EXPERT JURNAL MANAJEMEN SISTEM INFORMASI DAN TEKNOLOGI

Analisis Performansi Naïve Bayes pada Klasifikasi Plagiarisme Dokumen Berdasarkan Pembobotan Teks Menggunakan Algoritma TF-IDF

Erlangga [1], **Ari Kurniawan Saputra** [2*], **A. Herbantolo Nurendro Kushariantoko** [3]
[1] Sistem Informasi, Fakultas Ilmu Komputer, Universitas Bandar Lampung, Bandar Lampung
[2] Informatika, Fakultas Ilmu Komputer, Universitas Bandar Lampung, Bandar Lampung
[3] Informatika, Fakultas Ilmu Komputer, Universitas, Bandar Lampung
Lampung, Indonesia
[1] erlangga@ubl.ac.id, [2*] ari.kurniawan@ubl.ac.id, [3] a.herbantolo.22421057@student.ubl.ac.id

ABSTRACT – A thesis proposal is a research plan submitted by a student with guidance from a Supervisor, and is prepared following the rules of scientific writing. TF-IDF algorithm is a numerical statistical method that shows how important a word is in a document or corpus. Naive Bayes Classifiers method is a text classification technique that uses keyword probabilities to compare training documents with test documents. This research aims to analyze the performance of Naïve Bayes Classifier on plagiarism classification of thesis proposal documents based on text weighting using TF-IDF algorithm. Performance analysis is done using Confusion Matrix method. Based on testing and analysis conducted using the Naive Bayes Classifier performance test for the classification of Low Plagiarism, Moderate Plagiarism, and Severe Plagiarism classes with a total dataset of 85 thesis proposal documents divided into Training Data and Testing Data. The amount of Training Data amounted to 59 corpus and Testing Data amounted to 26 corpus. Based on the performance test conducted using the Confusion Matrix method, the results are shown in Table 6 with Split Data 70: 30, with an Accuracy value of 97.65%, a Precision value of 95.23%, and a Precall value of 98.74%. This shows that the Naive Bayes Classifier is at the excellent classification level. For future research, higher critical analysis in the dataset, the more accurate the prediction on the testing data. With a precision of 95.23%, recall of 98.74%, and accuracy of 97.65%, it can be concluded that the Naïve Bayes Classifier algorithm shows an excellent classification level.

Keywords: Performance; Naïve Bayes; TF-IDF; Term; Classification

ABSTRAK – Proposal skripsi adalah rencana penelitian yang diajukan oleh mahasiswa dengan bimbingan dari Dosen Pembimbing, dan disusun mengikuti aturan penulisan karya ilmiah. Algoritma TF-IDF adalah metode statistik numerik yang menunjukkan seberapa penting suatu kata dalam sebuah dokumen atau korpus. Metode *Naive Bayes Classifiers* adalah teknik klasifikasi teks yang menggunakan probabilitas kata kunci untuk membandingkan dokumen pelatihan dengan dokumen uji. Penelitian ini bertujuan untuk melakukan analisis performansi *Naïve Bayes Classifier* pada klasifikasi plagiarisme dokumen proposal skripsi berdasarkan pembobotan teks menggunakan algoritma TF-IDF. Analisis performansi dilakukan menggunakan metode *Confusion Matrix*. Berdasarkan pengujian dan analisis yang dilakukan menggunakan uji performansi *Naive Bayes Classifier* untuk klasifikasi kelas Plagiarisme Rendah, Plagiarisme Sedang, dan Plagiarisme Berat dengan jumlah dataset 85 dokumen proposal skripsi yang terbagi dalam *Data Training* dan *Data Testing*. Jumlah *Data Training* berjumlah 59 korpus dan *Data Testing* berjumlah 26 korpus. Berdasarkan uji performansi yang dilakukan mengguanakn metode *Confusion Matrix* didapatkan hasil yang ditunjukan pada Tabel 6 dengan *Split Data* 70 : 30, dengan nilai *Accuracy* 97,65%, nilai *Precision* 95,23%, dan nilai *Precall* 98,74%. Hal ini menunjukkan bahwa *Naive Bayes Classifier* berada pada tingkat *excellent classification*. Untuk penelitian berikutnya, analisis kritis lebih tinggi dalam *dataset*, maka prediksi pada data testing semakin akurat. Dengan *precision* sebesar 95,23%, *recall* sebesar 98,74%, dan *accuracy* sebesar 97,65%, dapat disimpulkan bahwa algoritma *Naïve Bayes Classifier* menunjukkan tingkat *excellent classification*.

Kata Kunci: Performansi; *Naïve Bayes*; TF-IDF; *Term*; Klasifikasi

1. PENDAHULUAN

Proposal skripsi adalah rencana penelitian yang diajukan oleh mahasiswa dengan bimbingan dari Dosen Pembimbing, dan disusun mengikuti aturan penulisan karya ilmiah [1]. Masalah plagiarisme dalam proposal skripsi adalah isu yang serius dan dapat merusak integritas akademik. Plagiarisme dalam penulisan ilmiah dianggap

melanggar hukum karena melibatkan pencurian karya orang lain [2]. Salah satu cara untuk mencegah plagiarisme adalah dengan mendeteksi kemiripan antar dokumen sebelum dokumen tersebut disetujui oleh pihak lain [3].

