Gmail

Hairani Hairani <hairani@universitasbumigora.ac.id>

Receipt ICOCSIM 2021 - PAPER ID 8

1 message

NUR SHAMSIAH BINTI ABDUL RAHMAN . <shamsiah@ump.edu.my> Wed, Aug 4, 2021 at 11:17 AM
To: hairani@universitasbumigora.ac.id

Dear author,
Thank you for joining the 4th International Conference on Computational Science and Information Management (ICoCSIM 2021) which will be held on 24-26 August 2021. Attached is the receipt payment for the PAPER ID 8. See you on conference day. Thank you.

NUR SHAMSIAH BINTI ABDUL RAHMAN (PhD)
Senior Lecturer
Faculty of Computing
College of Computing and Applied Sciences
Universiti Malaysia Pahang (UMP)
26600 Pekan, Pahang, Malaysia.
Phone : +609-4244702

www.ump.edu.my UMPMalaysia TEKNOLOGI UNTUK MASYARAKAT 5 STARS QS RATED FOR EXCELLENCE 2018 | 801-1000 QS WORLD UNIVERSITY RANKINGS 2022 | #133 ASIA QS WORLD UNIVERSITY RANKINGS 2021

"Think Green. Keep it on the screen.
If printing is necessary, please print it on both sides."

The information contained in this e-mail message and any accompanying files is or may be confidential. If you are not the intended recipient, any use, dissemination, reliance, forwarding, printing or copying of this e-mail or any attached files is unauthorized. This e-mail is subject to copyright. No part of it should be reproduced, adapted or communicated without the written consent of the copyright owner. If you have received this e-mail in error please advise the sender immediately by return e-mail or telephone and delete all copies. UMP does not guarantee the accuracy or completeness of any information contained in this e-mail or attached files. Internet communications are not secure, therefore UMP does not accept legal responsibility for the contents of this message or attached files.
Universiti Malaysia Pahang

Resit Paper ICOCSIM ID 8 (RP2108-0124).pdf
286K

Thesis Topic Classification Based on Abstract Using the Naïve Bayes Method

Hairani Hairani
Faculty of Engineering and Design
Bumigora University
Mataram, Indonesia
hairani@universitasbumigora.ac.id

Anthony Anggrawan
Faculty of Engineering and Design
Bumigora University
Mataram, Indonesia
anthony.anggrawan@universitasbumigora.ac.id

Ahmad Islahul Wathan
Faculty of Engineering and Design
Bumigora University
Mataram, Indonesia
wathanruna@gmail.com

Kurniadin Abd Latif
Faculty of Engineering and Design
Bumigora University
Mataram, Indonesia
kurniadin@universitasbumigora.ac.id

Khairan Marzuki
Faculty of Engineering and Design
Bumigora University
Mataram, Indonesia
khairan@universitasbumigora.ac.id

Muhammad Zulfikri
Faculty of Engineering and Design
Bumigora University
Mataram, Indonesia
mzulfikri@universitasbumigora.ac.id

***Abstract*— The thesis is a requirement for graduation from Bumigora university. The final year student's problem is determining the research topic because the undergraduate thesis collection of Computer Science is not grouped or classified based on student competencies. The purpose of this study was to compare the performance of the naïve Bayes method with TF-IDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The stages of this research are data collection, text pre-processing, term weighting with TF-IDF and without TF-IDF, Naïve Bayes method implementation, and result evaluation. Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces 88.69% accuracy, 89.76% precision, and 90.49% sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.**

Keywords—naïve bayes, TF-IDF weighting, asbtrak classification, text mining

I. Introduction

The thesis is one of the graduation requirements for undergraduate Computer Science students at Bumigora University. Students can start working on their thesis if the research topic has been approved through a synopsis exam. So far, students have difficulties in determining the proposed thesis topic. One of the difficulties is because the existing collection of an undergraduate thesis in Computer Science is not grouped or classified based on student competencies. Automatic thesis grouping or classification of topics is one solution that can make it easier for students to find references to research titles based on their competence. The competencies of students in the S1 Computer Science program at Bumigora university are computer networks, multimedia, and software engineering (RPL).

One of the solutions offered by this research is to use the concept of text mining. Previous research used various methods for text mining-based thesis document analysis such as the k-means method [1]–[4], K-Nearest Neighbor [5]–[7], Cosine Similiarity [8], [9], Decision Tree and Naïve Bayes [10], SVM and Naïve Bayes [11]. Research [10] compared Decision Trees, Naïve Bayes, and k-NN methods to predict thesis graduation. Based on the results of his research, the k-NN method has the best accuracy compared to the decision tree and naïve Bayes methods at 80.39%. Research [4] used the k-means method for grouping thesis titles. Before grouping, the first weighting of words is carried out using the TF - IDF method. Research [9] uses the cosine similarity method for the classification of thesis documents. Before grouping, the first weighting of words is carried out using the TF - IDF method.

Based on previous research, there is a difference made with this research, namely the research carried out a classification of thesis topics based on the abstract using the naïve Bayes method and also using the k-fold cross-validation test method. The aim is to compare the performance of the naïve Bayes method with TF-IDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The performance used in this study is accuracy, precision, and sensitivity.

II. Research Method

The stages used in this study are shown in Figure 1.

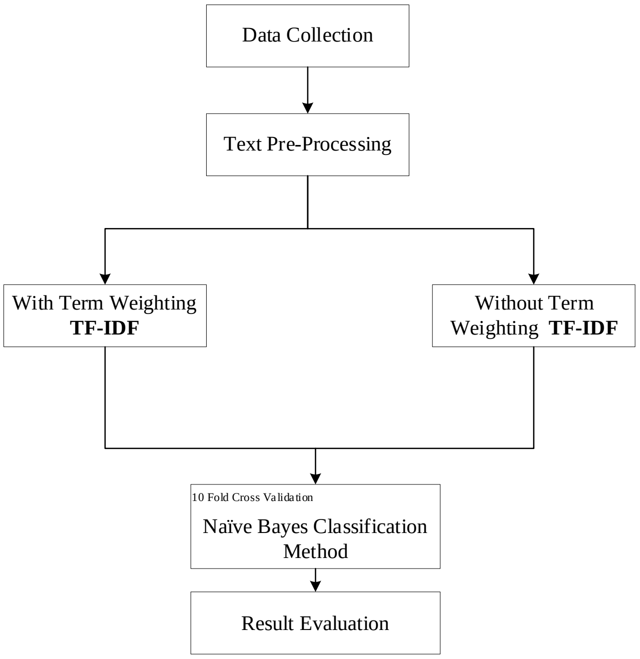
Data Collection
Text Pre-Processing
With Term Weighting **TF-IDF**
Without Term Weighting **TF-IDF**
10 Fold Cross Validation
Naïve Bayes Classification Method
Result Evaluation

Figure 1. Research Methodology

A. Data Collection

The data used in this study are the thesis abstracts of 2020 graduate computer science undergraduate students obtained

from www.repository.universitasbumigora.ac.id. The data collected were 115 thesis abstract data, consisting of 36 topics of computer networks, 23 multimedia, and 55 software engineering (RPL).

B. *Text Pre-Processing*

Text pre-processing used to get quality data. The classification was carried out using the naïve Bayes method. The text pre-processing technique used in this study consists of case folding, tokenization, stop word removal. and stemming [12]. Case folding is used to change text to lowercase. Tokenization is used to separate text into tokens. Stopword removal is used to remove unnecessary words such as conjunctions. Stemming is used to change all words that have affixes into basic words.

C. *Term Weighting TF-IDF*

The term weighting process is used to give a weight value to each word. The term weighting method used in this study is the Term Frequency - Inverse Document Frequency (TF-IDF). The TF-IDF method combines two concepts, namely TF and IDF. TF looks for the occurrence value of terms in related documents, the more occurrences of terms in the related document, the better. Meanwhile, the IDF concept is inversely proportional to the TF method, the less frequently the terms appear in all documents the better. TF - IDF method is calculated using equation (1) [13].

$$W_{ij} = tf_{ij} \ x \ idf_j = tf_{ij} \ x \ \log\left(\frac{N}{df_j}\right) \qquad (1)$$

W_{ij} is the weight of term j to document i. tf_{ij} is the number of occurrences of term j in the document d. N is the number of documents, and df_i is the number of occurrences of term j throughout the document.

D. *Naïve Bayes Method Classification*

At this stage, the classification is carried out using the naïve Bayes method. The basic concept of the naïve Bayes method is a probability-based classification method that assumes independence from the dependent variable and is also a conditional model based on the Bayes theorem [14][15]. The Naïve Bayes method used in this study is the multinomial Naïve Bayes which is calculated based on equation (2).

$$P(c \mid term\ document\ d) = P(c)\ x\ \mathrm{P(t_1|c)}\ x\ \mathrm{P(t_2|c)}\ x\ \mathrm{P(t_{\mathit{n}}|c)} \qquad (2)$$

$P(c)$ is the prior probability of class c. $P(c \mid term\ document\ d)$ is the probability of the appearance of a term in document d including class c. $P(t_n \mid c)$ is the probability of occurrence of term n known to class c.

The process of calculating the prior probability for class c uses equation (3).

$$P(c) = \frac{N_C}{N} \qquad (3)$$

N_c is the number of class c in all documents, while N is the total number of documents. The calculation of the probability of occurrence of term n is calculated using equation (4) involving the laplacian technique.

$$P(t_n \mid c) = \frac{count(t_n, c) + 1}{count(c) + |v|} \qquad (4)$$

$count(t_n, c)$ is the number of terms t_n appearing in the training data with class c. $count(c)$ is the number of terms in the class training data c. weighting is used to give weight to the value of each word. is the number of terms in the training data. V is the number of terms in the training data.

Data classified by the multinomial naïve Bayes method are grouped into training and testing data first. The distribution of training and testing data in this study uses the k-fold cross-validation method by dividing the data as much as the specified k. Each fold can be used as training and testing data in turn. This research uses 10 fold data validation method.

E. *Result Evaluation*

At this stage, the results are evaluated based on accuracy, precision, and sensitivity using the confusion matrix table shown in Table Tabel I.

