JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Pengaruh Preprocessing Regex dan Cosine Similarity terhadap Performa IndoBERT dalam Klasifikasi Berita Hoaks Berbahasa Indonesia

Minggar P. D. Ramadhan*, Zainal Abidin, M. Imamudin

Fakultas Sains dan Teknologi, Magister Informatika, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia
Email: [1*]minggar.ptra@gmail.com, [2]zainal@ti.uin-malang.ac.id, [3]imamudin@ti.uin-malang.ac.id
Email Penulis Korespondensi: minggar.ptra@gmail.com
Submitted **26-04-2026;** Accepted **21-05-2026;** Published **30-06-2026**

Abstrak

Perkembangan media daring telah mempermudah akses informasi, namun juga meningkatkan penyebaran berita hoaks di masyarakat. Dataset yang umum digunakan dalam penelitian deteksi berita hoaks bersumber dari platform fact-checking yang memuat struktur teks berupa klaim, narasi, serta kalimat klarifikasi yang menjelaskan bahwa suatu informasi tidak benar. Keberadaan kalimat klarifikasi ini berpotensi menimbulkan bias, yaitu kondisi ketika model mempelajari pola teks yang secara eksplisit mengindikasikan label kelas, sehingga mengurangi kemampuan model dalam memahami isi berita secara menyeluruh. Penelitian ini bertujuan untuk menganalisis pengaruh teknik preprocessing berbasis regular expression (regex) dan cosine similarity terhadap performa model IndoBERT dalam klasifikasi berita hoaks berbahasa Indonesia. Kedua pendekatan digunakan untuk mengidentifikasi dan menangani kalimat klarifikasi agar model lebih berfokus pada pemahaman konteks dan isi berita. Eksperimen dilakukan dengan membandingkan performa model pada dataset yang diproses menggunakan masing-masing teknik preprocessing. Hasil menunjukkan bahwa pendekatan cosine similarity menghasilkan performa yang lebih baik dengan nilai accuracy, precision, recall, dan f1-score sebesar 92.8%, dibandingkan dengan regular expression (regex) yang masing-masing memperoleh accuracy sebesar 90.7%, precision sebesar 91.3%, recall sebesar 90.7%, dan f1-score sebesar 90.6%. Penelitian ini berkontribusi dalam mengusulkan dan mengevaluasi strategi preprocessing berbasis regular expression (regex) dan cosine similarity sebagai pendekatan untuk mengurangi potensi bias pada dataset berita hoaks berbasis fact-checking, sekaligus memberikan bukti bahwa pemilihan teknik preprocessing yang tepat berpengaruh signifikan terhadap kemampuan model IndoBERT dalam klasifikasi berita hoaks berbahasa Indonesia.

Kata Kunci: Klasifikasi; Text Preprocessing; Regex; Cosine Similarity; IndoBERT

Abstract

The rapid growth of online media has significantly improved access to information, but it has also accelerated the spread of misinformation and hoax news. In hoax detection research, datasets are commonly derived from fact-checking platforms, which typically contain structured components such as claims, narratives, and clarification statements explicitly indicating that certain information is false. The presence of such clarification sentences has the potential to cause bias, a condition in which the model learns text patterns that explicitly indicate the label, thereby reducing the model's ability to fully understand the content of the news. This study aims to analyze the impact of preprocessing techniques based on regular expression (regex) and cosine similarity on the performance of the IndoBERT model for Indonesian hoax news classification. Both approaches are employed to identify and handle clarification sentences, enabling the model to focus more on contextual and semantic understanding of the news content. Experimental results show that the cosine similarity-based preprocessing outperforms the regex-based approach, achieving accuracy, precision, recall, and F1-score of 92.8%. In comparison, the regex-based method obtains an accuracy of 90.7%, precision of 91.3%, recall of 90.7%, and F1-score of 90.6%. These findings indicate that the semantic-based approach is more effective in handling linguistic variability and reducing potential bias caused by explicit clarification patterns. Overall, this study highlights the importance of appropriate preprocessing strategies in improving classification performance and provides insights into the impact of clarification statements in fact-checking datasets on transformer-based hoax detection models.

Keywords: Classification; Text Preprocessing; Regex; Cosine Similarity; IndoBERT

1. PENDAHULUAN

Dalam beberapa tahun terakhir, masyarakat Indonesia semakin terbiasa memperoleh berita melalui media daring. Akses informasi yang cepat dan mudah melalui berbagai platform digital telah memperluas penyebaran berita. Secara global, perkembangan media berita daring juga telah mengubah cara publik mengonsumsi informasi, sehingga konsumsi berita digital kini menjadi bagian penting dalam cara masyarakat menerima dan menyebarkan berita secara luas [1]. Pertumbuhan media daring tidak hanya menyediakan ruang bagi penyebaran berita, tetapi juga dapat menjadi tempat subur bagi penyebaran misinformasi atau berita hoaks [2]. Berita palsu atau hoaks merupakan informasi yang direkayasa dengan tujuan menyesatkan, serta dikemas dengan meniru bentuk berita yang kredibel [3]. Berita hoaks sering kali disajikan dengan tampilan dan gaya bahasa yang tampak meyakinkan dan menyerupai berita faktual, sehingga menyulitkan pembaca dalam membedakan informasi yang benar dan menyesatkan. Keberadaan berita hoaks dapat menimbulkan berbagai kesalahpahaman, memengaruhi pandangan masyarakat terhadap suatu isu, serta mengurangi kepercayaan terhadap informasi yang beredar di ruang digital.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Dalam konteks Indonesia, fenomena penyebaran informasi palsu atau hoaks semakin mendapat perhatian serius. Data dari Kementerian Komunikasi dan Informatika menunjukkan bahwa selama tahun 2024 terdapat 1.923 konten hoaks yang terdeteksi beredar di berbagai platform digital. Seiring dengan masifnya peredaran berita hoaks di media daring, proses identifikasi dan penyaringan berita hoaks secara manual menjadi semakin sulit untuk dilakukan secara konsisten. Verifikasi informasi memerlukan waktu, sumber daya, serta tingkat literasi informasi yang memadai, sementara arus penyebaran berita berlangsung sangat cepat. Kondisi tersebut mendorong perlunya pendekatan otomatis yang mampu membantu mengidentifikasi berita hoaks secara efisien, konsisten, dan sistematis. Salah satu pendekatan yang telah digunakan adalah pemanfaatan pemrosesan bahasa alami atau Natural Language Processing (NLP), yang merupakan bagian dari kecerdasan buatan dan berfokus pada pengolahan dan Pemahaman teks berbasis bahasa alami [4][5]. Pendekatan NLP dapat dikombinasikan dengan algoritma *machine learning*, *deep learning* bahkan model berbasis *transformer* untuk mengidentifikasi pola linguistik dan konteks semantik dalam pengolahan teks berita. Sejumlah penelitian menunjukkan bahwa model yang dilatih menggunakan pendekatan tersebut mampu mendeteksi berita hoaks secara otomatis dengan tingkat akurasi tertentu [6][7].

