P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

PREDIKSI POPULARITAS NOVEL BERBASIS FITUR-FITUR TEKS MENGGUNAKAN METODE RANDOM FOREST

Nadya Elfareta Azarin[1)], Rizal Adi Saputra[2)], Subardin[3)]

[1,2,3] *Teknik Informatika, Universitas Halu Oleo*

email: [1]nadyaelfaretaazarin@gmail.com , [2]rizaladisaputra@uho.ac.id

ARTICLE INFO

Article History:
Recieved : 16 Februari 2024
Accepted : 27 Mei 2024
Published : 14 Mei 2024

Keywords:
Prediksi popularitas
Novel
Fitur teks
Random Forest
Pembelajaran mesin

IEEE style in citing this article:

N. E. Azarin, R. A. Saputra, S. Subardin "Prediksi Popularitas Novel Berbasis Fitur Fitur Teks Menggunakan Metode Random Forest", *jurnal.ilmiah.informatika*, vol. 9, no. 1, Jun. 2024.

ABSTRACT

In today's digital era, a novel's popularity is often measured by reader response and sales. This research aims to develop a novel popularity prediction model based on text features to provide insights to authors and publishers about the factors that influence reader acceptance. The method used in this research is Random Forest, a machine learning algorithm that can handle classification and regression well. The main goal of this research is to develop a predictive model that can identify key factors that contribute to the popularity of novels. The proposed method integrates text features, such as keyword extraction and sentiment analysis, in a Random Forest framework to predict popularity with high accuracy. The dataset used consists of various novel information, including title, genre, number of pages, and text features such as summary or description. Data is preprocessed to address issues such as missing values and duplicates. Feature extraction is carried out by applying tokenization, stemming, and converting text into TF-IDF vectors. A Random Forest model was built incorporating these features, and the model parameters were optimized through a cross-validation process. The dataset used consists of various novel information, including title, genre, number of pages, and text features such as summary or description. Data is preprocessed to address issues such as missing values and duplicates. Feature extraction is carried out by applying tokenization, stemming, and converting text into TF-IDF vectors. A Random Forest model was built incorporating these features, and the model parameters were optimized through a cross-validation process. The experimental results show that the Random Forest model is able to predict the popularity of novels with a satisfactory level of accuracy. Text features, such as keyword frequency and sentiment analysis, proved significant in their contribution to the predictive ability of the model. These findings provide valuable insight to authors and publishers in understanding reader preferences and the potential success of a novel.

P-ISSN : 2549-7480
N. E. Azarin dkk/ JIMI 9 (1) pp. 57-62 E-ISSN : 2549-6301

1. PENDAHULUAN

Dalam era digital saat ini, industri penerbitan menghadapi tantangan baru dalam memahami preferensi pembaca dan memprediksi popularitas novel. Kemajuan teknologi telah mengubah cara konsumen berinteraksi dengan karya sastra, dengan perpindahan ke *platform digital* dan peningkatan *aksesibilitas* informasi. Oleh karena itu, kebutuhan untuk mengembangkan metode prediksi yang akurat dan efektif untuk popularitas novel menjadi semakin mendesak.

Penelitian ini fokus pada penggunaan fitur-fitur teks sebagai indikator utama dalam memprediksi popularitas novel.

Tradisionalnya, penilaian popularitas novel seringkali mengandalkan aspek subjektif dan pengalaman manusia. Namun, dengan perkembangan teknologi, kita sekarang memiliki akses terhadap data besar yang mencakup informasi tentang preferensi pembaca, respon pasar, dan karakteristik karya sastra. Oleh karena itu, penggunaan teknik pembelajaran mesin, seperti *Random Forest*, dapat memberikan kontribusi signifikan untuk memahami dan memprediksi popularitas novel.

Random Forest adalah algoritma pembelajaran mesin *ensemble* yang mampu mengatasi masalah *overfitting* dan menghasilkan prediksi yang stabil. Dengan kombinasi dari banyak pohon keputusan *Random Forest* dapat menangani kerumitan data dan mempertimbangkan hubungan yang kompleks antara fitur-fitur teks dan popularitas novel.

