BAB 2
LANDASAN TEORI

2.1 Hoaks (*Hoax*)

Hoaks (*hoax*) didefinisikan dalam literatur akademik sebagai konten informasi yang secara sengaja direkayasa untuk menyesatkan pembaca, dengan menyajikan klaim yang tidak berdasar seolah merupakan pelaporan berita yang sah dan faktual [1]. Definisi ini membedakan hoaks dari kesalahan jurnalistik yang tidak disengaja (*misinformation*). Hoaks selalu mengandung unsur intensionalitas (*disinformation*), yakni niat untuk memanipulasi persepsi publik melalui fabrikasi atau distorsi fakta. Ruang lingkup operasionalnya mencakup empat tipe utama: (a) artikel yang seluruh isinya direkayasa tanpa basis faktual, (b) konten visual yang dimanipulasi melalui penyuntingan digital, (c) peristiwa nyata yang dikutip di luar konteks sehingga makna aslinya terdistorsi, dan (d) informasi parsial yang dikombinasikan dengan judul atau narasi yang menyimpang [5].

Dari sudut pandang deteksi otomatis, hoaks diklasifikasikan berdasarkan atribut linguistik dan struktural yang membedakannya dari berita valid. Kajian terhadap berbagai pendekatan deteksi menemukan bahwa hoaks cenderung menggunakan diksi yang lebih emosional, klaim yang tidak terverifikasi, serta pola retorika tertentu yang dapat diidentifikasi oleh model pembelajaran mesin [5]. Hal ini menjadi dasar pengembangan sistem deteksi berbasis teks yang memproses konten artikel sebagai sekuens token dan mengekstraksi representasi semantiknya untuk keperluan klasifikasi.

2.2 Pemeriksaan Fakta Otomatis (*Automated Fact-Checking*)

Pemeriksaan fakta (*fact-checking*) merupakan proses verifikasi kebenaran suatu klaim melalui pencocokan terhadap sumber-sumber yang dapat dipertanggungjawabkan. Dalam konteks otomasi berbasis kecerdasan buatan, *automated fact-checking* didefinisikan oleh Guo et al. [17] sebagai tugas komputasional terstruktur yang terdiri dari empat sub-tahap berurutan: (1) *check-worthy claim detection*, yakni identifikasi pernyataan dalam suatu dokumen yang memerlukan verifikasi; (2) *evidence retrieval*, yakni pengambilan dokumen-dokumen pendukung dari sumber referensi terpercaya; (3) *claim verification*, yakni penalaran terhadap bukti yang diperoleh untuk menentukan veridict klaim;

dan (4) *justification*, yakni pembentukan penjelasan yang mendukung keputusan verifikasi yang dihasilkan.

Penting untuk dibedakan antara *automated fact-checking* dan klasifikasi hoaks (*hoax classification*) sebagai dua tugas yang berbeda secara fundamental, meskipun keduanya berkaitan erat. Klasifikasi hoaks beroperasi pada tingkat dokumen, di mana seluruh artikel diklasifikasikan secara holistik sebagai palsu atau valid berdasarkan fitur-fitur teks keseluruhannya [5]. Sebaliknya, *automated fact-checking* beroperasi pada tingkat klaim, dimana bagian ini mengidentifikasi dan memverifikasi pernyataan-pernyataan spesifik dalam teks dengan menyertakan bukti eksplisit yang mendukung keputusan verifikasinya [17]. Dalam penelitian ini, sistem yang dikembangkan mengintegrasikan kedua pendekatan, dimana model Bi-LSTM digunakan untuk mengklasifikasikan dokumen berita secara keseluruhan, sementara kerangka agentik yang melingkupinya memberikan dimensi proaktif yang merepresentasikan aspek pertama dari *automated fact-checking*, yaitu inisiasi mandiri proses verifikasi terhadap aliran berita nasional yang masuk secara berkelanjutan.

2.3 Sistem Agentik (*Agentic AI System*)

Sistem agentik (*agentic AI system*) didefinisikan dalam literatur modern sebagai sistem komputasional yang beroperasi secara mandiri untuk mencapai tujuan yang ditetapkan, dicirikan oleh empat kapabilitas utama yang membedakannya dari program komputer konvensional maupun sistem percakapan reaktif. Pertama, *otonomi* (*autonomy*), yaitu sistem menjalankan seluruh alurnya sendiri tanpa bergantung pada instruksi eksplisit per-langkah dari pengguna. Kedua, *perencanaan mandiri* (*self-directed planning*), yaitu sistem mampu menyusun rangkaian tindakan untuk mencapai tujuannya secara proaktif, bukan menunggu kueri dari luar ataupun masukan manual. Ketiga, *persepsi lingkungan* (*environmental reactivity*), yaitu sistem memantau kondisi lingkungannya, termasuk masukan data baru, dan menyesuaikan perilakunya secara adaptif terhadap perubahan yang terjadi. Keempat, *aktuasi* (*action execution*), yaitu sistem mengeksekusi tindakan konkret terhadap lingkungan sebagai keluaran dari proses penalaran internalnya [18].

