JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 10, No. 2, MEI, 2025, Pp. 137 – 144
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Perbandingan Kinerja Naïve Bayes dan Random Forest dalam Mendeteksi Berita Palsu

William [(1)], Teny Handhayani [(2)*]
Departemen Teknik Informatika, Universitas Tarumanagara, Jakarta, Indonesia
e-mail : william.535210013@stu.untar.ac.id, tenyh@fti.untar.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 26 Desember 2023, direvisi 30 Maret 2024, diterima 12 April 2024, dan dipublikasikan 31 Mei 2025.

Abstract

Fake news has become a serious problem in today's digital era. The existence of fake news can have various negative impacts, including the spread of misinformation, social unrest, and economic losses. This study compares the performance of Naïve Bayes and Random Forest classification methods in detecting fake news. Both methods were evaluated on a news dataset comprising 44,898 samples. It uses public data from the Kaggle repository. The news samples are represented by four features: title, news content, subject, and news date. This data is then subjected to cleaning, stemming, tokenization, and feature extraction. The results indicate that the Random Forest method outperforms the Naïve Bayes method. The Random Forest method has an accuracy of 99%, while the Naïve Bayes method has an accuracy of 96%. In general, this research demonstrates that the Random Forest method can be a viable alternative for detecting fake news.

Keywords: Naïve Bayes Algorithm, Random Forest, Text Classification, Fake News Detection, Machine Learning

Abstrak

Berita palsu menjadi salah satu masalah yang serius di era digital saat ini. Keberadaan berita palsu dapat menimbulkan berbagai dampak negatif, seperti penyebaran informasi yang salah, keresahan sosial, hingga kerugian ekonomi. Oleh karena itu, diperlukan metode yang efektif untuk mendeteksi berita palsu. Penelitian ini membandingkan kinerja metode klasifikasi Naïve Bayes dan Random Forest dalam mendeteksi berita palsu. Kedua metode tersebut diujicobakan terhadap *dataset* berita yang terdiri dari 44.898 sampel. *Dataset* yang digunakan merupakan data publik dari repositori Kaggle. Sampel berita diwakili empat fitur, yaitu judul, isi berita, tipe berita, dan tanggal berita. Data ini kemudian dilakukan *cleaning*, *stemming*, tokenisasi, dan ekstraksi fitur. Hasil penelitian menunjukkan bahwa metode Random Forest memiliki kinerja yang lebih baik dibandingkan metode Naïve Bayes. Metode Random Forest memiliki akurasi sebesar 99%, sedangkan metode Naïve Bayes memiliki akurasi sebesar 96%. Secara umum, penelitian ini menunjukkan bahwa metode Random Forest dapat menjadi alternatif yang efektif untuk mendeteksi berita palsu.

Kata Kunci: Algoritma Naïve Bayes, Random Forest, Klasifikasi Teks, Deteksi Berita Palsu, Pembelajaran Mesin

1. PENDAHULUAN

Penyebaran berita palsu (*fake news*) di media sosial dan internet telah menjadi masalah yang penting dalam era digital saat ini. Berita palsu dapat menimbulkan dampak negatif yang luas, seperti penyebaran informasi yang salah, polarisasi masyarakat, dan bahkan konflik (Fawaid et al., 2021). Oleh karena itu, penting untuk mengembangkan metode yang efektif untuk mendeteksi berita palsu. Berita palsu menjadi salah satu masalah yang serius di era digital saat ini. Keberadaan berita palsu dapat menimbulkan berbagai dampak negatif, seperti penyebaran informasi yang salah, keresahan sosial, hingga kerugian ekonomi.

