MENGATASI KETERBATASAN KONTEKS TEKS PANJANG PADA DETEKSI HOAKS MENGGUNAKAN INDOBERT DENGAN PENDEKATAN *SLIDING WINDOW*

SKRIPSI

Diajukan Untuk Memenuhi Sebagian Syarat Guna

Memperoleh Gelar Sarjana Komputer (S.Kom.)

Pada Program Studi Sistem Informasi

OLEH :

MOCHAMAD ABDUL AZIS

NPM: 2213030094

FAKULTAS TEKNIK DAN ILMU KOMPUTER (FTIK)

UNIVERSITAS NUSANTARA PERSATUAN GURU REPUBLIK INDONESIA

UN PGRI KEDIRI

2026

HALAMAN PERSETUJUAN

Skripsi oleh :

MOCHAMAD ABDUL AZIS

NPM : 2213030094

Judul

MENGATASI KETERBATASAN KONTEKS TEKS PANJANG PADA DETEKSI HOAKS MENGGUNAKAN INDOBERT DENGAN PENDEKATAN *SLIDING WINDOW*

Telah disetujui untuk diajukan Kepada

Panitia Ujian/Sidang Skripsi Program Studi Sistem Informasi

FTIK UN PGRI Kediri

Tanggal: 11 Juni 2026

Pembimbing I	Pembimbing II
Dr. Sucipto, M.Kom	Aidina Ristyawan, M.Kom
NIDN. 0721029101	NIDN. 0721018801

HALAMAN PENGESAHAN

Skripsi oleh :

MOCHAMAD ABDUL AZIS

NPM : 2213030094

Judul

MENGATASI KETERBATASAN KONTEKS TEKS PANJANG PADA DETEKSI HOAKS MENGGUNAKAN INDOBERT DENGAN PENDEKATAN *SLIDING WINDOW*

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi

Program Studi Sistem Informasi FTIK UN PGRI Kediri

Tanggal: 23 Juni 2026

Dan Dinyatakan telah Memenuhi Persyaratan

Panitia Penguji :

1. Ketua : Dr. Sucipto, S.Kom, M.Kom .....
2. Penguji 1 : Anita Sari Wardani, S.Kom, M.Kom .....
3. Penguji 2 : Aidina Ristyawan, S.Kom, M.Kom .....

Mengetahui,

Dekan Fakultas Teknik dan Ilmu Komputer

Dr. Sulistiono, M.Si

NIDN. 0007076801

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini saya,

Nama	: Mochamad Abdul Azis
Jenis Kelamin	: Laki-laki
Tempat/tgl. lahir	: Kediri/ 11 Maret 2004
NPM	: 2213030094
Fak/Prodi.	: FTIK/ S1-Sistem Informasi

menyatakan dengan sebenarnya, bahwa dalam Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya tidak terdapat karya tulis atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara sengaja dan tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Kediri, 11 Juni 2026

Yang Menyatakan

MOCHAMAD ABDUL AZIS

NPM: 2213030094

MOTTO

"Hanya Tidak Mudah, Bukan Tidak Mungkin"

ABSTRAK

Mochamad Abdul Azis: Penyebaran berita hoaks di media digital semakin meningkat seiring bertambahnya penggunaan platform berita dan media sosial sebagai sumber informasi utama masyarakat. Deteksi otomatis terhadap berita hoaks menjadi tantangan penting, terutama pada dokumen berita panjang yang melebihi batas maksimum token model Transformer seperti IndoBERT. Pendekatan *truncation* yang umum digunakan berpotensi menyebabkan hilangnya informasi penting pada bagian akhir dokumen sehingga dapat mempengaruhi kualitas klasifikasi model.

Penelitian ini bertujuan untuk menganalisis efektivitas pendekatan *sliding window* dalam mengatasi keterbatasan panjang token pada deteksi berita hoaks berbahasa Indonesia menggunakan IndoBERT. Penelitian dilakukan menggunakan metodologi CRISP-DM dengan tahapan *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Eksperimen dilakukan menggunakan empat skenario utama, yaitu *truncation* dan *sliding window* yang masing-masing dikombinasikan dengan *full fine-tuning* dan *partial fine-tuning*. Pada pendekatan *sliding window*, hasil prediksi antar segmen teks digabungkan menggunakan metode *mean aggregation*.

