Klasifikasi Soal Berdasarkan Kategori Topik Menggunakan Metode Algoritma Naïve Bayes Dan Algoritma C4.5

1st Luthfi Ahmad Muhaimin
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
luthfiahmadmuhaimin@telkomuniversity.ac.id

2nd Oktariani Nurul Pratiwi
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
onurulp@telkomuniversity.ac.id

3rd Riska Yanu Fa'rifah
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
riskayanu@telkomuniversity.ac.id

Klasifikasi dilakukan untuk mengelompokkan sekumpulan data ke dalam kelas-kelas yang telah ditentukan terlebih dahulu berdasarkan kesamaan karakteristik yang dimiliki. Klasifikasi soal berdasarkan topik membantu para siswa dan pengajar dalam mengambil keputusan untuk menentukan soal berdasarkan kategori topiknya. Pada penelitian ini, peneliti bermaksud untuk membuat suatu model klasifikasi soal Biologi kelas 11 SMA yang dikelompokkan menjadi sembilan kategori topik yaitu Sel, Jaringan Tumbuhan dan Hewan, Sistem Gerak Manusia, Sistem Peredaran Darah, Sistem Pencernaan, Sistem Pernapasan, Sistem Ekskresi, Sistem Koordinasi, dan Sistem Reproduksi Manusia. Soal-soal dan topik didapatkan dari buku bank soal yang berjudul "*Siap Pintar Belajar Mandiri*". Penelitian ini membandingkan nilai akurasi dan evaluasi performansi dari dua algoritma klasifikasi yaitu, Naive Bayes dan C4.5. Untuk evaluasi performansi peneliti menggunakan *cross validation* dan mencari nilai *precision, recall*, dan *f1-score* menggunakan *confusion matrix*. Dari hasil klasifikasi, diperoleh hasil akurasi algoritma Naive Bayes sebesar 72.72%, dan nilai akurasi evaluasi performansi menggunakan cross validation sebesar 73.09% dan nilai *precision* sebesar 73%, *recall* sebesar 73%, dan *F1-Score* sebesar 70%. Sedangkan algoritma C4.5 mendapat nilai akurasi sebesar 54.54%, dan nilai akurasi evaluasi performansi menggunakan cross validation sebesar 54.09% dan nilai *precision* sebesar 58%, *recall* sebesar 56%, dan *F1-Score* sebesar 55%.

***Kata kunci*— Klasifikasi Soal, Biologi, Naive Bayes, C4.5, Cross Validation**

I. PENDAHULUAN

E-Learning merupakan salah satu bentuk model pembelajaran yang difasilitasi dan didukung pemanfaatan teknologi informasi dan komunikasi [1]. Pendidikan merupakan hal yang penting bagi seluruh manusia. Ada beberapa daerah yang sudah menerapkan peraturan wajib belajar 12 tahun yaitu dari Sekolah Dasar (SD) sampai Sekolah Menengah Atas (SMA) atau Sekolah Menengah Kejuruan (SMK). Siswa kelas 10 SMA merupakan siswa yang masih dalam tahap transisi dari Sekolah Menengah Pertama (SMP) dan masih berfokus pada penjurusan kelas. Siswa kelas 12 SMA merupakan siswa yang sudah fokus pada Ujian Nasional (UN) dan sudah fokus pada mata pelajaran yang dipilihnya saja seperti Fisika, Ekonomi, Biologi, dan yang lainnya. Sedangkan kelas 11 merupakan kelas pertengahan yang tidak lagi berfokus pada penjurusan kelas dan UN [2]. Pada jenjang SMA dibagi menjadi dua jurusan yaitu Ilmu Pengetahuan Alam (IPA) dan Ilmu Pengetahuan Sosial (IPS). Pada jurusan IPA terdapat mata pelajaran wajib yang harus dipelajari yaitu Matematika, Fisika, Biologi, dan Kimia. Biologi adalah salah satu mata pelajaran yang memiliki berbagai macam istilah latin dan materi yang kompleks yang terkadang membuat siswa merasa sulit memahaminya, biologi juga memiliki cakupan topik materi yang luas seperti manusia, hewan, dan tumbuhan [3]. Buku Siap Pintar Belajar Mandiri (SPBM) merupakan suatu buku untuk pelajar atau guru SMA yang berisi materi dan soal-soal yang lengkap dan mudah dipahami yang memuat mata pelajaran penting seperti Biologi dan buku ini juga terdapat varian untuk setiap kelasnya.

