Buletin Sistem Informasi dan Teknologi Islam
Vol 7, No 2 (2026), pp 160-170
ISSN: 2721-0901
DOI: http://dx.doi.org/10.33096/busiti.v7i2.3466

Deteksi Berita Hoaks Berbahasa Indonesia yang Dapat Dijelaskan Menggunakan IndoBERT dan SHAP

Explainable Fake News Detection in Indonesian Language Using IndoBERT and SHAP

Muhammad Hasraddin Hasnan[a,1], Rizky Yusliana Bakti [a,2,], Muhammad Faisal[a,3*], Titik Khawa Abd Rahman[b,4], Nurnawaty[c,5], Muhammad Syafaat S Kuba[c,6], Titin Wahyuni[a,7]

[a] *Informatika, Universitas Muhammadiyah Makassar, Makassar, Indonesia*
[b]*School Science and Technology, Asia E University, Selangor, Malaysia*
[c]*Teknik Pengairan, Universitas Muhammadiyah Makassar, Makassar, Indonesia*
[1]*105841107722@student.unismuh.ac.id;* [2]*rizkyyusliana@unismuh.ac.id;* [3]*muhfaisal@unismuh.ac.id;*
[4]*titik.khawa@aeu.edu.my;*[5]*nurnawaty@unismuh.ac.id;* [6]*syafaat_skuba@unismuh.ac.id;* [7]*titinwahyuni@unismuh.ac.id*
**corresponding author*

Informasi Artikel

Diserahkan : 29 April 2026
Diterima : 18 Mei 2026
Direvisi : 28 Mei 2026
Diterbitkan : 31 Mei 2026

Kata Kunci:
Deteksi Berita Hoaks
IndoBERT
SHAP
Explainable AI
Klasifikasi Teks

ABSTRAK

Perkembangan teknologi informasi dan media sosial telah mempercepat penyebaran berita palsu, sehingga diperlukan sistem deteksi yang akurat, andal, dan mudah diinterpretasikan. Penelitian ini bertujuan mengembangkan sistem deteksi fake news berbahasa Indonesia dengan mengintegrasikan IndoBERT sebagai model klasifikasi teks dan SHAP sebagai pendekatan Explainable Artificial Intelligence (XAI) untuk menjelaskan kontribusi kata terhadap hasil prediksi. Dataset diperoleh dari TurnBackHoax dan Kaggle, kemudian melalui tahapan preprocessing berupa cleaning text, filtering bahasa Indonesia, tokenisasi, serta penyeimbangan data menggunakan random oversampling pada data latih. Dari 5.347 data awal, diperoleh 4.980 data setelah filtering bahasa Indonesia, terdiri atas 3.613 data valid dan 1.367 data hoaks. Data dibagi secara stratifikasi dengan rasio 80% untuk pelatihan dan 20% untuk pengujian. Setelah oversampling, data latih menjadi seimbang dengan masing-masing 2.890 sampel per kelas. Hasil eksperimen menunjukkan bahwa model baseline TF-IDF dan Logistic Regression memperoleh akurasi 77%, sedangkan IndoBERT mencapai akurasi 87%, dengan precision 0,87, recall 0,95, dan F1-score 0,91 pada kelas hoaks. Visualisasi SHAP menunjukkan token penting yang memengaruhi klasifikasi. Hasil ini membuktikan bahwa integrasi IndoBERT dan SHAP efektif meningkatkan deteksi berita palsu sekaligus memberikan transparansi model.

***Keywords*:**
Fake News Detection
IndoBERT
SHAP
Explainable AI
Text Classification

ABSTRACT

The development of information technology and social media has accelerated the spread of fake news, so an accurate, reliable, and easy-to-interpret detection system is needed. This study aims to develop an Indonesian fake news detection system by integrating IndoBERT as a text classification model and SHAP as an Explainable Artificial Intelligence (XAI) approach to explain the contribution of words to prediction results. The dataset was obtained from TurnBackHoax and Kaggle, then through preprocessing stages in the form of cleaning text, Indonesian filtering, tokenization, and data balancing using random oversampling on the training data. Of the initial 5,347 data, 4,980 data were obtained after filtering Indonesian, consisting of 3,613 valid data and 1,367 hoax data. The data is divided stratification with a ratio of 80% for training and 20% for testing. After oversampling, the training data became balanced with 2,890 samples per class each. The results of the experiment showed that the TF-IDF baseline model and Logistic Regression obtained an accuracy of 77%, while IndoBERT achieved an accuracy of 87%, with a precision of 0.87, a recall of 0.95, and an F1-score of 0.91 in the hoax class. The SHAP visualization shows the important tokens that affect the classification. These results prove that the integration of IndoBERT and SHAP effectively improves the detection of fake news while providing model transparency.

I. Pendahuluan

Kemajuan teknologi informasi serta media sosial telah mempercepat dan memperluas penyebaran informasi, termasuk berita hoaks yang sulit untuk diverifikasi keabsahannya. [1]. Kondisi ini menjadi tantangan serius karena dapat memengaruhi opini publik dan menimbulkan disinformasi. Berbagai pendekatan telah dikembangkan untuk mendeteksi *fake news*, mulai dari metode *machine learning* hingga *deep learning* yang menunjukkan peningkatan performa klasifikasi teks [2]. Penelitian terbaru menunjukkan bahwa pendekatan deteksi fake news modern tidak hanya berfokus pada peningkatan akurasi, tetapi juga pada kemampuan model dalam memberikan penjelasan terhadap hasil prediksi melalui pendekatan *explainable* AI [3]. Pendekatan tersebut mencakup berbagai metode komputasional yang dirancang untuk mendukung proses deteksi fake news secara real-time dalam skala besar [4]. Berbagai studi tinjauan sistematis juga telah merangkum perkembangan algoritma deteksi fake news, termasuk pengembangan model yang cepat dan skalabel untuk kebutuhan deteksi secara real-time [5],[6]. Model berbasis transformer seperti BERT menjadi pendekatan dominan karena kemampuannya memahami konteks semantik secara mendalam[1], dan dalam konteks bahasa Indonesia, IndoBERT terbukti memiliki kinerja yang lebih baik dibandingkan model lain seperti CNN-LSTM dan metode *ensemble* [7]. Meskipun demikian, model-model tersebut umumnya bersifat black-box sehingga sulit diinterpretasikan, yang menjadi kelemahan dalam aspek transparansi dan kepercayaan pengguna, sehingga diperlukan pendekatan *Explainable Artificial Intelligence* (XAI) untuk menjelaskan hasil prediksi model [8].

