G-Tech : Jurnal Teknologi Terapan
Vol. 8, No. 3, Juli 2024, hlmn. 1749-1758
E-ISSN: 2623-064X | P-ISSN: 2580-8737

Deteksi Berita *Hoax* dengan Perbandingan Website Menggunakan Pendekatan *Deep Learning* Algoritma BERT

Asep Ripa'i[1]✉, Firman Santoso[2], Farihin Lazim[3]

[1,2,3] Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Ibrahimy Sukorejo Situbondo, Indonesia

Informasi Artikel

Riwayat Artikel

Diserahkan : 08-06-2024
Direvisi : 12-06-2024
Diterima : 16-06-2024

ABSTRAK

Berita *hoax* merupakan informasi yang salah dan menyesatkan yang dapat menyebabkan provokasi dan kebencian bagi pembaca. Dengan adanya akses internet yang mudah membuat persebaran berita *hoax* semakin masif. Oleh karena itu, perlu ada metode yang dapat melakukan deteksi berita *hoax*. Penelitian menggunakan metode *deep learning* dengan mengintegrasikan *text mining* untuk mencari informasi dan pola berita yang berhubugan dengan *hoax*. Dengan menggunakan dataset dari situs kaggle berjumlah sekitar 2700-an kemudian dilakukan *text preprocessing* agar data lebih terstruktur untuk diolah lanjut. Kemudian membuat *feature engineering* dari BERT agar data dapat diproses oleh *machine learning* dengan tiga metode klasifikasi yaitu BERT, SVM dan *random forest* kemudian dilakukan pengujian dan evaluasi. Pada penelitian ini model yang menghasilkan performa yang paling tinggi yaitu BERT dengan (akurasi = 0.99, ROC-AUC = 0.99) dibanding model *machine learning* tradisional.

Kata Kunci:

Berita *Hoax*, BERT, *Deep Learning, Text Mining*

Keywords **:**

BERT , Deep Learning, Hoax News, Text Mining,

ABSTRACT

Hoax news is false and misleading information that can cause provocation and hatred for readers. With easy internet access, the spread of hoax news is getting more massive. Therefore, there needs to be a method that can detect hoax news. The research uses deep learning methods by integrating text mining to find information and news patterns related to hoaxes. By using a dataset from the kaggle site totaling around 2700 then text preprocessing is carried out so that the data is more structured for further processing. Then make feature engineering from BERT so that the data can be processed by machine learning with three classification methods namely BERT, SVM and random forest then testing and evaluation. In this study, the model that produces the highest performance is BERT with (accuracy = 0.99, ROC-AUC = 0.99) compared to traditional machine learning models.

Corresponding Author **:**
Asep Ripa'i
Teknologi Informasi, Fakultas Sain dan Teknologi, Universitas Ibrahimy Sukorejo, Indonesia
Jl. KHR. Syamsul Arifin No. 1-2, Sukorejo, Situbondo, Jawa Timur
Email: asepkhabril20@gmail.com

PENDAHULUAN

Salah satu permasalahan diera digital ini adalah maraknya penyebaran berita *hoax* atau palsu melalui media platform berbagi yang bisa diakses dimanapun membuat persebaran berita *hoax* menjadi cepat dan meluas (Andrian & Nur Asyikin, 2023). Berita *hoax* merupakan informasi

yang salah sengaja dibuat dan dapat menyesatkan pembaca (Munawar & Riadi Silitonga, 2019) berita *hoax* memiliki beberapa karakteristik seperti judul yang provokatif , konten berisi sara, ujaran kebencian, dan sumber media tidak jelas bahkan abal-abal dan terkadang mirip dengan media yang sudah terverifikasi. Pengaruh berita hoax terhadap bangsa Indonesia sangatlah berbahaya, dapat menimbulkan konflik antar masyarakat karena masyarakat Indonesia masih sering mempercayai berita yang tidak jelas sumbernya. (Rama dkk., 2022).

Menurut Dewan Pers, terdapat 61.800 website di Indonesia yang mengaku sebagai portal berita. Sekitar 1700 situs yang telah diverifikasi sebagai situs berita resmi (Munfarida, 2024) artinya terdapat setidaknya ada puluhan ribu situs yang berpotensi menyebarkan berita *hoax* di internet yang harus kita waspadai. Maka dari hal tersebut dibutuhkan suatu metode yang dapat mengidentifikasi atau mengklasifikasi berita *hoax,* melalui teknologi *machine learning* merupakan bagian dari dari kecerdasan buatan atau biasa disebut Artitificial Intelligence (Al) yang dapat otomatis belajar dan kemampuannya yang dapat meningkat berdasarkan data masa lalu tanpa harus diprogram secara eksplisit untuk menghasilkan prediksi dimasa mendatang (Daru Kusuma, 2020) Dengan mengintegrasikan dengan *text mining* kemampuannya dalam mengidentifikasi, mengekstrak, dan menganalisis informasi semantik dari berbagai sumber teks dengan tingkat kualitas yang lebih tinggi (Silwattananusarn & Kulkanjanapiban, 2022) yang dapat diarahkan untuk menemukan suatu pola-pola dan ciri kusus konten *hoax* berdasarkan pemahaman kontek sosial dan historis pada suatu berita.