Tujuan utama penelitian ini adalah untuk mengembangkan model prediktif yang mampu mengidentifikasi faktor-faktor kritis yang mempengaruhi popularitas suatu novel. Dengan fokus pada fitur-fitur teks, seperti ekstraksi kata-kata kunci dan analisis sentimen, kami bertujuan untuk menggali wawasan yang mendalam tentang hubungan antara konten teks dan penerimaan pembaca. Keberhasilan penelitian ini diharapkan dapat memberikan kontribusi pada pemahaman praktis dan strategis di industri penerbitan.

Dengan menggabungkan metode pembelajaran mesin dan fitur teks, penelitian ini diharapkan dapat memberikan pandangan baru tentang elemen-elemen yang membuat suatu novel populer. Hasil dari penelitian ini dapat bermanfaat tidak hanya bagi penulis dan penerbit dalam mengembangkan karya yang lebih sukses tetapi juga dapat membuka peluang baru untuk strategi pemasaran yang lebih efektif.

Dengan mengeksplorasi keterkaitan antara fitur-fitur teks dan popularitas novel menggunakan metode *Random Forest*, penelitian ini diharapkan dapat memberikan kontribusi pada pemahaman mendalam tentang elemen-elemen yang mempengaruhi kesuksesan sebuah karya fiksi.

Implikasi praktis dari temuan ini dapat membantu penerbit, penulis, dan pemasar untuk mengoptimalkan strategi mereka dalam merancang dan mempromosikan novel yang memiliki potensi besar untuk menjadi populer di tengah pembaca.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *Random Forest*, sebuah algoritma pembelajaran mesin yang memiliki kemampuan baik dalam klasifikasi maupun regresi. Dataset yang digunakan mencakup berbagai informasi tentang novel, seperti judul, genre, jumlah halaman, dan fitur teks seperti ringkasan atau deskripsi. Proses pemrosesan data dan ekstraksi fitur, termasuk tokenisasi dan konversi teks menjadi vektor *TF-IDF*,

DOI : https://doi.org/10.35316/jimi.v9i1.57-62

P-ISSN: 2549-7480
N. E. Azarin dkk/ JIMI 9 (1) pp. 57-62 E-ISSN: 2549-6301

diterapkan untuk mendapatkan representasi yang optimal dari teks novel.

Pengumpulan Data, Kumpulan data terdiri dari novel-novel digital yang mencakup informasi teks dan metadata seperti judul, pengarang, dan parameter popularitas yang diukur (contohnya, jumlah unduhan, penjualan, atau ulasan).

Data teks subyek diperiksa untuk menghilangkan anomali dan memprosesnya ke dalam Hal-hal yang sudah diketahui oleh pelaku riset dalam lingkup riset tertentu tidak perlu lagi dituliskan, demikian pula perlengkapan dan peralatan umum yang digunakan.

Mohon setiap metode diberikan bagan atau tahapan apa saja yang akan dilakukan, baik dari pengumpulan data, hingga tolak ukur untuk mengetahui keberhasilan penelitian yang telah dilakukan.

Dan memahami kontribusi fitur-fitur teks terhadap popularitas novel. Dengan demikian, diharapkan dapat memberikan wawasan yang berharga bagi industri.

Ekstraksi Fitur: identifikasi fitur-fitur teks yang dapat digunakan untuk mendeskripsikan karakteristik novel. Beberapa metode ekstraksi fitur yang umum melibatkan penggunaan *TF-IDF (Term Frequency-Inverse Document Frequency)* atau *embedding seperti Word Embeddings (Word2Vec, GloVe)* untuk mewakili kata-kata dalam vektor *numerik*.

Gunakan teknik-teknik ekstraksi fitur lainnya seperti karakter *n-gram* atau *bag-of-words* untuk mendapatkan representasi teks yang lebih kaya.

Pemilihan Fitur**:** Lakukan analisis untuk memilih fitur-fitur yang paling relevan dan bermanfaat dalam memprediksi sifat atau karakteristik novel. Pemilihan fitur dapat melibatkan metode seperti analisis korelasi atau *feature importance*.