Permasalahan utama dalam penelitian ini terletak pada tingginya kompleksitas data teks yang tidak terstruktur, keterbatasan model konvensional dalam memahami konteks bahasa, serta rendahnya *interpretabilitas* model berbasis *deep learning*. Beberapa penelitian sebelumnya telah mengembangkan metode ensemble learning untuk meningkatkan performa deteksi *fake news*, seperti kombinasi beberapa algoritma klasifikasi dan teknik hybrid yang mampu meningkatkan stabilitas serta akurasi prediksi model [7], [9], [10], Selain itu, penelitian lain juga mengembangkan pendekatan berbasis *reinforcement learning* untuk meningkatkan kemampuan pengambilan keputusan dalam pemrosesan bahasa alami [11], [12]. Pendekatan multimodal turut dikembangkan dengan memanfaatkan kombinasi teks, gambar, maupun informasi eksternal untuk meningkatkan representasi data dan akurasi deteksi *fake news* [13], [14]. Di sisi lain, model berbasis transformer khususnya BERT menjadi pendekatan yang dominan karena mampu memahami konteks semantik secara lebih mendalam. Berbagai pengembangan model seperti CredBERT, DistilBERT, dan modifikasi BERT lainnya telah diterapkan untuk meningkatkan performa deteksi *fake news*, termasuk melalui integrasi aspek kredibilitas informasi ke dalam representasi model [15], [16], [17]. Banyak pendekatan yang berhasil mendorong kemampuan model dalam hal klasifikasi menjadi semakin baik. Namun, sebagian besar riset yang ada saat ini masih lebih menekankan pada perbaikan nilai akurasi saja, dan belum memberikan penjelasan yang jelas mengenai bagaimana keputusan yang dihasilkan oleh model tersebut bisa diterima dan dipahami oleh pengguna. Selain itu, penelitian yang mengintegrasikan model bahasa Indonesia dengan pendekatan *Explainable Artificial Intelligence* (XAI) masih relatif terbatas, sehingga aspek transparansi dan kepercayaan pengguna terhadap hasil prediksi belum banyak dikaji secara mendalam.

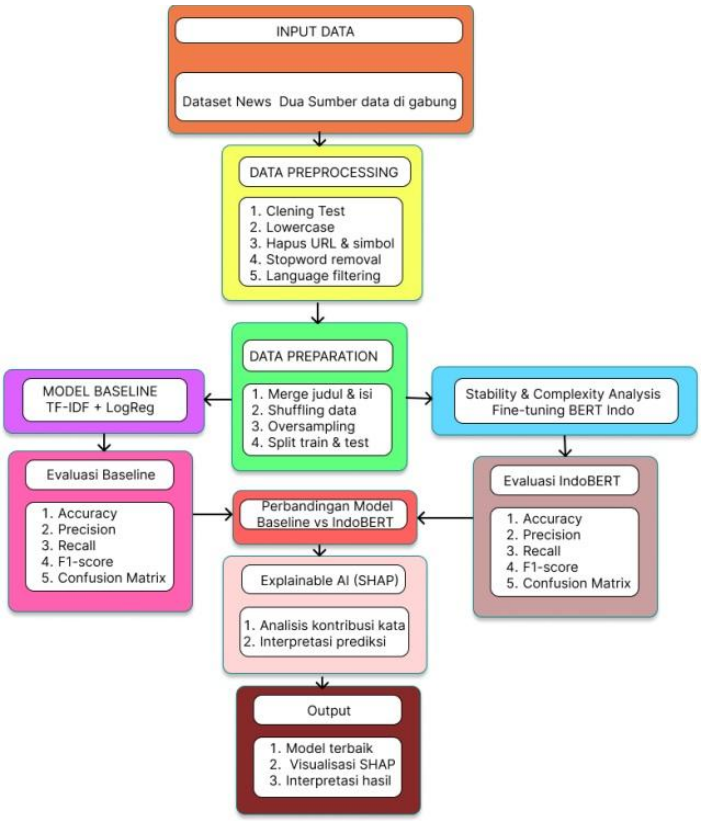
Menyikapi isu tersebut, penelitian ini bertujuan membangun model deteksi berita palsu menggunakan IndoBERT yang tidak hanya presisi tinggi, tapi juga transparan dan mudah diinterpretasikan. Metode yang diajukan menggabungkan model bahasa IndoBERT dengan teknik SHAP guna menilai peran setiap kata dalam proses prediksi yang dihasilkan oleh model. Dengan demikian, penelitian ini menawarkan kebaruan dalam bentuk integrasi deep learning dan XAI pada bahasa Indonesia, serta memberikan solusi yang lebih transparan dibandingkan pendekatan sebelumnya. Pendekatan ini juga melengkapi berbagai penelitian terkini yang mengintegrasikan berbagai komponen seperti model BERT, analisis gaya bahasa, dan verifikasi kredibilitas untuk meningkatkan akurasi deteksi *fake news* [18]. Berbeda dengan penelitian deteksi hoaks berbasis IndoBERT sebelumnya yang umumnya hanya berfokus pada peningkatan performa klasifikasi, penelitian ini secara eksplisit menambahkan aspek *Explainable Artificial Intelligence* (XAI) menggunakan SHAP untuk menjelaskan kontribusi fitur teks terhadap keputusan model. Integrasi ini menjadi penting karena sebagian besar model IndoBERT pada penelitian terdahulu masih bersifat black-box dan belum mampu memberikan interpretasi yang jelas kepada pengguna terkait alasan suatu berita diklasifikasikan sebagai hoaks atau valid. Oleh karena itu, penelitian ini tidak hanya menekankan aspek akurasi, tetapi juga transparansi dan interpretabilitas model dalam proses deteksi fake news berbahasa Indonesia. Metode ini diharapkan mampu menghasilkan sistem deteksi *fake news* yang memiliki performa tinggi sekaligus *interpretabilitas* yang baik sebagai bagian dari pengembangan *state of the art* dalam bidang ini.

II. Metode

A. Tahapan Penelitian

Dalam penelitian ini, pengerjaan dilakukan secara bertahap dan sistematis. Tahapan tersebut mencakup pengumpulan data along dengan proses *preprocessing*-nya, kemudian membangun model baseline sebagai pembanding, mengembangkan model IndoBERT, dan terakhir melakukan evaluasi terhadap tingkat interpretabilitas dengan memanfaatkan pendekatan SHAP. Alur penelitian dirancang untuk menghasilkan model deteksi *fake news* yang tidak hanya memiliki performa klasifikasi yang tinggi, tetapi juga mampu

memberikan transparansi terhadap proses pengambilan keputusan model. Berbagai penelitian sebelumnya menunjukkan bahwa metode deteksi *fake news* terus berkembang, termasuk pendekatan berbasis *ensemble learning* yang mengombinasikan beberapa model melalui teknik soft voting dan hard voting untuk meningkatkan akurasi klasifikasi [19]. Selain itu, pengembangan sistem deteksi *fake news* modern juga mempertimbangkan aspek efisiensi dan skalabilitas agar dapat diterapkan secara real-time dalam lingkungan digital yang dinamis [6]. Metode yang diterapkan dalam riset ini mengikuti tren pengembangan teknik berbasis deep learning serta (XAI) yang saat ini semakin berkembang. Pendekatan tersebut menitikberatkan pada keseimbangan antara kemampuan model dalam menghasilkan prediksi yang akurat serta kemampuan untuk menjelaskan atau menginterpretasikan hasil yang diperoleh. [8], [15], [3]. Kerangka penelitian digambarkan pada Gambar 1.



