Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Plagiarism Detection in Students' Theses Using The Cosine Similarity Method

Oppi Anda Resta[1)], Addin Aditya[2)*], Febry Eka Purwiantono[3)]
[1,2,3)]STIKI Malang, Indonesia
[1)]oppi.andar@gmail.com, [2)]addin@stiki.ac.id, [3)]febry@stiki.ac.id

Submitted : Mar 17, 2021 | **Accepted** : Apr 27, 2021 | **Published** : May 1, 2021

Abstract: The main requirement for graduation from students is to make a final scientific paper. One of the factors determining the quality of a student's scientific work is the uniqueness and innovation of the work. This research aims to apply data mining methods to detect similarities in titles, abstracts, or topics of students' final scientific papers so that plagiarism does not occur. In this research, the cosine similarity method is combined with the preprocessing method and TF-IDF to calculate the level of similarity between the title and the abstract of a student's final scientific paper, then the results will be displayed and compared with the existing final project repository based on the threshold value to make a decision whether scientific work can be accepted or rejected. Based on the test data and training data that has been applied to the TF-IDF method, it shows that the percentage level of similarity between the training data document and the test data document is 8%. This shows that the student thesis is still classified as unique and does not contain plagiarism content. The findings of this study can help the university in managing the administration of student theses so that plagiarism does not occur. Furthermore, it is necessary to study further adding methods to increase the accuracy of system performance so that when the process is run the system will work faster and optimally.

Keywords: Cosine Similarity, Student Theses, Text Mining, TF-IDF, Plagiarism

INTRODUCTION

Scientific work is always related to the academic world because human ideas or thoughts are written in scientifically structured articles. This work must be protected both ethically and legally (Kurniasar, 2016). According to Ministerial Regulation Number 17 of 2010 concerning Prevention and Handling of Plagiarism in Higher Education, Chapter 1 Article 1 states that plagiarism is an act intentionally or not in obtaining or trying to obtain credit or value from a scientific work by quoting part or all of the work of other parties recognized as a scientific work without mentioning its source precisely and adequately (*Permendiknas RI No 17 Tahun 2010*, 2010). Types of plagiarism can be categorized as follows:

1. Refer to and / or quote terms, words and / or sentences, data and / or information from a source without mentioning the source in the citation notes and / or without mentioning the source adequately.
2. Referring and / or randomly quoting terms, words and / or sentences, data and / or information from a source without mentioning the source in the citation notes and / or without mentioning the source adequately
3. Using the source of ideas, opinions, views, or theories without mentioning the source adequately.
4. Formulate in own words and / or sentences from the source of words and / or sentences, ideas, opinions, views or theory without citing credible sources.
5. Submit scientific works that have been produced and / or have been published by other parties as scientific works without state the source adequately

Therefore, it is necessary to have a detection system that is transparent and can assist students in researching thesis titles so that students can find out whether the chosen title has similarities or not with previously published final project data. Also, the abstract of a thesis will also be examined because there is a possibility that the title is the same but different from the abstraction. This research aims to implement the cosine similarity method in detecting the similarity of student thesis titles and abstracts to avoid the occurrence of similarities in titles and thesis abstracts.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

LITERATURE REVIEW

In higher education, plagiarism arises from the theory of social exchange between students and lecturers based on the concept of reward and punishments. Student plagiarism is an action based on the concept of reward and punishment, because in working on a project, students commit plagiarism in the hope of getting a good score as a 'reward' and avoiding the lecturer's anger as 'punishment' (Prihantini & Indudewi, 2016). Research conducted in Australia found that several factors that cause plagiarism (Biliæ-zulle, Frkoviæ, Turk, & Josip, 2005)(Ryan, Bonanno, Krass, Scouller, & Smith, 2009):

1. The attitude of students who consider academic dishonesty including plagiarism is common and there is no strict sanction for that.
2. Lack of knowledge of scientific writing methods.
3. Lack of ability to write references correctly and lack of knowledge about plagiarism also causes this to happen.
4. Lack of intellectuality of students.
5. Lecture activities are stressful and the number of lecture assignments are accompanied by easy internet access, many references on the internet cause students to use shortcuts by imitating other people's work to complete assignments quickly.

One of the important factors that influence the acceptance of a thesis at a university is the thesis topic itself and the method to be used. So far, the student thesis sorting process is still being carried out by the university administration and students are not involved in the process. As a result, students do not know whether the thesis title to be submitted has ever been worked on or published. So there is a possibility that students will work on a thesis with the same title and this action is called student thesis plagiarism.