Hasil penelitian menunjukkan bahwa pendekatan *sliding window* dengan *full fine-tuning* menjadi konfigurasi terbaik, khususnya pada dokumen dengan panjang melebihi 512 token. Pada pengujian dokumen panjang, model memperoleh nilai akurasi sebesar 0,9897, *precision* sebesar 1,0000, *recall* sebesar 0,9697, dan F1-*score* sebesar 0,9846. Hasil tersebut menunjukkan bahwa pendekatan *sliding window* mampu mempertahankan konteks informasi lebih baik dibandingkan *truncation*. Selain itu, model berhasil diimplementasikan ke dalam sistem berbasis FastAPI dan Telegram Bot untuk melakukan prediksi berita secara daring.

Penelitian ini menyimpulkan bahwa kombinasi IndoBERT, *sliding window*, dan *mean aggregation* efektif digunakan untuk deteksi berita hoaks pada teks panjang berbahasa Indonesia. Namun, penelitian ini masih memiliki keterbatasan pada penggunaan IndoBERT dengan batas token tetap serta metode agregasi yang masih sederhana. Penelitian selanjutnya disarankan untuk mengeksplorasi model Transformer khusus dokumen panjang seperti Longformer atau BigBird, serta mengembangkan metode agregasi dan sistem verifikasi fakta yang lebih adaptif dan kontekstual.

Kata Kunci: Deteksi Hoaks, IndoBERT, *Sliding Window*, *Mean Aggregation*, *Long Text Classification*.

KATA PENGANTAR

Puji Syukur Kami panjatkan Ke hadirat Allah Tuhan Yang Maha Kuasa, karena hanya atas perkenan-Nya penyusunan skripsi ini dapat diselesaikan. Skripsi dengan judul "Mengatasi Keterbatasan Konteks Teks Panjang pada Deteksi Hoaks Menggunakan IndoBERT dengan Pendekatan *Sliding Window*" ini ditulis guna memenuhi sebagian syarat untuk memperoleh gelar Sarjana Komputer, pada Program Studi Sistem Informasi Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri. Pada kesempatan ini diucapkan terima kasih dan penghargaan yang setulus-tulusnya kepada:

1. Dr. Zainal Afandi, M.Pd. Selaku Rektor UN PGRI Kediri.
2. Dr. Sulistiono, M.Si. selaku Dekan Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri yang selalu memberikan dorongan motivasi kepada mahasiswa.
3. Dr. Sucipto, M.Kom. selaku Ka Prodi Sistem Informasi Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri sekaligus pembimbing I yang selalu memberikan dorongan motivasi kepada mahasiswa.
4. Aidina Ristyawan, M.Kom. selaku Pembimbing skripsi II yang dengan sabar membimbing dan memotivasi dalam menyelesaikan skripsi ini.
5. Kedua orang tua penulis yang senantiasa memberikan doa, dukungan moral, dukungan materi, serta motivasi selama proses penyusunan skripsi ini. Segala bentuk kesabaran, kepercayaan, dan pengorbanan yang diberikan menjadi salah satu kekuatan utama bagi penulis untuk menyelesaikan penelitian ini dengan sebaik-baiknya.
6. Teman-teman seperjuangan yang telah memberikan dukungan, semangat, bantuan, serta ruang diskusi selama proses penyusunan skripsi ini. Kehadiran mereka membantu penulis melewati berbagai kendala akademik maupun teknis yang muncul selama penelitian.
7. Para peneliti dan pengembang Transformer, BERT, dan IndoBERT yang telah memberikan kontribusi besar dalam perkembangan bidang Natural Language Processing. Konsep dan model yang dikembangkan tersebut menjadi dasar penting dalam penelitian ini, khususnya dalam penerapan model berbasis bahasa untuk deteksi berita hoaks.