Gambar 1. Kerangka Penelitian

B. *Data Preprocessing*

Data yang dimanfaatkan dalam penelitian ini diperoleh dari dua platform berbeda, yaitu *TurnBackHoax* dan *Kaggle*. Kedua sumber data tersebut kemudian digabungkan menjadi satu kesatuan agar dapat menghasilkan data yang lebih beragam dan mewakili berbagai kondisi. Dataset TurnBackHoax diperoleh dari portal *fact-checking TurnBackHoax* milik Masyarakat Anti Fitnah Indonesia (MAFINDO) yang berisi berita hasil verifikasi fakta dari berbagai domain seperti politik, kesehatan, sosial, dan ekonomi. Data yang bersumber dari *TurnBackHoax* diperoleh melalui serangkaian proses verifikasi fakta terhadap informasi-informasi yang tersebar di berbagai platform media sosial dan juga situs berita online. Sementara itu, dataset *Kaggle* diperoleh dari repositori publik *Kaggle* yang memuat data berita berbahasa Indonesia dengan label hoaks dan valid yang dikumpulkan dari berbagai sumber berita digital. Kedua dataset terdiri atas teks berita dan label klasifikasi yang digunakan dalam proses pelatihan dan pengujian model.

Tahap *preprocessing* meliputi normalisasi teks, penghapusan URL, simbol non-alfabet, serta pengubahan huruf menjadi *lowercase*. Selain itu, dilakukan penghapusan *stopwords* untuk mengurangi noise pada data. Proses deteksi bahasa diterapkan untuk memastikan hanya teks berbahasa Indonesia yang digunakan, mengingat pentingnya konsistensi bahasa dalam model berbasis NLP [7]. Setelah tahap *preprocessing*, representasi teks menjadi faktor penting dalam menentukan performa model, di mana beberapa penelitian memanfaatkan pendekatan berbasis struktur semantik seperti semantic graph dan topic modeling untuk menangkap hubungan antar topik dalam deteksi *fake news* multibahasa [20].

Setelah proses pembersihan, dilakukan penggabungan antara judul dan isi berita untuk membentuk representasi teks yang lebih informatif. Langkah pertama yang dilakukan adalah mengacak (*shuffling*) dataset guna menghilangkan kemungkinan adanya bias yang muncul akibat urutan data. Setelah itu, data dibagi

menggunakan teknik *stratified sampling* dengan pembagian 80% untuk data latih dan 20% untuk data uji. Ketidakseimbangan jumlah sampel pada setiap kelas di data latih kemudian diatasi dengan teknik *random oversampling* yang berguna untuk meningkatkan kemampuan model dalam mengenali kelas minoritas [7]. Teknik *balancing* ini diterapkan hanya pada data latih, sedangkan data uji tetap menjaga distribusi aslinya untuk memastikan hasil evaluasi model tetap valid dan akurat. Hasil *preprocessing* menunjukkan jumlah data awal sebanyak 5.347 data dan setelah proses *filtering* bahasa Indonesia menjadi 4.980 data. Distribusi label sebelum *balancing* terdiri atas 3.613 data valid dan 1.367 data hoaks. Setelah proses pembagian data, distribusi data uji tanpa *balancing* terdiri atas 723 data valid dan 273 data hoaks. Sementara itu, data latih diseimbangkan menggunakan *random oversampling* sehingga masing-masing kelas berjumlah 2.890 data dan menghasilkan total 5.780 data latih setelah *balancing*.

C. Model Baseline (TF-IDF dan Logistic Regression)

Sebagai pembanding, digunakan model baseline berbasis TF-IDF dan *Logistic Regression*. Representasi TF-IDF diterapkan untuk mengonversi teks menjadi vektor numerik, dengan memerhatikan frekuensi kemunculan kata serta bobot pentingnya di dalam dokumen. Formula TF-IDF didefinisikan sebagai (1):

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

Selanjutnya, *Logistic Regression* digunakan sebagai model klasifikasi dengan fungsi probabilitas seperti (2):

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (2)$$

Model ini dipilih karena sederhana dan banyak digunakan sebagai baseline dalam penelitian klasifikasi teks, meskipun memiliki keterbatasan dalam menangkap konteks semantik [2]. Dalam penelitian lain, pendekatan yang lebih kompleks seperti model hybrid berbasis CNN dan BiLSTM telah dikembangkan untuk meningkatkan kemampuan ekstraksi fitur dan pemodelan konteks dalam deteksi *fake news* [21].

D. Model IndoBERT

Petunjuk Model utama yang digunakan dalam penelitian ini adalah IndoBERT, yang merupakan model berbasis transformer dan sudah menjalani proses pelatihan menggunakan korpus bahasa Indonesia. Tahap fine-tuning dilakukan dengan menambahkan lapisan klasifikasi khusus untuk menyelesaikan tugas klasifikasi biner. Teks-teks yang akan diproses diubah terlebih dahulu menjadi token menggunakan tokenizer dengan panjang maksimum 128 token agar proses komputasi tetap efisien. Pelatihan model dilakukan dengan menggunakan parameter learning rate sebesar 2×10^-5, ukuran batch yang kecil, dan jumlah epoch sebanyak tiga iterations. Fungsi aktivasi *softmax* digunakan untuk menghasilkan probabilitas kelas seperti (3):

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (3)$$

Pendekatan ini memungkinkan model memahami hubungan antar kata dalam konteks kalimat secara lebih mendalam dibandingkan metode berbasis fitur statis. Ini sejalan dengan studi terdahulu yang membuktikan bahwa penggabungan mekanisme attention pada model deep learning dapat memperbaiki representasi kontekstual serta akurasi dalam mendeteksi berita hoaks [22]. Ini selaras dengan riset sebelumnya yang menyoroti superioritas model BERT untuk tugas deteksi berita palsu [14], [15], serta efektivitas IndoBERT dalam konteks bahasa Indonesia[23], [7].

