JNATIA Volume 3, Nomor 2, Februari 2025 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya e-ISSN: 3032-1948

Klasifikasi Berita Berdasarkan Kategori Menggunakan Multinomial Naïve Bayes dengan K-Cross Validation dan Seleksi Fitur Chi-Squared

Febrian Valentino Agape[a1], Gst. Ayu Vida Mastrika Giri[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]febrian.valentino1402@gmail.com
[2]vida@unud.ac.id

Abstract

Classifying news articles based on categories is an important challenge in text analysis and natural language processing. Most categorization of online news articles is often done manually, making it a complex and time-consuming process. To address this issue, the development of an automatic system capable of classifying news articles into various categories such as technology, sports, and entertainment is needed. The system is built using an approach to classify news articles into several appropriate categories using the Naïve Bayes method with TF-IDF weighting and feature selection using Chi-Squared. The Naïve Bayes model training uses the reduced feature results of 10,000 features from 54,091 features. Evaluation results show that the Naïve Bayes approach is able to produce a news classification model with good accuracy, with accuracy, precision, recall, and f1-score values of 96%.

Keywords: *News Classification, Multinomial Naïve Bayes, Feature Weighting, TF-IDF (Term Frequency-Inverse Document Frequency), Text Analysis*

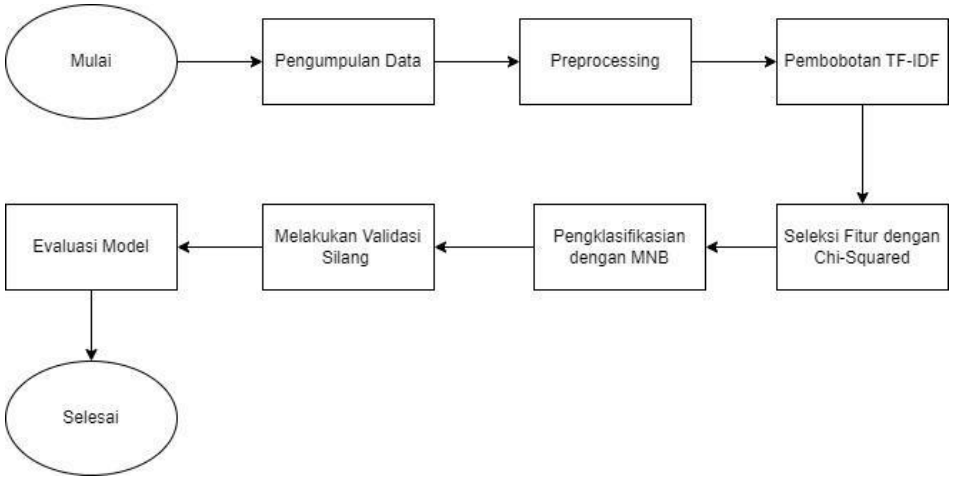
1. Pendahuluan

Di tengah banjir informasi yang kita hadapi hari ini, menimbulkan beragam topik berita yang bermunculan bersamaan. Berita mengandung elemen-elemen fakta dan pendapat yang penting untuk dikomunikasikan ke publik, namun tidak setiap elemen tersebut layak untuk dijadikan berita oleh media[1]. Seiring berjalannya waktu, media penyampaian berita telah berubah dari surat kabar, majalah, radio, dan televisi menjadi dominan di platform digital. Biasanya, konten berita di situs berita internet dikategorikan ke dalam berbagai segmen seperti politik, olahraga, ekonomi, hiburan, teknologi, kesehatan, dan sebagainya. Permasalahan timbul yaitu, Melimpahnya dokumen digital di internet dapat menjadi tantangan bagi masyarakat dalam mengakses informasi jika tidak ada sistem pengelolaan yang memadai. Cara umum untuk mengatur konten berita adalah dengan mengklasifikasikan setiap artikel berdasarkan kategori tertentu. Kategori ini bisa ditentukan berdasarkan situasi sosial yang berlaku atau sesuai dengan standar yang telah ditetapkan [2]. Kemampuan untuk mengklasifikasikan berita berdasarkan kategori menjadi sangat penting. Klasifikasi ini tidak hanya membantu dalam menyaring informasi yang relevan tetapi juga memudahkan pengguna dalam menemukan konten yang mereka cari.Teknik penambangan teks merupakan salah satu metode yang efektif untuk klasifikasi dokumen. Proses ini melibatkan penentuan kategori yang sesuai untuk dokumen berdasarkan isi teksnya [3]. Tujuan dari klasifikasi adalah untuk mengembangkan sebuah model atau fungsi yang dapat membedakan antara berbagai konsep atau kategori data, sehingga memungkinkan prediksi kategori untuk objek yang diberikan. Ini mengasumsikan adanya beragam kategori yang dapat diterapkan pada objek tersebut. Klasifikasi telah banyak dilakukan oleh para peneliti dengan menerapkan berbagai metode, salah satunya adalah Multinomial Naïve Bayes [4]. Terdapat cukup banyak penelitian terdahulu yang telah melakukan klasifikasi teks dengan metode ini. Metode Multinomial Naive Bayes telah menunjukkan efektivitasnya dalam memberikan hasil yang memadai untuk klasifikasi teks [5]. Sebagai ilustrasi, penelitian yang dilakukan oleh Wayan Firdaus Mahmudy dan Agus