Identifikasi kemiripan antar dokumen adalah salah satu metode yang dapat digunakan untuk mendeteksi plagiarisme dalam sebuah dokumen [4]. Kemiripan antar dokumen *(document similarity)* dapat dimanfaatkan sebagai metode untuk mencari informasi yang serupa, sehingga mempercepat proses pencarian. Fitur ini umumnya diterapkan dalam artikel berita dan jurnal akademik [5]. Dalam penelitian ini algoritma yang digunakan untuk mengukur bobot kemiripan dokumen adalah algoritma TF-IDF.

Algoritma TF-IDF *(Term Frequency-Inverse Document Frequency)* adalah metode statistik numerik yang menunjukkan seberapa penting suatu kata dalam sebuah dokumen atau korpus. Metode ini sering digunakan sebagai faktor bobot dalam pencarian informasi dan penambangan teks. Nilai TF-IDF meningkat seiring dengan frekuensi kemunculan kata dalam dokumen, namun juga memperhitungkan seberapa sering kata tersebut muncul dalam korpus secara keseluruhan. Variasi dari skema pembobotan TF-IDF sering dimanfaatkan oleh mesin pencari untuk menilai dan merangking relevansi dokumen berdasarkan permintaan pengguna [6]. Dalam penelitian ini, hasil dari pembobotan teks akan diklasifikasikan untuk menentukan persentase kemiripan atau plagiarisme. Klasifikasi ini mencakup: (1) plagiarisme ringan, di mana proporsi kata, kalimat, atau paragraf yang dijiplak tidak melebihi 30 persen (<30%), (2) plagiarisme sedang, dengan proporsi kata, kalimat, atau paragraf yang dijiplak berkisar antara 30-70 persen, dan (3) plagiarisme berat, di mana proporsi kata, kalimat, atau paragraf yang dijiplak melebihi 70 persen (>70%) [7]. Algoritma klasifikasi data yang digunakan adalah *Naïve Bayes Classifier*.

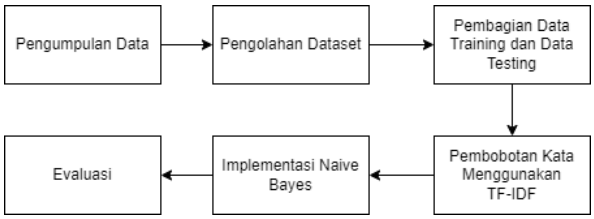
Metode *Naive Bayes Classifiers* adalah teknik klasifikasi teks yang menggunakan probabilitas kata kunci untuk membandingkan dokumen pelatihan dengan dokumen uji. Proses ini melibatkan beberapa tahap perbandingan, dan hasil akhir yang memiliki probabilitas tertinggi akan ditetapkan sebagai kategori untuk dokumen baru [8].

Berdasarkan penelitian yang dilakukan oleh Apriani, *et al* tahun 2021 menggunakan pembobotan TF-IDF dan metode cosine similarity pada 7 sampel data dapat menghasilkan tingkat akurasi sebesar 64,28% [9]. Penelitian lain yang dilakukan oleh Inda Sari Tomagola, *et al* tahun 2023, pengujian dilakukan menggunakan metode klasifikasi *Naïve Bayes* dan dianalisis menggunakan confusion matrix. Dari beberapa eksperimen, pembagian data 80:20 menghasilkan tingkat akurasi terbaik dibandingkan dua eksperimen lainnya, dengan akurasi TF-IDF *Naïve Bayes* sebesar 73% dan TF-RF *Naïve Bayes* sebesar 72%. Berdasarkan hasil tersebut, dapat disimpulkan bahwa TF-IDF *Naïve Bayes* memberikan akurasi yang lebih baik dibandingkan TF-RF *Naïve Bayes* [10]. Penelitian lain juga dilakukan oleh Rio Al Rasyid, *et al* tahun 2023, pengujian dilakukan pada 5 kueri yang berbeda, dan hasil pengujian presisi diperoleh dengan nilai rata-rata sebesar 83% [11].

Berdasarkan uraian penelitian-penelitian sebelumnya, penelitian ini bertujuan untuk melakukan analisis performansi *Naïve Bayes Classifier* pada klasifikasi plagiarisme dokumen proposal skripsi berdasarkan pembobotan teks menggunakan algoritma TF-IDF. Analisis performansi dilakukan menggunakan metode *Confusion Matrix*.

