ROBUSTNESS MODEL DETEKSI BERITA PALSU TERHADAP PENGURANGAN TEKS

Irfan Wibowo[1], Dandy HeragoTantoko[2], Popo Nugroho[3], Tungkot Siregar[4], Wahid Diyono[5]

[12345,] Program Studi Ilmu Komputer, Universitas Nusa Mandiri

Jl. Margonda No.545, RT.1/RW.7, Pondok Cina, Kecamatan Beji, Kota Depok, Jawa Barat 16424
[1] 14250029@nusamandiri.ac.id, [2] 14250026@nusamandiri.ac.id, [3] 14250016@nusamandiri.ac.id,
[4] 14250004@nusamandiri.ac.id, [5] 14250036@nusamandiri.ac.id

Abstract

The rapid spread of digital information has significantly increased the circulation of fake news, creating challenges for automatic detection systems. While recent deep learning models have achieved high classification accuracy, their robustness under incomplete textual information remains insufficiently explored. This study investigates the robustness of three text classification models, namely TF-IDF with Logistic Regression, BiLSTM, and DistilBERT, under progressive text reduction scenarios for fake news detection. A public fake news dataset was preprocessed and evaluated using three text reduction strategies: prefix truncation, random sentence removal, and source masking. Model performance was assessed using Accuracy, Macro Precision, Macro Recall, Macro F1-Score, Matthews Correlation Coefficient (MCC), ROC-AUC, Log Loss, Expected Calibration Error (ECE), performance retention, prediction consistency, and McNemar's statistical test. Experimental results show that DistilBERT achieved the best baseline performance with a Macro F1-Score of 0.9989 and maintained 100% performance retention under prefix truncation. However, TF-IDF with Logistic Regression demonstrated greater robustness under random sentence removal and source masking, indicating that robustness depends on the type of textual information removed. These findings provide a more comprehensive evaluation framework for robustness analysis and offer practical insights for developing more reliable fake news detection systems.

***Keywords** : fake news detection, deep learning, robustness, DistilBERT, text reduction*

Abstrak

Penyebaran informasi digital yang semakin cepat telah meningkatkan peredaran berita palsu sehingga menimbulkan tantangan bagi sistem deteksi otomatis. Meskipun model *deep learning* telah mencapai akurasi klasifikasi yang tinggi, ketahanannya (*robustness*) terhadap informasi tekstual yang tidak lengkap masih belum banyak diteliti. Penelitian ini bertujuan menganalisis *robustness* tiga model klasifikasi teks, yaitu TF-IDF dengan Logistic Regression, BiLSTM, dan DistilBERT, terhadap pengurangan teks pada deteksi berita palsu. Dataset berita palsu dipraproses dan dievaluasi menggunakan tiga skenario pengurangan teks, yaitu *prefix truncation*, *random sentence removal*, dan *source masking*. Evaluasi dilakukan menggunakan Accuracy, Macro Precision, Macro Recall, Macro F1-Score, Matthews Correlation Coefficient (MCC), ROC-AUC, Log Loss, Expected Calibration Error (ECE), *performance retention*, *prediction consistency*, serta uji statistik McNemar. Hasil penelitian menunjukkan bahwa DistilBERT memperoleh performa terbaik pada kondisi referensi dengan nilai *Macro F1-Score* sebesar 0,9989 dan mampu mempertahankan *performance retention* sebesar 100% pada skenario *prefix truncation*. Namun, TF-IDF dengan Logistic Regression menunjukkan ketahanan yang lebih baik pada skenario *random sentence removal* dan *source masking*. Temuan ini menunjukkan bahwa tingkat *robustness* model dipengaruhi oleh bentuk pengurangan informasi tekstual, sehingga evaluasi menggunakan berbagai skenario pengujian diperlukan untuk memperoleh gambaran ketahanan model yang lebih komprehensif.

Kata kunci : deteksi berita palsu, *deep learning*, *robustness*, DistilBERT, pengurangan teks

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mengubah cara masyarakat memperoleh, mengelola, dan menyebarkan informasi melalui berbagai *platform* digital. Kemudahan akses terhadap media daring, media sosial, dan aplikasi berbagi informasi memungkinkan suatu berita tersebar secara luas hanya dalam waktu yang singkat. Di sisi lain, kondisi tersebut juga meningkatkan penyebaran *fake news* yang sulit dibedakan dengan berita yang valid sehingga berpotensi memengaruhi opini publik, proses pengambilan keputusan, serta stabilitas sosial [1], [2], [3]. Berbagai penelitian menunjukkan bahwa deteksi berita palsu masih menjadi salah satu tantangan utama pada bidang Natural Language Processing (NLP) karena karakteristik bahasa, gaya penulisan, serta pola penyebaran informasi terus berkembang mengikuti dinamika media digital [1], [2], [4].

Untuk mengatasi permasalahan tersebut, berbagai pendekatan berbasis *machine learning* dan *deep learning* telah dikembangkan. Model klasifikasi berbasis representasi teks seperti Term Frequency–Inverse Document Frequency (TF-IDF) yang dikombinasikan dengan algoritma klasifikasi konvensional masih banyak digunakan sebagai *baseline* karena memiliki kompleksitas yang relatif rendah dan interpretasi yang baik [5], [6]. Seiring perkembangan teknologi, model berbasis *deep learning* mulai mendominasi penelitian deteksi berita palsu karena mampu mempelajari representasi semantik secara otomatis tanpa memerlukan proses *feature engineering* yang kompleks [3], [7], [8].

Arsitektur Recurrent Neural Network (RNN), khususnya LSTM dan BiLSTM, menjadi salah satu pendekatan yang banyak digunakan karena mampu menangkap hubungan kontekstual antar kata dalam dokumen berita. Berbagai penelitian menunjukkan bahwa pemanfaatan BiLSTM mampu meningkatkan performa klasifikasi dibandingkan pendekatan RNN konvensional, terutama pada teks dengan ketergantungan konteks yang panjang [7], [9], [10]. Pada penelitian lain, kombinasi BiLSTM dengan teknik *hyperparameter tuning* maupun metode *Explainable Artificial Intelligence* (XAI) seperti LIME juga memberikan peningkatan interpretabilitas tanpa mengurangi performa model secara signifikan [3], [9], [10].

Perkembangan *Transformer*-based *pre-trained language model* membawa peningkatan yang lebih signifikan pada tugas klasifikasi teks. Model seperti BERT, DistilBERT, maupun berbagai model hibrida berbasis *Transformer* mampu menghasilkan representasi bahasa yang lebih baik sehingga meningkatkan akurasi deteksi berita palsu pada berbagai dataset [11], [12], [13]. Selain itu, penelitian lain mengembangkan pendekatan berbasis GPT-BERT, *ensemble learning*, serta model *hierarchical attention* seperti 3HAN untuk meningkatkan kemampuan model dalam memahami hubungan antar kalimat dan struktur dokumen secara lebih komprehensif [4], [14], [15].