Wahyu Widodo menggunakan Naive Bayes Classifier yang belum dimodifikasi. Mereka mengevaluasi berbagai rasio pembagian data latih dan uji, yaitu 5:95, 10:90, 15:85, 20:80, 25:75, dan 30:70. Hasilnya, akurasi klasifikasi yang diperoleh secara bertahap adalah 54%, 65%, 65%, 69%, 71%, dan 76% [2]. Selain itu, dari hasil penelitian Bobby Suryo Prakoso dkk penelitian klasifikasi berita menggunakan metode Multinomial Naïve Bayes Dengan Seleksi Fitur Dan Boosting dimana menghasilkan tingkat akurasi, recall, dan presisi sebesar 73,2%. Tidak ada perbedaan antara model yang lebih rinci dan model Naive Bayes Classifier, yang berarti keduanya memiliki performa evaluasi yang sama. Ini menandakan bahwa kedua model tersebut memiliki kemampuan prediksi yang sebanding [6]. Dalam penelitian ini melakukan klasifikasi teks diperlakukan ekstraksi fitur untuk mengubah format tekstual yang tidak terstruktur menjadi terstruktur sehingga dapat diproses oleh model machine learning seperti Multinomial Naïve Bayes untuk mengklasifikasikan ke kelas yang telah ditentukan [7]. Pada penelitian ini data dibagi menjadi 80% data training dan 20% testing menggunakan Teknik ekstraksi fitur yaitu TF-IDF serta dilakukan optimalisasi model dengan cara melakukan seleksi fitur untuk mencegah masalah overfitting dengan menggunakan metode Chi-squared yang membedakan dengan penelitian penelitian sebelumnya karena dilihat dari hasil akurasi menghasilkan persentase lebih tinggi karena adanya metode optimalisasi tersebut.

2. Metode Penelitian

Bagian ini akan menggambarkan secara umum tahapan yang akan dilakukan oleh peneliti dalam penelitian yang bertujuan agar mempermudah peneliti dalam melakukan percobaan serta pengkajian data. Alur metodologi penelitian dapat dilihat pada gambar 1.



