J-COSINE (Journal of Computer Science and Informatics Engineering)
Vol. 6, No. 2, December 2022 E-ISSN:2541-0806
Accredited Sinta-3 by RISTEKDIKTI Decree No. 28/E/KPT/2019 P-ISSN:2540-8895

Term Weighting Based Indexing Class and Indexing Short Document for Indonesian Thesis Title Classification

Ana Tsalitsatun Ni’mah[1]*, Fahmi Syuhada[2]
[1] Informatics education, Faculty of Education, Universitas Trunojoyo Madura
Jl. Raya Telang, Telang Indah Housing, Telang, Kec. Kamal, Bangkalan, East Java 69162

[2] Computer Science, Faculty of Science and Technology, Universitas Qamarul Huda Badaruddin Bagu, West Nusa Tenggara
Turmuzi Badrudin, Bagu, Praya, Central Lombok, West Nusa Tenggara 83371
Email: ana.tsalits@trunojoyo.ac.id [1], fahmisy@uniqhba.ac.id [2]

Corresponding Author

Abstract. **Document classification nowadays is an easy thing to do because there are the latest methods to get maximum results. Document classification using the term weighting TF-IDF-ICF method has been widely studied. Documents used in this research generally use large documents. If the term weighting TF-IDF method is used in a short text document such as the Thesis Title, the document will not get a perfect score from the classification results. Because in the IDF will calculate the weight of words that always appear to be few, ICF will calculate the weight of words that often appear in the class to be few. While the word should have great weight to be the core of a short text document. Therefore, this study aims to conduct research on word weighting based on class indexation and short document indexation, namely TF-IDF-ICF-IDSF. This study uses a classification comparison Naïve Bayes and SVM. The dataset used is Thesis Title of Informatics Education student at Trunojoyo Madura University. The test results show that the classification results using the TF-IDF-ICF-IDSF term weighting method outperform other term weighting, namely getting 91% Precision, 93% Recall, 86% F1-Score, and 84% Accuracy on SVM.**

Key words*: *TF-IDF, TF-IDF-ICF, Term Weighting, Text Classification, Short Document *(not more than 5 words or word-phrases in the keywords for indexing the paper)*

I. Introduction

The number of documents today is abundant and very accessible for use as research material on text processing [1]. The development of text processing science continues to be carried out to obtain the best methods in order to achieve results that suit the needs [2]. Text processing requires text documents as materials to be studied. The changing times have made a lot of data in the form of documents presented on the internet [3]. These documents can be easily obtained with one click [4]. Documents have 2 types, namely documents with a lot of content or commonly called long text, and documents with little content called short text [5]. A long document is usually a news article or scientific article, it can also be a book [6]. Short documents are usually comments on social media, product reviews, and the title of a research or thesis [7]. Currently, the most commonly found are short documents or short text. Short text processing is a separate branch of science for finding new methods in its processing [8]. short texts do not have sufficient contextual information, which is part of the challenge in their classification [9]. Short texts require a special method for the classification process as well as in the weighting of the words [10]. Short text has different characteristics from Long text. Long text can generally be weighted using the usual term weighting method such as TF-IDF [11]. TF-IDF (Term Frequency Inverse Document Frequency) is the most widely used weighting scheme for keywords to facilitate the relevance of each document. [12]. The TF-IDF puts forward the number of words in the document, if the word often appears in many documents, the statistics for the word will be low [13]. TF-IDF developed again into TF-IDF-ICF where the method pays more attention to the weight of words in a class. The TF-IDF-ICF term weighting method is still not appropriate when applied to short texts [14]. So there is a need for research to find a term weighting method for developing TF-IDF-ICF specifically for short texts [15]. This study aims to analyze the term weighting based on indexing class and indexing short document for Indonesian thesis title classification which will be tested by comparing the classification results using Naïve Bayes and SVM [16]. The dataset used is the title of the thesis in the Informatics Education Study Program, Faculty of Education, Trunojoyo University, Madura in 2016 – 2021.

II. Related Works

Previous research on term weighting has been conducted in 2013 by Fuji Ren [15]. This study used the basic concept of TF IDF to conduct further research which resulted in several term weighting methods, namely TF-IDF-ICF and TF-IDF, ICSdF. This word weighting method has undergone many redevelopments, one of which is regarding the adaptation of gravity moment [17] in 2019.

This article describes the adaptation of the word weighting method by inserting the concept of gravity moment in the weighting. However, all of these studies were only applied to the Panjang document. There has been no research on the application of this TF-IDF word weighting method for short documents [18]. Previous research on short documents did not use this term weighting. Whereas the use of term weighting also has an impact on the results of the classification of short documents. This study also uses a new method of stemming. This new method is called Indonesian stemmer reconstruction (Fig. 1) [19]. The stemming process in this method does not go through checking the Indonesian dictionary like the previous method. Stemming has several advantages that have an impact on the results of the classification later. Because cleaning the data that will be used in text processing will also have an impact on the accuracy of the results later.