Penelitian mengenai deteksi berita hoaks telah banyak memanfaatkan arsitektur transformer, khususnya IndoBERT, yang dirancang untuk memahami karakteristik bahasa Indonesia melalui mekanisme *self-attention*. Hasil penelitian sebelumnya menunjukkan bahwa IndoBERT mampu melampaui kinerja model *machine learning* dan *deep learning* non-transformer [8]. Dengan menggunakan model IndoBERT terbukti mampu melampaui kinerja model *machine learning* seperti SVM dan Naïve Bayes dengan capaian skor precision, recall, dan F1-score sebesar 94.66% [9]. Penelitian lain lain juga telah menghasilkan bahwa IndoBERT mampu menghasilkan performa yang unggul dalam mengklasifikasikan berita hoaks berbahasa Indonesia, termasuk pada domain spesifik seperti berita politik [10]. Lebih lanjut, dengan memanfaatkan dataset berita Indonesia, IndoBERT terbukti secara signifikan melampaui model lainnya dengan capaian akurasi tertinggi sebesar 92.24% yang menegaskan efektivitas model berbasis transformer untuk deteksi berita hoaks berbahasa Indonesia [11]. Meskipun berbagai penelitian telah menunjukkan capaian performa yang baik, sebagian besar penelitian masih berfokus pada peningkatan nilai akurasi dan perbandingan antar model. Penelitian yang mengkaji potensi bias pada dataset serta pengaruhnya terhadap perilaku model relatif masih terbatas. Sebagian penelitian sebelumnya dalam deteksi berita hoaks berbahasa Indonesia menggunakan dataset yang bersumber dari turnbackhoax.id [10][9][11]. Turnbackhoax.id merupakan situs pengecekan fakta (*fact-checking*) yang dikelola oleh MAFINDO (Masyarakat Anti Fitnah Indonesia), sebuah organisasi pengecek fakta di Indonesia yang telah tersertifikasi oleh *International Fact-Checking Network* (IFCN) [12]. Pada dataset berita hoaks berbasis *fact-checking*, struktur dokumen umumnya memuat bagian klaim, narasi, serta kalimat klarifikasi atau kesimpulan yang secara eksplisit menyatakan bahwa informasi tersebut tidak benar atau menyesatkan. Keberadaan kalimat klarifikasi tersebut berpotensi menimbulkan bias, yaitu kondisi ketika model mempelajari pola yang secara langsung mengindikasikan label kelas, bukan memahami isi berita secara menyeluruh. Akibatnya, model dapat menunjukkan performa tinggi pada data pelatihan dan pengujian yang serupa, namun kurang *robust* ketika dihadapkan pada data baru yang tidak memiliki struktur klarifikasi yang sama. *Preprocessing* general NLP seperti penghapusan tanda baca, stopword, atau normalisasi huruf belum secara khusus dirancang untuk mengatasi permasalahan semacam ini. Oleh karena itu, diperlukan pendekatan *preprocessing* yang lebih terarah dan berbasis pada identifikasi pola klarifikasi yang berpotensi menyebabkan bias klasifikasi. Regular expression merupakan metode yang umum digunakan dalam pemrosesan teks untuk melakukan pencocokan pola (pattern matching) berdasarkan aturan yang telah ditentukan [13]. Dalam Penelitian ini, teknik *regular expression* (*regex*) dapat dimanfaatkan untuk mengenali pola kalimat atau frasa klarifikasi tertentu yang sering muncul pada dokumen *fact-checking*, sehingga bagian teks tersebut dapat diidentifikasi atau diproses lebih lanjut pada tahap pembersihan data. Selain itu, pendekatan berbasis *cosine similarity* dapat digunakan untuk mengukur tingkat kemiripan antar teks atau antar bagian dokumen [14], sehingga membantu dalam menganalisis kesamaan konten yang berpotensi muncul pada beberapa berita dengan struktur klarifikasi yang serupa. Penerapan kedua teknik tersebut dalam tahap *preprocessing* diharapkan dapat membantu mengurangi pengaruh pola teks yang secara eksplisit mengindikasikan label kelas, sehingga model IndoBERT lebih berfokus pada pemahaman konteks dan isi berita dalam proses klasifikasi.

Berdasarkan latar belakang tersebut, maka penelitian ini bertujuan untuk menganalisis pengaruh penerapan teknik *preprocessing* berbasis *regular expression* (*regex*) dan *cosine similarity* terhadap performa model IndoBERT dalam melakukan klasifikasi berita hoaks berbahasa Indonesia. Kedua teknik tersebut digunakan untuk membantu mengidentifikasi serta menangani keberadaan kalimat klarifikasi yang berpotensi dapat menyebabkan model mempelajari pola teks yang secara eksplisit berkaitan dengan label kelas. Evaluasi dilakukan dengan membandingkan performa model IndoBERT yang dilatih menggunakan data yang diproses dengan teknik *regex* dan data yang diproses menggunakan pendekatan *cosine similarity*. Dengan begitu, kontribusi penelitian ini adalah mengusulkan dan mengevaluasi dua pendekatan *preprocessing* yang dirancang untuk mengurangi potensi bias akibat keberadaan kalimat klarifikasi pada dataset berita hoaks berbasis *fact-checking*.

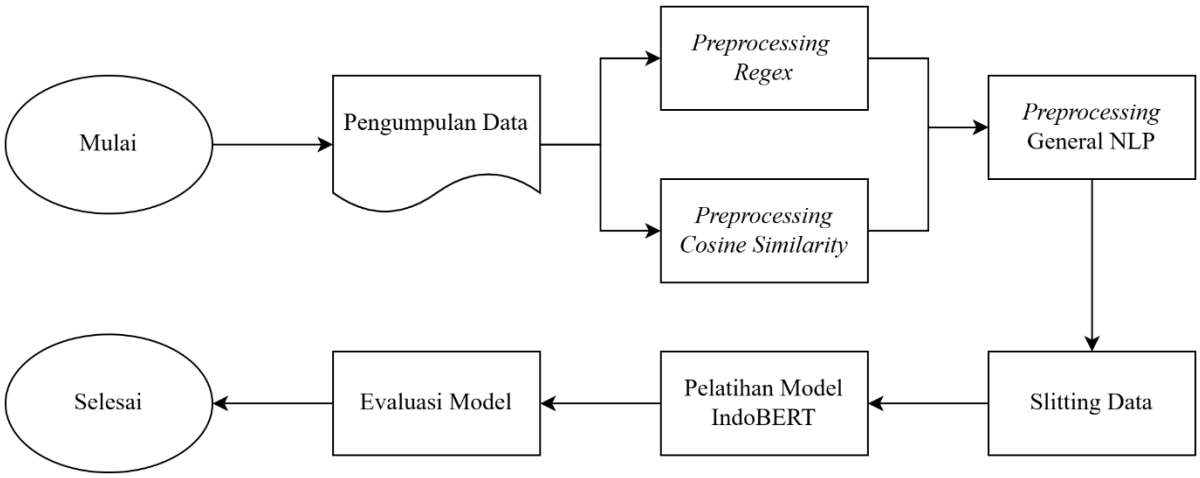
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian merupakan bagian yang secara rinci menjelaskan alur kerja yang dilakukan secara sistematis dan terarah dalam mencapai tujuan penelitian. Penyusunan tahapan ini bertujuan untuk memberikan gambaran yang jelas

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

mengenai proses yang dilalui, mulai dari tahap pengumpulan data hingga tahap evaluasi model yang digunakan. Dengan adanya tahapan yang terstruktur, setiap proses dalam penelitian dapat dilakukan secara ilmiah. Selain itu, tahapan penelitian juga dapat memudahkan dalam memahami hubungan antar proses yang dilakukan, khususnya dalam penerapan teknik *preprocessing* dan pemodelan dengan menggunakan IndoBERT pada klasifikasi berita hoaks. Berikut merupakan tahapan pada penelitian ini yang dilakukan sebagaimana ditunjukkan seperti Gambar 1.



Gambar 1. Tahapan penelitian

Berdasarkan Gambar 1 tahapan penelitian dimulai dengan proses Pengumpulan data yang kemudian dilanjutkan ke tahap *preprocessing*. Pada tahap *preprocessing*, data diproses melalui dua pendekatan yang berbeda, yaitu preprocessing berbasis *regular expression* (*regex*) dan *cosine similarity*. Masing-masing dari dua pendekatan tersebut dilakukan terlebih dahulu sebelum dilanjutkan dengan *preprocessing* standar yang meliputi *case folding*, *URL & email removal*, *cleaning special characters*, *whitespace removal*, *encoding label*, dan *tokenization*. Setelah seluruh tahapan preprocessing selesai, dataset kemudian dibagi/*splitting* menjadi data pelatihan, validasi, dan pengujian. Selanjutnya, dataset diproses dalam tahap pelatihan dengan menggunakan model IndoBERT untuk melakukan klasifikasi berita hoaks. Tahap akhir dari penelitian ini adalah evaluasi model untuk menilai performa klasifikasi yang dihasilkan.

2.2 Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini bersumber dari repositori data daring Kaggle, yang menyediakan berbagai dataset terbuka untuk keperluan riset dan pengembangan model kecerdasan buatan. Dataset yang dimanfaatkan berupa kumpulan berita berbahasa Indonesia yang dikompilasi dari empat sumber utama, yaitu CNN Indonesia, Kompas, Tempo, dan Turnbackhoax.id. Sumber-sumber berita tersebut mencerminkan dua kategori berita yang berbeda secara fundamental. CNN Indonesia, Kompas dan Tempo merupakan portal berita daring resmi di Indonesia yang menyajikan berita terverifikasi dan dapat dipercaya (nonhoaks). Turnbackhoax.id sebagai platform fact-checking yang berperan sebagai sumber data berita hoaks dalam penelitian ini. Turnbackhoax.id dikelola oleh MAFINDO (Masyarakat Anti Fitnah Indonesia atau *Indonesian Anti-Slander Society*), sebuah organisasi yang menjadi pionir dalam praktik *fact-checking* di Indonesia [15]. Setiap artikel pada turnbackhoax.id tersebut dipublikasikan merupakan hasil verifikasi fakta terhadap informasi yang beredar di masyarakat, termasuk misinformasi yang disebarkan melalui media sosial.