Gambar 1. Alur Metode Penelitian

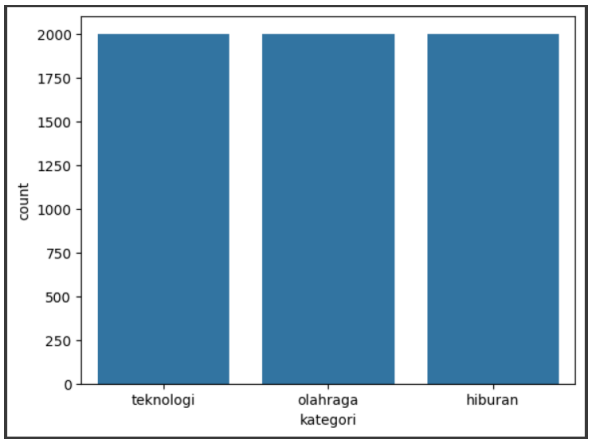
Penelitian akan dimulai dengan mengumpulkan data teks berupa dataset yang bersumber dari repository Github iwanlaudin0101 yang berisi kumpulan berita yang sudah dilabeli kategori nya. Setelah itu, akan dilakukan tahap preprocessing untuk mempersiapkan data tersebut. Selanjutnya, data akan diolah dengan menggunakan metode pembobotan kata menggunakan Term Frequency Inverse Document Frequency (TF-IDF). Setelah mendapatkan hasil pembobotan kata, langkah selanjutnya adalah melakukan proses seleksi fitur untuk menghindari masalah proses klasifikasi seperti overfitting selanjutnya dilanjutkan dengan proses pengklasifikasian dengan model Multinomial Naïve Bayes. Pada tahap akhir penelitian, dilakukan pengujian dan evaluasi terhadap kinerja metode yang digunakan.

2.1. Tahapan Pengumpulan Data

Pada proses pengumpulan data, data yang digunakan dalam penelitian ini merupakan data sekunder yang berasal dari repository GitHub dengan nama Iwanlaudin0101. Dataset ini terdiri dari tiga variabel utama, yaitu sumber, kategori, dan berita. Namun, untuk keperluan penelitian ini, peneliti hanya akan menggunakan dua variabel, yaitu kategori dan berita. Dataset ini terdiri

p-ISSN: 2986-3929
e-ISSN: 3032-1948

dari lima kategori berita yang berbeda, namun pada penelitian ini akan berfokus pada tiga kategori utama, yaitu teknologi, olahraga, dan hiburan. Oleh karena itu, dua kategori lainnya, yaitu showbiz dan tajuk utama, akan dihilangkan dari dataset. Sebagai hasilnya, penelitian ini menggunakan data set yang terdiri dari 6000 record dimana distribusi tiap kategori dapat dilihat pada gambar 2.



Gambar 2. Distribusi Data Per Kategori

Masing-masing kategori memiliki 2000 record dimana distribusi antara kategori teknologi, olahraga, dan hiburan adalah seimbang, dengan masing-masing kategori memiliki jumlah yang sama. Data ini akan dibagi menjadi data training dan data testing dengan perbandingan 80% : 20%. Oleh karena itu, sebanyak 4800 record akan digunakan sebagai data training, sedangkan 1200 record sisanya akan digunakan sebagai data testing.

2.2. Text Preprocessing

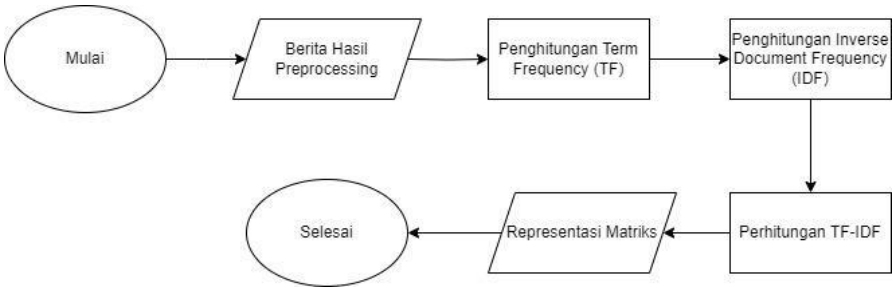
Text Preprocessing merupakan tahap awal dalam membangun sebuah model machine learning dalam text mining. Pada langkah ini, dilakukan pra-pemrosesan data teks yang telah dikumpulkan sebelumnya. Preprocessing teks adalah suatu proses untuk mengubah data teks yang tidak terstruktur menjadi data yang terstruktur, atau dengan kata lain, mengubah teks menjadi indeks kata sesuai kebutuhan [8]. Tujuan dari proses ini adalah untuk mempersiapkan teks agar siap digunakan dan diolah. Pra-pemrosesan melibatkan serangkaian langkah, meliputi cleaning, case folding, tokenization, stopwords removal, dan Lemetazing. Dari tahapan preprocessing text ditujukan untuk mengurangi informasi yang tidak relevan atau tidak dibutuhkan dalam data tersebut dengan menghilangkan kata atau teks yang tidak perlu. Semua langkah ini bertujuan untuk mempermudah proses pembobotan. Ilustrasi alur dari tahap pra-pemrosesan dapat dilihat pada Gambar 3
.