Keempat kapabilitas tersebut bersesuaian langsung dengan empat sifat karakteristik yang secara konsisten diidentifikasi dalam literatur sistem agentik otonom [18]: reaktivitas (merespons masukan dari lingkungan secara tepat

waktu), proaktivitas (mengambil inisiatif tanpa menunggu instruksi eksplisit), kemampuan sosial (berinteraksi dengan pengguna manusia untuk mengumpulkan koreksi dan masukan), serta otonomi (menjalankan seluruh alur kerjanya sendiri tanpa intervensi manual per-langkah). Dalam sistem yang dikembangkan penelitian ini, keempat sifat tersebut diimplementasikan melalui lima tahap pemrosesan yang berjalan berurutan dalam satu alur kerja, yaitu perumusan tujuan, penyusunan rencana internal, penggunaan alat bantu pencarian bukti, penalaran dan pengambilan keputusan, serta eksekusi dan refleksi hasil.

2.4 Estimasi Ketidakpastian Model (*Monte Carlo Dropout*)

Jaringan saraf tiruan konvensional menghasilkan prediksi deterministik yang tidak membedakan antara ketidakpastian yang bersumber dari variasi data (*aleatoric uncertainty*) dan ketidakpastian yang bersumber dari keterbatasan model itu sendiri (*epistemic uncertainty*). Kajian komprehensif terhadap metode estimasi ketidakpastian dalam sistem klasifikasi pembelajaran mendalam menunjukkan bahwa penerapan *dropout* pada saat inferensi, bukan hanya selama pelatihan, setara dengan melakukan aproksimasi *Bayesian* terhadap distribusi posterior bobot model [23]. Pendekatan ini dikenal sebagai *Monte Carlo Dropout* (MC Dropout).

Secara operasional, MC Dropout menjalankan satu masukan melalui model sebanyak T kali dengan *dropout* tetap aktif, menghasilkan T vektor probabilitas berbeda $\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_T\}$. Estimasi probabilitas akhir adalah rata-rata:

$$\bar{p} = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_t \tag{2.1}$$

di mana $\bar{p}$ menyatakan estimasi probabilitas akhir yang dihasilkan setelah proses rata-rata, dimana $\sum_{t=1}^{T}$ adalah notasi penjumlahan yang dimulai dari sampel pertama ($t = 1$) hingga sampel terakhir ke-T; $\frac{1}{T}$ adalah faktor pembagi yang menormalisasi hasil penjumlahan menjadi rata-rata; T adalah jumlah total sampel Monte Carlo (*forward pass*) yang dijalankan pada model, dimana pada penelitian ini ditetapkan sebanyak 15 kali; t adalah indeks yang menunjukkan sampel Monte Carlo ke berapa sedang diproses, dengan nilai berjalan dari 1 sampai T; dan $\hat{p}_t$ adalah vektor probabilitas keluaran model pada sampel ke-t, yaitu hasil satu kali *forward pass* dengan pola *dropout* yang aktif dan berbeda secara acak dari sampel lainnya.

Sedangkan ketidakpastian epistemik diestimasi dari variansi nilai

kepercayaan tertinggi di setiap sampel:

$$\sigma^2_{\text{epistemic}} = \text{Var}\left(\max_c \hat{p}_{t,c}\right)_{t=1}^{T} \tag{2.2}$$

di mana $\sigma^2_{\text{epistemic}}$ menyatakan nilai ketidakpastian epistemik yang diestimasi, yaitu ukuran seberapa besar variasi keyakinan model terhadap prediksinya sendiri di seluruh sampel Monte Carlo; $\text{Var}(\cdot)$ adalah operator variansi statistik yang mengukur sebaran nilai di dalam kurung terhadap nilai rata-ratanya; $\max_c$ adalah operator yang mengambil nilai probabilitas tertinggi di antara seluruh kelas c yang tersedia pada satu sampel tertentu; $\hat{p}_{t,c}$ adalah probabilitas keluaran model untuk kelas c pada sampel Monte Carlo ke-t; c adalah indeks kelas keluaran, dalam penelitian ini bernilai *fakta* atau *hoaks*; t adalah indeks sampel Monte Carlo yang berjalan dari 1 hingga T; dan T adalah jumlah total sampel Monte Carlo yang dijalankan, sama seperti pada Persamaan (2.1). Notasi $(\max_c \hat{p}_{t,c})_{t=1}^{T}$ secara keseluruhan menyatakan kumpulan T buah nilai kepercayaan tertinggi, satu nilai untuk setiap sampel Monte Carlo, yang kemudian dihitung variansinya.

2.5 Penanganan Ketidakseimbangan Kelas (*Class Imbalance*)

Ketidakseimbangan kelas (*class imbalance*) merupakan tantangan yang umum dalam dataset deteksi hoaks, di mana artikel faktual dari portal berita resmi jauh lebih banyak tersedia dibandingkan sampel hoaks yang telah terverifikasi [14]. Model yang dilatih pada distribusi kelas yang tidak seimbang cenderung bias ke kelas mayoritas, menghasilkan akurasi keseluruhan yang tinggi namun performa yang buruk pada kelas minoritas yang justru merupakan target deteksi utama.

