Prediksi Genre Novel Berdasarkan Sinopsis Menggunakan NLP dan LSTM

Dosen Pengampu: Dr. Basuki Rahmat, S.Si., MT.

Disusun Oleh:

Nabil Anshari 21081010143

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"

JAWA TIMUR

2024

BAB I

PENDAHULUAN

1.1 Latar Belakang

Menurut Kamus Besar Bahasa Indonesia, novel adalah sebuah karya prosa fiksi yang ditulis secara naratif yang berisi rangkaian cerita tentang kehidupan seseorang dan orang-orang di sekitarnya. Novel juga memiliki daya tarik tersendiri yang bersifat universal. Beragam genre novel, seperti romantis, horor, fiksi ilmiah, petualangan, dan misteri, telah menjadi konsumsi masyarakat luas di berbagai belahan dunia. Hal ini mencerminkan bahwa novel bukan hanya sekadar bentuk hiburan, tetapi juga media yang mampu menyampaikan nilai-nilai budaya, sosial, dan psikologis kepada pembacanya. Dalam beberapa dekade terakhir, meningkatnya jumlah novel yang diterbitkan setiap tahunnya menimbulkan banyak tantangan baru, bukan hanya bagi pembaca novel itu sendiri, tetapi juga bagi pustakawan, penerbit, dan pengembang aplikasi. Tantangan ini berkisar dari pengelompokan dan pengelolaan novel di perpustakaan hingga pengembangan sistem rekomendasi yang mampu membantu pembaca menemukan novel yang sesuai dengan minat atau preferensi mereka. Maka dari itu, sebuah sistem baru untuk merekomendasikan genre yang sesuai dengan preferensi pembaca sangat diperlukan untuk mengatasi permasalahan ini.

Dalam beberapa kasus penelitian terdahulu, pengklasifikasian berdasarkan sinopsis dari suatu film telah dilakukan menggunakan metode machine learning, seperti TF-IDF (Term Frequency Inverse Document Frequency) dan Naïve Bayes Classifier. Metode TF-IDF digunakan untuk menghitung bobot dari setiap kata dalam dokumen, yang hasilnya kemudian dimanfaatkan oleh metode Naïve Bayes Classifier untuk mengklasifikasikan sinopsis ke dalam genre tertentu. Berdasarkan evaluasi menggunakan confusion matrix, dengan 600 data untuk pelatihan dan 200 data untuk pengujian, diperoleh akurasi sebesar 80,5%. (Rahmayanti et al., 2019). Keberhasilan ini menunjukkan bahwa pendekatan berbasis machine learning mampu memberikan hasil yang cukup baik untuk tugas klasifikasi teks berbasis sinopsis. Namun, pendekatan ini masih memiliki keterbatasan, terutama ketika harus menangani data dengan jumlah yang lebih besar atau sinopsis yang kompleks, sehingga diperlukan metode yang lebih canggih.

Sinopsis adalah ringkasan singkat dari isi novel yang bertujuan memberikan gambaran kepada pembaca tentang alur cerita tanpa mengungkapkan detail secara menyeluruh. Sinopsis menjadi elemen penting dalam menentukan genre sebuah novel karena menggambarkan tema, suasana, dan elemen cerita secara keseluruhan. Analisis sinopsis menggunakan Natural Language Processing (NLP) memberikan peluang besar untuk memahami pola teks dan mengklasifikasikan genre novel secara otomatis. NLP adalah bidang kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Dengan menggunakan metode NLP,

seperti TF-IDF atau Word Embeddings, sinopsis dapat direpresentasikan secara numerik sehingga mempermudah proses analisis. Word Embeddings, misalnya, memungkinkan representasi kata dalam bentuk vektor yang mempertimbangkan konteksnya dalam kalimat, sehingga memberikan makna yang lebih mendalam dibandingkan metode berbasis frekuensi seperti TF-IDF.