Several studies discussing the prevention of plagiarism by utilizing information technology have been carried out. A study proposes an anti-plagiarism framework in higher education where the scope is stakeholder awareness of plagiarism, forming policies against plagiarism, and implementing plagiarism monitoring procedures (Ernawati, Nugroho, & Atmojo, 2014). In other studies, the application of the Rabin-Karp algorithm is considered capable of detecting plagiarism in the text so that it can be used to detect similarities in test samples taken from the final project proposal (Ramadhani, 2015).

METHOD

The main reason why cosine similarity is used is it measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis. In this research, the cosine similarity method will be combined with the pre-processing method and TF-IDF to measure the level of similarity of students' final assignments based on their titles and abstracts. Furthermore, it is displayed and compared with the available final project data based on a threshold value so that a decision can be made whether the student's scientific work can be accepted or rejected. Fig 1. Shows the application of plagiarism detection using cosine similarity:

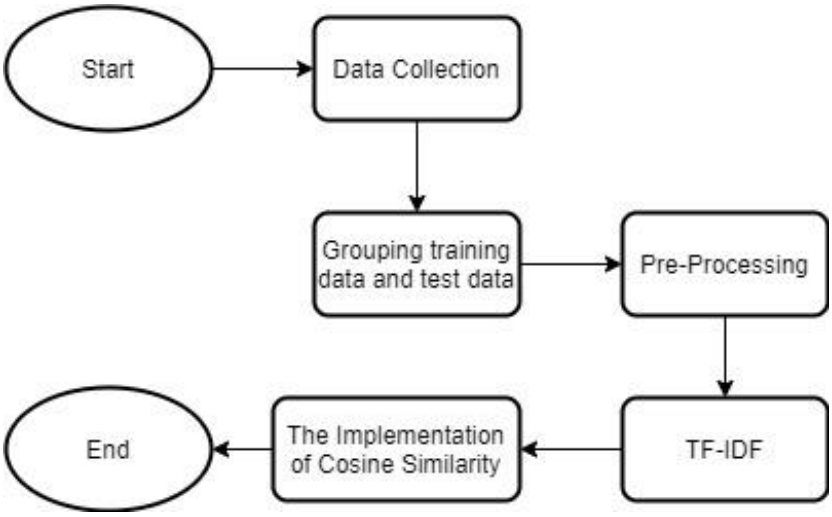

Fig. 1 Research Framework

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data collection was done by collecting data on the final project of STIKI Malang students which were then grouped into two parts. Data grouping is done based on training data and test data that will be used to make comparisons. There are 3 training data and 1 test data. The training data and test data used as examples of cases below are obtained based on sources from university libraries and google scholar. Table 1 and 2 shows the training data and test data that will be used.

Table 1.
Training Data

NO	TITLE	ABSTRACT
1	Sistem Pemesanan Online Guna Meningkatkan Pelayanan Pelanggan Pada UD STIKI	Untuk mengatasi permasalahan tersebut maka dengan mendirikan situs pemesanan online yang dapat menambah fleksibilitas dan kemudahan proses order dari pelanggan serta bentuk media untuk menyampaikan informasu produk perusahaan . Dengan adanya pengembangan situs pemesanan online ini maka lebih mempermudah pelanggan untuk melakukan order ataupu mendapatkan informasi seputar produk-produk yang ditawarkan oleh UD.Sukses - Jaya.
2	Sistem Informasi Pemantauan Kinerja Sales Berbasis Android	Tugas utama sales adalah melakukan pemasaran produk, dalam melaksanakan pekerjaannya, sales pada setiap harinya mendatangi toko-toko yang sudah terjadwal untuk dikunjungi, namun sebagian dari sales tidak mendatangi toko dan melakukan laporan palsu. Dengan memanfaatkan fitur GPS yang berada pada smartphone, dapat diketahui koordinat dari smartphone tersebut, data koordinat tersebut dapat dimanfaatkan untuk melakukan validasi laporan yang dilakukan sales dalam hal kunjungan toko. Hasil dari pembuatan sistem informasi pemantauan kinerja sales berbasis android ini adalah sebuah sistem yang dapat memantau pekerjaan yang dilakukan oleh sales, memberikan laporan hasil kerja sales kepada supervisor, dan memudahkan sales dalam melakukan pelaporan hasil kerja hariannya. Dengan menggunakan sistem informasi pemantauan kinerja sales ini pelaksanaan pemantauan tidak lagi harus melakukan inspeksi mendadak.
3	Pembuatan Sistem Informasi Pengguna Area Parkir Stiki Dengan Memanfaatkan QR Code	Dalam bidang keamanan, salah satu masalah penting adalah meningkatkan keamanan pada sistem parkir kendaraan bermotor. Semakin meningkatnya tingkat kejahatan kendaraan bermotor di Indonesia, dan modus kejahatan ini tidak hanya terjadi pada kendaraan bermotor yang terletak di luar area parkir. Banyak ditemukan kasus pencurian kendaraan bermotor ini di areal parkir. Di kota Malang untuk tahun 2015, mulai Januari–Mei, tercatat 539 kasus curanmor yang dilaporkan masyarakat. STIKI merupakan salah satu kampus di kota Malang, dan memiliki area parkir yang digunakan untuk kendaraan bermotor khususnya roda dua. Banyak akses jalan masuk orang, sehingga memungkinkan terjadi pencurian kendaraan yang parkir di dalam area kampus. Peluang curanmor semakin besar karena pada pos parkir, petugas masih menggunakan kartu untuk masuk dan keluar parkir

