Pendekatan Macine Learning dalam Evaluasi Label Berita Berdasarkan Judul: Studi Kasus Media Online

Rezky Yuranda[1]\*, Tata Sutabri[2], Delpiah Wahyuningsih[3]

Program Studi Magister Teknik Informatika Universitas Bina Darma[1], [2],

Program Studi Teknik Informatika ISB Atma Luhur [3]

Palembang[1][2], Indonesia

Pangkalpinang[3], Indonesia

[yurandarezky@gmail.com](mailto:yurandarezky@gmail.com)[1], [tata.sutabri@gmail.com](mailto:tata.sutabri@gmail.com)[2], delphibabel@atmaluhur.ac.id[3]

***Abstract*—** **In the current digital era, the availability of information is abundant. News serves as the primary source of up-to-date and reliable information for the public. However, with the increasing volume of information, a reliable evaluation method is needed to ensure accurate and trustworthy news labeling. This research utilizes a machine learning approach to evaluate news labels based on their headlines. The dataset used is obtained from Jakarta AI Research Space and covers various news topics. Three commonly used classification algorithms, namely Naive Bayes, SVM, and Random Forest, are employed in this study. Evaluation is conducted using accuracy, precision, recall, and F1-Score metrics. The results of this research indicate that the SVM algorithm performs the best, achieving an accuracy rate of *92.40%*, followed by Random Forest with an accuracy rate of *89.95%,* and Naive Bayes with an accuracy rate of *88.01%*. These findings provide insights into the effectiveness of machine learning approaches in news label evaluation. Further research is recommended to involve more classification algorithms and broader datasets to enhance the understanding of news label evaluation.**

***Keywords: News label evaluation, machine learning, Naive Bayes, SVM, Random Forest***

***Abstrak*—Dalam era digital saat ini, ketersediaan informasi sangat melimpah. Berita menjadi sumber informasi terkini dan terpercaya bagi masyarakat. Namun, dengan informasi yang semakin banyak, diperlukan sebuah metode evaluasi yang baik untuk memastikan label berita yang akurat dan dapat diandalkan. Penelitian ini menggunakan pendekatan machine learning untuk mengevaluasi label berita berdasarkan judul. Dataset yang digunakan berasal dari Jakarta AI Research Space dan mencakup berbagai topik berita. Tiga algoritma klasifikasi umum, yaitu *Naive Bayes*, *SVM*, dan *Random Forest* digunakan dalam penelitian ini. Evaluasi dilakukan dengan metrik *akurasi*, *presisi*, *recall*, dan *F1-Score*. Hasil penelitian ini menunjukan bahwa algoritma *SVM* memiliki kinerja terbaik, dengan tingkat akurasi *92.40%*, diikuti oleh *Random Forest* dengan tingkat akurasi *89.95%* dan Naive Bayes dengan tingkat akurasi *88.01%.* Temuan ini memberikan wawasan tentang efektivitas pendekatan machine learning dalam evaluasi label berita. Penelitian lebih lanjut dianjurkan untuk melibatkan lebih banyak algoritma klasifikasi dan dataset yang lebih luas guna meningkatkan pemahaman tentang evaluasi label berita**

***Kata Kunci: Evaluasi label berita, machine learning, naive bayes, svm, random forest***

# Pendahuluan

Dalam era digital saat ini, ketersediaan informasi sangat melimpah. Berita menjadi bagian penting dari informasi tersebut, karena berita menjadi sumber informasi terkini dan terpercaya bagi masyarakat [1].

Bersamaan dengan pertumbuhan teknologi saat ini, terjadi pula peningkatan jumlah informasi tekstual yang tersedia di dalamnya. Data dari *Indonesia Digital Association* (IDA) menunjukan bahwa *96%* masyarakat Indonesia mengkonsumsi berita secara online [2].

Pada sebuah berita, judul utama atau headline merupakan elemen penting dari setiap halaman web. Judul tersebut menentukan apakah pembaca akan tetap berinteraksi dengan konten Anda atau pergi ke halaman lain. Begitu juga dengan kategori atau label yang membantu pembaca untuk menemukan konten Anda berdasarkan topik dan memberikan konteks pada artikel yang Anda baca [3]. Oleh sebab itu, maka penting untuk memiliki metode evaluasi yang baik dalam mengelola label berita untuk memastikan informasi yang disajikan akurat dan dapat diandalkan bagi konsumen [2].

