Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 3107~3117
e-ISSN: 2548-6861

Real-Time News Authenticity Verification Using MPNet (Masked and Permuted Pre-training Network)-Based Sentence Embeddings on Digital News Portals

Ira Lestari [1], Herlinah [2], M.Adnan Nur [3]***
[1,2,3]* Teknik Informatika, Universitas Handayani Makassar
lestaryira8@gmail.com [1], linaherlinah@handayani.ac.id [2], adnan@handayani.ac.id [3]

Article Info

Article history:

Received 2026-06-04
Revised 2026-06-21
Accepted 2026-06-22

Keyword:

Hoax Detection,
MPNet,
News Verification,
Semantic Similarity.

ABSTRACT

The dissemination of fake news (hoaxes) on digital news portals represents a significant challenge in the digital era, as it may mislead the public and reduce trust in circulating information. The rapid and open nature of digital media enables unverified information to spread widely within a short period of time, while manual verification processes require substantial time and effort. This study proposes a semantic similarity-based approach to support real-time news verification using the Multilingual MPNet model. The proposed approach utilizes content text as input, followed by keyword extraction using KeyBERT to represent the core information of the news. The extracted keywords are employed in a news scraping process to obtain comparative news articles from digital news portals. A dataset consisting of 200 Indonesian news articles, including 100 factual news articles and 100 hoax news articles, was used for evaluation. Subsequently, semantic similarity measurement is conducted to evaluate the degree of semantic relevance between the test news and the scraped news. Evaluation metrics were applied to assess the effectiveness of the proposed approach. The findings demonstrate that semantic text representation using Multilingual MPNet effectively supports hoax detection and provides relevant supporting evidence in the form of semantically related news articles, enabling users to access comparative news sources that support the verification process. Experimental results show that the proposed approach achieved an accuracy of 83.5%, precision of 97.18%, recall of 69.0%, F1-score of 80.70%, and an AUC of 0.695, indicating that Multilingual MPNet can effectively support news verification through semantic similarity analysis.

I. PENDAHULUAN

Perkembangan media digital di Indonesia mendorong peningkatan konsumsi berita melalui portal berita daring yang memungkinkan penyebaran informasi secara cepat dan luas kepada masyarakat [1]. Kondisi ini berdampak pada meningkatnya jumlah informasi yang beredar tanpa melalui proses verifikasi yang memadai, sehingga memicu penyebaran berita palsu atau hoaks [2]. Penyebaran hoaks menjadi permasalahan serius karena sering kali menyebar lebih cepat dibandingkan klarifikasi resmi, khususnya pada media digital yang bersifat terbuka [3]. Proses verifikasi berita yang masih dilakukan secara manual membutuhkan waktu dan sumber daya yang besar, sehingga kurang efektif dalam menghadapi tingginya arus informasi digital [4].

Berbagai penelitian sebelumnya telah memanfaatkan pendekatan pembelajaran mesin dan *deep learning* untuk mendeteksi hoaks melalui analisis teks berita digital [5]. Pendekatan tersebut umumnya menggunakan metode klasifikasi berbasis *fine-tuning* model bahasa, yang memerlukan data latih berlabel dalam jumlah besar serta menghasilkan keluaran berupa label tanpa disertai penjelasan semantik yang memadai [6]. Keterbatasan ini menyebabkan hasil deteksi sulit diinterpretasikan oleh pengguna karena tidak menyediakan informasi pendukung yang menjelaskan kesesuaian makna antar teks [7].

Dalam beberapa tahun terakhir, model bahasa berbasis *Transformer* menunjukkan kemampuan yang baik dalam memahami konteks semantik teks secara mendalam [8]. Salah satu model yang banyak digunakan untuk tugas *semantic similarity* dan *information retrieval* adalah MPNet *(Masked and Permuted Pre-training Network)*. Model ini menggabungkan keunggulan *Masked Language Modeling* dan *Permuted Language Modeling* sehingga mampu menghasilkan representasi semantik yang lebih efektif dalam memahami konteks teks [9]. Selain itu, penggunaan *Sentence Transformer* memungkinkan kalimat direpresentasikan dalam bentuk *embedding* yang dapat dibandingkan secara langsung menggunakan *cosine similarity* untuk mengukur tingkat kemiripan semantik antar dokumen [10].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mendeteksi dan memverifikasi keaslian berita digital menggunakan pendekatan berbasis *semantic similarity*. Pendekatan ini mengukur tingkat kemiripan makna antara teks berita uji dan berita pembanding berdasarkan representasi semantik teks. Melalui pengukuran kemiripan semantik, hasil deteksi tidak hanya berupa keputusan akhir, tetapi juga dilengkapi dengan *evidence* yang membantu pengguna memahami tingkat kesesuaian makna antar dokumen. Penelitian ini diharapkan dapat meningkatkan efektivitas proses verifikasi berita serta membantu mengurangi penyebaran hoaks pada portal berita digital.

Dalam konteks penelitian berita digital, terdapat perbedaan mendasar antara fact-checking, hoax detection, dan semantic similarity. Fact-checking merupakan proses verifikasi kebenaran suatu informasi dengan membandingkannya terhadap sumber primer, data resmi, atau bukti yang dapat dipertanggungjawabkan [11]. Hoax detection berfokus pada proses identifikasi apakah suatu informasi termasuk kategori hoaks atau fakta berdasarkan karakteristik tertentu yang dipelajari oleh sistem [12]. Sementara itu, semantic similarity bertujuan mengukur tingkat kesamaan makna antara dua teks tanpa secara langsung menentukan kebenaran informasi yang terkandung di dalamnya [13]. Penelitian ini memanfaatkan pendekatan semantic similarity sebagai mekanisme untuk menemukan berita pembanding yang memiliki keterkaitan makna dengan berita uji sehingga dapat digunakan sebagai pendukung proses verifikasi informasi.

Dalam penelitian ini, istilah keaslian berita didefinisikan sebagai tingkat kesesuaian informasi suatu berita dengan informasi yang ditemukan pada berita pembanding dari sumber yang digunakan dalam penelitian. Berita dikategorikan sebagai fakta apabila memiliki keterkaitan semantik yang tinggi dengan berita dari sumber berita arus utama atau sumber klarifikasi fakta yang kredibel. Sebaliknya, berita dikategorikan sebagai hoaks apabila tidak ditemukan dukungan informasi yang memadai dari sumber pembanding atau memiliki keterkaitan dengan berita yang telah diklasifikasikan sebagai hoaks oleh sumber referensi yang digunakan. Definisi ini digunakan sebagai definisi operasional dalam ruang lingkup penelitian dan tidak dimaksudkan sebagai pengganti proses verifikasi fakta secara menyeluruh terhadap sumber primer [11].

Penelitian ini tidak mengasumsikan bahwa kemiripan semantik yang tinggi secara otomatis menjamin kebenaran suatu berita. Kemiripan semantik digunakan sebagai indikator untuk menemukan informasi pembanding yang relevan dari sumber yang dipilih berdasarkan tingkat kredibilitasnya. Oleh karena itu, hasil verifikasi yang dihasilkan sistem perlu dipahami sebagai dukungan terhadap proses evaluasi informasi dan bukan sebagai bukti mutlak mengenai kebenaran suatu berita. Pemilihan sumber berita pembanding yang memiliki reputasi dan proses editorial yang jelas dilakukan untuk meminimalkan risiko penggunaan informasi yang tidak terverifikasi sebagai dasar pembandingan [11][12].