Penelitian ini mengatasi permasalahan ini melalui pembobotan kelas berbasis invers-frekuensi (*inverse-frequency class weighting*) pada fungsi kerugian *CrossEntropy*:

$$w_c = \frac{N}{K \times n_c} \tag{2.3}$$

di mana N adalah total sampel pelatihan, K adalah jumlah kelas, dan n_c adalah jumlah sampel pada kelas c. Bobot ini diberikan pada fungsi kerugian sehingga kesalahan pada kelas minoritas mendapat penalti yang proporsional lebih besar, mendorong model untuk memprioritaskan akurasi pada kelas hoaks tanpa mengabaikan kelas faktual [5].

2.6 Algoritma *Long Short-Term Memory* (LSTM)

Long Short-Term Memory (LSTM) adalah arsitektur jaringan saraf berulang (*Recurrent Neural Network*/RNN) yang dirancang untuk mengatasi permasalahan *vanishing gradient* yang secara inheren dialami oleh RNN standar saat memproses sekuens panjang. Permasalahan *vanishing gradient* muncul karena pembaruan bobot melalui algoritma *backpropagation through time* (BPTT) menghasilkan gradien yang mengecil secara eksponensial seiring bertambahnya jarak temporal, sehingga informasi dari posisi awal sekuens gagal dipropagasikan secara efektif ke lapisan-lapisan selanjutnya [19].

LSTM mengatasi permasalahan ini melalui pengenalan mekanisme *cell state* ($\mathbf{c}_t$) sebagai jalur informasi jangka panjang yang dilindungi oleh tiga gerbang yang dapat dipelajari (*learnable gates*) [19]:

(a) **Gerbang Lupa** (*Forget Gate*, $\mathbf{f}_t$): memutuskan informasi apa dari *cell state* sebelumnya yang perlu dilupakan, dihitung sebagai:

$$\mathbf{f}_t = \sigma\big(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\big) \tag{2.4}$$

Persamaan (2.4) menghitung nilai gerbang lupa, di mana $\mathbf{f}_t$ adalah vektor keluaran gerbang lupa pada langkah waktu t yang bernilai antara 0 dan 1 untuk setiap elemen *cell state*; σ adalah fungsi aktivasi sigmoid yang memetakan nilai ke rentang $[0,1]$; $\mathbf{W}_f$ adalah matriks bobot yang dapat dipelajari untuk gerbang lupa; $\mathbf{h}_{t-1}$ adalah keadaan tersembunyi (*hidden state*) dari langkah waktu sebelumnya; $\mathbf{x}_t$ adalah vektor masukan pada langkah waktu t; $[\mathbf{h}_{t-1}, \mathbf{x}_t]$ adalah konkatenasi (penggabungan) vektor $\mathbf{h}_{t-1}$ dan $\mathbf{x}_t$; dan $\mathbf{b}_f$ adalah vektor bias yang dapat dipelajari untuk gerbang lupa.

(b) **Gerbang Masukan** (*Input Gate*, $\mathbf{i}_t$): memutuskan informasi baru apa yang akan ditambahkan ke *cell state*, bekerja bersama vektor kandidat $\tilde{\mathbf{c}}_t$:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i), \quad \tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \tag{2.5}$$

Persamaan (2.5) menghitung nilai gerbang masukan beserta vektor kandidat, di mana $\mathbf{i}_t$ adalah vektor keluaran gerbang masukan pada langkah waktu t yang menentukan proporsi informasi baru yang akan disimpan; $\tilde{\mathbf{c}}_t$ adalah vektor kandidat *cell state* baru yang berisi informasi kandidat yang berpotensi

ditambahkan; σ adalah fungsi aktivasi sigmoid; tanh adalah fungsi aktivasi tangen hiperbolik yang memetakan nilai ke rentang $[-1,1]$; $\mathbf{W}_i$ dan $\mathbf{W}_c$ adalah matriks bobot yang dapat dipelajari, masing-masing untuk gerbang masukan dan vektor kandidat; $\mathbf{b}_i$ dan $\mathbf{b}_c$ adalah vektor bias yang dapat dipelajari, masing-masing untuk gerbang masukan dan vektor kandidat; sedangkan $\mathbf{h}_{t-1}$ dan $\mathbf{x}_t$ adalah keadaan tersembunyi sebelumnya dan vektor masukan pada langkah waktu t sebagaimana didefinisikan pada Persamaan (2.4).

(c) **Pembaruan *Cell State***:

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \tag{2.6}$$

Persamaan (2.6) menghitung pembaruan *cell state*, di mana $\mathbf{c}_t$ adalah *cell state* pada langkah waktu t, yaitu jalur memori jangka panjang yang telah diperbarui; $\mathbf{c}_{t-1}$ adalah *cell state* dari langkah waktu sebelumnya; $\mathbf{f}_t$ adalah keluaran gerbang lupa (Persamaan 2.4) yang menentukan porsi $\mathbf{c}_{t-1}$ yang dipertahankan; $\mathbf{i}_t$ adalah keluaran gerbang masukan (Persamaan 2.5) yang menentukan porsi $\tilde{\mathbf{c}}_t$ yang ditambahkan; $\tilde{\mathbf{c}}_t$ adalah vektor kandidat *cell state* (Persamaan 2.5); dan $\odot$ adalah operator perkalian elemen demi elemen (*Hadamard product*).