III. Method

A. Text Preprocessing

TABLE I. Stopword Example

Stopword
Lalu
Kemudian
Apa
Entah
Mungkin
Seandainya
Dan
Itu
Ini

Text preprocessing is a series of steps to select text data to be more structured again by going through the stages of case folding, tokenizing, filtering (Fig. 2) and stemming. there are no definite rules about each stage in text preprocessing, it depends on the type and condition of the data to be processed. Text preprocessing is an implementation of text mining. Text mining itself is a data mining activity, where data is usually taken in the form of text sourced from documents that have the final result to find keywords that represent a set of documents so that later analysis of the relationship between these documents can be carried out. The first stage that is usually done is the case folding stage. This step is almost always included when doing text preprocessing. Let's take an example is the tweet data or a dataset of spam messages that must consist of sentences. The way to make the data analysis process easier, we have to break these sentences into words or called tokens. tokenizing can distinguish between word separators or not. The continuation of the tokenizing (Fig. 3) stage is the filtering stage which is used to retrieve important words from the token results. Common words that usually appear and have no meaning are called stop words. For example, the use of conjunctions such as “dan”, “mana”, “sebelum”, “iya”, and “sedangkan” (Table I). Removing this stopword can reduce index size and processing time. However, stopping does not always increase the retrieval value. Careless construction of a stopword list (called a stoplist) can worsen the performance of an Information Retrieval (IR) system. There is no definite conclusion that the use of stopping will always increase the retrieval value, because in several studies, the results obtained tend to vary.

The stemming stage is a step that is also needed to reduce the number of different indexes from one data so that a word that has a suffix or prefix will return to its basic form. Stemming also performs grouping of other words that have the same basic words and meanings but have different forms because they get different affixes. This research uses a new stemming method for Indonesian, namely Indonesian stemmer reconstruction which in the process does not go through a dictionary check (Fig. 3). In the previous stemming method, the affix removal process used ECS. ECS performs the morphological rules that exist in Indonesian. The order of the stemming process is to check the dictionary first, if the word is in the dictionary, the process will stop. However, if the word is not in the dictionary, then the affix removal process will be carried out.

Deletion of affixes begins by removing particles (-kah, -lah, -tah, -pun). Then it is matched back into the dictionary, if the word is already in the dictionary, the process of removing the affixes stops. However, if it is not in the dictionary then the process of removing the affix will continue. The next process of deleting affixes is deleting the affixes belonging to (-ku, -mu, -nya). The next process is checking the dictionary as in the previous process. The next process is checking for illegal affixes or affixes that are prohibited by standard Indonesian grammar. Then the process of removing suffixes and prefixes is carried out. Until the appropriate basic words in the Indonesian basic word dictionary are found in the system. The stemming process using this algorithm is still very dependent on the basic Indonesian dictionary, so it takes a long time to process. Because this process applies a word-for-word algorithm. However, in research related to Indonesian stemming, this technique is the technique that produces the most accurate basic words because the final result of this stemming will always be checked against the dictionary first.

B. TF-IDF

TF-IDF (Term Frequency Inverse Document Frequency) is the most basic word weighting method. This method has been widely used in many studies, especially in research on text processing. TF-IDF is a word weight calculation that takes into account the word weight in each document, namely TF, and takes into account the word weight in the entire document, namely IDF (Fig. 4). TF only calculates word weights in each document, while IDF pays more attention to word weights in all documents. The two values are combined which will produce a word weight value that pays attention to its weight in each document and also pays attention to its weight in all documents. The TF-IDF formula is described in Eq.1.

