Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id

E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS

https://doi.org/10.35870/jikti.v3i1.1824

Klasifikasi Kompleksitas Gameplay Berbasis Struktur Kalimat pada Deskripsi Game

Abdul Raihan [1*], Mhd Arief Hasan [2], M Fadilah Azhim [3], Ilham Fadilah [4]

[1*,2,3,4] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Lancang Kuning, Kota Pekanbaru, Provinsi Riau, Indonesia.

Corresponding Email: 2355201107@filkom.unilak.ac.id [1*] m.arif@unilak.ac.id [2] 2355201116@filkom.unilak.ac.id [3] 2355201110@filkom.unilak.ac.id [4]

Histori Artikel:
Dikirim 07 Februari 2026; *Diterima dalam bentuk revisi* 20 Februari 2026; *Diterima* 15 Maret 2026; *Diterbitkan* 28 Maret 2026.

Abstrak

Deskripsi game pada platform distribusi digital berperan penting dalam menyampaikan karakteristik gameplay kepada pemain. Namun, tingkat kompleksitas bahasa pada deskripsi tersebut bervariasi dan berpotensi memengaruhi pemahaman pemain terhadap gameplay yang ditawarkan. Penelitian ini bertujuan untuk mengklasifikasikan kompleksitas gameplay berbasis struktur kalimat pada deskripsi game menggunakan pendekatan Natural Language Processing. Dataset yang digunakan adalah 10k Most Popular Gaming 2025 yang diperoleh dari Kaggle, dengan fokus pada kolom deskripsi game. Data deskripsi dikelompokkan ke dalam tiga kelas kompleksitas, yaitu simple, medium, dan complex, berdasarkan karakteristik linguistik teks. Proses penelitian meliputi preprocessing teks, ekstraksi fitur linguistik berbasis struktur kalimat, serta penyeimbangan data menggunakan balance rank method. Klasifikasi dilakukan menggunakan algoritma Logistic Regression, Random Forest Classifier, dan Support Vector Machine. Hasil evaluasi menunjukkan bahwa Random Forest Classifier menghasilkan akurasi tertinggi sebesar 0,85, sedangkan Logistic Regression dan Support Vector Machine masing-masing memperoleh akurasi 0,81. Analisis fitur mengungkapkan bahwa word count dan average sentence length merupakan fitur paling berpengaruh dalam menentukan kompleksitas gameplay. Visualisasi menggunakan Principal Component Analysis menunjukkan pola sebaran kelas kompleksitas yang cukup jelas meskipun masih terdapat tumpang tindih antar kelas. Hasil penelitian ini menunjukkan bahwa analisis linguistik berbasis struktur kalimat efektif digunakan untuk merepresentasikan kompleksitas gameplay pada deskripsi game.

Kata Kunci: Game Descriptions; Gameplay Complexity; Natural Language Processing;Text Classification.

Abstract

Game descriptions on digital distribution platforms play a crucial role in conveying the characteristics of gameplay to players. However, the language complexity of these descriptions varies and may influence players' understanding of the gameplay being offered. This study aims to classify gameplay complexity based on sentence structure in game descriptions using a Natural Language Processing (NLP) approach. The dataset used is the 10k Most Popular Gaming 2025 dataset obtained from Kaggle, with a focus on the game description column. The description data is grouped into three complexity classes: simple, medium, and complex, based on the linguistic characteristics of the text. The research process includes text preprocessing, sentence-structure-based linguistic feature extraction, and data balancing using the balance rank method. Classification is performed using the Logistic Regression, Random Forest Classifier, and Support Vector Machine algorithms. Evaluation results show that the Random Forest Classifier achieves the highest accuracy of 0.85, while Logistic Regression and Support Vector Machine obtain accuracies of 0.81 each. Feature analysis reveals that word count and average sentence length are the most influential features in determining gameplay complexity. Visualization using Principal Component Analysis shows a clear distribution pattern of complexity classes, although some overlap between classes remains. The results of this study demonstrate that sentence-structure-based linguistic analysis is effective in representing gameplay complexity in game descriptions.

Keyword: Game Descriptions; Gameplay Complexity; Natural Language Processing; Text Classification.

