

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kebutuhan akan informasi terus mengalami peningkatan seiring dengan pesatnya perkembangan teknologi informasi. Perubahan ini menandai peralihan masyarakat global dari era tradisional dan industrial menuju era informasi, di mana kemajuan teknologi telah membuka akses yang luas terhadap berbagai sumber informasi dan komunikasi secara cepat dan efisien (Kustiawan dkk., 2022). Namun di balik kemajuan ini, kemudahan akses informasi juga memunculkan fenomena *information overload*, yaitu kondisi ketika individu menerima arus informasi yang sangat besar dan cepat melebihi kapasitas kognitifnya untuk memproses. Akibatnya, pengguna sering mengalami kesulitan dalam mengambil keputusan, memahami inti informasi berita, bahkan dapat mengalami tekanan psikologis seperti kecemasan dan stres (Tsabitah dkk., 2020).

Pada tingkat lokal, kebutuhan akan informasi yang akurat dan kontekstual menjadi hal yang penting bagi masyarakat. Namun, informasi lokal sering tersebar di berbagai sumber yang tidak terkoordinasi, sehingga menyulitkan pengguna dalam menemukan berita yang akurat & relevan. Hal ini menunjukkan perlunya pengembangan sistem penyajian informasi berbasis teknologi yang mampu mengolah dan menampilkan berita lokal secara terstruktur dan mudah diakses (Ramadhani & Setiawan, 2023). Khusus di wilayah Lhokseumawe, sebagai kota strategis di Provinsi Aceh, kebutuhan akan informasi lokal yang akurat, cepat, dan terstruktur semakin mendesak. Sayangnya, sebagian besar sistem pencarian berita masih bersifat umum dan belum memiliki mekanisme khusus untuk menyaring informasi berdasarkan lokasi geografis tertentu. Akibatnya masyarakat, jurnalis, dan lembaga lokal kesulitan memperoleh informasi terkini yang relevan dengan konteks wilayah mereka, yang justru seringkali tersebar dan tidak terorganisir di berbagai situs berita daring.

Untuk itu teknologi *Artificial Intelligence* (AI), khususnya *Natural Language Processing* (NLP), menawarkan solusi yang inovatif untuk *problem* ini. Penelitian ini mengusung integrasi beberapa metode NLP untuk membangun sebuah sistem.

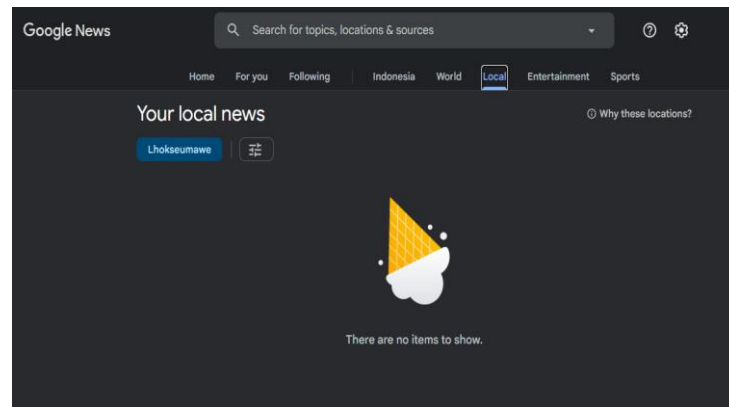
Sistem ini nantinya akan mengumpulkan data secara otomatis dari berbagai sumber berita di *google* melalui teknik *Web Scraping* lalu juga memprosesnya. Berdasarkan temuan penelitian yang dilakukan oleh Dewi & Hastjarjo(2025), saat ini jurnalis lebih cepat beradaptasi dengan teknologi AI karena mereka langsung merasakan peningkatan efisiensi dalam proses kerja, terutama dalam mempercepat pencarian serta pengolahan informasi.

Web scraping yang dipadukan dengan *Natural Language Processing* (NLP) menjadi salah satu cara yang efektif untuk mengambil dan mengolah informasi penting dari data teks yang tidak terstruktur (Pichiyan *et al.*, 2023). Langkah-langkah pemrosesannya nantinya meliputi penyaringan berita yang spesifik membahas hal yang berkaitan dengan inputan oleh pengguna dan berkaitan Kota Lhokseumawe. Lalu menggunakan teknik *Named Entity Recognition* (NER) untuk mengenali entitas lokasi Lhokseumawe, kemudian meringkasnya dengan *Extractive Summarization* untuk mengekstrak poin-poin penting dan yang terakhir menyusun ringkasan tersebut secara terstruktur berdasarkan tingkat kesesuaiannya dengan *query/input* pengguna dengan *Semantic Similarity Matching*.