Tidak hanya berfokus pada pengembangan model, beberapa penelitian juga mengombinasikan berbagai arsitektur *deep learning* untuk memperoleh performa klasifikasi yang lebih baik. Pendekatan berbasis *ensemble*, kombinasi BERT dengan BiLSTM, maupun pemanfaatan *transfer learning* pada bahasa Indonesia menunjukkan bahwa integrasi beberapa metode mampu meningkatkan *Accuracy* dan *F1-Score* dibandingkan penggunaan model tunggal [8], [16], [17]. Hasil tersebut menunjukkan bahwa penelitian deteksi berita palsu saat ini lebih banyak diarahkan pada peningkatan performa klasifikasi melalui pengembangan arsitektur model maupun strategi pelatihan.

Meskipun demikian, sebagian besar penelitian terdahulu mengevaluasi performa model menggunakan artikel berita yang lengkap (*complete text*) sehingga kondisi pengujian relatif ideal [4], [12], [13]. Dalam praktiknya, pengguna media digital sering kali hanya membaca judul, ringkasan, atau sebagian isi berita sebelum mengambil keputusan untuk mempercayai maupun menyebarkan informasi tersebut. Kondisi tersebut menunjukkan bahwa tingginya nilai *Accuracy* pada artikel lengkap belum tentu mencerminkan kemampuan model ketika menerima informasi yang tidak utuh atau telah mengalami pengurangan isi teks.

Seiring meningkatnya kompleksitas model *deep learning*, perhatian peneliti mulai bergeser dari sekadar meningkatkan akurasi menuju kemampuan model dalam mempertahankan performa pada berbagai kondisi masukan. Salah satu aspek yang banyak dikaji adalah *robustness*, yaitu kemampuan model menghasilkan prediksi yang konsisten ketika menerima perubahan atau gangguan pada data masukan. Berbagai penelitian menunjukkan bahwa model deteksi berita palsu masih rentan terhadap *adversarial attack*, manipulasi gaya bahasa, maupun perubahan distribusi data sehingga menghasilkan penurunan performa yang signifikan [18], [19], [20].

Selain aspek *robustness*, kemampuan *generalization* dan *calibration* juga menjadi perhatian dalam pengembangan *pre-trained language model*. Model yang memiliki nilai *Accuracy* tinggi pada data uji belum tentu mampu mempertahankan performa ketika dihadapkan pada distribusi data yang berbeda atau kondisi masukan yang tidak ideal. Penelitian terbaru menunjukkan bahwa peningkatan *robustness* tidak hanya berkaitan dengan akurasi, tetapi juga berhubungan dengan kemampuan model melakukan generalisasi dan menghasilkan tingkat kepercayaan prediksi yang lebih reliabel [21], [22].

Di sisi lain, pemilihan metrik evaluasi yang tepat juga berperan penting dalam menggambarkan performa model secara menyeluruh. Penggunaan *Accuracy* sebagai satu-satunya indikator evaluasi dinilai belum cukup karena tidak selalu mampu merepresentasikan kemampuan model pada berbagai karakteristik data. Oleh karena itu, beberapa penelitian merekomendasikan penggunaan kombinasi metrik seperti Precision, Recall, Macro F1-Score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic – Area Under Curve (ROC-AUC), maupun metrik reliabilitas untuk memperoleh interpretasi performa yang lebih komprehensif [22], [23], [24].

Berdasarkan telaah terhadap seluruh literatur yang digunakan pada penelitian ini, sebagian besar penelitian terdahulu masih berfokus pada pengembangan arsitektur *deep learning*, peningkatan performa klasifikasi, atau evaluasi terhadap gangguan berbasis *adversarial attack* dan perubahan gaya bahasa [18], [19], [25]. Meskipun penelitian tersebut memberikan kontribusi yang signifikan terhadap perkembangan sistem deteksi berita palsu, belum ditemukan penelitian yang secara sistematis mengevaluasi *robustness* model terhadap pengurangan teks (*text reduction*) menggunakan beberapa skenario evaluasi dalam satu kerangka eksperimen yang terintegrasi. Selain itu, penelitian sebelumnya juga belum mengombinasikan pendekatan *prefix truncation*, *random sentence removal*, dan *source masking* untuk mengidentifikasi perubahan performa model ketika informasi tekstual berkurang secara bertahap.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan menganalisis *robustness* model deteksi berita palsu terhadap pengurangan teks. Penelitian menggunakan TF-IDF dengan *Logistic Regression* sebagai model pembanding, BiLSTM, dan DistilBERT untuk membandingkan karakteristik model konvensional, model berbasis jaringan saraf berulang, dan *pre-trained language model*. Evaluasi dilakukan melalui tiga skenario pengurangan teks, yaitu *prefix truncation*, *random sentence removal*, dan *source masking*. Selanjutnya, performa model dievaluasi menggunakan Accuracy, Macro Precision, Macro Recall, Macro F1-Score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic–Area Under Curve (ROC-AUC), Log Loss, dan Expected Calibration Error (ECE), kemudian didukung oleh analisis *performance retention*, *prediction consistency*, serta uji statistik *McNemar* untuk memperoleh gambaran yang lebih komprehensif mengenai ketahanan model terhadap berkurangnya informasi tekstual.

Kontribusi utama penelitian ini adalah penyusunan kerangka evaluasi *robustness* terhadap pengurangan teks yang mengintegrasikan beberapa skenario pengujian dalam satu proses evaluasi. Pendekatan tersebut diharapkan mampu memberikan pemahaman yang lebih mendalam mengenai karakteristik model *deep learning* ketika menghadapi informasi yang tidak lengkap, sekaligus menjadi referensi bagi pengembangan sistem deteksi berita palsu yang lebih andal dan adaptif pada berbagai kondisi penggunaan di dunia nyata.

2. METODOLOGI PENELITIAN

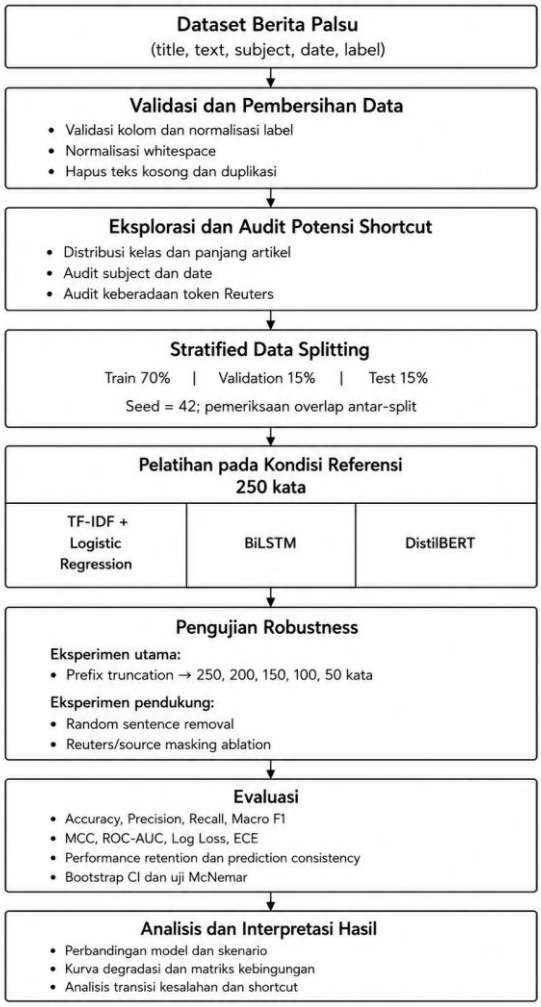
2.1 Skema Alur Penelitian

Penelitian ini menggunakan pendekatan eksperimental untuk menganalisis *robustness* model deteksi berita palsu terhadap pengurangan teks. Metodologi penelitian dirancang secara sistematis mulai dari proses persiapan data, pembangunan model klasifikasi, pengujian berbagai skenario pengurangan teks, hingga evaluasi *robustness*. Rancangan tersebut bertujuan menghasilkan proses eksperimen yang terstruktur, dapat direproduksi, serta memungkinkan perbandingan yang adil antar model klasifikasi. Skema alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.