Beberapa penelitian terdahulu tentang deteksi berita *hoax* dengan metode *machine learning* yang di lakukan oleh (Ula, 2020) menggunakan *machine learning* dengan menguji menggunakan 4 algoritma antara lain : *Multilayer Perceptron*, *Naïve Bayes, Support Vector Machine*, dan *Random Forest*. dengan total 150 artikel sebagai data yang terdiri atas 50 artikel *hoax* dan 100 artikel non *hoax* dan mendapatkan akurasi tertinggi menggunakan algoritma *Random Forest* dengan akurasi 75.37 %. Selanjutnya penelitian yang dilakukan oleh (Prasetya & Ferdiansyah, 2022) menggunakan algoritma*Naive Bayes* dan Pso untuk mengklasifikasi berita *hoax* seputar Covid-19 dengan dataset sebanyak 300 terdiri 50 data berita hoax dan 50 data berita valid, dengan akurasi 86.3%.

Selanjutnya penelitian dengan metode *deep learning* dilakukan oleh (Kurniawan & Mustikasari, 2020) menggunakan teknik *deep learning* yaitu algoritma *Convolutional Neural Network* (CNN) dan *Long Short Term Memory* (LSTM) dengan menggunakan 1786 data dengan total data berita valid sebanyak 802 dan data berita *hoax* 984. Penelitian tersebut menghasilkan performa *accuracy test*, *precision* dan *recall* sebesar 88% untuk CNN dan *accuracy test, prescison* dan *recall* sebesar 83% untuk LSTM. Selanjutnya penelitian yang dilakukan(Awalina dkk., 2021) melakukan pengujian dengan 4 metode yaitu CNN, BiLSTM, Hybrid CNN-BiLSTM dan BERT-Multilingual-Cased dengan dataset sebanyak 2216 dengan rincian 1055 berita valid dan berita *hoax* 1161 penelitan tersebut menghasilkan akurasi untuk CNN 74%, BilSTM 85%, Hybrid CNN-BiLSTM, dan BERT 90%

Berdasarkan penelitian terdahulu peneliti memahami bahwa metode *deep learning* lebih menggunguli dari metode *machine learning* yang tradisional. *Deep learning* adalah teknolgi modern terkini untuk pemrosesan gambar dan analisis data dan bagian dari *machine learning* yang berbasis jaringan saraf tiruan yang terinpirasi dari sistem otak manusia, sehingga dapat lebih optimal untuk menganalisis teks karena memiliki representasi yang lebih komplek dari pada *machine learning* biasa (Yunanto dkk., 2021).

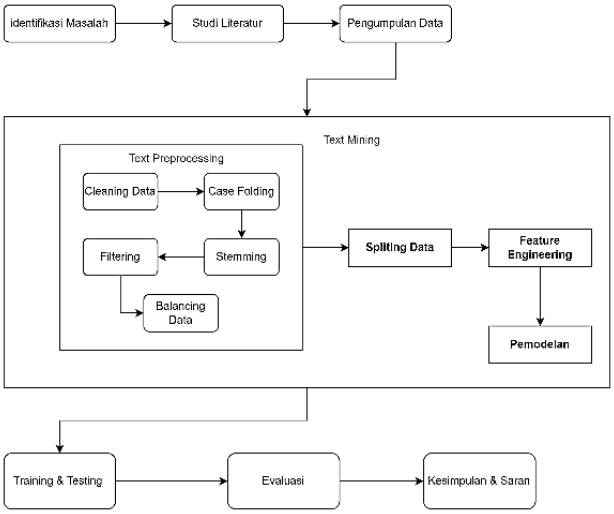
Pada penelitian ini menggunakan algoritma BERT (*Bidirectional Encoder Representations from Transformers*) merupakan algoritma *deep learning* yang berbasis NLP (*Natural Language Processing*) yang diluncurkan google pada tahun 2018 oleh Jacob Devlin dan rekan-rekannya dari *Goolge Research* berdasarkan arsitektur transformer yang yang dirancang untuk memahami konteks dari kata-kata sebuah teks secara dua arah (*bidirectional*) (Devlin dkk., 2019). Perbedaan penelitian ini dengan sebelumnya, peneliti memakai IndoBert sebagai *baseline* yang digunakan untuk *feature engineering* dimana dengan fitur tersebut akan dilakuan klasifikasi dengan tiga algoritma yaitu IndoBert, *Random Forest*, dan *Support Vector Machine* untuk klasifikasi berita *hoax* berbahasa

Indonesia dengan mengintegrasikan metode *text mining* untuk mengetahui performa dari masing-masing algoritma tersebut dan menemukan pola berita *hoax*.