8. Charlotte Jocelynne beserta rekan penulisnya, yang melalui artikel penelitiannya telah menjadi salah satu rujukan utama dalam penyusunan skripsi ini. Penelitian tersebut memberikan dasar pemahaman dan referensi penting bagi penulis dalam mengembangkan penelitian terkait deteksi berita hoaks menggunakan IndoBERT.
9. Seseorang dengan NPM 2314040008 yang telah memberikan dukungan, semangat, dan motivasi secara tidak langsung selama proses penyusunan skripsi ini. Kehadirannya menjadi salah satu dorongan bagi penulis untuk tetap bertahan dan menyelesaikan penelitian ini hingga akhir.

Disadari bahwa skripsi ini masih banyak kekurangan, maka diharapkan tegur sapa, kritik dan saran-saran, dari berbagai pihak sangat diharapkan. Akhirnya, disertai harapan semoga skripsi ini ada manfaatnya bagi kita semua, khususnya bagi dunia pendidikan, meskipun hanya ibarat setitik air bagi samudra yang luas.

Kediri, 11 Juni 2026

MOCHAMAD ABDUL AZIS

NPM: 2213030094

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

A. Latar Belakang

Media sosial menggantikan peran media tradisional sebagai kanal distribusi informasi utama masyarakat saat ini (Shu et al., 2017) Berdasarkan laporan digital terbaru, jumlah pengguna platform digital di Indonesia menyentuh angka 180 juta individu (Kemp, 2025). Peningkatan aksesibilitas ini berdampak pada perputaran data yang sangat cepat dan tidak terkendali di ruang siber. Ketergantungan terhadap sumber digital memicu munculnya risiko penyebaran berita tanpa melalui proses verifikasi fakta yang ketat. Populasi pengguna yang besar menjadi target potensial bagi sirkulasi narasi manipulatif yang merugikan publik. Kondisi tersebut menuntut adanya mekanisme pengawasan informasi yang lebih efisien, transparan, dan terukur.

Narasi palsu atau hoaks dirancang secara sengaja untuk mengelabui persepsi pembaca terhadap suatu fenomena tertentu (Tandoc et al., 2018). Distribusi hoaks sering kali ditujukan untuk memicu polarisasi politik serta mendapatkan keuntungan finansial bagi pihak tertentu. Ekosistem informasi digital menjadi terancam akibat banyaknya konten yang memuat data tidak akurat dan menyesatkan. Manipulasi opini publik melalui berita bohong dapat menyebabkan ketidakstabilan sosial dalam skala yang sangat luas. Karakteristik teks hoaks yang menyerupai berita asli menyulitkan proses identifikasi secara manual oleh masyarakat umum. Perlindungan terhadap kualitas informasi digital menjadi agenda esensial dalam menjaga ketertiban sirkulasi pengetahuan di ruang publik.

Kementerian Komunikasi dan Digital mencatat temuan ribuan konten hoaks yang tersebar masif selama periode 2024 (Kementerian Komunikasi dan Digital, 2025). Data resmi menunjukkan bahwa kategori penipuan mendominasi total konten menyesatkan yang berhasil diidentifikasi oleh pemerintah. Peningkatan jumlah temuan ini mengonfirmasi bahwa pola penyebaran berita palsu semakin terorganisir dan sistematis. Meskipun upaya pemblokiran dilakukan, kecepatan produksi konten baru tetap menjadi tantangan besar bagi sistem keamanan informasi. Penanganan secara manual oleh tim verifikatur tidak lagi sebanding

dengan pertumbuhan volume data digital yang masuk setiap detik. Oleh sebab itu, implementasi teknologi pemrosesan bahasa alami dipandang sebagai solusi teknis yang paling relevan.