Dalam beberapa tahun terakhir, konten *online* telah memainkan peran penting dalam mempengaruhi keputusan dan opini pengguna. Opini seperti ulasan *online* adalah sumber

informasi utama bagi pelanggan *e-commerce* untuk membantu mendapatkan wawasan tentang produk yang mereka rencanakan untuk dibeli. Dalam beberapa tahun terakhir, konten *online* telah memainkan peran krusial dalam memengaruhi opini dan keputusan pengguna. Opini, seperti ulasan produk, menjadi sumber informasi utama bagi konsumen *e-commerce* untuk mendapatkan wawasan mengenai produk yang akan mereka beli. Namun, perhatian terhadap spam opini telah meluas, tidak hanya terbatas pada ulasan produk, tetapi juga melibatkan penyebaran berita palsu dan artikel yang menyesatkan (Alsharif, 2022). Situs media sosial seperti Google Plus, Facebook, dan Twitter menjadi sumber utama penyebaran berita palsu.

Meskipun permasalahan berita palsu bukan hal baru, mendeteksinya menjadi tantangan kompleks, terutama karena manusia cenderung mempercayai informasi yang menyesatkan. Kemampuan manusia untuk secara manual mengidentifikasi berita palsu terbatas, memerlukan pengetahuan mendalam tentang tipe berita tersebut. Selain itu, dengan sifat terbuka web dan kemajuan dalam ilmu komputer, pembuatan dan penyebaran berita palsu semakin disederhanakan, membuatnya sulit untuk mengukur niat dan dampaknya. Membedakan berita palsu juga terbukti lebih sulit daripada ulasan produk palsu, karena berita palsu dapat menyebar dengan cepat melalui media sosial dan komunikasi mulut ke mulut.

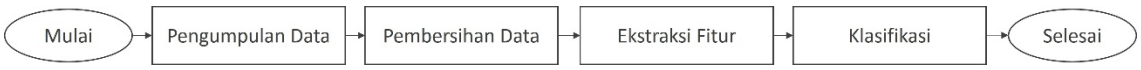
Terdapat banyak metode *machine learning* yang dapat digunakan untuk deteksi berita palsu (Hanum et al., 2024; Lazuardi et al., 2023; Praha et al., 2024). Metode klasifikasi adalah metode yang digunakan untuk mengelompokkan data ke dalam dua atau lebih kelas. Metode klasifikasi yang digunakan untuk memprediksi berita palsu adalah metode Naïve Bayes (Qubra & Saputra, 2024; Santoso et al., 2020) dan metode Random Forest (Ariatmanto & Rifai, 2024). Metode Naïve Bayes adalah metode *machine learning* yang sederhana dan mudah diterapkan. Metode ini menggunakan distribusi probabilitas untuk mengklasifikasikan berita sebagai asli atau palsu. Metode Random Forest adalah metode *machine learning* yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi.

Pada beberapa penelitian menunjukkan bahwa metode *machine learning*, terutama *deep learning*, merupakan metode yang efektif untuk deteksi berita palsu (Anand et al., 2023; Arora & Sikka, 2023; Nath et al., 2021). Metode *machine learning* dapat digunakan untuk mengekstrak fitur-fitur dari teks berita yang dapat digunakan untuk membedakan berita palsu dari berita asli. Hasil penelitian juga menunjukkan bahwa kombinasi metode *machine learning* dan *deep learning* dapat menghasilkan akurasi yang lebih tinggi dibandingkan dengan metode machine learning saja. Hal ini disebabkan oleh kekuatan masing-masing metode yang dapat saling melengkapi. Beberapa metode *machine learning* yang digunakan yaitu Naïve Bayes, Decision Tree, Random Forest, dan Long Short-Term Memory (LSTM).

Penelitian ini bertujuan untuk membandingkan kinerja algoritma Naïve Bayes dan Random Forest untuk mengklasifikasi berita palsu, dengan menggunakan data dari Kaggle. Hasil penelitian ini diharapkan dapat memberikan informasi yang bermanfaat bagi pengembangan metode deteksi berita palsu. Penelitian ini juga dapat menjadi referensi bagi para peneliti dan pengembang sistem deteksi berita palsu.