IndoBert merupakan versi Bahasa Indonesia yang di kembangkan oleh (Koto dkk., 2020) berdasarkan algoritma BERT dan mempunyai arsitektur yang sama, yang dilatih menggunakan lebih dari 220 juta kata-kata yang diambil dari 3 sumber utama yaitu Wikipedia (74 juta kata-kata), artikel berita seperti Kompas, Tempo, Liputan 6 (total 55 juta kata-kata), Web Corpus Indonesia (90 juta kata-kata).

METODE PENELITIAN

Pada penelitian ini menggunakan metode *deep learning* dengan BERT berbahasa Indonesia (IndoBert) yang di intergrasikan dengan *text mining*. BERT berfungsi untuk mentokenisasi dan *embedding* kata-kata juga bertugas untuk klasifikasi teks yang digunakan untuk mengidentifikasi pola dan karakteristik teks. Jenis penelitian yang dipakai adalah eksperimen studi kasus merupakan suatu pendekatan untuk mengungkapkan karakteristik dan keunikan dari suatu peristiwa dan untuk menguji hipotesis atau mengidentifikasi hubungan sebab-akibat melalui pengumpulan data empiris (Ridho, 2023). Berikut alur proses penelitian, terlihat pada gambar 1.



Gambar 1. Diagram Alur Tahapan Penelitian

Identifikasi Masalah

Pada tahapan ini menganalisa masalah yang akan di selesaikan berdasarkan latar belakang dan merumuskan masalah bertujuan mempermudah pemecahan masalah.

Studi literatur

Pada tahapan ini mencari informasi, teori dan konsep-konsep dasar materi yang berhubungan dengan penelitian. Pencarian informasi ini dilakukan dengan cara mempelajari dari jurnal, artikel, buku-buku, skripsi, video youtube dan sumber referensi lainnya.

Pengumpulan Data

Untuk penelitian ini data yang digunakan diambil dari website kaggle dengan *keyword* pencarian "*Indonesian Fact and Hoax Political News*". yang berisi dataset berita yang telah di *scrapping* dari website CNN, Kompas, Tempo untuk dataset berita valid, dan Turnbackhoax untuk dataset berita *hoax*, berikut informasi lengkap terlihat pada tabel 1.

Tabel 1. Hasil Kumpulan Dataset Berita Dari Kaggle

	Tanggal		
Sumber	**Awal**	**Akhir**	**Total Dataset**
Kompas	20 April 2017	21 Februari 2023	4749
Tempo	01 Januari 2021	04 Februari 2023	6591
CNN	01 Juni 2021	21 Februari 2023	10001
Turnbackhoax	08 September 2015	28 Februari 2023	10383

Text Mining

Text mining merupakan suatu teknik pengolahan data untuk menemukan informasi baru dan belum diketahui dengan cara mengekstraksi data tidak terstruktur berupa teks, mencari kata-kata yang dapat mewakili isi dokumen sehingga dapat dianalisis untuk menemukan pola(Yehia dkk., 2016).

Text Preprocessing

Text preprocessing adalah proses yang terpenting dalam *text mining*. Ini merupakan langkah pertama dalam proses *text mining* untuk menyeleksi data teks agar lebih terstruktur dan siap untuk diolah lanjut (Mohan, 2015). Berikut penjelasan tahapan-tahapan yang dilakukan.

1. *Cleaning Data*

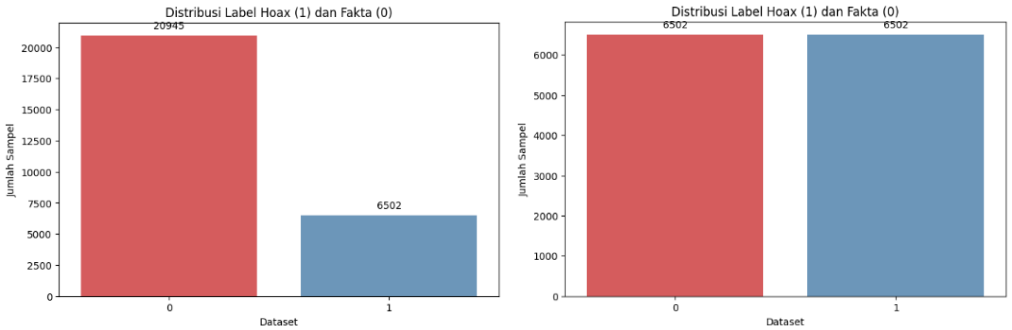
   Pada tahap ini data yang sudah diproleh akan dilakukan seleksi dan koreksi dataset yang salah seperti duplikat, inkosisten, *missing value,* salah format atau error lainya yang dapat menganggu proses analisis.
2. *Case folding*

