

BAB I

PENDAHULUAN

1.1 Latar Belakang

Indonesia kini tengah mengalami perkembangan pesat dalam teknologi informasi. Kemudahan akses internet turut mendorong kemajuan di berbagai sektor. Perkembangan ini memungkinkan informasi dapat diakses dengan lebih cepat dan efisien. Menurut data dari Kementerian Komunikasi dan Informatika, jumlah pengguna internet di Indonesia telah melebihi 80 juta orang, menjadikannya sebagai negara dengan jumlah pengguna internet terbesar kedelapan di dunia.[1]. Lonjakan penggunaan internet tentunya berpengaruh terhadap meningkatnya pemanfaatan media online sebagai sumber informasi, terutama dalam penyebaran berita digital.

Berita telah menjadi sebuah lembaga yang digunakan untuk menyebarkan informasi terkini yang dibutuhkan masyarakat. Melalui berbagai instansi, berita disampaikan dengan menggunakan media seperti Media *Online*, Televisi, Surat Kabar, Radio, dan berbagai media lainnya. Secara umum Informasi yang disajikan di media terbagi ke dalam berbagai kategori, seperti politik, olahraga, ekonomi, kesehatan, dan lainnya. [2]. Dan dengan seiring berkembangnya teknologi informasi dan komunikasi masyarakat Lebih banyak memanfaatkan media daring, khususnya platform media online, sebagai sarana untuk memperoleh informasi setiap saat, karena dapat diakses kapanpun dan dimanapun [3] [4] [5]. Akibat penggunaan dan akses media online tersebut maka berdampak pula pada peningkatan jumlah data dan informasi khususnya data teks [6]. Secara umum, berita disajikan dalam berbagai kategori, seperti ekonomi, teknologi, olahraga, kesehatan, dan sebagainya [7].

Hingga saat ini, pengelompokan berita ke dalam berbagai kategori masih dilakukan secara manual. Dalam prosesnya, berita harus dibaca dan dianalisis untuk menentukan kategori yang sesuai. Salah satu tantangan dalam pengelompokan manual adalah kesamaan isi berita, yang memerlukan ketelitian tinggi agar klasifikasi tetap akurat. Untuk mengatasi masalah ini, diperlukan suatu alat atau sistem yang dapat mengklasifikasikan secara otomatis. Salah satu

pendekatan yang efektif untuk mengatasi masalah ini adalah dengan menggunakan teknik *text mining* [7].

Text Mining merupakan proses yang terstruktur untuk mengekstrak atau membentuk token teks yang tersusun dengan rapi serta menggali informasi penting dari data teks. [8]. *Text mining* adalah proses pemahaman teks yang dilakukan secara otomatis oleh komputer untuk mengekstrak informasi atau menemukan pola-pola penting dalam suatu teks [9]. *Text mining* mengacu pada proses umum mengekstraksi pola atau pengetahuan yang terkandung dalam dokumen teks yang tidak terstruktur. Teks telah menjadi salah satu bentuk data paling alami yang disimpan, sehingga bidang *text mining* dipercaya sebagai bidang lanjutan dari data *mining*. Fakta yang muncul dari penelitian. Penelitian yang telah dijalankan menunjukkan bahwa 80% informasi perusahaan disajikan dalam bentuk dokumen teks. *Text mining adalah bidang multidisiplin yang meliputi pengambilan informasi, analisis teks, ekstraksi data, serta proses pengelompokan* [10]. Metode klasifikasi *text mining* merupakan salah satu metode yang bisa diterapkan sebagai proses klasifikasi.. Penambangan teks merupakan inovasi penambangan data yang berusaha menemukan pola-pola menarik dalam kumpulan teks berukuran besar. Selain dimanfaatkan untuk klasifikasi, *text mining* juga digunakan dalam menyelesaikan masalah yang berkaitan dengan pengelompokan (*clustering*), ekstraksi informasi, dan lainnya. pengambilan informasi [2].

Klasifikasi teks secara khusus bekerja mengelompokkan dokumen teks berdasarkan kategori, dan dalam lingkup dataset berita, kategori umumnya dibedakan menjadi politik, ekonomi, militer, olahraga dan lain-lain. Metode statistik merupakan salah satu metode yang paling umum digunakan dalam klasifikasi emosi teks [11]. Salah satu penerapan klasifikasi teks yang dapat ditemukan adalah pengklasifikasian teks berita berdasarkan kategori tertentu [12]. Klasifikasi teks berita memiliki relevansi dan urgensi yang tinggi dalam konteks informasi lokal dengan pentingnya mempertimbangkan aspek temporal dalam klasifikasi tersebut [13]. LSTM merupakan metode yang lebih unggul dibandingkan metode tradisional, metode LSTM sangat ideal untuk digunakan dalam analisis sentimen. dan klasifikasi teks [4] [14].