E. Evaluasi Model

Evaluasi terhadap kinerja model dilakukan dengan memakai metrik klasifikasi yang terdiri dari beberapa tolak ukur, antara lain *akurasi, presisi, recall*, dan *F1-score*. Berikut ini merupakan formula atau rumus perhitungan yang digunakan untuk setiap metrik tersebut terlihat pada (4) – (7):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

Penerapan metrik-metrik tersebut dimaksudkan untuk menyajikan evaluasi menyeluruh atas kinerja model, terutama dalam mengatasi masalah ketidakseimbangan kelas data [24]. Di samping itu, *confusion matrix* dimanfaatkan untuk mengkaji pola kesalahan klasifikasi dengan lebih rinci. Selanjutnya, pendekatan *Explainable Artificial Intelligence* (XAI) menggunakan SHAP diterapkan untuk meningkatkan interpretabilitas model.

Guna meningkatkan model agar lebih mudah dipahami, diterapkan teknik SHAP (*SHapley Additive exPlanations*) yang dapat memberikan penjelasan mengenai bagian setiap fitur dalam mempengaruhi hasil prediksi. SHAP memakai konsep nilai Shapley yang terdapat dalam bidang teori permainan, yang dinyatakan sebagai (8):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (8)$$

Dalam konteks penelitian ini, fitur yang dimaksud adalah token kata dalam teks. SHAP bekerja dengan menghapus atau memodifikasi kata tertentu dan mengamati perubahan probabilitas prediksi model. Dengan demikian, dapat diketahui kata-kata yang memiliki kontribusi signifikan terhadap keputusan klasifikasi.

Pendekatan ini memungkinkan interpretasi lokal terhadap setiap prediksi, sehingga model tidak lagi bersifat black-box. Integrasi SHAP dalam model deteksi *fake news* menjadi bagian penting dalam pengembangan sistem yang transparan dan dapat dipercaya [8]

F. Perancangan Sistem

Desain sistem pada penelitian ini mencakup sejumlah komponen pokok yang terhubung secara terpadu. Sistem menerima input berupa teks berita, kemudian dilakukan proses *preprocessing* untuk membersihkan dan menormalisasi data. Selanjutnya, data diproses melalui dua pendekatan, yaitu model baseline dan model IndoBERT. Hasil prediksi dari model IndoBERT kemudian dianalisis menggunakan SHAP untuk memberikan interpretasi terhadap keputusan model.

Secara umum, alur sistem dapat dirumuskan sebagai fungsi (9):

$$y = f(x) \quad (9)$$

di mana x adalah teks input dan y adalah label hasil klasifikasi. Untuk interpretasi, digunakan fungsi tambahan pada (10):

$$y = f(x) + \sum_{i=1}^{n} \phi_i \quad (10)$$

Dengan ϕi menggambarkan tingkat kontribusi yang diberikan oleh setiap fitur atau token dalam mempengaruhi hasil prediksi yang dihasilkan oleh model. Desain sistem pada penelitian ini mencakup sejumlah komponen pokok yang saling terhubung secara terpadu.

Oleh karena itu, sistem yang dikembangkan tidak hanya unggul dalam mengklasifikasikan berita dengan tepat, tapi juga menyediakan penjelasan atas prediksi yang dihasilkan, sehingga meningkatkan transparansi dan kepercayaan pengguna.

III. Hasil dan Pembahasan

A. Hasil dan Model Baseline

Evaluasi model baseline dilakukan menggunakan metode *TF-IDF dan Logistic Regression* terhadap data uji. Hasil evaluasi disajikan dalam bentuk classification report dan confusion matrix untuk memberikan gambaran performa model secara kuantitatif dan visual.

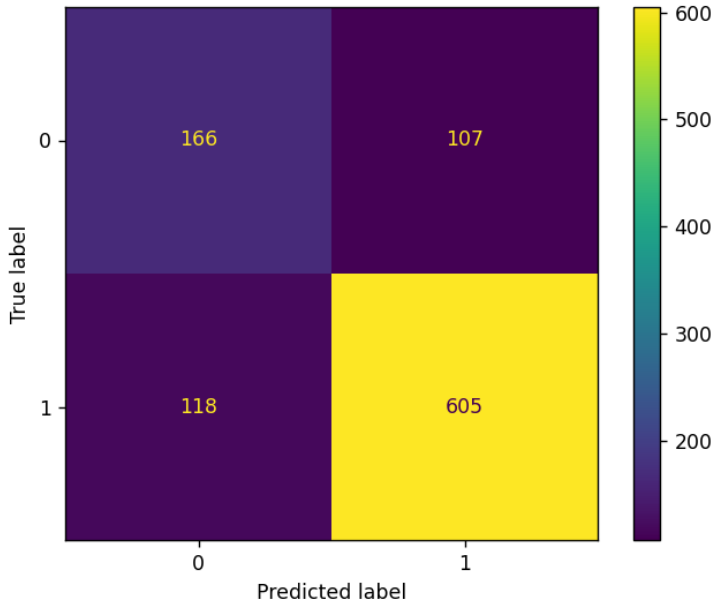
Hasil pengujian model baseline ditampilkan pada Tabel 1.

Tabel 1. Hasil Evaluasi Model Baseline

Kelas	Precision	Recall	F1 - Score	Support
0 (Valid)	0.58	0.61	0.60	273
1 (Hoaks)	0.85	0.84	0.84	723
Accuracy	-	-	0.77	996
Macro Avg	0.72	0.72	0.72	996
Weighted Avg	0.78	0.77	0.78	996
0 (Valid)	0.58	0.61	0.60	273

Berdasarkan *classification report*, model baseline memperoleh akurasi sebesar 77% pada 996 data uji. Hasil yang diperoleh model pada kelas hoaks (1) menunjukkan kinerja yang lebih tinggi dengan nilai *precision* 0,85, *recall* 0,84, dan *F1-score* 0,84, bila dibandingkan dengan kelas valid (0) yang hanya mencapai *F1-score* sebesar 0,60. Nilai *macro average* sebesar 0,72 dan *weighted average* sebesar 0,78 mengindikasikan bahwa

kemampuan model masih belum merata atau seimbang antara kedua kelas tersebut. Secara umum, model TF-IDF dan Logistic Regression cukup baik dalam mendeteksi berita hoaks, namun masih memiliki keterbatasan dalam memahami konteks semantik teks. Visualisasi confusion matrix diperlihatkan pada Gambar 2.