Dalam hal ini, penggunaan dataset berkualitas dan repesentatif sangat penting dalam melakukan penelitian dan pengembangan algoritma *AI* untuk evaluas label berita. Salah satu dataset yang digunakan dalam penelitian ini berasal dari *Jakarta AI Research Spaces*. *Jakarta AI Research Space* adalah lembaga riset *AI* terkemuka yang berfokus pada pengembangan teknologi *AI* di Indonesia.

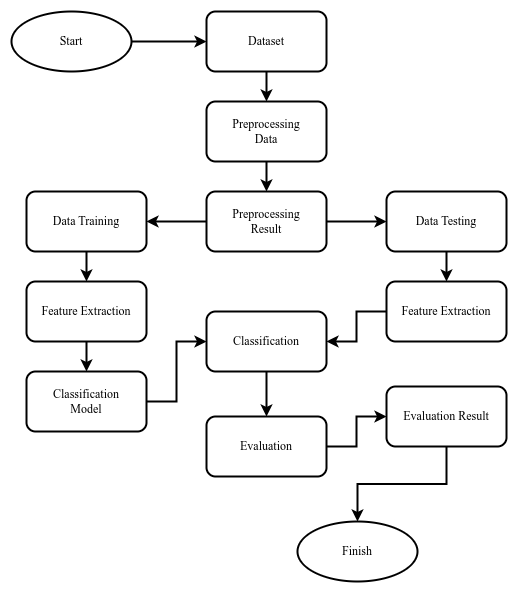
Dataset yang disediakan oleh *Jakarta AI Reserach Space* menawarkan akses ke beberapa media berita online yang mencakut topik yang beragam. Kumpulan data ini mencakup judul-judul berita yang dapat dijadikan dasar untuk melakuka analisis dan evaluasi label dari berita dengan menggunakan pendekatan machine learing. Dalam penelitian ini, kami akan memanfaatkan dataset tersebut untuk melihat sejauh mana pendekatan machine learning dapat memberikan hasil yang akurat dalam mengevaluasi label berita berdasarkan judul.

Pada penelitian ini, digunakan beberapa algoritma klasifikasi yang umum digunakan, seperti *Naive Bayes*, *Support Vector Machines* (*SVM*), dan *Random Forest*, untuk membangun model klasifikasi label berita berdasarkan judul. Algoritma-algoritma ini memiliki keunggukan masing-masing dalam memahami pola dan hubungan dalam data judul berita, yang kemudian akan digunakan untuk mengklasifikasi berita menjadi berbagai label atau kategori.

Dalam eksperimen ini akan dibandingkan performa ketiga algoritma klasifikas tersebut berdasarkan berbagai metrik evaluasi, seperti *akurasi*, *presisi*, *recall*, dan *F1-Score*. Hasil penelitian ini diharapkan dapat memberikan pemahaman lebih baik tentang efektivitas pendekatan machine learning dalam proses evaluasi label berita berdasarkan judul.

# Metodelogi Penelitian

Metodelogi penelitian untuk pendekatan machine learning dalam evaluasi label berita berdasarkan judul, ditampilkan pada gambar 1.

1. Alur penelitian

## Dataset

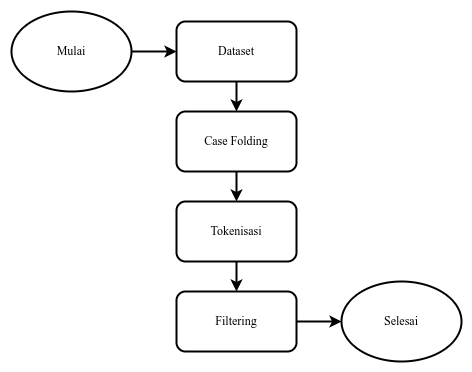
Langkah pertama mempersiapkan dataset yang akan digunakan untuk pelatihan dan pengujian. Dataset yang digunakan pada penelitian ini diambil dari *Jakarta AI Reserach Space* yang berisikan informasi judul berita beserta label atau kategorinya sebanyak 8754 data. Pada dataset ini, terdapat didefinisikan 5 label, yaitu bola, news, bisnis, tekno, dan otomotif. Untuk distribusi data per label dapat dilihat pada table 1.