Tahap pertama adalah pengumpulan dan validasi dataset berita palsu yang terdiri atas atribut *title*, *text*, *subject*, *date*, dan *label*. Pada penelitian ini hanya atribut *text* yang digunakan sebagai masukan model, sedangkan atribut *subject* dan *date* tidak digunakan karena berpotensi menghasilkan *shortcut learning*, yaitu kondisi ketika model memanfaatkan metadata sebagai dasar prediksi, bukan informasi linguistik yang terkandung di dalam isi berita. Setelah

proses validasi atribut, dilakukan normalisasi label, penghapusan data duplikat, penghapusan artikel dengan teks kosong, serta normalisasi karakter agar kualitas data tetap terjaga sebelum digunakan pada proses pelatihan model.

Tahap kedua merupakan eksplorasi data dan audit terhadap potensi *shortcut*. Analisis dilakukan terhadap distribusi kelas, panjang artikel, distribusi jumlah kata, serta keterkaitan atribut *subject* dengan label. Selain itu dilakukan identifikasi token yang berpotensi menjadi penanda sumber berita, seperti kata *Reuters*, untuk mengetahui kemungkinan model memanfaatkan informasi tersebut sebagai indikator klasifikasi. Tahap ini menjadi dasar dalam penyusunan eksperimen *source masking* sebagai salah satu skenario pengujian *robustness*.



Gambar 1. Skema Alur Penelitian

Tahap ketiga adalah pembagian dataset menggunakan metode *stratified split* sehingga proporsi kelas tetap terjaga pada data pelatihan, validasi, dan pengujian. Dataset dibagi menjadi 70% data pelatihan, 15% data validasi, dan 15% data pengujian menggunakan *random seed* yang sama pada seluruh eksperimen. Selain menjaga distribusi kelas, dilakukan pula pemeriksaan terhadap kemungkinan *data leakage* dengan memastikan tidak terdapat artikel yang muncul pada lebih dari satu subset data.

Tahap keempat adalah pembangunan model klasifikasi. Penelitian ini menggunakan tiga model yang mewakili pendekatan berbeda. Model pertama adalah TF-IDF yang dikombinasikan dengan Logistic Regression sebagai *baseline* berbasis *machine learning* karena memiliki kompleksitas yang relatif rendah serta masih banyak digunakan sebagai pembanding pada penelitian deteksi berita palsu [5], [6]. Model kedua menggunakan BiLSTM yang mampu mempelajari hubungan kontekstual dua arah pada urutan kata sehingga efektif dalam klasifikasi dokumen teks [7], [9], [10]. Model ketiga menggunakan DistilBERT sebagai *pre-trained Transformer model* yang menghasilkan representasi kontekstual dengan kompleksitas komputasi yang lebih rendah dibandingkan BERT, namun tetap mempertahankan performa klasifikasi yang tinggi [11], [12], [13].

Tahap kelima merupakan pengujian *robustness* model terhadap pengurangan informasi tekstual. Berbeda dengan penelitian terdahulu yang umumnya mengevaluasi model menggunakan artikel lengkap atau skenario *adversarial attack* [18], [19], [20], penelitian ini mengevaluasi ketahanan model melalui tiga skenario pengurangan teks, yaitu *prefix truncation*, *random sentence removal*, dan *source masking*. Pada skenario *prefix truncation*, artikel dipotong secara bertahap menjadi 250, 200, 150, 100, dan 50 kata. Skenario *random sentence removal* dilakukan dengan mempertahankan sejumlah kalimat secara deterministik hingga memenuhi batas jumlah kata yang sama. Sementara itu, *source masking* dilakukan dengan mengganti token yang mengindikasikan sumber berita menjadi token netral sehingga dapat diketahui apakah model bergantung pada identitas sumber dalam melakukan klasifikasi.

Tahap keenam adalah evaluasi performa model. Evaluasi dilakukan menggunakan beberapa metrik klasifikasi, yaitu Accuracy, Macro Precision, Macro Recall, Macro F1-Score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic–Area Under Curve (ROC-AUC), Log Loss, dan Expected Calibration Error (ECE). Penggunaan beberapa metrik tersebut bertujuan memberikan evaluasi yang lebih komprehensif

dibandingkan hanya menggunakan *Accuracy*, sebagaimana direkomendasikan pada penelitian mengenai evaluasi klasifikasi modern [21], [22], [23]. Selain itu, penelitian juga menghitung *performance retention*, *prediction consistency*, *normalized robustness area*, *bootstrap confidence interval*, serta melakukan uji statistik McNemar untuk mengevaluasi signifikansi perubahan performa model pada setiap skenario pengurangan teks.

Tahap terakhir adalah analisis hasil eksperimen. Seluruh hasil evaluasi dibandingkan untuk mengetahui tingkat ketahanan masing-masing model terhadap berkurangnya informasi tekstual. Analisis dilakukan berdasarkan perubahan nilai metrik evaluasi, pola degradasi performa, konsistensi prediksi, perubahan tingkat kepercayaan model, serta hasil uji statistik. Selanjutnya, hasil tersebut digunakan untuk mengidentifikasi model yang paling stabil pada berbagai kondisi pengurangan teks serta menjelaskan implikasinya terhadap pengembangan sistem deteksi berita palsu pada kondisi penggunaan di dunia nyata.

2.2 Pengumpulan Data

Penelitian ini menggunakan data sekunder berupa *Fake News Detection Dataset (6000 News Articles)* yang diperoleh dari platform Kaggle. Dataset terdiri atas artikel berita yang telah diberi label berita palsu (*fake news*) dan berita asli (*real news*), serta memuat atribut *title*, *text*, *subject*, *date*, dan *label*. Pada penelitian ini hanya atribut *text* yang digunakan sebagai masukan utama model, sedangkan atribut *subject* dan *date* tidak digunakan sebagai fitur karena berpotensi menimbulkan *shortcut learning* pada proses klasifikasi.

Proses pengumpulan data dilakukan dengan mengunduh dataset dari Kaggle, kemudian memvalidasi struktur data, tipe atribut, dan kelengkapan data sebelum digunakan pada tahap praproses. Selanjutnya dilakukan pembersihan data dengan menghapus artikel duplikat, menghilangkan data dengan teks kosong, serta melakukan normalisasi label agar seluruh data siap digunakan pada proses pelatihan dan pengujian model.