Efektivitas setiap komponen teknis yang digunakan juga telah dibuktikan melalui berbagai penelitian terdahulu. Metode *Named Entity Recognition* (NER) berbasis NLP terbukti mampu mengenali entitas lokasi dengan tingkat akurasi yang tinggi, serta model berbasis *transformer* yang menjadi dasar bagi metode *semantic similarity matching* dalam penelitian ini dinilai efektif dalam mengukur kesamaan semantik antar teks. Dengan mengintegrasikan metode-metode tersebut ke dalam satu alur sistem, penelitian ini diharapkan menghasilkan solusi yang inovatif dan andal untuk penyajian informasi berita lokal.

Kontribusi penelitian ini terletak pada integrasi seluruh komponen pembentuk AI. Mulai dari pengumpulan data, penyaringan lokasi, peringkasan, hingga penyusunan teks semantik ke dalam satu *pipeline* yang utuh. Sistem pencarian berita yang ada (seperti *Google News* atau website artikel lainnya) umumnya tidak mendeteksi dan menandai dengan sempurna berdasarkan entitas lokasi spesifik di tingkat kota/kabupaten. Fitur yang terdapat biasanya berdasarkan kategori negara/kota yang luas. Penelitian ini mengisi celah tersebut dengan filter

lokasi otomatis dan berbasis AI. Pendekatan terpadu semacam ini belum banyak diterapkan karena peneliti kesulitan menemukan judul dan penelitian yang menyerupai, khususnya untuk menangani masalah informasi berita lokal di Indonesia dan untuk kota Lhokseumawe.



Gambar 1.1 Halaman *Google News* Menampilkan Tidak Ada Data dengan Filter Kota Lhokseumawe

Berdasarkan penelitian oleh Pradhana, Daniati, dan Muzaki (2023), dalam jurnal “*Penerapan Bi-LSTM Untuk Named Entity Recognition Pada Teks Bahasa Indonesia*”, sistem *Named Entity Recognition* (NER) berbahasa Indonesia masih menghadapi tantangan dalam mengenali entitas pada teks non-formal yang memiliki struktur dan kosakata tidak baku. Model-model sebelumnya menunjukkan keterbatasan dalam memahami konteks linguistik Bahasa Indonesia yang kompleks, sehingga menurunkan akurasi ekstraksi entitas. Meskipun penelitian tersebut berhasil meningkatkan performa NER menggunakan arsitektur Bi-LSTM dengan hasil akurasi yang tinggi, tantangan dalam penerapan NER untuk teks dinamis seperti berita daring, khususnya dalam konteks berita lokal tetap menjadi isu penting yang akan dikembangkan lebih lanjut di penelitian ini.

Singkatnya, penelitian kali ini bertujuan untuk mengembangkan sebuah sistem cerdas berbasis AI yang mampu:

1. Mengambil data berita dari berbagai sumber berita *online* di *google* berdasarkan *input* kata kunci oleh pengguna.
2. Memfilter berita yang secara spesifik membahas seputar topik di Kota Lhokseumawe menggunakan analisis entitas lokasi.

3. Merangkum informasi penting dari teks berita secara otomatis.
4. Menyajikan hasil ringkasan berita yang telah difilter berdasarkan lokasi dalam format terstruktur dan relevan dengan permintaan pengguna.

Penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem penyajian berita lokal berbasis AI, serta menjadi penerapan praktis teknologi NLP untuk konteks informasi daerah.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijabarkan diatas, maka rumusan masalah yang akan dipecahkan melalui penelitian ini yaitu:

1. Bagaimana cara merancang dan mengimplementasikan sistem *web scraping* yang dapat mengumpulkan artikel berita dari berbagai sumber daring berdasarkan kata kunci pengguna secara efektif?
2. Bagaimana menerapkan metode *Named Entity Recognition* (NER) untuk menyaring dan mengidentifikasi berita yang secara spesifik membahas wilayah Kota Lhokseumawe?
3. Bagaimana melakukan proses *extractive summarization* untuk mengekstrak intisari atau kalimat-kalimat kunci dari kumpulan berita yang telah diperoleh?
4. Bagaimana mengevaluasi efektivitas integrasi *semantic similarity* dalam menyajikan ringkasan berita yang akurat, terstruktur, dan relevan dengan kebutuhan informasi pengguna?