Setiap data dalam dataset berita telah memiliki label yang mengategorikan ke dalam kelas berita hoaks maupun nonhoaks. Adapun susunan data atau struktur data terdiri atas dua bagian utama, yaitu isi teks berita sebagai data masukan / *input* dan label kelas kategori sebagai nilai target yang ingin diprediksi oleh model. Kelengkapan label pada setiap data memiliki aspek penting yang mendukung keandalan proses pelatihan model secara keseluruhan. Contoh representasi data yang digunakan dalam penelitian ini ditampilkan pada Tabel 1.

Tabel 1. Dataset Penelitian

Berita	Label
Terciduk Prabowo menggunakan kacamata canggih Google Glass terbaru yang bisa nyontek jawaban …	Hoaks
SBY Sempat Ingatkan Prabowo agar Kampanye Akbar Tak Tunjukkan Politik Identitas …	Nonhoaks
Info Pembuatan SIM Kolektif. Kabar gembira buat teman-teman yang belum memiliki Surat Izin Mengemudi (SIM) …	Hoaks
Habiskan Libur Tahun Baru 2023 di Pacitan, SBY Mampir Makan Siang di Colomadu sela acara, Ahad. 1 Januari 2023…	Nonhoaks

2.3 Preprocessing Teks

Tahapan *preprocessing* data memiliki peran penting dalam setiap proses pemrosesan bahasa alami (*Natural Language Processing*/NLP), karena kualitas data masukan sangat memengaruhi kinerja model yang digunakan [16]. Oleh karena

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

itu, pada penelitian ini dilakukan serangkaian proses pra-pemrosesan untuk menyesuaikan teks berita agar kompatibel dengan kebutuhan model IndoBERT. Pra-pemrosesan bertujuan untuk memperbaiki kualitas representasi teks dengan mereduksi unsur-unsur yang tidak relevan, seperti karakter khusus, simbol, maupun komponen lain yang berpotensi menimbulkan noise pada data [17]. Melalui tahapan ini, teks berita diharapkan dapat disajikan dalam bentuk yang lebih terstruktur dan konsisten sehingga mendukung proses pembelajaran model secara optimal. Pada penelitian ini, tahap preprocessing dilakukan secara bertahap dengan mengawali proses pada penerapan teknik berbasis *regular expression* (*regex*) dan *cosine similarity*. Kedua pendekatan tersebut digunakan untuk mengidentifikasi, menyaring, serta membersihkan dengan menghapus pola teks tertentu, khususnya kalimat klarifikasi yang terdapat dalam dataset berita. Proses ini bertujuan untuk mengurangi pengaruh pola teks yang berpotensi mengganggu pembelajaran model. Setelah melalui tahapan tersebut, data kemudian diproses lebih lanjut menggunakan teknik preprocessing general NLP untuk memastikan teks berada dalam kondisi yang lebih terstruktur dan siap digunakan pada tahap pemodelan.

2.3.1 Preprocessing Berbasis *Regular Expression* (*Regex*)

Regular expression atau yang sering disebut dengan *regex*, secara teori merupakan suatu urutan karakter terbatas yang mendefinisikan pola pencarian yang umumnya dimanfaatkan untuk menemukan, mengganti substring, serta memproses input teks secara sistematis [18]. Pada penelitian ini, teknik preprocessing berbasis *regular expression* (*regex*) digunakan untuk membantu mengidentifikasi serta menghapus pola teks yang berkaitan dengan kalimat klarifikasi pada dataset berita hoaks. *Regular expression* merupakan metode yang umum digunakan dalam pemrosesan teks untuk melakukan pencocokan pola (*pattern matching*) serta mengidentifikasi dan memodifikasi bagian tertentu dari string berdasarkan aturan yang telah ditentukan [13]. Secara konseptual, *regex* dapat direpresentasikan sebagai suatu ekspresi pola R, yang dibagun dari himpunan simbol alfabet Σ, dengan operasi dasar seperti konkatenasi, alternasi, dan pengulangan. Salah satu bentuk umum dari ekspresi tersebut dapat ditulis sebagai berikut:

$$R = (a \mid b)^* \tag{1}$$

Dimana:
R : pola *regular expression* yang digunakan untuk mencocokkan teks
a, b : elemen atau simbol dalam teks (dapat berupa karakter, kata, atau frasa tertentu)

Tabel 2. *Pseudocode Regex*

Pseudocode

```
START
LOAD dataset
DEFINE pattern_regex = [
"mafindo", "klaim", "fitnah", "palsu", "menyesatkan", ...
]
FOR each berita IN dataset:
    text = kolom berita
    text = lowercase(text)
    kalimat_list = split text into sentences
    hasil_kalimat = []

    FOR each kalimat IN kalimat_list:
        IF kalimat matches pattern_regex:
            hapus semua kata/frasa pada kalimat yang cocok dengan
pattern_gerex
        ELSE:
            ADD kalimat TO hasil_kalimat

    text_bersih = join hasil_kalimat
    SAVE text_bersih
END FOR

OUTPUT dataset hasil
END
```

Dalam *natural language processing, regular expression* merupakan pendekatan berbasis aturan yang memungkinkan proses pencocokan pola teks dilakukan secara langsung berdasarkan karakter atau susunan kata tertentu yang telah ditentukan sebelumnya [5]. Melalui pendekatan ini, bagian teks yang dianggap tidak relevan dapat dikenali dan diproses secara sistematis. Sesuai pada Tabel 2, proses penerapan *regex* dilakukan pada tingkat kalimat, sehingga setiap kalimat dalam dataset berita diperlakukan secara terpisah untuk meningkatkan ketepatan dalam mendeteksi pola yang tidak diinginkan. Sebelum tahap ini dilakukan, seluruh teks terlebih dahulu dikonversi ke dalam huruf kecil (*lower*

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

casing) guna menjaga konsistensi dalam proses pencocokan pola. Selanjutnya, *regex* digunakan untuk menghapus frasa atau kata tertentu yang mengandung indikasi klarifikasi, seperti "mafindo", "klaim", "fitnah", "palsu", "penipuan", "bohong", serta frasa seperti "berdasarkan hasil penelusuran menggunakan *google images*" dan ungkapan serupa lainnya. Proses penghapusan dilakukan secara langsung pada bagian teks yang terdeteksi tanpa menghilangkan keseluruhan kalimat, sehingga informasi utama dalam berita tetap terjaga. Melalui tahapan ini, diharapkan keberadaan pola teks yang secara eksplisit mengarah pada label tertentu dapat dikurangi, sehingga model IndoBERT lebih terdorong untuk memahami isi berita berdasarkan konteks yang lebih utuh. Perlu dicatat bahwa proses pencocokan pola dilakukan berdasarkan kata kunci yang telah ditentukan tanpa mempertimbangkan variasi bentuk kata secara mendalam, sehingga efektivitasnya sangat bergantung pada kelengkapan dan representativitas kata kunci yang digunakan.

2.3.2 Preprocessing Berbasis *Cosine Similarity*

Cosine similarity adalah suatu metode yang digunakan untuk menghitung jarak dan tingkat kemiripan antara dua teks [14]. Pada penelitian ini, pendekatan preprocessing berbasis *cosine similarity* digunakan untuk membantu mengidentifikasi dan menyaring kalimat klarifikasi yang terdapat dalam dataset berita hoaks. Berbeda dengan pendekatan berbasis aturan seperti *regex*, metode ini memanfaatkan representasi semantik teks untuk melihat tingkat kemiripan makna antar kalimat. Secara matematis, *cosine similarity* antara dua vektor teks A dan B didefinisikan sebagai berikut :

$$Cosine\ Similarity\ (A,B) = \frac{A\ .\ B}{||A|| \times ||B||} \quad (2)$$

Dimana :
A . B : hasil perkalian dot product
||A||×||B||: panjang vektor

Penentuan nilai ambang batas (*threshold*) pada *cosine similarity* dalam penelitian ini dilakukan berdasarkan analisis distribusi nilai kemiripan yang dihasilkan dari perbandingan antara teks kalimat berita dalam dataset dan template klarifikasi. Hasil pengukuran *cosine similarity* berada pada rentang nilai 0 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi antar teks, sedangkan nilai yang mendekati 0 mengindikasikan bahwa kedua teks memiliki tingkat perbedaan yang semakin besar [19]. Analisis ini dilakukan untuk memperoleh gambaran mengenai pola sebaran nilai *similarity*, sehingga penentuan nilai ambang batas (*threshold*) dapat dilakukan secara lebih objektif dan sesuai dengan karakteristik data.