   Pada tahap merupakan proses mengubah format teks kapital jadi hutuf kecil (lowercase). Tujuannya untuk menyamaratakan teks agar kata-kata tidak terdeteksi berbeda karena perbedaan huruf kapital.
3. *Stemming*

   Pada tahap ini dilakukan untuk mengurangi kata-kata ke bentuk dasar atau akar kata. bertujuan untuk menghilangkan infleksi kata yang tidak diperlukan, seperti imbuhan dan konjugasi, sehingga kata-kata dengan akar yang sama dapat dikenali
4. *Filtering*

   Pada tahap ini merupakan proses untuk membersihkan dan menyaring kata-kata dari elemen yang tidak relevan atau tidak diinginkan. Seperti menghapus kata-kata umum yang tidak bearti, tanda baca, whitspace, angka, karakter spesial, url dll.
5. *Balancing data*

   Setelah dilakukan cleaning data, case folding, stemming, dan filtering kemudian semua dataset di gabungkan, dari dataset tersebut terdapat ketidakseimbangan data antara berita valid dan hoax yaitu 20945 data valid dan 6502 data hoax, oleh karena itu dilakukan balancing data dengan teknik under sampling. Data yang tidak seimbang bisa menyebabkan model *overfiting*.



(a) **(b)**

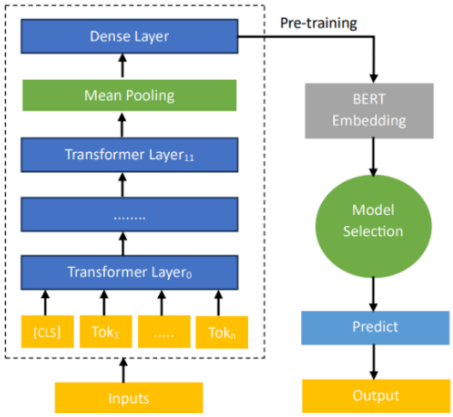
Gambar 2. Persebaran Dataset (a) Sebelum Balancing (b) Sesudah Balancing

Spliting Data

Setelah selesai tahapan *text processing* dan dataset sudah siap dipakai akan dilakukan *spliting data* untuk dilakukan proses pemodelan BERT dengan membagi *training data* 80% dan *testing data* 20%. Data tersebut akan digunakan untuk menguji dan mengevaluasi kinerja model BERT yang akan dibuat.

Feature Engineering

Pada tahap ini dilakukan untuk mengektrasi dan mengonversi data mentah ke dalam format yang dapat digunakan oleh algoritma *machine learning*. Supaya model dapat bekerja lebih baik dan mudah memahami, dengan melakukan pengembangan fitur baru, pemilihan fitur dan modifikasi fitur yang ada (Salehin dkk., 2024). Peneliti menggunakan model BERT sebagai *feature engineering* dengan memanfaatkan *embedding word* milik bert yang dihasilkan melalu token-token yang di tranformasi melalui 12 lapisan. setelah melewati transformer representasi dari setiap token teks digabungkan dengan teknik *mean polling*. Kemudian di akhir lapisan ada *dense layer* berfungsi untuk menggabungkan fitur-fitur telah dipelajari untuk melakukan tugas prediksi dan klasifikasi. Dapat dilihat pada gambar 3 untuk *feature engineering* BERT.

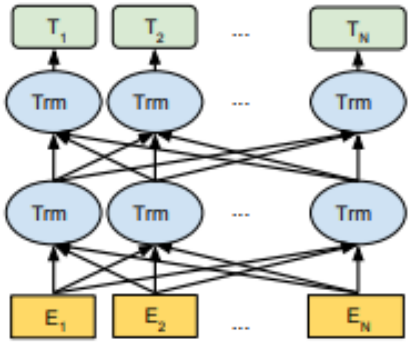


Gambar 3. *Feature Engineering* BERT

Pemodelan

1. BERT

Pada tahap ini peneliti menggunakan BERT (*Bidirectional Encoder Representations from Transformers*) sebagai model yang digunakan untuk *embedding* dan klasifikasi. BERT atau *Bidirectional Encoder Representations from Transformers* adalah algoritma *deep learning* yang berbasis NLP (*Natural Language Processing*) yang dibangun berdasarkan arsitektur transformer yang menerapkan pelatihan dua arah kemudian menggabungkan konteks dari lapisan kiri dan kanan yang diluncurkan google dan dikembangkan oleh (Devlin dkk., 2019) di google reseach. Berikut gambar arsitekur BERT.