2. METODE PENELITIAN

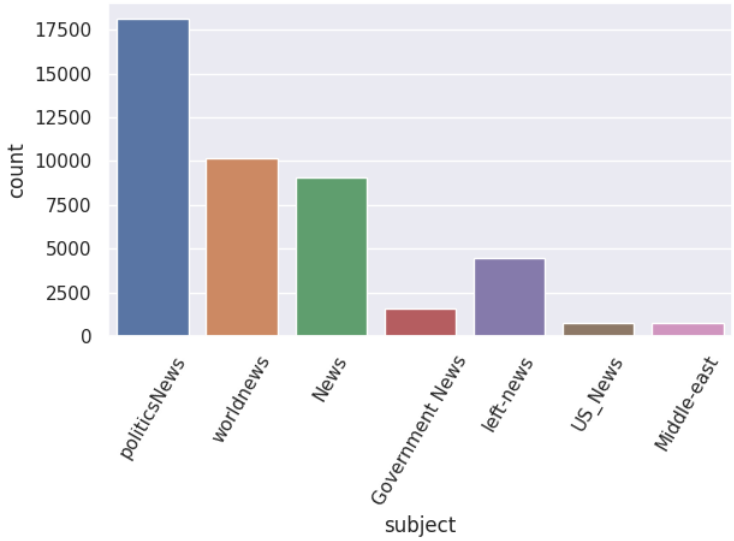
Metode penelitian ini menjelaskan tahapan-tahapan yang dilakukan dalam proses pembangunan dan evaluasi model klasifikasi. Dimulai dari penyajian data yang digunakan, kemudian dijabarkan metode yang diterapkan untuk membersihkan data (*data cleaning*) serta teknik ekstraksi fitur yang relevan. Selanjutnya, hasil dari proses tersebut digunakan sebagai input dalam pengembangan model klasifikasi yang kemudian dievaluasi untuk mengukur performanya. Gambar 1 menggambarkan alur keseluruhan metode penelitian yang digunakan dalam studi ini.



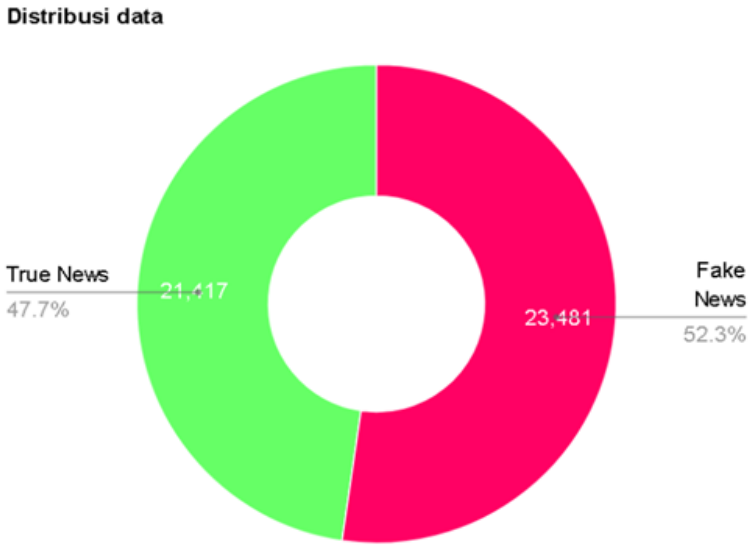
Gambar 1 Metode Penelitian

2.1 Data

Dataset yang digunakan merupakan data publik yang dapat diakses melalui tautan https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset?select=True.csv. Berita asli dikumpulkan dari Reuters.com (*website* berita). Item berita palsu dikumpulkan dari situs web tidak dapat diandalkan yaitu Politifact (pemeriksaan fakta organisasi di AS) telah bekerja sama dengan Facebook untuk memberantasnya. Penelitian ini berfokus pada artikel berita politik karena saat ini menjadi target utama para *spammer*. Artikel berita dari kategori palsu dan jujur terjadi di *timeline* yang sama, khususnya pada tahun 2016. Setiap artikel panjangnya lebih dari 200 karakter. Data ini memiliki fitur-fitur berupa judul berita, isi berita, tanggal berita, tipe berita, dan label data. Tipe berita dibagi menjadi tujuh tipe yaitu politicsNews, worldnews, News, Government News, left-news, US_News, Middle-east. Distribusi data pada tipe data yang terdiri dari berita palsu dan berita asli dapat dilihat pada Gambar 2. Jumlah berita asli adalah 21.417 data dan berita palsu adalah 23.481, perbandingan jumlah data berdasarkan berita palsu dan asli ini cukup seimbang, visualisasi pembagian *dataset* dapat dilihat pada Gambar 3.