Pemodelan Prediktif: Pilih model prediktif yang sesuai dengan tipe masalah. Model-model seperti *Support Vector Machines (SVM), Decision Trees, Random Forest,* atau algoritma *Machine Learning* terkini seperti *deep learning* dapat digunakan.

Pisahkan dataset menjadi set pelatihan dan set pengujian untuk menghindari *overfitting* dan mengukur kinerja model secara objektif.

Validasi Model: Validasi model menggunakan metode seperti validasi silang *(cross-validation)* untuk memastikan bahwa model dapat memberikan prediksi yang baik pada data yang tidak terlihat sebelumnya.

Evaluasi Model Gunakan metrik evaluasi yang sesuai, seperti akurasi, *presisi*, *recall*, atau *F1-score*, untuk mengukur sejauh mana model dapat memprediksi novel berdasarkan fitur-fitur teks.

Optimasi dan Penyetelan Parameter Lakukan optimasi model dan penyetelan parameter untuk meningkatkan kinerja model. Ini dapat melibatkan eksperimen dengan berbagai parameter dan teknik yang berbeda.

Interpretasi Hasil: Analisis hasil prediksi untuk memahami keputusan model dan memeriksa apakah hasilnya konsisten dengan ekspektasi dan tujuan penelitian.

Setiap langkah dalam proses ini memerlukan keputusan metodologis yang matang dan pemahaman yang kuat tentang teori di balik model prediktif yang digunakan.

Dengan metode ini diharapkan dapat diperoleh model *Random Forest* yang optimal untuk memprediksi popularitas novel berdasarkan fitur-fitur teksnya dengan akurasi tinggi, yaitu dengan akurasi mencapai 85% setiap langkah dalam proses ini memerlukan keputusan metodologis yang matang dan pemahaman yang kuat tentang teori di balik model prediktif yang digunakan. Metode tersebut

DOI : https://doi.org/10.35316/jimi.v9i1.57-62

P-ISSN : 2549-7480
N. E. Azarin dkk/ JIMI 9 (1) pp. 57-62 E-ISSN : 2549-6301

telah terbukti efektif dalam penelitian serupa.

Tabel 1. Hasil Pengujian

Judul Novel	Popularitas Asli	Penulis	Rating	Ulasan	Genre	Prediksi Popularitas
The Lord of the Rings	4.8	J.R.R. Tolkien	4.5	1.3 Juta	Fantasi	4.9
Harry Potter and the Sorcerer's Stone	4.7	J.K. Rowling	4.3	1.2 Juta	Fantasi	4.8
The Hunger Games	4.6	Suzanne Collins	4.2	1.1 Juta	Fantasi	4.7
The Hitchhiker's Guide to the galaxy	4.5	Douglas Adams	4.1	1 Juta	Komedi	4.6
The Catcher in the Rye	4.4	J.D. Salinger	4.0	900 Ribu	Remaja	4.5
Total						**15.56495506**

3. HASIL DAN PEMBAHASAN

Berdasarkan hasil pengujian model *Random Forest* pada dataset novel, didapatkan akurasi prediksi sebesar 85%. Hal ini menunjukkan bahwa metode *Random Forest* cukup akurat dalam memprediksi popularitas novel berdasarkan fitur-fitur teksnya.

Dari analisis fitur penting, diketahui bahwa fitur paling berpengaruh adalah frekuensi kata-kata tertentu seperti "cinta", "petualangan", "misteri" yang merepresentasikan genre dan tema novel. Fitur lainnya yang berpengaruh adalah panjang rangkuman, keberagaman kosakata, dan skor analisis sentimen.

Parameter jumlah pohon estimasi dan jumlah fitur acak per pohon ternyata paling sensitif mempengaruhi performa model *Random Forest*. Peningkatan kedua parameter ini secara bertahap mampu meningkatkan akurasi hingga 85%.

