishiyama, "Adaptive Use of Semantic Representations and

p-ISSN 2301–4156 | e-ISSN 2460–5719

Phonological Representations in Verbal Memory Maintenance," *J. Mem. Lang.*, Vol. 111, pp. 1–11, Apr. 2020, doi: 10.1016/j.jml.2019.104084.

[26] S. Inturi and S. Dusa, "Assessment of Descriptive Answers in Moodle-Based E-Learning Using Winnowing Algorithm," *J. Contemp. Issues Bus. Gov.*, Vol. 27, No. 3, pp. 2759–2769, 2021, doi: 10.47750/cibg.2021.27.03.331.

[27] E. Siswanto and Y.C. Giap, "Implementasi Algoritma Rabin-Karp dan Cosine Similarity untuk Pendeteksi Plagiarisme Pada Dokumen," *J. ALGOR*, Vol. 1, No. 2, pp. 16–22, May 2020.

[28] Y. Nurdiansyah, A. Andrianto, and L. Kamshal, "New Book Classification Based on Dewey Decimal Classification (DDC) Law Using TF-IDF and Cosine Similarity Method," *J. Phys. Conf. Ser.*, Vol. 1211, pp. 1–9, 2019, doi: 10.1088/1742-6596/1211/1/012044.

[29] R.N. Harahap and K. Muslim, "Peningkatan Akurasi pada Prediksi Kepribadian Mbti Pengguna Twitter Menggunakan Augmentasi Data," *J. Teknol. Inf., Ilmu Komput.*, Vol. 7, No. 4, pp. 815–822, Aug. 2020, doi: 10.25126/jtiik.2020743622.

[30] J. Xu, Y. Zhang, and D. Miao, "Three-Way Confusion Matrix for Classification: A Measure Driven View," *Inf. Sci. (Ny).*, Vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

[31] Y. Zhang and J.T. Yao, "Gini Objective Functions for Three-Way Classifications," *Int. J. Approx. Reason.*, Vol. 81, pp. 103–114, Feb. 2017, doi: 10.1016/j.ijar.2016.11.005.

[32] E. Sutoyo and A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 4, No. 1, pp. 95–101, Feb. 2020, doi: 10.29207/RESTI.V4I1.1502.