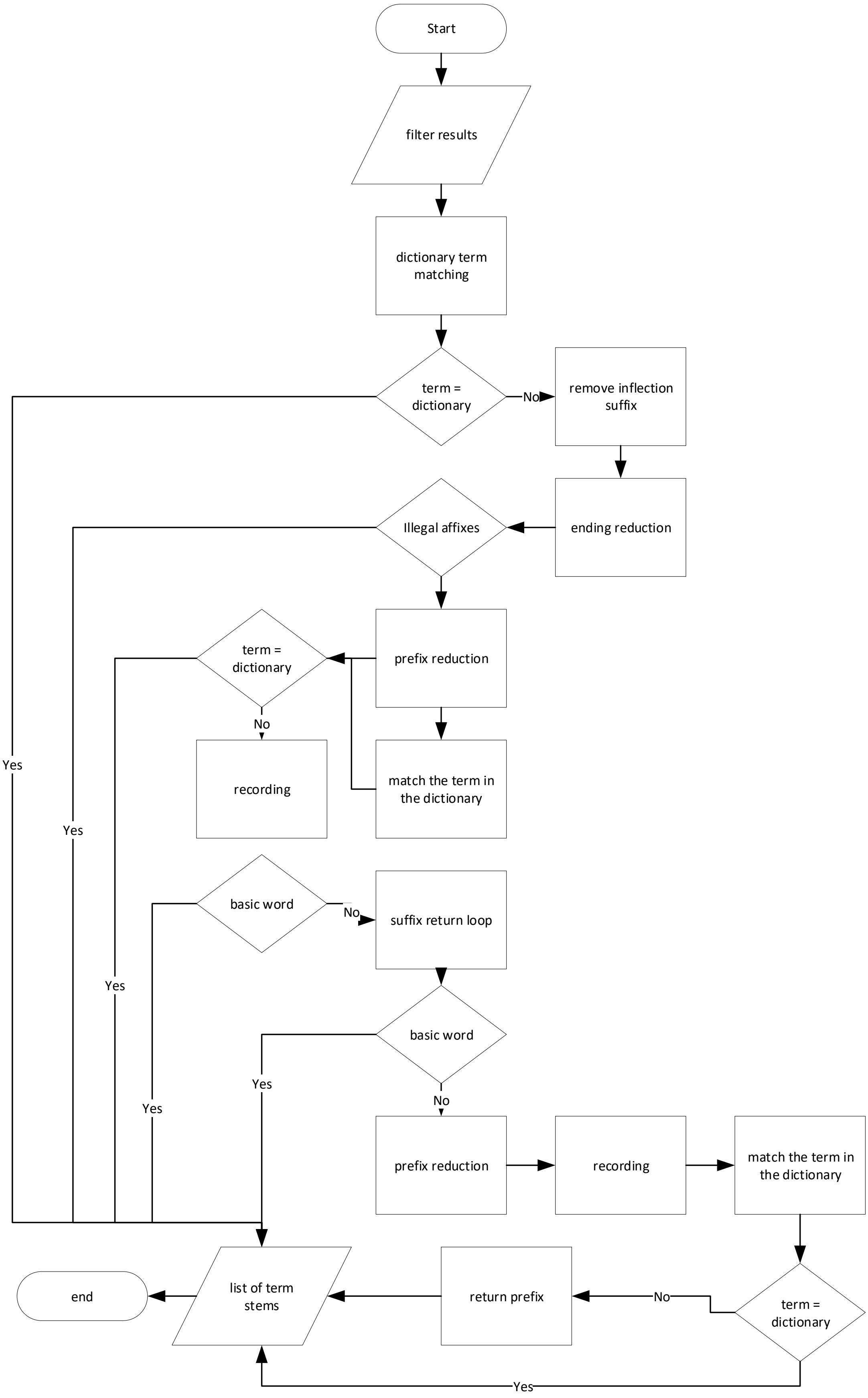

Fig. 1. Flowchart Indonesian Stemmer Reconstruction

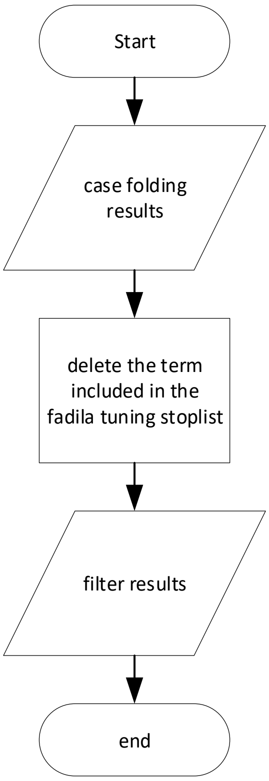

Fig. 2. Filtering

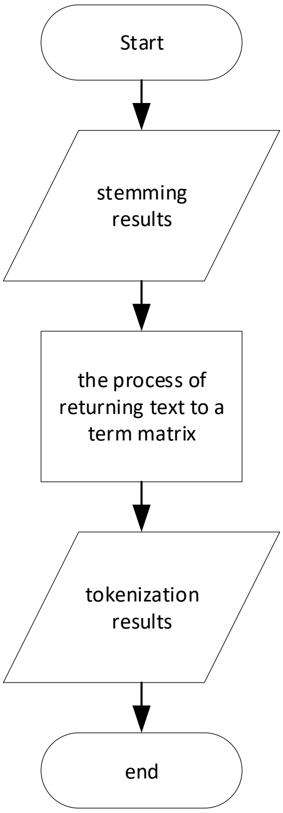

Fig. 3. Tokenization

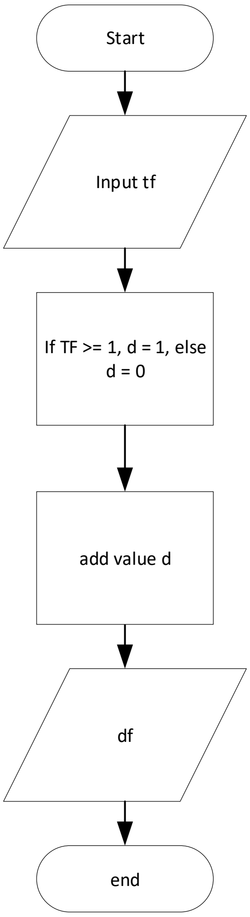

Fig. 4. IDF

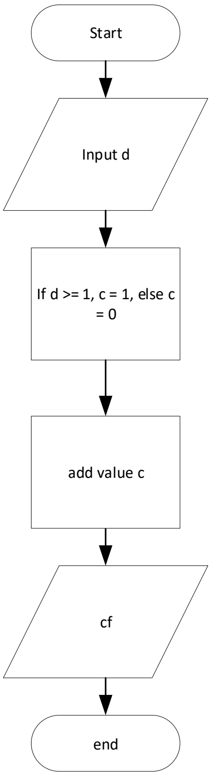

Fig. 5. ICF

$$TF.IDF\ (t,d,D) = TF(t,D)x\ IDF(t,D) \qquad (1)$$

where TF is Term Frequency, IDF is Inverse Document Frequency.