tabel 1. Distribusi data setiap label

|  |  |  |
| --- | --- | --- |
| No | ***Label*** | ***Jumlah*** |
| 1 | Bola | 2986 |
| 2 | News | 2897 |
| 3 | Bisnis | 1881 |
| 4 | Tekno | 790 |
| 5 | Otomotif | 200 |
| **Jumlah Total** | | **8754** |

## Preprocessing Data

Tahapan preprocessing yang sangat penting pada data mining, alasan utamanya adalah karena kualitas dari input data sangatmempengaruhi kualitas output analisis yang dihasilkan [4]. Tahapan Preprocessing data dapat dilihat pada gambar 2.

1. Preprocessing data
2. *Case folding*

Pada tahap ini, semua huruf yang ada pada judul berita diubah menjadi huruf kecil. Tujuan dari *case folding* adalah untuk menghilangkan variasi bentuk kata yang disebabkan oleh perbedaan kapitalisasi. Misalnya, kata “Data” dan “data” dapat dianggap sama setelah melewati proses *case folding*. Hal ini dapat membantu dalam konsistensi dan keseragaman dalam proses analisa data.

1. *Tokenisasi*

Tokenisasi adalah proses untuk memecah texts menjadi unit-unit kecil disebut token. Token bisa berupa kata, frasa, atau simbol tertentu tergantung pada konteks text yang diproses. Misalnya, kalimat “Kasus dugaan penganiayaan terhadap remaja” dapat di tokenisasi menjadi “[‘Kasus’, ‘dugaan’, ‘penganiayaan’, ‘terhadap’, ‘remaja’]”. Tokenisasi sangat penting di lakukan dalam preprocessing data [5]

1. *Filtering*

Pada proses *filtering*, karakter lain selain huruf, seperti tanda baca, simbol dan nomor akan dihapus dari teks judul aslinya [6]. Pada tahap ini juga dilakukan *filtering stop words*, yang dimana *stop words* adalah kata-kata umum yang sering muncul dalam teks dan cendrung tidak memberikan kontribusi sinifikan dalam pemrosesan teks. Contoh *stop words* pada bahasa indonesia adalah “di”, “ke”, “yang”, “dari”, dll. Proses ini akan membantu dalam mengurangi dimensi data dan meningkatkan efisiensi pemrosesan.

## Feature extraction

Pada tahap *feature extraction*, dilakukan extraksi fitur dar teks judul berita menggunakan metode *TF-IDF* (*Term Frequency-Inverse Document Frequency*). Metode ini digunakan untuk mengukur tingkat pentingnya sebuah kata dalam suatu dokumen atau dataset berdasarkan frekuensi kemunculan kata dalam dokumen tersebut [6]. Kata-kata terpilih dianggap lebih informatif dan efektif untuk dieksekusi pada model classifier [7][8].

Penerapan teknik *TF-IDF* dilakukan setelah *preprocessing* data dan dilakukan sebelum proses klasifikasi data. Secara lebih detail, *TF-IDF* menghitung skor atau bobot untuk setiap kata dalam dokumen dengan mempertimbangkan dua faktor yaitu *Term Frequency* (*TF*) dan *Inverse Document Frequency* (*IDF*) [9].

(1)

Pada persamaan (1), *fk* adalah jumlah kemunculan kata dalam suatu kalimat dan *ft* adalah total kata dalam suatu dokumen. *IDF* menggambarkan seberapa umum atau jarang kata tersebut muncul dalam seluruh dataset. *Inverse Document Frequency* dihitung dengan menggunakan rumus

(2)

Dimana *n* pada konteks ini merujuk pada total jumlah dokumen pada suatu korpus, sementara *df* adalah singkatan dari jumlah dokumen yang mengandung kata tertentu pada suatu korpus.