Pengembangan sistem klasifikasi otomatis telah melalui berbagai tahapan evaluasi algoritme pembelajaran mesin selama beberapa tahun terakhir (Rahmat et al., 2019). Metode tradisional seperti *Support Vector Machine* atau *Random Forest* sempat menjadi pilihan utama dalam riset deteksi teks berita. Namun, pendekatan tersebut memiliki keterbatasan fundamental dalam memahami hubungan makna asli antar kata yang kompleks (Sabir et al., 2025). Representasi kata berbasis frekuensi kemunculan sering kali mengabaikan urutan kalimat serta konteks semantik secara utuh. Model leksikal klasik cenderung gagal membedakan gaya bahasa sarkasme atau sindiran halus yang sering ditemukan dalam narasi hoaks. Kegagalan menangkap hubungan antar kata menurunkan akurasi prediksi pada dokumen berita yang memiliki struktur bahasa rumit.

Arsitektur Transformer hadir untuk memperbaiki kelemahan model pembelajaran mesin konvensional melalui mekanisme *self-attention* (De Santis et al., 2025). Teknologi ini memungkinkan mesin melakukan ekstraksi fitur teks secara otomatis tanpa membutuhkan intervensi manual yang berat. Representasi kata yang dihasilkan bersifat kontekstual sehingga mampu mengenali maksud kata berdasarkan kalimat di sekitarnya (Dhiman et al., 2024). Model ini terbukti sangat efektif untuk tugas-tugas klasifikasi teks yang membutuhkan penalaran bahasa yang mendalam. Keunggulan BERT terletak pada kemampuannya membaca teks secara dua arah guna memahami intensi penulis berita secara lebih akurat. Penggunaan arsitektur ini memberikan landasan baru bagi peningkatan performa sistem keamanan informasi berbasis kecerdasan buatan.

Efektivitas model BERT pada bahasa tertentu sangat dipengaruhi oleh sumber data yang digunakan saat proses pelatihan awal dilakukan. Model IndoBERT dikembangkan secara khusus untuk menangani karakteristik unik tata bahasa Indonesia secara lebih spesifik (Koto et al., 2020). Pelatihan dilakukan menggunakan miliaran kata dari sumber terpercaya seperti Wikipedia Indonesia, media massa, dan korpus web lokal. Hasil penelitian menunjukkan bahwa

IndoBERT memiliki performa yang lebih stabil dibandingkan dengan varian model multibahasa lainnya (Pekandi et al., 2025). Pemahaman terhadap nuansa budaya dan diksi lokal menjadikan model ini sangat unggul dalam menjalankan tugas klasifikasi teks domestik. Penggunaan model terlatih ini mempercepat proses adaptasi sistem deteksi hoaks terhadap tren bahasa masyarakat Indonesia terkini.

Meskipun memiliki performa tinggi, arsitektur IndoBERT dibatasi oleh kapasitas memori maksimal sebesar 512 token masukan pada setiap prosesnya. Banyak berita politik di Indonesia memiliki panjang narasi yang melampaui batasan teknis perangkat lunak tersebut secara signifikan. Implementasi standar biasanya menerapkan metode *truncation* untuk memangkas teks secara paksa tepat di batas akhir token maksimal. Pemotongan informasi tersebut mengakibatkan hilangnya data penting yang mungkin tersimpan pada bagian tengah atau akhir dokumen. Padahal, struktur penulisan berita sering kali menempatkan konklusi atau fakta pendukung pada paragraf penutup dokumen. Pengabaian terhadap sisa teks dokumen panjang berpotensi menurunkan kemampuan model dalam mendeteksi pola manipulasi informasi secara utuh.

Kelemahan IndoBERT dalam memproses narasi panjang telah diidentifikasi pada penelitian sebelumnya yang berfokus pada berita politik (C. J. L. Tobing et al., 2025b). Kesulitan dalam memahami konteks rumit muncul ketika model dihadapkan pada perubahan gaya bahasa yang tidak terduga dalam teks. Narasi hoaks sering kali menyisipkan klaim palsu di tengah dokumen guna mengecoh fokus utama dari pembaca berita. Fokus penelitian terdahulu yang terbatas pada metode pemotongan token menyisakan celah besar dalam penanganan teks dokumen panjang. Kurangnya strategi *fine-tuning* yang spesifik menyebabkan model belum optimal dalam melakukan klasifikasi terhadap berita yang sangat detail. Evaluasi terhadap integritas data pada dokumen yang dianalisis secara utuh menjadi kebutuhan fundamental untuk segera dikembangkan.