Gambar 2 Distribusi Data Berdasarkan Tipe Berita



Gambar 3 Perbandingan Jumlah Data Berita Asli dan Berita Palsu

2.2 Pembersihan Data

Data yang dikumpulkan perlu dibersihkan terlebih dahulu. Pembersihan data dilakukan untuk menghilangkan *noise* atau data yang tidak diperlukan. Dalam penelitian ini, pembersihan data dilakukan dengan menghilangkan *stopwords*, melakukan *stemming*, dan melakukan *word tokenizing* dengan menggunakan pustaka (*library*) nltk. *Stopwords* adalah komponen standar tugas pemrosesan bahasa alami untuk pengambilan informasi, pengindeksan, pemodelan tipe berita, dan klasifikasi teks. Mereka adalah komponen data yang tidak informatif yang sering kali dihapus selama langkah pra-pemrosesan. Kata-kata ini sering kali muncul di banyak dokumen bahasa alami atau bagian teks yang berbeda dalam sebuah dokumen, namun hanya membawa sedikit informasi tentang bagian teks tersebut. Meskipun para peneliti menggunakan daftar *stopwords* yang tersedia yang berasal dari sumber daya non-teknis, jargon teknis bidang teknik mengandung kata-kata mereka sendiri yang sangat sering dan tidak informatif dan tidak ada daftar *stopwords* standar untuk aplikasi pemrosesan bahasa teknis (Sarica & Luo, 2021).

Stemming adalah proses yang digunakan dalam pra-pemrosesan data untuk mengambil informasi dengan melacak kata-kata yang dibubuhkan kembali ke akarnya. Cara ini sudah digunakan sejak lama dan terbukti menghasilkan tingkat akurasi yang tinggi. Namun efektivitas *stemming* dapat berbeda-beda tergantung formalitas bahasa yang diproses. Misalnya, mungkin tidak banyak metode *stemming* untuk pemrosesan bahasa non-formal. Metode ini sering digunakan untuk meningkatkan akurasi model pengklasifikasi teks (Rianto et al., 2021).

Tokenisasi kata adalah langkah pra-pemrosesan mendasar untuk hampir semua tugas Pemrosesan Bahasa Alami (NLP) (Song et al., 2021). Tahapan ini adalah proses tokenisasi atau pemisahan *string*, yaitu mengubah teks menjadi daftar token. Token dapat dianggap sebagai unit terkecil dari suatu teks yang memiliki makna kontekstual. Misalnya, dalam sebuah kalimat, setiap kata bisa dianggap sebagai satu token. Selanjutnya, jika dilihat dari struktur yang lebih besar, satu kalimat dapat dianggap sebagai token dalam sebuah paragraf (Rai & Borah, 2021). Dalam tokenisasi, kalimat dipecah menjadi unit-unit bermakna yang lebih kecil yang dikenal sebagai token. Token merupakan satuan terkecil yang memiliki arti dalam konteks pengolahan bahasa seperti kata, tanda baca, atau karakter khusus. Tokenisasi dilakukan dengan mencari batasan kata dalam kalimat yaitu titik awal dan akhir dari setiap kata. Proses ini dikenal juga dengan istilah segmentasi, karena bertujuan untuk memisahkan teks menjadi segmen-segmen kecil yang dapat dianalisis lebih lanjut.