Data dalam penelitian ini diperoleh dari beberapa portal berita terkemuka di Indonesia, salah satunya adalah Kompas.com.. Berita tersebut dikumpulkan menggunakan teknik *Web Scraping* untuk memastikan representasi data yang mencerminkan dinamika berita di media maya. Dataset ini kemudian diproses untuk mengidentifikasi kategori berita berdasarkan fitur-fitur linguistik yang relevan. Teknik *embedding* digunakan untuk representasi data teks, dan *LSTM* diterapkan untuk membangun model klasifikasi. Pentingnya penelitian mengenai klasifikasi kategori berita telah mendorong beberapa studi terdahulu akan tetapi masih menghadapi keterbatasan pada performa model.

Berdasarkan studi literatur sebelumnya mengenai klasifikasi berita menggunakan berbagai metode, ditemukan bahwa kategori berita yang paling sering digunakan mencakup *health*, nasional, dan *sports*. Hal ini terlihat dari beberapa penelitian yang memasukkan kategori tersebut dalam dataset mereka, menunjukkan bahwa topik ini memiliki relevansi tinggi dalam dunia pemberitaan. Kategori *health* (kesehatan) sering muncul karena tingginya minat masyarakat terhadap informasi kesehatan, terutama dalam era digital yang memungkinkan akses cepat ke berita medis dan kebijakan kesehatan. Kategori nasional juga mendominasi, mengingat perannya yang krusial dalam dinamika sosial dan pemerintahan. Sementara itu, olahraga tetap menjadi salah satu kategori populer karena memiliki audiens yang luas serta tingginya permintaan informasi, baik di tingkat nasional maupun internasional. Oleh karena itu, penelitian ini akan berfokus pada tiga kategori berita utama, yaitu *health*, nasional, dan *sports*, dalam proses klasifikasi teks berita berbahasa Indonesia. [7] [15] [16] [17] [18].

Penelitian ini mengkaji penerapan model Long Short-Term Memory (LSTM) dalam klasifikasi berita berbahasa Mandarin. Proses penelitian meliputi tahap preprocessing data, ekstraksi fitur, dan desain model deep learning. Data berita diolah menggunakan teknik segmentasi kata dengan alat seperti jieba, kemudian dilakukan penghapusan stop words dan analisis frekuensi kata untuk meningkatkan kualitas data. Selanjutnya, kata-kata diubah menjadi vektor berdimensi 200 menggunakan metode Word2Vec, yang kemudian digunakan sebagai input model. Model LSTM digunakan untuk mengekstraksi fitur kontekstual dari data teks, dengan output dari jaringan LSTM diambil dari layer

terakhir sebagai representasi fitur yang terintegrasi. Hasil eksperimen menunjukkan bahwa model LSTM mencapai F1-score sebesar 93,16%, yang menunjukkan kemampuan yang baik dalam menangkap makna dan konteks dari teks berbahasa Mandarin. Penerapan teknik regularisasi dropout dan optimisasi menggunakan Adam membantu mencegah overfitting dan meningkatkan performa model. Evaluasi kinerja dilakukan dengan metrik seperti akurasi, presisi, recall, dan F1-score, yang menunjukkan bahwa penggunaan model LSTM efektif dalam meningkatkan akurasi klasifikasi berita. Temuan ini menegaskan keunggulan model LSTM dalam pengolahan teks berurutan dan konteks, sehingga mampu meningkatkan performa klasifikasi secara signifikan [45].

Penelitian ini mengembangkan sistem klasifikasi teks berita terkini di Manggarai dengan menerapkan metode Long Short Term Memory (LSTM). Data yang digunakan terdiri dari 4000 berita yang dibagi menjadi empat kategori, yaitu ekonomi, kecelakaan, politik, dan pariwisata. Model LSTM menunjukkan performa terbaik dengan rasio data latih dan uji 3600:400, mencapai akurasi 88,75%, presisi 88,79%, recall 88,75%, dan skor F1 88,76%. Penelitian ini menekankan pentingnya konteks lokal dalam klasifikasi berita dan menunjukkan bahwa LSTM efektif dalam mengklasifikasikan teks berita secara efisien dan akurat. Selain itu, penelitian ini juga mencatat bahwa overfitting dapat diatasi dengan beberapa strategi, seperti penambahan data dan pengurangan kompleksitas model. Meskipun terdapat variasi dalam ukuran data latih, model tetap menunjukkan kinerja yang stabil. Penelitian ini menyarankan evaluasi lebih lanjut terhadap teknik pengembangan model dan penanganan ketidakseimbangan kelas untuk penelitian selanjutnya. Diharapkan hasil ini dapat memberikan kontribusi bagi pengembangan sistem klasifikasi berita yang lebih baik di masa depan [40].