## Classification model

Dalam metode pendekatan machine learing dalam evaluasi label berita berdasarkan judul, kami menggunakan tiga algoritma klasifikasi yang berbeda yaitu *Naive Bayes*, *Support Vector Machine* (*SVM*), dan *Random Forest*.

*Naive Bayes* merupakan salah satu algoritma klasifikasi yang berdasarkan pada *Teorema Bayes* [10]. Algoritma ini memprediksi label berita berdarkan probalitas kelas yang diestimasikan dari fitur-fitur yang ada.

*SVM* (*Support Vector Machines*) adalah algoritma klasifikasi yang bekerja dengan mencari *hyperplane* terbaik yang memisahkan dua kelas berita berdasarkan fitur-fitur nya. *SVM* memiliki kemampuan untuk menangani data yang kompleks dan memiliki kelebihan dalam menangani dataset yang memiliki dimensi yang tinggi [11].

*Random Forest* merupakan algoritma klasifikasi ensemble yang membangun sejumlah pohon keputusan secara acak dan mengkombinasikan hasil prediksi dari setiap pohon untuk menghasilkan prediksi akhir. Dengan cara ini, *Random Forest* dapat mengatasi overfitting pada dataset dan menghasilkan hasil prediksi yang lebih stabil [11].

Dengan menggunakan kombinasi dari *Naive Bayes*, *SVM*, dan *Random Forest*, kami dapat membandingkan dan melakukan evaluasi performa dari ketiga algoritma ini dalam memprediksi dan mengklasifikasi label berita berdasarkan fitur-fitur yang diekstraksi.

## Classification

Pada tahap ini, model klasifikasi yang telah di latih akan digunakan untuk memprediksi label berita berdasarkan fitur-fitur yang telah di ekstraksi sebelumnya. Proses ini melibatkan penggunaan rumus atau algoritma tertentu yang dapat memberikan prediksi yang akurat.

Pada algoritma *Naive Bayes*, prediksi label berita dilakukan dengan menggunakan *Teorema Bayes*. Untuk setiap kategori *k,* nilai probalitas posterior dapat dihitung dengan persamaan 3.

(3)

Dimana P(*k∣x*) adalah probabilitas kondisional dari *k* terhadap *x*, P(*x∣k*) adalah probabilitas kondisional dari *x* terhadap *k*, P(*k*) adalah probabilitas prior dari *k*, dan P(*x*) adalah probabilitas margin dari *x*.

Pada *SVM*, prediksi label berita dilakukan dengan mencari hyperplane terbaik yang memisahkan dua kelas berdasarkan fitur-fitur yang ada. Rumus untuk memprediksi label pada SVM terdapat pada persamaan 4.

(4)

Dimana *x* adalah input sample yang ingin di prediksi, *xi* adalah training sample ke-*i*, *yi* adalah class label untuk training sample ke-*i* (mengambil nilai -1 atau 1), α*i*​ adalah koefisien untuk traiing sample ke-*i* yang dihitung oleh algoritma SVM selama proses training, K(*xi​,x*) adalah kernel function yang mengukur kesamaan anatar input sample *x* dan training sample ke-*i*, *b* adalah bias term yang juga dihitung oleh algoritma *SVM* selama proses training, dan *n* adalah jumlah total training sample.

Dalam *Random Forest*, prediksi label berita dilakukan dengan menggabungkan hasil prediksi dari banyak pohon keputusan yang terbentuk. Setiap pohon memberikan suara dalam memprediksi label, dan label yang paling banyak terpilih menjadi prediksi akhir. Rumus yang digunakan dalam prediksi Random Forest tidak memiliki matematis khusus, tetapi melibatkan agregasi hasil prediksi dari setiap cabang. .