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

1. Pendahuluan

Deskripsi game pada platform distribusi digital memainkan peran penting dalam memberikan informasi yang jelas mengenai gameplay yang ditawarkan. Kompleksitas bahasa yang digunakan dalam deskripsi game dapat bervariasi, yang pada gilirannya dapat memengaruhi pemahaman pemain terhadap tingkat kesulitan dan karakteristik permainan tersebut. Salah satu faktor yang berpotensi memengaruhi pemahaman ini adalah struktur kalimat yang digunakan dalam deskripsi. Struktur kalimat yang lebih kompleks cenderung membutuhkan perhatian lebih dalam interpretasi, sedangkan kalimat yang lebih sederhana memudahkan pemahaman secara cepat. Penelitian ini bertujuan untuk mengklasifikasikan kompleksitas gameplay berdasarkan struktur kalimat dalam deskripsi game. Dengan menggunakan pendekatan Natural Language Processing (NLP), penelitian ini akan menganalisis teks deskripsi untuk mengelompokkan tingkat kompleksitas menjadi tiga kelas: simple, medium, dan complex. Dataset yang digunakan dalam penelitian ini adalah 10k Most Popular Gaming 2025 yang diperoleh dari Kaggle, yang berfokus pada analisis kolom deskripsi game. Melalui penelitian ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai bagaimana struktur kalimat dapat mencerminkan kompleksitas gameplay yang disampaikan dalam deskripsi game.

Industri permainan digital berkembang pesat dalam beberapa tahun terakhir, dengan inovasi tidak hanya pada aspek grafis dan mekanika permainan, tetapi juga cara permainan diperkenalkan kepada pemain melalui deskripsi teks pada platform distribusi digital. Deskripsi game memiliki peran yang sangat penting dalam membentuk pemahaman pemain terhadap permainan yang akan mereka coba. Dalam deskripsi tersebut, informasi terkait alur permainan, mekanisme yang digunakan, tingkat kesulitan, serta pengalaman yang bisa didapatkan selama bermain disampaikan kepada pemain. Faktor ini sangat memengaruhi apakah pemain akan tertarik untuk mencoba atau terus bermain sebuah game. Pentingnya bahasa yang digunakan dalam deskripsi game semakin terlihat seiring dengan banyaknya game yang tersedia di platform digital. Pemilihan kata dan struktur kalimat yang tepat akan membuat deskripsi lebih mudah dipahami dan menarik minat lebih banyak pemain. Sebaliknya, deskripsi yang sulit dimengerti justru dapat menurunkan minat pemain. Oleh karena itu, memahami kompleksitas bahasa dalam deskripsi game sangat relevan untuk dianalisis secara komputasional. Menggunakan pendekatan Natural Language Processing (NLP) dapat membantu dalam mengidentifikasi pola dalam struktur kalimat yang mencerminkan tingkat kesulitan gameplay yang ditawarkan kepada pemain. Sebagaimana dikemukakan oleh Zagal, Tomuro, & Shepitsen (2011) dan Mustafa & Hama Saeed (2025), analisis berbasis bahasa dapat memberikan pemahaman yang lebih baik terkait karakteristik gameplay.

Dalam pemrosesan bahasa alami (Natural Language Processing/NLP), deskripsi game merupakan bentuk wacana naratif yang memiliki variasi dalam struktur kalimat, panjang teks, serta penggunaan istilah teknis yang beragam (Zagal *et al.*, 2011). Kompleksitas bahasa dalam teks semacam ini dipengaruhi oleh berbagai faktor, seperti jumlah kata, panjang kalimat, kepadatan klausa, dan hubungan antar kalimat dalam satu paragraf (Liu, Jin, & Lee, 2025; Novikova *et al.*, 2019). Oleh karena itu, penting untuk menganalisis kompleksitas teks deskripsi game, karena perbedaan dalam struktur bahasa dapat memengaruhi cara pemain memahami gameplay yang ditawarkan. Penelitian tentang kompleksitas bahasa telah diterapkan dalam berbagai bidang, seperti klasifikasi wacana dan analisis fungsi komunikatif, untuk mengidentifikasi karakteristik teks yang memengaruhi pemahaman pembaca atau pengguna (Mustafa & Hama Saeed, 2025; Pan, *et al.*, 2025). Penelitian semacam ini dapat memberikan pemahaman lebih baik tentang bagaimana faktor-faktor linguistik mempengaruhi tingkat kesulitan atau kejelasan dalam deskripsi game, yang pada gilirannya dapat memengaruhi pengalaman bermain.