Tabel 3. Statistik Deskriptif

Statistik	Nilai
Minimum	0.05
Kuartil 1 (Q1)	0.31
Median (Q2)	0.39
Kuartil 3 (Q3)	0.49
Maksimum	0.99

Tabel 4. Distribusi Persentil

Persentil	Nilai
P80	0.518
P85	0.551
P90	0.596
P95	0.672

Hasil analisis statistik pada Tabel 3, menunjukkan bahwa nilai median (Q2) berada pada 0.39, sedangkan kuartil ketiga (Q3) sebesar 0.49 yang mengindikasikan awal dari kelompok kalimat dengan tingkat kemiripan yang relatif tinggi. Di sisi lain pada Tabel 4, nilai percentil ke-85 (P85) sebesar 0.551 yang menggambarkan kelompok kalimat dengan tingkat kemiripan yang lebih kuat terhadap template klasifikasi. Rentang antara Q3 dan P85 tersebut dapat dipandang sebagai wilayah peralihan kemiripan sedang menuju kemiripan tinggi. Berdasarkan pertimbangan tersebut, nilai *threshold* pada penelitian ini sebesar 0.55 dipilih karena berada di atas Q3 dan mendekati P85, sehingga dianggap mampu merepresentasikan kalimat dengan tingkat kemiripan yang cukup tinggi tanpa menyebabkan penghapusan data secara berlebihan. Selain itu, pemilihan nilai *threshold* ini juga mempertimbangkan keseimbangan antara kemampuan dalam mendeteksi kalimat klasifikasi dan upaya mempertahankan informasi utama dalam teks berita. Dengan demikian, nilai threshold yang digunakan dalam penelitian ini dipilih berdasarkan analisis data secara empiris, sehingga lebih mencerminkan kondisi distribusi nilai *cosine similarity* yang sebenarnya.

Tabel 5. *Pseudocode Cosine Similarity*

Pseudocode
`START`
`LOAD dataset`

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

```
LOAD SentenceTransformer model
LOAD template klarifikasi

ENCODE all templates
template_embeddings = encode(templates klarifikasi)

SET SIM_THRESHOLD = 0.5
SET MIN_WORDS = 5
COPY dataset to cosine

FOR each row IN cosine:
    text = row["berita"]
    text = lowercase(text)
    SPLIT text into sentences

    filtered_sentences = []

    FOR each sentence s:
        s = trim(s)
        IF total_words(s) < MIN_WORDS:
            ADD s TO filtered_sentences
            CONTINUE

        sentence_embedding = encode(s)

        similarity_scores =
            cosine_similarity(
                sentence_embedding,
                template_embeddings)

        max_sim = maximum(similarity_scores)

        IF max_sim < SIM_THRESHOLD:
            ADD s TO filtered_sentences
        ELSE:
            REMOVE sentence s

    cleaned_text = join(filtered_sentences)
    SAVE cleaned_text
    TO row["text berita baru"]

END FOR
DELETE rows where text berita baru is empty

END
```

Berdasarkan pada Tabel 5, proses *cosine similarity* dilakukan pada tingkat kalimat, di mana setiap kalimat dalam teks berita dibandingkan dengan sejumlah template kalimat klarifikasi diperoleh dari kalimat klasifikasi yang terdapat pada dataset berita. Untuk merepresentasikan teks secara numerik, digunakan pendekatan sentence embedding berbasis *Sentence Transformer*, sehingga setiap kalimat dapat direpresentasikan dalam bentuk vektor. Tingkat kemiripan antara kalimat berita dan template klarifikasi dihitung menggunakan cosine similarity. Apabila nilai kemiripan yang diperoleh melebihi ambang batas (*threshold*) yang telah ditentukan, maka kalimat tersebut dianggap memiliki kemiripan makna dengan pola klarifikasi dan akan dihapus dari teks. Dalam proses ini, hanya kalimat dengan jumlah kata lebih dari lima yang dipertimbangkan, guna menghindari penghapusan pada teks yang terlalu pendek. Melalui pendekatan ini, diharapkan pola klarifikasi yang tidak selalu dapat terdeteksi secara langsung dapat diminimalkan, sehingga model IndoBERT dapat lebih fokus dalam memahami isi berita secara kontekstual.

2.3.3 Preprocessing General NLP

Preprocessing pada penelitian ini, dataset terlebih dahulu diproses menggunakan preprocessing berbasis *regex* dan *cosine similarity* untuk mengidentifikasi, menyaring dan membersihkan pola teks tertentu seperti teks klarifikasi yang ada pada dataset teks berita. Setelah dataset di *preprocessing* menggunakan *regex* dan cosine similarity, selanjutnya akan diproses dengan menggunakan preprocessing general NLP. Adapun tahapan yang diterapkan mencakup beberapa proses preprocessing general NLP yang umum digunakan dalam pengolahan teks meliputi:

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

a. *Case folding*, suatu proses yang dilakukan untuk mengubah seluruh karakter dalam teks menjadi huruf kecil. Proses ini bertujuan untuk menyeragamkan bentuk kata sehingga perbedaan penggunaan huruf besar dan kecil tidak memengaruhi hasil analisis teks.
b. *URL & Email Removal*, proses menghapus seluruh tautan (*Uniform Resource Locator* atau URL) dan alamat email karena tidak memberikan kontribusi terhadap isi semantik teks. Proses ini bertujuan untuk mengurangi noise pada data teks agar hasil klasifikasi menjadi lebih akurat.
c. *Cleaning Special Characters*, proses untuk menghapus simbol dan karakter non-alfanumerik, seperti emotikon, tanda khusus, serta karakter asing. Selain itu, angka dan elemen temporal seperti tanggal juga dieliminasi, karena tidak memberikan nilai informatif yang signifikan terhadap konteks berita.
d. *Whitespace Removal*, proses menghapus spasi ganda, tabulasi, serta newline characters yang berlebihan dihapus untuk menormalkan struktur teks.
e. *Encoding Label*, langkah untuk mengubah data pada kolom klasifikasi (label) kategorikal menjadi representasi numerik. Pada tahap ini, label hoaks dikodekan dengan label 1, sedangkan label nonhoaks dikodekan dengan label 0. Proses encoding ini bertujuan untuk memastikan bahwa label klasifikasi dapat diproses oleh algoritma machine learning termasuk model deep learning yang memerlukan input dalam bentuk numerik. Dengan penerapan label encoding ini, model dapat melakukan proses pelatihan dan evaluasi secara lebih efektif dan terstruktur.
f. *Tokenization*, mengubah teks berita menjadi unit-unit terkecil (token) yang kemudian dapat diproses oleh model.

2.4 Splitting Dataset

Pada penelitian ini, dataset dibagi menjadi tiga bagian utama, yaitu data pelatihan (*training*), validasi (*validation*), dan pengujian (*testing*) dengan perbandingan 70:15:15. Pembagian ini bertujuan untuk memastikan bahwa model dapat dilatih secara optimal, sekaligus dapat dievaluasi secara objektif terhadap data yang tidak digunakan selama proses pelatihan. Proses pembagian data dilakukan setelah seluruh tahapan *preprocessing* selesai dilakukan, sehingga setiap variasi dataset, baik hasil *preprocessing* berbasis *regex* maupun *cosine similarity* telah diproses terlebih dahulu sebelum digunakan dalam tahap pelatihan model. Metode pembagian data dilakukan dengan mempertahankan proporsi distribusi label pada setiap subset data. Dengan mempertahankan representasi distribusi dataset asli dalam setiap subset merupakan hal yang sangat penting, karena pembagian yang tidak merata berpotensi menghasilkan subset dengan fitur dan kategori yang bias, yang pada akhirnya dapat memengaruhi performa model secara signifikan [20]. Dengan demikian, komposisi data antara kelas hoaks dan nonhoaks tetap seimbang pada data pelatihan, validasi, maupun pengujian. Selain itu, untuk memastikan konsistensi hasil dan kemudahan dalam reproduksi eksperimen, proses pembagian data dilakukan dengan menggunakan nilai *random seed* yang tetap. Pendekatan ini diharapkan dapat menghasilkan pembagian data yang representatif serta mendukung evaluasi model secara lebih adil dan stabil.