Salah satu metode deep learning yang efektif untuk analisis teks adalah Long Short-Term Memory (LSTM). LSTM, sebagai jenis dari Recurrent Neural Network (RNN), dirancang untuk menangkap hubungan jangka panjang dalam data sekuensial seperti teks. Keunggulan LSTM terletak pada kemampuannya untuk menangani masalah vanishing gradient yang sering ditemui pada RNN biasa, sehingga memungkinkan model untuk memahami konteks yang lebih kompleks dalam teks. Dalam klasifikasi genre novel, LSTM mampu memanfaatkan pola dalam sinopsis untuk memberikan prediksi yang lebih akurat dibandingkan metode konvensional. Selain itu, keunggulan LSTM dalam menangkap urutan data memberikan peluang untuk memahami hubungan antara elemen-elemen cerita dalam sinopsis yang mungkin relevan dalam menentukan genre. (Yudi Widhiyasana et al., 2021)

Dalam dunia nyata, sebuah novel sering kali tidak hanya tergolong dalam satu genre, melainkan memiliki elemen dari beberapa genre sekaligus, seperti novel romantis yang juga mengandung elemen misteri atau petualangan. Hal ini mencerminkan bahwa genre novel sering kali bersifat multidimensional dan kompleks, sehingga pendekatan klasifikasi konvensional yang hanya memberikan satu label tidak mampu mencerminkan keragaman genre secara akurat. Oleh karena itu, metode multilabel classification menjadi sangat relevan dalam penelitian ini. Multilabel classification memungkinkan model untuk memprediksi lebih dari satu genre secara bersamaan, memberikan representasi yang lebih realistis terhadap keragaman genre dalam novel. Dalam konteks ini, sistem multilabel classification tidak hanya membantu dalam memberikan hasil yang lebih akurat, tetapi juga meningkatkan relevansi rekomendasi bagi pembaca.(Arham, 2018)

Penggunaan multilabel classification memberikan beberapa manfaat signifikan. Sistem ini mampu mengenali bahwa sebuah novel dapat memiliki kombinasi genre yang beragam, sehingga memberikan hasil yang lebih akurat dan relevan bagi pembaca. Selain itu, pendekatan ini meningkatkan pengalaman pengguna karena rekomendasi novel yang dihasilkan dapat disesuaikan dengan preferensi pembaca secara lebih spesifik. Sistem ini juga menjadi lebih efisien dalam memproses data sinopsis novel yang kompleks, sekaligus membuka peluang untuk analisis lebih lanjut, seperti mengidentifikasi tren dalam genre novel atau mengelompokkan novel berdasarkan kombinasi genre tertentu. Misalnya, sistem ini dapat membantu penerbit untuk memahami kebutuhan pasar dan menghasilkan karya yang sesuai dengan preferensi pembaca di berbagai segmen.
Dengan menggabungkan metode NLP modern, seperti TF-IDF atau Word Embeddings, dengan LSTM dan pendekatan multilabel classification, diharapkan sistem yang dikembangkan mampu mengklasifikasikan genre novel berdasarkan

sinopsis secara lebih akurat, efisien, dan sesuai dengan kebutuhan pengguna. Sistem ini tidak hanya memberikan manfaat dalam dunia literatur, tetapi juga memiliki potensi untuk diterapkan dalam berbagai bidang lain yang memanfaatkan analisis teks, seperti sistem rekomendasi di platform e-commerce, analisis sentimen di media sosial, dan pengelolaan data di perpustakaan digital. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam pengembangan teknologi berbasis kecerdasan buatan untuk analisis teks yang kompleks.

1.2 Rumusan Masalah

1. Bagaimana memanfaatkan teknologi NLP modern (seperti TF-IDF atau Word Embeddings) untuk merepresentasikan sinopsis buku/novel secara numerik yang efektif?
2. Bagaimana kinerja metode deep learning seperti LSTM pada klasifikasi novel/buku menggunakan sinopsis?
3. Berapa tingkat akurasi dari model LSTM untuk prediksi genre novel berdasarkan sinopsis?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah untuk menggunakan teknologi NLP modern dalam merepresentasikan sinopsis novel secara numerik, mengevaluasi performa metode deep learning, khususnya LSTM, dalam klasifikasi genre novel berdasarkan sinopsis, serta mengukur tingkat akurasi model LSTM dalam memprediksi genre novel dari sinopsis yang tersedia.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini adalah memberikan kontribusi bagi pengembangan teknologi analisis teks berbasis deep learning, khususnya bagi peneliti yang tertarik pada pengolahan bahasa alami. Selain itu, penelitian ini dapat diaplikasikan oleh pengembang aplikasi dalam menciptakan sistem rekomendasi yang lebih efektif untuk pengguna. Bagi masyarakat umum, hasil penelitian ini diharapkan dapat memudahkan pembaca dalam memilih novel yang sesuai dengan preferensi genre mereka.

1.5 Sistematika Penulisan Tugas Akhir

Sistematika dari penulisan Tugas Akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini berisi tentang gambaran umum dari penulisan Tugas Akhir ini yang meliputi latar belakang masalah, perumusan masalah, batasan masalah, tujuan, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada bab ini berisi tentang teori-teori keilmuan untuk menunjang proses penelitian. Mencakup teori-teori keilmuan dan penelitian yang pernah dilakukan.