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

yang tidak dilengkapi dengan data pengguna maupun data kendaraan. Dengan kondisi tersebut, peneliti akan membuat sebuah sistem informasi berbasis desktop untuk pencatatan pengguna dan menggunakan QR code atau singkatan dari Quick Response Code sebagai kartu identitas yang digunakan untuk membantu petugas parkir mengecek semua kendaraan yang akan masuk atau keluar di area parkir saat ini. QR code merupakan sebuah marker pola dari pengembangan barcode yang digunakan sebagai penanda nomor.

Pre-Processing

Preprocessing is the initial process applied to text data that aims to generate numerical data. The preprocessing process is a stage where the description is handled until it is ready to be processed entering the text mining stage (Wahyuni, Prastiyanto, & Supraptono, 2017). After all data has been collected, the next step is to preprocess with the following steps (Purwiantono & Aditya, 2020):

1. Data Integration
   Data integration is the stage for combining data from several sources into one file (Prianto & Bunyamin, 2020). In this study, data integration was used to combine training data obtained from the results of scrapping the Google Scholar API (Application Programming Interface) into 1 file in CSV format. The resulting file is a BibTex file which is then converted into JSON format and displayed as csv to generate data in spreadsheet format.
2. Case Folding
   Case folding is a mechanism for transforming uppercase letters into lowercase letters (Wahyuni et al., 2017). In this process, case folding is used to convert all letters in the training document and test documents to lowercase
3. Tokenizing
   Tokenizing is the process of dividing data into several tokens or words (Mahfud, Mudawamah, & Hariyanto, 2020). In this process, the data generated from the case folding process will be parsed into a collection of words. Furthermore, certain characters such as punctuation or hyphens will be removed
4. Stopword/Filtering
   Filtering is used to remove meaningless characters or words. Based on the results obtained from the tokenization process, the data is reprocessed into a stopword / filtering process. This process is used to eliminate non-descriptive words contained in the document.
5. Stemming
   Stemming is the process of finding the root word of each filtered word (Francis & Flynn, 2010). The data generated from the stopword / filtering process is converted into the root word that will be used to perform TF-IDF calculations

Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency and Inverse Document Frequency are weighting methods used in information retrieval and text mining (Turney, 2010). Term Frequency is a simple weighting method where the importance of a word is considered equal or proportional to the number of occurrences of the word in the document, while Inverse Document Frequency is a weighting method that measures the urgency of a word in a document seen in documenting globally.

Cosine Similarity

Cosine similarity is a method used to calculate the degree of similarity between two objects. Generally, the calculation of this method is based on the size of the similarity of the vector space. This method calculates the similarity between two objects (for example D1 and D2) expressed in two vectors using keywords from the document as benchmarks (Pradnyana & ER, 2012).

RESULT

The scrapping technique on Google Scholar was carried out to obtain valid data that was used as training data. Next, the preprocessing stage will be carried out where the results will be processed using the TF-IDF method to

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

get the terms that will be used to calculate document similarities using the cosine similarity method. Table 3 shows the calculation of the cosine similarity method in this study.