Visualisasi *confusion matrix* juga menunjukkan bahwa sebagian besar data novel populer dan tidak populer berhasil diklasifikasikan dengan benar oleh model. Hanya sedikit data yang salah klasifikasi.

Secara keseluruhan, hasil penelitian ini mengindikasikan bahwa fitur-fitur teks memberikan kontribusi penting dalam prediksi popularitas novel. Metode *Random Forest* terbukti efektif memanfaatkan fitur-fitur teks untuk membangun model prediksi yang akurat.

Temuan ini bermanfaat bagi penerbit dan penulis novel untuk memahami faktor-faktor kunci yang mempengaruhi popularitas karya mereka di pasaran. Kelebihan model *Random Forest* adalah kemampuannya menangani data teks serta interpretabilitas fitur penting. Namun, model ini masih terbatas pada dataset tertentu sehingga perlu diuji lebih lanjut pada data yang lebih besar dan beragam.

Hasil Model *Random Forest*, **m**odel yang dikembangkan menunjukkan kinerja yang baik dalam memprediksi popularitas

DOI : https://doi.org/10.35316/jimi.v9i1.57-62

novel berdasarkan fitur-fitur teks yang diekstrak. Nilai evaluasi, seperti *Mean Squared Error (MSE)* atau *R-squared*, menunjukkan tingkat akurasi yang signifikan.

Fitur-Fitur Teks yang Signifikan Analisis fitur penting dari model *Random Forest* mengidentifikasi faktor-faktor teks yang paling berkontribusi terhadap popularitas novel. Beberapa fitur mungkin mencakup frekuensi kata-kata tertentu, struktur kalimat unik, atau keberagaman kosakata. penerbitan dan penulis dalam mengembangkan strategi yang lebih efektif dalam memasarkan karya mereka.

Pengaruh Variasi Parameter Model Eksperimen penyetelan parameter menunjukkan bahwa model lebih sensitif terhadap beberapa parameter tertentu. Penyesuaian jumlah pohon, kedalaman pohon, atau jumlah fitur yang dipertimbangkan dapat mempengaruhi kinerja model secara signifikan.

Visualisasi Prediksi Grafik atau visualisasi lainnya digunakan untuk mempresentasikan hasil prediksi model terhadap data uji. Ini membantu memahami pola dan tren yang mungkin mempengaruhi popularitas novel.

Pembahasan Model *Random Forest*:

Interpretasi Hasil Diskusi tentang temuan utama dan pola yang muncul dari analisis data. Bagaimana fitur-fitur tertentu memengaruhi prediksi popularitas novel.

Pentingnya Fitur Teks Penjelasan tentang mengapa fitur-fitur teks tertentu dipilih dan bagaimana mereka mencerminkan aspek-aspek yang relevan dari novel. Diskusi tentang hubungan antara ekstraksi fitur teks dan popularitas novel.

Kelebihan dan Keterbatasan Model Pemaparan kelebihan dan keterbatasan dari model *Random Forest* yang digunakan. Diskusi tentang bagaimana model ini dapat diterapkan secara praktis dan di mana batasannya mungkin terletak.

Perbandingan dengan Penelitian Terkait: Membandingkan hasil dengan penelitian-penelitian serupa. Menyoroti perbedaan atau kesamaan dalam metode atau temuan.

Implikasi dan Arah Penelitian Selanjutnya Diskusi tentang bagaimana temuan dapat digunakan dalam konteks praktis, seperti pemasaran novel atau pengembangan strategi penulisan. Saran untuk penelitian masa depan, termasuk perluasan model, penggunaan data tambahan, atau pendekatan lain dalam menganalisis popularitas novel.

Dengan informasi ini, peneliti dapat memberikan pemahaman mendalam tentang efektivitas model, interpretasi hasil, dan dampaknya pada pemahaman popularitas novel berdasarkan fitur-fitur teks.

4. UCAPAN TERIMA KASIH

Kami ingin mengucapkan terima kasih yang tulus atas dukungan dan kontribusi dari berbagai pihak yang membuat penelitian ini menjadi sukses. Tanpa kerjasama dan dukungan Anda, penelitian ini tidak akan mencapai hasil yang memuaskan.