2.3 Ekstraksi Fitur

Data yang telah dibersihkan kemudian digunakan untuk mengekstrak fitur. Ekstraksi fitur dilakukan untuk mengubah data menjadi format yang dapat diolah oleh model. Dalam penelitian ini, ekstraksi fitur dilakukan dengan menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF telah terintergrasi dalam *library* sklearn. TF-IDF adalah ukuran statistik yang mengevaluasi seberapa relevan suatu kata dengan dokumen dalam kumpulan dokumen. Perhitungan TF-IDF dilakukan dengan mengalikan dua metrik: frekuensi kemunculan sebuah kata dalam sebuah dokumen (TF) dan kebalikan dari frekuensi dokumen yang mengandung kata tersebut dalam seluruh dokumen (IDF). Nilai TF-IDF meningkat secara proporsional dengan berapa kali sebuah kata muncul dalam dokumen dan menurun seiring dengan jumlah dokumen dalam korpus yang memuat kata tersebut. Metode ini sering digunakan sebagai faktor pembobotan dalam pencarian pengambilan informasi, penambangan teks, dan pemodelan pengguna.

2.4 Algoritma Klasifikasi

Dalam studi ini, digunakan *library* Scikit-learn (sklearn) untuk menerapkan algoritma *machine learning*. Scikit-learn adalah pustaka *machine learning* berbasis Python yang menyediakan berbagai algoritma dan *tools* untuk klasifikasi, regresi, klastering, reduksi dimensi, pemilihan fitur, dan praproses data (Zollanvari, 2023). Dalam konteks penelitian ini, sklearn digunakan khususnya untuk membangun dan mengevaluasi model klasifikasi secara efisien dan terstruktur.

2.4.1 Algoritma Naïve Bayes

Penelitian ini menggunakan algoritma Naïve Bayes, lebih spesifiknya Multinomial Naïve Bayes, yang merupakan salah satu algoritma populer untuk klasifikasi teks. Algoritma ini mengasumsikan bahwa fitur-fitur tersebut independen secara kondisional, yang mungkin tidak selalu benar dan dapat memengaruhi kinerjanya. Untuk mengatasi masalah ini, para peneliti telah mengusulkan berbagai perluasan pada Multinomial Naïve Bayes. Salah satu perluasan tersebut adalah Extended Multinomial Naïve Bayes Structure, yang menggabungkan ketergantungan fitur menggunakan penduga satu ketergantungan (Solanki & Saxena, 2020).

Multinomial Naïve Bayes adalah pengklasifikasi probabilistik yang digunakan dalam berbagai aplikasi, seperti mendeteksi berita palsu, analisis sentimen, dan identifikasi kejahatan. Dalam pendekatan ini fitur dianggap sebagai variabel diskrit, sedangkan label kelas dapat memiliki lebih dari dua nilai kategori (Yerlekar et al., 2021). Dengan kesederhanaan dan efisiensinya, algoritma ini tetap menjadi pilihan yang kuat untuk tugas-tugas klasifikasi berbasis teks.

2.4.2 Algoritma Random Forest

Random Forest adalah algoritma pembelajaran mesin berbasis *ensemble* yang menggunakan kumpulan *decision tree* untuk meningkatkan akurasi dan ketahanan model. Algoritma ini termasuk dalam metode *bagging*, yaitu menggabungkan beberapa *decision tree* untuk mengurangi *overfitting* dan meningkatkan kinerja generalisasi (Abdullah & Prasetyo, 2020). Random Forest dapat digunakan untuk berbagai tugas seperti klasifikasi, regresi, dan lainnya. Dalam proses pelatihannya, algoritma ini beroperasi dengan membangun sejumlah besar pohon keputusan pada waktu pelatihan. Untuk prediksi klasifikasi, hasil akhir ditentukan berdasarkan modus dari kelas (klasifikasi) atau rata-rata prediksi (regresi) dari pohon individu (Breiman, 2001). Dalam Random Forest, setiap *decision tree* dilatih berdasarkan *subset* acak dari data pelatihan dan subset acak fitur. Keacakan ini membantu mengurangi korelasi antar pepohonan dan meningkatkan keanekaragaman *ensemble*.