(d) **Gerbang Keluaran** (*Output Gate*, $\mathbf{o}_t$): menentukan keadaan tersembunyi $\mathbf{h}_t$ yang diteruskan ke langkah berikutnya:

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o), \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{2.7}$$

Persamaan (2.7) menghitung nilai gerbang keluaran beserta keadaan tersembunyi, di mana $\mathbf{o}_t$ adalah vektor keluaran gerbang keluaran pada langkah waktu t yang menentukan porsi *cell state* yang akan diteruskan sebagai *hidden state*; $\mathbf{h}_t$ adalah keadaan tersembunyi (*hidden state*) pada langkah waktu t yang diteruskan ke langkah berikutnya sekaligus dapat digunakan sebagai keluaran; $\mathbf{W}_o$ adalah matriks bobot yang dapat dipelajari untuk gerbang keluaran; $\mathbf{b}_o$ adalah vektor bias yang dapat dipelajari untuk gerbang keluaran; $\mathbf{c}_t$ adalah *cell state* hasil pembaruan (Persamaan 2.6); sedangkan σ, tanh, dan $\odot$ berturut-turut adalah fungsi sigmoid, fungsi tangen hiperbolik, dan operator perkalian elemen demi elemen sebagaimana didefinisikan sebelumnya.

2.7 Algoritma *Bidirectional Long Short-Term Memory* (Bi-LSTM)

Bidirectional RNN (BiRNN) merupakan perluasan arsitektur RNN searah yang memproses sekuens masukan dari dua arah secara bersamaan [20]. Ide dasarnya bertumpu pada observasi bahwa prediksi pada posisi t dalam suatu sekuens sering kali tidak hanya bergantung pada konteks masa lalu $(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$, tetapi juga pada konteks masa depan $(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T)$ [20].

Dalam arsitektur BiRNN, dua lapisan RNN terpisah dioperasikan secara paralel pada sekuens yang sama, di mana lapisan *forward* memproses sekuens dari kiri ke kanan menghasilkan keadaan tersembunyi $\overrightarrow{\mathbf{h}}_t$, sementara lapisan *backward* memproses sekuens dari kanan ke kiri menghasilkan keadaan tersembunyi $\overleftarrow{\mathbf{h}}_t$. Representasi akhir pada setiap posisi t diperoleh melalui konkatenasi kedua keadaan tersembunyi [20]:

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t \,\|\, \overleftarrow{\mathbf{h}}_t \right] \tag{2.8}$$

di mana $\mathbf{h}_t$ menyatakan representasi akhir (keadaan tersembunyi gabungan) pada posisi atau langkah waktu t; $\overrightarrow{\mathbf{h}}_t$ adalah keadaan tersembunyi yang dihasilkan oleh lapisan *forward*, yaitu hasil pemrosesan sekuens dari kata pertama menuju kata ke-t (arah kiri ke kanan); $\overleftarrow{\mathbf{h}}_t$ adalah keadaan tersembunyi yang dihasilkan oleh lapisan *backward*, yaitu hasil pemrosesan sekuens dari kata terakhir mundur menuju kata ke-t (arah kanan ke kiri), dan operator $\|$ menyatakan konkatenasi (penggabungan) kedua vektor tersebut menjadi satu vektor representasi tunggal yang memuat konteks dari kedua arah baca secara bersamaan.

Ketika komponen RNN digantikan dengan sel LSTM, arsitektur ini dikenal sebagai *Bidirectional LSTM* (Bi-LSTM). Bi-LSTM mewarisi seluruh mekanisme gerbang LSTM untuk menangani dependensi jangka panjang, sekaligus memperoleh pemahaman konteks bidireksional dari BiRNN. Dalam konteks klasifikasi teks berita, keunggulan ini bersifat langsung operasional, di mana makna sebuah klaim dalam berita sering kali tidak dapat ditentukan hanya dari arah baca kiri ke kanan. Frasa penentu seperti penyangkalan, modalitas, atau kuantifikasi kerap muncul setelah klaim yang harus dievaluasinya. Bi-LSTM mampu mengintegrasikan kedua arah konteks ini dalam satu representasi terpadu [20].

Dalam implementasi yang dikembangkan pada penelitian ini, lapisan Bi-LSTM yang dikonfigurasi untuk memproses sekuens dari dua arah secara bersamaan menghasilkan tensor keadaan tersembunyi akhir **hn** berukuran

$[2, N_{\text{batch}}, D_h]$, di mana $D_h = 64$ adalah dimensi unit tersembunyi per arah. Representasi dokumen diperoleh dengan mengambil $\mathbf{hn}[-2]$ sebagai keadaan *forward* akhir dan $\mathbf{hn}[-1]$ sebagai keadaan *backward* akhir yang kemudian digabungkan menjadi vektor berdimensi $2 \times 64 = 128$. Vektor gabungan ini selanjutnya diproses oleh lapisan linier penuh untuk menghasilkan probabilitas kelas [4].

Sejumlah penelitian komparatif yang menerapkan Bi-LSTM pada teks berita palsu menunjukkan keunggulannya atas LSTM unidireksional dan pendekatan berbasis statistik konvensional, baik pada teks berbahasa Inggris [4] maupun pada pendekatan hibrida dengan pembelajaran lemah [8]. Pada konteks hoaks berbahasa Indonesia secara spesifik, temuan [6] turut mengonfirmasi bahwa arsitektur bidireksional secara khusus mendapat manfaat dari tidak diterapkannya *stemming*, karena morfologi aglutinasi Bahasa Indonesia menyimpan informasi gramatikal dalam afiksasi yang akan hilang bila dilakukan pengupasan [7].