Dataset yang telah melalui proses validasi dan pembersihan kemudian dibagi menjadi data pelatihan, data validasi, dan data pengujian menggunakan metode *stratified split* dengan rasio 70:15:15. Pembagian tersebut bertujuan mempertahankan proporsi kelas pada setiap subset sehingga proses pelatihan dan evaluasi dapat dilakukan secara adil dan konsisten.

2.3 Analisis Data

Analisis data dilakukan untuk mengevaluasi tingkat *robustness* model deteksi berita palsu terhadap pengurangan teks. Analisis ini dilakukan karena performa model yang tinggi pada artikel lengkap belum tentu mampu dipertahankan ketika informasi yang tersedia semakin terbatas. Selain itu, evaluasi *robustness* menjadi salah satu aspek penting dalam pengembangan sistem klasifikasi modern untuk mengetahui kemampuan model dalam menghadapi perubahan karakteristik data masukan [1], [24].

Objek analisis berupa artikel berita yang telah melalui proses validasi, pembersihan, dan pembagian dataset menjadi data pelatihan, validasi, dan pengujian. Penelitian membandingkan tiga model klasifikasi yang mewakili pendekatan berbeda, yaitu TF-IDF dengan Logistic Regression sebagai *baseline* berbasis *machine learning*, BiLSTM sebagai model *deep learning* berbasis *Recurrent Neural Network*, serta DistilBERT sebagai *pre-trained Transformer model*. Ketiga model tersebut dipilih karena memiliki karakteristik representasi teks yang berbeda sehingga dapat memberikan gambaran mengenai tingkat *robustness* pada setiap pendekatan klasifikasi [7], [15], [19].

Analisis dilakukan setelah seluruh model selesai dilatih menggunakan data pelatihan pada kondisi referensi sebanyak 250 kata. Model yang sama kemudian digunakan untuk melakukan inferensi pada berbagai kondisi pengurangan teks tanpa dilakukan proses pelatihan ulang. Pendekatan ini bertujuan untuk memastikan bahwa perubahan performa yang diamati benar-benar disebabkan oleh berkurangnya informasi tekstual, bukan oleh perubahan parameter atau proses pelatihan model.

Pengujian dilakukan melalui tiga skenario, yaitu *prefix truncation*, *random sentence removal*, dan *source masking*. Skenario *prefix truncation* digunakan untuk mensimulasikan kondisi ketika pembaca hanya memperoleh sebagian awal isi artikel. Sementara itu, *random sentence removal* merepresentasikan hilangnya informasi yang tersebar pada berbagai bagian artikel, sedangkan *source masking* digunakan untuk mengevaluasi ketergantungan model terhadap identitas sumber berita. Ketiga skenario tersebut dirancang untuk memberikan gambaran yang lebih komprehensif mengenai ketahanan model terhadap berbagai bentuk pengurangan informasi tekstual.

Hasil klasifikasi kemudian dianalisis menggunakan beberapa metrik evaluasi, yaitu Accuracy, Macro Precision, Macro Recall, Macro F1-Score, Matthews Correlation Coefficient (MCC), ROC-AUC, Log Loss, dan Expected Calibration Error (ECE). Penggunaan kombinasi metrik tersebut bertujuan memberikan evaluasi yang lebih menyeluruh terhadap kemampuan klasifikasi, kualitas probabilitas prediksi, serta reliabilitas model dibandingkan hanya menggunakan *Accuracy* sebagai indikator utama [20], [23], [24]. Selain itu, penelitian ini juga menganalisis *performance retention*, *prediction consistency*, *bootstrap confidence interval*, serta uji statistik McNemar untuk mengidentifikasi perubahan performa dan tingkat ketahanan masing-masing model terhadap pengurangan informasi tekstual.

2.4 Tahapan Penelitian

2.4.1 Data *Preprocessing*

Tahap *data preprocessing* bertujuan menyiapkan dataset agar memiliki kualitas yang baik sebelum digunakan pada proses pelatihan model. Seluruh proses praproses dilakukan secara berurutan untuk mengurangi potensi *noise*, menjaga konsistensi data, serta memastikan bahwa setiap model menerima data masukan dengan format yang sama.

Proses diawali dengan memuat dataset berita palsu dalam format CSV, kemudian dilakukan validasi terhadap atribut yang digunakan pada penelitian, yaitu *title*, *text*, *subject*, *date*, dan *label*. Selanjutnya dilakukan normalisasi label menjadi dua kelas, yaitu 0 untuk *fake news* dan 1 untuk *real news*, sehingga seluruh model menggunakan representasi kelas yang seragam selama proses pelatihan dan evaluasi.

Tahap berikutnya adalah pembersihan data dengan menghapus artikel yang memiliki teks kosong (*empty text*) serta menghilangkan data duplikat berdasarkan isi artikel (*text*). Langkah ini bertujuan mengurangi redundansi data sekaligus mencegah terjadinya *data leakage* yang dapat memengaruhi hasil evaluasi model.

Setelah proses pembersihan selesai, dilakukan normalisasi karakter dengan menghilangkan spasi berlebih (*whitespace normalization*) sehingga struktur teks menjadi lebih konsisten. Pada tahap praproses umum, penelitian ini tidak melakukan *stemming*, *stopword removal*, maupun *lemmatization* agar informasi tekstual asli tetap dipertahankan. *Stopword removal* hanya diterapkan secara internal pada pembentukan fitur TF-IDF, sedangkan masukan BiLSTM dan DistilBERT tidak melalui proses tersebut. Penghilangan kata atau perubahan bentuk kata dikhawatirkan dapat mengurangi informasi semantik yang diperlukan selama proses pembelajaran model.

Selanjutnya dilakukan analisis panjang artikel untuk mengetahui distribusi jumlah kata pada dataset. Berdasarkan hasil analisis tersebut, panjang maksimum artikel pada kondisi referensi ditetapkan sebanyak 250 kata. Batas tersebut digunakan sebagai masukan standar pada proses pelatihan seluruh model sehingga setiap model memperoleh informasi tekstual dengan panjang yang sama. Penetapan kondisi referensi ini juga menjadi dasar dalam penyusunan eksperimen pengurangan teks yang dijelaskan pada Subbab 2.4.3.

Setelah seluruh tahapan praproses selesai dilakukan, diperoleh sebanyak 5.828 artikel yang kemudian dibagi menjadi data pelatihan, validasi, dan pengujian menggunakan metode *stratified split* dengan rasio 70:15:15 dan *random seed* 42. Pembagian menghasilkan 4.079 artikel pada data pelatihan yang terdiri atas 1.983 *fake news* dan 2.096 *real news*, 874 artikel pada data validasi yang terdiri atas 425 *fake news* dan 449 *real news*, serta 875 artikel pada data pengujian yang terdiri atas 425 *fake news* dan 450 *real news*. Metode *stratified split* digunakan untuk mempertahankan proporsi kelas pada setiap *subset*. Selanjutnya, dilakukan pemeriksaan *overlap* berdasarkan isi artikel dan tidak ditemukan artikel yang sama antara data pelatihan, validasi, maupun pengujian, sehingga potensi *data leakage* antar *subset* dapat diminimalkan.