Ada berbagai macam metode klasifikasi, diantaranya adalah algoritma Naive Bayes dan algoritma C4.5. Algoritma Naive Bayes dan algoritma C4.5 merupakan metode klasifikasi pada *text mining*. Pada penelitian yang dilakukan dalam [4] yang membandingkan algoritma Naive Bayes, Support Vector Machine, C4.5, dan K-Nearest Neighbour. Pada penelitian mereka, data yang digunakan adalah ulasan pada aplikasi bibit yang ada di *play store*. Data yang digunakan terdiri dari 464 *dataset* yang terbagi menjadi 310 data sentimen positif dan 154 sentimen negatif. Hasil dari penelitian mereka adalah algoritma Naïve Bayes memiliki nilai akurasi tertinggi sebesar 84.91%, C4.5 sebesar 76.94%, Support Vector Machine sebesar 71.77%, dan K-Nearest Neighbour sebesar 66,81. Adapun penelitian yang dilakukan dalam [5] yang membandingkan algoritma Support Vector Machine, Naive Bayes, dan C4.5. Data yang digunakan adalah *Short Message Service* (SMS) spam dan non spam. dataset terdiri dari 100 SMS spam dan 100 SMS non spam. Hasil dari penelitian tersebut adalah didapat nilai akurasi tertinggi oleh C4.5 sebesar 95.5%, Naïve Bayes sebesar 95%, sedangkan Support Vector Machine sebesar 76%.

Algoritma Naive Bayes dan C4.5 digunakan karena dianggap cukup mudah dipahami dan merupakan algoritma yang sangat populer dalam melakukan klasifikasi. Naive Bayes merupakan algoritma klasifikasi dengan rumus yang sederhana dan mudah untuk diaplikasikan, serta mampu menghasilkan klasifikasi yang akurat [6]. Sedangkan algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterprestasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik [7]. Sehingga pada tugas akhir ini peneliti bermaksud melakukan perbandingan hasil akurasi antara algoritma Naive Bayes dan algoritma C4.5 dalam melakukan klasifikasi topik soal pada mata pelajaran biologi SMA kelas 11 dengan menggunakan *dataset* soal biologi yang terdapat pada buku SPBM kelas 11 SMA

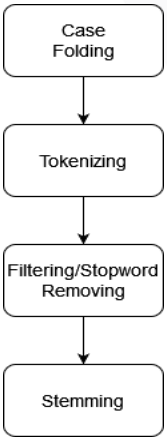
II. KAJIAN TEORI

A. *Text Mining*

Text mining merupakan proses untuk mengekstraksi pengetahuan yang terkandung dalam data yang bersifat teks. *Text mining* merupakan proses yang bisa menyelesaikan masalah informasi yang berlebih dengan menggunakan teknik *data mining, machine learning, natural language processing* (NLP*), information retrieval* (IR), dan *knowledge management* [9] . *Text classification, clustering*, dan *association* merupakan jenis tugas yang dapat dilakukan pada proses *text mining* [8]. *Text mining* berbeda dengan *data mining*. Untuk *text mining* data yang digunakan berupa dokumen atau kumpulan teks. *Text mining* melakukan proses untuk menemukan hubungan atara suatu teks dengan teks lainnya dengan standar yang telah ditentukan.

B. *Text Preprocessing*

Text preprocessing adalah suatu proses menyeleksi atau menyiapkan data berupa teks yang tidak terstruktur menjadi terstruktur. Proses *preprocessing* dilakukan agar data yang digunakan bebas dari noise, memiliki dimensi yang lebih kecil, serta lebih terstruktur, sehingga dapat diproses lebih lanjut [9]. Ada 4 tahapan dalam text preprocessing [9].