2. METODOLOGI

Penelitian ini menggunakan metodologi penelitian kuantitatif eksperimental. Metode ini melibatkan percobaan yang dirancang untuk mengevaluasi pengaruh variabel independen (perlakuan atau *treatment*) terhadap variabel lainnya. Metode penelitian kuantitatif ini bertujuan untuk menguji dampak perlakuan tertentu dalam kondisi yang terkontrol [12]. Penelitian ini bertujuan untuk melakukan uji performansi berdasarkan klasifikasi *Naive Bayes Classifier* dari hasil pembobotan algoritma TF-IDF. Berikut merupakan tahapan-tahapan pengujian algortima TF-IDF dan klasifikasi *Naive Bayes Classifier* [13]:



Gambar 1. Tahapan Penelitian

1. Pengumpulan Data
   Pengumpulan data merupakan proses mengumpulkan informasi yang memungkinkan peneliti untuk membuat kesimpulan serta mengambil keputusan atau tindakan yang tepat [14]. Pada penelitian ini tahapan pengumpulan data penelitian dilakukan dengan mengumpulkan 85 dokumen proposal yang terdiri dari skripsi dan tesis, berasal dari 11 program studi jenjang S1 dan 4 program studi jenjang S2.

2. Pengolahan *Dataset*
   Dataset adalah objek yang merepresentasikan data dan relasinya di memori [15]. Pada penelitian ini pengolahan *dataset* dilakukan dengan memisahkan *cover* dan daftar pustaka untuk mempermudah tahap *pre-proccessing*.

3. Pembagian *Data Training* dan *Data Testing*
   Data Training merupakan sampel data yang digunakan dalam proses penyusunan pohon keputusan yang telah diuji tingkat akurasinya. *Data Testing* merupakan

DOI http://dx.doi.org/10.36448/expert.v14i2.3952
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 14 No. 2
December 31, 2024 – Hal.116

sekumpulan data yang digunakan sebagai acuan dalam proses klasifikasi data [16].

4. Pembobotan Kata Menggunakan TF-IDF
   Pembobotan TF-IDF merupakan metode yang memberikan bobot nilai pada kata-kata *(term)* dalam sebuah dokumen berdasarkan hubungan kata tersebut dengan keseluruhan isi dokumen [17]. Pada penelitian ini, penerapan TF-IDF digunakan untuk menghitung bobot kata yang muncul. Sebelum melakukan pembobotan teks, terlebih dahulu melakukan *pre-processing*, tahapan ini bertujuan untuk menghapus atau membersihkan teks yang tidak diperlukan dalam sebuah dokumen, langkah-langkah yang dilakukan meliputi *Tokenizing, Cleaning, Stemming,* dan *Filtering* [18][19].

5. Implementasi *Naive Bayes*
   Naive Bayes beroperasi berdasarkan konsep frekuensi istilah, yang menunjukkan seberapa sering suatu kata muncul dalam sebuah dokumen [20]. Pada penelitian ini, implementasi *Naive Bayes* dilakukan setelah hasil dari pembobotan kata TF-IDF, untuk kemudian dilakukan klasifikasi dokumen *Dataset* masuk dalam klasifikasi Plagiarisme Ringan, Sedang, atau Berat.

6. Evaluasi
   Evaluasi yang digunakan adalah evaluasi performansi. Untuk evaluasi performansi hasil klasifikasi *Naive Bayes* berdasarkan hasil pembobotan teks algoritma TF-IDF menggunakann *Cunfusion Matrix*. *Confusion Matrix* adalah sebuah metode evaluasi dalam bentuk tabel matriks yang berfungsi untuk menilai performa *machine learning* dan dapat digunakan sebagai alat visual dalam menganalisis hasil klasifikasi [21][22]. Dalam penelitian ini, tahapan evaluasi dilakukan berdasarkan hasil pembobotan teks menggunakan algoritma TF-IDF, untuk kemudian dilakukan labeling klasifikasi setiap pembobotan teks yaitu Plagiarisme Ringan, Sedang, dan Berat.

3. HASIL DAN PEMBAHASAN

Penerapan tahapan yang dilakukan dalam penelitian ini mengacu pada alur Gambar 1 pada metodologi penelitian yang terdiri dari penerapan tahapan:

1. Pengumpulan Data
   Dalam penerapan penelitian ini, tahap pengumpulan data dilakukan dengan mengumpulkan dokumen proposal skripsi dari 17 program studi yang terdiri dari program studi strata 1 (S1) yaitu Sistem Informasi, Informatika, Pendidikan Bahasa Inggris, Akuntansi, Manajemen, Arsitektur, Teknik Mesin, Teknik Sipil, Hukum, Administrasi Bisnis, Administrasi Publik, dan Ilmu Komunikasi. Untuk program studi strata 2 (S2) yaitu Magister Manajemen, Magister Hukum, Magister Teknik, dan Magister Ilmu Administrasi. Untuk program studi strata 3 (S3) yaitu Manajemen. Dalam tahapan pengumpulan data penelitian ini, dokumen proposal yang dijadikan sebagai data korpus berjumlah 85 dokumen proposal untuk kemudian data ini akan di olah pada tahapan pengolahan *dataset* menjadi *data training*.