Berdasarkan telaah terhadap penelitian-penelitian terdahulu, sebagian besar penelitian deteksi berita hoaks berbahasa Indonesia masih berfokus pada peningkatan performa model klasifikasi menggunakan IndoBERT tanpa memberikan perhatian khusus terhadap permasalahan panjang dokumen yang

melebihi batas maksimum 512 token. Evaluasi yang dilakukan umumnya menggunakan representasi dokumen hasil *truncation* sehingga informasi pada bagian tengah maupun akhir dokumen berpotensi tidak ikut diproses oleh model. Di sisi lain, penelitian yang mengimplementasikan teknik penanganan dokumen panjang pada tugas klasifikasi teks sebagian besar masih dilakukan pada korpus berbahasa asing dan belum banyak dievaluasi pada kasus deteksi hoaks berbahasa Indonesia. Selain itu, kajian yang membandingkan secara langsung pendekatan *truncation* dan *sliding window* pada model IndoBERT, khususnya dalam kombinasi strategi *full fine-tuning* dan *partial fine-tuning*, masih sangat terbatas. Oleh karena itu, diperlukan penelitian yang secara khusus mengevaluasi efektivitas pendekatan *sliding window* dalam mempertahankan konteks informasi pada dokumen berita panjang berbahasa Indonesia.

Penelitian ini mengusulkan penerapan teknik *sliding window* untuk mengatasi hambatan batasan kapasitas token pada model IndoBERT. Seluruh bagian dokumen berita panjang akan dipecah menjadi beberapa segmen teks kecil yang memiliki irisan tertentu. Proses ini menjamin tidak ada satu pun informasi dalam teks yang terbuang selama tahapan pemrosesan data dilakukan oleh mesin. Nilai prediksi dari setiap potongan segmen kemudian digabungkan menggunakan metode *mean aggregation* untuk menghasilkan label tingkat dokumen. Penggunaan rata-rata probabilitas memungkinkan model memberikan penilaian akhir yang lebih adil dan representatif terhadap keseluruhan isi teks. Pendekatan ini diharapkan mampu merekonstruksi pemahaman kontekstual yang hilang pada penggunaan metode pemotongan teks konvensional.

Masalah ketidakseimbangan jumlah data hoaks dan fakta diantisipasi melalui modifikasi fungsi kerugian selama tahapan pelatihan berlangsung. Implementasi *class weights* digunakan untuk memberikan beban penalti yang lebih besar terhadap kesalahan prediksi pada kelas minoritas. Strategi ini dijalankan tanpa harus melakukan rekayasa atau manipulasi terhadap struktur makna asli dari teks berita yang diteliti. Prapemrosesan teks juga dilakukan secara ketat untuk menormalisasi penggunaan bahasa tidak baku melalui penggunaan kamus khusus. Penelitian ini bertujuan mengevaluasi perbandingan kinerja antara metode *sliding window* dan

truncation pada berbagai skema pelatihan model. Hasil akhir diharapkan memberikan kontribusi nyata bagi pengembangan standar baru dalam sistem klasifikasi teks panjang di Indonesia.

