Naïve Bayes, Neural Network dan K-Nearest Neighbor untuk Klasifikasi Topik Tugas Akhir

Ari Putra Wibowo*[1], Widiyono[2], Anas Saifudin[3], Arief Soma Darmawan[4]
Eko Budihartono[5]
[1234] STMIK Widya Pratama Pekalongan
[5]Politeknik Harapan Bersama
E-mail: ***[1]ariputra.stmikwp@gmail.com**, [2]widdyono@gmail.com, [3]anzt@gmail.com,
[4]ariefsoma24@gmail.com, [5]tara.niscita@gmail.com

Abstrak

Pemilihan topik atau judul skripsi menentukan mahasiswa dalam menyelesaikan pengerjaan skripsi tepat waktu, hal ini juga berpengaruh dalam kebutuhan akreditasi program studi. Namun penentuan topik atau judul skripsi menjadi hal yang cukup sulit untuk mahasiswa, beberapa penelitian mengenai klasifikasi topik skripsi telah banyak dilakukan untuk mengelolompokkan topik atau judul skripsi sesuai dengan konsentrasi keahliannya sehingga memberikan informasi yang dapat membantu mahasiswa. Pada penelitian ini dilakukan perbandingan model klasifikasi untuk mengetahui model klasifikasi terbaik dalam klasifikasi topik atau judul skripsi. Ada tiga model klasifikasi yang dibangun dalam penelitian ini dengan menggunakan algoritma Naïve Bayes, Neural Network dan K-Nearest Neighbor. Evaluasi hasil dilakukan dengan metode cofusion matrix untuk mengetahui nilai akurasi, presisi, recall dan f-score. Dari hasil eksperimen menunjukkan bahwa model klasifikasi dengan algoritma Neural Network memiliki nilai akurasi paling tinggi dengan nilai 94,1% sedangkan nilai akurasi paling rendah adalah model klasifikasi Naïve Bayes dengan nilai 79% .

Kata Kunci: *tugas akhir; naïve bayes; neural network; k-nearest neighbor*

1. PENDAHULUAN

Salah satu syarat bagi mahasiswa untuk bisa menyelesaikan pendidikan program Sarjana adalah dengan membuat karya tulis ilmiah tugas akhir atau skripsi [1]. Skripsi merupakan bukti mahasiswa memiliki kemampuan akademik dalam melakukan penelitian sesuai bidang keahlian yang ditempuh [2]. Pemilihan topik atau judul skripsi menentukan mahasiswa dalam menyelesaikan pengerjaan skripsi tepat waktu, hal ini juga berpengaruh dalam kebutuhan akreditasi program studi [3]. Namun penentuan topik atau judul skripsi menjadi hal yang cukup sulit untuk mahasiswa, beberapa penelitian mengenai klasifikasi topik skripsi telah banyak dilakukan untuk mengelolompokkan topik atau judul skripsi sesuai dengan konsentrasi keahliannya sehingga memberikan informasi yang dapat membantu mahasiswa.