2. Pengolahan *Dataset*
   Dalam penerapan penelitian ini, tahap pengolahan *dataset* dilakukan dengan memisahkan *cover* dan daftar pustaka pada setiap dokumen proposal. Hal ini bertujuan untuk mempermudah dalam tahap *pre-proccessing*.

3. Pembagian *Data Training* dan *Data Testing*
   Dalam penerapan penelitian ini, tahap pembagian *data training* dan *data testing*. Pembagian *data training* dan *data testing* bertujuan untuk melatih model agar dapat mengenali pola data serta menguji kinerjanya dalam memprediksi data baru, sehingga membantu menghindari *overfitting* dan memastikan kemampuan generalisasi model. Dalam penelitian ini, data sebanyak 85 akan dibagi menggunakan metode *cross-validation*. Pada metode *cross-validation*, data akan dibagi menjadi 5 bagian yang masing-masing terdiri dari 17 data. Setiap iterasi, model akan dilatih menggunakan 4 bagian (68 data) dan di uji menggunakan 1 bagian (17 data). Proses ini akan diulang sebanyak 5 kali, sehingga setiap bagian bergantian menjadi *data testing*. Hasil evaluasi akhir merupakan rata-rata performa dari kelima iterasi tersebut.

4. Pembobotan Kata Menggunakan TF-IDF
   Dalam penulisan artikel ilmiah ini hanya menggunakan beberapa teks atau kata-kata sampel yang berasal dari 2 data korpus yang sama untuk dijadikan sebagai dokumen uji, terdapat beberapa tahapan yang dilakukan sebelum melakukan pembobotan kata yaitu *pre-processing* yang terdiri dari:

   a. Tokenizing
   Pada tahap *Tokenizing*, setiap kalimat dalam dokumen proposal skripsi akan diuraikan menjadi kata per kata (*token*), seperti yang ditunjukkan pada Tabel 1 berikut ini.

Tabel 1. Tahap *Tokenizing*

Input	Output

DOI http://dx.doi.org/10.36448/expert.v14i2.3952
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 14 No. 2
December 31, 2024 – Hal.117

Pada era sekarang ini pemanfaatan *text mining* sangatlah diperlukan untuk mevisualkan atau mengevaluasi pengetahuan dari kumpulan besar dari teks dokumen.	Pada, era, sekarang, ini, pemanfaatan, *text*, *mining*, sangatlah, diperlukan, untuk, mevisualkan, atau, mengevaluasi, pengetahuan, dari, kumpulan, besar, dari, teks, dokumen.
Pada penelitian ini akan dianalisis dan dibandingkan algoritma TF-IDF, TF.RF, dan WIDF.	Pada, penelitian, ini, akan, dianalisis, dan, dibandingkan, algoritma, TF-IDF, TF.RF, dan, WIDF.
Untuk metode pengujiannya akan digunakan metode klasifikasi *naïve* bayes dan analisis perbandingannya menggunakan *confusion matrix*.	Untuk, metode, pengujiannya, akan, digunakan, metode, klasifikasi, *naïve*, *bayes*, dan, analisis, perbandingannya, menggunakan, *confusion*, *matrix*.
Dengan *dataset* yang digunakan sebanyak 130 dokumen yang mana 100 data *training* dan 30 data uji.	Dengan, *dataset*, yang, digunakan, sebanyak, 130, dokumen, yang, mana, 100, data, training, dan, 30, data, uji.
Berdasarkan analisa pada hasil klasifikasi yang telah dilakukan, dapat disimpulkan bahwa pembobotan TF.RF dengan klasifikasi *Naïve bayes* lebih baik...	Berdasarkan, analisa, pada, hasil, klasifikasi, yang, telah, dilakukan, dapat, disimpulkan, bahwa, pembobotan, TF.RF, dengan, klasifikasi, *Naïve*, *bayes*, lebih, baik...

b. Cleaning

Pada tahap ini bertujuan untuk membersihkan data dari elemen-elemen yang tidak relevan dalam proses penelitian, seperti yang ditunjukan pada Tabel 2 berikut ini