Gambar 3. Alur Text Preprocessing

Gambar 3 menunjukkan bahwa proses pra-pemrosesan meliputi beberapa tahapan. Pada tahap awal, pembersihan teks dilakukan untuk mengeliminasi karakter non-alfabetis seperti simbol, emotikon, dan angka. Selanjutnya, proses case folding diaplikasikan untuk menyeragamkan semua teks menjadi huruf kecil. Proses berikutnya adalah tokenisasi, yang membagi teks menjadi kata-kata individu berdasarkan spasi. Kemudian, kata-kata yang tidak memberikan informasi signifikan, atau stop words, dihilangkan. Tahap akhir adalah lemmatisasi, di mana kata-kata dikembalikan ke bentuk dasar mereka, memfasilitasi analisis teks yang lebih konsisten dan umum.

2.3. Ekstraksi Fitur TF-IDF

Setelah proses pra-pemrosesan selesai, langkah selanjutnya adalah tahap pembobotan atau ekstraksi fitur. Dalam analisis teks, konversi kata menjadi format numerik adalah esensial karena komputer hanya mampu mengolah data numerik. Teknik TF-IDF (Term Frequency-Inverse Document Frequency) merupakan salah satu metode ekstraksi fitur yang sering digunakan. TF-IDF berfungsi untuk mengevaluasi seberapa signifikan sebuah kata dalam dokumen atau keseluruhan korpus. Metode ini memanfaatkan kata-kata yang sudah diproses pada tahap pra-pemrosesan sebagai masukan, dan proses ekstraksi fitur TF-IDF ini diilustrasikan pada gambar 4.



Gambar 4. Alur Ekstraksi Fitur TF-IDF

Berdasarkan gambar 4, proses ekstraksi fitur menggunakan metode TF-IDF melibatkan dua tahapan, yaitu term frequency (TF) dan Inverse Document Frequency (IDF). Dengan menggunakan TF-IDF, nilai penting atau bobot dari setiap kata dalam dokumen dapat ditentukan berdasarkan seberapa sering kata tersebut muncul dalam dokumen tersebut dan di seluruh korpus secara keseluruhan [9].

2.3.1 Term Frequency (TF)

Tahap awal ini bertujuan untuk menghitung frekuensi kemunculan setiap kata (term) dalam setiap dokumen. Frekuensi ini diukur dengan membagi jumlah kemunculan kata tersebut dengan jumlah total kata dalam dokumen tersebut. Persamaan matematis untuk menghitung TF ini dapat dirumuskan sebagai berikut (1) [10]:

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \frac{\text{Jumlah kemunculan kata (t) dalam dokumen (d)}}{\text{Total jumlah kata dalam dokumen (d)}} \quad (1)$$

Dimana TF(d,t) adalah term frequency, t adalah Kata yang sedang dievaluasi, d adalah Dokumen tempat kata t muncul.

2.3.2 Inverse Document Frequency (IDF)

Tahap kedua ini bertujuan untuk mengukur seberapa penting atau unik sebuah kata terhadap seluruh korpus atau kumpulan dokumen. Ini membantu mengurangi bobot kata yang muncul secara umum di seluruh dokumen. Penulisan matematis dari penghitungan TF ini diberikan dalam persamaan (2)[10].

$$\mathrm{IDF(t, D)} = \log\left(\frac{\text{Total jumlah dokumen dalam korpus (D)}}{\text{Jumlah dokumen yang mengandung kata (t)}}\right) \quad (2)$$

Kemudian setelah mendapat kan sail IDF dengan persamaan (2) selanjutnya adalah mendapat kan jumlah keseluruhan TF-IDF dengan persamaan yang terdapat dalam persamaan (3) [10].

$$\mathrm{TF - IDF(t, d, D)} = \mathrm{TF(t, d)} \times \mathrm{IDF(t, D)} \quad (3)$$