2.8 Definisi dan Ruang Lingkup Berita Nasional

Penelitian ini beroperasi dalam domain berita nasional berbahasa Indonesia, sehingga diperlukan pembatasan konseptual yang tegas mengenai apa yang dimaksud dengan *berita nasional* dalam konteks deteksi hoaks berbantuan kecerdasan buatan. Dua dimensi pembatasan yang relevan adalah dimensi kelembagaan (di mana dan oleh siapa berita diproduksi) serta dimensi linguistik (dalam bahasa apa berita tersebut disajikan).

Dari dimensi kelembagaan, penelitian-penelitian terdahulu tentang berita palsu berbahasa Indonesia merupakan istilah payung yang turut mencakup hoaks sebagai salah satu subkategorinya [24] yang secara konsisten mendefinisikan ruang lingkup studinya berdasarkan sumber konten yang terdaftar dan beroperasi di bawah yurisdiksi Indonesia. Rohman et al. [10] dalam tinjauan literatur sistematis terhadap 31 studi deteksi berita palsu berbahasa Indonesia mengidentifikasi bahwa sumber data yang paling umum digunakan adalah portal berita Indonesia (23%), situs pemantauan disinformasi resmi pemerintah seperti *turnbackhoax.id* (26%), yakni lembaga pemeriksa fakta hoaks yang menjadi sumber label *hoax* pada dataset penelitian ini serta platform media sosial yang beroperasi di ekosistem digital Indonesia (20%). Berdasarkan informasi tersebut, *berita nasional* dalam penelitian ini merujuk pada konten jurnalistik yang diproduksi dan disebarluaskan oleh entitas media yang sudah terdaftar serta beroperasi di bawah yurisdiksi hukum Indonesia,

terlepas dari wilayah geografis peristiwa yang diliput. Pembatasan kelembagaan ini penting karena hal ini memastikan bahwa konten yang dianalisis tunduk pada regulasi dan mekanisme pertanggungjawaban yang berlaku di Indonesia, termasuk kewajiban etika jurnalistik yang diatur oleh Dewan Pers.

Dari dimensi linguistik, berita nasional dalam penelitian ini secara operasional dibatasi pada konten yang disajikan dalam Bahasa Indonesia. Pembatasan ini bertumpu pada dua justifikasi yang saling melengkapi. *Pertama*, justifikasi mediatik, di mana seluruh penelitian deteksi berita palsu dalam konteks Indonesia menggunakan Bahasa Indonesia sebagai kriteria pembatas dataset yang bersifat non-negosiabel. Hal ini mengingat bahwa konten yang menyasar khalayak Indonesia diproduksi dan dikonsumsi terutama dalam Bahasa Indonesia [10]. *Kedua*, justifikasi teknis-linguistik. Bahasa Indonesia memiliki karakteristik morfologi aglutinasi dan struktur linguistik yang unik sehingga model NLP yang dilatih pada korpus bahasa lain tidak dapat digeneralisasi langsung [7]. Hal ini dikonfirmasi oleh sejumlah penelitian yang mengembangkan model bahasa khusus Bahasa Indonesia yang menunjukkan bahwa model multibahasa umum tidak mampu merepresentasikan struktur linguistik Bahasa Indonesia secara memadai tanpa proses adaptasi atau pelatihan ulang menyeluruh [11]. Penelitian deteksi hoaks berbasis *machine learning* yang berfokus pada konteks Indonesia secara konsisten menggunakan Bahasa Indonesia sebagai kriteria pembatas dataset, baik yang menggunakan pendekatan Naïve Bayes [12] maupun model LSTM dan CNN [13], keduanya secara eksplisit menyasar deteksi *hoax* sesuai judul studi aslinya. Cakupan ini juga konsisten dengan penelitian deteksi berita palsu berbasis arsitektur Transformer pada teks berita populer Indonesia secara lebih umum [14].

Atas dasar itu, dalam penelitian ini *berita nasional* didefinisikan secara operasional sebagai konten jurnalistik berbahasa Indonesia yang diproduksi oleh entitas media yang terdaftar dan beroperasi di bawah yurisdiksi hukum Indonesia, serta disebarluaskan melalui platform digital yang ditujukan bagi khalayak Indonesia. Definisi ini mencakup berita dari portal daring nasional, situs media massa terakreditasi Dewan Pers, maupun konten yang tervalidasi pada sistem pemantauan disinformasi resmi pemerintah seperti *www.komdigi.go.id* [6]. Sebaliknya, konten berbahasa Inggris yang diproduksi oleh media berinduk Indonesia, misalnya *The Jakarta Post*, berada di luar cakupan penelitian ini, mengingat perbedaan struktur morfologi aglutinasi dan tantangan pemrosesan linguistik yang secara fundamental berbeda dari bahasa-bahasa yang umumnya digunakan sebagai basis pelatihan model multibahasa [11]. Demikian pula, konten

berbahasa Indonesia yang diproduksi oleh entitas media asing tidak termasuk dalam kategori berita nasional sebagaimana dimaksud dalam penelitian ini, sesuai dengan pembatasan kelembagaan yang digunakan secara konsisten dalam literatur deteksi berita palsu Indonesia [10].