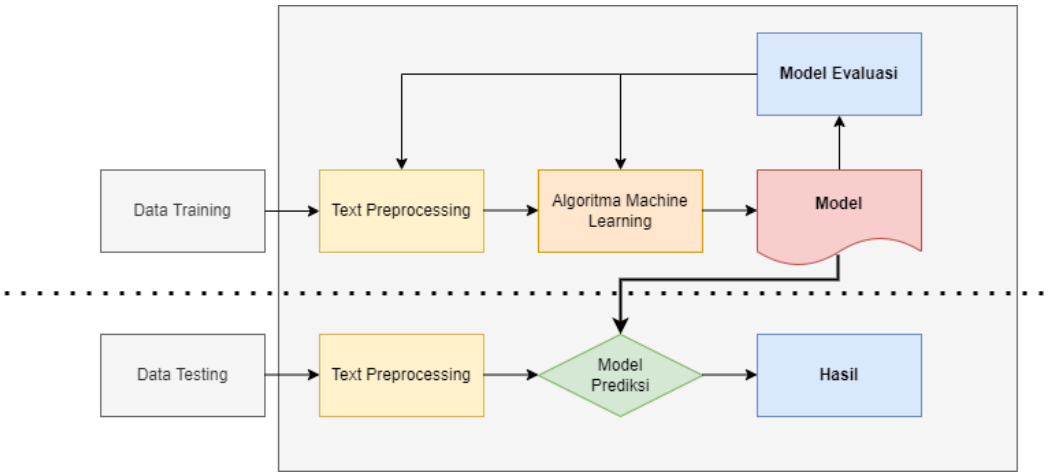
Adapun penelitan yang dilakukan oleh M. Priyantono dkk menggunakan algoritma Naïve Bayes untuk optimasi sistem pelabelan topik skripsi dengan pendekatan design thinking. Pada penelitian ini diperoleh hasil akurasi sebesar 87,5%, presisi sebesar 93% dan recall sebesar 83% dengan jumlah data yang digunakan 20 topik atau judul skripsi dalam bentuk dokumen teks [4]. Selanjutnya penelitian yang dilakukan oleh Arif Rahman dkk mengungkapkan tema Tugas Akhir dari mahasiswa yang kurang sesuai dengan kemampuan kompetensi, sehingga dilakukan penelitian dengan menggunakan algoritma Neural Network. Hasil penelitian ini diperoleh nilai akurasi sebesar 91,24% terjadi peningkatan akurasi setelah dilakukan optimasi dengan menggunakan algoritma Particle Swarm Optimization (PSO) menjadi 92.70% [5]. Klasifikasi Tugas Akhir juga dilakukan oleh Kitami Akromunnisa dkk, pada penelitiannya digunakan algoritma K-Nearest Neighbor dengan menggunakan data judul skripsi sebanyak 504 judul dari

program studi Teknik Informatika. Dari penelitian yang dilakukan algoritma K-Nearest Neighbor bisa digunakan untuk melakukan klasifikasi dengan hasil yang baik [6].

Berdasarkan uraian diatas pada penelitian ini akan dilakukan perbandingan terhadap kinerja pada tiga algoritma Naïve Bayes, Neural Network dan K-Nearest Neighbor. Perbandingan akan dilakukan dengan pengujian kinerja klasifikasi dengan menggunakan data judul Tugas Akhir atau Skripsi.

2. METODE PENELITIAN

Penelitian ini menggunakan metode eksperimental dengan penjelasan tahapan seperti pada gambar 1 berikut ini :



Gambar 1. Alur Penelitian

2.1. Dataset Tugas Akhir

Dataset yang digunakan pada penelitian ini berupa dokumen teks judul Tugas Akhir atau Skripsi mahasiswa STMIK Widya Pratama Pekalongan dari tahun 2018-2020 yang terdiri dari Skripsi program studi Sistem Informasi sebanyak 180 judul dan Skripsi program studi Teknik Informatika sebanyak 180 judul total ada 360 judul dokumen teks Skripsi [7]. Dalam melakukan eksperimen dataset yang digunakan akan dibagi menjadi dua sebagai data training dan data testing.

Tabel 1. Dataset Skripsi

Dataset	Sistem Informasi	Teknik Informatika
Tahun 2018	60	60
Tahun 2019	60	60
Tahun 2020	60	60

2.2. Text Preprocessing

Text preprocessing merupakan langkah awal dalam membangun sebuah model *machine learning* dalam text mining [8]. Ada berbagai tahapan dalam *text preprocessing*, namun pada penelitian ini akan dilakukan empat tahap didalam *text preprocessing* yang paling umum digunakan yaitu *lowercase conversion*, *tokenization*, *stemming* dan *stop-word removal* [9].

Langkah pertama dalam *text preprocessing* adalah *lower case conversion* yaitu mengubah karakter huruf menjadi huruf kecil semua, walaupun tidak terdapat perbedaan dalam penggunaan huruf besar atau huruf kecil namun pada penelitian klasifikasi teks biasanya digunakan huruf kecil. *Tokenization* adalah proses pemecahan kalimat menjadi bentuk frasa atau kata, sering juga disebut dengan segmentasi teks. Segmentasi dilakukan pada karakter alfabet atau alfanumerik yang dipisahkan oleh karakter non-alfanumerik (seperti: tanda baca, spasi). *Stemming* merupakan proses untuk mencari bentuk kata dasar pada setiap kata yang terdapat pada dokumen teks. *Stop-word removal* menghilangkan atau menghapus kata yang kurang relevan dalam melakukan proses klasifikasi. Penggunaan kata yang biasa ditemui di dalam dokumen teks tanpa adanya ketergantungan pada topik tertentu (misalnya, konjungsi, preposisi, artikel, dll) [10].