1.3 Tujuan Penelitian

Sesuai dengan rumusan masalah yang telah disebutkan diatas, tujuan dari penelitian ini adalah untuk:

1. Merancang dan membangun sistem berbasis *Artificial Intelligence* (AI) dengan fungsi utama melakukan *web scraping* yang handal untuk mengumpulkan data berita dari hasil pencarian *Google* secara otomatis berdasarkan kata kunci yang dimasukkan pengguna.
2. Mengimplementasikan metode *Named Entity Recognition* (NER) guna untuk mengidentifikasi dan memfilter berita yang berkaitan dengan Kota Lhokseumawe.

3. Menerapkan metode *extractive summarization* terhadap berita yang telah dikumpulkan untuk menampilkan inti atau bagian paling relevan dari isi berita.
4. Mengevaluasi kinerja sistem secara keseluruhan dengan fokus pada akurasi filter lokasi dan relevansi semantik (*semantic similarity*) dari ringkasan yang dihasilkan dengan kata kunci pengguna.

1.4 Manfaat Penelitian

Berikut adalah manfaat yang diharapkan penulis pada penelitian yang dilakukan:

1. Bagi peneliti :
 - a) Memberikan kontribusi dalam pengembangan aplikasi *Natural Language Processing* (NLP) untuk Bahasa Indonesia, khususnya dalam konteks identifikasi lokasi dan peringkasan teks.
 - b) Menjadi studi kasus penerapan metode NER, *extractive summarization*, dan *semantic similarity* secara terintegrasi dalam satu sistem pencarian dan pemrosesan berita berbasis AI.
2. Bagi pengguna, hasil penelitian diharapkan dapat:
 - a) Mempermudah masyarakat, jurnalis, dan instansi di wilayah Lhokseumawe dalam memperoleh ringkasan berita lokal yang relevan, cepat, dan terstruktur berdasarkan topik tertentu.
 - b) Menyediakan alat bantu untuk menyaring dan menganalisis berita secara otomatis dari berbagai sumber, tanpa perlu membaca seluruh isi artikel.
 - c) Meningkatkan efektivitas pencarian informasi lokal yang biasanya tersebar dan tidak terorganisir di berbagai situs berita daring.

1.5 Ruang Lingkup Dan Batasan Penelitian

Agar penulis dapat melakukan penelitian dengan fokus dan terarah, terdapat ruang lingkup dan batasan pada penelitian ini yaitu:

1. Sumber berita yang digunakan terfokus pada website berita daring yang dapat diakses secara publik (tanpa autentikasi) pada *platform google*, dan memiliki struktur HTML yang memungkinkan untuk dilakukan *scraping*

menggunakan *framework Selenium*.

2. Data yang diambil berupa data artikel yang telah di *scraping* oleh sistem. Sebagai keabsahan data artikel yang telah di ringkas, akan ditampilkan pula 3 link *website* sumber teratas di bawah tampilan hasil ringkasan.
3. Pemrosesan bahasa terbatas pada berita berbahasa Indonesia. Jika terdapat berita atau artikel berbahasa daerah (bahasa Aceh), artikel tersebut akan dilewatkan.
4. Lokasi target untuk filter berita difokuskan hanya pada “Kota Lhokseumawe” sebagai inti studi kasus wilayah lokal.
5. Metode peringkasan yang digunakan adalah *extractive summarization*, yaitu pemilihan kalimat-kalimat penting dari teks asli (meringkaskan), bukan membuat ulang kalimat baru (*abstractive*).
6. Antarmuka sistem akan dibangun dalam bentuk *web interface* menggunakan *framework Flask* berbasis bahasa pemrograman *Python*.
7. Evaluasi sistem akan mencakup aspek kualitatif dan kuantitatif. Secara kuantitatif, akurasi NER akan diukur menggunakan metrik *precision*, *recall*, dan *F1-score* dengan target nilai *F1-score* minimal sebesar 70% sebagai ambang batas validitas sistem. Secara kualitatif, akan dilakukan analisis terhadap kualitas ringkasan dan relevansi hasil akhir.
8. Untuk mempercepat proses pengambilan data, maka sistem akan berhenti melakukan *scraping* jika sudah ditemukan 3 data artikel valid.
9. Masukan berupa kata kunci yang terdiri atas minimal satu kata dan maksimal satu kalimat. Kata kunci dapat berbentuk kalimat pertanyaan maupun frasa yang merepresentasikan objek atau informasi yang ingin dicari.