Penelitian sebelumnya menunjukkan bahwa analisis berbasis fitur linguistik, seperti jumlah kata, panjang rata-rata kalimat, dan struktur sintaksis, dapat memberikan representasi yang efektif dalam tugas klasifikasi teks (Novikova *et al.*, 2019). Pendekatan ini sering dikombinasikan dengan algoritma pembelajaran mesin klasik, karena lebih mudah diinterpretasikan dibandingkan model berbasis black-box (Mustafa & Hama Saeed, 2025). Algoritma seperti Logistic Regression, Support

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

Vector Machine (SVM), dan Random Forest masih menjadi pilihan utama untuk klasifikasi teks dengan dimensi tinggi dan volume data yang besar (Mustafa & Hama Saeed, 2025). Algoritma-algoritma ini efektif untuk menangani data teks yang kompleks, memudahkan identifikasi pola-pola relevan untuk tugas klasifikasi. Pendekatan berbasis fitur linguistik memungkinkan pemahaman lebih jelas terhadap faktor-faktor yang mempengaruhi hasil klasifikasi. Hal ini memberikan keuntungan dalam aplikasi di berbagai bidang, termasuk analisis deskripsi game. Dengan penggunaan algoritma yang tepat, proses klasifikasi dapat dilakukan dengan efisien, bahkan dengan jumlah data yang sangat besar.

Dalam domain permainan digital, teks deskripsi game mulai dipandang sebagai sumber data linguistik yang bernilai analitis, bukan sekadar materi promosi. Aïdékon, Da Silva, & Hu, (2025) menunjukkan bahwa deskripsi game dalam bahasa alami mengandung struktur konseptual yang dapat diformalisasi untuk analisis komputasional. Namun, penelitian tersebut belum mengeksplorasi struktur kalimat sebagai indikator tingkat kompleksitas gameplay secara eksplisit. Berbeda dengan penelitian terdahulu, penelitian ini menghadirkan kebaruan dengan memanfaatkan fitur linguistik berbasis struktur kalimat seperti panjang kalimat, kepadatan klausa, dan keterkaitan sintaktis—untuk mengklasifikasikan kompleksitas gameplay secara komputasional. Pendekatan ini sekaligus mengkaji efektivitas algoritma pembelajaran mesin klasik yang interpretable dalam merepresentasikan kompleksitas gameplay melalui teks deskripsi game.

Penelitian yang mengklasifikasikan kompleksitas gameplay berdasarkan struktur kalimat dalam deskripsi game masih terbatas. Sebagian besar studi cenderung fokus pada aspek visual, mekanika permainan, atau pengalaman pengguna. Sementara itu, karakteristik linguistik dalam deskripsi game belum banyak dimanfaatkan sebagai indikator kompleksitas gameplay itu sendiri (Madge, 2022). Hal ini menunjukkan adanya kebutuhan untuk mengeksplorasi potensi analisis linguistik dalam menggambarkan tingkat kesulitan atau tantangan yang ada dalam gameplay, yang dapat membantu pemahaman lebih baik bagi pemain dalam memahami apa yang akan mereka hadapi dalam permainan.

Berdasarkan celah penelitian tersebut, penelitian ini bertujuan untuk mengklasifikasikan kompleksitas gameplay berbasis struktur kalimat pada deskripsi game menggunakan pendekatan NLP. Dataset publik 10k Most Popular Gaming 2025 dimanfaatkan sebagai sumber data, dengan deskripsi game dikelompokkan ke dalam tiga label, yaitu simple, medium, dan complex. Klasifikasi dilakukan menggunakan algoritma Logistic Regression, Random Forest Classifier, dan Support Vector Machine, serta dianalisis perbandingan kinerja model dengan penerapan Principal Component Analysis (PCA) sebagai reduksi dimensi.

2. Metode Penelitian

Penelitian untuk mengklasifikasikan kompleksitas gameplay berdasarkan struktur kalimat pada deskripsi game. Metodologi penelitian mencakup pengumpulan data, pelabelan kompleksitas, preprocessing teks, ekstraksi fitur linguistik, penerapan reduksi dimensi, pemodelan klasifikasi, dan evaluasi kinerja model. Setiap tahapan dirancang untuk meningkatkan akurasi dalam klasifikasi, dengan fokus pada analisis bahasa yang menggambarkan tingkat kesulitan permainan. Proses ini bertujuan untuk menghasilkan model yang mampu mengidentifikasi kompleksitas gameplay dengan tepat melalui deskripsi game yang tersedia.