Penelitian selanjutnya, menerapkan metode Artificial Neural Network (ANN) sebagai klasifikasi. Dataset yang digunakan terdiri dari 304 sampel yang diambil dari platform Kaggle. Proses penelitian melibatkan beberapa tahap, termasuk penambangan data, desain model, dan evaluasi kinerja model. Evaluasi dilakukan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model ANN memperoleh akurasi sebesar 73.77%, presisi 80.43%, recall 84.09%, serta F1-score 82.22% [41].

Penelitian selanjutnya, menggunakan algoritma Naive Bayes untuk mengklasifikasikan data. Metodologi yang diterapkan meliputi pengumpulan data melalui teknik scraping, diikuti dengan tahapan preprocessing mencakup mengonversi teks ke huruf kecil, menghapus stopwords, dan stemming untuk mengubah kata ke bentuk dasarnya. Dataset yang digunakan terdiri dari 300 artikel, dengan distribusi 250 artikel diklasifikasikan sebagai kategori tertentu dan 50 artikel sebagai kategori lainnya. Model Naive Bayes yang diterapkan dalam penelitian ini dievaluasi menggunakan teknik cross-validation. Hasil dari evaluasi menunjukkan bahwa model ini mencapai tingkat akurasi sebesar 86.33% [39].

Penelitian selanjutnya, menerapkan algoritma *Naïve Bayes* untuk mengklasifikasikan suatu dataset. Proses penelitian dimulai dengan pengolahan teks yang melibatkan penghapusan kata-kata umum dan proses stemming untuk menyederhanakan bentuk kata. Setelah itu, dilakukan pembobotan menerapkan teknik *TF-IDF* untuk menilai tingkat kepentingan setiap kata dalam dokumen. Dataset yang digunakan mencakup 235 data yang terbagi ke dalam beberapa kelas. Pengujian dilakukan melalui empat percobaan dengan menggunakan data uji yang beragam. Hasil penelitian ini mengungkapkan bahwa akurasi tertinggi yang diperoleh mencapai 81,48%, sedangkan akurasi terendah mencapai 79,63%. Penelitian ini menyoroti pentingnya seleksi data latih yang tepat dan tahap preprocessing dalam mempengaruhi akurasi sistem klasifikasi [42].

Penelitian selanjutnya dengan judul “Perancangan Aplikasi *Mobile* Menggunakan *Machine Learning* Untuk Menentukan Klasifikasi Kategori Berita”. Pada penelitian ini, dilakukan klasifikasi artikel berita menggunakan metode regresi logistik dalam sebuah aplikasi *mobile*. Penelitian ini memanfaatkan dataset sekitar 2000 judul berita dengan jumlah kategori berita mencakup berita terkini, olahraga, politik, ekonomi, sosial, kesehatan, dan lainnya. yang diterbitkan pada tahun 2020 dan menerapkan teknik *preprocessing* seperti tokenisasi dan penggunaan TF-IDF untuk vektorisasi, guna meningkatkan akurasi klasifikasi. Hasil eksperimen menunjukkan bahwa aplikasi ini efektif untuk klasifikasi berita, dengan tingkat akurasi di atas 85%. Penelitian ini bertujuan untuk membantu editor dan penulis berita dalam mengkategorikan berita secara efisien, mengingat meningkatnya volume artikel berita yang tersedia

secara digital. Penelitian selanjutnya akan fokus pada pengembangan dataset yang lebih *robust* dan peningkatan model untuk mencapai kinerja klasifikasi yang lebih baik [16].

Mengacu pada permasalahan yang telah dijelaskan, penelitian ini. mengusulkan **Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Metode *Long Short Term Memory (LSTM)*** . Dengan studi kasus kumpulan berita yang terdapat di *platform* Portal Media Massa Kompas.com

1.2 Rumusan Masalah

1.2.1 Permasalahan

Banyaknya berita yang beredar di berbagai platform digital membuat kita sulit untuk menentukan kategori berita secara otomatis. Berita-berita ini seringkali memiliki karakteristik yang beragam dan kompleksitas dalam struktur penulisannya, sehingga sulit untuk diklasifikasikan secara manual. Penentuan kategori berita memerlukan analisis yang mendalam terhadap konten, konteks, dan pola bahasa yang digunakan dalam teks berita. Kurangnya alat klasifikasi otomatis berbasis teknologi terkini, seperti *Long Short Term Memory (LSTM)*, menyebabkan proses klasifikasi menjadi lambat, tidak konsisten, dan kurang akurat.