BAB III METODOLOGI PENELITIAN

Pada bab ini dibahas tentang langkah – langkah dan metode yang digunakan menyelesaikan Tugas Akhir ini. untuk

BAB IV PERANCANGAN DAN IMPLEMENTASI

Pada bab ini akan dijelaskan tentang model dan desain dari sistem yang akan dibentuk. Hal -hal tersebut meliputi visualisasi data, tahap pra proses data, transformasi data dengan TF-IDF atau Word Embeddings, pembuatan model LSTM sebagai acuan dalam implementasi sistem.

BAB V UJI COBA DAN EVALUASI SISTEM Pada bab ini akan dibahas tentang pengujian sistem yang telah terimplementasi dengan melakukan proses verifikasi dan validasi beserta pengujian kinerja dari sistem yang dibuat.

BAB VI KESIMPULAN DAN SARAN Bab ini berisi kesimpulan yang diperoleh dari pembahasan masalah sebelumnya serta saran yang diberikan selanjutnya.

BAB II

TINJAUAN PUSTAKA

2.1 Novel

Novel adalah salah satu bentuk karya sastra berbentuk prosa fiksi yang disusun secara naratif dan berisi rangkaian cerita. Menurut Kamus Besar Bahasa Indonesia (KBBI), novel mengisahkan kehidupan seseorang atau lebih dengan penekanan pada karakter, konflik, dan perkembangan psikologisnya. Novel memiliki daya tarik universal karena kemampuannya mengangkat tema yang beragam dan relevan dengan pengalaman pembaca, mulai dari cinta, petualangan, hingga misteri. Sebagai karya seni, novel juga menjadi cerminan budaya dan kondisi sosial masyarakat pada masanya.

2.2 Genre

Genre dalam novel adalah pengelompokan cerita berdasarkan tema, gaya penulisan, atau elemen tertentu dalam narasi. Genre membantu pembaca untuk memilih novel sesuai dengan preferensinya. Beberapa genre populer meliputi romantis, horor, fiksi ilmiah, petualangan, dan misteri. Selain itu, kombinasi antar-genre, seperti romantis-misteri atau fiksi ilmiah-petualangan, menciptakan dimensi baru yang menarik dalam dunia sastra. Keberagaman genre ini tidak hanya memberikan pilihan yang luas bagi pembaca tetapi juga menjadi tantangan dalam klasifikasi otomatis novel.

2.3 Sinopsis

Sinopsis adalah ringkasan singkat dari sebuah cerita yang bertujuan memberikan gambaran kepada pembaca mengenai isi buku tanpa mengungkapkan seluruh detail alur cerita. Sinopsis sering kali digunakan sebagai alat promosi dan panduan awal bagi pembaca untuk memutuskan apakah cerita tersebut sesuai dengan minat mereka. Dalam penelitian ini, sinopsis menjadi sumber data utama untuk analisis dan klasifikasi genre novel secara otomatis. Sinopsis yang baik mampu merepresentasikan inti cerita secara padat dan jelas.

2.4 Dataset

Dataset adalah kumpulan data yang digunakan untuk melatih dan menguji model dalam penelitian ini. Dataset yang digunakan terdiri dari sinopsis berbagai novel dengan beragam genre. Dataset ini dapat diperoleh dari berbagai sumber, seperti platform online, perpustakaan digital, atau penyedia dataset terbuka. Kualitas dataset sangat penting untuk memastikan model dapat mengenali pola dan menghasilkan prediksi yang akurat. Oleh karena itu, dataset yang digunakan harus mencakup representasi genre yang seimbang dan bervariasi.

2.5 Preprocessing Data

Preprocessing data adalah langkah awal yang krusial dalam analisis teks. Tahapan ini bertujuan untuk mengubah data mentah menjadi format yang lebih terstruktur dan siap digunakan oleh model machine learning. Preprocessing melibatkan beberapa proses seperti pembersihan data, tokenisasi, dan normalisasi teks. Langkah ini memastikan bahwa data yang digunakan memiliki konsistensi dan relevansi yang tinggi.

2.6 Data cleaning

Preprocessing data adalah langkah awal yang krusial dalam analisis teks. Tahapan ini bertujuan untuk mengubah data mentah menjadi format yang lebih terstruktur dan siap digunakan oleh model machine learning. Preprocessing melibatkan beberapa proses seperti pembersihan data, tokenisasi, dan normalisasi teks. Langkah ini memastikan bahwa data yang digunakan memiliki konsistensi dan relevansi yang tinggi.