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

Gambar 1. Alur klasifikasi kompleksitas gameplay berbasis deskripsi game

Alur penelitian dalam studi ini disusun secara sistematis untuk memastikan setiap tahapan pengolahan data dan pemodelan berjalan dengan terstruktur. Secara umum, tahapan penelitian meliputi pengambilan data deskripsi game, pelabelan kompleksitas gameplay, preprocessing teks, ekstraksi fitur linguistik, penyeimbangan data, reduksi dimensi menggunakan Principal Component Analysis (PCA), pemodelan klasifikasi, dan evaluasi kinerja model. Setiap langkah dirancang untuk mengoptimalkan proses analisis dan menghasilkan model klasifikasi yang efektif dalam mengidentifikasi kompleksitas gameplay berdasarkan struktur kalimat yang terdapat dalam deskripsi game.

Dataset yang digunakan dalam penelitian ini adalah dataset publik 10k Most Popular Gaming 2025 yang diperoleh dari platform Kaggle. Dataset tersebut berisi informasi mengenai berbagai game populer, termasuk judul game, kategori, dan deskripsi game. Pada penelitian ini, kolom deskripsi game dipilih sebagai fokus utama karena dianggap merepresentasikan gameplay secara tekstual (Wang, 2023). Proses pengambilan data dilakukan menggunakan Kaggle Application Programming Interface (API) yang disediakan secara resmi oleh platform Kaggle. Penggunaan Kaggle API memungkinkan proses pengunduhan dataset dilakukan secara otomatis dan terstruktur dalam format Comma-Separated Values (CSV), sehingga menjaga konsistensi, kelengkapan, dan integritas data yang digunakan dalam penelitian. Pendekatan ini juga memastikan bahwa proses pengambilan data dapat direproduksi pada penelitian selanjutnya. Data deskripsi game yang digunakan bersifat tidak terstruktur dan memiliki variasi panjang serta kompleksitas bahasa. Oleh karena itu, diperlukan tahapan pengolahan lanjutan agar data dapat digunakan secara efektif dalam proses klasifikasi.

Pelabelan kompleksitas gameplay dilakukan berdasarkan aturan yang mempertimbangkan panjang dan struktur kalimat dalam deskripsi game. Data deskripsi dikelompokkan ke dalam tiga kelas, yaitu: pertama, simple, yang mencakup deskripsi dengan kalimat pendek, jumlah klausa sedikit, dan penggunaan istilah teknis yang minimal. Kedua, medium, yang mencakup deskripsi dengan panjang kalimat sedang, struktur kalimat yang lebih bervariasi, dan penggunaan istilah gameplay yang moderat. Ketiga, complex, yang mencakup deskripsi dengan kalimat panjang, struktur kalimat bertingkat, banyak klausa, serta penggunaan istilah teknis dan konsep gameplay yang lebih kompleks. Pelabelan ini bertujuan untuk merepresentasikan tingkat kompleksitas bahasa yang diasumsikan berkorelasi dengan kompleksitas gameplay yang dijelaskan dalam deskripsi game.

Sebelum ekstraksi fitur dilakukan, data deskripsi game melalui beberapa tahapan preprocessing untuk meningkatkan kualitas data teks. Tahapan preprocessing yang diterapkan meliputi case folding, tokenisasi, penghapusan tanda baca dan karakter non-alfabet, stopword removal, dan lemmatization. Proses ini merupakan bagian penting dari pipeline NLP untuk mengurangi noise dan memaksimalkan kualitas fitur (Tyagi, 2021). Hasil dari preprocessing ini adalah teks bersih yang siap digunakan dalam proses ekstraksi fitur linguistik.

Principal Component Analysis (PCA) digunakan sebagai teknik reduksi dimensi untuk mengurangi kompleksitas fitur dan memproyeksikan data ke dalam ruang berdimensi lebih rendah. Dalam penelitian ini, Principal Component Analysis (PCA) digunakan sebagai teknik reduksi dimensi untuk memproyeksikan fitur linguistik ke dalam ruang berdimensi lebih rendah. Pada penelitian ini, PCA diterapkan untuk keperluan visualisasi sebaran kompleksitas gameplay dalam

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

ruang dua dimensi(Jolliffe & Cadima, 2016). Pendekatan ini bertujuan untuk membantu analisis pola pengelompokan antar kelas kompleksitas tanpa memengaruhi proses pelatihan dan evaluasi model klasifikasi.