2.4.2 Pembangunan Model

Tahap pembangunan model bertujuan memperoleh model klasifikasi yang akan digunakan pada seluruh skenario pengujian *robustness*. Penelitian ini menggunakan tiga model dengan karakteristik yang berbeda, yaitu TF-IDF dengan *Logistic Regression* sebagai *baseline*, BiLSTM sebagai model *deep learning* berbasis *Recurrent Neural Network*, serta DistilBERT sebagai *pre-trained Transformer model* [5], [6], [7]. Ketiga model dilatih menggunakan data pelatihan yang sama pada kondisi referensi sebanyak 250 kata sehingga perbedaan performa yang diamati berasal dari karakteristik model, bukan dari perbedaan data pelatihan.

Model pertama menggunakan representasi TF-IDF yang mengubah dokumen menjadi vektor numerik berdasarkan frekuensi kemunculan kata

dan tingkat kepentingannya pada seluruh korpus. TF-IDF dikonfigurasi menggunakan *unigram* dan *bigram* dengan ngram_range=(1,2), min_df=2, max_df=0,98, maksimum 50.000 fitur, serta *sublinear term frequency*. Teks dikonversi menjadi huruf kecil dan *stopword* bahasa Inggris diabaikan pada tahap pembentukan fitur TF-IDF. Representasi tersebut selanjutnya diklasifikasikan menggunakan Logistic Regression dengan max_iter=2.000, class_weight=balanced, dan random_state=42. Model ini digunakan sebagai *baseline* karena memiliki kompleksitas komputasi yang relatif rendah serta masih banyak digunakan sebagai pembanding pada penelitian klasifikasi teks [5], [6].

Model kedua menggunakan arsitektur BiLSTM yang memproses urutan kata dari dua arah, yaitu maju (*forward*) dan mundur (*backward*), sehingga mampu mempelajari hubungan kontekstual antarkata secara lebih baik dibandingkan Recurrent Neural Network konvensional [7], [9], [10]. Teks ditokenisasi dengan *vocabulary* maksimum 30.000 token dan diubah menjadi *sequence* dengan panjang maksimum 250 token menggunakan *padding*. Arsitektur menggunakan *embedding dimension* 128, satu lapisan BiLSTM *bidirectional* dengan *hidden dimension* 128, dan *dropout* sebesar 0,30. Keluaran *hidden state* dari kedua arah digabungkan dan diteruskan ke lapisan linear dengan satu output logit. Model dilatih menggunakan BCEWithLogitsLoss dan *optimizer* AdamW dengan *learning rate* 1×10^{-3}, *weight decay* 1×10^{-4}, *training batch size* 32, *validation batch size* 64, serta maksimum 10 *epoch*. *Early stopping* diterapkan dengan *patience* 2 berdasarkan *validation loss* dan *gradient clipping* sebesar 1,0. Pada tahap inferensi, probabilitas kelas diperoleh menggunakan fungsi sigmoid dengan *threshold* 0,5.

Model ketiga menggunakan DistilBERT dengan *pre-trained* model distilbert-base-uncased untuk klasifikasi biner. Setiap artikel terlebih dahulu dibatasi pada kondisi referensi 250 kata dan diproses menggunakan DistilBERT *tokenizer* dengan panjang maksimum 384 token. *Dynamic padding* diterapkan melalui *data collator* pada setiap *batch*. Proses *fine-tuning* dilakukan menggunakan *learning rate* 2×10^{-5}, *batch size* 8 untuk pelatihan dan validasi, *weight decay* 0,01, serta maksimum 3 *epoch*. Evaluasi dan penyimpanan model dilakukan pada setiap *epoch*, dengan Macro F1-Score sebagai metrik pemilihan model terbaik. *Early stopping* diterapkan dengan *patience* 1 dan model dengan Macro F1-Score validasi terbaik dimuat kembali pada akhir proses pelatihan. *Random seed* dan *data seed* ditetapkan sebesar 42 untuk menjaga konsistensi eksperimen.

Untuk menjaga konsistensi eksperimen, setiap model hanya dilatih satu kali pada kondisi referensi sebanyak 250 kata. Model yang telah dilatih selanjutnya digunakan kembali pada seluruh skenario pengurangan teks tanpa dilakukan pelatihan ulang. Pendekatan ini memastikan bahwa perubahan performa yang diamati benar-benar disebabkan oleh berkurangnya informasi tekstual, bukan oleh perubahan parameter selama proses pelatihan.

2.4.3 Skenario Pengurangan Teks

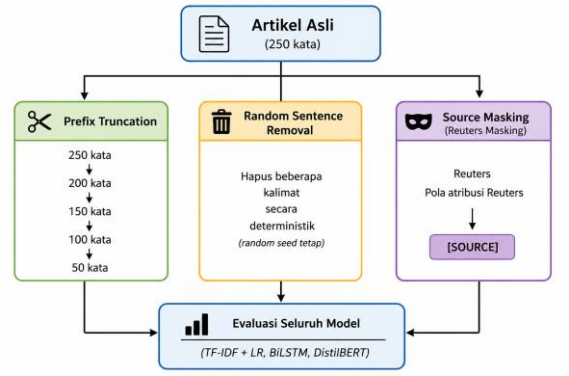
Setelah seluruh model selesai dilatih menggunakan artikel pada kondisi referensi sebanyak 250 kata, dilakukan pengujian *robustness* melalui beberapa skenario pengurangan teks. Pengujian ini bertujuan untuk mengevaluasi kemampuan model dalam mempertahankan performa ketika informasi tekstual yang tersedia semakin berkurang. Berbeda dengan penelitian sebelumnya yang umumnya mengevaluasi model menggunakan artikel lengkap atau skenario *adversarial attack* [18], [19], [20], penelitian ini berfokus pada pengurangan informasi tekstual secara bertahap sehingga dapat menggambarkan kondisi penggunaan di dunia nyata.

Skenario pertama adalah *prefix truncation*, yaitu mempertahankan bagian awal artikel dengan panjang tertentu, sedangkan bagian sisanya dihilangkan. Pada penelitian ini digunakan lima tingkat panjang teks, yaitu 250, 200, 150, 100, dan 50 kata**.** Seluruh model yang telah dilatih pada kondisi referensi kemudian diuji kembali pada setiap tingkat panjang teks tanpa dilakukan pelatihan ulang. Pendekatan ini digunakan untuk mensimulasikan kondisi ketika pengguna hanya membaca sebagian awal isi berita sebelum mengambil keputusan.

Skenario kedua adalah *random sentence removal*, yaitu mengurangi panjang artikel dengan memilih sejumlah kalimat secara acak dari dokumen. Pemilihan kalimat dilakukan secara acak dengan *random seed* tetap sehingga hasil eksperimen bersifat deterministik dan dapat direproduksi. Sejumlah kalimat dipilih secara acak hingga mendekati *word budget*, kemudian urutan kalimat asli dipertahankan dan hasil akhir dibatasi hingga maksimum sesuai panjang target. Artikel yang panjangnya lebih pendek dari *word budget* tetap dipertahankan tanpa penambahan teks.

Skenario ini digunakan untuk mengevaluasi sensitivitas model terhadap hilangnya informasi yang tersebar pada berbagai bagian artikel.