2.7 Tokenization

Tokenization adalah proses memecah teks menjadi unit-unit kecil yang disebut token. Token biasanya berupa kata, frasa, atau karakter tergantung pada kebutuhan analisis. Tokenization memungkinkan model untuk memahami struktur teks dengan lebih baik dan menjadi dasar untuk langkah analisis selanjutnya.

2.8 Stemming/lemmatization

Stemming dan lemmatization adalah teknik untuk mengubah kata menjadi bentuk dasarnya. Stemming menggunakan pendekatan yang lebih sederhana dengan menghapus akhiran kata, sedangkan lemmatization lebih kompleks dengan mempertimbangkan konteks gramatikal kata. Kedua teknik ini membantu mengurangi variasi kata yang tidak diperlukan sehingga meningkatkan efisiensi analisis teks.

2.9 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode statistik yang digunakan untuk menghitung pentingnya suatu kata dalam dokumen relatif terhadap kumpulan dokumen. TF-IDF memberikan bobot lebih tinggi pada kata-kata yang sering muncul dalam dokumen tertentu tetapi jarang muncul di dokumen lainnya. Teknik ini sangat efektif dalam merepresentasikan teks dalam format numerik yang dapat digunakan oleh algoritma machine learning.(Rahmayanti et al., 2019)

2.10 Word Embeddings

Word embeddings adalah teknik representasi kata dalam bentuk vektor berdimensi tetap. Berbeda dengan TF-IDF yang bersifat frekuensi, word embeddings seperti Word2Vec, GloVe, atau FastText mempertimbangkan hubungan semantik antar kata dalam konteks tertentu. Teknik ini memungkinkan model untuk menangkap makna kata yang lebih mendalam dan memahami hubungan antara kata-kata dalam teks.

2.11 Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah cabang kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP mencakup berbagai teknik untuk memproses, menganalisis, dan memahami teks, seperti tokenization, stemming, dan word embeddings. Dalam penelitian ini, NLP digunakan untuk mengolah sinopsis novel sehingga dapat direpresentasikan dalam bentuk yang dapat dipahami oleh model machine learning.

2.12 LSTM

Long Short-Term Memory (LSTM) adalah salah satu varian dari Recurrent Neural Network (RNN) yang dirancang untuk menangkap hubungan jangka panjang dalam data sekuensial. LSTM memiliki mekanisme unik yang memungkinkan model untuk menyimpan dan menghapus informasi sesuai kebutuhan, sehingga dapat mengatasi masalah vanishing gradient yang sering ditemui pada RNN konvensional. Dalam tugas klasifikasi genre novel, LSTM digunakan untuk menganalisis pola dalam sinopsis dan memberikan prediksi genre yang lebih akurat.(Arham, 2018)

2.13 Multilabel Classification

Multilabel classification adalah pendekatan dalam machine learning di mana sebuah data dapat diklasifikasikan ke dalam lebih dari satu kategori atau label secara bersamaan. Dalam konteks klasifikasi genre novel, pendekatan ini relevan karena banyak novel yang memiliki elemen dari berbagai genre sekaligus, seperti novel romantis yang juga mengandung unsur misteri atau petualangan. Multilabel classification tidak hanya memberikan representasi yang lebih realistis dari genre novel tetapi juga meningkatkan akurasi dan relevansi sistem rekomendasi. Metode ini memungkinkan sistem untuk memahami kompleksitas data teks secara lebih mendalam dan menyediakan hasil yang lebih spesifik bagi pembaca. Dengan pendekatan ini, pengguna dapat menerima rekomendasi novel yang lebih sesuai dengan preferensi mereka berdasarkan kombinasi genre yang diminati.

2.14 Model Deep Learning

Dalam penelitian ini, model deep learning digunakan untuk menangkap pola kompleks dalam data teks sinopsis. Kombinasi metode NLP, seperti TF-IDF atau word embeddings, dengan arsitektur LSTM memberikan dasar yang kuat untuk membangun model klasifikasi genre. Model ini dirancang untuk mempelajari hubungan kontekstual antar kata dalam sinopsis sehingga mampu menghasilkan prediksi genre yang lebih akurat. Dengan mengintegrasikan multilabel classification, model dapat memberikan

hasil yang mencakup lebih dari satu genre per novel, yang mencerminkan sifat alami dari banyak karya sastra.