Gambar 4. Arsitektur BERT
(Sumber : Devlin dkk., 2019)

BERT memiliki dua arsitektur model yaitu BERT BASE model ini dibangun dari 12 transformer block, 12 *atention layer* dan 768 *hiden layer*, dan BERT LARGE model ini memilki *layer* dan *atantion layer* lebih banyak dari BERT BASE dengan hasil yang lebih banyak yaitu 24 tranformers blok, 16 *attention layer* dan 1024 *hiden layer.* Untuk melakukan klasifikasi pada BERT dilakukan dengan *fine-tuning* yaitu dengan menambahkan *layer* yang digunakan untuk klasifikasi.

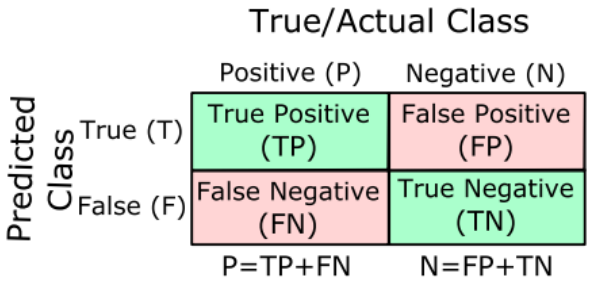
2. *Random Forest*

   Random forest merupakan algoritma pengembangan dari *decision tree* yang memiliki beberapa pohon keputusan dari sekumpulan pohon tersebut terbentuk hutan (*forest)* semakin banyak pohon yang tumbuh maka hasilnya akan akurat dan tidak akan *overfiting*. Kemudian setiap keputusan pohon akan diambil berdasarkan terbanyak (Lin dkk., 2017).
3. *Support Vector Machine*

   Support vector machine atau SVM merupakan algoritma *machine learning* dengan metode pembelajaran *supervised* terkait menganalisis data yang digunakan untuk klasifikasi dan regresi. SVM pada dasarnya merupakan algoritma klasifikasi *linier*, namun dapat digunakan untuk klasifikasi *non-linier* dengan *kernel trick.* Cara kerja algoritma ini adalah dengan mencari *hyperplane* yang digunakan untuk memisahkan dimensional data menjadi dua kelas (Mahesh, 2018).

Evaluasi Model

Proses ini dilakukan untuk mengukur dan menguji kinerja model yang dibuat untuk mengidentifikasi kelemahan dan kelebihan. Peneliti menggunakan metode *Confusion Matrix* adalah metode untuk menilai keakuratan model dengan menentukan klasifikasi benar dan salah. Perhitungan dapat dilakukan untuk menentukan *accuracy*, *precision*, *recall* dan f1-*score* berikut gambar *confusion matrix* (Tharwat, 2018).



Gambar 5 .*Confusion Matrix*
(Sumber : Tharwat, 2018)

1. *Accuracy* merupakan teknik evalasi untuk mengukur tingkat akurat model dalam memprediksi nilai benar dan salah, berikut rumus *accuracy*

   $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
2. *Precision* merupakan teknik evaluasi untuk mengukur tingkat ketepatan akurasi model dalam memprediksi kelas, berikut rumus *precision*

   $Precision = \frac{TP}{TP+FP}$
3. *Recall* merupakan teknik evaluasi untuk mengukur kemampuan model dalam menemukan nilai positif yang sesuai dengan data yang ada, berikut rumus *recall.*

   $Recall = \frac{TP}{TP+FN}$
4. F1-*Score* merupakan teknik evaluasi untuk mengukur kinerja klasifikasi model dengan menggabungkan skor rata-rata *prescision* dan *recall* yang dibobotkan, berikut rumus f1-*score*

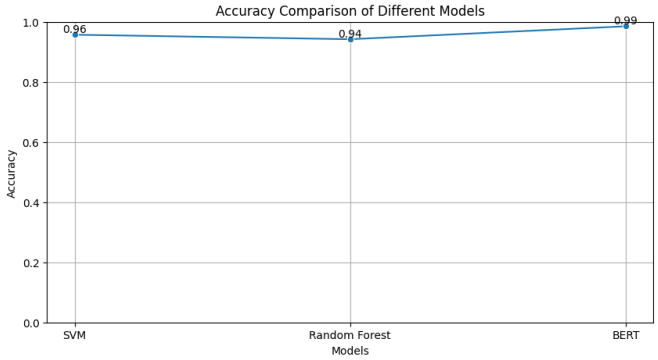
   F1-*score* $= \frac{2 \times precission \times recall}{recall + precision}$
5. ROC-AUC adalah metrik evaluasi untuk mengukur performa model klasifikasi pada semua batas klasifikasi yang mengambarkan antara hubungan *True Positif Rate* (TPR) dan *False Positif Rate* (FPR), berikut rumus TPR dan FPR.