GAMBAR 1. TAHAPAN TEXT PREPROCESSING

1. *Case Folding*
   Case folding merupakan proses dalam *text preprocessing* yang dilakukan untuk menyeragamkan karakter pada suatu data. Proses *case folding* adalah proses mengubah seluruh huruf menjadi huruf kecil. Pada proses ini karakter-karakter ‘A’-‘Z’ yang terdapat dalam suatu data diubah menjadi karakter ‘a’-‘z’. Karakter-karakter selain huruf ‘a’ sampai ‘z’ (tanda baca dan angka) akan dihilangkan dari data dan dianggap sebagai *delimeter*. *Delimeter* merupakan urutan satu atau lebih karakter yang digunakan untuk menentukan batas pemisah.
2. *Tokenizing*
   Sebelum data/teks dapat diolah ke tahap selanjutnya, data tersebut harus disegmentasi menjadi kata-kata, proses ini disebut *tokenizing*. Tahap *tokenizing* merupakan tahap pemotongan string input berdasarkan kata-kata yang menyusunnya atau bisa disebut pemecahan kalimat menjadi kata. Strategi umum yang dilakukan pada tahap *tokenizing* adalah memotong kata pada *white space* atau spasi dan membuang karakter tanda baca. Tahap *tokenizing* membagi urutan karakter menjadi kalimat dan kalimat menjadi *token.*
3. *Filtering/Stopword Removal*
   Setelah tahap *tokenizing*, dilakukanlah tahap *filtering* atau *stopword removal* yaitu dengan menghilangkan kata-kata yang sangat umum. Kata yang termasuk dalam *stopword* contohnya adalah yang, dan, di, itu, dengan, untuk, tidak, dari, dalam, akan, pada, ini, juga, saya, serta, adalah, bahwa, lain, kamu, dan lain lain.
4. *Stemming*
   Stemming merupakan tahapan dari *text preprocessing* yang bertujuan untuk mengubah istilah ke bentuk akar katanya. *Stem* (akar kata) adalah bagian dari kata yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran)

C. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan skema yang cukup poluler di bidang information retrieval. TF-IDF memerlukan referensi ke seluruh teks yang disebut corpus. Bobot kata yang dihitung dengan skema TF-IDF sebanding dengan jumlah kemunculan suatu teks tertentu, tetapi berbanding terbalik dengan teks lainnya [8]. Untuk melakukan perhitungan bobot pada dokumen atau teks, dapat dilakukan perhitungan dengan formula seperti di sebagai berikut.

$$w_{ij} = \log\frac{N}{DF_i}(1 + \log TF_i) \quad (1)$$

Keterangan:

a) w_{ij} : *Term Frequency-Inverse Document Frequency* (TF-IDF)
b) N : Jumlah kalimat yang berisi term
c) DF_i : *Inverse Document Frequency*
d) TF_i : *Term Frequency*

D. Algoritma Naïve Bayes

Pengklasifikasi bayes adalah pengklasifikasi statistikyang dapat memprediksi probabilitas keanggotaan kelas dari suatu data tuple yang akan

masuk ke dalam kelas tertentu sesuai dengan perhitungan probabilitas. Pengklasifikasi *Bayes* didasarkan pada teorema *bayes*, yang ditemukan oleh Thomas Bayes pada abad ke-18. Dalam sebuah studi perbandingan algoritma klasifikasi telah ditemukan simple bayesian atau yang biasa dikenal dengan Naive Bayes *classifier.* Naive Bayes *classifier* memberikan nilai akurasi dan kecepatan yang tinggi ketika diterapkan pada *database* yang besar. Metode ini banyak digunakan dalam menyelesaikan masalah dalam bidang *machine learning* karena metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana [20]. Di bawah ini adalah bentuk umum dari teorema bayes.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (2)$$

Keterangan:

a) X : Data yang belum diketahui kelasnya
b) H : Hipotesis pada data X yang merupakan suatu class spesifik
c) $P(H|X)$: probabilitas akhir bersyarat suatu hipotesis H terjadi jika diberikan bukti X terjadi
d) $P(H)$: Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun
e) $P(X|H)$: Probabilitas sebuah bukti X terjadi akan mempengaruhi hipotesis H.
f) $P(X)$: Probabilitas awal (priori) bukti X terjadi tanpa memandang hipotesis/ bukti yang lain

Dengan menggunakan persamaan di atas, nilai yang telah diperoleh dapat diproses dengan algoritma Naïve Bayes untuk penilaian data yang akan diklasifikasikan.

E. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang dapat digunakan untuk membuat pohon keputusan (*decision tree*). Algoritma C4.5 merupakan salah satu algoritma dalam induksi *decision tree* yaitu ID3 (*Iterative Dichotomiser 3*) yang dikembangan oleh J. Ross Quinlan. Dalam prosedur algoritma ID3, input adalah sampel *training*, label *training* dan atribut. Algoritma C4.5 ini juga merupakan pengembangan dari ID3. Ide dasar dari algoritma ini adalah membuat pohon keputusan dengan memilih atribut yang memiliki prioritas tertinggi atau dapat disebut memiliki nilai *gain* tertinggi berdasarkan nilai entropy atribut tersebut sebagai poros atribut klasifikasi. Kemudian secara rekursif cabang-cabang pohon diperluas sehingga seluruh pohon terbentuk [17]. Terdapat empat langkah dalam proses pembuatan pohon keputusan pada algoritma C4.5, yaitu:

a Memilih atribut sebagai akar.
b Membuat cabang untuk masing-masing nilai.
c Membagi setiap kasus dalam cabang.
d Mengulangi proses dalam setiap cabang sehingga semua kasus dalam cabang memiliki kelas yang sama.

Berikut dibawah ini merupakan rumus untuk memperoleh data yang bisa diproses algoritma C4.5 untuk membuat *decision tree* sebagai berikut:

1. Pertama dilakukan perhitungan untuk mencari nilai *entropy*.
   Berikut ini rumus untuk mencari nilai *entropy*.

$$Entropy(S) = \sum_{j=1}^{k} -pj \ \log 2 \ pj \quad (3)$$

Keterangan:

a) S : Himpunan kasus
b) k : Jumlah partisi S
C) pj : Jumlah kasus pada partisi ke-j

2. Kedua dilakukan perhitungan *gain* setelah mendapat nilai *entropy*.
   Berikut ini rumus untuk mencari nilai *gain*.

$$Gain(A) = Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (4)$$

Keterangan:

a) A : Atribut dari dataset
b) k : Jumlah partisi S
c) S : Himpunan Kasus

Dengan menggunakan 2 persamaan di atas, nilai yang telah diperoleh dapat diproses dengan algoritma C4.5 untuk membuat *decision tree.*

F. Cross Validation

Metode *cross validation* merupakan metode yang digunakan untuk mengevaluasi kinerja model maupun algoritma. Salah satu variasi dari teknik pengujian *cross validation* yaitu *k-fold*. Metode ini dilakukan dengan membagi sampel data menjadi k bagian yang rata untuk digunakan sebagai *data training set* dan *data test set*, secara berulang-ulang k kali [18].

G. Confusion Matrix

Confusion matrix merupakan tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [14]. Berikut merupakan contoh tabel dari *confusion matrix* untuk klasifikasi biner.

TABEL 1. CONFUSION MATRIX

		Kelas Prediksi	
		1	0
	1	TP	FN

Kelas Sebernarnya	0	FP	TN

Keterangan:

a) TP (*True Positive*) = jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1
b) TN (*True Negative*) = jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0
c) FP (*False Positive*) = jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1
d) FN (*False Negative*) = jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0

Berikut merupakan rumus *confusion matrix* untuk menghitung *accuracy, precision*, dan *recall* seperti berikut.