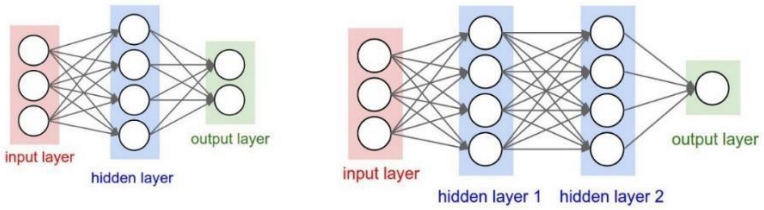
2.3. Algoritma Klasifikasi

Proses klasifikasi pada dataset dokumen teks judul skripsi menggunakan algoritma Naïve Bayes, Neural Network dan K-Nearest Neighbor. Naïve Bayes merupakan metode klasifikasi berdasarkan teorema bayes dengan asumsi karakteristik independensi kondisional [11]. Algoritma Naïve Bayes salah satu algoritma *supervised learning* yang biasanya digunakan untuk mengatasi klasifikasi seperti analisis emosional, *feature classification* dan masalah multi-klasifikasi lainnya. Salah satu keunggulan dari algoritma Naïve Bayes adalah lebih efisien dalam melakukan perhitungan [12]. Persamaan dari algoritma Naiva Bayes dapat di tulis pada rumusan (1) berikut ini :

$$\diamond\diamond(C|\diamond\diamond) = \diamond\diamond(\diamond\diamond)\frac{P(A|C)}{\diamond\diamond(\diamond\diamond)} \qquad (1)$$

dimana A = {a1, a2, ..., ak} memiliki *eigenvectors* berdimensi k, C adalah label dari A. P(C), P(C|A), dan P(A|C)/P(A) prior probability, *probabilitas posterior*, dan *likehood respectively*.

Algoritma Neural Network bekerja layaknya jaringan neuron yang terdapat pada otak manusia, algoritma ini merupakan salah satu algoritma dengan kinerja klasifikasi yang sangat baik. Namun algoritma Neural Network ini membutuhkan waktu yang lebih lama dalam melakukan perhitungan, karena banyaknya lapisan dari node-node yang dibentuk [13].



Gambar 2. *Neural Network Architectures*

Algoritma K-Nearest Neighbor adalah salah satu algoritma *machine learning* yang paling sederhana. Algoritma K-Nearest Neighbor bekerja berdasarkan aturan yang memilih jarak minimum dari data uji ke sampel pelatihan untuk menentukan K tetangga terdekat. Setelah mendefinisikan K tetangga terdekat, selanjutnya akan digunakan untuk memprediksi nilai instance baru [14]. Untuk mencari kedekatan antar titik pada nilai k, dihitung dengan melakukan perhitungan jarak euclidean dengan perasamaan (2) berikut ini :

$$\diamond = \sqrt{(\diamond_2 - \diamond_1)^2 + (\diamond_2 - \diamond_1)^2} \qquad (2)$$

2.4. Evaluasi

Tahap evaluasi dilakukan untuk mengetahui performa dari model klasifikasi, adapun evaluasi yang digunakan pada panelitian ini adalah menggunakan metode *k-fold cross validation* dimana akan dilakukan pemisahan dataset menjadi data training dan data testing.

Tabel 3. *K-fold Cross Validation*

Fold 1	*Testing*	*Training*	
Fold 2	*Training*	*Testing*	*Training*
...			
Fold 10	*Training*		*Testing*

Pada penelitian ini juga akan dilakukan evaluasi pada model klasifikasi dengan menggunakan metode *cofusion matrix* yang ditampilkan pada sebuah tabel. Tabel ini akan menampilkan dan membandingkan nilai aktual dengan nilai prediksi model klasifikasi yang bisa digunakan untuk menghasilkan matrix evaluasi seperti akurasi, presisi dan *recall*.