2.5 Skema Eksperimen

Sebelum data digunakan untuk pelatihan model, data dibersihkan terlebih dahulu untuk menghilangkan *noise* atau data yang tidak diperlukan. Proses pembersihan data ini meliputi, menghilangkan *stopwords*, *stemming*, dan *word tokenizing*. Setelah data dibersihkan, fitur-fitur yang penting untuk klasifikasi berita palsu diekstraksi menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency). Data yang telah dibersihkan dan diekstraksi fiturnya kemudian dibagi menjadi dua set data, yaitu set data pelatihan (*training set*) sebesar 80% dan set data pengujian (*testing set*) sebesar 20%. Set data pelatihan digunakan untuk melatih model klasifikasi, sedangkan set data pengujian digunakan untuk mengevaluasi kinerja model klasifikasi. Pembagian data tersebut untuk memastikan bahwa model klasifikasi tidak terlalu dilatih pada set data pelatihan dan dapat generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Kinerja model klasifikasi dievaluasi dengan menggunakan beberapa metrik, yaitu akurasi, presisi, *recall*, dan *F1-Score*.

3. HASIL DAN PEMBAHASAN

Penelitian ini membandingkan kinerja dua metode klasifikasi, yaitu Naïve Bayes dan Random Forest, dalam mendeteksi berita palsu menggunakan dataset berisi 44.898 sampel. Sebelum diklasifikasikan, fitur-fitur seperti judul, isi, tipe, dan tanggal berita diproses melalui tahap pembersihan dan ekstraksi. Random Forest menunjukkan akurasi yang lebih tinggi berkat kemampuannya menggabungkan banyak pohon keputusan, menjadikannya efektif untuk deteksi berita palsu. Namun, Naïve Bayes tetap kompetitif karena waktu pelatihan yang cepat dan efisiensi komputasi, sehingga cocok digunakan dalam situasi dengan sumber daya terbatas.

Hasil pada Tabel 1 menunjukkan bahwa kedua model mampu menghasilkan performa klasifikasi yang tinggi dilihat dari metrik akurasi, presisi, *recall*, dan *F1-Score*. Meskipun keduanya

memberikan hasil yang baik, Random Forest unggul dengan akurasi sebesar 99%, sedangkan Naïve Bayes mencapai 96%. Selain itu, nilai recall, precision, dan *F1-Score* pada Random Forest juga lebih tinggi daripada Naïve Bayes, menandakan kemampuan yang lebih baik dalam mengidentifikasi berita palsu dengan tepat.

Tabel 1 Hasil Evaluasi Klasifikasi

Metrik	Naïve Bayes	Random Forest
Akurasi	96%	99%
Presisi	96%	99%
Recall	97%	99%
F1-Score	96%	99%

Dalam evaluasi model klasifikasi, *recall* mengukur seberapa baik model dapat mendeteksi berita palsu, maka semakin tinggi *recall*, semakin baik model dalam mendeteksi berita palsu. Akurasi mengukur seberapa akurat model dalam memprediksi label berita, baik berita palsu maupun bukan berita palsu, maka semakin tinggi akurasi, semakin akurat model dalam memprediksi label berita. *F1-Score* adalah gabungan dari *recall* dan akurasi, maka semakin tinggi *F1-Score*, semakin baik model dalam mendeteksi berita palsu dan memprediksi label berita secara akurat. Presisi mengukur seberapa akurat model dalam memprediksi berita palsu sebagai palsu, maka semakin tinggi presisi, semakin akurat model dalam memprediksi berita palsu sebagai palsu.