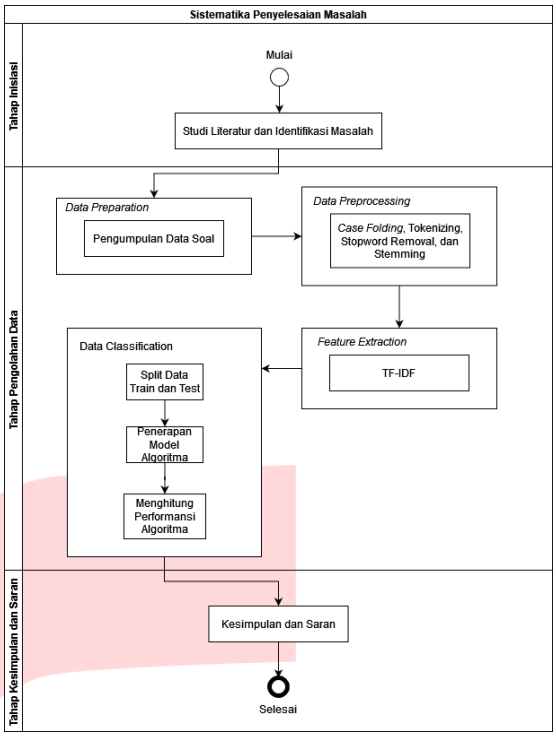
$$accuracy = \frac{TP + TN}{Total} \quad (5)$$

$$precision = \frac{TP}{TP + FP} \quad (6)$$

$$recall = \frac{TP}{TP + FN} \quad (7)$$

III. METODE

Penelitian ini menggunakan beberapa tahapan untuk mendapatkan hasil klasifikasi menggunakan kedua algoritma. Tahapan pertama dimulai dari inisiasi yang dilanjutkan dengan pengolahan data dan diakhiri dengan kesimpulan dan saran. Tahapan penelitian ditunjukkan pada GAMBAR 1, dan selanjutnya menjelaskan masing-masing tahapan yang dilakukan.



GAMBAR 2.
SISTEMATIKA PENELITIAN

A. Tahap Inisiasi

Pada tahap ini peneliti mengidentifikasi masalah terkait topik yang telah diambil. Untuk topik yang diambil peneliti adalah klasifikasi topik soal pada mata pelajaran biologi SMA kelas 11. Setelah mengidentifikasi masalah, kemudian peneliti melakukan studi literatur sebagai bahan referensi untuk dijadikan acuan dalam mengerjakan Tugas Akhir sesuai dengan topik yang telah diambil. Selanjutnya peneliti menentukan solusi dari permasalahan yang telah diambil, berdasarkan studi literatur terkait identifikasi masalah. Setelah menentukan solusi, peneliti menentukan batasan masalah penelitian yang akan diselesaikan. Kemudian peneliti menentukan tujuan dan manfaat dari penelitian yang diambil.

B. Tahap Pengolahan Data

Pada tahap pengolahan data terdapat empat bagian yaitu *Data Preparation*, *Data Preprocessing*, *Feature Extraction* dan *Data Classification*. Pada bagian pertama tahap pengolahan data adalah *Data Preparation*, pada bagian ini pertama peneliti melakukan proses pengumpulan data. Data yang dikumpulkan adalah soal-soal mata pelajaran biologi SMA kelas 11. Setelah itu soal-soal tersebut dikelompokkan berdasarkan kategori topik, serta menentukan atribut-atribut data yang akan digunakan sebagai *dataset*.

Setelah melakukan proses pembuatan *dataset* dilakukan, masuk kebagian *Data Preprocessing*. Pada tahap ini peneliti melakukan proses *data cleansing* pada soal sesuai dengan batasan masalah yang telah ditentukan, misalnya soal hanya bersifat pilihan ganda dan tidak terdapat gambar. Pada bagian ini juga peneliti melakukan proses *text preprocessing*. Adapun tahap-tahap pada proses *text preprocessing* yaitu *case folding, tokenizing, stopwords removal*, dan *stemming*. Pada *case folding* proses yang dilakukan adalah menyeragamkan kata-kata menjadi bentuk yang sama, misal

huruf kapital dan huruf kecil. Kemudian, masuk ke tahap *tokenizing* yaitu proses membagi teks menjadi beberapa bagian dalam bentuk kata-kata, frasa atau elemen bermakna lainnya. Lalu dilanjutkan dengan proses *stopwords removal* untuk menghilangkan kata-kata yang tidak menjadi fokus penelitian. Setelah itu masuk proses *stemming* dilakukan untuk menghilangkan imbuhan pada kata sehingga didapatkan dasar katanya. Setelah itu masuk ke tahap *Feature Extraction* menggunakan TF-IDF. Dengan menggunakan TF-IDF peneliti memberikan bobot nilai pada setiap kata yang telah melalui proses *text preprocessing*.