1.2.2 Solusi Permasalahan

Penelitian ini mengusulkan penggunaan metode Long Short Term Memory (LSTM) diterapkan untuk mengklasifikasikan kategori berita.. secara otomatis. Dengan pendekatan ini, diharapkan klasifikasi berita dapat dilakukan lebih cepat, akurat, dan konsisten tanpa mengabaikan kompleksitas konten berita serta pola bahasa yang beragam. Model LSTM digunakan karena kemampuannya dalam menangkap urutan kata serta memahami konteks dalam teks. sekuensial, sehingga dapat memberikan performa yang optimal dalam mengolah data berita yang memiliki struktur dan tema yang berbeda-beda.

1.2.3 Pertanyaan Penelitian

1. Berapa akurasi model *Long Short Term Memory (LSTM)* dalam mengklasifikasikan kategori berita?
2. Bagaimana kombinasi parameter terhadap performa model Long Short Term Memory (LSTM) dalam klasifikasi kategori berita?

1.3 Batasan Masalah

Untuk menjaga fokus dan keberhasilan penelitian, batasan berikut akan diterapkan:

1. Penelitian ini memanfaatkan data dari Kompas.com dengan total 750 dataset.
<https://docs.google.com/spreadsheets/d/1DNxYNGAsMMY0u9hfTGv0BmWQXihMGD1L/edit?usp=sharing&ouid=111545919223301629132&rtpof=true&sd=true>
2. Dataset berita akan diproses menggunakan teknik web scraping atau menggunakan dataset publik yang tersedia secara daring
3. Dataset berita dari periode Februari 2025
4. Penelitian ini hanya berfokus pada analisis teks berita berbahasa Indonesia.
5. Kategori berita meliputi kategori *Health*, *Sports*, Nasional
6. Ekstraksi fitur menggunakan metode TF-IDF (Term Frequency - Inverse Document Frequency).
7. Model Long Short Term Memory (LSTM) digunakan sebagai metode utama dalam klasifikasi kategori berita.
8. Performa model akan dinilai menggunakan metrik akurasi, presisi, recall, dan F1-score.

1.4 Tujuan Dan Manfaat

1.4.1 Tujuan Penelitian

1. Mengetahui akurasi Long Short Term Memory (LSTM) dalam mengklasifikasikan kategori berita.

2. Mengetahui pengaruh kombinasi parameter terhadap akurasi model Long Short Term Memory (LSTM) dalam klasifikasi kategori berita.

1.4.2 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat dengan menghadirkan solusi otomatis dalam klasifikasi berita yang lebih cepat dan akurat, sehingga proses kategorisasi berita bagi penyedia informasi digital menjadi lebih efisien. Selain itu, penggunaan metode otomatis ini dapat mengurangi kesalahan klasifikasi yang sering terjadi dalam proses manual, sehingga meningkatkan konsistensi dalam pengelompokan berita. Penelitian ini juga berkontribusi dalam pengembangan wawasan mengenai implemberitaentasi metode *deep learning* dalam *text mining*, khususnya dalam klasifikasi teks berita, yang dapat menjadi referensi ntuk mendukung penelitian lebih lanjut dalam bidang yang serupa.

1.5 Sistematika Laporan

Sistematika penulisan proposal skripsi ini terdiri dari beberapa bab yaitu sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang, perumusan masalah yang meliputi permasalahan, metode yang diusulkan sebagai solusi permasalahan dan pertanyaan penelitian, batasan-batasan penelitian, kemudian selanjutnya terdapat tujuan dan manfaat penelitian, serta sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi penjelasan mengenai teori-teori pendukung dari berbagai sumber yang berhubungan dengan penelitian ini dan penelitian terdahulu yang membahas topik serupa.

BAB III METODOLOGI PENELITIAN

Bab ini berisi penjelasan tentang data penelitian, arsitektur sistem dan skenario uji.

BAB IV HASIL DAN PENELITIAN

Bab ini menjelaskan hasil implementasi sistem, mulai dari proses scraping, preprocessing data, ekstraksi fitur dengan TF-IDF, pelatihan model LSTM, hingga evaluasi menggunakan berbagai skenario uji. Juga disajikan tampilan aplikasi menggunakan Streamlit dan analisis performa sistem.

BAB V KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan dari hasil penelitian terkait performa model LSTM dalam klasifikasi teks berita, serta saran untuk pengembangan lebih lanjut pada aspek data, metode, dan sistem.