B. Batasan Masalah

1. Penelitian ini difokuskan pada deteksi berita hoaks berbahasa Indonesia menggunakan model IndoBERT (indobenchmark/indobert-base-p1) sebagai model utama klasifikasi teks.
2. Dataset yang digunakan berasal dari TurnBackHoax sebagai sumber data hoaks serta CNN Indonesia, Kompas, dan Tempo sebagai sumber data non-hoaks.
3. Penelitian ini membatasi penanganan teks panjang pada perbandingan antara pendekatan *truncation* dan *sliding window* untuk dokumen yang melebihi batas maksimum 512 token pada IndoBERT.
4. Metode agregasi pada pendekatan *sliding window* dibatasi pada penggunaan *mean aggregation* tanpa membandingkannya dengan metode agregasi lain, seperti *max pooling*, *majority voting*, maupun *attention-based aggregation*.
5. Penelitian ini menerapkan *class weight* sebagai pendekatan *cost-sensitive learning* untuk mengurangi pengaruh ketidakseimbangan distribusi kelas berdasarkan rekomendasi literatur, tanpa melakukan perbandingan dengan metode penanganan *class imbalance* lainnya, seperti *oversampling*, *undersampling*, maupun *Synthetic Minority Over-sampling Technique* (SMOTE).
6. Eksperimen penelitian dibatasi pada skenario *full fine-tuning* dan *partial fine-tuning* tanpa membandingkannya dengan model Transformer lain yang dirancang khusus untuk teks panjang, seperti Longformer atau BigBird.
7. Implementasi sistem dibatasi pada layanan prediksi berbasis FastAPI dan Telegram Bot yang di-*deploy* menggunakan Railway.
8. Fitur pencarian berita terkait dibatasi pada pemanfaatan sumber berita berbasis RSS dari beberapa media terpercaya dan belum menerapkan *semantic search* maupun sistem verifikasi fakta otomatis berbasis penelusuran web secara menyeluruh.

C. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, pertanyaan penelitian ini dirumuskan sebagai berikut:

1. Seberapa besar perbedaan skor evaluasi antara metode *truncation* 512 token dan metode *sliding window*?
2. Seberapa besar perbedaan kinerja (akurasi, presisi, *recall*, f1-*score*) antara skema *full fine-tuning* dan *partial fine-tuning* pada arsitektur IndoBERT?
3. Bagaimana pengaruh pendekatan sliding window terhadap kinerja IndoBERT pada dokumen berita yang melebihi batas maksimum 512 token?

D. Tujuan Penelitian

Tujuan yang ingin dicapai dalam pelaksanaan penelitian ini dijabarkan sebagai berikut:

1. Menganalisis perbedaan skor evaluasi antara metode *truncation* 512 token dan metode *sliding window* dalam proses klasifikasi teks menggunakan IndoBERT.
2. Membandingkan kinerja model (akurasi, presisi, recall, f1-*score*) antara skema *full fine-tuning* dan *partial fine-tuning* pada arsitektur IndoBERT dalam tugas klasifikasi teks.
3. Menganalisis pengaruh pendekatan *sliding window* terhadap kemampuan model dalam mempertahankan konteks informasi pada teks panjang.

E. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan kontribusi baik secara teoritis maupun praktis. Secara teoritis, penelitian ini berkontribusi terhadap pengembangan metode pengolahan dokumen panjang dalam bidang pemrosesan bahasa alami, khususnya pada konteks bahasa Indonesia. Selain itu, penelitian ini juga diharapkan dapat menjadi referensi akademis terkait efektivitas teknik agregasi rata-rata dalam meningkatkan sensitivitas model terhadap deteksi berita hoaks. Secara praktis, penelitian ini memberikan kontribusi melalui pengembangan pendekatan yang dapat digunakan sebagai sarana verifikasi berita yang lebih mudah diakses oleh masyarakat luas. Lebih lanjut, hasil penelitian ini diharapkan dapat mendukung upaya dalam menekan penyebaran informasi palsu serta meningkatkan kualitas ekosistem informasi digital di tingkat nasional.

DAFTAR PUSTAKA

Abodayeh, A., Shihadeh, L., Hejazi, R., Latif, R., & Najjar, W. (2023). *Web Scraping for Data Analytics: A BeautifulSoup Implementation*. https://doi.org/https://doi.org/10.1109/WiDS-PSU57071.2023.00025

Aliyah Salsabila, N., Ardhito Winatmoko, Y., Akbar Septiandri, A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *2018 International Conference on Asian Language Processing (IALP)*, 226–229. https://doi.org/https://doi.org/10.1109/IALP.2018.8629151

Alva Principe, R., Chiarini, N., & Viviani, M. (2025). Long Document Classification in the Transformer Era: A Survey on Challenges, Advances, and Open Issues. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 15, Number 2). John Wiley and Sons Inc. https://doi.org/https://doi.org/10.1002/widm.70019