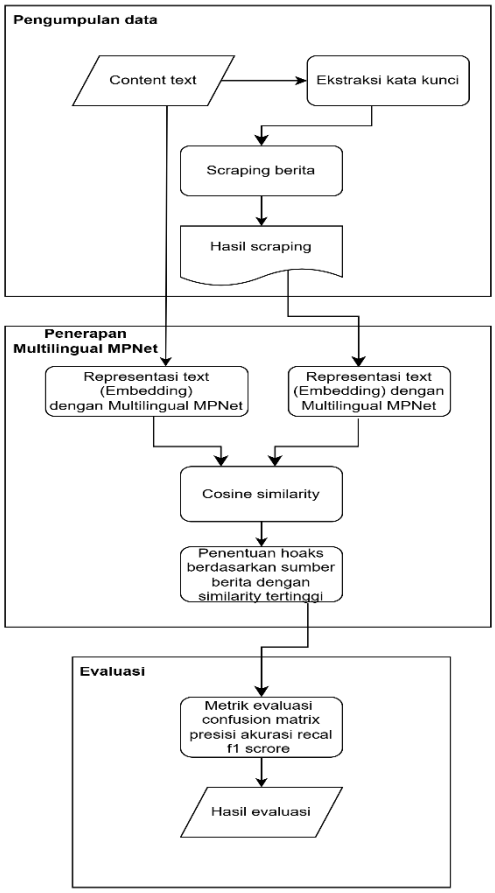
Berbagai pendekatan telah digunakan dalam penelitian deteksi hoaks, seperti klasifikasi berbasis *transformer*, *knowledge graph, claim verification,* maupun *retrieval-augmented fact-checking.* Pendekatan klasifikasi umumnya berfokus pada pemberian label fakta atau hoaks berdasarkan pola yang dipelajari dari data pelatihan [14]. Sementara itu, *claim verification* memanfaatkan bukti dan sumber eksternal untuk melakukan verifikasi terhadap suatu klaim [11]. Perkembangan terbaru juga menunjukkan penggunaan *retrieval-augmented fact-checking* yang menggabungkan proses pencarian informasi dan verifikasi berbasis bukti untuk meningkatkan kualitas verifikasi informasi [15]. Berbeda dengan pendekatan tersebut, penelitian ini memanfaatkan *semantic similarity* untuk menemukan berita pembanding yang relevan sehingga proses verifikasi dapat didukung oleh informasi pembanding yang memiliki keterkaitan makna dengan berita yang diuji.

Berbeda dengan penelitian sebelumnya yang umumnya berfokus pada klasifikasi hoaks menggunakan model pembelajaran mesin atau *transformer*, penelitian ini mengintegrasikan ekstraksi kata kunci menggunakan KeyBERT, proses *scraping* berita secara otomatis dari berbagai portal berita digital, serta pengukuran *semantic similarity* berbasis *Multilingual* MPNet untuk memperoleh berita pembanding yang relevan. Kebaruan penelitian ini terletak pada penyediaan *evidence* berupa berita pembanding yang diperoleh secara otomatis dan dianalisis berdasarkan tingkat kemiripan semantik. Dengan demikian, sistem yang diusulkan tidak hanya menghasilkan keputusan verifikasi berupa kategori fakta atau hoaks, tetapi juga menyediakan informasi pendukung yang memungkinkan pengguna menelusuri dasar hasil verifikasi yang diberikan.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengembangkan sistem verifikasi berita berbasis *semantic similarity* menggunakan *Multilingual* MPNet yang dikombinasikan dengan ekstraksi kata kunci menggunakan KeyBERT dan *scraping* berita otomatis dari portal berita digital. Sistem yang diusulkan diharapkan mampu membantu proses verifikasi informasi dengan menyediakan berita pembanding yang relevan secara semantik.

II. METODE

Penelitian ini menggunakan pendekatan penelitian berbasis analisis teks untuk mendukung proses deteksi hoaks dan verifikasi keaslian berita digital dengan tahapan berikut.



Gambar 1. Metode penelitian yang digunakan

Metode penelitian diawali dengan pengumpulan *content text* sebagai berita uji, kemudian dilakukan ekstraksi kata kunci menggunakan KeyBERT. Kata kunci tersebut digunakan dalam proses *scraping* berita untuk memperoleh berita pembanding dari portal berita digital. Selanjutnya, *content text* dan berita hasil scraping direpresentasikan ke dalam bentuk *embedding* menggunakan *Multilingual MPNet* dan dibandingkan menggunakan *cosine similarity*. Berita dengan nilai *similarity* tertinggi digunakan sebagai dasar penentuan label fakta atau hoaks. Kinerja sistem kemudian dievaluasi menggunakan *confusion matrix, accuracy, precision, recall,* dan *F1-score*.

A. Pengumpulan data

1) Content text

Tahap awal penelitian dimulai dari *content text*, yaitu teks berita yang digunakan sebagai data uji dalam proses deteksi hoaks. Dataset yang digunakan dalam penelitian ini terdiri dari 200 berita berbahasa Indonesia yang dikumpulkan pada tahun 2026 dan diperoleh dari tiga sumber utama, yaitu 100 berita dari portal pemeriksa fakta TurnBackHoax, 50 berita dari Kompas.com, dan 50 berita dari Detik.com.

TABEL I
DISTRIBUSI DATASET

Sumber Data	Jumlah	Label
TurnBackHoax	100	Hoaks
Detik.com	50	Fakta
Kompas.com	50	Fakta
Total	200	-

Penggunaan dataset berita digital sebagai objek penelitian telah banyak diterapkan dalam studi deteksi hoaks untuk merepresentasikan penyebaran informasi pada media daring [16]. Data dari TurnBackHoax digunakan sebagai referensi berita yang telah terverifikasi, sedangkan data dari Kompas dan Detik digunakan sebagai representasi berita dari portal berita digital arus utama.

Untuk meningkatkan variasi bentuk kalimat dan menghindari kemiripan literal antar teks, sebagian data dilakukan proses parafrase. Proses parafrase dilakukan menggunakan model bahasa berbasis generatif, yaitu ChatGPT (GPT-4), yang telah digunakan dalam berbagai tugas pemrosesan bahasa alami termasuk generasi dan parafrase teks [17]. Penggunaan model ini bertujuan untuk menghasilkan variasi struktur kalimat tanpa mengubah makna utama berita, sehingga tetap mempertahankan konteks semantik teks. Proses parafrase dilakukan secara terkontrol dengan memastikan bahwa informasi inti dalam berita tidak mengalami distorsi makna.

2) Ekstraksi kata kunci

Setelah *content text* diperoleh, tahap berikutnya adalah ekstraksi kata kunci untuk merepresentasikan informasi utama dari berita. Ekstraksi dilakukan menggunakan metode KeyBERT, yaitu pendekatan berbasis *embedding* yang memanfaatkan representasi semantik teks untuk mengidentifikasi kata atau frasa yang paling relevan terhadap dokumen [18].

