

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dalam dua dekade terakhir telah membawa perubahan signifikan terhadap cara masyarakat memperoleh dan mengonsumsi informasi. Internet dan perangkat digital memungkinkan akses cepat terhadap berita dari berbagai sumber, sehingga media massa online berkembang menjadi salah satu kanal informasi utama di Indonesia. Karakteristiknya yang *real time*, interaktif, dan mudah diakses menjadikan media massa online semakin dominan dalam penyebaran informasi kepada publik.

Peningkatan volume berita yang dipublikasikan setiap hari menimbulkan kebutuhan akan sistem otomatis yang mampu mengklasifikasikan konten secara akurat dan efisien. Dalam konteks ini, pendekatan klasifikasi multi-label menjadi semakin penting karena satu berita sering kali memuat lebih dari satu isu atau topik secara bersamaan. Fenomena keterkaitan lintas sektor misalnya antara politik dan hukum, atau ekonomi dan bisnis menjadikan pendekatan klasifikasi tunggal tidak lagi memadai untuk merepresentasikan kompleksitas berita media massa modern.

Berbagai model pembelajaran mesin telah digunakan dalam klasifikasi teks, seperti *Support Vector Machine* (SVM), *Convolutional Neural Network* (CNN), dan *Recurrent Neural Network* (RNN). Namun, sejumlah penelitian menunjukkan bahwa model-model tersebut masih memiliki keterbatasan dalam menangkap konteks panjang dan hubungan dua arah dalam teks (Kowsher et al., 2021; Nurul et al., 2025). *Bidirectional Long Short-Term Memory* (BiLSTM) menjadi salah satu model yang banyak digunakan karena kemampuannya membaca teks dari arah maju dan mundur sekaligus, sehingga dapat menangkap konteks yang lebih lengkap. Berbagai penelitian menunjukkan keunggulan BiLSTM dalam klasifikasi multi label, analisis sentimen, dan deteksi ujaran kebencian (Salsabila, 2023; Yonathan, 2023). Penerapan model hybrid BiLSTM juga terbukti mampu meningkatkan akurasi, terutama ketika dikombinasikan dengan representasi embedding modern

Dalam konteks bahasa Indonesia, perkembangan model transformer seperti IndoBERT memberikan representasi semantik yang jauh lebih kaya dibandingkan embedding tradisional seperti Word2Vec atau FastText. Penelitian menunjukkan bahwa IndoBERT unggul secara signifikan dalam ekstraksi fitur dan klasifikasi teks Indonesia (Alkahfi et al., 2024) bahkan ketika digunakan sebagai embedding sebelum masuk ke model LSTM atau BiLSTM (Iswavigra, 2025; Widodo & Nugraheni, 2024). Oleh karena itu, penelitian ini memanfaatkan embedding IndoBERT sebagai fitur masukan untuk model BiLSTM guna meningkatkan kemampuan klasifikasi multi label berita.

Berbeda dengan penelitian terdahulu yang umumnya membandingkan banyak jenis embedding atau banyak model sekaligus, penelitian ini fokus pada eksperimen konfigurasi internal embedding IndoBERT, khususnya variasi panjang token (*token length*) yang mempengaruhi kedalaman representasi teks, serta pengaturan nilai *Threshold* pada proses multi label classification, yang berperan menentukan sensitivitas model dalam memprediksi lebih dari satu label.

Berbagai penelitian sebelumnya telah menunjukkan efektivitas penggunaan Bi-LSTM dan BERT dalam tugas klasifikasi teks. Namun, masih sedikit penelitian yang secara khusus mengkombinasikan kedua pendekatan ini untuk klasifikasi multi-label berita berbahasa Indonesia. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi empiris dan praktis dalam pengembangan sistem analisis berita berbasis NLP di Indonesia.

Berdasarkan uraian tersebut, penelitian ini berfokus pada pembangunan sistem klasifikasi multi-label berita media massa online menggunakan metode Bi-LSTM dengan embedding IndoBERT. Diharapkan hasil penelitian ini dapat memberikan kontribusi nyata dalam bidang pengolahan teks berita media massa berbahasa Indonesia .

1.2 Perumusan Masalah

Bab ini menjelaskan tentang masalah yang diidentifikasi, metode usulan yang digunakan, serta pertanyaan penelitian.

1.2.1 Masalah

Berdasarkan uraian latar belakang, masalah yang dapat diidentifikasi dalam penelitian ini adalah:

- a. Banyaknya berita online menyebabkan sulitnya proses identifikasi kategori berita secara manual.
- b. Beberapa judul berita memiliki sifat singkat dan dapat mengandung lebih dari satu kategori, sehingga diperlukan klasifikasi multi-label.
- c. Diperlukan metode yang mampu menghasilkan representasi teks Bahasa Indonesia secara kontekstual.