Bagian terakhir dari tahap pengolahan data yang dilakukan pada penelitian ini adalah *Data Classification*. Pada bagian ini hasil TF-IDF pertama akan dibagi menjadi *data training* dan *data testing* dan setelah itu akan diklasifikasikan berdasarkan kategori topik. Setelah itu dihitung hasil akurasi performansi algoritma yang digunakan.

C. Hasil dan Kesimpulan

Setelah selesai melakukan tahap pengumpulan dan pengolahan data, peneliti melakukan evaluasi perbandingan performansi akurasi algoritma Naive Bayes dan C4.5. Hasil perbandingan akurasi dibandingkan dalam bentuk grafik. Setelah itu peneliti menarik kesimpulan dan saran dari analisis hasil yang telah dilakukan.

III. HASIL DAN PEMBAHASAN

Data preprocessing dilakukan melalui beberapa tahap agar data tersebut dapat diterima atau digunakan oleh model. Berikut merupakan hasil *data preprocessing* pada tahap pengolahan data.

TABEL 2.
DATA PREPROCESSING

Preprocessing	Hasil *Preprocessing*
Sebelum *Preprocessing*	Organel berikut yang dimiliki oleh sel hewan dan tumbuhan adalah…: plastida, lisosom, dan nucleus; lisosom, sentrosom, dan ribosom; sentrosom, sentriol, dan plastida; mitokondria, nukleus, dan ribosom; inti sel, kloroplas, dan ribosom
Case Folding	organel berikut yang dimiliki oleh sel hewan dan tumbuhan adalah plastida lisosom dan nucleus lisosom sentrosom dan ribosom sentrosom sentriol dan plastida mitokondria nukleus dan ribosom inti sel kloroplas dan ribosom
Tokenizing	['organel', 'berikut', 'yang', 'dimiliki', 'oleh', 'sel', 'hewan', 'dan', 'tumbuhan', 'adalah', 'plastida', 'lisosom', 'dan', 'nucleus', 'lisosom', 'sentrosom', 'dan', 'ribosom', 'sentrosom', 'sentriol', 'dan', 'plastida', 'mitokondria', 'nukleus', 'dan', 'ribosom', 'inti', 'sel', 'kloroplas', 'dan', 'ribosom']
Stopword Removal	['organel', 'dimiliki', 'sel', 'hewan', 'tumbuhan', 'plastida', 'lisosom', 'nucleus', 'lisosom', 'sentrosom', 'ribosom', 'sentrosom', 'sentriol', 'plastida', 'mitokondria', 'nukleus', 'ribosom', 'inti', 'sel', 'kloroplas', 'ribosom']
Stemming	['organel', 'milik', 'sel', 'hewan', 'tumbuh', 'plastida', 'lisosom', 'nucleus', 'lisosom', 'sentrosom', 'ribosom', 'sentrosom', 'sentriol', 'plastida', 'mitokondria', 'nukleus', 'ribosom', 'inti', 'sel', 'kloroplas', 'ribosom']

Setelah melalui tahap *preprocessing* peneliti masuk kedalam tahap *feature extraction* dengan menggunakan TF-IDF pada tahap ini setiap kata pada dokumen diberikan bobot nilai. Perhitungan bobot nilai diambil dari hasil *stemming*, dibawah ini menunjukan beberapa hasil kata yang telah diberi bobot nilai.