KeyBERT bekerja dengan menghitung kemiripan antara *embedding* dokumen dan *embedding* kandidat kata kunci menggunakan representasi berbasis model bahasa. Dalam penelitian ini, tiga kata kunci utama diekstraksi dari setiap berita untuk digunakan sebagai dasar pencarian berita pembanding pada tahap berikutnya. Ekstraksi kata kunci bertujuan untuk memfokuskan proses pencarian pada istilah yang paling merepresentasikan isi berita.

KeyBERT dipilih karena mampu menghasilkan kata kunci yang mempertimbangkan konteks semantik dokumen sehingga lebih representatif dibandingkan metode berbasis frekuensi kata. karena relevansi kata kunci yang dihasilkan akan memengaruhi kualitas dan keterkaitan berita yang diperoleh. Dengan memanfaatkan representasi semantik berbasis *embedding*, KeyBERT dapat menghasilkan kata kunci yang lebih sesuai dengan topik utama berita.

3) Scraping berita

Tahap selanjutnya adalah proses *scraping* berita dari portal berita digital menggunakan kata kunci hasil ekstraksi. Proses scraping dilakukan dengan memanfaatkan teknik pengambilan data berbasis *HTTP request* dan *parsing HTML* untuk memperoleh elemen penting berita seperti judul, isi berita, tanggal publikasi, dan *URL* sumber [19].

Untuk setiap satu *content text* yang diuji, dilakukan proses *scraping* maksimal 10 berita per sumber portal berita. Dengan demikian, jika digunakan tiga sumber berita daring sebagai pembanding, maka satu berita uji akan dibandingkan dengan maksimal 30 berita hasil *scraping*. Proses scraping dilakukan setiap kali berita uji dimasukkan ke dalam sistem, sehingga berita pembanding yang diperoleh menyesuaikan informasi yang tersedia pada portal berita digital saat proses verifikasi berlangsung. Pembatasan jumlah berita ini bertujuan untuk menjaga efisiensi proses komputasi serta memastikan bahwa berita pembanding tetap relevan dengan kata kunci yang dihasilkan.

Dalam penelitian ini, istilah *real-time* mengacu pada kemampuan sistem untuk melakukan proses pencarian berita pembanding, pembentukan representasi *embedding*, dan pengukuran kemiripan semantik secara langsung setelah berita uji dimasukkan oleh pengguna tanpa memerlukan proses pelatihan ulang model. Dengan mekanisme tersebut, hasil verifikasi dapat diperoleh secara otomatis berdasarkan informasi yang tersedia pada saat proses pengujian dilakukan.

Berita-berita hasil *scraping* kemudian digunakan sebagai data pembanding dalam tahap pengukuran *semantic similarity* untuk menilai tingkat kesesuaian makna antara berita uji dan berita dari portal berita digital. Pendekatan pengukuran kemiripan semantik berbasis representasi *embedding* telah banyak digunakan dalam berbagai tugas pemrosesan bahasa alami, khususnya untuk mengevaluasi kesesuaian makna antar dokumen teks secara kontekstual [20].

Portal berita yang digunakan dipilih berdasarkan ketersediaan informasi, tingkat kredibilitas sumber, serta kemudahan akses data untuk proses *scraping*. Kompas.com dan Detik.com digunakan sebagai representasi media berita arus utama yang memiliki proses editorial, sedangkan TurnBackHoax digunakan sebagai sumber referensi berita hoaks dan klarifikasi fakta. Untuk menjaga relevansi berita pembanding, hanya berita yang mengandung kata kunci hasil ekstraksi dan memiliki keterkaitan topik dengan berita uji yang digunakan dalam proses pengukuran *semantic similarity*.

Selain proses pengambilan data, hasil *scraping* juga menunjukkan bahwa sistem berhasil memperoleh berita pembanding dari beberapa portal berita digital berdasarkan kata kunci hasil ekstraksi. Portal berita yang digunakan dalam penelitian ini meliputi Detik.com, Kompas.com, dan TurnBackHoax. Setiap berita hasil *scraping* terdiri atas informasi judul berita, isi berita, tanggal publikasi, *URL* sumber, serta label kategori berita.

Pada portal Detik.com dan Kompas.com, berita hasil *scraping* digunakan sebagai representasi berita fakta dari media digital arus utama. Sementara itu, berita dari TurnBackHoax digunakan sebagai referensi berita hoaks dan klarifikasi fakta. Seluruh berita hasil *scraping* kemudian digunakan sebagai data pembanding dalam proses pengukuran *semantic similarity* menggunakan *Multilingual MPNet*. Berikut contoh data hasil *scraping*.

TABEL II
HASIL SCRAPING BERITA PEMBANDING

No	Judul Berita Pembanding	Portal Berita	Label Asli
1	UI Lakukan Riset Program MBG di 5 SD Jakarta, Temuannya Mengejutkan!	Detik.com	Fakta
2	*Double Degree* ITS 2026 Resmi Dibuka, Bisa Daftar dengan Beasiswa Garuda!	Kompas.com	Fakta
3	Video yang disebut memperlihatkan Israel dilanda banjir setelah serangan Iran.	TurnBackHoax	Hoaks

B. Penerapan Multilingual MPNet

1) Representasi Teks menggunakan MPNet

Pada tahap ini, *content text* dan berita hasil *scraping* direpresentasikan ke dalam bentuk vektor numerik (*embedding*) menggunakan model bahasa *Multilingual MPNet*. Representasi teks berbasis *sentence embedding* telah banyak digunakan dalam berbagai tugas pemrosesan bahasa alami seperti *semantic similarity*, information retrieval, dan pencocokan dokumen. Pendekatan ini memungkinkan teks direpresentasikan ke dalam bentuk vektor numerik sehingga hubungan semantik antar dokumen dapat diukur menggunakan *cosine similarity* [21].

Dalam penelitian ini digunakan model *paraphrase-multilingual-mpnet-base-v2* melalui pustaka *Sentence Transformers* untuk menghasilkan representasi semantik teks. Pendekatan berbasis *sentence embedding* memungkinkan dokumen direpresentasikan ke dalam bentuk vektor sehingga tingkat kemiripan semantik antar dokumen dapat diukur menggunakan *cosine similarity* [22]. Model MPNet telah banyak digunakan pada tugas *semantic similarity* dan *information retrieval* karena mampu menghasilkan representasi semantik yang baik pada tingkat kalimat maupun dokumen [23]. Selain itu, model *multilingual* berbasis MPNet juga menunjukkan performa yang kompetitif pada berbagai tugas *semantic textual similarity* lintas bahasa sehingga cocok digunakan untuk membandingkan dokumen dengan konteks semantik yang beragam [24].

Dalam penelitian ini, model *Multilingual* MPNet digunakan dalam bentuk *pre-trained* melalui model *paraphrase-multilingual-mpnet-base-v2* tanpa proses *fine-tuning* tambahan pada dataset berita berbahasa Indonesia. Pemanfaatan model *pre-trained* dilakukan karena model tersebut telah dilatih pada korpus multibahasa dalam skala besar dan mampu menghasilkan representasi semantik yang baik untuk berbagai tugas pemrosesan bahasa alami, termasuk pengukuran kemiripan teks.

Representasi *embedding* memungkinkan teks dengan makna yang serupa memiliki jarak vektor yang lebih dekat dalam ruang semantik.