2.5 Pelatihan Model IndoBERT

Penelitian ini menggunakan IndoBERT sebagai model untuk melakukan proses klasifikasi berita hoaks berbahasa Indonesia. IndoBERT merupakan versi khusus dari model *Bidirectional Encoder Representations from Transformers* (BERT) yang dikembangkan dan dipralatih secara spesifik untuk korpus berbahasa Indonesia [21]. BERT dirancang untuk memahami konteks sebuah kata melalui pemrosesan dua arah, yaitu kanan ke kiri dan atau kiri ke kanan dalam suatu rangkaian teks. Kemampuan tersebut menjadikan IndoBERT relevan untuk merepresentasikan pola dan struktur teks berbahasa Indonesia. Secara arsitektur, IndoBERT terdiri dari atas 12 lapisan *transformer* dengan ukuran lapisan tersembunyi (*hidden size*) sebesar 768, dilengkapi 12 *self-attention heads* [9]. Dalam tugas klasifikasi teks, BERT dapat menerima masukan/*input* berupa rangkaian teks dengan panjang maksimum 512 token, yang selanjutnya direpresentasikan ke dalam bentuk vektor [22]. Konfigurasi ini memungkinkan proses atensi berjalan secara optimal dalam membangun representasi kontekstual yang kaya, sehingga hubungan antar kata dalam suatu teks dapat dipahami secara lebih mendalam. Pada tahap pemrosesan masukan, setiap teks berita direpresentasikan dalam bentuk vektor laten (*hidden state*). Representasi token khusus [CLS] pada lapisan akhir digunakan sebagai ringkasan informasi keseluruhan teks dan dimanfaatkan sebagai masukan utama dalam proses klasifikasi [23]. Pendekatan ini memungkinkan model IndoBERT menangkap makna keseluruhan teks secara efektif untuk membedakan berita hoaks dan berita nonhoaks.
Selama proses pelatihan, konfigurasi hyperparameter penelitian ini ditetapkan untuk memastikan proses pembelajaran berlangsung secara optimal dan stabil. Parameter model dioptimalkan menggunakan *AdamW Optimizer* dengan *learning rate* sebesar 2e-5. Batch size yang digunakan sebesar 16, yang merupakan keseimbangan antara efisiensi komputasi dan kestabilan estimasi gradien selama proses pelatihan. Panjang maksimum token masukan (max length) dibatasi sebesar 256 token, dengan pertimbangan bahwa sebagian besar teks berita dalam dataset telah tercakup dalam panjang tersebut sehingga dapat mengurangi beban komputasi tanpa kehilangan informasi yang signifikan. Proses pelatihan dilakukan selama maksimum 20 epoch dengan penerapan early stopping berdasarkan nilai validation loss dengan nilai patience sebesar 2, artinya pelatihan akan dihentikan secara otomatis apabila tidak terdapat peningkatan performa pada data validasi selama 2 epoch berturut-turut. Mekanisme early stopping ini diterapkan untuk mencegah terjadinya overfitting serta memastikan model yang disimpan merupakan model dengan performa terbaik pada data validasi.

2.7 Evaluasi Model

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Evaluasi kinerja model dilakukan untuk mengukur kemampuan IndoBERT dalam mengklasifikasikan berita secara akurat. Penelitian ini termasuk ke dalam tugas klasifikasi biner karena data dikelompokkan ke dalam dua kelas, yaitu hoaks dan non-hoaks. Proses evaluasi model pada penelitian ini dengan memanfaatkan sejumlah metrik evaluasi yang umum digunakan pada permasalahan klasifikasi biner, meliputi accuracy, precision, recall, dan F1-score [24]. Pemilihan metrik tersebut bertujuan untuk memberikan gambaran yang lebih utuh mengenai kinerja model, baik dari sisi ketepatan prediksi secara keseluruhan maupun kemampuan model dalam mengidentifikasi masing-masing kelas secara tepat. Melalui kombinasi keempat metrik ini, evaluasi tidak hanya berfokus pada tingkat akurasi, tetapi juga mempertimbangkan keseimbangan antara kesalahan prediksi dan keberhasilan klasifikasi pada kelas berita hoaks maupun berita nonhoaks, sehingga analisis performa model dapat dilakukan secara lebih objektif dan komprehensif.

a. *Precision*, merepresentasikan perbandingan antara jumlah prediksi positif yang benar dengan seluruh prediksi yang diberikan model sebagai kelas positif. Metrik ini digunakan untuk menggambarkan tingkat ketepatan model ketika menetapkan suatu berita sebagai hoaks, serta menunjukkan sejauh mana model mampu meminimalkan kesalahan prediksi pada kelas positif. Dengan demikian, precision memberikan informasi mengenai kualitas prediksi positif yang dihasilkan oleh model. *Precision* dapat dituliskan dalam persamaan (3):

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

b. *Recall*, mengukur proporsi data positif yang berhasil dikenali dengan benar oleh model dari keseluruhan data yang memang termasuk dalam kelas positif. Metrik ini mencerminkan kemampuan model dalam mendeteksi seluruh berita hoaks yang terdapat dalam dataset, sehingga penting untuk menilai sejauh mana model mampu menghindari kegagalan dalam mengenali data positif. Recall dapat dituliskan dalam persamaan (4):

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

c. *F1-score*, merupakan metrik evaluasi yang mengombinasikan nilai precision dan recall dalam satu ukuran kinerja. Metrik ini digunakan untuk memberikan gambaran yang lebih seimbang terhadap performa model, khususnya ketika terdapat perbedaan distribusi jumlah data antar kelas. Dengan mempertimbangkan kedua metrik tersebut secara bersamaan, *F1-score* membantu menilai efektivitas model dalam menghasilkan prediksi yang konsisten dan proporsional. *F-score* dapat dituliskan dalam persamaan (5):

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

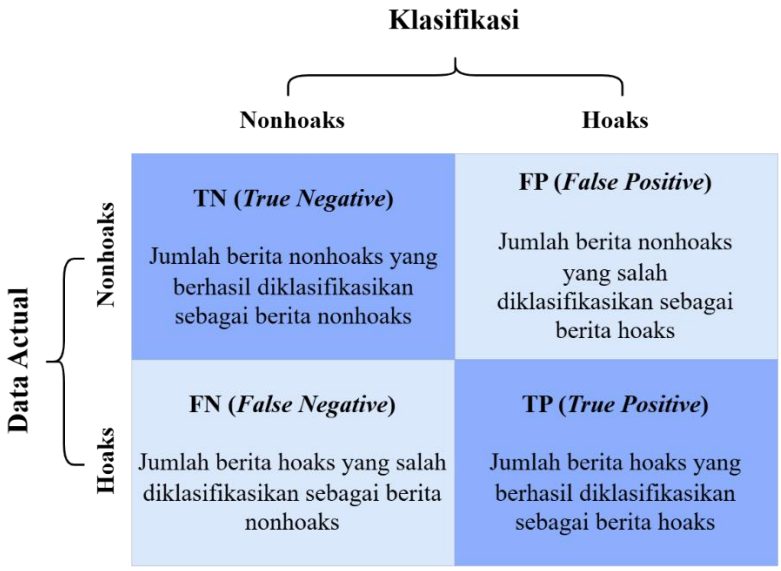
d. Accuracy, menunjukkan proporsi keseluruhan prediksi model yang sesuai dengan label sebenarnya pada data uji. Metrik ini memberikan gambaran umum mengenai tingkat ketepatan model dalam mengklasifikasikan berita ke dalam kategori hoaks maupun non-hoaks. Nilai akurasi yang tinggi mengindikasikan bahwa sebagian besar prediksi yang dihasilkan model berada pada kelas yang benar. Namun, dalam kondisi distribusi kelas yang tidak seimbang, metrik ini perlu dianalisis bersama metrik lain seperti precision, recall, dan F1-score agar evaluasi kinerja model dapat dilakukan secara lebih objektif. Accuracy dapat dituliskan dalam persamaan (6):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Dimana:
TP (*True Positive*) : Jumlah data berita hoaks yang berhasil diklasifikasikan sebagai berita hoaks.
TN (*True Negative*) : Jumlah data berita nonhoaks yang berhasil diklasifikasikan sebagai berita nonhoaks.
FP (*False Positive*) : Jumlah data berita nonhoaks yang salah diklasifikasikan sebagai berita hoaks.
FN (*False Negative*) : Jumlah data berita hoaks yang salah diklasifikasikan sebagai berita non.