   TPR $= \frac{TP}{TP+FN}$, FPR $= \frac{FP}{FP+TN}$

HASIL DAN PEMBAHASAN

Pada penelitian ini menggunakan metode *feature engineering* untuk mengektrasi data mentah agar dapat diolah dan dimengerti oleh mesin untuk proses *text mining* dengan menggunakan arsitektur model indobert. Indobert merupakan model *pre-trained* menggunakan transformer mengikuti konfigurasi BERTbase dan dikembangkan untuk bahasa Indonesia. Untuk proses *feature engineering* Indobert melibatkan ekstraksi fitur dengan mengambil representasi vektor teks dari token-token berupa *input ids* dan *layer* terakhir atau *layer* terdekatnya yaitu *dense layer* kemudian mengambil token cls untuk dijadikan *embedding*.

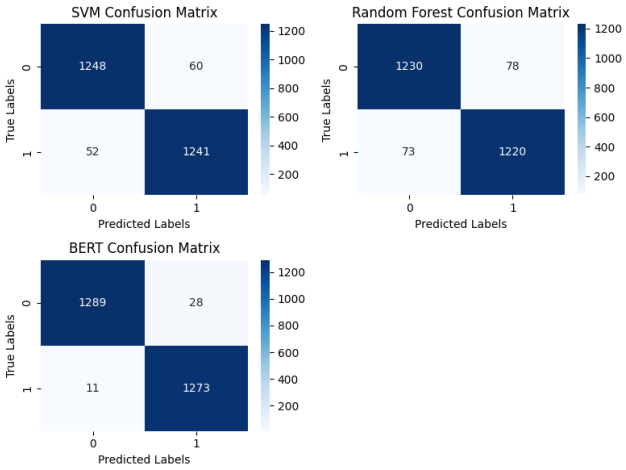
Selanjutnya setelah mendapatkan *embedding* dilakukan proses *training* dengan tiga algoritma klasifikasi yaitu indobert menggunakan metode *fine-tuning* untuk melakukan klasfikasi, dengan memakai *hyperparameter* berupa *batch size train* dan *eval* 16, epoch 5, dan *learning rate* 2e-5. Untuk *support vector machine* menggunakan parameter berupa *probability true,* gamma 0.1, dan kernel *linier.* Untuk *random forest* menggunakan parameter berupa *n estimator* 100 dan *random state* 42. Semua model dilatih dengan dataset yang telah di *text preprocessing* dengan jumlah dataset sebanyak 13004 yang sebelumya berjumlah 27447 data mentah. Dari hasil *training* ketiga model tersebut mendapatkan akurasi sebesar 99% untuk indobert, 96% untuk *support vector machine*, dan 94% untuk *random forest*. Berikut gambar 6 grafik hasil perbandingan akurasi.



Gambar 6. Perbandingan Hasil Akurasi Model BERT, SVM dan *Random Forest*

Evaluasi Model

Setelah melakukan percobaan *training* pada ketiga model yaitu BERT, SVM dan *random forest* dan mendapatkan hasil terbaik, maka tahapan berikutnya mengevaluasi model dengan data baru berupa data *testing* yang telah di *split* dari dataset dengan ukuran 20%. Tujuan evaluasi model adalah untuk mengukur seberapa baik model pada data baru, berikut diagram *confusion matrix* dari percobaan semua model pada gambar 7.



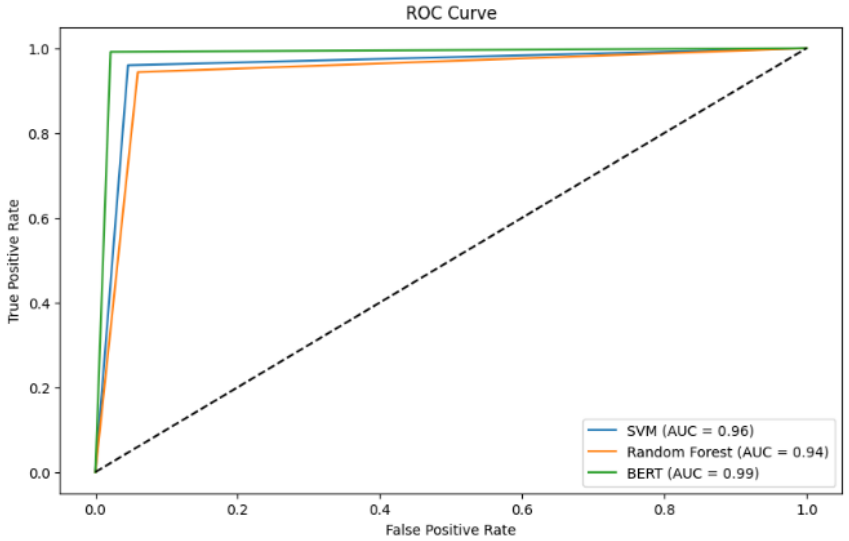
Gambar 7. Diagram Hasil *Confusion Matrix*

Pada gambar 7. dari hasil pengujian *confusion matrix* menunjukan hasil prediksi tepat dan salah dari semua model yang diuji. hasil *matrix* BERT pada baris pertama pada label valid ada 1289 prediksi benar dan 28 prediksi salah, di baris kedua pada label *hoax* ada 1273 prediksi benar dan 11 prediksi salah. Kemudian pada *matrix* SVM pada baris pertama pada label valid ada 1248 prediksi benar dan 60 prediksi salah, di baris kedua pada label *hoax* ada 1241 prediksi benar dan 52 prediksi salah. Dan kemudian pada *random forest* di baris pertama pada label valid ada 1230 prediksi benar dan 78 prediksi salah, di baris kedua pada label *hoax* ada 1220 prediksi benar dan 73 prediksi salah.