Gambar 2. Confusion Matrix Model Baseline

Berdasarkan hasil confusion matrix, diperoleh nilai True Negative (TN) sebanyak 166, yang menunjukkan bahwa data valid berhasil diklasifikasikan dengan benar sebagai valid. Nilai False Positive (FP) sebesar 107 menunjukkan bahwa terdapat data valid yang keliru diklasifikasikan sebagai hoaks. Selanjutnya, nilai False Negative (FN) sebanyak 118 menunjukkan adanya data hoaks yang salah diklasifikasikan sebagai valid. Sementara itu, nilai True Positive (TP) sebesar 605 menunjukkan bahwa data hoaks berhasil diklasifikasikan dengan benar sebagai hoaks.

Berdasarkan hasil evaluasi, model baseline memperoleh kemampuan klasifikasi dengan nilai akurasi sebesar 77%. Untuk kelas hoaks, model mendapatkan *precision* 85%, *recall* 84%, dan *F1-score* 84%. Di sisi lain, kinerja pada kelas valid masih berada pada level yang lebih rendah, yaitu *precision* 58%, *recall* 61%, dan *F1-score* 60%. Temuan ini menunjukkan bahwa model memiliki kemampuan yang lebih baik dalam mendeteksi beritahoaks apabila dibandingkan dengan mengenali berita valid.

Kesalahahaman dalam proses klasifikasi masih dapat dilihat dari besarnya angka *false positive* dan *false negative* yang dihasilkan oleh model. Kondisi ini menunjukkan bahwa model TF-IDF dan Logistic Regression masih mengalami kesulitan dalam membedakan karakteristik teks valid dan hoaks yang memiliki pola bahasa serupa. Representasi TF-IDF hanya mempertimbangkan frekuensi kemunculan kata tanpa memahami hubungan konteks antar kata secara mendalam, sehingga model kurang optimal dalam menangkap makna semantik dari sebuah kalimat.

Selain itu, nilai *recall* pada kelas valid yang masih rendah mengindikasikan bahwa sebagian berita valid masih sering terdeteksi sebagai hoaks. Oleh karena itu, diperlukan pendekatan yang lebih kompleks seperti model berbasis *transformer* untuk meningkatkan kemampuan pemahaman konteks dan menghasilkan performa klasifikasi yang lebih baik.

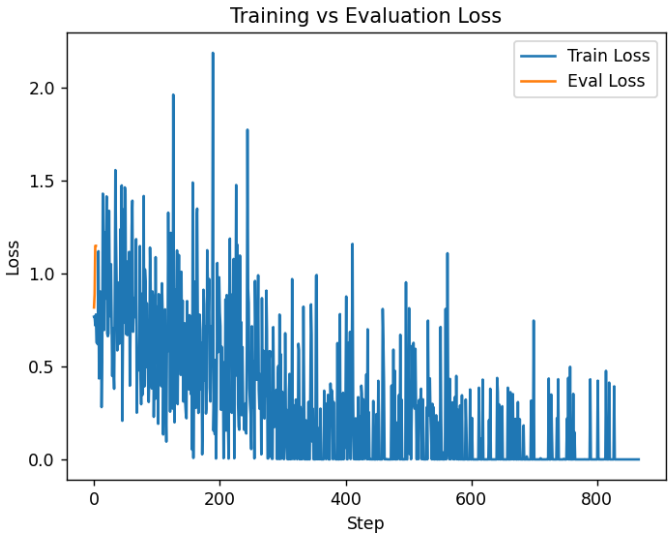
B. Hasil Model IndoBERT

Dalam proses pelatihan model IndoBERT, diterapkan data latih yang sudah melewati tahap *preprocessing* serta proses *balancing*. Kemudian, evaluasi dilakukan terhadap data uji dengan memakai metrik klasifikasi dan visualisasi proses pelatihan agar dapat memahami bagaimana perilaku model sewaktu menjalani training. Berdasarkan hasil evaluasi yang diperoleh dari proses training, model IndoBERT menghasilkan performa seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Hasil Evaluasi Model IndoBERT

Kelas	Precision	Recall	F1 - Score	Support
0 (Valid)	0.83	0.64	0.72	273
1 (Hoaks)	0.87	0.95	0.91	723
Accuracy	-	-	0.87	996
Macro Avg	0.85	0.79	0.82	996
Weighted Avg	0.86	0.87	0.86	996

Berdasarkan classification report yang ditampilkan pada tabel 3, model IndoBERT berhasil mencapai nilai akurasi sebesar 87% ketika diuji menggunakan 996 data uji. Untuk kelas valid (0), model memperoleh *precision* 0,83, *recall* 0,64, dan *F1-score* 0,72. Sementara untuk kelas hoaks (1), model menunjukkan hasil yang sangat memuaskan dengan *precision* 0,87, *recall* 0,95, dan *F1-score* 0,91. Nilai *macro average* sebesar 0,82 dan *weighted average* sebesar 0,86 mengindikasikan bahwa model memiliki kemampuan klasifikasi yang lebih merata dan juga lebih tinggi bila dibandingkan dengan model baseline. Secara umum, hasil ini membuktikan bahwa IndoBERT mampu memahami konteks semantik teks dengan lebih efektif sehingga menghasilkan performa deteksi *fake news* yang lebih akurat [25]. Visualisasi Training dan *Evaluation Loss* ditnjukkan pada Gambar 3.



Gambar 3. Training vs Evaluation Loss

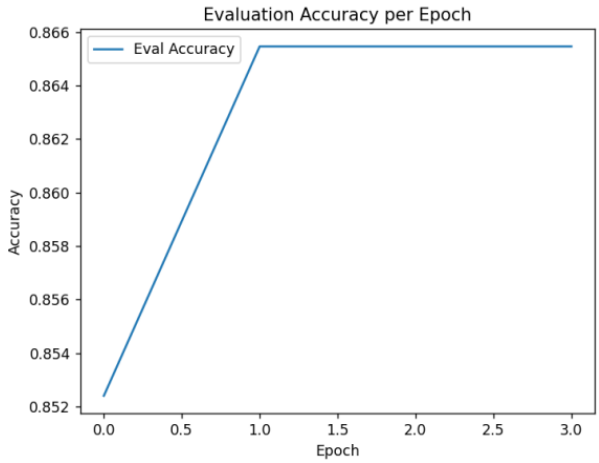
Berdasarkan Gambar 3. terlihat bahwa nilai *training loss* mengalami penurunan secara bertahap seiring bertambahnya langkah (*step*) pelatihan. Pada tahap awal proses training, grafik menunjukkan fluktuasi loss yang cukup tinggi dengan beberapa lonjakan nilai. Namun, setelah proses pelatihan berlangsung, nilai loss cenderung menurun dan semakin mendekati nol. Situasi ini mengindikasikan bahwa model IndoBERT secara bertahap dapat memahami pola yang ada dalam data pelatihan dengan lebih optimal di setiap tahap iterasi pelatihan.