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1). https://doi.org/https://doi.org/10.1186/s40537-025-01119-4

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What Does BERT Look At? An Analysis of BERT's Attention*. https://doi.org/https://doi.org/10.48550/arXiv.1906.04341

Daniati, E., Wibawa, A. P., Irianto, W. S. G., & Hernandez, L. (n.d.). *Few-Shot-BERT-RNN Narrative Structure Analysis for Andersen's Stories*. https://doi.org/http://dx.doi.org/10.62527/joiv.9.4.3932

De Santis, E., Martino, A., Ronci, F., & Rizzi, A. (2025). From Bag-of-Words to Transformers: A Comparative Study for Text Classification in Healthcare Discussions in Social Media. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *9*(1), 1063–1077. https://doi.org/https://doi.org/10.1109/TETCI.2024.3423444

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/https://doi.org/10.48550/arXiv.1810.04805

Dhiman, P., Kaur, A., Gupta, D., Juneja, S., Nauman, A., & Muhammad, G. (2024). GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection. *Heliyon*, *10*(16). https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e35865

Hamdikatama, B. (2025). Beyond Algorithms: An Integrated Approach to Fake News Detection Using Machine Learning Techniques. *Jurnal Ilmu Pengetahuan Dan Teknologi Komputer (JITK)*, *10*(3), 609–622. https://doi.org/https://doi.org/10.33480/jitk.v10i3.6061

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. https://doi.org/https://doi.org/10.48550/arXiv.1801.06146

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization*. https://doi.org/https://doi.org/10.48550/arXiv.2003.11080

Imron, A., Daniati, E., Firliana, R., Fauzan, M., & Mufid, A. (2025). ***Analisis Data Penjualan Pada UMKM Konveksi Menggunakan Metode K-Means Clustering dengan Menerapkan CRISP-DM*****. https://doi.org/https://doi.org/10.33005/sitasi.v5i1.2542**

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., & Zhang, W. (2019). *A Comparative Study on Transformer vs RNN in Speech Applications*. https://doi.org/https://doi.org/10.1109/ASRU46091.2019.9003750

Kementerian Komunikasi dan Digital. (2025). *Komdigi Identifikasi 1.923 Konten Hoaks Sepanjang Tahun 2024*. https://www.komdigi.go.id/berita/siaran-pers/detail/komdigi-identifikasi-1923-konten-hoaks-sepanjang-tahun-2024

Kemp, S. (2025). *Digital 2026: Top Digital and Social Media Trends in Indonesia*. https://wearesocial.com/id/blog/2025/11/digital-2026-top-digital-and-social-media-trends-in-indonesia/

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. https://doi.org/https://doi.org/10.48550/arXiv.2011.00677

Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization*. https://doi.org/https://doi.org/10.48550/arXiv.1711.05101

Lu, H., Xu, Y., Ye, M., Yan, K., Gao, Z., & Jin, Q. (2019). Learning misclassification costs for imbalanced classification on gene expression data. *BMC Bioinformatics*, *20*. https://doi.org/https://doi.org/10.1186/s12859-019-3255-x

Masood, M. A., Abbasi, R. A., & Keong, N. W. (2020). Context-Aware Sliding Window for Sentiment Classification. *IEEE Access*, *8*, 4870–4884. https://doi.org/https://doi.org/10.1109/ACCESS.2019.2963586

Nugroho, A., & Harini, D. (2024). Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang. *JSITIK*, *2*(2). https://doi.org/https://doi.org/10.53624/jsitik.v2i2.XX

Pappagari, R., Zelasko, P. ˙, Villalba, J., Carmiel, Y., & Dehak, N. (2019). *HIERARCHICAL TRANSFORMERS FOR LONG DOCUMENT CLASSIFICATION*. https://doi.org/https://doi.org/10.48550/arXiv.1910.10781