TABLE I. CONFUSION MATRIX

Actual	**Predicted**			**Total**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	True Jaringan	Error	Error	Total (Jaringan)
Multimedia	Error	True Multimedia	Error	Total (Multimedia)
RPL	Error	Error	True RPL	Total (RPL)
	Predicted (Jaringan)	Predicted (Multimedia)	Predicted (RPL)	

Evaluation of results based on accuracy, precision, and sensitivity using equations (5), (6), and (7).

$$\text{Accuracy}=\frac{\text{True Jaringan+True Multimedia+True RPL}}{\text{Total (Jaringan)+Total (Multimedia)+Total (RPL)}} \qquad (5)$$

$$\text{Precision}_{\text{(Jaringan)}} = \frac{\text{True Jaringan}}{\text{Prediksi (Jaringan)}} \qquad (6)$$

$$\text{Sensitivity}_{\text{(Jaringan)}} = \frac{\text{True Jaringan}}{\text{Total (Jaringan)}} \qquad (7)$$

III. RESULT AND DISCUSSION

A. *Data Collection*

The data used in this study are the thesis abstracts of 2020 graduate computer science undergraduate students obtained from www.repository.universitasbumigora.ac.id. The data collected were 115 thesis abstract data, consisting of 36 topics of computer networks, 23 multimedia, and 55 software engineering (RPL). The sample abstract data of this research thesis is shown in Table II.

TABLE II. THESIS ABSTRACT DATASET

No	***Abstract***	***Topic***
1.	Perkembangan teknologi informasi sangat cepat seperti Internet of Things (IoT), dimana seseorang dapat melakukan segala aktivitasnya dengan mudah dengan mengandalkan sistem Internet of Things (IoT). Seiring dengan perkembangan zaman maka semakin canggih teknologi yang dihasilkan baik digunakan sebagai hal yang positif maupun melakukan hal yang negatif, tak terkecuali pada system peternakan sehingga perlu mengembangkan teknologi untuk manajemen pakan ternakkhususnya hewan ternak ayam	Jaringan

No	***Abstract***	***Topic***
	broiler.Pengembangkan sistem menggunakan sistem Internet of Things dan sistem penjadwalan otomasi dimana sistem Internet of Things (IoT) adalah sistem yang berfungsi melakukan kontroller pada alat alat elektronik. Metodelogi penelitian yang digunakan adalah Network Development Life Cycle (NDLC), terdiri dari; analisis, desain, prototype dan ujicoba. Pada tahap analisis memuat tentang pengumpulan data, tahap desain memuat rancangan sistem pemberian pakan ternak, prototyping memuat instalasi konfigurasi dan membangun kerangka sistem pakan ternak. Ujicoba memuat tentang pengujian sistem pemberian pakan ternak secara otomatis atau terjadwal. Kesimpulan dari penelitian ini adalah menginplementasi Sever VPS dengan sistem nodemcu dalam pemberian pakan ternak berbasis Internet of Things (IoT) untuk efisiensi dalam pemberian pakan ternak ayam.	
2.	Pemanfaatan teknologi Augmented Reality (AR) sebagai media pembelajaran tentang sendi gerak tubuh manusia bertujuan sebagai alat bantu dalam proses belajar dan mengajar alternatif antara guru dan siswa dengan cara memvisualisasikan objek 3D secara realtime. Aplikasi Visualisasi sendi gerak tubuh manusia menggunakan teknologi Augmented Reality berbasis mobile dengan mengacu pada materi dalam buku IPA SMA sederajat kelas XI Semester kurikulum 2013 revisi tahun 2016. Metodologi yang digunakan dalam pengembangan aplikasi ini adalah metode pengembangan Luther Sutopo. Dimana metode ini terdiri dari 6 (Enam) tahap yaitu concept, design, material collecting, assembly, testing, dan distribution. Hasil atau keluaran dari aplikasi yang penulis bangun ini adalah sebuah aplikasi android dengan memanfaatkan teknologi Augmented Reality untuk memperlihatkan bentuk dari proses pergerakan sendi pada tulang manusia secara realtime. Kesimpulan dari penelitian ini adalah secara keseluruhan respon dari end user terhadap aplikasi ini sudah cukup baik. dimana diketahui dari responden yang menyatakan Sangat Setuju (SS) = 40%, yang menyatakan Setuju (S) = 57% yang meyatakan Netral (N) = 2% dan yang menyatakan tidak setuju (ST) = 0,6% Berdasarkan hasil tersebut menunjukkan bahwa aplikasi dapat digunakan sebagai media pembelajaran dalam memahami materi sendi gerak tubuh manusia	Multimedia
….	………………………………………………….	………
11 5	Saat ini penyakit Telinga Hidung dan Tenggorokan (THT) telah menjadi suatu penyakit yang cukup banyak diderita oleh masyarakat dunia. Di Indonesia, penderita penyakit THT berjumlah sekitar 190-230 per 1000 penduduk. Aplikasi ini menggunakan metode Forward Chaining dan Certainty Factor. Jenis penyakit yang diteliti pada penelitian ini adalah Ortitis Media Serosa, Polip Hidung, Faringtis Akut, Abses Retrofaring, dan Karsinoma Nafosaring. Tujuan pembuatan sistem pakar ini adalah untuk memudahkan pasien untuk mengetahui penyakit apa yang dideritanya, serta memudahkan tenaga medis dalam menangani pasien THT. Tahapan pengembangan sistem pakar pada penelitian ini terdiri dari identifikasi masalah untuk analisis domain permasalahan dan analisis kebutuhan fungsional, akuisisi pengetahuan digunakan untuk mendapatkan nilai MB dan MD tiap-tiap gejala pada penyakit THT dengan metode wawancara, perancangan digunakan untuk merancang representasi pengetahuan seperti tabel keputusan dan mesin inferensi. Dengan adanya sistem pakar diagnosis penyakit THT dapat mempermudah dokter mengambil keputusan, atau diagnosa yang tepat terhadap suatu gejala – gejala yang timbul pada penyakit THT, sehingga diperoleh pengobatan yang tepat dan minimalisir terjadinya kesalahan diagnosa	RPL

B. *Text Pre-Processing*

Text pre-processing used to get quality data. The classification was carried out using the naïve Bayes method. The text pre-processing technique used in this study consists of case folding, tokenization, stop word removal. and stemming. The examples of text pre-processing stages are shown in Table III.

TABLE III. EXAMPLE OF TEXT PREPROCESSING

Pre-processing	Result
Data Original	Tujuan pembuatan sistem pakar diagnosis jenis penyakit THT adalah memudahkan masyarakat umum untuk mengetahui jenis penyakit THT diderita tanpa perlu datang ke dokter spesialis THT
Case Folding	tujuan pembuatan sistem pakar diagnosis jenis penyakit tht adalah memudahkan masyarakat umum untuk mengetahui jenis penyakit tht diderita tanpa perlu datang ke dokter spesialis tht
Tokenization	[‘tujuan’, ‘pembuatan’, ‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘penyakit’, ‘tht’, ‘adalah’, ‘memudahkan’, ‘masyarakat’, ‘umum’, ‘untuk’, ‘mengetahui’, ‘jenis’, ‘penyakit’, ‘tht’, ‘diderita’, ‘tanpa’, ‘perlu’, ‘datang’, ‘ke’, ‘dokter’, ‘spesialis’, ‘tht’]
stop word removal	[‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘penyakit’, ‘tht’, ‘masyarakat’, ‘mengetahui’, ‘jenis’, ‘penyakit’, ‘tht’, ‘diderita’, ‘dokter’, ‘spesialis’, ‘tht’]
stemming	[‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘sakit’, ‘tht’, ‘masyarakat’,‘tahu’, ‘jenis’, ‘sakit’, ‘tht’, ‘derita’, ‘dokter’, ‘spesialis’, ‘tht’]

C. *Term Weighting TF-IDF*

The term weighting process is used to give weight to the value of each word. The term or word weighting method used in this study is TF-IDF. The example of the TF-IDF calculation process using the documents in Tabel III, the stemming section, is shown in Table IV.

TABLE IV. RESULT OF WEIGHTING TERM TF-IDF

Term	tf	D	D/df	log (IDF)+1	W= tf * (IDF+1)
	D1				D1
datang	1	1	1	1	1
derita	1	1	1	1	1
diagnosis	1	1	1	1	1
dokter	1	1	1	1	1
jenis	2	1	1	1	2
masyarakat	1	1	1	1	1
pakar	1	1	1	1	1
sakit	2	1	1	1	2
sistem	1	1	1	1	1
spesialis	1	1	1	1	1

tht	3	1	1	1	3

D. *Naïve Bayes Method Classification*

At this stage, the classification is carried out using the naïve Bayes multinomial method by comparing the performance using TF-IDF weighting and without TF-IDF weighting using equation (2).

E. *Result Evaluation*

At this stage, results are evaluated based on accuracy, precision, and sensitivity using the confusion matrix table shown in Table V, VI, and VII.

TABLE V. CONFUSION MATRIX OF MULTINOMIAL NAÏVE BAYES WITH TF - IDF

Actual	**Predicted**			**Sensitivity**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	31	0	6	83.78%
Multimedia	0	18	9	66.67%
RPL	4	1	45	90%
Precision	88.57%	94.74%	75%	

TABLE VI. CONFUSION MATRIX OF MULTINOMIAL NAÏVE BAYES WITHOUT TF - IDF

Actual	**Predicted**			**Sensitivity**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	33	0	4	89.19%
Multimedia	0	26	1	96.29%
RPL	5	2	43	86%
Precision	86.84%	92.86%	89.58%	

TABLE VII. PERFORMANCE RESULT OF MULTINOMIAL NAÏVE BAYES METHOD

Performance	With TF - IDF	Without TF - IDF
Accuracy	81.74%	**88.69%**
Precision	86.1%	**89.76%**
Sensitivity	80.15%	**90.49%**

Based on the results of the tests shown in Table VII, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces **88.69%** accuracy, **89.76%** precision, and **90.49%** sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.

IV. CONCLUSION

Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces **88.69%** accuracy, **89.76%** precision, and **90.49%** sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract. The suggestions for further research can use feature selection methods such as chi-square to improve the performance of the naïve Bayes method.