2.9 Penelitian Terdahulu

Bagian ini memaparkan penelitian-penelitian terdahulu yang relevan secara langsung dengan topik penelitian ini. Perlu dicatat bahwa sebagian studi yang ditinjau menggunakan istilah *fake news* sesuai cakupan aslinya yang lebih luas, sementara penelitian ini secara spesifik menyasar hoaks sebagai salah satu subkategorinya [24, 25], dimana peninjauan tetap dilakukan karena arsitektur dan teknik yang diusulkan bersifat dapat ditransfer (*transferable*) ke tugas klasifikasi hoaks. Dalam beberapa tahun terakhir, penelitian di bidang deteksi berita palsu berbasis pembelajaran mendalam mengalami perkembangan pesat seiring dengan meningkatnya ketersediaan dataset berlabel dan kemajuan arsitektur jaringan saraf berulang. Model konvensional berbasis statistik statis seperti TF-IDF dan *Count Vectorizer* telah banyak digunakan untuk merepresentasikan teks berita, namun kinerjanya sering kali terbatas dalam menangkap pola kebahasaan yang halus dan bergantung pada konteks. Oleh karena itu, penelitian terbaru mulai berfokus pada arsitektur bidireksional dan hibrida yang tidak hanya memproses fitur leksikal, tetapi juga memahami dependensi kontekstual jangka panjang dalam urutan teks. Penelitian ini mengacu pada tiga kelompok studi yang disajikan pada Tabel 2.1, yaitu studi dengan arsitektur Bi-LSTM dan *attention* untuk klasifikasi berita palsu, studi *Named Entity Recognition* (NER) pada berita palsu berbahasa Indonesia, serta pendekatan pembelajaran lemah hibrida dan pembelajaran mendalam untuk deteksi disinformasi.

Tabel 2.1. Penelitian terdahulu berdasarkan metode yang digunakan

No	Tahun	Masalah	Dataset	Model	Result (Akurasi/F1)	Main Findings
Studi Literatur Arsitektur Bi-LSTM dan Attention untuk Klasifikasi Berita Palsu						

No	Tahun	Masalah	Dataset	Model	Result (Akurasi/F1)	Main Findings
1	2024	Klasifikasi berita palsu menggunakan model pembelajaran mendalam berbasis LSTM	WELFake (72.134 artikel; gabungan Kaggle, McIntire, Reuters, BuzzFeed)	Bi-LSTM; *Attention-based* Bi-LSTM	Akurasi 97,49% (Bi-LSTM); 97,66% (Att-BiLSTM); F1 97,48%; 97,62%	*Attention-based* Bi-LSTM melampaui seluruh model pembanding, dimana mekanisme atensi memberikan fokus pada frasa paling indikatif terhadap keaslian berita [4].
Studi Literatur NER pada Berita Palsu Berbahasa Indonesia						
2	2025	*Named Entity Recognition* pada teks berita palsu berbahasa Indonesia menggunakan PoS *tagging* dan model BERT selama kampanye Pemilu Presiden 2024	1.083 dokumen berita palsu tervalidasi dari *www.komdigi.go.id* (Oktober 2023–Mei 2024)	BiLSTM-CRF; BiGRU; BERT (*IndoBERT Base*)	F1 81,26% (BiLSTM-CRF + PoS); 79,46% (BiGRU + PoS); 87,38% (BERT + PoS)	Integrasi PoS *tagging* meningkatkan deteksi entitas; *stemming* justru menurunkan performa pada seluruh arsitektur; BERT+PoS memberikan hasil terbaik [6].
Studi Literatur Pembelajaran Lemah Hibrida dan Pembelajaran Mendalam						

UNIVERSITAS MULTIMEDIA NUSANTARA

No	Tahun	Masalah	Dataset	Model	Result (Akurasi/F1)	Main Findings
3	2023	Deteksi berita palsu dari propaganda siber menggunakan pembelajaran lemah hibrida pada data Twitter yang sebagian besar tidak berlabel	195.943 *tweet* (6 topik: pemilu AS, COVID-19, ledakan Beirut, Israel-Palestina, pembelajaran jarak jauh, banjir Sudan)	*Weakly Supervised* SVM + Bi-LSTM; *Weakly Supervised* SVM + Bi-GRU	Akurasi 89,95% (Bi-LSTM); 89,60% (Bi-GRU); F1 89% (keduanya)	Kombinasi Bi-LSTM dan Bi-GRU dengan SVM berbasis *transductive learning* unggul atas LSTM searah (86,19%) dan CNN (82,06%) dalam kondisi data berlabel terbatas [8].