TF IDF ICF is the development of the IDF TF term weighting method. This method adds attention to the weight of words in each class. The TF-IDF can be used for unsupervised documents that do not use classes in each document, the TF-IDF-ICF must use supervised documents that already have classes in each document. The addition of ICF (Fig. 5) makes the TF-IDF-ICF formula pay attention to word weights in all classes, not only in all documents. The equation of TF-IDF-ICF can be seen in Eq. 2.

$$TF.IDF.ICF(t,d,D,C) = \\ TF(t,D)x\ IDF(t,D)x\ ICF(t,D,C) \qquad (2)$$

where TF is Term Frequency, IDF is Inverse Document Frequency, ICF is Inverse Class Frequency.

C. TF-IDF-ICF-IDSF

TF IDF ICF pays attention to the class of word occurrences, then if the document is short, an adaptation of a new formula is needed to overcome it. Because if you only pay attention to the appearance of words in the document and class, then the words that should be the core of the short text document will get a minimum value, even though these words are important in the thesis title for the classification process. Therefore, this research was conducted to develop TF-IDF-ICF that can adapt to short text. The formula (Eq. 3) is found with the assumption that when added IDSF will increase the weight of important words from a thesis title when weighting words. The appearance of equation must be shown as presented in Eq. 3.

$$TF.IDF.ICF.IDSF(t,d,D,C,DS) = \\ TF(t,D)x\ IDF(t,D)x\ ICF\ (t,C)x\ IDSF(t,DS) \quad (3)$$

where TF is Term Frequency, IDF is Inverse Document Frequency, ICF is Inverse Class Frequency, IDSF is Inverse Document Short Frequency

$$IDSF(t,d,D,C,DS) = \log(D/DSF) \qquad (4)$$

where D is Number of Documents, DSF is number of selected word frequencies. DSF calculates the word weight by taking into account the number of occurrences of all documents. this is different from DF which pays attention to the many occurrences of words in many documents. DSF takes into account more if the word appears frequently in several documents by doing an inverse comparison to increase the weight of the word that appears frequently. Because the short text in this case is the title of the thesis, the words that appear frequently have a major role to characterize as material for classification.

D. Confusion Matrix

Confusion Matrix is a method to measure classification ability that can be used for at least 2 classes. Confusion Matrix generally uses a table with 4 combinations containing four terms which describe the results of the classification process, namely True Negative, True Positive, False Negative, and False Positive.

TABLE II. Thesis Title Example

Title
Pengembangan media pembelajaran sistem bilangan menggunakan augmented reality berbasis android untuk smk
Pengaruh media pembelajaran e-learning berbasis moodle terhadap hasil belajar siswa pada materi javascript kelas x tkj smk negeri 1 labang
Profil berpikir kritis siswa kelas x tkj ditinjau dari kemampuan awal siswa materi struktur kontrol percabangan di smkn 3 bangkalan
Pengembangan game edukasi 3d struktur algoritma pemrograman menggunakan unity berbasis android sebagai media pembelajaran pemrograman dasar di smk bahrul ulum surabaya
Pengembangan sistem pencarian informasi buku berbasis web menggunakan moving contracting window pattern algorithm
Pengembangan media pembelajaran trainer komputer dalam mata pelajaran produktif materi perakitan komputer di smkn 1 sepulu
Pengembangan game educative berbasis android untuk pada mata pelajaran matematika materi bangun ruang untuk siswa sekolah dasar
Pengembangan media pembelajaran e-learning berbasis web untuk siswa kelas x teknik komputer dan jaringan
Pengaruh penerapan sublime text 3 terhadap minat belajar siswa pada mata pelajaran pemrograman web bahasan style halaman web kelas x tjk di smk al – hikam
Pengembangan media pembelajaran perakitan komputer berbasis augmented reality
Pengaruh metode simulasi berbantu media vmware terhadap hasil belajar pada materi instalasi sistem operasi linux di kelas x smkn 1 labang
Aplikasi kamus pemrograman mobile menggunakan metode term frequency inverse document frequency (tf-idf) untuk smartphone android
Pengembangan media pembelajaran game edukatif berbasis rpg pada materi algoritma pemrograman di smk nurul amanah
Pengembangan media pembelajaran game edukasi berbasis flash pada mata pelajaran merakit komputer untuk kelas x tkj di smkn 1 kamal
Pengaruh model pembelajaran berbasis proyek terhadap hasil belajar siswa kelas xi pada mata pelajaran keterampilan komputer dan pengelolaan informasi di smk siding puri sumenep
Pengembangan sistem informasi monitoring siswa berbasis web untuk memonitoring kegiatan siswa di smk negeri 1 kamal
Pengembangan media simulasi instalasi sistem operasi menggunakan unity di smk al-hikam burneh
Pengaruh model problem based learning terhadap motivasi belajar siswa kelas x tkj pada materi pengantar subnetting jaringan komputer di smk agung mulia
Pengembangan media e-book menggunakan flipbook maker berbasis web pada mata pelajaran perakitan komputer kelas x tkj di smk agung mulia bangkalan
Efektivitas model project based learning pada materi cascading style sheet (css) kelas x
Pengembangan aplikasi m-learning berbasis android menggunakan unity pada materi jaringan lan untuk siswa kelas xi tkj di smkn 2 lamongan.
Pengembangan sistem informasi perpustakaan berbasis web menggunakan barcode scanner di man bangkalan
Pengaruh pembelajaran menggunakan media e-learning berbasis web (gnomio) terhadap hasil belajar siswa pada mata pelajaran perakitan komputer di kelas x rpl smkn 2 bangkalan
Pengaruh metode pembelajaran artikulasi terhadap hasil belajar siswa pada mata pelajaran jaringan dasar kelas x rpl di smk negeri 2 bangkalan