Dalam proses verifikasi digunakan nilai ambang batas *(threshold*) sebesar 0,55 untuk membantu menentukan tingkat kemiripan semantik antara berita uji dan berita pembanding. Nilai *threshold* tersebut diperoleh melalui serangkaian pengujian awal dengan membandingkan beberapa nilai ambang batas, yaitu 0,55, 0,66, 0,86, dan 0,88. Berdasarkan hasil evaluasi terhadap dataset penelitian, nilai 0,55 menghasilkan performa klasifikasi terbaik dibandingkan nilai *threshold* lainnya, sehingga digunakan pada proses pengujian akhir. Penggunaan *threshold* bertujuan untuk membantu mengidentifikasi tingkat keterkaitan semantik antara berita uji dan berita pembanding yang diperoleh melalui proses *scraping.*

2) Pengukuran cosine similarity

Setelah diperoleh representasi *embedding* dari *content text* dan berita hasil *scraping*, tahap berikutnya adalah pengukuran tingkat kemiripan menggunakan metode *cosine similarity*. *Cosine similarity* merupakan metode yang mengukur kesamaan dua vektor berdasarkan sudut kosinus di antara keduanya dalam ruang vektor [25].

Secara matematis, *cosine similarity* dihitung menggunakan persamaan berikut:

$$Cosine\ Similarity = \frac{A \cdot B}{|A||B|}$$

Nilai *cosine similarity* berada pada rentang -1 hingga 1, namun dalam konteks embedding teks umumnya berada pada rentang 0 hingga 1. Semakin mendekati nilai 1, semakin tinggi tingkat kesamaan makna antara dua teks. Pendekatan ini banyak digunakan dalam tugas *semantic textual similarity* untuk membandingkan makna antar dokumen atau kalimat secara kontekstual [26].

C. Metrik evaluasi

Tahap akhir penelitian adalah evaluasi hasil deteksi hoaks menggunakan metrik evaluasi berbasis klasifikasi untuk menilai kinerja pendekatan *semantic similarity* dalam mengidentifikasi keaslian berita. Evaluasi dilakukan menggunakan *confusion matrix*, yang digunakan untuk menganalisis hasil prediksi terhadap data aktual dalam bentuk *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [27].

Berdasarkan *confusion matrix*, dihitung beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* digunakan untuk mengukur proporsi keseluruhan prediksi yang benar terhadap total data uji. *Precision* mengukur tingkat ketepatan prediksi positif terhadap seluruh prediksi positif yang dihasilkan model. *Recall* mengukur kemampuan pendekatan dalam mendeteksi seluruh data positif yang sebenarnya ada dalam dataset. Sementara itu, *F1-score* merupakan rata-rata harmonis antara precision dan recall yang digunakan untuk menilai keseimbangan kinerja model dalam tugas klasifikasi [28].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Nilai masing-masing metrik berada pada rentang 0 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan performa deteksi yang semakin baik. Penggunaan kombinasi metrik tersebut memberikan evaluasi yang lebih komprehensif dibandingkan hanya menggunakan satu metrik, terutama dalam kasus klasifikasi biner seperti deteksi hoaks.

III. HASIL DAN PEMBAHASAN

A. Hasil Scraping Berita

Proses scraping dilakukan menggunakan kata kunci yang diperoleh dari hasil ekstraksi KeyBERT pada setiap berita uji. Kata kunci tersebut digunakan untuk melakukan pencarian berita pembanding pada portal Detik.com, Kompas.com, dan TurnBackHoax. Hasil *scraping* menunjukkan bahwa sistem berhasil memperoleh berita pembanding dari ketiga sumber tersebut. Berita yang diperoleh selanjutnya digunakan sebagai data pembanding dalam proses pengukuran *semantic similarity* untuk menentukan tingkat kemiripan antara berita uji dan berita hasil *scraping*.

B. Hasil Semantic Similarity

Setelah proses *scraping* dilakukan menggunakan kata kunci hasil ekstraksi KeyBERT, berita pembanding yang diperoleh dari portal Detik.com, Kompas.com, dan TurnBackHoax digunakan dalam proses pengukuran *semantic similarity*. Setiap berita pembanding direpresentasikan ke dalam bentuk embedding menggunakan model *Multilingual MPNet*, kemudian dihitung nilai *cosine similarity* terhadap berita uji. Berita dengan nilai *similarity* tertinggi dipilih sebagai hasil terbaik dan digunakan sebagai dasar dalam proses penentuan label prediksi berita.

Hasil pengujian menunjukkan bahwa sistem mampu menemukan berita pembanding yang memiliki keterkaitan makna dengan berita uji meskipun terdapat perbedaan susunan kata dan struktur kalimat. Kemampuan ini menunjukkan bahwa representasi semantik yang dihasilkan oleh *Multilingual MPNet* dapat digunakan untuk mengidentifikasi kesesuaian makna antar berita secara kontekstual. Berikut merupakan contoh hasil pengukuran

semantic similarity antara berita uji dan berita pembanding yang diperoleh dari proses *scraping*.

TABEL III
HASIL SEMANTIC SIMILARITY

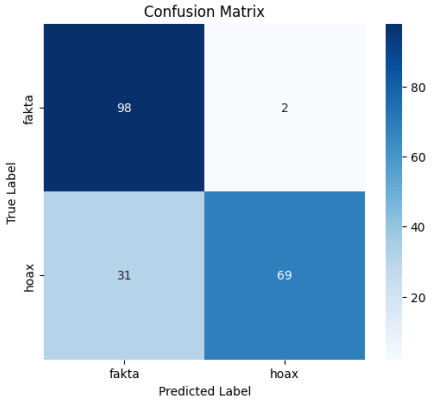
No	Judul Asli	Berita Pembanding	Portal Berita	Similarity Score	Label Prediksi
1	Riset yang dilakukan UI pada program MBG di 5 sekolah dasar Jakarta menghasilkan temuan yang mengejutkan.	UI Lakukan Riset Program MBG di 5 SD Jakarta, Temuannya Mengejutkan!	Detik.com	0.7601	Fakta
2	ITS membuka program *double degree* melalui Beasiswa Garuda 2026 yang memungkinkan kuliah tanpa biaya.	*Double Degree* ITS 2026 Resmi Dibuka, Bisa Daftar dengan Beasiswa Garuda!	Kompas.com	0.7731	Fakta
3	Video yang disebut memperlihatkan Israel dilanda banjir setelah serangan Iran.	Dokumentasi "Israel Banjir Akibat Serangan Iran"	TurnBackHoax	0.8720	Hoaks

Berdasarkan hasil pada Tabel III, sistem berhasil menemukan berita pembanding yang relevan dengan topik berita uji. Berita yang berasal dari portal TurnBackHoax menghasilkan label prediksi hoaks, sedangkan berita yang berasal dari Kompas.com dan Detik.com menghasilkan label prediksi fakta. Nilai *similarity score* yang relatif tinggi menunjukkan bahwa model *Multilingual MPNet* mampu menangkap hubungan semantik antar berita dengan baik sehingga dapat digunakan dalam proses verifikasi berita digital secara otomatis.