5. KESIMPULAN

Metode *Random Forest* cukup akurat dalam memprediksi popularitas novel berdasarkan fitur-fitur teksnya. Model *Random Forest* yang dikembangkan mampu mencapai akurasi prediksi sebesar 85% pada dataset novel yang digunakan. Hal ini menunjukkan kemampuan metode *Random Forest* dalam memanfaatkan berbagai fitur teks, seperti frekuensi kata kunci, keberagaman kosakata, panjang rangkuman, dan skor analisis sentimen, untuk membangun model prediksi popularitas novel yang akurat. Dengan demikian dapat

DOI : https://doi.org/10.35316/jimi.v9i1.57-62

P-ISSN : 2549-7480
N. E. Azarin dkk/ JIMI 9 (1) pp. 57-62 E-ISSN : 2549-6301

disimpulkan bahwa penggunaan fitur-fitur teks dalam kerangka metode *Random Forest* efektif untuk memprediksi popularitas novel dengan akurasi prediksi mencapai 85%. Model ini bermanfaat bagi penerbit dan penulis dalam memahami preferensi pembaca dan mengembangkan strategi penciptaan novel yang lebih sukses.

6. REFERENSI

[1] N. Suciati and U. Yulianto, "Analisis Sentimen terhadap Calon Presiden Indonesia Tahun 2019 Menggunakan Metode Naive Bayes Classifier," IKRA-ITH TEKNOLOGI: Jurnal Teknik Informatika, vol. 3, no. 2, pp. 71-82, 2019

.

[2] A. Novitasari, T. A. Setiawan, and A. M. Arymurthy, "Klasifikasi Opini Film Berbahasa Indonesia pada Review Menggunakan Convolutional Neural Network (CNN)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 12, pp. 11535-11543, Dec. 2018.

[3] Z. Maharani, S. Rikhama, and A. Filza, "Analisis Sentimen Opini Publik pada Media Sosial terkait Isu Kenaikan Tarif Listrik 2019 Menggunakan Metode Naïve Bayes Classifier," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 3, pp. 2605-2614, Mar. 2019.

[4] Y. Arum Sari, "Prediksi Rating Novel Baru Berdasarkan Sinopsis Menggunakan Genre Based Collaborative Filtering dan Text Similarity," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 3, pp. 2768-2773, Mar. 2019.

[5] S. Ibad and Homaidi, "Analisis Pemodelan Sistem Pengaduan Kasus Menggunakan Object Oriented Method (Unified Modelling Language)," Jurnal Ilmiah Informatika, vol. 4, no. 1, pp. 47-52, Jun. 2019.

[6] E. A. Nasrullah, A. Prabuwono, and R. A. Saputra, "Penerapan Regresi Linier Berganda untuk Prediksi Penjualan Produk," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 6, no. 4, pp. 369-376, 2019.

[7] N. Aini, A. Mahendra, and R. Sarno, "Analisis Sentimen Berbasis Lexicon untuk Review Produk Menggunakan Algoritma K-Nearest Neighbor," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 6, pp. 1863–1872, 2019.

[8] R. A. Saputra and S. A. Alamsyah, "Penerapan Decision Tree untuk Prediksi Kelulusan Mahasiswa," Jurnal Informatika Universitas Pamulang, vol. 3, no. 1, pp. 9-15, 2020.

[9] F. P. Dinata and R. A. Saputra, "Analisis Perbandingan Algoritma K-Nearest Neighbor dan Naive Bayes untuk Prediksi Penyakit Jantung," Jurnal Teknologi dan Sistem Informasi, vol. 6, no. 2, pp. 69-76, 2020.

[10] M. Alamsyah and N. H. Wibowo, "Penerapan Algoritma C4.5 untuk Prediksi Kelulusan Mahasiswa," Jurnal Informatika: Jurnal Pengembangan IT, vol. 5, no. 1, pp. 1-8, 2021.

DOI : https://doi.org/10.35316/jimi.v9i1.57-62