ACKNOWLEDGMENT

Thank you to the Bumigora University Computer Science Undergraduate Study Program for providing research data to the author.

REFERENCES

[1] D. Adhe, C. Rachman, R. Goejantoro, and D. Tisna, “Implementation Of Text Mining For Grouping Thesis Documents Using K-Means Clustering,” *J. EKSPONENSIAL*, vol. 11, no. 2, pp. 167–174, 2020.

[2] A. A. Jalal and B. H. Ali, “Text documents clustering using data mining techniques,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, Feb. 2021, doi: 10.11591/ijece.v11i1.pp664-670.

[3] M. Sholehhudin, M. Fauzi Ali, and S. Adinugroho, “Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya),” vol. 2, no. 11, pp. 5518–5524, 2018.

[4] L. Zahrotun, N. H. Putri, and A. Nur Khusna, “The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesis Titles,” in *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Oct. 2018, pp. 1–4, doi: 10.1109/TSSA.2018.8708817.

[5] D. M. U. Atmaja and R. Mandala, “Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor,” *IT Soc.*, vol. 4, no. 2, pp. 1–6, Aug. 2020, doi: 10.33021/itfs.v4i2.1182.

[6] M. Eminağaoğlu and Y. Gökşen, “A New Similarity Measure for Document Classification and Text Mining,” *KnE Soc. Sci.*, vol. 2019, pp. 353–366, Jan. 2020, doi: 10.18502/kss.v4i1.5999.

[7] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, “Using text mining techniques for extracting information from research articles,” *Stud. Comput. Intell.*, vol. 740, no. November, pp. 373–397, 2018, doi: 10.1007/978-3-319-67056-0_18.

[8] R. Rismanto, A. Rachmad Syulistyo, and B. P. Citra Agusta, “Research Supervisor Recommendation System Based on Topic Conformity,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 12, no. 1, pp. 26–34, Feb. 2020, doi: 10.5815/ijmecs.2020.01.04.

[9] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, “Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi,” *J. Tek. Elektro*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/view/10955/6659.

[10] A. Solichin, “Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation,” in *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Sep. 2019, pp. 217–222, doi: 10.23919/EECSI48112.2019.8977081.

[11] S. Sulova, L. Todoranova, B. Penchev, and R. Nacheva, “Using Text Mining to Classify Research Papers,” in *International Multidisciplinary Scientific GeoConference Surveying Geology*

and Mining Ecology Management, SGEM, Jun. 2017, vol. 17, no. 21, pp. 647–654, doi: 10.5593/sgem2017/21/S07.083.

[12] A. Rahman and A. Doewes, “Online News Classification Using Multinomial Naive Bayes,” *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.

[13] A. Prabhat and V. Khullar, “Sentiment classification on big data using Naïve bayes and logistic regression,” in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, 2017, pp. 1–5, doi: 10.1109/ICCCI.2017.8117734.

[14] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano, “Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers,” *Neurocomputing*, vol. 92, pp. 98–115, Sep. 2012, doi: 10.1016/j.neucom.2012.01.030.

[15] X. Glorot, A. Bordes, and Y. Bengio, “Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach,” in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 513–520.

ICSECS 2021
ADVANCING DIGITAL SOCIETY WITH ADVANCED TECHNOLOGIES

UMP
اونيۏرسيتي مليسيا ڤهڠ
UNIVERSITI MALAYSIA PAHANG

4th International ICoCSIM 2021
Computational Information Science & Management

CERTIFICATE
OF PARTICIPATION

This is to certify that

Hairani Hairani, Anthony Anggrawan, Ahmad Islahul Wathan, Kurniadin Abd Latif, Khairan Marzuki and Muhammad Zulfikri

presented a paper titled

The Abstract of Thesis Classifier by Using Naive Bayes Method

in a Joint Conference of

7th International Conference on
Software Engineering & Computer Systems (ICSECS 2021)

and

4th International Conference on
Computational Science and Information Management (ICoCSIM 2021)

August 24 - 25, 2021
Universiti Malaysia Pahang, Pekan, Malaysia

DR. JAMALUDIN BIN SALLIM
General Chair
ICSECS-ICoCSIM 2021

The Abstract of Thesis Classifier by Using Naive Bayes Method

By Hairani Hairani

WORD COUNT	2596	TIME SUBMITTED	19-JUL-2021 01:53PM
		PAPER ID	74615405

The Abstract of Thesis Classifier by Using Naive Bayes Method

Hairani Hairani
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
hairani@universitasbumigora.ac.id

Anthony Anggrawan
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
anthony.anggrawan@universitasbumigora.ac.id

Ahmad Islahul Wathan
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
wathanruna@gmail.com

Kurniadin Abd Latif
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
kurniadin@universitasbumigora.ac.id

Khairan Marzuki
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
khairan@universitasbumigora.ac.id

Muhammad Zulfikri
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
mzulfikri@universitasbumigora.ac.id

***Abstract*— The thesis is a requirement for graduation from Bumigora university. The final year student's problem is determining the research topic because the undergraduate thesis collection of Computer Science is not grouped or classified based on student competencies. The purpose of this study was to compare the performance of the naïve Bayes method with TF-IDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The stages of this research are data collection, text pre-processing, term weighting with TF-IDF and without TF-IDF, Naïve Bayes method implementation, and result evaluation. Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces 88.69% accuracy, 89.76% precision, and 90.49% sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.**

Keywords—naïve bayes, TF-IDF weighting, asbtract classification

I. Introduction

The thesis is one of the graduation requirements for undergraduate Computer Science students at Bumigora University. Students can start working on their thesis if the research topic has been approved through a synopsis exam. So far, students have difficulties in determining the proposed thesis topic. One of the difficulties is because the existing collection of an undergraduate thesis in Computer Science is not grouped or classified based on student competencies. Automatic thesis grouping or classification of topics is one solution that can make it easier for students to find references to research titles based on their competence. The competencies of students in the S1 Computer Science program at Bumigora university are computer networks, multimedia, and software engineering (RPL).

One of the solutions offered by this paper is to use the concept of text mining. Previous research used various methods for text mining-based thesis document analysis such as the k-means method [1]–[4], K-Nearest Neighbor [5]–[7], Cosine Similiarity [8], [9], Decision Tree and Naïve Bayes [10], SVM and Naïve Bayes [11]. Research [10] compared Decision Trees, Naïve Bayes, and k-NN methods to predict thesis graduation. Based on the results of his research, the k-NN method has the best accuracy compared to the decision tree and naïve Bayes methods at 80.39%. Research [4] used the k-means method for grouping thesis titles. Before grouping, the first weighting of words is carried out using the TF - IDF method. Research [9] uses the cosine similarity method for the classification of thesis documents. Before grouping, the first weighting of words is carried out using the TF - IDF method.

Based on previous research, there is a difference made with this research, namely the research carried out a classification of thesis topics based on the abstract using the naïve Bayes method and also using the k-fold cross-validation test method. The aim is to compare the performance of the naïve Bayes method with TF-IDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The performance used in this study is accuracy, precision, and sensitivity.

II. Research Method

The stages used in this study are shown in Figure 1.

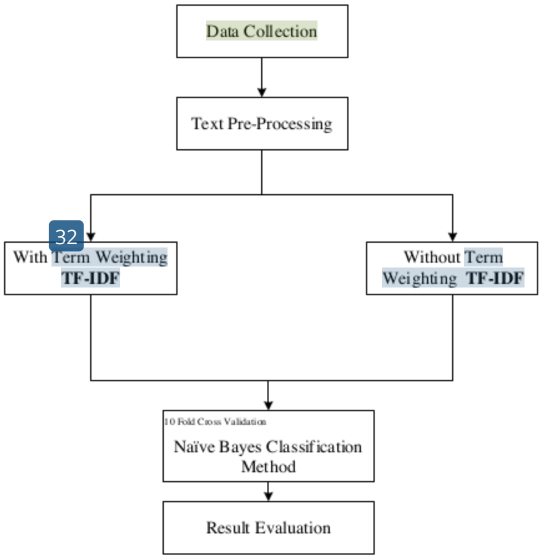
Data Collection
Text Pre-Processing
With Term Weighting TF-IDF
Without Term Weighting TF-IDF
10 Fold Cross Validation
Naïve Bayes Classification Method
Result Evaluation

Figure 1. Research Methodology

A. Data Collection

The data used in this study are the thesis abstracts of 2020 graduate computer science undergraduate students obtained

from www.repository.universitasbumigora.ac.id. The data collected were 115 thesis abstract data, consisting of 36 topics of computer networks, 23 multimedia, and 55 software engineering (RPL).

B. Text Pre-Processing

Text pre-processing used to get quality data. The classification was carried out using the naïve Bayes method. The text pre-processing technique used in this study consists of case folding, tokenization, stop word removal. and stemming [12]. Case folding is used to change text to lowercase. Tokenization is used to separate text into tokens. Stopword removal is used to remove unnecessary words such as conjunctions. Stemming is used to change all words that have affixes into basic words.

C. Term Weighting TF-IDF

The term weighting process is used to give a weight value to each word. The term weighting method used in this study is the Term Frequency - Inverse Document Frequency (TF-IDF). The TF-IDF method combines two concepts, namely TF and IDF. TF looks for the occurrence value of terms in related documents, the more occurrences of terms in the related document, the better. Meanwhile, the IDF concept is inversely proportional to the TF method, the less frequently the terms appear in all documents the better. TF - IDF method is calculated using equation (1) [13].

$$W_{ij} = tf_{ij}\ x\ idf_j = tf_{ij}\ x\ \log\left(\frac{N}{df_j}\right) \qquad (1)$$

W_{ij} is the weight of term j to document i. tf_{ij} is the number of occurrences of term j in the document d. N is the number of documents, and df_i is the number of occurrences of term j throughout the document.