Model Random Forest menghasilkan performa klasifikasi yang sangat baik dengan nilai presisi, *recall*, dan *F1-Score* di atas 98% untuk setiap kelas. Hal ini menunjukkan kemampuan model dalam mengklasifikasikan data positif dan negatif dengan tingkat akurasi yang tinggi. Akurasi keseluruhan mencapai 99%, menandakan bahwa model ini dapat diandalkan untuk tugas klasifikasi pada *dataset* yang digunakan. Model Naïve Bayes juga memberikan hasil yang baik, meskipun sedikit di bawah performa Random Forest. Dengan nilai presisi, *recall*, dan *F1-Score* di sekitar 96%, Naïve Bayes tetap dapat melakukan klasifikasi dengan tingkat akurasi yang tinggi. Akurasi keseluruhan sebesar 96% menunjukkan bahwa Naïve Bayes adalah pilihan yang layak untuk tugas klasifikasi pada *dataset* ini.

Perlu dicatat bahwa *dataset* yang digunakan dalam penelitian ini mayoritas terkait dengan topik politik. Hal ini dapat mempengaruhi generalisasi hasil penelitian terutama ketika diterapkan pada konteks berita palsu di luar ranah politik. Keterkaitan dominan dengan topik politik dapat menyebabkan model lebih terlatih untuk mengenali pola-pola khusus yang mungkin muncul dalam berita politik, sementara mungkin kurang efektif dalam mendeteksi berita palsu dalam konteks topik yang berbeda. Penting untuk diingat bahwa dinamika dan ciri khas berita palsu dapat bervariasi tergantung pada subjek atau topiknya. Oleh karena itu, hasil penelitian ini mungkin tidak secara langsung dapat diterapkan pada berbagai konteks berita palsu di luar topik politik.

Selain itu, untuk meningkatkan generalisasi dan reliabilitas model perlu dilakukan penelitian lebih lanjut pada topik lainnya. Dengan pemahaman yang lebih mendalam tentang batasan dan keunggulan model dalam konteks tertentu, dapat dikembangkan pendekatan yang lebih holistik dan aplikatif untuk mendeteksi berita palsu di berbagai bidang. Dalam pengembangan model untuk mendeteksi berita palsu, penting untuk mempertimbangkan variasi topik dan konteks yang lebih luas agar model dapat memberikan hasil yang konsisten dan dapat diandalkan di berbagai situasi.

4. KESIMPULAN

Berdasarkan hasil eskperimen, dapat disimpulkan bahwa Random Forest memiliki performa yang lebih unggul dibandingkan Naïve Bayes, dengan akurasi sebesar 99%, sementara Naïve Bayes memperoleh nilai akurasi sebesar 96%. Meskipun demikian, Naïve Bayes unggul dari segi kecepatan komputasi, menjadikannya cocok untuk skenario dengan keterbatasan sumber daya. Naïve Bayes dan Random Forest memberikan hasil yang baik dalam melakukan klasifikasi

penyebaran berita palsu dan dapat disesuaikan dengan kebutuhan deteksi. Penelitian ini dapat menjadi landasan untuk pengembangan lebih lanjut dalam deteksi berita palsu selanjutan, termasuk eksplorasi dengan menerapkan teknik *deep learning* untuk mendeteksi berita palsu dalam Bahasa Indonesia.

DAFTAR PUSTAKA

Abdullah, S., & Prasetyo, G. (2020). Easy Ensemble with Random Forest to Handle Imbalanced Data in Classification. *Journal of Fundamental Mathematics and Applications (JFMA)*, *3*(1), 39–46. https://doi.org/10.14710/jfma.v3i1.7415

Alsharif, N. (2022). Fake Opinion Detection in an E-Commerce Business Based on a Long-Short Memory Algorithm. *Soft Computing*, *26*(16), 7847–7854. https://doi.org/10.1007/s00500-022-06806-5

Anand, A., Kulkarni, R., & Agrawal, P. (2023). Fake News Identification: An Effective Combined Approach Using ML and DL Techniques. *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, 1–6. https://doi.org/10.1109/PCEMS58491.2023.10136087

Ariatmanto, D., & Rifai, A. M. (2024). The Impact of Feature Extraction in Random Forest Classifier for Fake News Detection. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *8*(6), 730–736. https://doi.org/10.29207/resti.v8i6.6017