BAB III

METODE PENELITIAN

Bab ini berisi pendekatan dan langkah-langkah sistematis yang akan digunakan dalam penelitian ini agar bisa mencapai hasil dan tujuan yang diharapkan. Metodologi penelitian mencakup desain penelitian, sumber data, tahapan preprocessing data, metode yang digunakan, dan evaluasi model.

3.1 Pengumpulan Data

Dataset yang akan digunakan dalam penelitian ini terdiri dari beberapa synopsis atau dalam beberapa kasus terdiri dari rangkuman isi novel itu sendiri, untuk dataset yang ditemukan berupa data banyak buku yang dicari dari situs Kaggle. Dan berdasarkan dataset ini setiap sinopsis dilabeli dengan satu atau lebih genre berdasarkan informasi yang tersedia. Dataset ini dibagi menjadi tiga bagian:

- Data Latih: Digunakan untuk melatih model.
- Data Validasi: Digunakan untuk mengoptimalkan hyperparameter model.
- Data Uji: Digunakan untuk mengukur kinerja model.

3.2 Preprocessing Data

Preprocessing data adalah tahap penting untuk memastikan data siap digunakan oleh model. Tahapan preprocessing meliputi:

- **Data Cleaning**: Menghapus tanda baca, angka, dan simbol yang tidak relevan.
- **Tokenization**: Memecah teks menjadi unit kata atau token.
- **Stemming/Lemmatization**: Mengubah kata menjadi bentuk dasar untuk mengurangi variasi kata yang tidak perlu.
- **Stopword Removal**: Menghapus kata-kata umum yang tidak memberikan informasi signifikan (contoh: "dan", "yang", "di").
- **Vectorization**: Merepresentasikan teks dalam bentuk numerik menggunakan metode TF-IDF atau Word Embeddings.

3.3 Natural Language Processing (NLP)

NLP digunakan untuk mengolah sinopsis novel menjadi representasi numerik yang dapat digunakan oleh model deep learning. Metode TF-IDF dan Word Embeddings digunakan untuk merepresentasikan teks secara efektif.

3.4 Model Deep Learning

Model yang digunakan dalam penelitian ini adalah Long Short-Term Memory (LSTM). LSTM dipilih karena kemampuannya dalam menangkap hubungan jangka panjang dalam data sekuensial. Struktur model meliputi:

1. **Input Layer**: Menerima representasi numerik teks.
2. **Embedding Layer**: Menerjemahkan kata ke dalam representasi vektor untuk konteks semantik.
3. **LSTM Layer**: Menangkap pola dan hubungan dalam data teks.
4. **Dense Layer**: Menghasilkan output berupa probabilitas untuk setiap genre.

3.5 **Multilabel Classification**

Karena sebuah novel dapat memiliki lebih dari satu genre, penelitian ini menggunakan pendekatan multilabel classification. Model dirancang untuk menghasilkan beberapa label genre dengan probabilitas tertentu untuk setiap sinopsis.

3.6 Evaluasi Model

Model dievaluasi menggunakan metrik berikut:

- **Precision, Recall, dan F1-Score**: Mengukur performa model untuk setiap label.
- **Hamming Loss**: Mengukur jumlah kesalahan label dalam prediksi model.
- **Accuracy**: Mengukur proporsi prediksi yang benar secara keseluruhan.

3.7 Alat dan Perangkat Lunak

Penelitian ini menggunakan alat dan perangkat lunak berikut:

- **Bahasa Pemrograman**: Python
- **Perpustakaan**: TensorFlow, Keras, Scikit-learn, dan NLTK
- **Lingkungan Pengembangan**: Google Colab atau Jupyter Notebook

DAFTAR PUSTAKA

- Arham, A. Z. (2018). *Klasifikasi Ulasan Buku Menggunakan Algoritma Convolutional Neural Network – Long Short Term Memory*. https://repository.its.ac.id/51134/1/06111340000118-Undergraduate_Theses.pdf
- Rahmayanti, V., Basuki, S., & Hilman, H. (2019). Klasifikasi sinopsis novel menggunakan metode naïve bayes classifier. *Jurnal Repositor*, *1*(2), 125. https://doi.org/10.22219/repositor.v1i2.799
- Yudi Widhiyasana, Transmissia Semiawan, Ilham Gibran Achmad Mudzakir, & Muhammad Randi Noor. (2021). Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(4), 354–361. https://doi.org/10.22146/jnteti.v10i4.2438