Skenario ketiga adalah *source masking*, yaitu menyamarkan token Reuters dan pola atribusi sumber Reuters pada artikel menjadi token netral [*SOURCE*]. Pendekatan ini digunakan untuk mengevaluasi sensitivitas model terhadap informasi sumber yang berpotensi menjadi *shortcut feature*. Dengan menghilangkan informasi tersebut, penelitian dapat mengidentifikasi tingkat ketergantungan masing-masing model terhadap atribut yang tidak berkaitan langsung dengan konten berita. Perhatikan Gambar 2.



Gambar 2. Skenario Pengurangan Teks

2.4.4 Evaluasi *Robustness*

Evaluasi *robustness* dilakukan untuk mengukur kemampuan setiap model dalam mempertahankan performa klasifikasi pada berbagai skenario pengurangan teks. Seluruh model dievaluasi menggunakan data pengujian yang sama sehingga perubahan performa yang diamati hanya dipengaruhi oleh berkurangnya informasi tekstual. Untuk memperoleh hasil evaluasi yang komprehensif, penelitian ini menggunakan metrik klasifikasi, metrik *robustness*, serta pengujian statistik [22], [23], [24].

Seluruh metrik evaluasi yang digunakan pada penelitian ini disajikan pada Tabel 1.

TABEL I. METRIK EVALUASI PENELITIAN

Kelompok	Metrik	Tujuan
Klasifikasi	*Accuracy*	Mengukur ketepatan klasifikasi
Klasifikasi	*Macro Precision*	Mengukur rata-rata Precision antar kelas
Klasifikasi	*Macro Recall*	Mengukur rata-rata Recall antar kelas
Klasifikasi	Macro F1-Score	Mengukur rata-rata F1-Score antar kelas
Klasifikasi	MCC	Mengukur kualitas klasifikasi secara keseluruhan
Klasifikasi	ROC-AUC	Mengukur kemampuan membedakan dua kelas
Klasifikasi	Log Loss	Mengukur kualitas probabilitas prediksi
Klasifikasi	ECE	Mengukur *calibration* model
Robustness	*Performance Retention*	Mengukur kemampuan mempertahankan performa
Robustness	*Prediction Consistency*	Mengukur konsistensi prediksi antar skenario
Statistik	*Bootstrap Confidence Interval*	Mengestimasi rentang kepercayaan metrik
Statistik	Uji McNemar	Menguji signifikansi perbedaan hasil klasifikasi

3. HASIL DAN PEMBAHASAN

3.1 Hasil Evaluasi Model

Tahap pertama pengujian dilakukan untuk mengevaluasi performa ketiga model klasifikasi pada kondisi referensi serta beberapa variasi jumlah kata sebagai dasar analisis *robustness*. Model yang dibandingkan terdiri atas TF-IDF dengan Logistic Regression, BiLSTM, dan DistilBERT. Evaluasi dilakukan menggunakan metrik Accuracy, Macro Precision, Macro Recall, Macro F1-Score, MCC, ROC-AUC, Log Loss, dan

Expected Calibration Error (ECE). Ringkasan hasil evaluasi disajikan pada Tabel 2.

Berdasarkan Tabel 2, DistilBERT menunjukkan performa terbaik pada seluruh metrik evaluasi. Model ini memperoleh nilai *Macro F1-Score* sebesar 0,9989, MCC sebesar 0,9977, serta ROC-AUC mendekati 1,0000, yang menunjukkan kemampuan klasifikasi dan pemisahan kelas yang sangat baik. Selain itu, DistilBERT juga menghasilkan nilai *Log Loss* dan ECE yang paling rendah, sehingga probabilitas prediksi yang dihasilkan lebih terkalibrasi dibandingkan model lainnya.

BiLSTM juga memberikan performa yang sangat baik dengan *Macro F1-Score* sebesar 0,9954 dan MCC sebesar 0,9909. Sementara itu, TF-IDF dengan *Logistic Regression* menghasilkan performa yang lebih rendah dibandingkan kedua model *deep learning*, meskipun masih mampu mencapai nilai *Macro F1-Score* di atas 0,96. Hasil ini menunjukkan bahwa representasi kontekstual yang dipelajari oleh BiLSTM dan terutama DistilBERT lebih efektif dalam membedakan berita palsu dan berita asli dibandingkan representasi berbasis frekuensi kata.

TABEL II. HASIL EVALUASI KETIGA MODEL PADA BERBAGAI *WORD BUDGET*

	model	word_budget	accuracy	macro_precision	macro_recall	macro_f1	mcc	roc_auc	log_loss	ece
0	BiLSTM	50	0.9954	0.9953	0.9956	0.9954	0.9909	0.9999	0.0209	0.0030
1	BiLSTM	100	0.9954	0.9953	0.9956	0.9954	0.9909	0.9999	0.0209	0.0030
2	BiLSTM	150	0.9954	0.9953	0.9956	0.9954	0.9909	0.9999	0.0209	0.0030
3	BiLSTM	200	0.9954	0.9953	0.9956	0.9954	0.9909	0.9999	0.0209	0.0030
4	BiLSTM	250	0.9954	0.9953	0.9956	0.9954	0.9909	1.0000	0.0207	0.0030
5	DistilBERT	50	0.9989	0.9989	0.9988	0.9989	0.9977	1.0000	0.0088	0.0004
6	DistilBERT	100	0.9989	0.9989	0.9988	0.9989	0.9977	1.0000	0.0088	0.0005
7	DistilBERT	150	0.9989	0.9989	0.9988	0.9989	0.9977	1.0000	0.0088	0.0005
8	DistilBERT	200	0.9989	0.9989	0.9988	0.9989	0.9977	1.0000	0.0088	0.0005
9	DistilBERT	250	0.9989	0.9989	0.9988	0.9989	0.9977	1.0000	0.0089	0.0005
10	TFIDF_LogReg	50	0.9726	0.9737	0.9720	0.9725	0.9457	0.9985	0.2672	0.1945
11	TFIDF_LogReg	100	0.9760	0.9763	0.9758	0.9760	0.9521	0.9977	0.2561	0.1891
12	TFIDF_LogReg	150	0.9749	0.9750	0.9747	0.9748	0.9497	0.9968	0.2492	0.1808
13	TFIDF_LogReg	200	0.9714	0.9715	0.9713	0.9714	0.9428	0.9965	0.2405	0.1693
14	TFIDF_LogReg	250	0.9646	0.9647	0.9644	0.9645	0.9291	0.9962	0.2344	0.1565

Performa dasar tersebut kemudian digunakan sebagai acuan pada seluruh eksperimen *robustness* yang dibahas pada subbab berikutnya. Dengan menggunakan model yang sama tanpa pelatihan ulang, perubahan performa yang diamati pada setiap skenario pengurangan teks dapat diinterpretasikan sebagai dampak langsung dari berkurangnya informasi tekstual.

3.2 Analisis *Robustness* Model terhadap Pengurangan Teks

Setelah diperoleh performa dasar masing-masing model pada kondisi referensi, dilakukan pengujian *robustness* untuk mengevaluasi kemampuan model dalam mempertahankan performa ketika informasi tekstual dikurangi. Pengujian dilakukan menggunakan tiga skenario, yaitu *prefix truncation*, *random sentence removal*, dan *source masking*. Seluruh eksperimen menggunakan model yang sama tanpa proses pelatihan ulang sehingga perubahan performa yang diamati dapat dikaitkan dengan perubahan informasi pada data masukan.