D. Naïve Bayes Method Classification

At this stage, the classification is carried out using the naïve Bayes method. The basic concept of the naïve Bayes method is a probability-based classification method that assumes independence from the dependent variable and is also a conditional model based on the Bayes theorem [14][15]. The Naïve Bayes method which is calculated based on equation (2).

$$P(c \mid term\ document\ d) = P(c)\ x\ \mathrm{P}(\mathrm{t}_1|\mathrm{c})\ x\ \mathrm{P}(\mathrm{t}_2|\mathrm{c})\ x\ \mathrm{P}(\mathrm{t}_n|\mathrm{c})\ (2)$$

$P(c)$ is the prior probability of class c. $P(c \mid term\ document\ d)$ is the probability of the appearance of a term in document d including class c. $P(t_n \mid c)$ is the probability of occurrence of term n known to class c.

The process of calculating the prior probability for class c uses equation (3).

$$P(c) = \frac{N_C}{N} \qquad (3)$$

N_c is the number of class c in all documents, while N is the total number of documents. The calculation of the probability of occurrence of term n is calculated using equation (4) involving the laplacian technique.

$$P(t_n \mid c) = \frac{count(t_n, c) + 1}{count(c) + |v|} \qquad (4)$$

$count(t_n, c)$ is the number of terms t_n appearing in the training data with class c. $count(c)$ is the number of terms in the class training data c. weighing is used to give weight to the value of each word. is the number of terms in the training data. V is the number of terms in the training data.

Data classified by the multinomial naïve Bayes method are grouped into training and testing data first. The distribution of training and testing data in this study uses the k-fold cross-validation method by dividing the data as much as the specified k. Each fold can be used as training and testing data in turn. This research uses 10 fold data validation method.

E. Result Evaluation

At this stage, the results are evaluated based on accuracy, precision, and sensitivity using the confusion matrix table shown in Table Tabel I.

TABLE I. CONFUSION MATRIX

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Evaluation of results based on accuracy, precision, and sensitivity using equations (5), (6), and (7).

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \qquad (5)$$

$$\text{Precision}=\frac{TP}{TP+FP} \qquad (6)$$

$$\text{Sensitivity} = \frac{TP}{\text{TP + FN}} \qquad (7)$$

III. RESULT AND DISCUSSION

A. Data Collection

The data used in this study are the thesis abstracts of 2020 graduate computer science undergraduate students obtained from www.repository.universitasbumigora.ac.id. The data collected were 115 thesis abstract data, consisting of 36 topics of computer networks, 23 multimedia, and 55 software engineering (RPL). The sample abstract data of this research thesis is shown in Table II.

TABLE II. THESIS ABSTRACT DATASET

No	***Abstract***	***Topic***
1.	Perkembangan teknologi informasi sangat cepat seperti Internet of Things (IoT), dimana seseorang dapat melakukan segala aktivitasnya dengan mudah dengan mengandalkan sistem Internet of Things (IoT). Seiring dengan perkembangan zaman maka semakin canggih teknologi yang dihasilkan baik digunakan sebagai hal yang positif maupun melakukan hal yang negatif, tak terkecuali pada system peternakan sehingga perlu mengembangkan teknologi untuk manajemen pakan ternakkhususnya hewan ternak ayam broiler.Pengembangkan sistem menggunakan sistem Internet of Things dan sistem penjadwalan otomasi dimana sistem Internet of Things (IoT) adalah sistem yang berfungsi melakukan kontroller pada alat alat elektronik. Metodelogi	Jaringan

No	*Abstract*	*Topic*
	penelitian yang digunakan adalah Network Development Life Cycle (NDLC), terdiri : analisis, desain, prototype dan ujicoba. Pada tahap analisis memuat tentang pengumpulan data, tahap desain memuat rancangan sistem pemberian pakan ternak, prototyping memuat instalasi konfigurasi dan membangun kerangka sistem pakan ternak. Ujicoba memuat tentang pengujian sistem pemberian pakan ternak secara otomatis atau terjadwal. Kesimpulan dari penelitian ini adalah menginplementasi Sever VPS dengan sistem nodemcu dalam pemberian pakan ternak berbasis Internet of Things (IoT) untuk efisiensi dalam pemberian pakan ternak ayam.	

B. *Text Pre-Processing*

Text pre-processing used to get quality data. The classification was carried out using the naïve Bayes method. The text pre-processing technique used in this study consists of case folding, tokenization, stop word removal. and stemming. The examples of text pre-processing stages are shown in Table III.

TABLE III. EXAMPLE OF TEXT PREPROCESSING

Pre-processing	Result
Data Original	Tujuan pembuatan sistem pakar diagnosis jenis penyakit THT adalah memudahkan masyarakat umum untuk mengetahui jenis penyakit THT diderita tanpa perlu datang ke dokter spesialis THT
Case Folding	tujuan pembuatan sistem pakar diagnosis jenis penyakit tht adalah memudahkan masyarakat umum untuk mengetahui jenis penyakit tht diderita tanpa perlu datang ke dokter spesialis tht
Tokenization	[‘tujuan’, ‘pembuatan’, ‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘penyakit’, ‘tht’, ‘adalah’, ‘memudahkan’, ‘masyarakat’, ‘umum’, ‘untuk’, ‘mengetahui’, ‘jenis’, ‘penyakit’, ‘tht’, ‘diderita’, ‘tanpa’, ‘perlu’, ‘datang’, ‘ke’, ‘dokter’, ‘spesialis’, ‘tht’]
stop word removal	[‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘penyakit’, ‘tht’, ‘masyarakat’, ‘mengetahui’, ‘jenis’, ‘penyakit’, ‘tht’, ‘diderita’, ‘dokter’, ‘spesialis’, ‘tht’]
stemming	[‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘sakit’, ‘tht’, ‘masyarakat’,‘tahu’, ‘jenis’, ‘sakit’, ‘tht’, ‘derita’, ‘dokter’, ‘spesialis’, ‘tht’]

C. *Term Weighting TF-IDF*

The term weighting process is used to give weight to the value of each word. The term or word weighting method used in this study is TF-IDF. The example of the TF-IDF calculation process using the documents in Tabel III, the stemming section, is shown in Table IV.

TABLE IV. RESULT OF WEIGHTING TERM TF-IDF

Term	tf	D	D/df	log (IDF)+1	W= tf * (IDF+1)
	D1				D1
datang	1	1	1	1	1
derita	1	1	1	1	1
diagnosis	1	1	1	1	1
dokter	1	1	1	1	1
jenis	2	1	1	1	2
masyarakat	1	1	1	1	1
pakar	1	1	1	1	1
sakit	2	1	1	1	2
sistem	1	1	1	1	1
spesialis	1	1	1	1	1
tht	3	1	1	1	3

D. *Naïve Bayes Method Classification*

At this stage, the classification is carried out using the naïve Bayes method by comparing the performance using TF-IDF weighting and without TF-IDF weighting using equation (2).

E. *Result Evaluation*

At this stage, results are evaluated based on accuracy, precision, and sensitivity using the confusion matrix table shown in Table V, VI, and VII.

TABLE V. CONFUSION MATRIX OF NAÏVE BAYES WITH TF - IDF

Actual	**Predicted**			**Sensitivity**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	31	0	6	83.78%
Multimedia	0	18	9	66.67%
RPL	4	1	45	90%
Precision	88.57%	94.74%	75%	

TABLE VI. CONFUSION MATRIX OF NAÏVE BAYES WITHOUT TF - IDF

Actual	**Predicted**			**Sensitivity**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	33	0	4	89.19%
Multimedia	0	26	1	96.29%
RPL	5	2	43	86%
Precision	86.84%	92.86%	89.58%	

TABLE VII. PERFORMANCE RESULT OF NAÏVE BAYES METHOD

Performance	With TF - IDF	Without TF - IDF
Accuracy	81.74%	**88.69%**
Precision	86.1%	**89.76%**
Sensitivity	80.15%	**90.49%**

Based on the results of the tests shown in Table VII, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces **88.69%** accuracy, **89.76%** precision, and **90.49%** sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.

IV. CONCLUSION

Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces **88.69%** accuracy, **89.76%** precision, and **90.49%** sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract. The suggestions for further research can use feature selection methods such as chi-square to improve the performance of the naïve Bayes method.

REFERENCES

[1] D. Adhe, C. Rachman, R. Goejantoro, and D. Tisna, "Implementation Of Text Mining For Grouping Thesis Documents Using K-Means Clustering," *J. EKSPONENSIAL*, vol. 11, no. 2, pp. 167–174, 2020.

[2] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, Feb. 2021, doi: 10.11591/ijece.v11i1.pp664-670.

[3] M. Sholehhudin, M. Fauzi Ali, and S. Adinugroho, "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya)," vol. 2, no. 11, pp. 5518–5524, 2018.

[4] L. Zahrotun, N. H. Putri, and A. Nur Khusna, "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesis Titles," in *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Oct. 2018, pp. 1–4, doi: 10.1109/TSSA.2018.8708817.

[5] D. M. U. Atmaja and R. Mandala, "Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 2, pp. 1–6, Aug. 2020, doi: 10.33021/itfs.v4i2.1182.

[6] M. Eminağaoğlu and Y. Gökşen, "A New Similarity Measure for Document Classification and Text Mining," *KnE Soc. Sci.*, vol. 2019, pp. 353–366, Jan. 2020, doi: 10.18502/kss.v4i1.5999.

[7] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," *Stud. Comput. Intell.*, vol. 740, no. November, pp. 373–397, 2018, doi: 10.1007/978-3-319-67056-0_18.

[8] R. Rismanto, A. Rachmad Syulistyo, and B. P. Citra Agusta, "Research Supervisor Recommendation System Based on Topic Conformity," *Int. J. Mod. Educ. Comput. Sci.*, vol. 12, no. 1, pp. 26–34, Feb. 2020, doi: 10.5815/ijmecs.2020.01.04.

[9] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/view/10955/6659.

[10] A. Solichin, "Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation," in *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Sep. 2019, pp. 217–222, doi: 10.23919/EECSI48112.2019.8977081.

[11] S. Sulova, L. Todoranova, B. Penchev, and R. Nacheva, "Using Text Mining to Classify Research Papers," in *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*, Jun. 2017, vol. 17, no. 21, pp. 647–654, doi: 10.5593/sgem2017/21/S07.083.

[12] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.