Sementara itu, *evaluation loss* juga menunjukkan kecenderungan yang stabil tanpa peningkatan yang signifikan selama proses evaluasi. Meskipun nilai *eval loss* masih berada di atas *training loss*, pola grafik tidak menunjukkan divergensi ekstrem yang mengarah pada *overfitting* berat. Hal ini menunjukkan bahwa model masih dapat mempertahankan performa yang memadai saat diuji menggunakan data yang tidak digunakan dalam proses pelatihan.

Penurunan *training loss* yang konsisten serta kestabilan *evaluation loss* menunjukkan bahwa proses fine-tuning IndoBERT berjalan dengan baik dan model berhasil mencapai konvergensi. Berdasarkan hasil tersebut, model menunjukkan kemampuan generalisasi yang memadai dalam mengidentifikasi *fake news* pada data baru yang belum pernah diproses sebelumnya. Visualisasi Evaluation Akurasi per Epoch ditunjukkan pada Gambar 4.

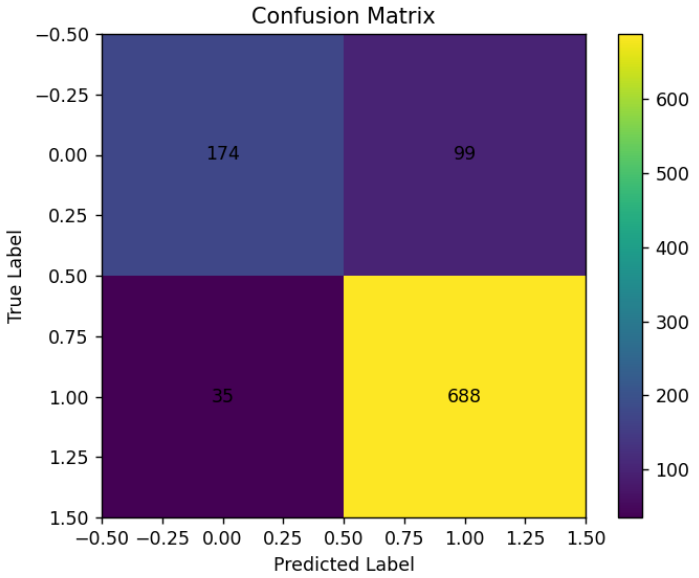
Berdasarkan Gambar 4. terlihat bahwa nilai *evaluation accuracy* mengalami peningkatan pada awal proses pelatihan, yaitu dari sekitar 85,2% pada epoch pertama menjadi sekitar 86,5% pada epoch berikutnya. Setelah mengalami kenaikan, perolehan akurasi tetap konstan hingga akhir proses pelatihan. Hal ini menunjukkan

bahwa model IndoBERT dapat memahami pola data dengan baik serta mencapai titik keseimbangan tanpa mengalami penurunan performa yang berarti.



Gambar 4. Evaluation Accuracy per Epoch

Visualisasi confusion matrix diperlihatkan pada Gambar 2.



Gambar 5. Confusion Matrix Model IndoBERT

Berdasarkan hasil confusion matrix, diperoleh nilai True Negative (TN) sebesar 174, yang menunjukkan bahwa data valid berhasil diklasifikasikan dengan benar sebagai valid. Nilai False Positive (FP) sebesar 99 menunjukkan bahwa terdapat data valid yang salah diklasifikasikan sebagai hoaks. Selanjutnya, nilai False Negative (FN) sebesar 35 menunjukkan adanya data hoaks yang keliru diklasifikasikan sebagai valid. Sementara itu, nilai True Positive (TP) sebesar 688 menunjukkan bahwa data hoaks berhasil diklasifikasikan dengan benar sebagai hoaks.

Berdasarkan hasil pengujian, model IndoBERT menunjukkan kemampuan klasifikasi yang sangat optimal dalam membedakan antara *fake news* dan valid, dengan jumlah *true positive* yang tinggi serta *false negative* yang rendah. Dibandingkan model baseline berbasis TF-IDF dan Logistic Regression, IndoBERT menghasilkan distribusi kesalahan yang lebih kecil, khususnya pada klasifikasi *fake news*. Hal ini menunjukkan bahwa arsitektur *transformer* mampu memahami konteks semantik dan hubungan antar kata secara lebih mendalam melalui *mekanisme contextual embedding*, sehingga performa klasifikasi menjadi lebih akurat, stabil, dan efektif pada data uji. Visualisasi Distribusi Label ditunjukkan pada Gambar 6.

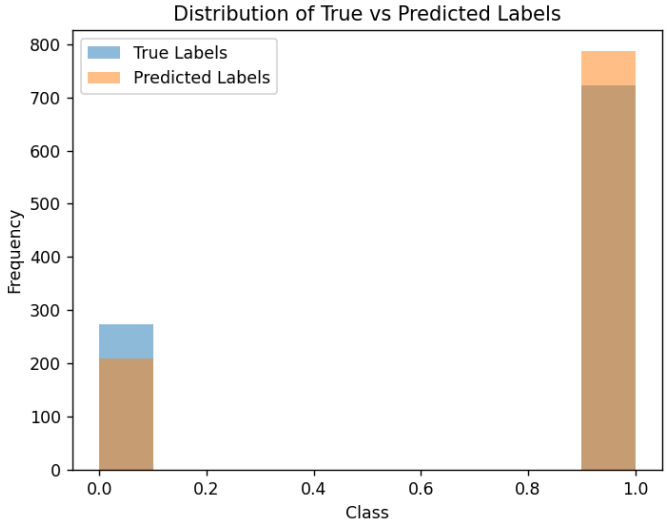
Berdasarkan grafik *Distribution of True vs Predicted Labels* pada gambar 6, terlihat bahwa distribusi label prediksi yang dihasilkan oleh model IndoBERT memiliki pola yang cukup mendekati distribusi label sebenarnya. Pada kelas valid (label 0), jumlah data aktual berada pada kisaran 273 data, sedangkan model

memprediksi sekitar 209 data sebagai valid. Sementara itu, pada kelas hoaks (label 1), jumlah data aktual sekitar 723 data dan hasil prediksi model mencapai sekitar 787 data. Perbedaan jumlah tersebut menunjukkan bahwa model cenderung lebih banyak memprediksi data ke dalam kelas hoaks dibandingkan kelas valid.

Kondisi ini menunjukkan bahwa model memiliki *sensitivitas* yang tinggi terhadap pendeteksian *fake news*, sehingga sebagian besar *fake news* berhasil dikenali dengan baik. Akan tetapi, kecenderungan prediksi yang lebih banyak pada kelas hoaks juga menunjukkan adanya beberapa data valid yang terklasifikasi secara keliru sebagai hoaks atau *false positive*. Secara keseluruhan, grafik ini menunjukkan bahwa model IndoBERT dapat mempertahankan distribusi prediksi yang hampir menyerupai distribusi data actuality, menandakan kemampuan generalisasi yang memadai dalam tugas klasifikasi *fake news*.