Hasil pengujian *prefix truncation* menunjukkan bahwa DistilBERT merupakan model dengan tingkat *robustness* terbaik. Berdasarkan Tabel 3, nilai *Macro F1-Score* DistilBERT tetap berada pada 0,9989 pada seluruh *word budget*, sehingga menghasilkan *performance retention* sebesar 100%. BiLSTM juga menunjukkan performa yang sangat stabil dengan Macro F1-Score sebesar 0,9954, sedangkan TF-IDF dengan Logistic Regression menunjukkan variasi performa yang lebih besar ketika jumlah kata dikurangi. Menariknya, Macro F1-Score TF-IDF meningkat dari 0,9645 pada kondisi referensi 250 kata menjadi 0,9725 pada 50 kata, sehingga pengurangan panjang teks tidak selalu menyebabkan degradasi performa pada model berbasis fitur leksikal. Hasil ini menunjukkan bahwa representasi kontekstual yang dipelajari DistilBERT mampu mempertahankan informasi diskriminatif meskipun panjang teks berkurang. Stabilitas tersebut mengindikasikan bahwa informasi yang digunakan model untuk membedakan kelas telah tersedia pada bagian awal artikel, sehingga pemotongan bagian akhir hingga tersisa 50 kata tidak mengubah prediksi secara berarti. BiLSTM menunjukkan pola serupa, sedangkan TF-IDF dengan Logistic Regression menunjukkan variasi performa karena perubahan jumlah kata secara langsung mengubah distribusi fitur leksikal yang tersedia pada representasi dokumen. Perhatikan Gambar 3.

Analisis lebih lanjut menunjukkan bahwa DistilBERT dan BiLSTM memiliki *prediction consistency* yang tinggi pada seluruh tingkat pengurangan teks. Temuan tersebut didukung oleh hasil *bootstrap confidence interval* yang menunjukkan rentang kepercayaan *Macro F1-Score* relatif sempit. Selain itu, hasil uji McNemar pada DistilBERT dan BiLSTM tidak menunjukkan perubahan prediksi yang signifikan pada seluruh skenario *prefix truncation*, sedangkan analisis *error transition* menunjukkan tidak terdapat artikel yang berubah dari prediksi benar pada kondisi referensi menjadi salah pada kondisi 50 kata. Hasil tersebut mengindikasikan bahwa kedua model *deep learning* memiliki tingkat stabilitas yang sangat baik terhadap pengurangan panjang teks secara bertahap.

TABEL III. PERBANDINGAN *PERFORMANCE RETENTION* KETIGA MODEL PADA BERBAGAI *WORD BUDGET*

Performance retention:

	model	word_budget	macro_f1	performance_retention	absolute_f1_drop
0	BiLSTM	50	0.9954	1.0000	0.0000
1	BiLSTM	100	0.9954	1.0000	0.0000
2	BiLSTM	150	0.9954	1.0000	0.0000
3	BiLSTM	200	0.9954	1.0000	0.0000
4	BiLSTM	250	0.9954	1.0000	0.0000
5	DistilBERT	50	0.9989	1.0000	0.0000
6	DistilBERT	100	0.9989	1.0000	0.0000
7	DistilBERT	150	0.9989	1.0000	0.0000
8	DistilBERT	200	0.9989	1.0000	0.0000
9	DistilBERT	250	0.9989	1.0000	0.0000
10	TFIDF_LogReg	50	0.9725	1.0083	-0.0080
11	TFIDF_LogReg	100	0.9760	1.0119	-0.0114
12	TFIDF_LogReg	150	0.9748	1.0107	-0.0103
13	TFIDF_LogReg	200	0.9714	1.0071	-0.0069
14	TFIDF_LogReg	250	0.9645	1.0000	0.0000

Berbeda dengan hasil *prefix truncation*, eksperimen *random sentence removal* menunjukkan penurunan performa yang lebih besar karena informasi dihilangkan dari berbagai bagian artikel. TF-IDF dengan Logistic Regression menjadi model yang paling stabil, dengan rentang perubahan Macro F1-Score yang lebih kecil dibandingkan BiLSTM dan DistilBERT. Perbedaan ini dapat dijelaskan oleh karakteristik representasi masing-masing model. TF-IDF mengklasifikasikan dokumen berdasarkan distribusi fitur leksikal tanpa memodelkan urutan kata secara eksplisit, sehingga penghapusan beberapa kalimat tidak selalu menghilangkan keseluruhan fitur diskriminatif selama kata atau frasa penting masih tersedia. Sebaliknya, BiLSTM dan DistilBERT memanfaatkan hubungan kontekstual antartoken. Penghapusan kalimat yang tersebar dapat mengubah konteks dan hubungan semantik yang sebelumnya digunakan model, sehingga prediksi menjadi lebih sensitif dibandingkan pada *prefix truncation*.

Pengujian *source masking* menghasilkan perbedaan yang paling mencolok. Penggantian identitas sumber berita dengan token netral menyebabkan penurunan Macro F1-Score sebesar −0,6685 pada BiLSTM dan −0,6335 pada DistilBERT, dengan *prediction consistency* sekitar 49–50%. Sebaliknya, TF-IDF dengan Logistic Regression hanya mengalami penurunan sebesar −0,0205 dan mempertahankan *prediction consistency* sebesar 97,03%. Penurunan yang besar pada BiLSTM dan DistilBERT mengindikasikan bahwa kedua model memanfaatkan informasi sumber sebagai salah satu sinyal diskriminatif selama proses klasifikasi. Ketika sinyal tersebut dihilangkan, representasi yang dipelajari model berubah secara signifikan sehingga keputusan klasifikasi menjadi tidak stabil. Sebaliknya, TF-IDF dengan Logistic Regression tampak lebih bergantung pada distribusi fitur leksikal yang tersebar pada keseluruhan dokumen, sehingga penghilangan token sumber memberikan dampak yang relatif kecil. Temuan ini juga menunjukkan adanya potensi *shortcut learning* pada model kontekstual, yaitu model dapat memanfaatkan korelasi antara identitas sumber dan label sebagai jalan pintas dalam melakukan prediksi, bukan hanya mengandalkan isi semantik berita.

Secara keseluruhan, hasil penelitian menunjukkan bahwa *robustness* tidak hanya ditentukan oleh kemampuan klasifikasi dasar model, tetapi juga oleh jenis informasi yang dihilangkan dan karakteristik representasi yang digunakan. DistilBERT memiliki performa dan ketahanan terbaik terhadap *prefix truncation*, tetapi keunggulan tersebut tidak bertahan ketika struktur informasi diubah melalui *random sentence removal* atau ketika informasi sumber dihilangkan. Sebaliknya, TF-IDF dengan Logistic Regression memiliki performa dasar lebih rendah, tetapi menunjukkan ketahanan yang lebih baik pada kedua skenario tersebut. Hasil ini menegaskan bahwa performa klasifikasi yang tinggi pada kondisi referensi tidak secara otomatis menunjukkan *robustness* yang lebih tinggi, sehingga evaluasi ketahanan model perlu dilakukan menggunakan beberapa bentuk gangguan tekstual yang berbeda.