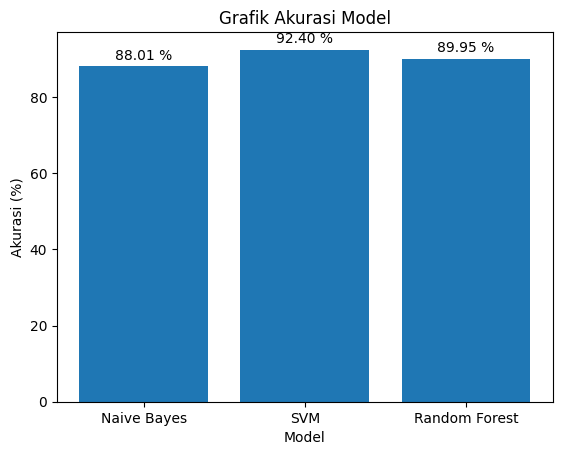
## Evaluasi

Pada tahap ini, dilakukan pengukuran performa model klasifikasi yang telah dibangun menggunakan metrik-metrik seperti *akurasi*, *presisi*, *recall*, dan *F1-Score*.

*Akurasi* adalah metrik yang mengukur sejauh mana model mampu memberikan prediksi yang benar secara keseluruhan. *Presisi* mengukur sejauh mana prediksi positif yang diberikan oleh model adalah benar, sementara *recall* mengukur sejauh mana model mampu mendeteksi semua kasus positif yang sebenarnya. *F1-Score* adalah metrik yang menggabungkan presisi dan recal untuk memberikan gambaran yang lebih holistik tentang performa model.

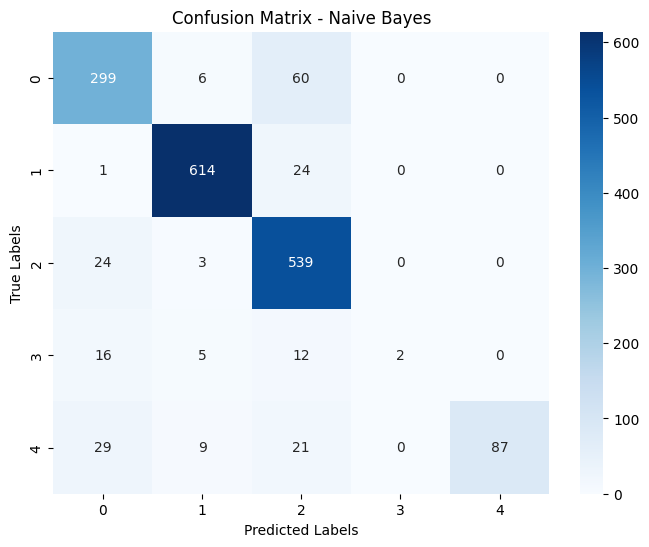
Dalam tahapan evaluasi, kita menganalisis hasil prediksi model klasifikasi dengan membandingkan prediksi dengan label sebenarnya. Dengan menghitung *akurasi*, *presisi*, *recal* dan *F1-score*, kita dapat memperoleh pemahaman yang lebih mendalam tentang kualitas model klasifikasi dalam tugas evaluasi label berita berdasarkan judul. Metrik-metrik ini akan membantu kita dalam hal mengevaluasi tingkat keberhasilan model dalam memprediksi label berita dengan akurasi tinggi, dan juga membantu kita memahami seberapa baik model dapat menghindari kesalahan prediksi yang mungkin terjadi [12].

# Hasil dan pembahasan

Proses training menggunakan 3 Algoritma klasifikasi yaitu *Naive Bayes*, *SVM*, dan *Random Forest*. Berdasarkan hasil training terhadap 3 Algoritma tersebut, penelitian ini mendapatkan hasil akurasi seperti gambar 3.