Table 2.
Cosine Similarity Calculation Result

TERM	D1
android	0
area	0
basis	0
informasi	0
layan	1
manfaat	0
online	1
pantau	0
parkir	0
pelanggan	1
pesan	0
qr	0
sales	0
sistem	1
stiki	1
tingkat	1
ud	1
TF	**7**
IDF	0,0579919
TF*IDF	0,4059436
Vector Q -> D	2,0102119
Vector Q	2,6457513
SIMILARITY	0,537251722

Table 3.
Similarity Test Based on Title

TERM	**D1**
android	0
area	0
Basis	0
informasi	0
Layan	0
manfaat	0
Online	0
pantau	0
parkir	0
pelanggan	0
Pesan	0
qr	0
Sales	0
Sistem	1
stiki	1
tingkat	0
Ud	0
TF	2
IDF	0,1760913

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

***TF**IDF**	0,3521825
Vector Q -> D	2,0310081
Vector Q	1,4142136
SIMILARITY	0,089758726 or 8%

$$\frac{\sum_{d}^{i}=_1(q_{id}.d_{id})}{\sqrt{\sum_{d}^{t}=_1(q_{id})^2}.\sqrt{\sum_{d}^{t}=_1(d_{id})^2}} \quad (1)$$

$q_{\rightarrow}d$ $= (q_{term1} * d_{term1}) + (q_{term2} * d_{term2}) + (q_{term3} * d_{term3}) *+ \ldots$ (2)

$q_{\rightarrow}d$ $= (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 1) + (0 * 0) +$
$(0 * 1) + (0 * 0) + (0 * 0) + (0 * 1) + (0 * 0) + (0 * 0) +$
$(0 * 0) + (1 * 1) + (1 * 1) + (0 * 1) + (0 * 1)$
$= 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 1 + 0 + 0$
$= 2$
$= 2{,}031008$

$$Q = \sqrt{\begin{matrix}(0*0)+(0*0)+(0*0)+(0*0)+\\(1*1)+(0*0)+(1*1)+\\(0*0)+(0*0)+(1*1)+\\(0*0)+(0*0)+(0*0)+\\(1*1)+(1*1)+(1*1)+(1*1)\end{matrix}}$$

$$\sqrt{\begin{matrix}0+0+0+0+1+0+1+0+0+1+0\\+0+0+1+1+1+1\end{matrix}}$$
$= \sqrt{7}$
$= 2{,}6457513$

$$Sim = \frac{2{,}0102119}{2{,}6457513 * 1{,}4142136}$$

$= 0{,}537251722$

Description:
q = Test Document Data
d = Training Document Data
i $= tf_{ij} \cdot idf_j$
t = Words/Term
$q_{\rightarrow}d$ = Distance between q and d
$\sum_{d}^{t}$ = Sigma Notation on t and d

DISCUSSIONS

The design of a plagiarism detection system is divided into several features with their respective functions. Figure 2 shows the main application page interface which is the login page. This login page is used by the user to log into the system.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

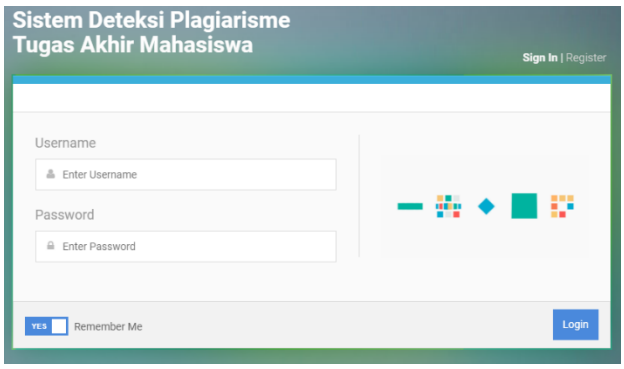

Fig. 2 Login Page

After the user has successfully logged into the system, the user will be automatically directed to the start menu page. The homepage will display a list of theses that have been submitted by students as well as thesis data that have been received and validated by the Head of Study Program and BAA (Academic Administration Bureau) STIKI Malang.