Visualisasi Menggunakan SHAP

SHAP (*SHapley Additive exPlanations*) digunakan untuk mengetahui kata-kata yang paling memengaruhi keputusan model IndoBERT dalam mengklasifikasikan berita hoaks dan valid. Visualisasi SHAP membantu menjelaskan alasan model memberikan prediksi pada suatu teks.



Gambar 6. Distribusi True Label dan Predicted Label

Berdasarkan visualisasi SHAP pada Gambar 7, analisis interpretabilitas dilakukan terhadap dua sampel data uji, yaitu satu sampel *fake news* dan satu sampel berita valid yang diambil dari dataset pengujian (*test_dataset.csv*). Pemilihan sampel dilakukan secara otomatis menggunakan data pertama pada masing-masing kelas (label == 1 untuk hoaks dan label == 0 untuk valid). Oleh karena itu, sampel yang digunakan bersifat r*epresentatif* awal dari masing-masing kategori, namun belum mewakili keseluruhan variasi data dalam dataset. Untuk mempercepat proses interpretasi dan visualisasi, teks dipotong menjadi 25 kata pertama dan nilai SHAP dihitung menggunakan parameter *max_evals=100*.

Visualisasi SHAP menunjukkan kontribusi setiap token terhadap hasil prediksi model IndoBERT. Cara membaca visualisasi ini adalah token berwarna biru menunjukkan kata yang meningkatkan keyakinan model terhadap kelas prediksi yang sedang ditampilkan, sedangkan token berwarna merah muda menunjukkan kata yang menurunkan keyakinan model terhadap prediksi tersebut. Semakin panjang panah pada token, maka semakin besar pengaruh token tersebut terhadap keputusan model.

Pada sampel pertama, token seperti *"menko", "indonesia", dan "kontroversi"* memberikan kontribusi besar terhadap prediksi model, sedangkan token seperti *"isu" dan "sara"* cenderung mengurangi keyakinan model terhadap kelas tersebut. Pada sampel kedua, token seperti *"guru", "lulusan", "madrasah",* dan *"baca"* menjadi kata yang paling dominan memengaruhi hasil klasifikasi. Hal ini menunjukkan bahwa model IndoBERT mampu menangkap hubungan kontekstual antar kata, bukan hanya frekuensi kemunculan kata seperti pada pendekatan TF-IDF.

Secara umum, pola kontribusi token yang ditemukan cukup konsisten dengan karakteristik data pada dataset, di mana kata-kata yang berkaitan dengan isu kontroversial, narasi provokatif, lowongan palsu, atau informasi sensasional cenderung memiliki pengaruh besar terhadap prediksi kelas hoaks. Namun, karena analisis SHAP pada penelitian ini hanya dilakukan pada beberapa sampel representatif, maka interpretasi yang dihasilkan bersifat lokal (*local interpretability*) dan belum sepenuhnya menggambarkan pola global seluruh

dataset. Oleh karena itu, analisis pada lebih banyak sampel dapat dilakukan pada penelitian selanjutnya untuk memperoleh pemahaman yang lebih menyeluruh mengenai perilaku model.

Gambar 7. Visualisasi SHAP

IV. Kesimpulan dan saran

Penelitian ini berhasil mengembangkan sistem deteksi *fake news* berbahasa Indonesia menggunakan model IndoBERT dan pendekatan (XAI) berbasis SHAP. Hasil *preprocessing* menunjukkan jumlah data awal sebanyak 5.347 data dan setelah proses *filtering* bahasa Indonesia menjadi 4.980 data. Selanjutnya, data dipilah menggunakan teknik *stratified sampling* dengan perbandingan 80% untuk bagian pelatihan dan 20% untuk bagian uji, serta diterapkan random *oversampling* pada data pelatihan guna menangani ketidakseimbangan kelas. Berdasarkan hasil pengujian, model baseline yang menggunakan TF-IDF dan Logistic Regression mencatat perolehan akurasi 77% dengan *weighted F1-score* sebesar 0,78. Di sisi lain, model IndoBERT menunjukkan hasil yang lebih superior dengan akurasi 87%, *precision* 0,87, *recall* 0,95, serta *F1-score* 0,91 pada kategori hoaks. Hasil confusion matrix juga menggambarkan bahwa IndoBERT dapat meminimalkan jumlah kesalahan klasifikasi bila dibandingkan dengan model baseline, khususnya pada nilai false negative. Visualisasi *training loss* dan *evaluation loss* menunjukkan bahwa proses fine-tuning IndoBERT berjalan dengan stabil tanpa indikasi *overfitting* yang signifikan. Selain itu, visualisasi SHAP membuktikan bahwa model mampu memberikan interpretasi terhadap prediksi dengan menunjukkan token-token yang paling berpengaruh dalam proses klasifikasi. Dengan demikian, integrasi IndoBERT dan SHAP terbukti efektif dalam meningkatkan performa deteksi *fake news* sekaligus menyediakan transparansi terhadap keputusan model. Penelitian mendatang dapat memanfaatkan kumpulan data yang lebih luas serta mencakup domain yang lebih bervariasi sehingga model dapat memiliki kemampuan generalisasi yang lebih optimal. Selain itu, dapat diterapkan *teknik balancing* lain seperti SMOTE untuk membandingkan pengaruh metode *oversampling* terhadap performa model. Pengembangan model juga dapat dilakukan dengan membandingkan IndoBERT dengan arsitektur *transformer* lain seperti RoBERTa atau XLM-RoBERTa. Dari sisi *interpretabilitas*, penelitian berikutnya dapat mengombinasikan SHAP dengan metode *Explainable AI* lainnya seperti LIME atau *attention visualization* untuk menghasilkan analisis yang lebih komprehensif.

Daftar Pustaka

[1] M. U. G. Khan, A. Mehmood, M. Elhadef, and S. A. Chaudhry, “Fake News Classification: Past, Current, and Future,” *Comput. Mater. Contin.*, vol. 77, no. 2, pp. 2225–2249, Nov. 2023, doi: 10.32604/CMC.2023.038303.

[2] A. Saeed and E. Al Solami, “Fake News Detection Using Machine Learning and Deep Learning Methods,” *Comput. Mater. Contin.*, vol. 77, no. 2, pp. 2079–2096, Nov. 2023, doi: 10.32604/CMC.2023.030551.

[3] H. D. Nguyen and T. Le, “Explainable Fake News Detection via Multi-Aspect Semantic Discovery,” *Procedia Comput. Sci.*, vol. 270, pp. 3211–3220, 2025, doi: 10.1016/j.procs.2025.09.446.