Tabel 4. *Cofusion Matrix* Topik Skripsi

Eksperimen	Kelas Prediksi	
Kelas Aktual	Positif	Negatif
Positif	TP	TN
Negatif	FP	FN

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (3)$$

$$\text{Presisi} = \frac{TP}{TP + FP} \times 100\% \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

2.5. Peralatan

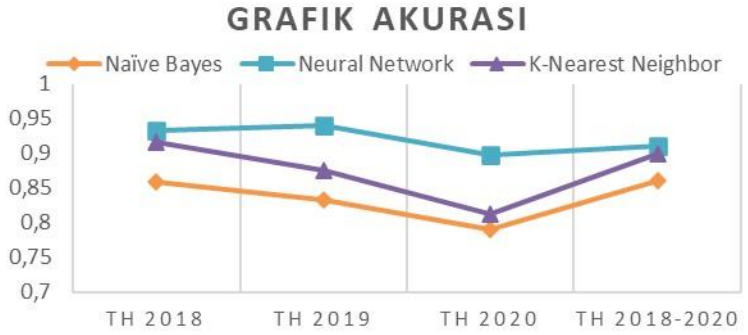
Penelitian ini menggunakan perangkat dengan spesifikasi sebagai berikut : laptop dengan prosesor intel core i3 generasi 4 @1.90GHz, SSD 240 GB dan RAM 8 GB. Menggunakan Sistem Operasi Windows 10 Pro dengan infrastruktur 64 bit. Sedangkan untuk Pengolahan data menggunakan tool RapidMiner Studio versi 9.10.

3. HASIL DAN PEMBAHASAN

Dari eksperimen model algoritma klasifikasi yang dilakukan diperoleh hasil penelitian sebagai berikut ini:

3.1. Akurasi Model Klafikasi

Hasil perhitungan akurasi pada model klasifikasi dapat dilihat pada gambar 3 berikut ini:



Gambar 3 Grafik akurasi model klasifikasi

Berdasarkan gambar 3, hasil eksperimen menunjukkan bahwa algoritma Neural Network memiliki nilai akurasi tertinggi untuk dataset tahun 2019 dengan nilai akurasi sebesar 94,1%. Sedangkan nilai akurasi terendah ditunjukan oleh algoritma Naïve Bayes untuk dataset tahun 2020 dengan nilai akurasi 79%. Untuk keseluruhan penggunaan dataset kinerja model yang paling bagus ditunjukan oleh algoritma Neural Network dan kinerja paling rendah ditunjukkan oleh algoritma Naïve Bayes.

3.2. Precision, Recall dan F-Score Model Klasifikasi

Evaluasi model berikutnya dilakukan dengan melakukan perhitungan *precision, recall* dan *f-score* seperti pada tabel 5 dibawah ini:

Tabel 5 perhitungan presisi, recall dan f-score

Dataset	Naïve Bayes			Neural Network			K-Nearest Neighbor		
	recall	*precision*	*f-score*	*recall*	*precision*	*f-score*	*recall*	*precision*	*f-score*
th 2018	0,95	0,85	0,86	0,95	0,91	0,93	0,91	0,91	0,91
th 2019	0,85	0,82	0,83	0,93	0,94	**0,94**	0,86	0,88	0,87
th 2020	0,79	**0,78**	0,79	0,86	0,92	0,89	0,83	0,80	0,81
th 2018-2020	0,88	0,84	0,86	0,88	0,94	0,91	0,90	0,89	0,90

Dari hasil perhitungan yang ditunjukkan pada tabel 5 dapat diketahui bahwa perolehan *f-score* dengan algoritma Neural Network memiliki nilai yang paling baik dibandingan dengan algoritma Naïve Bayes dan K-Nearest Neighbor. Diketahui nilai *f-score* tertinggi adalah 94% diperoleh algoritma Neural Network, sedangkan nilai terkecil adalah 78% dihasilkan oleh algoritma Naïve Bayes.

3.3. Perhitungan Recall dan Precision tiap Kelas

Pada tabel 6 berikut ini menunjukkan hasil perhitungan *recall* dan *precision* untuk masing-masing kelas dokumen label.