Tabel 2. Tahap *Cleaning*

Input	Output
Pada, era, sekarang, ini, pemanfaatan, *text*, *mining*, sangatlah, diperlukan, untuk, mevisualkan, atau, mengevaluasi, pengetahuan, dari, kumpulan, besar, dari, teks, dokumen.	era, sekarang, pemanfaatan, *text*, *mining*, diperlukan, mevisualkan, mengevaluasi, pengetahuan, kumpulan, besar, teks, dokumen
Pada, penelitian, ini, akan, dianalisis, dan, dibandingkan, algoritma, TF-IDF, TF.RF, dan, WIDF.	penelitian, dianalisis, dibandingkan, algoritma, TF-IDF, TF.RF, WIDF
Untuk, metode, pengujiannya, akan, digunakan, metode, klasifikasi, *naïve*, *bayes*, dan, analisis, perbandingannya, menggunakan, *confusion*, *matrix*.	metode, pengujian, digunakan, metode, klasifikasi, *naïve*, *bayes*, analisis, perbandingan, menggunakan, *confusion*, *matrix*
Dengan, *dataset*, yang, digunakan, sebanyak, 130, dokumen, yang, mana, 100, *data*, *training*, dan, 30, data, uji.	*dataset*, digunakan, 130, dokumen, 100, *data*, *training*, 30, data, uji
Berdasarkan, analisa, pada, hasil, klasifikasi, yang, telah, dilakukan, dapat, disimpulkan, bahwa, pembobotan, TF.RF, dengan, klasifikasi, *Naïve*, *bayes*, lebih, baik...	analisa, hasil, klasifikasi, disimpulkan, pembobotan, TF.RF, klasifikasi, *Naïve*, *bayes*, baik...

c. Stemming

Pada tahap ini, data teks dalam *dataset* diubah menjadi kata dasar dengan menghilangkan awalan dan akhiran. Selanjutnya, dilakukan koreksi terhadap kata-kata yang salah eja, selanjutnya memperbarui kata-kata dalam bahasa Indonesia yang masih menggunakan ejaan lama, serta mengembangkan kata-kata yang disingkat. Tahap ini ditunjukan pada Tabel 3 berikut ini.

DOI http://dx.doi.org/10.36448/expert.v14i2.3952
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 14 No. 2
December 31, 2024 – Hal.118

Tabel 3. Tahap Stemming

Input	Output
era, sekarang, pemanfaatan, *text, mining*, diperlukan, mevisualkan, mengevaluasi, pengetahuan, kumpulan, besar, teks, dokumen	era, sekarang, manfaat, teks, mining, perlu, visual, evaluasi, tahu, kumpul, besar, teks, dokumen
penelitian, dianalisis, dibandingkan, algoritma, TF-IDF, TF.RF, WIDF	teliti, analisis, banding, algoritma, TF-IDF, TF.RF, WIDF
metode, pengujian, digunakan, metode, klasifikasi, *naïve*, *bayes*, analisis, perbandingan, menggunakan, *confusion*, *matrix*	metode, uji, guna, metode, klasifikasi, *naïve*, *bayes*, analisis, banding, guna, *confusion*, *matrix*
dataset, digunakan, 130, dokumen, 100, *data*, *training*, 30, data, uji	data, guna, 130, dokumen, 100, data, latih, 30, data, uji
analisa, hasil, klasifikasi, disimpulkan, pembobotan, TF.RF, klasifikasi, *Naïve*, *bayes*, baik...	analisa, hasil, klasifikasi, simpul, bobot, TF.RF, klasifikasi, *naïve*, *bayes*, baik...

d. Filtering

Tahap ini adalah bagian dari proses *pre-processing* data teks dalam *dataset* yang bertujuan untuk menghapus kata-kata yang tidak relevan atau tidak memiliki makna dalam penelitian. Tahap ini ditunjukan pada Tabel 4 berikut ini

Tabel 4. Tahap *Filtering*

Input	Output
era, sekarang, manfaat, teks, *mining*, perlu, visual, evaluasi, tahu, kumpul, besar, teks, dokumen	manfaat, teks, mining, visual, evaluasi, tahu, kumpul, teks, dokumen
teliti, analisis, banding, algoritma, TF-IDF, TF.RF, WIDF	analisis, banding, algoritma, TF-IDF, TF.RF, WIDF
metode, uji, guna, metode, klasifikasi, *naïve*, *bayes*, analisis, banding, guna, *confusion*, *matrix*	metode, uji, klasifikasi, *naïve*, *bayes*, analisis, banding, *confusion*, *matrix*
data, guna, 130, dokumen, 100, data, latih, 30, data, uji	data, dokumen, latih, uji
analisa, hasil, klasifikasi, simpul, bobot, TF.RF, klasifikasi, *naïve*, *bayes*, baik...	analisa, klasifikasi, simpul, bobot, TF.RF, klasifikasi, *naïve*, *bayes*

Pada tahap pembobotan teks (*term*), dilakukan proses untuk menetapkan nilai atau bobot pada setiap dokumen kata dalam proposal skripsi dengan menggunakan metode TF-IDF. Tujuan dari proses ini adalah untuk memberikan nilai pada term yang akan digunakan saat entri dalam proses klasifikasi. Hal ini ditunjukan pada Tabel 5 berikut ini