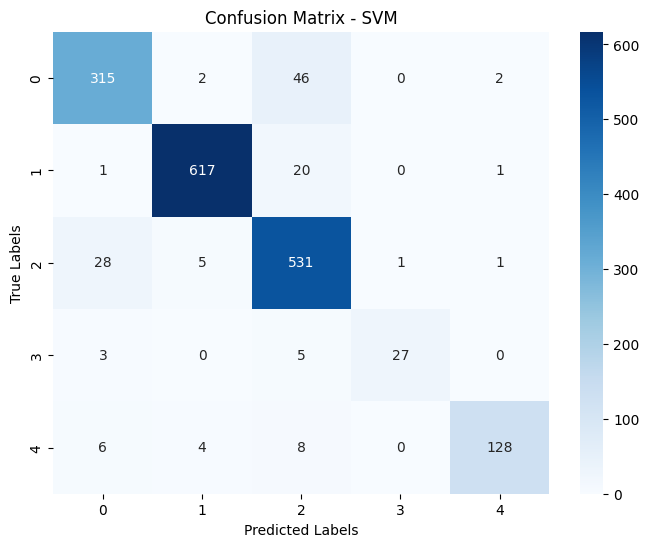
1. Grafik akurasi

Berdasarkan hasil analisa pada gambar 3, dapat diketahui bahwa *SVM* menghasilkan akurasi tertinggi dengan nilai *92,40%*, diikuti oleh *Random forest* dengan akuarsi *89,95%* dan *Naive Bayes* dengan akurasi *88,01%.* Hasil ini menunjukan bahwa *SVM* adalah alogritma yang paling efektif dalam hal akurasi untuk memprediksi label berita berdasarkan judul. *SVM* mampu membangun model klasifikasi yang dapat memberikan prediksi yang lebih akurat dengan memanfaatkan hiperplane yang optimal untuk memisahkan kelas-kelas yang berbeda.

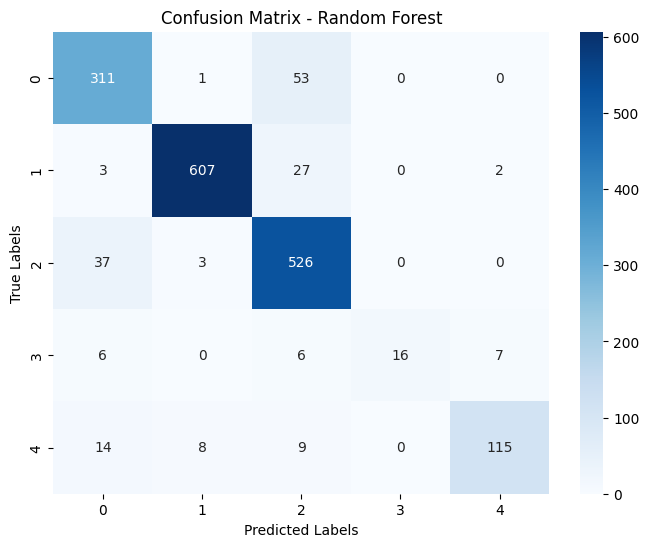


1. *Confusion matrix Naive Bayes*

Hasil *confusion matrix* model *Naive Bayes* pada gambar 4 menggunakan sebanyak 8753 judul artikel berita terhadap setiap label berita. Label berita yang di prediksi terdiri dari bola, news, bisnis, tekno, dan otomotif yang diberi label 0-4 secara berurutan. Label 0 sebagai kategori bola menghasilkan nilai prediksi sebanyak 299, untul label 1 dengan kategori news menghasilkan nilai prediksi 614, untuk label 2 dengan kategori bisnis menghasilkan nilai prediksi 539, untuk label 3 dengan kategori tekno mengasilkan nilai prediksi 2, dan untuk label 4 dengan kategori dengan nilai prediksi 87.

1. C*onfusion matrix SVM*

Pada gambar 5, hasil *confusion matrix* pada *SVM* pada label 0 dengan kategori bola menghasilkan nilai prediksi 315, untul label 1 dengan kategori news menghasilkan hasil prediksi 617, untuk label 2 dengan kategori bisnis menghasilkan hasil prediksi 531, untuk label 3 dengan kategori tekno mengasilkan nilai prediksi 27 dan untuk label 4 dengan kategori otomotif menghasilkan nilai prediksi 128.

1. *Confusion matrix Random Forest*

Pada gambar 6, hasil *confusion matrix* pada *Random Forest* pada label 0 dengan kategori bola menghasilkan nilai prediksi 311, untul label 1 dengan kategori news menghasilkan hasil prediksi 607, untuk label 2 dengan kategori bisnis menghasilkan hasil prediksi 526, untuk label 3 dengan kategori tekno mengasilkan nilai prediksi 16 dan untuk label 4 dengan kategori otomotif menghasilkan nilai prediksi 115

Tabel 2. Rangkuman perhitungan rasio

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | ***Model*** | ***Presisi*** | ***Recall*** | ***F1-score*** |
| 1 | Naive Bayes | 0.889628 | 0.880069 | 0.870617 |
| 2 | SVM | 0.926065 | 0.924043 | 0.924158 |
| 3 | Random Forest | 0.903712 | 0.899486 | 0.898386 |