[13] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, 2017, pp. 1–5, doi: 10.1109/ICCCI.2017.8117734.

[14] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano, "Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers," *Neurocomputing*, vol. 92, pp. 98–115, Sep. 2012, doi: 10.1016/j.neucom.2012.01.030.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 513–520.

The Abstract of Thesis Classifier by Using Naive Bayes Method

ORIGINALITY REPORT

23%
SIMILARITY INDEX

PRIMARY SOURCES

1	publishing-widyagama.ac.id Internet	46 words — 2%
2	Bambang Harjito, Ardhi Wijayanto, Kuni Nur Aini, Budi Murtiyas. "Comparison of Multinomial Naïve Bayes with K-Nearest Neighbors, Support Vector Machine and Random Forest for Classification of “Network Attacks” Document", 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019 Crossref	42 words — 1%
3	iptek.its.ac.id Internet	40 words — 1%
4	Wang, Zhaoxia, Victor Joo, Chuan Tong, Xin Xin, and Hoong Chor Chin. "Anomaly Detection through Enhanced Sentiment Analysis on Social Media Data", 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, 2014. Crossref	36 words — 1%
5	Fakhruddin Farid Irfani, M. Ali Fauzi, Yuita Arum Sari. "News Classification on Twitter Using Naive Bayes and Hypernym-Hyponym Based Feature Expansion", 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018 Crossref	34 words — 1%

6	Lucan Yance Nanlohy, Eko Mulyanto Yuniarno, Supeno Mardi Susiki Nugroho. "Classification of Public Complaint Data in SMS Complaint Using Naive Bayes Multinomial Method", 2020 4th International Conference on Vocational Education and Training (ICOVET), 2020 Crossref	32 words — 1%
7	Joel Reed, Yu Jiao, Thomas Potok, Brian Klump, Mark Elmore, Ali Hurson. "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams", 2006 5th International Conference on Machine Learning and Applications (ICMLA'06), 2006 Crossref	31 words — 1%
8	Ishfaq Ali, Muhammad Asif, Muhammad Shahbaz, Adnan Khalid, Mariam Rehman, Aziz Guergachi. "Text Categorization approach for Secure Design Pattern Selection using Software Requirement Specification", IEEE Access, 2018 Crossref	30 words — 1%
9	www.hindawi.com Internet	29 words — 1%
10	Saesarinda Rahmike Juwita, Sukmawati Nur Endah. "Classification of Indonesian Music Using the Convolutional Neural Network Method", 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), 2019 Crossref	28 words — 1%
11	Lecture Notes in Computer Science, 2002. Crossref	26 words — 1%
12	repository.unika.ac.id Internet	26 words — 1%

13 Internet

25 words — 1%

14 Ghosh, Madhumala, Devkumar Das, Chandan Chakraborty, and Ajoy Kumar Ray. "Quantitative characterisation of Plasmodium vivax in infected erythrocytes: a textural approach", International Journal of Artificial Intelligence and Soft Computing, 2013.
Crossref

21 words — 1%

15 Y B N D Artissa, I Asror, S A Faraby. "Personality Classification based on Facebook status text using Multinomial Naïve Bayes method", Journal of Physics: Conference Series, 2019
Crossref

21 words — 1%

16 e-journal.president.ac.id
Internet

17 words — 1%

17 ijair.id
Internet

16 words — 1%

18 pure.ulster.ac.uk
Internet

16 words — 1%

19 M. Dolfi, I. Colzi, S. Morosi, E. Masi, S. Mancuso, E. Del Re, F. Francini, R. Magliacani. "Plant electrical activity analysis for ozone pollution critical level detection", 2015 23rd European Signal Processing Conference (EUSIPCO), 2015
Crossref

15 words — 1%

20 Lertnattee, V.. "Class normalization in centroid-based text categorization", Information Sciences, 20060622
Crossref

14 words — < 1%

21	María Cora Urdaneta-Ponte, Amaia Mendez-Zorrilla, Ibon Oleagordia-Ruiz. "Recommendation Systems for Education: Systematic Review", Electronics, 2021 Crossref	14 words — < 1%
22	Utomo Pujianto, Muhammad Fahmi Hidayat, Harits Ar Rosyid. "Text Difficulty Classification Based on Lexile Levels Using K-Means Clustering and Multinomial Naive Bayes", 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019 Crossref	13 words — < 1%
23	ceur-ws.org Internet	13 words — < 1%
24	Achmad Jauhari, Ika Oktavia Suzanti, Yoga Dwitya Pramudita, Husni, Nourma Pangestika Wulan Diantisari. "Enhanced Confix Stripping Stemmer And Cosine Similarity For Search Engine in The Holy Qur'an Translation", 2020 6th Information Technology International Seminar (ITIS), 2020 Crossref	12 words — < 1%
25	arxiv.org Internet	11 words — < 1%
26	knepublishing.com Internet	11 words — < 1%
27	www.slideshare.net Internet	11 words — < 1%
28	J. Guruprakash, Srinivas Koppu. "EC-ElGamal and Genetic Algorithm-Based Enhancement for	9 words — < 1%

Lightweight Scalable Blockchain in IoT Domain", IEEE Access, 2020
Crossref

29 Riska Haerani, Lalu Zazuli Azhar Mardedi. "Analisa Penerapan Hierarchical Tokken Buket Untuk Optimalisasi Management Bandwith Pada Server Ubuntu", Jurnal Bumigora Information Technology (BITe), 2020
Crossref
9 words — < 1%

30 www.sgem.org
Internet
9 words — < 1%

31 Hairani Hairani, Muhammad Innuddin, Majid Rahardi. "Accuracy Enhancement of Correlated Naive Bayes Method by Using Correlation Feature Selection (CFS) for Health Data Classification", 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020
Crossref
8 words — < 1%

32 www.coursehero.com
Internet
8 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		

ICoCSIM 2021 (author)

Help / Log out

Submission 8 Conference News EasyChair

ICoCSIM 2021 Submission 8

Submission information updates are disabled.

For all questions related to processing your submission you should contact the conference organizers. Click here to see information about this conference.

All **reviews sent to you** can be found at the bottom of this page.

Submission 8	
Title	Thesis Topic Classification Based on Abstract Using the Naïve Bayes Method
Paper:	(Jun 08, 02:42 GMT) (previous versions)
Author keywords	naïve bayes TF-IDF weighting asbtrak classification text mining
EasyChair keyphrases	naive baye method (364), tf idf weighting (221), naive baye (215), text pre processing (142), baye method (115), design bumigora university (95), bumigora university mataram (95), text mining (80), thesis topic (80), weighting tf idf (63), tf idf method (63), pemberian pakan ternak (63), multinomial naive baye (63), term weighting tf (63), computer science (60), term weighting (50), teknologi augmented reality (47), sistem pakar diagnosis (47), penelitian ini adalah (47), pre processing (46), naive baye method classification (40), data collection (40), decision tree (40), actual sensitivity jaringan multimedia (40), multinomial naive baye method (40), tujuan pembuatan sistem pakar (40), bumigora university (40), true jaringan (40), predicted actual sensitivity jaringan (40)
Abstract	The thesis is a requirement for graduation from Bumigora university. The final year student's problem is determining the research topic because the undergraduate thesis collection of Computer Science is not grouped or classified based on student competencies. The purpose of this study was to compare the performance of the naïve Bayes method with TFIDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The stages of this research are data collection, text pre-processing, term weighting with TF-IDF and without TF-IDF, Naïve Bayes method implementation, and result

	evaluation. Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces 88.69% accuracy, 89.76% precision, and 90.49% sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.
Submitted	May 28, 02:02 GMT
Last update	May 28, 02:02 GMT

Authors

first name	last name	email	country	affiliation	Web page	corresponding?
Hairani	Hairani	hairani@universitasbumigora.ac.id	Indonesia	Universitas Bumigora		✓
Ahmad Islahul	Wathan	wathanruna@gmail.com	Indonesia	Universitas Bumigora		
Kurniadin Abd	Latif	kurniadin@universitasbumigora.ac.id	Indonesia	Universitas Bumigora		
Khairan	Marzuki	khairan@universitasbumigora.ac.id	Indonesia	Universitas Bumigora		
Muhammad	Zulfikri	mzulfikri@universitasbumigora.ac.id	Indonesia	Universitas Bumigora		
Anthony	Anggrawan	anthony.anggrawan@universitasbumigora.ac.id	Indonesia	Universitas Bumigora		

Reviews

Review 1

Detailed Comments

It's quite a nice sequence and a good idea for a beginner to study. I would like you to follow the comments, to be more understandable for readers.

Title: Which one you would like to classifier the thesis or Abstract? I suggest for you this title.
The Abstract of thesis classifier by using Naïve Bayes method

Abstract: thesis topics based on the abstract…. If you try to extract the topic will have a lot of features that would be a huge dataset used, but in this work seen only used the abstract as three specific topics, Why?…. Please recheck again.

With TF-IDF and without TF-IDF…. You mentioned twice…. Please recheck again and rewrite.

Need to explain more about thesis topic what is it?
Keywords: check what you mean by asbtrak? Is it Abstract? Also, why use text mining as long as you didn't use it in your work properly? Please recheck….

Introduction: The end of it adds…. Paper organized by….
One of the solutions offered by this research??? In this research or Paper?? Please rewrite probably………. is to use the concept of text mining. Previous research used various methods for text mining-based thesis document analysis such as the k-means method……..…… In this Paragraph are you used text mining to cover the words as Vector from the topic??

Research methodology: Please describe your work steps how are going on, and refigure to show the pre-processing steps as you mentioned in this paper.
Usually, 10 Fold Cross Validation used with TF-IDF by using dataset an example 200 will be 150 training and 50 testings? So how you applied to get good accuracy with NB please prove...

Text Pre-Processing: Please make a subsection for each step to show your work off about pre-processing. 1. Tokenization 2. Stop word removal 3. Stemming including a figure for each one to show your work how to process it…. Prove it...

C. Term Weighting TF-IDF
Does the TF-IDF method combine two concepts?... What you mean by that, are you trying using concepts, but you extract from were to combine which tool are you used to combine from? Because you work extract topics as words… Please Prove that….
Data classified by the naïve Bayes method are grouped into training and testing data first. Multinomial (What you mean by multinomial) because this usually used for word pairs as using Multinomial logistic regression?