Selanjutnya setelah mendapat hasil *confusion matrix* dilakukan pengukuran performa model dengan menghitung *precision, recall,* f1-*score* dan roc-auc.

Tabel 2. Hasil Pengukuran *Matrix*

Matrix	BERT	SVM	*Random Forest*
Precision	0.98	0.95	0.93
Recall	0.98	0.95	0.94
F1-*Score*	0.98	0.95	0.94
ROC-AUC	0.99	0.96	0.94

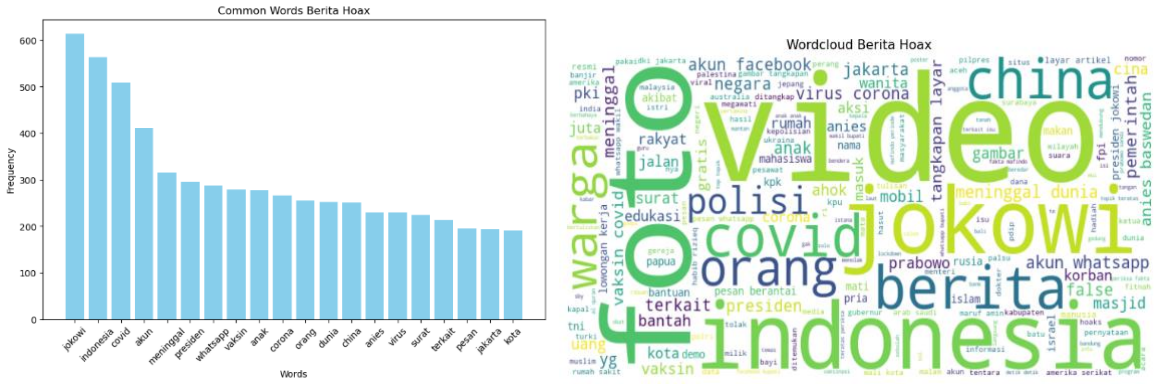


Gambar 8. Grafik Hasil Kurva ROC-AUC

Pada tabel 2 berdasarkan hasil pengujian *matrix* model yang telah dibuat memliki performa yang sangat baik dan gambar 8 pada hasil grafik kurva roc auc pada sumbu y yang menggambarkan prediksi benar sesuai dengan label terdapat garis yeng membentuk ruang diatas garis diagonal dan nilai auc yang mendekati 1 menunjukkan tingkat klasifikasi model bekerja sangat baik. Dari perbandingan tiga model yang di uji model BERT memiliki performa yang lebih baik dari pada model svm dan *random forest* dikarenakan BERT merupakan algoritma berbasis *deep learning* yang memiliki banyak *layer* (*deep neural network*) akan tetapi memiliki kekurangan yaitu membutuhkan biaya komputasi lebih besar dan waktu yang lama.

Analisis Teks

Pada tahap ini dilakukan analisa untuk mengetahui informasi dan pola berita dari dataset yang diperoleh dengan memvisualisikan menggunakan grafik *common word* untuk menemukan umum kata-kata yang sering muncul dalam dataset dan *word cloud* untuk menemukan keterkaitan kata-kata yang sering muncul.



(a) **(b)**

Gambar 9. Visualisasi Analisis (a) Hasil *Common Word* (b) Hasil *Word Cloud*

Berdasarkan gambar 9 pada grafik *common word* kata yang sering muncul pada berita dataset yang dimiliki adalah kata jokowi, Indonesia dan covid ada pada urutan tiga besar menunjukkan berita *hoax* yang sering dibahas dan pada *word cloud* terdapat kata yang menonjol yaitu video dan foto yang mengindikasikan berita *hoax* sering menggunakan foto atau video yang tidak sesuai dengan isi berita untuk membuat isu dan memprovokasi pembaca agar percaya.

KESIMPULAN DAN SARAN

Dapat disimpulkan hasil dari penelitian tentang deteksi berita *hoax* dengan melakukan pengujian menggunakan tiga metode BERT, SVM dan *Random Forest* untuk klasifikasi berita *hoax* menunjukkan performa yang sangat baik dan metode yang paling mengungguli adalah BERT dengan performa (akurasi = 0.99, ROC-AUC = 0.99) dikarenakan BERT merupakan algoritma berbasis *deep learning* yang mempunyai banyak *layer* sehingga dapat belajar dan menganalisa lebih dalam. Pada anlisa teks dapat diketahui bahwa berita *hoax* sering menggunakan foto dan video yang dimanipulasi tidak sesuai dengan isi konten.