Title
Pengaruh model pembelajaran discovery learning terhadap hasil belajar sistem komputer siswa kelas x smk al-hikam
Pengaruh model pembelajaran problem based learning terhadap hasil belajar siswa pada mata pelajaran perakitan komputer di smkn 2 bangkalan
Pengembangan media pembelajaran berbasis augmented reality pada materi pengenalan komponen elektronika di kelas x jurusan tkj di smks ibnu cholil bangkalan
Pengembangan sistem informasi absensi siswa berbasis java desktop di sma darul kholil bangkalan
Pengaruh model pembelajaran kooperatif metode time token pada materi struktur sistem operasi open source terhadap hasil belajar siswa
Pengaruh metode pembelajaran probing promting terhadap hasil belajar kognitif siswa kelas x tki pada materi instalasi sistem operasi open-source di smkn 1 sumenep.
Pengaruh pembelajaran menggunakan media vmware workstation terhadap hasil belajar siswa materi pelajaran instalasi sistem operasi smk al-hikam bangkalan.
Pengaruh model pembelajaran cooperative learning tipe snowball throwing terhadap hasil belajar sistem operasi siswa kelas x smkn 1 kamal
Efektivitas model pembelajaran project based learning pada mata pelajaran pemrograman dasar di smkn 1 sumenep
Pengaruh media aplikasi virtualbox terhadap hasil belajar kognitif siswa pada materi proxy server kelas xi tkj smk alhikam.
Pengaruh model pembelajaran problem based learning (pbl) terhadap hasil belajar siswa pada materi instalasi program aplikasi di smkn 1 kamal
Proses berpikir kreatif siswa pada materi perkembangan sistem operasi ditinjau dari kemampuan awal siswa
Pengaruh metode problem solving menggunakan media pembelajaran cbi (computer based instruction) terhadap hasil belajar desain grafis di smk agung mulia
Penerapan sistem informasi manajemen berkas berbasis web menggunakan owncloud di smk negeri 1 kamal

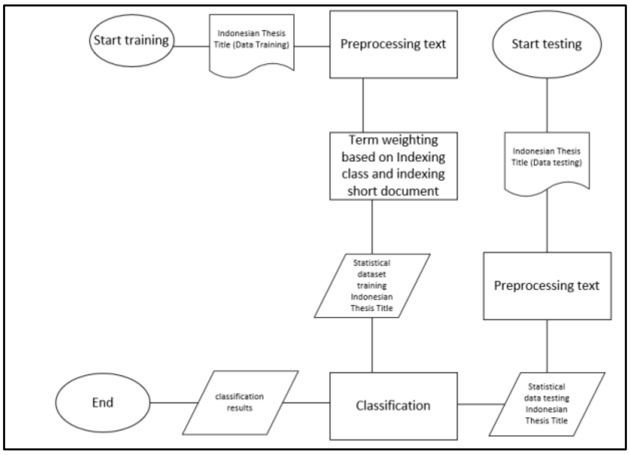
Start training
Indonesian Thesis Title (Data Training)
Preprocessing text
Start testing
Term weighting based on Indexing class and indexing short document
Indonesian Thesis Title (Data testing)
Statistical dataset training Indonesian Thesis Title
Preprocessing text
End
classification results
Classification
Statistical data testing Indonesian Thesis Title

Fig. 6. System Process

E. *Naïve Bayes*

Naive Bayes is a classification method that is generally used for binary and multiclass data [20]. This method applies supervised classification, which is to determine the class label first to the record using conditional probabilities.

F. *SVM*

Support Vector Machine (SVM) is a classification method that is often used in supervised learning. SVM has advantages over the previous classification. This classification is more stable and gets more accurate results.

IV. Result and Discussion

The data used in this study is the Indonesian thesis title data taken from the thesis of the Informatics Education student at Trunojoyo Madura University. The training data is taken from the title of the alumni thesis for 2016 – 2020. The testing data uses the title of the student thesis for 2021 (Table II). the research process using text data tends to be more complicated and takes a lot of time. This is indicated by the number of steps that must be taken.