Tabel 5. Tahap Pembobotan Teks Algortima TF-IDF

Jenis Dokumen		*Term Dictionary*	*Term Frequency* (TF)	*Inverse Document Frequency* (IDF)	TF-IDF
Data Training	*Data Testing*				
Dokumen 1	Dokumen 2	era	1	1.5	1.5
Dokumen 1	Dokumen 2	pemanfaatan	1	2.0	2.0
Dokumen 1	Dokumen 2	*text mining*	1	2.2	2.2
Dokumen 1	Dokumen 2	evaluasi	1	1.8	1.8
Dokumen 1	Dokumen 2	dokumen	1	1.7	1.7
Dokumen 1	Dokumen 2	analisis	1	1.9	1.9
Dokumen 1	Dokumen 2	dibandingkan	1	2.1	2.1

DOI http://dx.doi.org/10.36448/expert.v14i2.3952
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 14 No. 2
December 31, 2024 – Hal.119

Dokumen 1	Dokumen 2	algoritma	1	1.6	1.6
Dokumen 1	Dokumen 2	TF-IDF	1	2.3	2.3
Dokumen 1	Dokumen 2	klasifikasi	1	1.8	1.8
Dokumen 1	Dokumen 2	*naïve bayes*	1	2.4	2.4
Dokumen 1	Dokumen 2	*confussion*	1	2.0	2.0
Dokumen 1	Dokumen 2	*matrix*	1	2.5	2.5
Dokumen 1	Dokumen 2	*dataset*	1	1.7	1.7
Dokumen 1	Dokumen 2	*training*	1	2.2	2.2
Dokumen 1	Dokumen 2	uji	1	1.5	1.5
Dokumen 1	Dokumen 2	*Accuracy*	1	2.0	2.0
Dokumen 1	Dokumen 2	*Precision*	1	2.3	2.3
Dokumen 1	Dokumen 2	*Recall*	1	2.1	2.1

5. Implementasi *Naive Bayes*
   Berdasarkan hasil pembobotan teks menggunaakn algoritma TF-IDF pada Tabel 5, selanjutnya dilakukan klasifikasi *Naive Bayes* yang dalam hal uji coba ini hanya menggunakan beberapa teks pada Dokumen 1 *Data Training* yaitu kata: "era", "pemanfaatan", "*text mining*". Berikut langkah-langkah dalam penerapan *Naive Bayes*:

   a. Menghitung *Likelihood*
      Likelihood merupakan metode estimasi parameter yang didasarkan pada pendekatan distribusi dengan cara mengoptimalkan fungsi *likelihood* [23]. Pada langkah ini akan menghitung probabilitas *likelihood* (kemunculan dokumen dalam suatu kelas) dengan mengalikan probabilitas masing-masing kata dalam dokumen untuk setiap kelas.
      Plagiarisme Ringan

$$P(Dokumen\ 1|Ringan) = 0.5 \times 0.4 \times 0.6 = 0.12$$

      Plagiarisme Sedang

$$P(Dokumen\ 1 \mid Sedang) = 0.3 \times 0.6 \times 0.4 = 0.072$$

      Plagiarisme Berat

$$P(\ Dokumen\ 1 \mid Berat\ ) = 0.8 \times 0.7 \times 0.9\ 0.504$$

   b. Mengalikan *Likelihood* dengan Probabilitas *Prior*
      Pada langkah ini menghitung probabilitas *posterior* (probabilitas kelas setelah melihat dokumen) dengan mengalikan hasil *likelihood* dengan probabilitas *prior* untuk setiap kelas.
      Plagiarisme Ringan

$$P(Ringan \mid Dokumen\ 1) = 0.12 \times 0.4 = 0.048$$

      Plagiarisme Sedang

$$P(Sedang \mid Dokumen\ 1) = 0.072 \times 0.3 = 0.0216$$

      Plagiarisme Berat

$$P(Berat \mid Dokumen\ 1) = 0.504 \times 0.3 = 0.1512$$

   c. Normalisasi Probabilitas
      Pada langkah ini, normalisasi probabilitas bertujuan Untuk mendapatkan probabilitas akhir dan membandingkan probabilitas kelas.

$$Total = 0.048 + 0.0216 + 0.1512 = 0.2208$$

      Probabilitas ter-normalisasi untuk setiap kelas:
      Plagiarisme Ringan

$$P(\ Ringan \mid Dokumen\ 1\ )0.22080.048\ 0.217$$

      Plagiarisme Sedang

$$P(\ Sedang \mid Dokumen\ 1\ ) = 0.22080.0216\ 0.098$$

      Plagiarisme Berat

$$P(\ Berat \mid Dokumen\ 1\ ) = 0.22080.1512\ 0.685$$

   d. Hasil Klasifikasi
      Berdasarkan hasil perhitungan normalisasi probabilitas, didapatkan hasil sebagai berikut: Plagiarisme Ringan: 21.7%, Plagiarisme Sedang: 9.8%, Plagiarisme Berat: 68.5%. Nilai probabilitas tertinggi ditunjukan pada kelas klasifikasi Plagiarisme Berat dengan nilai 68.5%. Maka Dokumen 1 (*Data Testing*) diklasifikasikan sebagai Plagiarisme Berat.