Berdasarkan Tabel 2.1, berbagai penelitian terdahulu menunjukkan perkembangan signifikan dalam deteksi berita palsu menggunakan pendekatan pembelajaran mendalam, baik untuk teks berbahasa Inggris maupun berbahasa Indonesia. Pada kelompok studi dengan arsitektur Bi-LSTM dan mekanisme atensi, penelitian [4] memperlihatkan kemajuan berarti dalam peningkatan akurasi klasifikasi artikel berita. Model Bi-LSTM yang diusulkan mencapai akurasi 97,49 persen, sedangkan model *attention-based* Bi-LSTM mencapai akurasi tertinggi sebesar 97,66 persen pada dataset WELFake yang terdiri dari 72.134 artikel dari empat sumber. Keunggulan model berbasis atensi ini terletak pada kemampuannya memberikan bobot lebih tinggi pada frasa atau pola dalam konten berita yang paling indikatif terhadap keaslian informasi, sehingga model tidak hanya bergantung pada pola leksikal permukaan, tetapi juga pada konteks semantik yang lebih dalam. Penelitian tersebut juga mengonfirmasi bahwa keunggulan Bi-LSTM atas RNN dan LSTM konvensional bertumpu pada kemampuan pemrosesan konteks dua arah sebagaimana dijelaskan dalam kajian arsitektur RNN terkini [20], bahwa lapisan *forward* menangkap dependensi dari kiri ke kanan, sementara lapisan *backward* menangkap dependensi dari kanan ke kiri, sehingga setiap token memperoleh representasi yang mencerminkan konteks dari kedua sisinya secara simultan.

Pada kelompok studi NER untuk berita palsu berbahasa Indonesia, penelitian [6] mengembangkan pendekatan baru yang mengintegrasikan penandaan kelas kata (*Part-of-Speech*/PoS *tagging*) sebagai tahap pra-pemrosesan sebelum deteksi entitas dilakukan. Evaluasi komparatif terhadap tiga arsitektur, yaitu BiLSTM-CRF, BiGRU, dan BERT berbasis *IndoBERT Base* menunjukkan bahwa model BERT tanpa *stemming* namun dengan

penyaringan PoS mencapai F1-*Score* tertinggi sebesar 87,38 persen. Temuan penting dari penelitian ini adalah bahwa penerapan *stemming* pada teks berbahasa Indonesia secara konsisten menurunkan performa deteksi entitas pada seluruh arsitektur yang diuji, karena proses pengupasan afiksasi menghilangkan informasi gramatikal yang krusial bagi identifikasi entitas dalam bahasa yang bersifat aglutinasi [7]. Hal ini secara langsung relevan dengan tantangan morfologi Bahasa Indonesia yang dihadapi dalam penelitian ini dan menjadi landasan keputusan untuk tidak menerapkan *stemming* pada tahap pra-pemrosesan.

Pada kelompok studi pembelajaran lemah hibrida, penelitian [8] mengatasi permasalahan fundamental dalam pengembangan sistem deteksi otomatis, yaitu kelangkaan data berlabel di lingkungan media sosial berskala besar. Pendekatan yang diusulkan menggabungkan *transductive weakly supervised learning* berbasis SVM, dengan akurasi anotasi tertinggi mencapai 85 persen dibandingkan model pembelajaran mesin lainnya, dengan dua arsitektur pembelajaran mendalam secara bersamaan, yaitu Bi-LSTM dan Bi-GRU menggunakan representasi kata GloVe berukuran 100 dimensi. Hasil eksperimen menunjukkan bahwa Bi-LSTM mencapai akurasi 89,95 persen dan Bi-GRU mencapai 89,60 persen, keduanya secara signifikan melampaui LSTM searah sebesar 86,19 persen dan CNN sebesar 82,06 persen. Perbedaan struktural utama antara Bi-LSTM dan Bi-GRU terletak pada mekanisme gerbang internalnya [19], dimana Bi-LSTM menggunakan tiga gerbang dengan komponen memori *cell state* yang terpisah, sementara Bi-GRU menggunakan dua gerbang tanpa memori terpisah sehingga lebih ringan secara komputasi namun berpotensi kurang optimal untuk dependensi jarak sangat jauh.

Penelitian-penelitian yang menggunakan arsitektur Bi-LSTM dan variannya untuk deteksi berita palsu umumnya masih berfokus pada teks berbahasa Inggris dengan dataset berlabel yang sudah tersedia, atau menerapkan model NER secara terisolasi tanpa mengintegrasikannya ke dalam sistem pemeriksaan fakta yang proaktif. Meskipun penelitian [6] telah mendemonstrasikan penerapan model bidireksional pada teks berbahasa Indonesia, pendekatan tersebut masih terbatas pada tugas deteksi entitas statis, bukan sistem yang mampu secara mandiri mengambil keputusan verifikasi fakta terhadap berita nasional yang masuk secara berkelanjutan sebagaimana dicirikan dalam konsep sistem agentik yang otonom dan proaktif [18]. Di sisi lain, penelitian [8] menunjukkan bahwa pendekatan pembelajaran lemah dapat mengatasi keterbatasan data berlabel, namun diterapkan pada konteks media sosial berbahasa Inggris dan belum menyentuh ekosistem media digital Indonesia. Selain itu, belum ada penelitian terdahulu yang secara eksplisit mengukur kalibrasi kepercayaan model melalui metrik *hallucination rate* dan *groundedness* sebagaimana didefinisikan dalam kerangka *automated fact-checking* [17]. Oleh karena itu, penelitian ini mengusulkan penerapan algoritma Bi-LSTM dalam pengembangan sistem *fact-checking* agentik untuk berita nasional berbahasa Indonesia, yang memadukan kemampuan pemrosesan konteks bidireksional dari arsitektur Bi-LSTM [20] dengan

pendekatan yang proaktif dan otonom [18], bukan sekadar sistem klasifikasi reaktif untuk mendukung verifikasi fakta pada ekosistem berita digital Indonesia yang belum banyak dieksplorasi secara sistematis.