Berdasarkan table II Rangkuman perhitungan rasio, kita dapat melakukan analisis terhadap performa ketiga model klasifikasi yang di uji, yaitu *Naive Bayes*, *SVM*, dan *Random Forest*

1. *Naive Bayes*

*Naive Bayes* memiliki nilai presisi sebesar 0.889628. Hal ini menunjukan bahwa model *Naive bayes* mampu mengidentifikasi dengan benar sekitar *88.96%* dari total prediksi positif yang dilakukan. Presisi yang tinggi menandakan bahwa model ini cendrung memberikan sedikit *false positif* (prediksi positif yang sebenarnya negatif). Untuk *recall* atau tingkat keberhasilan pengenalan adalah sebesar 0.880069. Hal ini menunjukan bahwa model *Naive Bayes* dapat mengenali sekitar *88.01%* dari keseluruhan kasus positif yang ada dalam dataset. *Recall* yang tinggi menandakan kemampuan model untuk mengidentifikasi sebanyak mungkin kasus positif yang sebenarnya. Untuk *F1-Score* sebesar 0.870617 menggambarkan nilai rata-rata harmonik antara *presisi* dan *recall*. *F1-Score* yang tinggi menunjukan performa yang seimbang antara *presisi* dan *recall*.

1. *SVM*

Pada model *SVM*, menghasilkan *presisi* sebesar 0.926065. Hal ini menunjukan model *SVM* mampu mengidentifikasi dengan benar sebesar *92.61%* dari total prediksi positif yang dilakukan. *Presisi* yang tinggi menandakan kemampuan model untuk memberikan sedikit *false positif*. Untuk hasil *recall*, menghasilkan 0.924043 yang menunjukan bahwa model *SVM* dapat mengenali sekitar *92.40%* dari keseluruhan kasus positif yang ada dalam dataset. *Recall* yang tinggi juga menandakan kemampuan model untuk mengidentifikasi sebanyak mungkin kasus positif yang sebenarnya. Untuk *F1-Score* dengan hasil 0.924158 pada model *SVM* menunjukan keseimbangna yang baik antara *presisi* dan *recall*, menandakan performa yang baik secara keseluruhan.

1. *Random Forest*

Pada *Random Forest*, menghasilkan presisi sebesar 0.903712. Hal ini menunjukan bahwa model *Random Forest* mampu mengidentifikasi dengan benar sekitar *90.37%* dari total prediksi positif yang dilakukan. Untuk *recall* menghasilkan nilai 0.899486 menunjukan bahwa model *Random Forest* dapat mengenali sekitar *89.95%* dari keseluruhan kasus positif yang ada dalam dataset. Dan untuk *F1-Score* mengasilkan nilai sebesar 0.898386 menunjukan keseimbangan yang baik antara *presisi* dan *recall* pada model *Random Forest*.

Berdasarkan analisa ini dapat disimpulkan bahwa ketiga model klasifikasi memiliki performa yang sangat baik dengan nilai *akurasi*, *presisi*, *recall*, dan *F1-Score* yang tinggi. Model *SVM* menunjukan performa yang sedikit lebih unggu dengan *presisi* dan *recall* tertinggi, diikuti *Random Forest* dan *Naive* *Bayes*. Namun, perbandingan kinerja yang lebih komprehensif dapat dilakukan dengna melihat faktor-faktor lain seperti kecepatan komputasi, komplksitas model, dan karakteristik dataset yang lebih mendalam.

# Kesimpulan

Berdasarkan hasil penelitian yang dilakukan terhadap tiga model kelasifikasi, yaitu *Naive Bayes*, *SVM*, dan *Random Forest* dapat ditarik kesimpulan bahwa model *SVM* memiliki performa yang paling baik dalam hal *akurasi*, *presisi*, *recall* dan *F1-Score*. Model *SVM* mampu mengenali dan mengklasifikasikan label berita berdasarkan judul pada dataset dengan akurat. Di sisi lain, model *Naive Bayes* menunjukan performa yang cukup baik dan menjadi pilihan yang baik jika kecepatan komputasi menjadi faktor penting, karena waktu pelatihan model ini relatif lebih cepat. Model *Random Forest* juga menunjukan performa yang baik, dengan presisi, recall, dan *F1-Score* yang tinggi.