2.3.3. Seleksi Fitur Chi-Squared

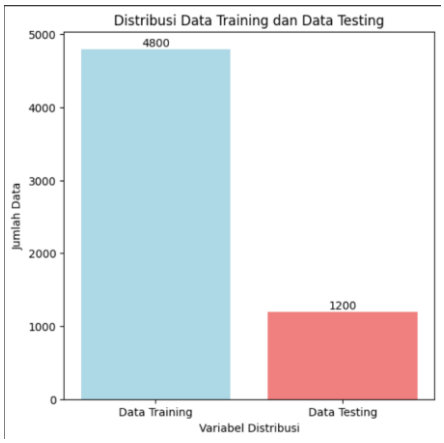
Metode seleksi fitur Chi-Squared memungkinkan kita untuk menentukan fitur mana yang memiliki keterkaitan paling kuat dengan variabel target dalam dataset. Proses ini melibatkan perhitungan seberapa besar ketergantungan antara fitur dan kelas yang ingin diprediksi. Khususnya dalam analisis sentimen atau klasifikasi teks, teknik ini efektif untuk menyoroti fitur yang memiliki pengaruh signifikan terhadap prediksi klasifikasi. Model matematis untuk seleksi fitur Chi-Squared dapat diringkas seperti persamaan (4)[11].

$$\mathrm{x}^2 = \sum \frac{(\mathrm{o}-\mathrm{E})^2}{\mathrm{E}} \quad (4)$$

Dimana chi^2 adalah nilai statistik Chi-Squared, O adalah frekuensi observasi, dan E adalah frekuensi yang diharapkan. Nilai (\chi^2) yang tinggi menunjukkan bahwa fitur tersebut memiliki ketergantungan yang signifikan dengan kelas target dan oleh karena itu penting untuk model klasifikasi.

2.3.4. Pemisahan Data

Langkah berikutnya adalah memisahkan data menjadi data latih dan data uji. Dalam penelitian ini, digunakan 80% data sebagai data latih dan 20% data sebagai data uji dari total 6000 record.



Gambar 5. Proporsi Data Latih dan Data Uji

Berdasarkan ilustrasi pada Gambar 5, terdapat 4800 data yang ditetapkan sebagai data pelatihan, sementara 1200 data lainnya diidentifikasi sebagai data uji. Proses pembagian antara data pelatihan dan data uji dilaksanakan secara acak dengan tujuan mempertahankan proporsi yang seimbang antara kelas-kelas yang terlibat.

2.3.5. Klasifikasi Multinomial Naive Bayes

Metode Naive Bayes Multinomial merupakan algoritma yang bergantung pada prinsip teorema Bayes dan umumnya dipakai dalam penugasan klasifikasi teks. Algoritma ini mengoperasikan asumsi bahwa setiap fitur dalam dokumen, seperti kata-kata, adalah independen satu sama lain. Hal ini memungkinkan perhitungan probabilitas kelas berdasarkan frekuensi kata menjadi lebih

sederhana. Penghitung probabilitas sebuah dokumen d terhadap kelas C yang ditunjukkan pada persamaan (5)[12].

$$P(C) = \frac{N_C}{N} \quad (5)$$

Dimana N_c adalah jumlah kelas C pada seluruh dokumen dan N adalah jumlah seluruh dokumen. Untuk probabilitas dari kata ke-n ditentukan dengan menggunakan persamaan (6)[12]t:

$$P(X_n|C) = \frac{N_{x_n,c}+\alpha}{N(C)+V} \quad (6)$$

Dimana $N_{x_n,c}$ mencerminkan jumlah kemunculan term X_n dalam seluruh data pelatihan pada kelas C dan $N(C)$ menyatakan total kemunculan term dalam seluruh data pelatihan pada kelas C, dan α merupakan parameter laplace smoothing, V adalah jumlah total kata pada data untuk melatih model. Sementara rumus Multinomial yang digunakan dalam pembobotan TF-IDF adalah sebagai berikut:

$$P(X_n|C) = \frac{\sum \quad tf(X_n,d\in C)+\alpha}{\sum \quad N_{d\in C}+V} \quad (7)$$

Dimana $\sum \quad tf(X_n, d \in C)$ adalah jumlah pembobotan kata $X_n$dari seluruh data training pada kelas C dan $\sum \quad N_{d\in C}$ adalah jumlah bobot seluruh term pada data training pada kelas C.