Selain itu, hasil pengujian menunjukkan bahwa pendekatan *semantic similarity* tidak hanya membandingkan kesamaan kata secara literal, tetapi juga mempertimbangkan kesesuaian konteks dan makna antar berita. Dengan demikian, berita yang memiliki informasi serupa meskipun ditulis menggunakan struktur kalimat yang berbeda tetap dapat dikenali sebagai berita yang relevan.

C. *Confusion Matrix*

Untuk mengevaluasi performa sistem dalam mengklasifikasikan berita fakta dan hoaks, dilakukan pengujian menggunakan *confusion matrix* berdasarkan hasil prediksi yang diperoleh dari proses *semantic similarity*. Evaluasi dilakukan terhadap 200 data berita yang terdiri atas 100 berita fakta dan 100 berita hoaks.



Gambar 2. *Confusion Matrix*

Berdasarkan hasil pengujian diperoleh nilai *True Negative* (TN) sebanyak 98 data, *False Positive* (FP) sebanyak 2 data, *False Negative* (FN) sebanyak 31 data, dan *True Positive* (TP) sebanyak 69 data. Hasil tersebut menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh sistem.

Nilai *True Negative* sebesar 98 menunjukkan bahwa hampir seluruh berita fakta berhasil diprediksi dengan benar sebagai fakta. Selain itu, nilai *True Positive* sebesar 69 menunjukkan bahwa sistem mampu mengidentifikasi sebagian besar berita hoaks dengan benar. Sementara itu, terdapat 2 berita fakta yang diprediksi sebagai hoaks dan 31 berita hoaks yang diprediksi sebagai fakta.

Jumlah *False Negative* yang mencapai 31 data menunjukkan bahwa masih terdapat sejumlah berita hoaks yang belum berhasil dikenali oleh sistem. Kondisi ini terjadi karena beberapa berita hoaks memiliki topik, kata kunci, dan konteks yang sangat mirip dengan berita fakta yang diperoleh dari proses *scraping*. Akibatnya, nilai *semantic similarity* yang dihasilkan antara berita hoaks dan berita pembanding fakta menjadi tinggi sehingga sistem cenderung memberikan prediksi fakta.

Selain itu, proses penentuan label prediksi berdasarkan berita dengan nilai *similarity* tertinggi juga memengaruhi hasil klasifikasi. Pada beberapa kasus, sistem lebih mudah menemukan berita fakta yang relevan dari portal Kompas.com dan Detik.com dibandingkan berita klarifikasi hoaks dari TurnBackHoax. Kondisi tersebut menyebabkan berita hoaks memperoleh pasangan berita pembanding dari sumber fakta dengan nilai kemiripan yang lebih tinggi sehingga diklasifikasikan sebagai fakta.

Berdasarkan hasil pengamatan terhadap data yang mengalami kesalahan klasifikasi, beberapa kasus *False Negative* ditemukan pada berita hoaks yang memiliki topik serupa dengan berita fakta, seperti isu pendidikan, bantuan sosial, dan kebijakan pemerintah. Kondisi tersebut menyebabkan nilai *semantic similarity* yang tinggi terhadap berita pembanding dari sumber fakta sehingga sistem menghasilkan prediksi fakta. Karena memiliki konteks yang sangat mirip dengan berita fakta, model *Multilingual MPNet* menghasilkan nilai kemiripan yang tinggi terhadap berita

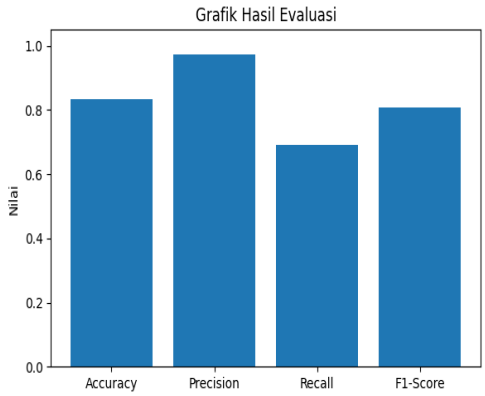
pembanding sehingga sistem memberikan prediksi fakta meskipun label sebenarnya adalah hoaks.

Di sisi lain, jumlah *False Positive* yang hanya sebanyak 2 data menunjukkan bahwa sistem memiliki kemampuan yang sangat baik dalam menghindari kesalahan klasifikasi berita fakta sebagai hoaks. Hal ini menunjukkan bahwa pendekatan *semantic similarity* berbasis *Multilingual MPNet* mampu mempertahankan tingkat ketepatan yang tinggi dalam mengidentifikasi berita fakta.

Secara keseluruhan, hasil *confusion matrix* menunjukkan bahwa pendekatan *semantic similarity* berbasis *Multilingual MPNet* mampu membedakan berita fakta dan hoaks dengan baik. Meskipun masih terdapat sejumlah berita hoaks yang diprediksi sebagai fakta, sistem berhasil menghasilkan jumlah prediksi benar yang jauh lebih tinggi dibandingkan prediksi yang salah, sehingga dapat digunakan untuk mendukung proses verifikasi berita digital secara otomatis.

D. Evaluasi Kinerja Sistem

Untuk mengetahui performa sistem dalam melakukan verifikasi berita, dilakukan evaluasi menggunakan metrik *accuracy, precision, recall,* dan *F1-score* berdasarkan hasil klasifikasi yang diperoleh dari *confusion matrix*. Hasil evaluasi kinerja sistem ditunjukkan pada gambar 3.



Gambar 3. Hasil Grafik Evaluasi

Pengujian dilakukan terhadap 200 berita yang terdiri atas 100 berita fakta dan 100 berita hoaks. Setiap berita diproses melalui tahapan ekstraksi kata kunci, *scraping* berita pembanding, pembentukan *embedding* menggunakan *Multilingual* MPNet, dan pengukuran *semantic similarity*. Berita pembanding dengan nilai *similarity* tertinggi digunakan sebagai dasar dalam proses penentuan hasil verifikasi.

Validasi dilakukan dengan membandingkan label prediksi yang dihasilkan sistem terhadap label aktual (*ground truth*) pada dataset penelitian. Hasil perbandingan tersebut kemudian digunakan untuk membentuk *confusion matrix* sebagai dasar perhitungan metrik evaluasi.

Penelitian ini tidak menggunakan proses pelatihan model (*training*) maupun pembagian data *training* dan *testing* karena *Multilingual* MPNet digunakan dalam bentuk *pre-trained*. Seluruh dataset digunakan sebagai data evaluasi untuk mengukur kemampuan sistem dalam melakukan verifikasi berita berdasarkan kemiripan semantik.

Berdasarkan *confusion matrix*, nilai *accuracy, precision, recall,* dan *F1-score* dihitung untuk mengukur performa sistem. *Accuracy* digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data, sedangkan *precision, recall,* dan *F1-score* digunakan untuk mengevaluasi kemampuan sistem dalam mendeteksi berita hoaks secara lebih rinci.

Berdasarkan hasil evaluasi pada Gambar 3, sistem memperoleh nilai *accuracy* sebesar 0,835 atau 83,5%. Hasil tersebut menunjukkan bahwa sistem mampu mengklasifikasikan sebagian besar data berita dengan benar berdasarkan kesesuaian makna antara berita uji dan berita hasil *scraping*. Nilai *accuracy* yang diperoleh menunjukkan bahwa pendekatan *semantic similarity* berbasis *Multilingual MPNet* dapat digunakan secara efektif dalam proses verifikasi berita digital.