Untuk peneliti yang akan datang bisa melakukan *fine-tuning* untuk algoritma *machine learning* tradisional untuk meningkatkan performa dan juga berbagai metode yang lain.

UCAPAN TERIMA KASIH

Ucapan terima kasih kepada dosen pembimbing tugas akhir dan semua pihak yang terlibat sudah meluangkan, waktu, tenaga, pikiran dan arahan sehingga dapat menyelasaikan penelitian ini.

REFERENSI

Andrian, S., & Nur Asyikin, N. (2023). Literasi Digital Dalam Tindak Pidana Penyebaran Berita Bohong dan Menyesatkan Berdasarkan Undang-Undang Nomor 19 Tahun 2016 Tentang Informasi dan Transaksi Elektronik. *Ameena Journal*, *1*(4), 340–350. https://ejournal.ymal.or.id/index.php/aij/article/view/38

Awalina, A., Fawaid, J., Krisnabayu, R. Y., & Yudistira, N. (2021, Juli 14). Indonesia's Fake News Detection using Transformer Network. *SIET '21: 6th International Conference on Sustainable Information Engineering and Technology 2021*. https://doi.org/10.1145/3479645.3479666

Daru Kusuma, P. (2020). *Machine Learning Teori, Program, Dan Studi Kasus*. Sleman : Deepublish.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, Oktober 10). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics.*

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020, November 1). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *COLING 2020 - The 28th International Conference on Computational Linguistics*.

Kurniawan, A. A., & Mustikasari, M. (2020). Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia. *Jurnal Informatika Universitas Pamulang*, *5*(4), 2622–4615. https://doi.org/10.32493/informatika.v5i4.7760

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access*, *5*, 16568–16575. https://doi.org/10.1109/ACCESS.2017.2738069

Mahesh, B. (2018). Machine Learning Algorithms-A Review. *International Journal of Science and Research*. https://doi.org/10.21275/ART20203995

Mohan, V. (2015). Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16. https://www.researchgate.net/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview

Munawar, & Riadi Silitonga, Y. (2019). Sistem Pendeteksi Berita Hoax di Media Sosial dengan Teknik Data Mining Scikit Learn. Dalam *Jurnal Ilmu Komputer* (Vol. 4).

Munfarida, B. (2024, Maret 1). *Dewan Pers: Baru 1.700 Media yang Sudah Terverifikasi* . SindoNews. https://nasional.sindonews.com/read/1331869/15/dewan-pers-baru-1700-media-yang-sudah-terverifikasi-1709283799

Prasetya, F., & Ferdiansyah, F. (2022). Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes. *Jurnal Sistem Komputer dan Informatika (JSON)*, *4*(1), 132. https://doi.org/10.30865/json.v4i1.4852

Rama, M., Sulistyo, D., Fatma, &, & Najicha, U. (2022). PENGARUH BERITA HOAX TERHADAP KESATUAN DAN PERSATUAN BANGSA INDONESIA. *Jurnal Kewarganegaraan*, *6*(1).

Ridho, U. (2023). *METODE PENELITIAN STUDI KASUS: TEORI DAN PRAKTIK* (A. Royani, Ed.; Pertama). Publica Indonesia Utama.

Salehin, I., Islam, Md. S., Saha, P., Noman, S. M., Tuni, A., Hasan, Md. M., & Baten, Md. A. (2024). AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, *2*(1), 52–81. https://doi.org/10.1016/j.jiixd.2023.10.002

Silwattananusarn, T., & Kulkanjanapiban, P. (2022). A text mining and topic modeling based bibliometric exploration of information science research. *IAES International Journal of Artificial Intelligence*, *11*(3), 1057–1065. https://doi.org/10.11591/ijai.v11.i3.pp1057-1065

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Ula, M. (2020). ANALISA DAN DETEKSI KONTEN HOAX PADA MEDIA BERITA INDONESIA MENGGUNAKAN MACHINE LEARNING. *Jurnal Teknologi Terapan and Sains 4.0*, *1*(2), 229. https://doi.org/10.29103/tts.v1i2.3263

Yehia, A. M., Ibrahim, L. F., & Abulkhair, M. F. (2016). Text Mining and Knowledge Discovery from Big Data: Challenges and Promise. *International Journal of Computer Science Issues*, *13*(3), 54–61. https://doi.org/10.20943/01201603.5461

Yunanto, R., Purfini, A. P., & Prabuwisesa, A. (2021). Jurnal Manajemen Informatika (JAMIKA) Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning. *Jurnal Manajemen Informatika (JAMIKA)*, *11*(2), 118–130. https://doi.org/10.34010/jamika.v11i2.5362