6. Evaluasi
   Pada langkah ini, uji performansi menggguanakn metode *Cunfusion Matrix* yang akan menghasilkan nilai

DOI http://dx.doi.org/10.36448/expert.v14i2.3952
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 14 No. 2
December 31, 2024 – Hal.120

Accuracy, *Precision*, dan *Recall*. Untuk pembagian *Data Traning* dan *Data Testing* digunakan perbandingan 70% : 30%. Hasil uji performansi *Naïve Bayes Classifier* ditunjukan pada Tabel 6 berikut ini

Tabel 6. Hasil Uji Performansi *Naive Bayes Classifier*

Data Training	*Data Testing*	*Split Data*	*Accuracy*	*Precision*	*Recall*
59	26	70:30	97,65%	95,23%	98,74%

Berdasarkan perhitungan *precision*, *recall*, dan *acuracy* pada tiga kategori kelas, yaitu Plagiarisme Ringan, Sedang, dan Berat, dapat disimpulkan bahwa akurasi model dipengaruhi oleh data pelatihan dan pengujian. Semakin banyak kata referensi yang ada dalam *data training*, yang menghasilkan frekuensi kata lebih tinggi dalam *dataset*, maka prediksi pada data testing semakin akurat. Dengan *precision* sebesar 95,23%, *recall* sebesar 98,74%, dan *accuracy* sebesar 97,65%, dapat disimpulkan bahwa algoritma *Naïve Bayes Classifier* menunjukkan tingkat *excellent classification*.

4. KESIMPULAN

Berdasarkan pengujian dan analisis yang dilakukan menggunakan uji performansi *Naive Bayes Classifier* untuk klasifikasi kelas Plagiarisme Rendah, Plagiarisme Sedang, dan Plagiarisme Berat dengan jumlah *dataset* 85 dokumen proposal skripsi yang terbagi dalam *Data Training* dan *Data Testing*. Jumlah *Data Training* berjumlah 59 korpus dan *Data Testing* berjumlah 26 korpus. Berdasarkan uji performansi yang dilakukan mengguanakn metode *Confusion Matrix* didapatkan hasil yang ditunjukan pada Tabel 6 dengan *Split Data* 70 : 30, dengan nilai *Accuracy* 97,65%, nilai *Precision* 95,23%, dan nilai *Precall* 98,74%. Hal ini menunjukkan bahwa *Naive Bayes Classifier* berada pada tingkat *excellent classification*. Untuk penelitian berikutnya, analisis kritis dengan pendekatan kualitatif juga dapat dikembangkan untuk memberikan pemahaman yang lebih menyeluruh.

DAFTAR PUSTAKA

[1] Fakultas Ekonomi, "Pedoman Penulisan Skripsi dan Proposal Skripsi 2021 Fakultas Ekonomi Universitas Negeri Jakarta," 2021.

[2] A. K. Saputra, R. Y. Endra, F. Ariani, T. Tanjung, and A. Prakarsya, "Implementasi Algoritma Rabin-Karp pada Pendeteksian Plagiarisme," *Expert J. Manaj. Sist. Inf. dan Teknol.*, vol. 13, no. 1, p. 23, Jun. 2023, doi: 10.36448/expert.v13i1.3161.

[3] Y. vita Via and R. Mumpuni, "Deteksi Kemiripan Dokumen Publikasi Skripsi Mahasiswa Menggunakan Algoritma Modifikasi Cosine Similarity," *J. Inf. Eng. Educ. Technol.*, vol. 3, no. 2, pp. 57–61, Dec. 2019, doi: 10.26740/jieet.v3n2.p57-61.

[4] S. Yuliyanti and Rizky, "Implementasi Algoritma Rabin Karp Untuk Mendeteksi Kemiripan Dokumen Stmik Bandung," *J. Bangkit Indones.*, vol. 10, no. 02, p. 1, Oct. 2020, doi: 10.52771/bangkitindonesia.v10i02.124.

[5] L. Mayola, M. Hafizh, and D. M. Putra, "Perancangan Aplikasi Similarity Deteksi Kemiripan Judul Disertasi Berbasis Web," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 6, no. 2, pp. 452–257, Apr. 2024, doi: 10.47233/jteksis.v6i2.1164.

[6] S. Chamira, "Implementasi Metode Text Mining Frequency-Invers Document Frequency (Tf-Idf) Untuk Monitoring Diskusi Online," *J. Informatics, Electr. Electron. Eng.*, vol. 1, no. 3, pp. 97–102, Mar. 2022, doi: 10.47065/jieee.v1i3.353.

[7] A. K. Saputra, K. Muludi, and T. Thamrin, "Comparative Analysis between Rabin Karp Algorithm , Winnowing , and Turnitin Applications for Detecting Plagiarized Words," *Proceeding 6th ICITB 2020* –, no. December, pp. 40–49, 2020, [Online]. Available: https://jurnal.darmajaya.ac.id/index.php/icitb/article/view/2505

[8] F. Handayani, D. Feddy, and S. Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *J. Tek. Elektro*, vol. 7, no. 1, pp. 19–24, 2015.