[4] C. Zhang, A. Gupta, X. Qin, and Y. Zhou, “A computational approach for real-time detection of fake news,” *Expert Syst. Appl.*, vol. 221, p. 119656, Jul. 2023, doi: 10.1016/J.ESWA.2023.119656.

[5] A. J. Dal Forno, G. P. Richetti, and V. H. Knaesel, “Fake news detection algorithms – A systematic literature review,” *Data Knowl. Eng.*, vol. 158, p. 102441, Jul. 2025, doi: 10.1016/J.DATAK.2025.102441.

[6] R. K. Gurjwar, A. Kumar, and U. P. Rao, "EPRVFL: A fast and scalable model for real-time fake news detection," *Pattern Recognit. Lett.*, vol. 196, pp. 267–273, Oct. 2025, doi: 10.1016/J.PATREC.2025.06.006.

[7] L. A. Pekandi, R. G. Widjaja, A. Ananta, J. Harefa, and K. Jingga, "Evaluating IndoBERT for Indonesian Hoax News Detection: A Comparative Study with Ensemble and CNN-LSTM Models," *Procedia Comput. Sci.*, vol. 269, pp. 1625–1633, 2025, doi: 10.1016/j.procs.2025.09.105.

[8] A. B. Athira, S. D. M. Kumar, and A. M. Chacko, "A systematic survey on explainable AI applied to fake news detection," *Eng. Appl. Artif. Intell.*, vol. 122, p. 106087, Jun. 2023, doi: 10.1016/J.ENGAPPAI.2023.106087.

[9] Z. Zhang *et al.*, "GBCA: Graph Convolution Network and BERT combined with Co-Attention for fake news detection," *Pattern Recognit. Lett.*, vol. 180, pp. 26–32, Apr. 2024, doi: 10.1016/J.PATREC.2024.02.014.

[10] K. Yu, S. Jiao, and Z. Ma, "Fake News Detection Based on BERT Multi-domain and Multi-modal Fusion Network," *Comput. Vis. Image Underst.*, vol. 252, p. 104301, Feb. 2025, doi: 10.1016/J.CVIU.2025.104301.

[11] Q. Zhang, X. Weng, G. Zhou, Y. Zhang, and J. X. Huang, "ARL: An adaptive reinforcement learning framework for complex question answering over knowledge base," *Inf. Process. Manag.*, vol. 59, no. 3, p. 102933, May 2022, doi: 10.1016/J.IPM.2022.102933.

[12] G. Folino, M. Guarascio, L. Pontieri, and P. Zicari, "Discovering ensembles of small language models out of scarcely labelled data for fake news detection," *Appl. Soft Comput.*, vol. 171, no. January, p. 112794, 2025, doi: 10.1016/j.asoc.2025.112794.

[13] M. Han, J. Li, Y. Chen, L. Xu, and L. Tao, "EC-Fake: A fake news detection model based on external knowledge and contrast-driven feature augmentation," *Neurocomputing*, vol. 653, p. 131214, Nov. 2025, doi: 10.1016/J.NEUCOM.2025.131214.

[14] M. Han, J. Li, S. Ou, L. Xu, and L. Tao, "ET-Fake: A multi-modal fake news detection model augmented with external knowledge and two-stage feature fusion," *Appl. Soft Comput.*, vol. 188, p. 114392, Feb. 2026, doi: 10.1016/J.ASOC.2025.114392.

[15] Amandeep and S. Suresh, "Transforming Fake News Detection: Leveraging DistilBERT Models for Enhanced Accuracy," *Procedia Comput. Sci.*, vol. 260, pp. 283–290, 2025, doi: 10.1016/j.procs.2025.03.203.

[16] B. Xie and Q. Li, "Detecting fake news by RNN-based gatekeeping behavior model on social networks," *Expert Syst. Appl.*, vol. 231, p. 120716, Nov. 2023, doi: 10.1016/J.ESWA.2023.120716.

[17] R. Anju and N. Pervin, "CredBERT: Credibility-aware BERT model for fake news detection," *Data Knowl. Eng.*, vol. 160, p. 102461, Nov. 2025, doi: 10.1016/J.DATAK.2025.102461.

[18] R. N. Tanaja, Johnny, M. A. Rafif, and A. A. S. Gunawan, "Fake News Detection using Machine Learning: Integrating FakeBERT Classification, Style Analysis, and Credibility Verification," *Procedia Comput. Sci.*, vol. 269, pp. 1067–1076, 2025, doi: 10.1016/j.procs.2025.09.048.

[19] A. Rahman, S. Zaman, S. Parvej, P. C. Shill, M. S. Salim, and D. Das, "Fake News Detection: Exploring the Efficiency of Soft and Hard Voting Ensemble," *Procedia Comput. Sci.*, vol. 252, pp. 748–757, 2025, doi: 10.1016/j.procs.2025.01.035.

[20] R. Mohawesh, X. Liu, H. M. Arini, Y. Wu, and H. Yin, "Semantic graph based topic modelling framework for multilingual fake news detection," *AI Open*, vol. 4, no. June, pp. 33–41, 2023, doi: 10.1016/j.aiopen.2023.08.004.

[21] E. Raja, B. Soni, and S. K. Borgohain, "Fake news detection in Dravidian languages using multiscale residual CNN_BiLSTM hybrid model," *Expert Syst. Appl.*, vol. 250, p. 123967, Sep. 2024, doi: 10.1016/J.ESWA.2024.123967.

[22] H. Jin and P. Wang, "Fake news detection on social media using triple-attention mechanism optimized by advanced tailor optimization algorithm," *Egypt. Informatics J.*, vol. 32, no. November, p. 100815, 2025, doi: 10.1016/j.eij.2025.100815.

[23] P. Dhiman, A. Kaur, D. Gupta, S. Juneja, A. Nauman, and G. Muhammad, "GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection," *Heliyon*, vol. 10, no. 16, p. e35865, 2024, doi: 10.1016/j.heliyon.2024.e35865.

[24] K. S. Sreekar Datta, G. Narasimha Naidu, S. Abhishek, and T. Anjali, "Enhancing Veracity: Empirical Evaluation of Fake News Detection Techniques," *Procedia Comput. Sci.*, vol. 233, pp. 97–107, Jan. 2024, doi: 10.1016/J.PROCS.2024.03.199.

[25] J. Alghamdi, Y. Lin, and S. Luo, "ABERT: Adapting BERT model for efficient detection of human and AI-generated fake news," *Int. J. Inf. Manag. Data Insights*, vol. 5, no. 2, p. 100353, 2025, doi: 10.1016/j.jjimei.2025.100353.