Please provide a figure that shows the input and output of your TF-IDF by using NB and how it affects your topics...

Table 1: What you mean by Confusion Matrix did you coding as a table? Please prove or add your epscode

The equations (5), (6), (7) I didn't see your equation are you apply your method, is that your own equation created?... Its blur does not show anything…. Please recheck.

I didn't see any section of Related Work of previous studies table?? What is the Research Gap in your work? Please must provide a table and section explain about your related work too….

Table 2: Why you extract only on the topic? How about others? And why you used Cross-Validation to get only one topic??.... Please explain?.

D. Naïve Bayes Method Classification
At this stage, the classification is carried out using the naïve Bayes multinomial?? In your paper used NB and now you

	mentioned the naïve Bayes multinomial? What are you trying to do with multinomial? Please explain…. Conclusion: please add your future work...chi-square is used for two pairs of words as a topic, did you? Please remove your Acknowledgment this is for funds if you have to add….. The paper needs proofreading. Good Luck!

Review 2	
Detailed Comments	- it needs proofreading. - References should be relevant, recent and readily retrievable

Gmail

Hairani Hairani <hairani@universitasbumigora.ac.id>

IEEE ICoCSIM 2021 notification for paper 8

1 message

ICoCSIM 2021 <icocsim2021@easychair.org> Fri, Jun 25, 2021 at 10:24 AM
To: Hairani Hairani <hairani@universitasbumigora.ac.id>

Dear Dr. Hairani Hairani:

Apologies for the delay in announcing the acceptance for the submitted paper. We have received overwhelming numbers of the submissions, which made the selection quite competitive.

The review process for the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) has been completed. Over 200 international experts volunteered to perform reviews with a minimum of 2 reviews per paper and a maximum of 5 reviews per reviewers.

Based on the recommendations of the reviewers and the Program Committee, I am very pleased to inform you that your paper “Thesis Topic Classification Based on Abstract Using the Naïve Bayes Method” for ICSECS-ICOCSIM 2021 has been accepted for the presentation and proceedings publication.

You are cordially invited to present the paper at ICSECS-ICOCSIM 2021 to be held online between 24-26 August, 2021. All accepted papers MUST be presented through ORAL presentation subjected to the final arrangement.

This notification email serves as our formal acceptance of your paper as well as an invitation to present your work at ICSECS-ICOCSIM 2021.

Kindly read comments from the reviewers and make the necessary corrections where appropriate as suggested for the camera-ready submission. The reviewers' comments are included at the end of this notification email. Please list down the comments and the correction you have done in a TABLE using a SEPARATE FILE and upload it together with the camera-ready. Your paper may still be REJECTED if these are not followed.

Please also read instructions from the conference website for all the necessary requirements for registration at https://icocsim.ump.edu.my/. After completing the conference fee payment, please complete your registration process at the following link. WE ACCEPT THE REGISTRATION PAYMENT UNTIL 15th JULY 2021.
Registration link: https://forms.gle/Dj8hmWf8hq1YvHpGA

The acceptance of your paper is made with the understanding that at least one of the authors from the respective paper REGISTER with the necessary registration fee and attend the conference to present the paper. Without proof of payment, your registration will not be processed. Should you have any question regarding this matter, please email to ismalina@ump.edu.my

I would like to take this opportunity to thank you for choosing ICSECS-ICOCSIM 2021 to present your research results. We are looking forward to seeing you virtually on the conference day.

Regards,
Jamaluddin Salim
General Chair of ICSECS-ICOCSIM 2021

===============

Reviews
===============

General Comments:

1. Maximum number of pages should be 6 (SIX) only.
2. Maximum plagiarism (Turnitin) similarities should be below 30%. (Make sure to set Turnitin class to “no repository” before you upload into Turnitin.)
3. Please make sure you follow exactly the IEEE format for the camera-ready paper. The template can be found here: https://www.ieee.org/conferences/publishing/templates.html
4. Upload your final camera-ready paper (Word or pdf format) and fill the information related to IEEE copyright form by 4 July 2021 following this link: https://forms.gle/aYN3QyVWRnQd6fk29
5. The link of copyright form will be sent to you after the registration and payment have been made.

SUBMISSION: 8
TITLE: Thesis Topic Classification Based on Abstract Using the Naïve Bayes Method

----------------------- REVIEW 1 ---------------------
SUBMISSION: 8
TITLE: Thesis Topic Classification Based on Abstract Using the Naïve Bayes Method
AUTHORS: Hairani Hairani, Ahmad Islahul Wathan, Kurniadin Abd Latif, Khairan Marzuki, Muhammad Zulfikri and Anthony Anggrawan

----------- Detailed Comments -----------
It's quite a nice sequence and a good idea for a beginner to study. I would like you to follow the comments, to be more understandable for readers.

Title: Which one you would like to classifier the thesis or Abstract? I suggest for you this title.
The Abstract of thesis classifier by using Naïve Bayes method

Abstract: thesis topics based on the abstract…. If you try to extract the topic will have a lot of features that would be a huge dataset used, but in this work seen only used the abstract as three specific topics, Why?…. Please recheck again.

With TF-IDF and without TF-IDF…. You mentioned twice…. Please recheck again and rewrite.
Need to explain more about thesis topic what is it?
Keywords: check what you mean by asbtrak? Is it Abstract? Also, why use text mining as long as you didn’t use it in your work properly? Please recheck….

Introduction: The end of it adds…. Paper organized by….
One of the solutions offered by this research??? In this research or Paper?? Please rewrite probably………. is to use the concept of text mining. Previous research used various methods for text mining-based thesis document analysis such as the k-means method……..…… In this Paragraph are you used text mining to cover the words as Vector from the topic??

Research methodology: Please describe your work steps how are going on, and refigure to show the pre-processing steps as you mentioned in this paper.
Usually, 10 Fold Cross Validation used with TF-IDF by using dataset an example 200 will be 150 training and 50 testings? So how you applied to get good accuracy with NB please prove...

Text Pre-Processing: Please make a subsection for each step to show your work off about pre-processing. 1. Tokenization 2. Stop word removal 3. Stemming including a figure for each one to show your work how to process it…. Prove it...

C. Term Weighting TF-IDF
Does the TF-IDF method combine two concepts?... What you mean by that, are you trying using concepts, but you extract from were to combine which tool are you used to combine from? Because you work extract topics as words… Please Prove that….
Data classified by the naïve Bayes method are grouped into training and testing data first. Multinomial (What you mean by multinomial) because this usually used

for word pairs as using Multinomial logistic regression?

Please provide a figure that shows the input and output of your TF-IDF by using NB and how it affects your topics...

Table 1: What you mean by Confusion Matrix did you coding as a table? Please prove or add your epscode

The equations (5), (6), (7) I didn't see your equation are you apply your method, is that your own equation created?... Its blur does not show anything…. Please recheck.

I didn't see any section of Related Work of previous studies table?? What is the Research Gap in your work? Please must provide a table and section explain about your related work too….

Table 2: Why you extract only on the topic? How about others? And why you used Cross-Validation to get only one topic??.... Please explain?.

D. Naïve Bayes Method Classification
At this stage, the classification is carried out using the naïve Bayes multinomial?? In your paper used NB and now you mentioned the naïve Bayes multinomial? What are you trying to do with multinomial? Please explain….

Conclusion: please add your future work...chi-square is used for two pairs of words as a topic, did you?

Please remove your Acknowledgment this is for funds if you have to add…..

The paper needs proofreading.

Good Luck!

----------------------- REVIEW 2 ---------------------
SUBMISSION: 8
TITLE: Thesis Topic Classification Based on Abstract Using the Naïve Bayes Method
AUTHORS: Hairani Hairani, Ahmad Islahul Wathan, Kurniadin Abd Latif, Khairan Marzuki, Muhammad Zulfikri and Anthony Anggrawan

----------- Detailed Comments -----------
- it needs proofreading.

- References should be relevant, recent and readily retrievable

CONFERENCE PROCEEDINGS

Proceedings

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management

ICSECS-ICOCSIM 2021

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) | 978-1-6654-1407-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSECS52883.2021.00001

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) | 978-1-6654-1407-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSECS52883.2021.0

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)

ICSECS-ICOCSIM 2021

Table of Contents

1A: IoT, Big Data Analytics, Artificial Intelligence

2D: Software Engineering, Knowledge Engineering

2E: IoT, Artificial Intelligence, Image Processing, E- Learning

2F: Information Management, E-Learning

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) | 978-1-6654-1407-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSECS52883.2021.00006

Committee Members

GENERAL CHAIR
DR. JAMALUDIN SALLIM

CHAIR
DR. FAUZIAH ZAINUDDIN (ICSECS)
DR. MOHD AZWAN HAMZA (ICoCSIM)

SECRETARY
DR. MOHD ZAMRI OSMAN
DR. NOOR AZIDA SAHABUDIN
DR. NUR HAFIEZA BINTI ISMAIL

TREASURER
DR. NUR SHAMSIAH ABDUL RAHMAN

PROCEEDING COMMITTEE
ASSOC. PROF. DR. ARAFATUR RAHMAN
DR. SYIFAK IZHAR HISHAM

MEMBERS
DR. ROZLINA MOHAMED
DR. YUSNITA MUHAMAD NOOR
DR. MOHD IZHAM MOHD JAYA
TS. DR. MRITHA RAMALINGAM
TS. DR. KOHBALAN A/L MOORTHY
DR. ZAHIAN BINTI ISMAIL

TECHNICAL COMMITTEE
DR. ZAFRIL RIZAL AZMI
TS. DR. TAHA HUSSEIN ALAALDEEN RASSEM

MEMBERS
DR. LIEW SIAU CHIN
TS. DR MOHD ARFIAN ISMAIL
DR. NOR SYAHIDATUL NADIAH ISMAIL
DR. SALWANA MOHAMAD @ ASMARA
DR. ABDULLAH NASSER