2.3.6. Pengujian dan Evaluasi

Evaluasi dilakukan untuk menilai kinerja model yang telah dikembangkan. Kinerja ini bisa diukur menggunakan tabel confusion matrix, yang merupakan alat visual yang sering digunakan untuk menampilkan performa model klasifikasi pada sejumlah data uji yang memiliki nilai sebenarnya yang diketahui. Kinerja model dapat dihitung dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score dengan rumus-rumus yang sesuai seperti di bawah ini [13].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$F1-\text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (11)$$

3. Hasil dan Diskusi

3.1. Hasil Preprocessing

Hasil dari seluruh proses preprocessing pada dataset disajikan dalam bentuk tabel yang dapat ditemukan pada Gambar di bawah ini. Berikut adalah contoh beberapa kalimat yang diambil dari dataset.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Hasil Preprocessing

```
df.head()
```

	kategori	berita	Hasil
0	teknologi	Uber pada hari Jumat mengatakan akan menguak d...	uber jumat menguak data perjalanan paris publi...
1	teknologi	Menyusul jejak NES Classic Edition , SNES Clas...	menyusul jejak ne classic edition snes classic...
2	teknologi	MDI Ventures , perusahaan modal ventura yang d...	mdi venture perusahaan modal ventura didukung ...
3	teknologi	Mazda masih menutup rapat informasi soal sport...	mazda menutup rapat informasi sportscar anyar ...
4	teknologi	Sampai di akhir tahun 2017 ini , frasa " print...	tahun frasa printer mencetak video terdengar m...

Gambar 6. Hasil Preprocessing

Gambar 6 menunjukan contoh kalimat yang belum dilakukan preprocessing dan hasil akhirnya contohnya pada berita berindex 1 dari kategori berita yang berisi kalimat "Menyusul jejak NES Classic Edition," setelah dilakukan preprocessing menjadi "menyusul jejak ne classic edition snes classic"

3.2. Hasil Pembobotan TF-IDF

Untuk mengonversi teks berita menjadi bentuk yang dapat diproses oleh model klasifikasi, peneliti menggunakan teknik ekstraksi fitur TF-IDF. Teknik ini memungkinkan peneliti untuk mengukur pentingnya setiap kata dalam dokumen berdasarkan frekuensi kemunculannya dalam dokumen tersebut dan dalam keseluruhan korpus.

Hasil Setelah diberi Bobot Nilai dan diubah ke bentuk matriks

```
[ ] X_tf_idf = tf_idf.transform(x).toarray()
    X_tf_idf

    array([[0., 0., 0., ..., 0., 0., 0.],
           [0., 0., 0., ..., 0., 0., 0.],
           [0., 0., 0., ..., 0., 0., 0.],
           ...,
           [0., 0., 0., ..., 0., 0., 0.],
           [0., 0., 0., ..., 0., 0., 0.],
           [0., 0., 0., ..., 0., 0., 0.]])
```

Gambar 7. Hasil Pembobotan TF-IDF

3.3. Hasil Seleksi Fitur dengan Chi-Squared

Untuk meningkatkan kinerja model klasifikasi, dilakukan seleksi fitur dengan metode chi-squared untuk memilih fitur-fitur yang paling relevan dari dataset yang sudah dilakukan preprocessing. Peneliti melakukan evaluasi terhadap berbagai nilai k, yang mewakili jumlah fitur yang akan dipilih dari dataset. Hasil evaluasi menunjukkan bahwa menggunakan nilai k=10000 dari jumlah asli fitur data menghasilkan akurasi yang lebih tinggi dibandingkan dengan menggunakan nilai k=5000, k = 1000, k = 2000.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

# 5000 dengan highest chi-squared statistics are selected
chi2_features = SelectKBest(chi2, k=10000)
X_kbest_features = chi2_features.fit_transform(X, y)

# Reduksi features
print('Angka Original Feature :', X.shape[1])
print('Angka Reduksi Feature :', X_kbest_features.shape[1])
```