Nilai *precision* sebesar 0,9718 menunjukkan bahwa sebagian besar berita yang diprediksi sebagai hoaks benar-benar merupakan berita hoaks. Tingginya nilai *precision* menunjukkan bahwa sistem memiliki tingkat kesalahan yang rendah dalam mengklasifikasikan berita fakta sebagai hoaks. Hasil ini sejalan dengan nilai *False Positive* yang hanya berjumlah 2 data pada *confusion matrix*.

Sementara itu, nilai *recall* sebesar 0,6900 menunjukkan bahwa sistem berhasil mendeteksi 69% dari seluruh berita hoaks yang terdapat pada dataset pengujian. Nilai *recall* yang lebih rendah dibandingkan *precision* menunjukkan bahwa masih terdapat beberapa berita hoaks yang belum berhasil dikenali oleh sistem. Kondisi ini dipengaruhi oleh adanya berita hoaks yang memiliki kemiripan semantik tinggi dengan berita fakta sehingga menghasilkan nilai *similarity* yang relatif besar terhadap berita pembanding dari sumber fakta.

Nilai *F1-score* sebesar 0,8070 menunjukkan bahwa sistem memiliki keseimbangan yang baik antara *precision* dan *recall*. Hasil tersebut mengindikasikan bahwa pendekatan yang digunakan tidak hanya mampu menghasilkan prediksi yang akurat, tetapi juga cukup konsisten dalam mendeteksi berita hoaks pada berbagai topik berita yang diuji.

Berdasarkan *confusion matrix*, sistem mampu mengklasifikasikan sebagian besar berita dengan benar pada kedua kelas. Nilai *True Positive* dan *True Negative* yang relatif tinggi menunjukkan bahwa pendekatan yang digunakan mampu membedakan berita fakta dan hoaks secara efektif. Meskipun demikian, masih terdapat sejumlah *False Negative* yang menunjukkan bahwa beberapa berita hoaks belum berhasil terdeteksi oleh sistem.

Selain itu, evaluasi juga dilakukan menggunakan ROC *Curve* dan *Area Under Curve* (AUC) untuk mengukur kemampuan sistem dalam membedakan berita fakta dan hoaks pada berbagai nilai *threshold*. Hasil ROC *Curve* menunjukkan nilai AUC sebesar 0,695, yang mengindikasikan bahwa sistem memiliki kemampuan diskriminasi yang baik dalam membedakan kedua kelas. Nilai

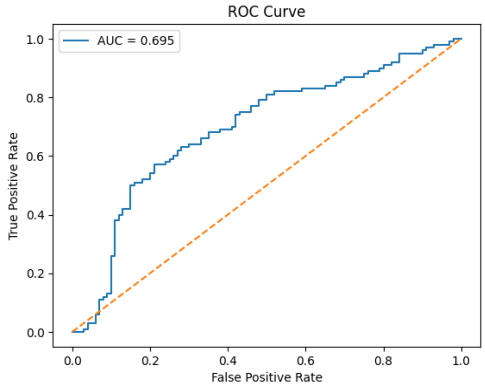
AUC yang diperoleh menunjukkan bahwa pendekatan *semantic similarity* berbasis *Multilingual* MPNet mampu memberikan performa klasifikasi yang stabil pada dataset penelitian.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa penggunaan model *Multilingual MPNet* sebagai representasi *embedding* teks mampu mendukung proses pengukuran *semantic similarity* secara efektif. Dengan tingkat *accuracy* sebesar 83,5%, *precision* sebesar 97,18%, *recall* sebesar 69,0%, dan *F1-score* sebesar 80,70%, sistem mampu memberikan performa yang baik dalam membedakan berita fakta dan hoaks berdasarkan kesesuaian makna antar berita.

E. ROC Curve dan AUC

Untuk melengkapi evaluasi kinerja sistem, dilakukan pengukuran menggunakan *Receiver Operating Characteristic* (ROC) *Curve* dan *Area Under Curve* (AUC). Pengukuran dilakukan berdasarkan nilai *similarity score* yang dihasilkan pada proses penghitungan kemiripan semantik antara berita uji dan berita hasil scraping.

Berdasarkan hasil pengujian diperoleh nilai AUC sebesar 0,695. Nilai tersebut menunjukkan bahwa sistem memiliki kemampuan yang cukup baik dalam membedakan berita fakta dan berita hoaks berdasarkan nilai *similarity score* yang dihasilkan oleh model *Multilingual* MPNet.



Gambar 4. ROC *Curve* Hasil Evaluasi

Kurva ROC pada Gambar 4 berada di atas garis diagonal yang merepresentasikan klasifikasi acak. Hal ini menunjukkan bahwa sistem memiliki kemampuan klasifikasi yang lebih baik dibandingkan prediksi secara acak dalam membedakan kedua kelas berita yang diuji.

Meskipun nilai AUC yang diperoleh belum mendekati nilai maksimum, hasil tersebut tetap menunjukkan bahwa pendekatan *semantic similarity* mampu menangkap perbedaan karakteristik antara berita fakta dan berita hoaks. Nilai AUC yang lebih rendah dibandingkan *accuracy* dipengaruhi oleh adanya sejumlah berita hoaks yang memiliki kemiripan semantik tinggi terhadap berita fakta sehingga menghasilkan nilai *similarity score* yang berdekatan. Kondisi ini juga terlihat pada *confusion matrix* yang menunjukkan masih terdapat 31 data *False Negative*.

Secara keseluruhan, hasil ROC *Curve* dan AUC memperkuat hasil evaluasi sebelumnya yang menunjukkan bahwa pendekatan semantic similarity berbasis *Multilingual MPNet* mampu digunakan untuk mendukung proses verifikasi berita digital secara otomatis dengan performa yang cukup baik.

Hasil penelitian menunjukkan bahwa kualitas berita pembanding yang diperoleh melalui proses *scraping* memiliki pengaruh terhadap performa sistem. Ketika jumlah berita pembanding yang ditemukan relatif sedikit atau memiliki keterkaitan topik yang rendah dengan berita uji, nilai *semantic similarity* yang dihasilkan cenderung menurun sehingga dapat memengaruhi hasil verifikasi. Selain itu, kualitas dan kredibilitas sumber berita juga berpengaruh terhadap kualitas informasi yang digunakan sebagai pembanding. Oleh karena itu, performa sistem sangat bergantung pada ketersediaan berita yang relevan dan kredibel pada saat proses verifikasi dilakukan.

F. Perbandingan Dengan Metode Baseline

Untuk menilai efektivitas pendekatan yang diusulkan, penelitian ini melakukan perbandingan dengan metode baseline *TF-IDF* yang dikombinasikan dengan *cosine similarity*. Metode *TF-IDF* dipilih karena merupakan salah satu pendekatan yang umum digunakan dalam pengukuran kemiripan teks berbasis representasi frekuensi kata. Perbandingan dilakukan menggunakan dataset dan skenario pengujian yang sama sehingga perbedaan performa yang diperoleh dapat menggambarkan kemampuan masing-masing metode dalam mendukung proses verifikasi berita. Hasil perbandingan performa kedua metode ditunjukkan pada Tabel IV.