e. *Confusion matrix*, merupakan bentuk gambaran dan visualisasi dari ringkasan keempat metrik evaluasi terhadap data testing [25]. *Confusion matrix* adalah representasi tabular sekaligus visual yang memetakan hasil klasifikasi model terhadap label aktual data, sehingga memberikan gambaran yang komprehensif mengenai distribusi prediksi pada setiap kelas. Metriks ini menyajikan empat nilai utama, yaitu TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), dan FN (*False Negative*), yang menjadi dasar perhitungan seluruh metrik evaluasi yang telah dijelaskan sebelumnya. Adapun *confusion matrix* untuk klasifikasi berita hoaks pada penelitian ini sesuai pada Gambar 2.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 2. Contoh Confusion *Matrix* Klasifikasi Berita Hoaks

Confusion matrix pada umumnya umumnya divisualisasikan dalam bentuk *heatmap*, yaitu suatu representasi visual berbasis gradasi warna yang dapat memudahkan pembacaan pola distribusi prediksi antar kelas secara informatif. Pendekatan visualisasi berbasis warna ini sejalan dengan prinsip desain visualisasi informasi, bahwa penggunaan warna sebagai elemen visual memiliki peran yang sangat penting dalam membantu pembaca mengidentifikasi pola dan hubungan antar data secara lebih cepat dan akurat [26]. Pada visualisasi confusion matrix dalam bentuk *heatmap*, di mana gradasi warna membantu peneliti mengidentifikasi pola kesalahan klasifikasi secara lebih cepat, misalnya kecenderungan model dalam mengklasifikasikan berita hoaks sebagai nonhoaks atau sebaliknya, sehingga analisis performa model dapat dilakukan secara lebih sistematis.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen yang diperoleh dari penerapan model IndoBERT dalam klasifikasi berita hoaks, serta pembahasan terhadap temuan yang dihasilkan. Evaluasi dilakukan untuk mengukur performa model berdasarkan metrik yang relevan, seperti akurasi, precision, recall, dan f1-score. Selain itu, analisis juga difokuskan pada pengaruh perbedaan teknik preprocessing, yaitu berbasis *regex* dan *cosine similarity*, terhadap kinerja model.

3.1 Hasil Model IndoBERT

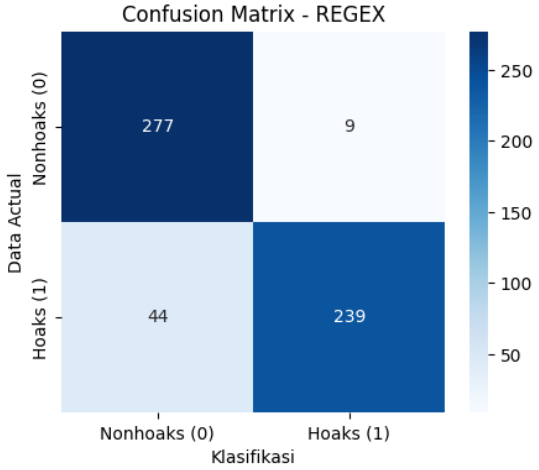
Pada tahap ini, dilakukan evaluasi terhadap performa model IndoBERT dalam melakukan klasifikasi berita hoaks pada dataset yang telah melalui proses *preprocessing* berbasis *regular expression* (*regex*) dan *cosine similarity*. Model yang digunakan adalah IndoBERT *base p1*, dengan proses pelatihan hingga maksimum 20 *epoch* dan menerapkan mekanisme *early stopping* dengan nilai *patience* sebesar 2 untuk mencegah terjadinya *overfitting*. Proses pelatihan menggunakan *optimizer* AdamW dengan *learning rate* sebesar 2e-5.

Tabel 6. Hasil Evaluasi Model

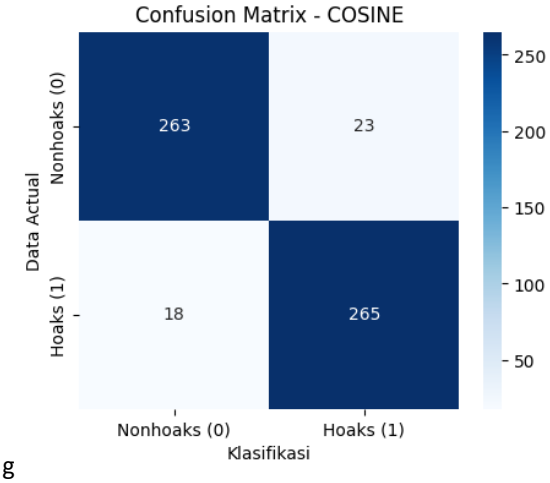
Eksperimen	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Regular Expression (*Regex*)	90.7%	91.3%	90.7%	90.6%
Cosine Similarity	92.8%	92.8%	92.8%	92.8%

Evaluasi model dilakukan dengan menggunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, dan *f1-score*. Berdasarkan hasil pengujian yang ditunjukkan pada Tabel 6, model yang dilatih dengan menggunakan dataset *preprocessing* berbasis *regular expression* (*regex*) memperoleh nilai *accuracy* sebesar 90.7%, *precision* sebesar 91.3%, *recall* sebesar 90.7%, dan *f1-score* sebesar 90.6%. Sementara itu, model yang menggunakan dataset *preprocessing* berbasis *cosine similarity* menunjukkan performa yang lebih tinggi dan lebih konsisten, dengan nilai *accuracy*, *precision*, *recall*, dan *f1-score* masing-masing sebesar 92.8%.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 3. *Confusion Matrix* Dataset *Regular Expression (Regex)*



g

Gambar 4. *Confusion Matrix* Dataset *Cosine Similarity*

Untuk memperoleh gambaran yang lebih rinci terkait performa model, dilakukan analisis menggunakan *confusion matrix* pada masing-masing dataset. Berdasarkan hasil *confusion matrix* pada dataset *preprocessing* berbasis *regular expression* (*regex*) yang ditunjukkan pada Gambar 3, model mampu mengklasifikasikan 277 data nonhoaks dan 239 data hoaksGar dengan benar. Namun, masih terdapat kesalahan klasifikasi, yaitu sebanyak 9 data nonhoaks yang diprediksi sebagai hoaks (*false positive*) dan 44 data hoaks yang diprediksi sebagai nonhoaks (*false negative*). Hal ini menunjukkan bahwa model cenderung lebih sering gagal mendeteksi berita hoaks dibandingkan kesalahan dalam mengklasifikasikan berita nonhoaks. Sementara itu, pada dataset berbasis *cosine similarity* yang ditunjukkan pada Gambar 4, model berhasil mengklasifikasikan 263 data nonhoaks dan 265 data hoaks dengan benar. Jumlah kesalahan yang terjadi relatif lebih kecil, yaitu 23 data nonhoaks yang diprediksi sebagai hoaks dan 18 data hoaks yang diprediksi sebagai nonhoaks. Dibandingkan dengan pendekatan *regex*, model pada dataset yang diproses menggunakan *preprocessing* berbasis *cosine similarity* menunjukkan kemampuan yang lebih baik dalam mengidentifikasi berita hoaks, yang ditunjukkan dengan jumlah *false negative* yang lebih rendah. Secara umum, hasil ini menunjukkan bahwa kedua pendekatan *preprocessing* mampu menghasilkan performa yang baik pada model IndoBERT, namun dengan pendekatan *cosine similarity* memberikan hasil yang lebih optimal.