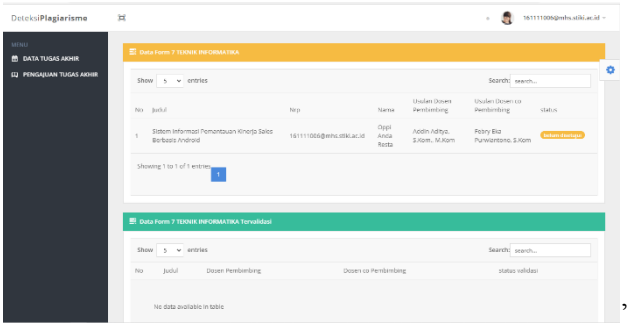

Fig. 3 Dashboard Menu

Figure 4 explains that students input data into the input-form provided by the system then make submissions by pressing the save button. After the data is successfully saved, the plagiarism check button will automatically appear.

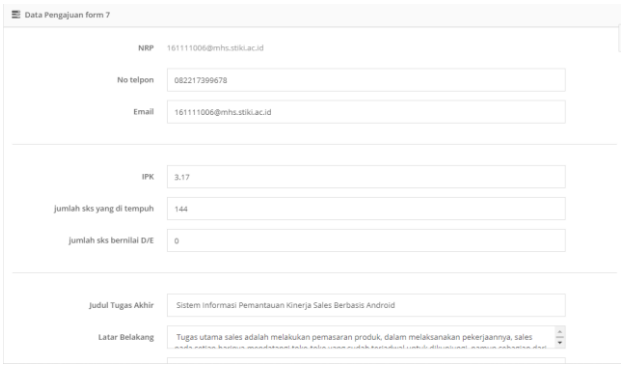

Fig. 4 Submission of student thesis

After students press the check plagiarism button, they will be directed to the student detail page. On this page, the system will perform calculations on the final project data or what is called test data and then displays the calculation results on that page. The system will automatically perform calculations by comparing the test

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 2, April 2021
DOI : https://doi.org/10.33395/sinkron.v5i2.10909
e-ISSN : 2541-2019
p-ISSN : 2541-044X

documents with each training document to determine the similarity level of the submitted documents. After the level of data similarity in the final project is detected, the results will be displayed and sorted according to the document similarity ranking percentage. Details of the final project data that have been submitted will also be displayed here (see figure 5). f the similarity of documents is more than 80%, the university has the right to refuse the final project that has been submitted. Conversely, if the university receives the submitted thesis data, it will automatically update the status of the submitted thesis data on the thesis data page to be 'accepted'.

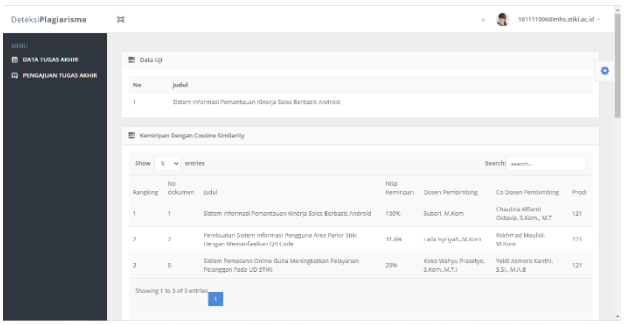

Fig. 5 Plagiarism Result using Cosine Similarity

To calculate the submitted thesis data or what is called test data, comparative data or training data are needed to determine the similarity with the submitted documents. The training data is filled in by the university and then displayed on the training data list page. Figure 6 shows the application not only displays a list of training data, but universities can also process training data, where the data is processed and converted into frequency terms.

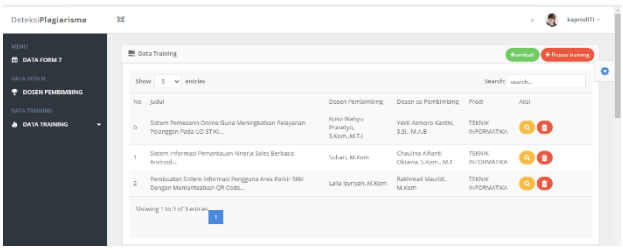

Fig. 6 List of Training Data

Figure 7 shows the frequency term list page where the training data input is processed using preprocessing and TF-IDF to produce terms from each training data document. The terms generated from the preprocessing process and TF-IDF are used to make comparisons with the submitted test data documents, which later on the system will perform calculations and get the results of the similarities of each document.