```
Angka Original Feature : 54091
Angka Reduksi Feature : 10000
```

Gambar 8. Hasil Seleksi Fitur

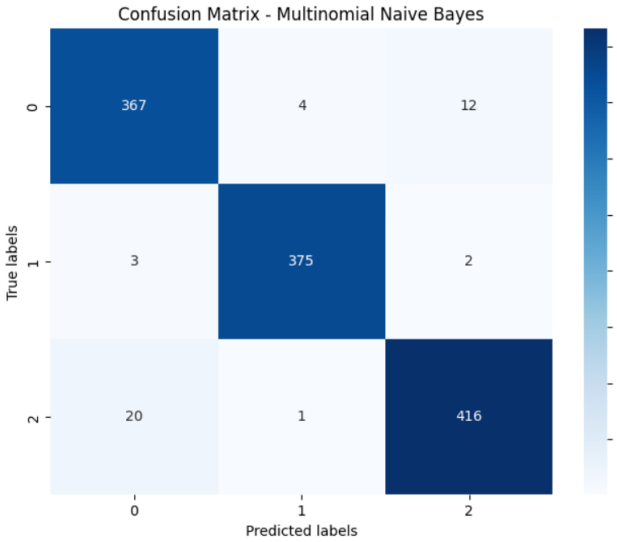
Berdasarkan gambar 8 menunjukkan bahwa peneliti mempertahankan jumlah fitur sebanyak nilai k yaitu k=10000 dari jumlah asli fitur data yaitu sebanyak 54091. Ini menunjukkan bahwa mempertahankan lebih banyak fitur memberikan informasi tambahan yang diperlukan bagi model untuk membedakan antara kelas-kelas target dengan lebih baik. Seleksi fitur menggunakan nilai k=10000 telah meningkatkan kinerja model dengan meningkatkan akurasi klasifikasi. Dengan mempertahankan lebih banyak fitur, dapat dipastikan bahwa model memiliki akses ke informasi yang lebih banyak dan relevan untuk melakukan klasifikasi dengan lebih baik.

3.4. Evaluasi Model Multinomial Naïve Bayes

Model Multinomial Naive Bayes dibangun dengan menggunakan bantuan library sklearn dalam bahasa pemrograman Python. Setelah itu model akan dilatih dengan data training yang telah dilakukan ekstraksi fitur TF-IDF dan seleksi fitur dengan Chi-squared untuk kemudian dilakukan evaluasi menggunakan data testing. Pelatihan Model dan Prediksi Model Multinomial Naïve Bayes dilatih menggunakan data latih yang telah diseleksi fitur-fiturnya. Hasil akurasi validasi silang menunjukkan bahwa model memiliki kinerja yang stabil dan konsisten dalam mengklasifikasikan data seperti dibuktikan pada gambar 9 dan 10.