INTERNATIONAL AFFAIRS
ASSOC. PROF. DR. ABDUL RAHMAN ALSEWARI (ICSECS)
TS. DR. FERDA ERNAWAN (ICSECS)
DR. AHMAD FAKHRI BIN AB. NASIR (ICoCSIM)

INDUSTRIAL & ALUMNI
DR. ZALILI MUSA
DR. SYAFIQ FAUZI KAMARULZAMAN

SPONSORSHIP COMMITTEE
DR. AHMAD FIRDAUS ZAINAL ABIDIN
DR. AZLEE ZABIDI

MEMBERS
DR. RAHMAH MOKHTAR
DR. SURAYA ABU BAKAR
ASSOC. PROF. DR MOHAMED ARIFF AMEEDEEN
DR NGAHZAIFA AB GHANI
TS DR AWANIS ROMLI
ARIFIN SALLEH

HOSPITALITY, LOGISTIC & TRANSPORTATION COMMITTEE
DR. LUHUR BAYUAJI
DR. ANIS FARIHAN MAT RAFFEI
AMINATUL NOR MOHAMED SAID

MEMBERS
ROSLINA MOHD SIDEK

MEDIA & PUBLICITY COMMITTEE
DR. DANAKORN NINCAREAN

MEMBERS
NURLINDA KADRI
IZYAN HUSNA IBRAHIM

PROMOTION COMMITTEE
DR. NUR SHAZWANI KAMARUDIN
DR. NOR SARADATUL AKMAR

EVENT MANAGEMENT
DR. MOHD FAIZAL AB RAZAK
NORANIZA SAMAT

SECRETARIAT
ISMALINA MOHD ISAH
NORAINI MD ZAMRI
NOOR ASHIKIN RAMLY
NOOR AFTALINA OMAR
DARWINA RASTAM TAN
SYARIFAH AZLIN SYED YUSOP
SURIANA ABIDIN

TECHNICAL SUPPORT
CHE YAHAYA YAAKOB
DR. MUHAMMAD ‘ARIF BIN MOHAMAD
TS. DR. HOH WEI SIANG
IMRAN EDZEREIQ
SYAHRULANUAR NGAH
MAHMUD ABDUL SAMAD
AMIRUL HUSNI ABDUL GHAFFAR
MOHD AMERUL SHUIB
MOHD NAIM GATI @MOHD GATI
MUHAMMAD TAUFIK MOHAMAD REFFIN
MOHD FAIZUL GHAFAR
ABDUL RAHMAN ABDUL KARIM
WAN MD NAHARRUDDIN WAN ZULKIFLI
AHMAD MUSTAQIM MOHAMAD GHANI
KHAIRIL CHAIRY
MASTURA SARKON
RUZAINAH ABDULLAH
KHAIRUN NISSAK ABDULLAH

PROGRAM & PROTOCOL COMMITTEE
DR. JUNAIDA SULAIMAN
TS. DR SURYANTI AWANG
DR. WAN ISNI SOFIAH WAN DIN
DR. ABDUL SAHLI FAKHARUDIN
DR. ZURIANI MUSTAFFA
DR NOORLIN MOHD ALI
TS. DR NOORHUZAIMI@KARIMAH
DR. BARIAH YUSOF
DR. NABILAH FILZAH BINTI MOHD RADZUAN
DR. SITI SUHAILA BINTI ABDUL HAMID

The Abstract of Thesis Classifier by Using Naive Bayes Method

Hairani Hairani
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
hairani@universitasbumigora.ac.id

Anthony Anggrawan
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
anthony.anggrawan@universitasbumigora.ac.id

Ahmad Islahul Wathan
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
wathanruna@gmail.com

Kurniadin Abd Latif
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
kurniadin@universitasbumigora.ac.id

Khairan Marzuki
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
khairan@universitasbumigora.ac.id

Muhammad Zulfikri
Faculty of Engineering and Design
Universitas Bumigora
Mataram, Indonesia
mzulfikri@universitasbumigora.ac.id

***Abstract*— The thesis is a requirement for graduation from Bumigora university. The final year student's problem is determining the research topic because the undergraduate thesis collection of Computer Science is not grouped or classified based on student competencies. The purpose of this study was to compare the performance of the naïve Bayes method with TF-IDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The stages of this research are data collection, text pre-processing, term weighting with TF-IDF and without TF-IDF, Naïve Bayes method implementation, and result evaluation. Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces 88.69% accuracy, 89.76% precision, and 90.49% sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.**

Keywords—naïve bayes, TF-IDF weighting, asbtract classification

I. Introduction

The thesis is one of the graduation requirements for undergraduate Computer Science students at Bumigora University. Students can start working on their thesis if the research topic has been approved through a synopsis exam. So far, students have difficulties in determining the proposed thesis topic. One of the difficulties is because the existing collection of an undergraduate thesis in Computer Science is not grouped or classified based on student competencies. Automatic thesis grouping or classification of topics is one solution that can make it easier for students to find references to research titles based on their competence. The competencies of students in the S1 Computer Science program at Bumigora university are computer networks, multimedia, and software engineering (RPL).

One of the solutions offered by this paper is to use the concept of text mining. Previous research used various methods for text mining-based thesis document analysis such as the k-means method [1]–[4], K-Nearest Neighbor [5]–[7], Cosine Similiarity [8], [9], Decision Tree and Naïve Bayes [10], SVM and Naïve Bayes [11]. Research [10] compared Decision Trees, Naïve Bayes, and k-NN methods to predict thesis graduation. Based on the results of his research, the k-NN method has the best accuracy compared to the decision tree and naïve Bayes methods at 80.39%. Research [4] used the k-means method for grouping thesis titles. Before grouping, the first weighting of words is carried out using the TF - IDF method. Research [9] uses the cosine similarity method for the classification of thesis documents. Before grouping, the first weighting of words is carried out using the TF - IDF method.

Based on previous research, there is a difference made with this research, namely the research carried out a classification of thesis topics based on the abstract using the naïve Bayes method and also using the k-fold cross-validation test method. The aim is to compare the performance of the naïve Bayes method with TF-IDF weighting and without TF-IDF weighting for the classification of thesis topics based on the abstract. The performance used in this study is accuracy, precision, and sensitivity.

II. Research Method

The stages used in this study are shown in Figure 1.

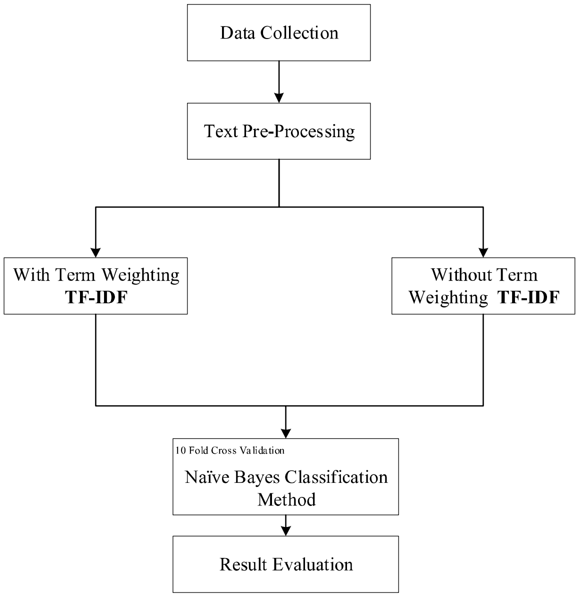
Data Collection
Text Pre-Processing
With Term Weighting TF-IDF
Without Term Weighting TF-IDF
10 Fold Cross Validation
Naïve Bayes Classification Method
Result Evaluation

Figure 1. Research Methodology

A. Data Collection

The data used in this study are the thesis abstracts of 2020 graduate computer science undergraduate students obtained

978-1-6654-1407-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICSECS52883.2021.00063

from www.repository.universitasbumigora.ac.id. The data collected were 115 thesis abstract data, consisting of 36 topics of computer networks, 23 multimedia, and 55 software engineering (RPL).

B. *Text Pre-Processing*

Text pre-processing used to get quality data. The classification was carried out using the naïve Bayes method. The text pre-processing technique used in this study consists of case folding, tokenization, stop word removal. and stemming [12]. Case folding is used to change text to lowercase. Tokenization is used to separate text into tokens. Stopword removal is used to remove unnecessary words such as conjunctions. Stemming is used to change all words that have affixes into basic words.

C. *Term Weighting TF-IDF*

The term weighting process is used to give a weight value to each word. The term weighting method used in this study is the Term Frequency - Inverse Document Frequency (TF-IDF). The TF-IDF method combines two concepts, namely TF and IDF. TF looks for the occurrence value of terms in related documents, the more occurrences of terms in the related document, the better. Meanwhile, the IDF concept is inversely proportional to the TF method, the less frequently the terms appear in all documents the better. TF - IDF method is calculated using equation (1) [13].

$$W_{ij} = tf_{ij}\ x\ idf_j = tf_{ij}\ x\ \log\left(\frac{N}{df_j}\right) \qquad (1)$$

W_{ij} is the weight of term j to document i. tf_{ij} is the number of occurrences of term j in the document d. N is the number of documents, and df_i is the number of occurrences of term j throughout the document.

D. *Naïve Bayes Method Classification*

At this stage, the classification is carried out using the naïve Bayes method. The basic concept of the naïve Bayes method is a probability-based classification method that assumes independence from the dependent variable and is also a conditional model based on the Bayes theorem [14][15]. The Naïve Bayes method which is calculated based on equation (2).

$$P(c \mid term\ document\ d) = P(c)\ x\ \mathrm{P}(\mathrm{t}_1|\mathrm{c})\ x\ \mathrm{P}(\mathrm{t}_2|\mathrm{c})\ x\ \mathrm{P}(\mathrm{t}_n|\mathrm{c}) \qquad (2)$$

$P(c)$ is the prior probability of class c. $P(c \mid term\ document\ d)$ is the probability of the appearance of a term in document d including class c. $P(t_n \mid c)$ is the probability of occurrence of term n known to class c.