Pada akhirnya, kesenjangan utama yang masih tampak pada penelitian-penelitian terdahulu adalah belum adanya integrasi yang eksplisit antara klasifikasi berita palsu berbasis Bi-LSTM dengan kerangka agentik yang proaktif, kalibrasi kepercayaan model, serta mekanisme pembelajaran *human-in-the-loop* [21, 22]. Sebagian penelitian memang telah menunjukkan keunggulan Bi-LSTM untuk klasifikasi teks dan sebagian lain telah membahas pemeriksaan fakta otomatis, tetapi integrasi penuh antara aspek, evaluasi kepercayaan, dan mekanisme koreksi manusia masih belum diterangkan secara menyeluruh [4, 8, 17, 18]. Oleh karena itu, penelitian ini memposisikan diri sebagai pengembangan sistem pemeriksaan fakta berbasis sistem agentik yang melakukan lebih dari sekadar klasifikasi, dimana sistem ini turut mengukur kepercayaan prediksi, menyimpan bukti evaluasi, dan mendukung pembelajaran ulang dari umpan balik manusia [21, 22].

2.10 Kesenjangan Penelitian (Research Gap)

Berdasarkan tinjauan pada Penelitian Terdahulu, ditemukan tiga kelompok penelitian terdahulu yang relevan, masing-masing memiliki kontribusi penting namun juga menyisakan kesenjangan spesifik terhadap konteks penelitian ini.

1. **Kesenjangan pada studi Bi-LSTM dan *attention* untuk klasifikasi berita palsu.** Penelitian [4] menunjukkan performa klasifikasi yang sangat tinggi menggunakan *attention-based* Bi-LSTM, namun model tersebut beroperasi sebagai pengklasifikasi pasif berbahasa Inggris, tanpa mekanisme otonom untuk mencari bukti eksternal, mengukur ketidakpastian prediksinya sendiri, atau menyatakan tingkat kepercayaan yang dapat ditindaklanjuti pengguna.

2. **Kesenjangan pada studi NER berita palsu berbahasa Indonesia.** Penelitian [6] berhasil menerapkan arsitektur bidireksional (BiLSTM-CRF, BiGRU, BERT) pada teks berbahasa Indonesia, tetapi terbatas pada tugas deteksi entitas statis (*Named Entity Recognition*), bukan sistem yang mampu secara mandiri mengambil keputusan verifikasi *hoaks-atau-fakta* terhadap berita yang masuk secara berkelanjutan.

3. **Kesenjangan pada studi pembelajaran lemah hibrida.** Penelitian [8] mengatasi keterbatasan data berlabel melalui *weakly supervised learning* dikombinasikan dengan Bi-LSTM/Bi-GRU, namun diterapkan pada media sosial berbahasa Inggris dan tidak menyertakan mekanisme kalibrasi kepercayaan maupun metrik halusinasi yang relevan untuk sistem *fact-checking* otonom.

Ketiga kesenjangan tersebut dirangkum pada Tabel 2.2. Secara umum, belum ditemukan penelitian yang mengintegrasikan **keempat** aspek berikut secara bersamaan: (1) klasifikasi berbasis Bi-LSTM pada teks **berbahasa Indonesia**, (2) ditempatkan di dalam kerangka **sistem agentik yang proaktif dan otonom** (bukan pengklasifikasi reaktif), (3) dilengkapi **estimasi ketidakpastian** melalui *Monte Carlo Dropout* sebagai sinyal kendali proses verifikasi, dan (4) diukur menggunakan metrik **tingkat halusinasi** dan *groundedness* sebagaimana didefinisikan dalam kerangka *automated fact-checking* [17].

Tabel 2.2. Ringkasan kesenjangan penelitian terdahulu

Aspek	**Padalko et al. (2024)**	**Cahyo et al. (2025)**	**Syed et al. (2023)**	**Penelitian Ini**
Bahasa	Inggris	Indonesia	Inggris	Indonesia
Arsitektur bidireksional	Ya (Bi-LSTM+Attn)	Ya (BiLSTM-CRF/ BiGRU/BERT)	Ya (Bi-LSTM/ Bi-GRU)	Ya (Bi-LSTM)
Kerangka agentik proaktif/otonom	Tidak	Tidak	Tidak	Ya
Estimasi ketidakpastian	Tidak	Tidak	Tidak	Ya (MC Dropout)
Metrik halusinasi/ groundedness	Tidak	Tidak	Tidak	Ya

Dengan demikian, penelitian ini memposisikan diri untuk mengisi kesenjangan tersebut dengan mengembangkan sistem *fact-checking* agentik berbasis Bi-LSTM untuk berita berbahasa Indonesia, yang tidak hanya mengklasifikasikan teks, tetapi juga secara otonom mencari bukti pendukung, mengukur kepercayaan dan ketidakpastian prediksinya sendiri, serta menyediakan metrik evaluasi yang secara spesifik dirancang untuk konteks sistem agentik.

UNIVERSITAS MULTIMEDIA NUSANTARA