3.2 Analisis Performa Preprocessing Regex dan Cosine Similarity

Perbandingan performa dilakukan untuk menganalisis pengaruh dua pendekatan preprocessing, yaitu berbasis *regular expression* (*regex*) dan *cosine similarity*, terhadap kinerja model IndoBERT dalam klasifikasi berita hoaks. Berdasarkan hasil eksperimen yang telah diperoleh, terlihat adanya perbedaan performa antara kedua pendekatan tersebut pada seluruh metrik evaluasi yang digunakan. Model yang dilatih menggunakan dataset hasil preprocessing berbasis cosine similarity menunjukkan performa yang lebih tinggi dibandingkan dengan regex. Hal ini ditunjukkan oleh nilai accuracy, precision, recall, dan f1-score yang secara konsisten mencapai 92.8%, sementara pada pendekatan regex nilai metrik berada pada kisaran 90.6% hingga 91.3%. Perbedaan ini mengindikasikan bahwa pendekatan cosine similarity mampu menghasilkan representasi data yang lebih sesuai untuk proses klasifikasi oleh model.

Berdasarkan hasil analisis terhadap *confusion matrix* dari masing-masing pendekatan yaitu *preprocessing* berbasis regex dan *preprocessing* berbasis *cosine similarity*, hasil menunjukkan bahwa keunggulan *preprocessing* berbasis *cosine*

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9730
Hal 823-835
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

similarity terutama berasal dari peningkatan kemampuan dalam mengidentifikasi berita hoaks. Jumlah prediksi benar pada kelas hoaks meningkat dari 239 pada pendekatan *regex* menjadi 265 pada *cosine similarity*, atau mengalami kenaikan sekitar 10.88%. Selain itu, kesalahan klasifikasi hoaks (*false negative*) berkurang secara signifikan, dari 44 menjadi 18 kasus, yang setara dengan penurunan sebesar 59.09%. Temuan ini menunjukkan bahwa pendekatan *preprocessing* berbasis *cosine similarity* lebih efektif dalam menangkap variasi makna dan pola klarifikasi serta klaim yang sering muncul pada teks berita hoaks.

Secara keseluruhan, hasil perbandingan ini menunjukkan bahwa *preprocessing* berbasis *cosine similarity* lebih adaptif dalam menangani kompleksitas bahasa dan variasi pola pada dataset berbasis fact-checking, khususnya dalam mengidentifikasi kalimat yang mengandung klaim serta kalimat klarifikasi yang tidak sesuai dengan fakta. Di sisi lain, pendekatan regex tetap memberikan keunggulan dalam menjaga kestabilan klasifikasi pada data nonhoaks, namun kurang fleksibel dalam menghadapi variasi pola berita hoaks. Pendekatan berbasis kemiripan semantik memiliki potensi yang lebih baik dalam menangani variasi bahasa pada teks berita, dibandingkan dengan pendekatan berbasis pola yang bersifat lebih kaku. Dengan demikian, teknik *preprocessing* tidak hanya bergantung pada capaian metrik evaluasi, tetapi juga pada kebutuhan sistem dalam menyeimbangkan sensitivitas deteksi hoaks dan ketepatan klasifikasi secara keseluruhan.

3.3 Analisis Dampak Preprocessing

Perbedaan performa yang dihasilkan dari penerapan *preprocessing* berbasis *regular expression* (*regex*) dan *cosine similarity* menunjukkan bahwa tahapan preprocessing tidak hanya berfungsi sebagai proses pembersihan data, tetapi juga memiliki peran yang sangat penting dalam membentuk kualitas representasi teks yang akan dipelajari oleh model. Dalam konteks klasifikasi berita hoaks, karakteristik data yang kompleks terutama keberadaan kalimat klarifikasi serta klaim yang tidak sesuai dengan fakta menjadikan proses preprocessing sebagai faktor yang dapat secara langsung memengaruhi arah dan kualitas pembelajaran model.

Secara umum, berita hoaks tidak hanya ditandai oleh adanya informasi yang salah, tetapi juga sering kali memuat klaim yang terlihat meyakinkan namun pada kenyataannya tidak sepenuhnya didukung oleh narasi yang benar ataupun fakta yang valid. Klaim semacam ini biasanya disusun sedemikian rupa agar menyerupai informasi yang kredibel, sehingga sulit dibedakan secara kasat mata. Dalam banyak kasus, dataset berbasis *fact-checking* juga menyertakan bagian klarifikasi yang secara eksplisit menjelaskan bahwa klaim tersebut tidak benar. Kombinasi antara klaim yang menyesatkan dan kalimat klarifikasi inilah yang dapat menimbulkan bias dalam proses pembelajaran model apabila tidak ditangani dengan tepat pada tahap preprocessing.

Pendekatan *regular expression* (*regex*) dalam penelitian ini diterapkan dengan memanfaatkan pola-pola tertentu yang telah ditentukan sebelumnya, seperti kata kunci atau frasa yang sering muncul dalam kalimat klarifikasi. Metode ini memiliki keunggulan dalam hal kesederhanaan dan kemudahan implementasi, serta cukup efektif dalam menghapus bagian teks yang secara eksplisit mengandung indikasi klarifikasi. Namun demikian, sifatnya yang berbasis aturan (*rule-based*) membuat pendekatan ini cenderung kaku dan bergantung pada kelengkapan pola yang dirancang. Variasi kalimat klarifikasi yang tidak sesuai dengan pola yang telah ditentukan berpotensi tidak terdeteksi, sehingga masih tersisa dalam dataset dan berpotensi memengaruhi proses pembelajaran model. Sebaliknya, pendekatan *cosine similarity* memanfaatkan representasi semantik dari teks untuk mengidentifikasi kemiripan antara kalimat dalam dataset dan template klarifikasi. Dengan menggunakan representasi berbasis *embedding*, metode ini mampu menangkap hubungan makna antar kalimat, tidak hanya sekedar kesamaan kata. Hal ini memungkinkan sistem untuk mengenali kalimat klarifikasi yang memiliki makna serupa meskipun disampaikan dengan struktur atau pilihan kata yang berbeda. Selain itu, pendekatan ini juga lebih adaptif dalam menghadapi variasi bahasa yang umum ditemukan pada teks berita, termasuk dalam mengidentifikasi klaim yang disusun secara implisit namun tetap memiliki kesamaan konteks dengan pola klarifikasi.

Dampak dari perbedaan pendekatan tersebut terlihat pada hasil klasifikasi model. Berdasarkan analisis *confusion matrix*, model yang dilatih menggunakan dataset hasil *preprocessing* berbasis *cosine similarity* menunjukkan jumlah *false negative* yang lebih rendah dibandingkan dengan pendekatan *regular expression* (*regex*). Hal ini menunjukkan bahwa model lebih mampu mengenali berita hoaks secara tepat, termasuk yang mengandung klaim yang tidak sesuai dengan fakta namun tidak selalu disertai dengan pola klarifikasi yang eksplisit. Sebaliknya, pendekatan *regular expression* (*regex*) masih menunjukkan kecenderungan kesalahan dalam bentuk *false negative*, yang mengindikasikan bahwa beberapa berita hoaks dengan variasi klaim dan kalimat klarifikasi tertentu belum dapat diidentifikasi secara optimal. Kondisi ini juga berkaitan dengan potensi terjadinya bias dalam dataset. Kalimat klarifikasi yang secara eksplisit menyatakan bahwa suatu informasi adalah hoaks dapat menjadi sinyal langsung bagi model dalam menentukan label, tanpa benar-benar memahami isi atau konteks klaim yang disampaikan. Jika kalimat tersebut tidak dihapus secara efektif, model cenderung mempelajari pola yang bersifat dangkal. Dalam hal ini, pendekatan *cosine similarity* yang lebih mampu mendeteksi variasi kalimat klarifikasi berkontribusi dalam mengurangi keberadaan pola eksplisit tersebut, sehingga model terdorong untuk lebih fokus pada pemahaman isi klaim dan kesesuaiannya dengan narasi yang benar.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa pemilihan teknik *preprocessing* memiliki dampak yang cukup signifikan terhadap performa model klasifikasi. Pendekatan berbasis *cosine similarity* terbukti lebih efektif dalam menangani kompleksitas bahasa pada dataset berita hoaks, terutama dalam mengidentifikasi kalimat klarifikasi pada dataset berita. Sementara itu, pendekatan *regular expression* (*regex*) tetap memiliki keunggulan dari sisi kesederhanaan, namun memerlukan perancangan pola yang lebih komprehensif agar dapat menangkap variasi pola pada teks berita yang lebih luas. Dengan demikian, pemahaman terhadap karakteristik data, khususnya terkait kalimat