TABLE III. Naïve Bayes Classifier

Confusion Matrix	TF-IDF	TF-IDF-ICF	TF-IDF-ICF-IDSF
Precission	0,85	0,93	**0,93**
Recall	0,75	**0,96**	0,92
F1-Score	0,78	0,94	**0,95**
Accuracy	0,83	0,86	**0,92**

TABLE IV. SVM Classifier

Confusion Matrix	TF-IDF	TF-IDF-ICF	TF-IDF-ICF-IDSF
Precission	0,81	0,83	**0,91**
Recall	0,66	0,71	**0,93**
F1-Score	0,71	0,75	**0,86**
Accuracy	0,77	0,83	**0,84**

Text pre-processing uses several existing methods. In the steming section, this research uses a new method, namely Indonesian steming which does not use a dictionary. The next process is the calculation of word weights using the term weighting method, namely TF-IDF, TF-IDF.ICF, and the proposed method is TF-IDF-ICF-IDSF. This research first processes more training data and then enters testing data (Fig. 6). The test results are presented in Tables 1 and 2. The table summarizes how large the comparison of the classification results from the 3 weighting terms tested in this study. From the results of the summary, it is stated that the TF-IDF method with the Naive Bayes classification engine obtains results, namely Precision of 85%, Recall of 75%, F1-Score of 78%, and Accuracy of 83%. TF-IDF-ICF with the Naive Bayes classification engine got the results, namely Precision of 93%, Recall of 96%, F1-Score of 94%, and Accuracy of 86%. TF-IDF-ICF-IDSF with the Naive Bayes classification engine gets the results. Precision is 93%, Recall is 92%, F1-Score is 95%, and Accuracy is 92%. TF-IDF with the SVM classification engine got the results, namely Precision of 81%, Recall of 66%, F1-Score of 71%, and Accuracy of 77%. TF-IDF-ICF with the Naive Bayes classification engine got the results, namely Precision of 83%, Recall of 71%, F1-Score of 75%, and Accuracy of 83%. TF-IDF-ICF-IDSF with the Naive

Bayes classification engine got results, namely Precision of 91%, Recall of 93%, F1-Score of 86%, and Accuracy of 84%. It is shown in Table II that the best and stable percentage results using the proposed term weighting method, namely TF-IDF-ICF-IDSF Precision of 91%, recall 93%, F1-Score 86%, and Accuracy 84%. Furthermore, the research data are summarized in the curve in Fig. 7. And the results of the classification of the titles are presented in Table V.