1.2.2 Metode Usulan

Terkait menyelesaikan permasalahan tersebut, penelitian ini mengusulkan:

- a. Menggunakan IndoBERT sebagai metode *embedding* untuk mendapatkan representasi kontekstual dari judul berita .
- b. Menggunakan model Bi-LSTM sebagai metode klasifikasi multi-label.
- c. Melakukan evaluasi performa model menggunakan matriks seperti *Precision*, *recall*, *F1-score*, dan *hamming loss*.

1.2.3 Pertanyaan Penelitian

Berdasarkan masalah dan metode usulan, pertanyaan penelitian yang akan dijawab adalah:

- a. Bagaimana kinerja *embedding* IndoBERT sebagai representasi kontekstual pada data judul berita
- b. Bagaimana kinerja model Bi-LSTM pada klasifikasi multi label judul berita media massa online?.
- c. Bagaimana pengaruh variasi panjang token pada embedding IndoBERT serta variasi *Threshold* pada klasifikasi terhadap klasifikasi multi-label berita online?

1.3 Tujuan dan Manfaat

Bab ini menjelaskan tentang tujuan yang ingin dicapai dan manfaat yang diharapkan dari hasil penelitian ini

1.3.1 Tujuan

Tujuan penelitian ini adalah:

- a. Membangun model klasifikasi multi-label untuk judul berita online.
- b. Menggunakan gabungan *embedding* IndoBERT dan model Bi-LSTM pada klasifikasi multi-label.
- c. Melakukan evaluasi performa model menggunakan metrik pengukuran yang relevan.

1.3.2 Manfaat

Hasil penelitian ini diharapkan memberikan manfaat sebagai berikut:

- a. Pada bidang akademis, diharapkan penelitian ini bisa menjadi referensi dalam penelitian pengembangan sistem klasifikasi multi-label Bahasa Indonesia menggunakan model transformer dan Bi-LSTM.
- b. Pada bidang teknis, diharapkan penelitian ini bisa menyediakan sistem yang mampu mengenali kategori berita secara otomatis
- c. Penelitian ini diharapkan mampu membantu masyarakat, analis media, dan pelaku industri dalam mengklasifikasikan berita media massa online.

1.4 Batasan Penelitian

Agar penelitian terfokus dan dapat diselesaikan sesuai lingkup, batasan dalam penelitian ini adalah:

- a. Data yang digunakan berupa headline berita dari media massa online dalam Bahasa Indonesia.
- b. Data yang dibutuhkan berasal dari website tempo, liputan 6, dan kompas.
- c. Penelitian hanya akan melakukan klasifikasi multi-label
- d. Model Deep Learning yang digunakan adalah IndoBERT untuk *embedding* dan Bi-LSTM untuk klasifikasi.

1.5 Sistematika penulisan

Sistematika penulisan ditunjukkan untuk memberikan pembahasan secara ringkas atau uraian secara garis besar, sistematika yang telah disesuaikan dengan format penulisan skripsi teknik informatika Universitas Trunojoyo Madura sebagai berikut:

1. BAB I: Pendahuluan

Pada bab ini dijelaskan latar belakang, rumusan masalah, batasan, tujuan, manfaat, dan sistem penulisan yang akan diusulkan.

2. BAB II: Kajian Pustaka

Pada bab ini berkaitan dengan penelitian terdahulu sebagai tinjauan pustaka yang akan membantu untuk mendasari penelitian berupa teori yang terkait dengan klasifikasi teks, NLP, klasifikasi multi label, *embedding* IndoBERT, Bi-LSTM, dan penelitian terkait.

3. BAB III: Metode Usulan

Pada bab ini menerangkan alur dari penelitian, bagaimana data itu diproses melalui split data, melakukan embedding dengan IndoBERT, pelatihan model dengan Bi-LSTM, evaluasi model, prediksi kategori, serta deteksi *Trending* topik.

4. BAB IV: Hasil dan Pembahasan

Pada bab ini membahas hasil dan pembahasan penelitian yang meliputi implementasi sistem, implementasi model Bi-LSTM dengan embedding IndoBERT, hasil pengujian model menggunakan berbagai skenario pengujian, analisis hasil berdasarkan metrik evaluasi, serta pembahasan mengenai performa model

5. BAB V: Penutup

Bab ini berisi penutup yang terdiri atas kesimpulan penelitian berdasarkan hasil yang diperoleh serta saran yang dapat dijadikan acuan untuk pengembangan penelitian dan sistem pada masa mendatang.