klarifikasi pada teks berita yang dapat menyebabkan bias pada proses klasifikasi berita hoaks, menjadi aspek penting dalam merancang strategi *preprocessing* yang optimal.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa teknik *preprocessing* memiliki peran yang signifikan dalam memengaruhi performa model IndoBERT dalam klasifikasi berita hoaks berbahasa Indonesia. Perbandingan antara pendekatan berbasis *regular expression* (*regex*) dan *cosine similarity* memperlihatkan adanya perbedaan kinerja yang konsisten pada berbagai metrik evaluasi. Berdasarkan hasil eksperimen, model yang menggunakan *preprocessing* berbasis *regular expression* (*regex*) memperoleh nilai *accuracy* sebesar 90.7%, *precision* sebesar 91.3%, *recall* sebesar 90.7%, dan *f1-score* sebesar 90.6%. Sementara itu, pendekatan berbasis *cosine similarity* menunjukkan performa yang lebih tinggi dengan nilai *accuracy*, *precision*, *recall*, dan *f1-score* masing-masing sebesar 92.8%. Temuan ini menunjukkan bahwa pendekatan berbasis *cosine similarity* lebih efektif dalam menangani kompleksitas bahasa pada teks berita, khususnya dalam mengidentifikasi kalimat klarifikasi dan pola klaim yang terdapat pada dataset berita hoaks berbasis *fact-checking*. C*osine similarity* mampu mendeteksi variasi kalimat yang memiliki makna serupa meskipun berbeda secara struktur, sehingga proses pembersihan data menjadi lebih adaptif. Sebaliknya, pendekatan *regex* yang berbasis pola menunjukkan keterbatasan dalam menangkap variasi bahasa yang lebih luas, sehingga masih menyisakan potensi bias dalam dataset. Selain itu, hasil penelitian juga menunjukkan bahwa penghapusan kalimat klarifikasi yang lebih efektif dapat membantu mengurangi potensi bias pada dataset, sehingga model tidak hanya bergantung pada pola eksplisit dalam menentukan label, tetapi juga terdorong untuk memahami isi dan konteks berita secara lebih mendalam. Hal ini berdampak pada peningkatan kemampuan model dalam mengklasifikasikan berita hoaks, terutama pada kasus yang tidak memiliki indikator yang jelas. Secara keseluruhan, kontribusi utama penelitian ini terletak pada analisis terhadap dua pendekatan *preprocessing* dalam konteks klasifikasi berita hoaks menggunakan IndoBERT, serta penekanan pada pentingnya penanganan kalimat klarifikasi sebagai sumber potensi bias dalam dataset. Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan metode *preprocessing* yang lebih efektif, khususnya dalam penelitian yang berkaitan dengan deteksi hoaks dan pemrosesan teks berbahasa Indonesia.

REFERENCES

[1] F. L. Gaol, A. Maulana, and T. Matsuo, “News consumption patterns on Twitter: fragmentation study on the online news media network,” *Heliyon*, vol. 6, no. 10, Oct. 2020, doi: 10.1016/j.heliyon.2020.e05169.

[2] D. Gaozhao, “Flagging fake news on social media: An experimental study of media consumers’ identification of fake news,” *Gov. Inf. Q.*, vol. 38, no. 3, p. 101591, Jul. 2021, doi: 10.1016/j.giq.2021.101591.

[3] S. Chen, L. Xiao, and A. Kumar, “Spread of misinformation on social media: What contributes to it and how to combat it,” *Comput. Human Behav.*, vol. 141, p. 107643, Apr. 2023, doi: 10.1016/j.chb.2022.107643.

[4] S. A. Althubiti, F. Alenezi, and R. F. Mansour, “Natural Language Processing with Optimal Deep Learning Based Fake News Classification,” *Computers, Materials and Continua*, vol. 73, no. 2, pp. 3529–3544, 2022, doi: 10.32604/cmc.2022.028981.

[5] A. Tabassum and R. R. Patil, “A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing,” *International Research Journal of Engineering and Technology*, 2020, [Online]. Available: www.irjet.net

[6] F. Gereme, W. Zhu, T. Ayall, and D. Alemu, “Combating fake news in ‘low-resource’ languages: Amharic fake news detection accompanied by resource crafting,” *Information (Switzerland)*, vol. 12, no. 1, pp. 1–9, Jan. 2021, doi: 10.3390/info12010020.

[7] A. Pardamean and H. F. Pardede, “Tuned bidirectional encoder representations from transformers for fake news detection,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1667–1671, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1667-1671.

[8] M. Y. Ridho and E. Yulianti, “From Text to Truth: Leveraging IndoBERT and Machine Learning Models for Hoax Detection in Indonesian News,” *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 10, no. 3, pp. 544–555, Sep. 2024, doi: 10.26555/jiteki.v10i3.29450.

[9] S. M. Isa, G. Nico, and M. Permana, “IndoBERT for Indonesian Fake News Detection,” *ICIC Express Letters*, vol. 16, no. 3, pp. 289–297, Mar. 2022, doi: 10.24507/icicel.16.03.289.

[10] C. Jocelynne, L. Tobing, I. Lanang Wijayakusuma, L. Putu, and I. Harini, “Detection of Political Hoax News Using Fine-Tuning IndoBERT,” 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] 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.

[12] Syarifah Ema Rahmaniah, Septiaji Eko Nugroho, Rupita, and Nikodemus Niko, “The Disinfodemic Mitigation Strategy of MAFINDO in Indonesia,” *International Journal of Social Science*, vol. 1, no. 6, pp. 879–888, Apr. 2022, doi: 10.53625/ijss.v1i6.1903.

[13] Q. Chen, A. Banerjee, Ç. Demiralp, G. Durrett, and I. Dillig, “Data Extraction via Semantic Regular Expression Synthesis,” *Proceedings of the ACM on Programming Languages*, vol. 7, no. OOPSLA2, Oct. 2023, doi: 10.1145/3622863.

[14] 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.

[15] D. Rahmawan, I. Garnesia, and R. Hartanto, "Content Analysis of MAFINDO's Fact Check articles during the 2015-2020 period: Classification of Themes, Channels, and Content Types," *Jurnal ASPIKOM*, vol. 8, no. 2, Jul. 2023, doi: 10.24329/aspikom.v8i2.1267.

[16] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi, "A Hybrid Transformer-Based Model for Optimizing Fake News Detection," *IEEE Access*, vol. 12, pp. 160822–160834, 2024, doi: 10.1109/ACCESS.2024.3476432.

[17] A. Kunaefi, Z. Abidin, and R. Kusumawati, "Klasifikasi Berita Hoaks Bahasa Indonesia Menggunakan IndoBERT Fine-Tuning dengan Pendekatan Focal Loss pada Data Tidak Seimbang," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 2, pp. 1706–1714, May 2025, doi: 10.29100/jipi.v10i2.7811.

[18] Z. Nagy, *Regex Quick Syntax Reference*. Berkeley, CA: Apress, 2018. doi: 10.1007/978-1-4842-3876-9.

[19] D. Iskandar and A. Kurniawati, "Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 133–144, Feb. 2025, doi: 10.25126/jtiik.2025129143.

[20] K. M. Kahloot and P. Ekler, "Algorithmic Splitting: A Method for Dataset Preparation," *IEEE Access*, vol. 9, pp. 125229–125237, 2021, doi: 10.1109/ACCESS.2021.3110745.

[21] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, 2024, doi: 10.1155/2024/2826773.

[22] L. F. Simanjuntak, R. Mahendra, and E. Yulianti, "We Know You Are Living in Bali: Location Prediction of Twitter Users Using BERT Language Model," *Big Data and Cognitive Computing*, vol. 6, no. 3, Sep. 2022, doi: 10.3390/bdcc6030077.

[23] F.-E. Lagrari and Y. Elkettani, "Customized BERT with Convolution Model A New Heuristic Enabled Encoder for Twitter Sentiment Analysis," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020, [Online]. Available: www.ijacsa.thesai.org

[24] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[25] A. J. Keya, Md. A. H. Wadud, M. F. Mridha, M. Alatiyyah, and Md. A. Hamid, "AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification," *Applied Sciences*, vol. 12, no. 17, p. 8398, Aug. 2022, doi: 10.3390/app12178398.

[26] Z. Abidin, R. Munir, S. Akbar, R. Mandala, and D. H. Widyantoro, "Storychart: A Character Interaction Chart for Visualizing the Activities Flow," *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION*, Dec. 2023, [Online]. Available: www.joiv.org/index.php/joiv