```
Algorithm: Multinomial Naive Bay
Confusion Matrix:
 [[367   4  12]
 [  3 375   2]
 [ 20   1 416]]
Accuracy:  0.965
Precision:  0.965
Recall Score:  0.965
F1 Score:  0.965
```

Gambar 9. Hasil Akurasi Validasi silang



Gambar 10. Confusion Matrix Validasi silang

Berdasarkan Gambar 9 waktu yang diperlukan untuk melatih model adalah 0.337 detik, sedangkan waktu yang dibutuhkan untuk melakukan prediksi terhadap data uji adalah 0.057 detik. Dilakukan validasi silang menggunakan metode K-Fold Cross Validation dengan 5 lipatan. Hasil validasi silang menunjukkan akurasi rata-rata sebesar 97.562% dengan deviasi standar sebesar 0.252%. Selanjutnya evaluasi dilakukan menggunakan Confusion Matrix untuk mengukur akurasi yang diperoleh dari metode yang digunakan. Pada pemodelan menggunakan Multinomial Naïve Bayes, diperoleh akurasi yang tinggi pada tahap pelatihan dan pengujian. Informasi evaluasi yang lebih rinci disajikan dalam Tabel 1.

Tabel 1. Hasil Evaluasi

	Precision	**Recall**	**F-1 Score**	**Akurasi**
Teknologi	95.19%	95.19%	95.19%	95.19%
Olahraga	98.68%	98.68%	98.68%	98.68%
Hiburan	95.82%	95.82%	95.82%	95,9%
Average	96.57%	96.57%	96.57%	96.57%

4. Kesimpulan

Berdasarkan penelitian yang sudah dilakukan, dapat disimpulkan bahwa metode Multinomial Naïve Bayes berhasil mengklasifikasikan berita berdasarkan kategori dengan tingkat akurasi tinggi. Dataset teks berita yang dikumpulkan berasal dari repository GitHub dan diolah dengan beberapa tahap yaitu tahap preprocessing, pembobotan TF-IDF, dan seleksi fitur chi-squared. Model klasifikasi yang dikembangkan menunjukkan akurasi yang tinggi dengan nilai sebesar 97.562% dalam validasi silang. Hasil evaluasi menunjukkan bahwa pendekatan ini efektif dalam memproses data teks dan menghasilkan prediksi yang akurat, menunjukkan potensi aplikatifnya dalam pengelolaan berita secara otomatis. Dengan demikian, penelitian ini memberikan kontribusi dalam pengembangan sistem yang dapat membantu dalam pengelolaan dan analisis berita secara efisien.

Daftar Pustaka

[1] D. N. Chandra, G. Indrawan, dan I. N. Sukajaya, “Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram,” Jurnal Ilmiah Teknologi dan Informasi Asia (JITIKA), vol. 10, hlm. 11, 2016.

[2] W. F. Mahmudy dan A. W. Widodo, “Klasifikasi Artikel Berita Secara Otomatis Menggunakan Metode Naive Bayes Classifier Yang Dimodifikasi,” TEKNO, vol. 21, Mar 2014.

[3] A. F. Hidayatullah dkk., “Penerapan Text Mining dalam Klasifikasi Judul Skripsi,” 2016.

[4] S. Kumar, A. Sharma, B. K. Reddy, S. Sachan, V. Jain, dan J. Singh, “An intelligent model based on integrated inverse document frequency and multinomial Naive Bayes for current affairs news categorisation. ,” International Journal of System Assurance Engineering and Management, vol. 13, hlm. 1–15, Nov 2021.

[5] A. Sabrani, I. W. Gede Putu Wirarama Wedashwara, dan F. Bimantoro, “Metode Multinomial Naïve Bayes Untuk Klasifikasi Artikel Online Tentang Gempa Di Indonesia (Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia).” [Daring]. Tersedia pada: http://jtika.if.unram.ac.id/index.php/JTIKA/

[6] S. K. Dirjen dkk., “Terakreditasi SINTA Peringkat 2 Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifier Dengan Seleksi Fitur Dan Boosting,” masa berlaku mulai, vol. 1, no. 3, hlm. 227–232, 2017.

[7] irwan Budiman, R. F. M, dan D. T. Nugrahadi, “Studi Ekstraksi Fitur Berbasis Vektor Word2vec Pada Pembentukan Fitur Berdimensi Rendah,” Jurnal Komputasi, vol. 8, 2020.

[8] E. Indrayuni, “Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes,” Jurnal Khatulistiwa Informatika, vol. 7, no. 1, hlm. 29–36, 2019.

[9] P. M. Prihatini, “Implementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia,” Jurnal Manajemen Teknologi dan Informatika , vol. 6, no. 3, 2016.
[10] N. Komang dkk., “Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia,” MERPATI, vol. 6, no. 2, 2018.
[11] S. Goswami, “Using the Chi-Squared test for feature selection with implementation,” Nov 2020.
[12] B. Harjito, K. N. Aini, dan B. Murtiyasa, “Klasifikasi Dokumen berkonten Serangan jaringan menggunakan Multinomial Naive Bayes,” Seminar Nasional Teknologi Informasi dan Komunikasi (SEMNASTIK), vol. 1, no. 1, hlm. 112–118, 2018.
[13] E. Mas’udah, E. Wahyuni, dan A. Anjani, “Analisis sentimen: Pemindahan ibu kota Indonesia pada twitter,” Jurnal Informatika dan Sistem Informasi, vol. 1, no. 2, hlm. 397–401, 2020.