The process of calculating the prior probability for class c uses equation (3).

$$P(c) = \frac{N_C}{N} \qquad (3)$$

N_c is the number of class c in all documents, while N is the total number of documents. The calculation of the probability of occurrence of term n is calculated using equation (4) involving the laplacian technique.

$$P(t_n \mid c) = \frac{count(t_n, c) + 1}{count(c) + |v|} \qquad (4)$$

$count(t_n, c)$ is the number of terms t_n appearing in the training data with class c. $count(c)$ is the number of terms in the class training data c. weighting is used to give weight to the value of each word. is the number of terms in the training data. V is the number of terms in the training data.

Data classified by the multinomial naïve Bayes method are grouped into training and testing data first. The distribution of training and testing data in this study uses the k-fold cross-validation method by dividing the data as much as the specified k. Each fold can be used as training and testing data in turn. This research uses 10 fold data validation method.

E. *Result Evaluation*

At this stage, the results are evaluated based on accuracy, precision, and sensitivity using the confusion matrix table shown in Table Tabel I.

TABLE I. CONFUSION MATRIX

Actual	**Predicted**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Evaluation of results based on accuracy, precision, and sensitivity using equations (5), (6), and (7).

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \qquad (5)$$

$$\text{Precision}=\frac{TP}{TP+FP} \qquad (6)$$

$$\text{Sensitivity}=\frac{TP}{\text{TP + FN}} \qquad (7)$$

III. RESULT AND DISCUSSION

A. *Data Collection*

The data used in this study are the thesis abstracts of 2020 graduate computer science undergraduate students obtained from www.repository.universitasbumigora.ac.id. The data collected were 115 thesis abstract data, consisting of 36 topics of computer networks, 23 multimedia, and 55 software engineering (RPL). The sample abstract data of this research thesis is shown in Table II.

TABLE II. THESIS ABSTRACT DATASET

No	***Abstract***	***Topic***
1.	Perkembangan teknologi informasi sangat cepat seperti Internet of Things (IoT), dimana seseorang dapat melakukan segala aktivitasnya dengan mudah dengan mengandalkan sistem Internet of Things (IoT). Seiring dengan perkembangan zaman maka semakin canggih teknologi yang dihasilkan baik digunakan sebagai hal yang positif maupun melakukan hal yang negatif, tak terkecuali pada system peternakan sehingga perlu mengembangkan teknologi untuk manajemen pakan ternakkhususnya hewan ternak ayam broiler.Pengembangkan sistem menggunakan sistem Internet of Things dan sistem penjadwalan otomasi dimana sistem Internet of Things (IoT) adalah sistem yang berfungsi melakukan kontroller pada alat alat elektronik. Metodelogi	Jaringan

No	*Abstract*	*Topic*
	penelitian yang digunakan adalah Network Development Life Cycle (NDLC), terdiri dari; analisis, desain, prototype dan ujicoba. Pada tahap analisis memuat tentang pengumpulan data, tahap desain memuat rancangan sistem pemberian pakan ternak, prototyping memuat instalasi konfigurasi dan membangun kerangka sistem pakan ternak. Ujicoba memuat tentang pengujian sistem pemberian pakan ternak secara otomatis atau terjadwal. Kesimpulan dari penelitian ini adalah menginplementasi Sever VPS dengan sistem nodemcu dalam pemberian pakan ternak berbasis Internet of Things (IoT) untuk efisiensi dalam pemberian pakan ternak ayam.	

B. *Text Pre-Processing*

Text pre-processing used to get quality data. The classification was carried out using the naïve Bayes method. The text pre-processing technique used in this study consists of case folding, tokenization, stop word removal. and stemming. The examples of text pre-processing stages are shown in Table III.

TABLE III. EXAMPLE OF TEXT PREPROCESSING

Pre-processing	Result
Data Original	Tujuan pembuatan sistem pakar diagnosis jenis penyakit THT adalah memudahkan masyarakat umum untuk mengetahui jenis penyakit THT diderita tanpa perlu datang ke dokter spesialis THT
Case Folding	tujuan pembuatan sistem pakar diagnosis jenis penyakit tht adalah memudahkan masyarakat umum untuk mengetahui jenis penyakit tht diderita tanpa perlu datang ke dokter spesialis tht
Tokenization	[‘tujuan’, ‘pembuatan’, ‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘penyakit’, ‘tht’, ‘adalah’, ‘memudahkan’, ‘masyarakat’, ‘umum’, ‘untuk’, ‘mengetahui’, ‘jenis’, ‘penyakit’, ‘tht’, ‘diderita’, ‘tanpa’, ‘perlu’, ‘datang’, ‘ke’, ‘dokter’, ‘spesialis’, ‘tht’]
stop word removal	[‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘penyakit’, ‘tht’, ‘masyarakat’, ‘mengetahui’, ‘jenis’, ‘penyakit’, ‘tht’, ‘diderita’, ‘dokter’, ‘spesialis’, ‘tht’]
stemming	[‘sistem’, ‘pakar’, ‘diagnosis’, ‘jenis’ ‘sakit’, ‘tht’, ‘masyarakat’,‘tahu’, ‘jenis’, ‘sakit’, ‘tht’, ‘derita’, ‘dokter’, ‘spesialis’, ‘tht’]

C. *Term Weighting TF-IDF*

The term weighting process is used to give weight to the value of each word. The term or word weighting method used in this study is TF-IDF. The example of the TF-IDF calculation process using the documents in Tabel III, the stemming section, is shown in Table IV.

TABLE IV. RESULT OF WEIGHTING TERM TF-IDF

Term	tf	D	D/df	log (IDF)+1	W= tf * (IDF+1)
	D1				D1
datang	1	1	1	1	1
derita	1	1	1	1	1
diagnosis	1	1	1	1	1
dokter	1	1	1	1	1
jenis	2	1	1	1	2
masyarakat	1	1	1	1	1
pakar	1	1	1	1	1
sakit	2	1	1	1	2
sistem	1	1	1	1	1
spesialis	1	1	1	1	1
tht	3	1	1	1	3

D. *Naïve Bayes Method Classification*

At this stage, the classification is carried out using the naïve Bayes method by comparing the performance using TF-IDF weighting and without TF-IDF weighting using equation (2).

E. *Result Evaluation*

At this stage, results are evaluated based on accuracy, precision, and sensitivity using the confusion matrix table shown in Table V, VI, and VII.

TABLE V. CONFUSION MATRIX OF NAÏVE BAYES WITH TF - IDF

Actual	**Predicted**			**Sensitivity**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	31	0	6	83.78%
Multimedia	0	18	9	66.67%
RPL	4	1	45	90%
Precision	88.57%	94.74%	75%	

TABLE VI. CONFUSION MATRIX OF NAÏVE BAYES WITHOUT TF - IDF

Actual	**Predicted**			**Sensitivity**
	Jaringan	***Multimedia***	***RPL***	
Jaringan	33	0	4	89.19%
Multimedia	0	26	1	96.29%
RPL	5	2	43	86%
Precision	86.84%	92.86%	89.58%	

TABLE VII. PERFORMANCE RESULT OF NAÏVE BAYES METHOD

Performance	With TF - IDF	Without TF - IDF
Accuracy	81.74%	**88.69%**
Precision	86.1%	**89.76%**
Sensitivity	80.15%	**90.49%**

Based on the results of the tests shown in Table VII, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces **88.69%** accuracy, **89.76%** precision, and **90.49%** sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract.

IV. CONCLUSION

Based on the results of the tests that have been done, the naïve Bayes method with TF-IDF has an accuracy of 81.74%, a precision of 86.1%, and a sensitivity of 80.15%. While the naïve Bayes method without TF-IDF weighting produces **88.69%** accuracy, **89.76%** precision, and **90.49%** sensitivity. Thus, the naïve Bayes method without TF-IDF weighting has better performance than TF-IDF weighting for the classification of thesis topics based on the abstract. The suggestions for further research can use feature selection methods such as chi-square to improve the performance of the naïve Bayes method.

References

[1] D. Adhe, C. Rachman, R. Goejantoro, and D. Tisna, "Implementation Of Text Mining For Grouping Thesis Documents Using K-Means Clustering," *J. EKSPONENSIAL*, vol. 11, no. 2, pp. 167–174, 2020.

[2] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, Feb. 2021, doi: 10.11591/ijece.v11i1.pp664-670.

[3] M. Sholehhudin, M. Fauzi Ali, and S. Adinugroho, "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya)," vol. 2, no. 11, pp. 5518–5524, 2018.

[4] L. Zahrotun, N. H. Putri, and A. Nur Khusna, "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesis Titles," in *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Oct. 2018, pp. 1–4, doi: 10.1109/TSSA.2018.8708817.

[5] D. M. U. Atmaja and R. Mandala, "Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 2, pp. 1–6, Aug. 2020, doi: 10.33021/itfs.v4i2.1182.

[6] M. Eminağaoğlu and Y. Gökşen, "A New Similarity Measure for Document Classification and Text Mining," *KnE Soc. Sci.*, vol. 2019, pp. 353–366, Jan. 2020, doi: 10.18502/kss.v4i1.5999.

[7] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," *Stud. Comput. Intell.*, vol. 740, no. November, pp. 373–397, 2018, doi: 10.1007/978-3-319-67056-0_18.

[8] R. Rismanto, A. Rachmad Syulistyo, and B. P. Citra Agusta, "Research Supervisor Recommendation System Based on Topic Conformity," *Int. J. Mod. Educ. Comput. Sci.*, vol. 12, no. 1, pp. 26–34, Feb. 2020, doi: 10.5815/ijmecs.2020.01.04.

[9] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/view/10955/6659.

[10] A. Solichin, "Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation," in *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Sep. 2019, pp. 217–222, doi: 10.23919/EECSI48112.2019.8977081.

[11] S. Sulova, L. Todoranova, B. Penchev, and R. Nacheva, "Using Text Mining to Classify Research Papers," in *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*, Jun. 2017, vol. 17, no. 21, pp. 647–654, doi: 10.5593/sgem2017/21/S07.083.

[12] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.

[13] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, 2017, pp. 1–5, doi: 10.1109/ICCCI.2017.8117734.

[14] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano, "Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers," *Neurocomputing*, vol. 92, pp. 98–115, Sep. 2012, doi: 10.1016/j.neucom.2012.01.030.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 513–520.