TABEL IV
PERBANDINGAN HASIL EVALUASI TF-IDF DAN MULTILINGUAL MPNET

Metode	Accuracy	Precision	Recall	F1-Score
TF-IDF + Cosine Similarity	0.7035	0.9762	0.4141	0.5816
Multilingual MPNet	0.835	0.9718	0.69	0.8070

Untuk mengevaluasi efektivitas metode yang diusulkan, dilakukan perbandingan dengan metode *baseline TF-IDF* yang dikombinasikan dengan *cosine similarity* menggunakan dataset yang sama. Hasil pengujian menunjukkan bahwa metode *Multilingual* MPNet menghasilkan *accuracy* sebesar 83,5%, lebih tinggi dibandingkan metode *TF-IDF* yang memperoleh *accuracy* sebesar 70,35%.

Perbedaan performa yang paling signifikan terlihat pada nilai *recall*. Metode *TF-IDF* memperoleh *recall* sebesar 41,41%, sedangkan *Multilingual* MPNet mencapai *recall* sebesar 69,0%. Hasil tersebut menunjukkan bahwa *Multilingual* MPNet mampu mendeteksi lebih banyak berita hoaks dibandingkan *TF-IDF*. Hal ini disebabkan karena MPNet menghasilkan representasi semantik yang mampu

menangkap hubungan makna antar kalimat, sedangkan *TF-IDF* hanya mengandalkan kemunculan dan frekuensi kata.

Meskipun nilai *precision* kedua metode relatif tinggi, *Multilingual* MPNet menghasilkan keseimbangan yang lebih baik antara *precision* dan *recall* yang ditunjukkan oleh nilai *F1-score* sebesar 80,70%, lebih tinggi dibandingkan *TF-IDF* yang hanya mencapai 58,16%. Dengan demikian, pendekatan *semantic similarity* berbasis *Multilingual* MPNet terbukti lebih efektif dalam mendukung proses verifikasi berita dibandingkan metode berbasis *TF-IDF*.

G. Keterbatasan Penelitian

Pendekatan yang digunakan dalam penelitian ini juga memiliki keterbatasan pada kasus berita baru (*breaking news*) yang belum banyak dipublikasikan oleh media lain. Pada kondisi tersebut, sistem mungkin mengalami kesulitan menemukan berita pembanding yang relevan sehingga berpotensi menghasilkan hasil verifikasi yang kurang akurat. Akibatnya, berita yang sebenarnya valid dapat diklasifikasikan sebagai hoaks karena minimnya sumber pembanding yang tersedia pada saat proses verifikasi dilakukan. Oleh karena itu, hasil verifikasi pada berita yang masih berkembang perlu diinterpretasikan secara hati-hati dan tidak digunakan sebagai satu-satunya dasar dalam pengambilan keputusan.

Penerapan sistem verifikasi berita secara otomatis juga memiliki implikasi etis yang perlu diperhatikan. Kesalahan klasifikasi yang dihasilkan sistem dapat memengaruhi persepsi pengguna terhadap suatu informasi. Berita fakta yang diklasifikasikan sebagai hoaks berpotensi menurunkan kepercayaan terhadap informasi yang benar, sedangkan berita hoaks yang diklasifikasikan sebagai fakta dapat berkontribusi terhadap penyebaran informasi yang tidak akurat. Oleh karena itu, hasil verifikasi yang diberikan sistem sebaiknya digunakan sebagai alat bantu pendukung dan tetap memerlukan evaluasi lebih lanjut oleh pengguna sebelum dijadikan dasar pengambilan keputusan.

Penelitian ini juga memiliki keterbatasan yang berkaitan dengan ketergantungan terhadap ketersediaan berita pembanding dari portal berita digital. Karena proses verifikasi dilakukan berdasarkan berita yang diperoleh melalui *scraping*, jumlah dan kualitas berita pembanding yang tersedia dapat berubah dari waktu ke waktu sesuai dengan kondisi portal berita yang digunakan. Perubahan struktur situs, penghapusan berita, maupun keterbatasan ketersediaan berita dengan topik yang relevan berpotensi memengaruhi hasil verifikasi yang dihasilkan oleh sistem. Oleh karena itu, performa sistem dapat berbeda apabila diterapkan pada periode waktu atau sumber berita yang berbeda.

Penelitian ini juga belum secara khusus mengevaluasi ketahanan (*robustness*) sistem terhadap variasi gaya penulisan berita, penggunaan sinonim, parafrase, maupun informasi yang bersifat parsial. Meskipun *Multilingual* MPNet dirancang untuk menangkap kesamaan makna secara semantik, pengaruh variasi bahasa tersebut terhadap performa sistem masih perlu dikaji lebih lanjut melalui skenario pengujian yang lebih beragam.

IV. KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan *semantic similarity* berbasis *Multilingual MPNet* untuk mendukung proses deteksi dan verifikasi berita pada portal berita digital. Sistem yang dikembangkan memanfaatkan ekstraksi kata kunci menggunakan KeyBERT, proses *scraping* berita dari portal Detik.com, Kompas.com, dan TurnBackHoax, serta pengukuran *cosine similarity* untuk menentukan tingkat kesesuaian makna antara berita uji dan berita pembanding.

Berdasarkan hasil pengujian terhadap 200 data berita yang terdiri atas 100 berita fakta dan 100 berita hoaks, sistem memperoleh nilai *accuracy* sebesar 83,5%, *precision* sebesar 97,18%, *recall* sebesar 69,0%, *F1-score* sebesar 80,70%, dan AUC sebesar 0,695. Hasil tersebut menunjukkan bahwa pendekatan yang digunakan mampu mengidentifikasi berita fakta dan hoaks dengan tingkat ketepatan yang baik.

Hasil penelitian juga menunjukkan bahwa representasi *embedding* menggunakan *Multilingual MPNet* mampu menangkap hubungan semantik antar berita sehingga sistem tidak hanya membandingkan kesamaan kata secara literal, tetapi juga mempertimbangkan kesesuaian konteks dan makna antar dokumen. Selain menghasilkan prediksi fakta atau hoaks, sistem mampu menyediakan berita pembanding yang relevan sebagai pendukung proses verifikasi berita.

Meskipun demikian, hasil penelitian ini perlu dipahami sesuai dengan ruang lingkup eksperimen yang dilakukan. Pendekatan yang digunakan berfokus pada pengukuran kemiripan semantik antara berita uji dan berita pembanding yang diperoleh dari portal berita digital, sehingga hasil verifikasi yang diberikan belum dapat sepenuhnya menggantikan proses verifikasi fakta secara menyeluruh terhadap sumber primer atau data resmi. Oleh karena itu, sistem lebih tepat digunakan sebagai alat bantu untuk mendukung proses verifikasi informasi berdasarkan kesesuaian makna antar berita. Keterbatasan ini menyebabkan sistem masih berpotensi menghasilkan kesalahan klasifikasi ketika informasi yang dibandingkan memiliki kemiripan konteks yang tinggi namun belum tentu memiliki tingkat kebenaran yang sama.