TABEL 3
HASIL TF-IDF

Preprocessing	TF-IDF
'organel'	0.1362952800442604
'milik'	0.1289547714810576
'sel'	0.1532798011842
'hewan'	0.16930228864235303

Setelah dilakukan *tahap feature extraction* menggunakan TF-IDF, peneliti melakukan perhitungan akurasi pada algoritma Naïve Bayes dan C4.5. Tahap ini juga peneliti melakukan *train test split* untuk memisahkan *data training* dan *data testing* sebelum melakukan implementasi algoritma Naïve Bayes dan C4.5, dengan rasio yang akan di uji adalah 90:10, 80:20, 70:30, dan 60:40. Nantinya rasio dengan hasil terbaik akan digunakan sebagai nilai akurasi dapat.

TABEL 4
AKURASI NAÏVE BAYES

	90:10	80:20	70:30	60:40
Naïve Bayes	72.72%	57.14%	37.50%	35.71%

TABEL 5
AKURASI C4.5

	90:10	80:20	70:30	60:40
C4.5	54.54%	52.38%	40.62%	40.47%

Selanjutnya peneeneliti melakukan evaluasi kinerja model dari algoritma Naive Bayes dan C4.5 menggunakan metode *K-Fold Cross Validation*. Pengujian *Cross Validation* menggunakan algoritma Naive Bayes dan C4.5 dengan menerapkan *k-fold* sebanyak 10. Hasil skor dari 10 *fold Cross Validation* untuk algoritma Naive Bayes.

TABEL 6.
AKURASI CROSS VALIDATION NAÏVE BAYES

Fold ke-	Akurasi
1	63.63%
2	72.72%
3	81.81%
4	90.90%
5	81.81%
6	70%
7	70%
8	70%
9	60%
10	70%

Dari penerapan tersebut diperoleh rata-rata skor pada algoritma Naive Bayes sebesar 73.09%. Untuk hasil skor dari 10 *fold Cross Validation* untuk algoritma C4.5.

TABEL 7.
AKURASI CROSS VALIDATION C4.5

Fold ke-	Akurasi
1	54.54%
2	72.72%
3	45.45%
4	72.72%
5	45.45%
6	50%
7	50%
8	40%
9	70%
10	40%

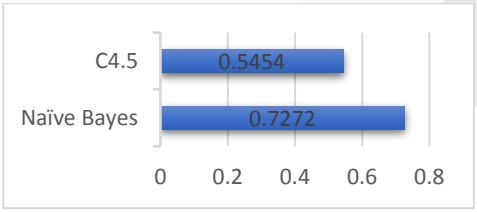
Dari penerapan *cross validation* dengan 10 *fold* juga diperoleh rata-rata skor pada algoritma C4.5 sebesar 54.09%. Selain itu juga, peneliti juga menghitung nilai *recall, precision* dan *f1-score*. Nilai-nilai ini didapat dari rata-rata nilai hasil *confusion matrix* menggunakan *cross validation.*

TABEL 8.
F1-MEASURE

Metode	Precision	Recall	F1-Score	Accuracy
Naïve Bayes	73%	73%	70%	73%
C4.5	58%	56%	55%	57%

nilai *recall, precision* dan *f1-score* diambil berdasarkan nilai pada masing-masing hasil *confusion matrix.* Berdasarkan Tabel 7 juga dapat dilihat hasil evaluasi performansi *f1-measure* algoritma Naïve Bayes memberikan nilai yang lebih baik dibandingkan algoritma C4.5

Adapun grafik perbandingan antara algoritma Naïve Bayes dan C4.5 yang dapat dilihat dibawah ini.



GAMBAR 3.
PERBANDINGAN AKURASI

Berdasarkan hasil akurasi yang telah didapat pada penelitian ini, algoritma Naive Bayes merupakan metode terbaik dalam melakukan klasifikasi soal berdasarkan topik kategori.