Arora, Y., & Sikka, S. (2023). Reviewing Fake News Classification Algorithms. In *Proceedings of the Third International Conference on Information Management and Machine Intelligence* (pp. 425–429). Springer. https://doi.org/10.1007/978-981-19-2065-3_46

Breiman, L. (2001). Random Forests. In *Machine Learning* (Vol. 45, Issue 1, pp. 5–32). Springer. https://doi.org/10.1023/A:1010933404324/METRICS

Fawaid, J., Awalina, A., Krisnabayu, R. Y., & Yudistira, N. (2021). Indonesia's Fake News Detection Using Transformer Network. *6th International Conference on Sustainable Information Engineering and Technology 2021*, 247–251. https://doi.org/10.1145/3479645.3479666

Hanum, A. R., Zetha, I. A., Putri, S. C., Wulandari, R. A., Andina, S. P., Fajrina, J. N., & Yudistira, N. (2024). Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *11*(3), 537–546. https://doi.org/10.25126/jtiik.938093

Lazuardi, M. F., Hiunarto, R., Ramadhani, K. F., Noviandi, N., Widayanti, R., & Arfian, M. H. (2023). Hoax News Detection Using Passive Aggressive Classifier and TfidfVectorizer. *Jurnal Teknik Informatika*, *16*(2), 185–193. https://doi.org/10.15408/jti.v16i2.34084

Nath, K., Soni, P., Anjum, Ahuja, A., & Katarya, R. (2021). Study of Fake News Detection Using Machine Learning and Deep Learning Classification Methods. *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 434–438. https://doi.org/10.1109/RTEICT52294.2021.9573583

Praha, T. C., Widodo, W., & Nugraheni, M. (2024). Indonesian Fake News Classification Using Transfer Learning in CNN and LSTM. *JOIV: International Journal on Informatics Visualization*, *8*(3), 1213–1221. https://doi.org/10.62527/joiv.8.2.2126

Qubra, R., & Saputra, R. A. (2024). Classification of Hoax News Using the Naïve Bayes Method. *International Journal Software Engineering and Computer Science (IJSECS)*, *4*(1), 40–48. https://doi.org/10.35870/ijsecs.v4i1.2068

Rai, A., & Borah, S. (2021). Study of Various Methods for Tokenization. In *Applications of Internet of Things* (Vol. 137, pp. 193–200). Springer. https://doi.org/10.1007/978-981-15-6198-6_18

Rianto, R., Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the Accuracy of Text Classification Using Stemming Method: A Case of Non-Formal Indonesian Conversation. *Journal of Big Data*, *8*(1), Article ID: 26. https://doi.org/10.1186/s40537-021-00413-1

Santoso, H. A., Rachmawanto, E. H., Nugraha, A., Nugroho, A. A., Rosal Ignatius Moses Setiadi, D., & Basuki, R. S. (2020). Hoax Classification and Sentiment Analysis of Indonesian News Using Naive Bayes Optimization. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *18*(2), 799–806. https://doi.org/10.12928/telkomnika.v18i2.14744

Sarica, S., & Luo, J. (2021). Stopwords in Technical Language Processing. *PLOS ONE*, *16*(8), Article ID: e0254937. https://doi.org/10.1371/journal.pone.0254937

Solanki, A., & Saxena, R. (2020). Text Classification Using Self-Structure Extended Multinomial Naive Bayes. In *Handbook of Research on Emerging Trends and Applications of Machine Learning* (pp. 107–129). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-5225-9643-1.ch006

Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2021). Fast WordPiece Tokenization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2089–2103. https://doi.org/10.18653/v1/2021.emnlp-main.160

Yerlekar, A., Mungale, N., & Wazalwar, S. (2021). A Multinomial Technique for Detecting Fake News Using the Naive Bayes Classifier. *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, 1–5. https://doi.org/10.1109/ICCICA52458.2021.9697244

Zollanvari, A. (2023). Supervised Learning in Practice: The First Application Using Scikit-Learn. In *Machine Learning with Python* (pp. 111–131). Springer International Publishing. https://doi.org/10.1007/978-3-031-33342-2_4