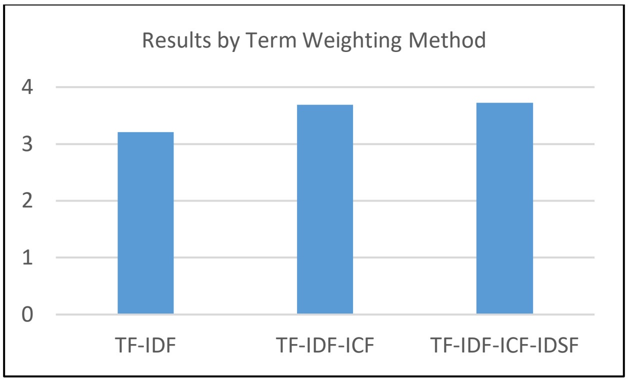

Fig. 7. Results Confussion Matrix

TABLE V. RESULTS

Title	Class
Pengembangan media pembelajaran sistem bilangan menggunakan augmented reality berbasis android untuk smk	Informatics education
Pengaruh media pembelajaran e-learning berbasis moodle terhadap hasil belajar siswa pada materi javascript kelas x tkj smk negeri 1 labang	Informatics education
Profil berpikir kritis siswa kelas x tkj ditinjau dari kemampuan awal siswa materi struktur kontrol percabangan di smkn 3 bangkalan	Education
Pengembangan game edukasi 3d struktur algoritma pemrograman menggunakan unity berbasis android sebagai media pembelajaran pemrograman dasar di smk bahrul ulum surabaya	Informatics education
Pengembangan sistem pencarian informasi buku berbasis web menggunakan moving contracting window pattern algorithm	Informatics education
Pengembangan media pembelajaran trainer komputer dalam mata pelajaran produktif materi perakitan komputer di smkn 1 sepulu	Informatics education
Pengembangan game educative berbasis android untuk pada mata pelajaran matematika materi bangun ruang untuk siswa sekolah dasar	Informatics education
Pengembangan media pembelajaran e-learning berbasis web untuk siswa kelas x teknik komputer dan jaringan	Informatics education
Pengaruh penerapan sublime text 3 terhadap minat belajar siswa pada mata pelajaran pemrograman web bahasan style halaman web kelas x tjk di smk al – hikam	Education
Pengembangan media pembelajaran perakitan komputer berbasis augmented reality	Informatics education
Pengaruh metode simulasi berbantu media vmware terhadap hasil belajar pada materi instalasi sistem operasi linux di kelas x smkn 1 labang	Education
Aplikasi kamus pemrograman mobile menggunakan metode term frequency inverse document frequency (tf-idf) untuk smartphone android	Informatics
Pengembangan media pembelajaran game edukatif berbasis rpg pada materi algoritma pemrograman di smk nurul amanah	Informatics education
Pengembangan media pembelajaran game edukasi berbasis flash pada mata pelajaran merakit komputer untuk kelas x tkj di smkn 1 kamal	Informatics education
Pengaruh model pembelajaran berbasis proyek terhadap hasil belajar siswa kelas xi pada mata pelajaran keterampilan komputer dan pengelolaan informasi di smk siding puri sumenep	Education
Pengembangan sistem informasi monitoring siswa berbasis web untuk memonitoring kegiatan siswa di smk negeri 1 kamal	Informatics education
Pengembangan media simulasi instalasi sistem operasi menggunakan unity di smk al-hikam burneh	Informatics education
Pengaruh model problem based learning terhadap motivasi belajar siswa kelas x tkj pada materi pengantar subnetting jaringan komputer di smk agung mulia	Informatics education
Pengembangan media e-book menggunakan flipbook maker berbasis web pada mata pelajaran perakitan komputer kelas x tkj di smk agung mulia bangkalan	Informatics education
Efektivitas model project based learning pada materi cascading style sheet (css) kelas x	Informatics education
Pengembangan aplikasi m-learning berbasis android menggunakan unity pada materi jaringan lan untuk siswa kelas xi tkj di smkn 2 lamongan.	Informatics education
Pengembangan sistem informasi perpustakaan berbasis web menggunakan barcode scanner di man bangkalan	Informatics education
Pengaruh pembelajaran menggunakan media e-learning berbasis web (gnomio) terhadap hasil belajar siswa pada mata pelajaran perakitan komputer di kelas x rpl smkn 2 bangkalan	Education
Pengaruh metode pembelajaran artikulasi terhadap hasil belajar siswa pada mata pelajaran jaringan dasar kelas x rpl di smk negeri 2 bangkalan	Education
Pengaruh model pembelajaran discovery learning terhadap hasil belajar sistem komputer siswa kelas x smk al-hikam	Education
Pengaruh model pembelajaran problem based learning terhadap hasil belajar siswa pada mata pelajaran perakitan komputer di smkn 2 bangkalan	Education
Pengembangan media pembelajaran berbasis augmented reality pada materi pengenalan komponen elektronika di kelas x jurusan tkj di smks ibnu cholil bangkalan	Informatics education
Pengembangan sistem informasi absensi siswa berbasis java desktop di sma darul kholil bangkalan	Informatics
Pengaruh model pembelajaran kooperatif metode time token pada materi struktur sistem operasi open source terhadap hasil belajar siswa	Education
Pengaruh metode pembelajaran probing promting terhadap hasil belajar kognitif siswa kelas x tki pada materi instalasi sistem operasi open-source di smkn 1 sumenep.	Education

Title	Class
Pengaruh pembelajaran menggunakan media vmware workstation terhadap hasil belajar siswa materi pelajaran instalasi sistem operasi smk al-hikam bangkalan.	Education
Pengaruh model pembelajaran cooperative learning tipe snowball throwing terhadap hasil belajar sistem operasi siswa kelas x smkn 1 kamal	Education
Efektivitas model pembelajaran project based learning pada mata pelajaran pemrograman dasar di smkn 1 sumenep	Informatics
Pengaruh media aplikasi virtualbox terhadap hasil belajar kognitif siswa pada materi proxy server kelas xi tkj smk alhikam.	Education
Pengaruh model pembelajaran problem based learning (pbl) terhadap hasil belajar siswa pada materi instalasi program aplikasi di smkn 1 kamal	Education
Proses berpikir kreatif siswa pada materi perkembangan sistem operasi ditinjau dari kemampuan awal siswa	Education
Pengaruh metode problem solving menggunakan media pembelajaran cbi (computer based instruction) terhadap hasil belajar desain grafis di smk agung mulia	Education

V. Conclusion

The conclusion of this study is that the TF-IDF-ICF-IDSF term weighting method has been able to provide the best results and outperform other term weighting in its application to the classification engine with the Indonesian thesis title dataset. The results obtained on the Naive Bayes classification engine get the results of 85% Precission, 75% Recall, 78% F1-Score, and 83% Accuracy. TF-IDF-ICF with the Naive Bayes classification engine got the results, namely Precision of 93%, Recall of 96%, F1-Score of 94%, and Accuracy of 86%. TF-IDF-ICF-IDSF with the Naive Bayes classification engine gets the results. Precision is 93%, Recall is 92%, F1-Score is 95%, and Accuracy is 92%. TF-IDF with the SVM classification engine got the results, namely Precision of 81%, Recall of 66%, F1-Score of 71%, and Accuracy of 77%. TF-IDF-ICF with the Naive Bayes classification engine got the results, namely Precision of 83%, Recall of 71%, F1-Score of 75%, and Accuracy of 83%. TF-IDF-ICF-IDSF with the Naive Bayes classification engine got results, namely Precision of 91%, Recall of 93%, F1-Score of 86%, and Accuracy of 84%. This research has succeeded in placing the title of the thesis against its classification, namely education, informatics, and informatics education. So that the grouping of student interest areas will be more structured so that the selection of supervisors can also be directed more correctly. Suggestions for further research is to conduct research on comparisons with more term weighting, classification machines, and application to datasets other than the Indonesian thesis title.

References

[1] T. Dogan and A. K. Uysal, "On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification," *Arab. J. Sci. Eng.*, 2019, doi: 10.1007/s13369-019-03920-9.