Penelitian ini memiliki kelebihan dalam melibatkan tiga model klasifikasi yang berbeda, memberikan perbandingan yang komprehensif dalam hal performa. Metode pemrosesan teks yang baik, termasuk tokenisasi, pengahpusan *stop words*, dan vektorisasi *TF-IDF*, juga digunakan dalam penelitian ini.

Namun penelitian ini memiliki beberapa kekurangan. Keterbatasan pada dataset yang digunakan, seperti jumlah sample, distribusi label yang tidak seimbang dan representasi kelas, sehingga tidak dapat memberikan pemahaman yang lebih baik tentang performa model. Penelitian ini hanya menggunakan algoritma klasifikasi yang umum digunakan dan tidak mempertimbangkan model kasifikasi lainnya yang mungkin memberikan performa yang lebih baik.

Dalam penelitian selanjutnya, disarankan untuk memperluas analisis dengan melibatkan lebih banyak model klasifikasi dan mempertimbangkan variasi parameter untuk meningkatkan performa. Selain itu, dataset yang lebih beragam dan representatif juga dapat membantu untuk mengevaluasi dan membandingkan model secara lebih akurat.

##### References

1. D. Rani dan S. D. Setiawati, “Penyajian Jurnalistik Online Infobdg untuk Menjadi Sumber Informasi Kredibel,” J. Jurnalisa, vol. 6, no. 2, pp. 233–247, 2020.
2. F. S. Nurfikri dan M. S. Mubarok, “Klasifikasi Topik Berita Menggunakan,” vol. 5, no. 1, pp. 1579–1588, 2018.
3. Briggs, Mark. Journalism next: A practical guide to digital reporting dan publishing. CQ Press, 2013.
4. H. Junaedi, H. Budianto, I. Maryati, dan Y. Melani, “Data Transformation pada Data Mining,” Pros. Konf. Nas. Inov. dalam Desain dan Teknol., vol. 7, pp. 93–99, 2011.
5. T. Jamaluddin dan dkk, “Perbdaningan Algoritma Sentencepiece BPE dan Unigram Pada Tokenisasi Artikel Bahasa Indonesia Pendahuluan Studi Terkait,” e-Proceeding Eng., vol. 7, no. 2, pp. 8323–8331, 2020.
6. M. J. Lavin, Z. Leblanc, dan Q. Dombrowski, “The Programming Historian Analyzing Documents with TF-IDF,” Program. Hist., pp. 1–21, 2020.
7. D. Wahyuningsih dan E. Patima, “Penerapan Naive Bayes Untuk Penerimaan Beasiswa,” Telematika, vol. 11, no. 1, p. 135, 2018, doi: 10.35671/telematika.v11i1.665.
8. Y. Zhai, W. Song, X. Liu, L. Liu, dan X. Zhao, “A Chi-square Statistics Based Feature Selection,” 2018 IEEE 9th Int. Conf. Softw. Eng. Serv. Sci., pp. 160–163, 2018.
9. Schutze, Hinrich, Christopher D. Manning, dan Prabhakar Raghavan. Introduction to information retrieval. Cambridge University Press, 2008.
10. R. Wati, “Penerapan Algoritma Naive Bayes Dan Particle Swarm Optimization Untuk Klasifikasi Berita Hoax Pada Media Sosial,” JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer), vol. 5, no. 2, pp. 159–164, 2020, doi: 10.33480/jitk.v5i2.1034.
11. Han, Jiawei, Jian Pei, dan Hanghang Tong. Data mining: concepts and techniques. Morgan kaufmann, 2022.
12. S. Sudianto, A. D. Sripamuji, I. Ramadhanti, R. R. Amalia, J. Saputra, dan B. Prihatnowo, “Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasisifikasi Topik Berita,” J. Nas. Pendidik. Tek. Inform. JANAPATI, vol. 11, no. 2, pp. 84–91, 2022.