IV. KESIMPULAN

Berdasarkan hasil penelitian ini, peneliti mendapatkan beberapa kesimpulan. Klasifikasi soal membantu para siswa dan pengajar dalam mengambil keputusan untuk menentukan soal berdasarkan kategori topiknya. Berdasarkan hasil akurasi algoritma Naïve Bayes dan C4.5 dalam melakukan klasifikasi topik soal biologi SMA kelas 11 berdasarkan kategori topik, algoritma Naïve Bayes memiliki hasil yang lebih baik dari algoritma C4.5. Setelah melakukan evaluasi performansi dengan *cross validation* dan *confusion matrix* juga, algoritma Naïve Bayes menunjukkan hasil yang lebih baik dari algoritma C4.5.

REFERENSI

[1] N. S. Hanum, "Keefetifan e-learning sebagai media pembelajaran (studi evaluasi model pembelajaran e-learning SMK Telkom Sandhy Putra Purwokerto)," *J. Pendidik. Vokasi*, vol. 3, no. 1, pp. 90–102, 2013.

[2] P. Winarni, H. H. Pranoto, and L. D. Afriani, "Hubungan antara Pengetahuan Tentang Gizi Seimbang dengan Perilaku Pemenuhan Gizi Seimbang pada SiswaKelas XI SMA Negeri 1 Ungaran," *J. Gizi Dan Kesehat.*, vol. 7, no. 15, pp. 1–8, 2015.

[3] N. A. Yulistiawati, "Pentingnya Motivasi Peserta Didik terhadap Hasil Belajar Biologi," *Semin. Nasioal Biol. VI*, pp. 1–4, 2019.

[4] A. Z. Kamalia, A. A. Zaroni, and M. Wangsadanureja, "Analisis Sentimen Pada Ulasan Aplikasi Bibit Di Play Store Dengan Metode Naive Bayes, Support Vector Machine, C4.5 Dan K-Nearest Neighbor," vol. 13, no. 1, pp. 9–25, 2019.

[5] R. Sari, "Komparasi Algoritma Support Vector Machine, Naïve Bayes Dan C4.5 untuk Klasifikasi SMS," *IJCIT(Indonesia J. Comput. Infomation Technol.*, vol. 2, no. 2, pp. 7–13, 2017.

[6] Maryamah, F. Asikin, D. Kurniawaty, S. K. Sari, and I. Cholissodin, "Implementasi Metode Naïve Bayes Classifier Untuk Seleksi Asisten Praktikum Pada Simulasi Hadoop Multinode Cluster," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 4, pp. 273–278, 2016.

[7] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *J. Ultim.*, vol. 6, no. 1, pp. 15–20, 2014.

[8] T. Jo, *Text Mining Concepts, Implementation, and Big Data Challenge*. Seoul, 2021.

[9] R. Feldman and J. Sanger, *The Text Mining Handbook-Advanced Approaches in Analyzing Unstructured Data*. New york, 2007.

[10] K. L. Kohsasih and Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular," *J. Inform.*, vol. 9, no. 1, pp. 13–17, 2022.

[11] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 1, pp. 19–25, 2017.

[12] Y. I. Kurniawan, "Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 455, 2018.

[13] M. Listiana, Sudjalwo, and D. Guanawan, "Perbandingan Algoritma Decision Tree (C4.5) Dan Naive Bayes Pada Data Mining Untuk Identifikasi Tumbuh Kembang Anak Balita (Studi Kasus Puskesmas Kartasura)," *Syria Stud.*, vol. 7, no. 1, pp. 37–72, 2015.
[14] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform.*, vol. 5, no. 2, pp. 697–711, 2021.
[15] D. Prajarini, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit," *Informatics J.*, vol. 1, no. 3, p. 137, 2016.
[16] A. H. Yunial, "Analisa Perbandingan Klasifikasi Support Vector Machine, Decession Tree dan Naive Bayes," vol. 5, pp. 169–185, 2020.
[17] Purushottam, K. Saxena, and R. Sharma, "Efficient heart disease prediction system using decision tree," *Int. Conf. Comput. Commun. Autom.*, 2015.
[18] S. Bahri, D. M. Midyanti, and R. Hidayati, "Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak," *Semin. Nas. Apl. Teknol. Inf.*, pp. 24–31, 2018.
[19] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data*. 2014.
[20] F. Handayani and F. 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.