[2] A. Qazi and R. H. Goudar, "ScienceDirect An Ontology-based Term Weighting Technique for Web Document Categorization," *Procedia Comput. Sci.*, vol. 133, pp. 75–81, 2018, doi: 10.1016/j.procs.2018.07.010.

[3] M. Belazzoug, M. Touahria, F. Nouioua, and M. Brahimi, "An improved sine cosine algorithm to select features for text categorization," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 4, pp. 454–464, 2020, doi: 10.1016/j.jksuci.2019.07.003.

[4] S. K. Vishwakarma, K. I. Lakhtaria, D. Bhatnagar, and A. K. Sharma, "Monolingual Information Retrieval using Terrier : FIRE 2010 Experiments based on n-gram indexing," *Procedia - Procedia Comput. Sci.*, vol. 57, pp. 815–820, 2015, doi: 10.1016/j.procs.2015.07.484.

[5] C. Li, S. Chen, and Y. Qi, "Filtering and Classifying Relevant Short Text with a Few Seed Words," *Data Inf. Manag.*, vol. 3, no. 3, pp. 165–186, 2019, doi: 10.2478/dim-2019-0011.

[6] L. Yang, C. Li, Q. Ding, and L. Li, "Combining Lexical and Semantic Features for Short Text Classification," vol. 22, pp. 78–86, 2013, doi: 10.1016/j.procs.2013.09.083.

[7] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, "Short text classification for Arabic social media tweets," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6595–6604, 2022, doi: 10.1016/j.jksuci.2022.03.020.

[8] R. Jiménez, B. Redondo, R. Molina, M. Á. Martínez-domingo, J. Hernández-andrés, and J. Vera, "Short-term e ff ects of text-background color combinations on the dynamics of the accommodative response," *Vision Res.*, vol. 166, no. December 2019, pp. 33–42, 2020, doi: 10.1016/j.visres.2019.11.006.

[9] A. M. Saeed, S. R. Hussein, C. M. Ali, and T. A. Rashid, "Medical dataset classification for Kurdish short text over social media," *Data Br.*, vol. 42, p. 108089, 2022, doi: 10.1016/j.dib.2022.108089.

[10] T. Ek, C. Kirkegaard, H. Jonsson, and P. Nugues, "Pacific Association for Computational Linguistics (PACLING 2011) Named Entity Recognition for Short Text Messages," vol. 27, no. Pacling, pp. 178–187, 2011, doi: 10.1016/j.sbspro.2011.10.596.

[11] Y. Wang, S. Wu, D. Li, S. Mehrabi, and H. Liu, "A Part-Of-Speech term weighting scheme for biomedical information retrieval," *J. Biomed. Inform.*, vol. 63, pp. 379–389, 2016, doi: 10.1016/j.jbi.2016.08.026.

[12] M. Kumari, A. Jain, and A. Bhatia, "Synonyms Based Term Weighting Scheme : An Extension to TF . IDF," *Procedia - Procedia Comput. Sci.*, vol. 89, pp. 555–561, 2016, doi: 10.1016/j.procs.2016.06.093.

[13] A. Z. Arifin and T. Informatika, "Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis," vol. 13, no. 2, pp. 172–180, 2023.

[14] A. Alwehaibi, M. Bikdash, M. Albogmi, and K. Roy, "A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches," *J. King Saud Univ. - Comput. Inf. Sci.*, vol.

34, no. 8, pp. 6140–6149, 2022, doi: 10.1016/j.jksuci.2021.07.011.

[15] F. Ren and M. G. Sohrab, “Class-indexing-based term weighting for automatic text classification,” vol. 236, pp. 109–125, 2013.

[16] Y. Man, “Feature Extension for Short Text Categorization Using Frequent Term Sets,” *Procedia - Procedia Comput. Sci.*, vol. 31, pp. 663–670, 2014, doi: 10.1016/j.procs.2014.05.314.

[17] T. Dogan and A. K. Uysal, “Improved inverse gravity moment term weighting for text classification,” *Expert Syst. Appl.*, vol. 130, pp. 45–59, 2019, doi: 10.1016/j.eswa.2019.04.015.

[18] I. Agustien, R. Pahlevi, and M. Kautsar, “ScienceDirect ScienceDirect Combination of Term Weighting and Integrated Color Intensity Co- occurrence Matrix for Two-Level Image Retrieval on Social Media Data,” *Procedia Comput. Sci.*, vol. 157, pp. 329–336, 2019, doi: 10.1016/j.procs.2019.08.174.

[19] A. Tsalitsatun, D. Ari, and A. Zainal, “Autonomy Stemmer Algorithm for Legal and Illegal Affix Detection Use Finite-State Automata Method,” vol. 2, no. 1, pp. 46–55, 2019, doi: 10.25042/epi-ije.022019.09.

[20] R. Setiawan, “ScienceDirect ScienceDirect ScienceDirect Performance Performance Comparison Comparison and and Optimization Optimization of of Text Text Document Document Classification using Naïve Bayes Classification Classification using k-NN and Naïve Bayes Classific,” *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017, doi: 10.1016/j.procs.2017.10.017.

.