Penelitian ini mengimplementasikan proses verifikasi berita secara *real-time* melalui mekanisme *scraping* dan pencarian berita pembanding secara langsung pada saat berita diuji. Namun demikian, penelitian ini belum melakukan pengukuran kuantitatif terhadap waktu proses pada setiap tahapan sistem, seperti waktu *scraping*, waktu pembentukan *embedding*, waktu pengukuran kemiripan semantik, maupun total waktu respons sistem. Evaluasi performa waktu secara rinci akan menjadi bagian dari pengembangan penelitian selanjutnya.

Meskipun demikian, masih terdapat sejumlah berita hoaks yang diprediksi sebagai fakta akibat tingginya kemiripan semantik dengan berita pembanding yang berasal dari sumber

berita fakta. Kondisi ini menunjukkan bahwa pendekatan *semantic similarity* masih memiliki keterbatasan dalam membedakan kebenaran informasi ketika dua berita memiliki kesamaan konteks yang tinggi.

DAFTAR PUSTAKA

[1] L. Triyono, R. Gernowo, M. Rahaman, and T. R. Yudantoro, "International Journal On Informatics Visualization journal homepage : www.joiv.org/index.php/joiv International Journal On Informatics Visualization Indonesian Fake News Detection Using Various Machine Learning Technique," Sep. 2023. [Online]. Available: www.joiv.org/index.php/joiv

[2] V. Prisscilya and A. S. Girsang, "Classification of Indonesia False News Detection Using Bertopic and Indobert," *Jurnal Indonesia Sosial Teknologi*, vol. 5, no. 8, 2024, [Online]. Available: http://jist.publikasiindonesia.id/

[3] E. Effendi, "User behaviour and hoax information on social media case of Indonesia," *Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies)*, vol. 7, no. 3, pp. 930–943, Nov. 2023, doi: 10.25139/jsk.v7i3.7402.

[4] M. Dicky Desriansyah and I. Utna Sari, "Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks: Pada Berita Digital Berbahasa Indonesia," *JISKA: Jurnal Sistem Informasi Dan Informatika*, vol. 3, no. 2, p. 63, 2025, [Online]. Available: http://jurnal.unidha.ac.id/index.php/jiska

[5] A. Mu, amar Wahid, K. Adi Nugroho, T. Safitri, and F. Setyo Utomo, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis TF-IDF," *Jurnal Pendidikan dan Teknologi Indonesia (JPTI)*, vol. 4, no. 8, pp. 381–392, 2024, doi: 10.52436/1.jpti.602.

[6] 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," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1625–1633. doi: 10.1016/j.procs.2025.09.105.

[7] A. Fardhina, R. M. Siregar, M. R. W. Br Sibarani, I. C. Br Ginting, and A. Pratama, "Sistem Deteksi Berita Hoaks berbasis Algoritma Natural Language Processing (NLP) menggunakan BERT," *Jurnal Manajemen Informatika, Sistem Informasi dan Teknologi Komputer (JUMISTIK)*, vol. 4, no. 1, pp. 450–461, Jun. 2025, doi: 10.70247/jumistik.v4i1.156.

[8] C. J. L. Tobing, IGN Lanang Wijayakusuma, and Luh Putu Ida Harini, "Perbandingan Kinerja IndoBERT dan MBERT Untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia," *JST (Jurnal Sains dan Teknologi)*, vol. 14, no. 1, pp. 114–123, May 2025, doi: 10.23887/jstundiksha.v14i1.92126.

[9] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and Permuted Pre-training for Language Understanding," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2004.09297

[10] Pulung Hendro Prastyo, Eddy Tungadi, and Shaifudin Zuhdi, "Indonesian Automated Essay Scoring: A Comparative Study of Pretrained Transformer Models," *Information Technology Education Journal*, pp. 120–130, Jun. 2025, doi: 10.59562/intec.v4i2.8069.

[11] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A Survey on Automated Fact-Checking", doi: 10.1162/tacl.

[12] I. Ali, M. N. Bin Ayub, P. Shivakumara, and N. F. B. M. Noor, "Fake News Detection Techniques on Social Media: A Survey," 2022, *Hindawi Limited*. doi: 10.1155/2022/6072084.

[13] W. Mu and K. H. Lim, "Modelling Text Similarity: A Survey," in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2023*, Association for Computing Machinery, Inc, Nov. 2023, pp. 698–705. doi: 10.1145/3625007.3627305.

[14] S. Fitria, N. Azizah, H. D. Cahyono, S. W. Sihwi, and W. Widiarto, "Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection." [Online]. Available: https://github.com/Shafna81/fakenewsdetection.git

[15] Ö. Sevgili, I. Nikishina, S. M. Yimam, M. Semmann, and C. Biemann, "UHH at AVeriTeC: RAG for Fact-Checking with Real-World Claims," 2024. [Online]. Available: https://fever.ai/task.html

[16] A. R. Hanum *et al.*, "Analisis Kinerja Algoritma Klasifikasi Teks Bert Dalam Mendeteksi Berita Hoaks," vol. 11, no. 3, pp. 537–546, 2024, doi: 10.25126/jtiik938093.

[17] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2303.08774

[18] Z. H. Amur, Y. K. Hooi, G. M. Soomro, H. Bhanbhro, S. Karyem, and N. Sohu, "Unlocking the Potential of Keyword Extraction: The Need for Access to High-Quality Datasets," 2023, doi: 10.3390/app.

[19] Y. A. Hafiz and E. Sudarmilah, "Implementasi Web Scraping Pada Portal Berita Online."

[20] S. Sannigrahi, J. van Genabith, and C. Espana-Bonet, "Are the Best Multilingual Document Embeddings simply Based on Sentence Embeddings?," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.14796

[21] A. Pannadhika Putra, D. Purnami Singgih Putri, and Aak. Cahyawan Wiranatha, "Scientific Paper Recommendation System: Application of Sentence Transformers and Cosine Similarity Using arXiv Data," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[22] M. Abdul, H. Fathuddin, E. Prakarsa Mandyartha, and A. L. Nurlaili, "Penerapan Sentence-Bert dan Cosine Similarity untuk Pencarian Semantik Dokumen Skripsi dalam Format PDF," *R2J*, vol. 8, no. 1, 2025, doi: 10.38035/rrj.v8i1.

[23] M. T. Colangelo, M. Meleti, S. Guizzardi, E. Calciolari, and C. Galli, "A Comparative Analysis of Sentence Transformer Models for Automated Journal Recommendation Using PubMed Metadata," *Big Data and Cognitive Computing*, vol. 9, no. 3, Mar. 2025, doi: 10.3390/bdcc9030067.

[24] M. Siino, "All-MPNet at SemEval-2024 Task 1: Application of MPNet for Evaluating Semantic Textual Relatedness." [Online]. Available: https://semantic-textual-relatedness.github.

[25] N. Muennighoff, "SGPT: GPT Sentence Embeddings for Semantic Search," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2202.08904

[26] Z. H. Amur, Y. Kwang Hooi, H. Bhanbhro, K. Dahri, and G. M. Soomro, "Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives," Mar. 01, 2023, *MDPI*. doi: 10.3390/app13063911.

[27] A. Awalina, J. Fawaid, R. Yunus Krisnabayu, and N. Yudistira, "Indonesia's Fake News Detection using Transformer Network." [Online]. Available: https://github.com/JibranFawaid/turnbackhoax-dataset.

[28] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2104.08663