Pekandi, L. A., Widjaja, R. G., Ananta, A., Harefa, J., & Jingga, K. (2025). Evaluating IndoBERT for Indonesian Hoax News Detection: A Comparative Study with Ensemble and CNN-LSTM Models. *Procedia Computer Science*, *269*, 1625–1633. https://doi.org/https://doi.org/10.1016/j.procs.2025.09.105

Peltarion. (2020). *Structure of BERT*. https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/bert-encoder

Pes, B., & Lai, G. (2021). Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. *PeerJ Computer Science*, *7*. https://doi.org/https://doi.org/10.7717/PEERJ-CS.832

Rahmat, M. A., Indrabayu, & Areni, I. S. (2019). *Hoax Web Detection For News in Bahasa Using Support Vector Machine*. https://doi.org/https://doi.org/10.1109/ICOIACT46704.2019.8938425

Ridho, M. Y., & Yulianti, E. (2024). From Text to Truth: Leveraging IndoBERT and Machine Learning Models for Hoax Detection in Indonesian News. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *10*(3), 544–555. https://doi.org/https://doi.org/10.26555/jiteki.v10i3.29450

Ripa'i, A., Santoso, F., & Lazim, F. (2024). Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan Deep Learning Algoritma BERT. *G-Tech: Jurnal Teknologi Terapan*, *8*(3), 1749–1758. https://doi.org/https://doi.org/10.33379/gtech.v8i3.4541

Ristyawan, A., Nugroho, A., & Amarya, T. K. (2025). Optimasi Preprocessing Model Random Forest Untuk Prediksi Stroke. *Jurnal Teknik Informatika Dan Sistem Informasi (JATISI)*, *12*(1), 29–44. https://doi.org/10.25008/jatisi.v12i1.9587

Sabir, M., Farooq Khan, T., & Azam, M. (2025). A Comparative Study of Traditional and Hybrid Models for Text Classification. In *Journal of Computers and Intelligent Systems* (Vol. 3, Number 1). https://journals.iub.edu.pk/index.php/JCIS/

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks. *SBC Reviews on Computer Science*, *4*(1), 28–43. https://doi.org/https://doi.org/10.5753/reviews.2024.3757

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, *19*(1), 22–36. https://doi.org/https://doi.org/10.1145/3137597.3137600

Sucipto, S., Dwi Prasetya, D., & Widiyaningtyas, T. (2024). Educational Data Mining: Multiple Choice Question Classification in Vocational School. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *23*(2), 379–388. https://doi.org/https://doi.org/10.30812/matrik.v23i2.3499

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). *How to Fine-Tune BERT for Text Classification?* https://doi.org/https://doi.org/10.48550/arXiv.1905.05583

Tammam, A., Sucipto, & Indriati, R. (2018). Hoax Detection At Social Media With Text Mining Clarification System-Based . ***Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika*****, *3*. https://doi.org/https://doi.org/10.29100/jipi.v3i2.837**

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/https://doi.org/10.1080/21670811.2017.1360143

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline*. 4593–4601. https://doi.org/https://doi.org/10.18653/v1/P19-1452

Thölke, P., Mantilla-Ramos, Y. J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*. https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.120253

Tobing, C. J. L., Wijayakusuma, I. L., & Harini, L. P. I. (2025a). Detection of Political Hoax News Using Fine-Tuning IndoBERT. *Journal of Applied Informatics and Computing*, *9*(2), 354–360. https://doi.org/https://doi.org/10.30871/jaic.v9i2.8989

Tobing, C. J. L., Wijayakusuma, I. L., & Harini, L. P. I. (2025b). Perbandingan Kinerja IndoBERT dan MBERT Untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia. *JST (Jurnal Sains Dan Teknologi)*, *14*(1), 114–123. https://doi.org/https://doi.org/10.23887/jstundiksha.v14i1.92126

Tobing, E. (2020). *NormalisasiKata: Kamus Normalisasi Kata Tidak Baku Bahasa Indonesia*. Github. https://github.com/evrintobing17/NormalisasiKata

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). *Attention Is All You Need*. https://doi.org/https://doi.org/10.48550/arXiv.1706.03762