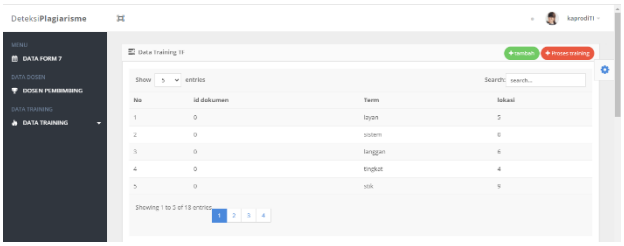

Fig. 7 Term Frequencies

CONCLUSION

The results of this study can make it easier for students and the university to detect the similarities between each thesis submitted by students and help the university, especially the faculty, in making decisions whether the proposed thesis meets the requirements or not. However, this system is limited because it only uses the dataset from Google Scholar, where the dataset is quite limited and has a data scrap mechanism that is quite complicated

and takes a long time. Further research can be carried out to discuss the possibility of using a dataset that is more concise and has an easy data scrapping mechanism. This study also suggests adding methods to improve the accuracy of system performance so that when the process is executed the system can work optimally.

REFERENCES

Biliæ-zulle, L., Frkoviæ, V., Turk, T., & Josip, A. (2005). Prevalence of Plagiarism among Medical Students. *Croat Med J.*, *46*(1), 126–131.

Ernawati, E., Nugroho, R., & Atmojo, P. (2014). SISTEM PENDETEKSI PLAGIARISME UNTUK TUGAS AKHIR MAHASISWA DI UNIVERSITAS BINA NUSANTARA : *Jurnal Humaniora*, *5*(1), 541–549.

Francis, L., & Flynn, M. (2010). *Text Mining Handbook*.

Kurniasar. (2016). Upaya pencegahan dan penanggulangan plagiarisme di perguruan tinggi. *Jurnal Bhinneka Tunggal Ika*, *3*(2), 125–134.

Mahfud, F. K. R., Mudawamah, N. S., & Hariyanto, W. (2020). Sentiment Analysis of Perpustakaan Nasional Republik Indonesia Through Social Media Twitter. *Matics: Jurnal Ilmu Komputer Dan Teknologi Informasi*, *12*(1), 90–93. https://doi.org/10.18860/mat.v12i1.8973

Permendiknas RI No 17 Tahun 2010. (2010). Kementerian Pendidikan Nasional.

Pradnyana, G. A., & ER, N. A. S. (2012). PERANCANGAN DAN IMPLEMENTASI AUTOMATED DOCUMENT INTEGRATION DENGAN MENGGUNAKAN ALGORITMA COMPLETE LINKAGE AGGLOMERATIVE HIERARCHICAL CLUSTERING. *Jurnal Ilmu Komputer*, *5*(2), 1–10.

Prianto, C., & Bunyamin, S. (2020). *Pembuatan Aplikasi Clustering Gangguan Jaringan Menggunakan Metode K-Means Clustering* (1st ed.). Bandung: Kreatif Industri Nusantara.

Prihantini, F. N., & Indudewi, D. (2016). Kesadaran dan Perilaku Plagiarisme dikalangan Mahasiswa (Studi pada Mahasiswa Fakultas Ekonomi Jurusan Akuntansi Universitas Semarang). *Jurnal Dinamika Sosial Budaya*, *18*(1), 68–75.

Purwiantono, F. E., & Aditya, A. (2020). KLASIFIKASI SENTIMEN SARA, HOAKS DAN RADIKAL PADA POSTINGAN MEDIA SOSIAL MENGGUNAKAN ALGORITMA NAIVE BAYES MULTINOMIAL TEXT. *Jurnal TeknoKompak*, *14*(2), 68–73.

Ramadhani, S. (2015). Sistem Pencegahan Plagiarism Tugas Akhir Menggunakan Algoritma Rabin-Karp (Studi Kasus : Sekolah Tinggi Teknik Payakumbuh). *Jurnal Teknologi Informasi & Komunikasi Digital ZOne*, *6*(1), 44–52.

Ryan, G., Bonanno, H., Krass, I., Scouller, K., & Smith, L. (2009). Undergraduate and Postgraduate Pharmacy Students ' Perceptions of Plagiarism and Academic Honesty. *American Journal of Pharmaceutical Education*, *73*(6), 1–8.

Turney, P. D. (2010). From Frequency to Meaning : Vector Space Models of Semantics From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research2*, *37*(March 2010), 141–188. https://doi.org/10.1613/jair.2934

Wahyuni, R. T., Prastiyanto, D., & Supraptono, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro*, *9*(1), 18–23.