[9] A. Apriani, H. Zakiyudin, and K. Marzuki, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta," *J. Bumigora Inf. Technol.*, vol. 3, no. 1, pp. 19–27, Jul. 2021, doi: 10.30812/bite.v3i1.1110.

[10] I. S. Tomagola, A. Id Hidiana, and P. Nurul Sabrina, "Analisis Sentimen Terhadap Pangan Nasional Pada Media Sosial Twitter Menggunakan Algoritma Naïve Bayes," *Jati*

(Jurnal Mhs. Tek. Inform., vol. 7, no. 5, pp. 3350–3356, Jan. 2024, doi: 10.36040/jati.v7i5.7473.

[11] R. Al Rasyid and D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 1, pp. 170–178, Jan. 2024, doi: 10.35870/jtik.v8i1.1416.

[12] S. N. Zyra, T. P. Alamsyah, and R. Yuliana, "Penggunaan E-Learning Berbasis Edmodo Terhadap Hasil Belajar Kelas 4 Sekolah Dasar," *J. PGSD J. Ilm. Pendidik. Guru Sekol. Dasar*, vol. 15, no. 2, pp. 97–106, Nov. 2022, doi: 10.33369/pgsd.15.2.97-106.

[13] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *J. Manaj. Inform.*, vol. 13, no. 1, pp. 84–93, Apr. 2023, doi: 10.34010/jamika.v13i1.9424.

[14] R. Zulfirman, "Implementasi Metode Outdoor Learning dalam Peningkatan Hasil Belajar Siswa pada Mata Pelajaran Agama Islam di MAN 1 Medan," *J. Penelitian, Pendidik. dan Pengajaran JPPP*, vol. 3, no. 2, pp. 147–153, Aug. 2022, doi: 10.30596/jppp.v3i2.11758.

[15] Y. Yahya and M. Hamonangan Nasution, "Penggunaan Algoritma K-Means Untuk Menganalisis Pelanggan Potensial Pada Dealer SPS Motor Honda Lombok Timur Nusa Tenggara Barat," *Infotek J. Inform. dan Teknol.*, vol. 2, no. 2, pp. 32–41, Feb. 2019, doi: 10.29408/jit.v3i1.1814.

[16] Y. Wulandari, E. Haerani, S. K. Gusti, and S. Ramadhani, "Klasifikasi Berita Menggunakan Algoritma C4.5," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 2, pp. 279–289, Apr. 2022, doi: 10.32672/jnkti.v5i2.4194.

[17] H. Sari, G. L. Ginting, T. Zebua, and Mesran, "Penerapan Algoritma Text Mining dan TF-IDF untuk Pengelompokan Topik Skripsi pada Aplikasi Repository STMIK Budi Darma," *TIN Terap. Inform. Nusant.*, vol. 2, no. 7, pp. 414–432, 2021.

[18] R. Kurniawan R and I. Zufria, "Penerapan Text Mining Pada Sistem Penyeleksian Judul Skripsi Menggunakan Algoritma Latent Dirichlet Allocation(LDA)," *Indones. J. Comput. Sci.*, vol. 11, no. 3, pp. 1036–1052, Dec. 2022, doi: 10.33022/ijcs.v11i3.3120.

[19] B. R. Atmadja, "Analisis Sentimen Bahasa Indonesia Pada Tempat Wisata Di Kabupaten Sukabumi Dengan Naive Bayes Classifier," *Elkom J. Elektron. dan Komput.*, vol. 15, no. 2, pp. 371–382, Dec. 2022, doi: 10.51903/elkom.v15i2.872.

[20] N. Hidayah, "Implementasi Algoritma Multinomial Naïve Bayes, TF-IDF dan Confusion Matrix dalam Pengklasifikasian Saran Monitoring dan Evaluasi Mahasiswa Terhadap Dosen Teknik Informatika Universitas Dayanu Ikhsanuddin," *J. Akad. Pendidik. Mat.*, vol. 10, no. 1, pp. 8–15, 2024.

[21] R. Indransyah, Y. H. Chrisnanto, P. N. Sabrina, and S. Kom, "Klasifikasi Sentimen Pergelaran MotoGP di Indonesia Menggunakan Algoritma Correlated Naive Bayes Clasifier," *INFOTECH J.*, vol. 8, no. 2, pp. 60–66, 2022, doi: https://doi.org/10.31949/infotech.v8I2.3103.

[22] Karsito and S. Santi, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia," *J. Teknol. Pelita Bangsa*, vol. 9, pp. 43–48, 2019.

[23] S. Angnitha Purba, "Estimasi Parameter Data Berdistribusi Normal Menggunakan Maksimum Likelihood Berdasarkan Newton Raphson," *J. Sains Dasar*, vol. 9, no. 1, pp. 16–18, Feb. 2021, doi: 10.21831/jsd.v9i1.38564.