Tabel 6 Hasil perhitungan *recall* dan *precision* Naïve Bayes

Kelas Label	Dataset							
	2018		2019		2020		2018-2022	
	Recall	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*
Teknik Informatika	0,87	0,85	0,85	0,82	0,80	0,78	0,88	0,84

Sistem Informasi	0,85	0,87	0,82	0,84	0,78	0,80	0,84	0,88

Tabel 7 Hasil perhitungan *recall* dan *precision* Neural Network

Kelas Label	Dataset							
	2018		2019		2020		2018-2022	
	Recall	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*
Teknik Informatika	0,95	0,92	0,94	0,95	0,86	0,93	0,88	0,91
Sistem Informasi	0,92	0,95	0,95	0,93	0,93	0,88	0,91	0,92

Tabel 8 Hasil perhitungan *recall* dan *precision* K-Nearest Neighbor

Kelas Label	Dataset							
	2018		2019		2020		2018-2022	
	Recall	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*
Teknik Informatika	0,92	0,92	0,97	0,88	0,83	0,80	0,91	0,90
Sistem Informasi	0,92	0,92	0,88	0,87	0,80	0,83	0,90	0,91

4. KESIMPULAN

Pada penelitian ini telah dibangun model untuk klasifikasi Tugas Akhir, ada tiga model yang dibangun dengan menggunakan algoritma yang berbeda yaitu Naive Bayes, Neural Network dan K-Nearest Neighbor. Berdasarkan eksperimen yang dilakukan, model Naive Bayes memiliki akurasi yang peling rendah dengan perbedaan yang cukup tinggi dengan algoritma Neural Network. Hal ini kemungkinan terjadi karena belum diterapkannya seleksi fitur pada tahap preprosesing. Dari ketiga model ini memiliki pola yang hampir sama untuk proses perhitungan precision, recall dan f-score. Dari hasil perhitungan model klasifikasi dengan algoritma Neural Network memiliki nilai tertinggi dibandingkan dengan model yang lainnya.

DAFTAR PUSTAKA

[1] D. Patmawati, "Pedoman Penulisan Skripsi," no. 59, pp. 96–144, 2016.
[2] Unsika, "Panduan Penulisan Skripsi," no. December, 2015.
[3] M. R. Muttaqin and M. Defriani, "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 121–129, 2020.
[4] M. B. Priyantono, M. Ahnan, M. A. Widhianto, and D. D. Prasetyo, "Optimasi Sistem Pelabelan Topik Skripsi menggunakan Algoritma Naive Bayes dengan Pendekatan Design Thinking," vol. 8, no. 1, pp. 168–174, 2022.
[5] A. Rahman and A. Haqiqi, "Rekomendasi Tema Tugas Akhir Menggunakan Metode Clasifikasi Supervised Learning," *Smart Comp*, vol. 11, pp. 535–540, 2022.
[6] K. Akromunnisa and R. Hidayat, "Klasifikasi Dokumen Tugas Akhir (Skripsi) Menggunakan K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 4, no. 1,

p. 69, 2019.
[7] W. Darmawan, a p Wibowo, and B. Ismanto, "Klasifikasi Teks Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Forward Selection," *Ic-Tech*, vol. XVI, no. 1, pp. 1–7, 2021.
[8] S. I. G. Situmeang, "Impact of Text Preprocessing on Named Entity Recognition Based on Conditional Random Field in Indonesian Text Samuel Indra Gunawan Situmeang," vol. 6, no. 36, pp. 423–430, 2022.
[9] S. Bal and E. S. Gunal, "The Impact of Features and Preprocessing on Automatic Text Summarization," vol. 25, no. 2, pp. 117–132, 2022.
[10] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.
[11] D. B. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, *Bayesian Data Analysis (3rd ed.)*. 2013.
[12] G. Liang, Y. G. Yan, M. Wang, X. L. Lian, M. S. Li, and W. H. Tang, "Classification for Text Data from the Power System Based on Improving Naive Bayes," *Asia-Pacific Power Energy Eng. Conf. APPEEC*, vol. 2020-September, 2020.
[13] M. Ghiassi, M. Olschimke, B. Moon, and P. Arnaudo, "Automated text classification using a dynamic artificial neural network model," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10967–10976, 2012.
[14] T. Yu and K. T. Nwet, "Sentiment analysis system for myanmar news using k nearest neighbor and naïve bayes," *WCSE 2020 2020 10th Int. Work. Comput. Sci. Eng.*, no. Wcse, pp. 512–516, 2020.