4. KESIMPULAN DAN SARAN

Penelitian ini berhasil menganalisis *robustness* tiga model klasifikasi teks, yaitu TF-IDF dengan *Logistic Regression*, BiLSTM, dan DistilBERT terhadap pengurangan informasi tekstual pada deteksi berita palsu. Hasil penelitian menunjukkan bahwa DistilBERT memberikan performa terbaik pada kondisi referensi dengan *Macro F1-Score* sebesar 0,9989 serta mampu mempertahankan performa secara konsisten pada skenario *prefix truncation* hingga menghasilkan *performance retention* sebesar 100%.

Hasil eksperimen juga menunjukkan bahwa tingkat *robustness* model bergantung pada bentuk pengurangan informasi yang diberikan. Pada skenario *random sentence removal* dan *source masking*, TF-IDF dengan *Logistic Regression* menunjukkan ketahanan yang lebih baik dibandingkan BiLSTM dan DistilBERT. Temuan ini menunjukkan bahwa model dengan performa

klasifikasi terbaik tidak selalu menjadi model yang paling tahan terhadap seluruh bentuk pengurangan informasi tekstual.

Berdasarkan hasil tersebut, penelitian ini menunjukkan bahwa evaluasi *robustness* sebaiknya tidak hanya dilakukan menggunakan satu skenario pengujian, tetapi perlu mempertimbangkan berbagai bentuk pengurangan informasi tekstual agar karakteristik ketahanan model dapat dievaluasi secara lebih komprehensif. Penelitian selanjutnya dapat memperluas evaluasi menggunakan dataset dari berbagai domain maupun bahasa yang berbeda, serta mengombinasikan skenario pengurangan teks dengan *adversarial attack* atau teknik augmentasi data sehingga kemampuan generalisasi dan ketahanan model dapat dianalisis secara lebih menyeluruh.

Daftar Pustaka:

[1] M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 114435–114459, 2024, doi: 10.1109/ACCESS.2024.3435497.

[2] M. Mostafa, A. S. Almogren, M. Al-Qurishi, and M. Alrubaian, "Modality Deep-learning Frameworks for Fake News Detection on Social Networks: A Systematic Literature Review," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–50, Mar. 2025, doi: 10.1145/3700748.

[3] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Front. Big Data*, vol. 6, p. 1320800, Jan. 2024, doi: 10.3389/fdata.2023.1320800.

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

[5] D. Baker, "Çakı et al. / JAIDA vol (2024) 22-32," 2024.

[6] R. Austin and A. Syahrian, "ANALISIS PERFORMA ALGORITMA SUPERVISED LEARNING TERHADAP DATA DESKRIPSI DENGAN REPRESENTASI DAN PARAMETER TUNING," *Jurnal Informatika*.

[7] H. Çetiner, "Fake News Detection and Classification with Recurrent Neural Network Based Deep Learning Approaches," *OKÜ Fen Bil. Ens. Dergisi ((OKU Journal of Nat. & App. Sci)*, vol. 7, no. 3, pp. 973–993, Jun. 2024, doi: 10.47495/okufbed.1199738.

[8] J. Reddy, S. Mundra, and A. Mundra, "Ensembling Deep Learning Models for Fake News Classification," *Procedia Computer Science*, vol. 235, pp. 2766–2774, 2024, doi: 10.1016/j.procs.2024.04.261.

[9] M. Tan and H. Bakır, "Fake News Detection Using BERT and Bi-LSTM with Grid Search Hyperparameter Optimization," *Bilişim Teknolojileri Dergisi*, vol. 18, no. 1, pp. 11–28, Jan. 2025, doi: 10.17671/gazibtd.1521520.

[10] S. G. Sneha, A. Sen, S. Malik, A. Uddin, and Md. J. Hossen, "BiLSTM-LIME: integrating NLP and advanced machine learning models for fake news detection," *Discov Artif Intell*, vol. 6, no. 1, p. 125, Jan. 2026, doi: 10.1007/s44163-026-00852-w.

[11] M. Farokhian, V. Rafe, and H. Veisi, "Fake news detection using parallel BERT deep neural networks," *Multimed Tools Appl*, vol. 83, no. 15, pp. 43831–43848, Oct. 2023, doi: 10.1007/s11042-023-17115-w.

[12] N. Raza *et al.*, "Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach," *PLoS One*, vol. 20, no. 9, p. e0330954, Sep. 2025, doi: 10.1371/journal.pone.0330954.

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

[14] S. Singhania, N. Fernandez, and S. Rao, "3HAN: A Deep Neural Network for Fake News Detection," vol. 10635, 2017, pp. 572–581. doi: 10.1007/978-3-319-70096-0_59.

[15] M. Al-alshaqi, D. B. Rawat, and C. Liu, "Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning," *Sensors*, vol. 24, no. 18, p. 6062, Sep. 2024, doi: 10.3390/s24186062.

[16] D. F. Brianna, M. A. Saputra, and M. A. Hapiz, "Model Deteksi Berita Palsu Menggunakan BERT dan Bi-LSTM Berbasis Discriminative Approach," vol. 08, no. 3, 2025.

[17] T. C. Praha, W. Widodo, and M. Nugraheni, "Indonesian Fake News Classification Using Transfer Learning in CNN and LSTM," *JOIV : Int. J. Inform. Visualization*, vol. 8, no. 3, p. 1213, Sep. 2024, doi: 10.62527/joiv.8.2.2126.

[18] J. Wu, J. Guo, and B. Hooi, "Fake News in Sheep's Clothing: Robust Fake News

Detection Against LLM-Empowered Style Attacks," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Aug. 2024, pp. 3367–3378. doi: 10.1145/3637528.3671977.

[19] S. Maham, A. Tariq, M. U. G. Khan, F. S. Alamri, A. Rehman, and T. Saba, "ANN: adversarial news net for robust fake news classification," *Sci Rep*, vol. 14, no. 1, p. 7897, Apr. 2024, doi: 10.1038/s41598-024-56567-4.

[20] A. Koren, C. Underwood, E. Serra, and F. Spezzano, "Evaluating the robustness of fake news detectors to adversarial attacks with real user comments," *Int J Data Sci Anal*, vol. 20, no. 6, pp. 5689–5702, Nov. 2025, doi: 10.1007/s41060-025-00790-3.

[21] X. Pu, H. Wu, X. Bi, Y. Wu, and X. Gao, "DEAR: Disentangled Event-Agnostic Representation Learning for Early Fake News Detection".

[22] J. Jukic and J. ˇSnajder, "From Robustness to Improved Generalization and Calibration in Pre-trained Language Models".

[23] P. Vickers, L. Barrault, E. Monti, and N. Aletras, "We Need to Talk About Classification Evaluation Metrics in NLP," Jan. 08, 2024, *arXiv*: arXiv:2401.03831. doi: 10.48550/arXiv.2401.03831.

[24] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 820–836, Jun. 2024, doi: 10.1162/tacl_a_00675.

[25] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake News Detection: Comparative Evaluation of BERT-like Models and Large Language Models with Generative AI-Annotated Data," Dec. 20, 2024, *arXiv*: arXiv:2412.14276. doi: 10.48550/arXiv.2412.14276.