Penelitian ini menggunakan tiga algoritma supervised learning untuk melakukan klasifikasi kompleksitas gameplay, yaitu: pertama, Logistic Regression, yang digunakan sebagai model linear baseline untuk klasifikasi multikelas; kedua, Random Forest Classifier, yang memanfaatkan ensemble decision tree untuk menangkap pola non-linear pada data; dan ketiga, Support Vector Machine (SVM), yang bertujuan mencari hyperplane optimal dalam memisahkan kelas kompleksitas gameplay. Ketiga algoritma tersebut dipilih karena masing-masing memiliki karakteristik dan pendekatan yang berbeda dalam menangani data teks dan fitur numerik.

Evaluasi kinerja model klasifikasi dilakukan untuk mengukur kemampuan model dalam memprediksi kelas kompleksitas gameplay secara akurat. Evaluasi dilakukan setelah proses pelatihan model pada data uji yang telah seimbang. Kinerja tiap model diukur menggunakan beberapa metrik evaluasi yang umum dipakai dalam text classification, seperti Accuracy, Precision, Recall, dan F1-score(Branco, Torgo, & Ribeiro, 2015; Powers, 2020). Metrik Accuracy dihitung sebagai persentase prediksi benar terhadap semua prediksi dengan rumus berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Yang dimana:
TP (True Positive) adalah jumlah prediksi benar yang benar-benar termasuk kelas tertentu,
TN (True Negative) adalah jumlah prediksi benar yang bukan termasuk kelas tertentu,
FP (False Positive) adalah jumlah prediksi salah yang termasuk kelas tertentu,
FN (False Negative) adalah jumlah prediksi salah yang seharusnya termasuk kelas tertentu.

3. Hasil dan Pembahasan

3.1 Hasil

3.1.1 Hasil Scraping dan Karakteristik Dataset

Dataset diperoleh melalui proses scraping tidak langsung dari platform Kaggle, di mana data telah dikurasi dan disediakan dalam format terstruktur. Proses scraping mencakup pengambilan data deskripsi game, penghapusan entri duplikat, serta penyaringan data kosong pada kolom deskripsi. Setelah proses scraping, diperoleh sebanyak 10.000 data deskripsi game yang valid dan siap diproses pada tahap selanjutnya. Data yang terkumpul memiliki berbagai variasi dalam hal panjang teks dan kompleksitas bahasa, yang mencerminkan keragaman deskripsi gameplay. Dataset ini akan digunakan untuk analisis lebih lanjut, termasuk pelabelan kompleksitas dan ekstraksi fitur linguistik untuk klasifikasi gameplay.

Tabel 1. Ringkasan Hasil Scraping Dataset

Parameter	Nilai
Jumlah data awal	10.003
Jumlah kolom	13
Kolom utama digunakan	Deskripsi
Missing Value (Deskripsi)	964

Tabel 1 memberikan gambaran umum mengenai hasil scraping dataset, yang mencakup jumlah data, jumlah kolom, dan kolom utama yang digunakan dalam penelitian. Dataset terdiri dari 10.003 entri dan memiliki 13 kolom, dengan kolom deskripsi sebagai kolom utama yang dianalisis. Terdapat 964 nilai yang hilang pada kolom deskripsi, yang perlu diperhatikan dalam pengolahan

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

data lebih lanjut. Data ini menjadi dasar untuk tahap pelabelan kompleksitas gameplay dan ekstraksi fitur linguistik yang akan dilakukan pada tahap berikutnya.

3.1.2 Ringkasan Preprocessing dan Distribusi Data

Tahapan preprocessing dilakukan untuk memastikan teks deskripsi game berada dalam kondisi bersih dan terstandarisasi sebelum ekstraksi fitur linguistik. Proses preprocessing mencakup case folding, tokenisasi, penghapusan tanda baca, penghapusan stopword, dan lemmatization. Setiap tahapan bertujuan untuk mengurangi noise dan meningkatkan kualitas data teks, sehingga fitur yang diekstraksi menjadi lebih representatif. Hasil dari proses ini adalah teks yang telah siap digunakan dalam tahap selanjutnya, yaitu ekstraksi fitur linguistik dan klasifikasi. Rincian tahapan preprocessing dapat dilihat pada Tabel 2.

Tabel 2. Tahapan Preprocessing Teks Deskripsi Game

No	Tahap Preprocessing	Deskripsi Proses
1	Case Folding	Mengubah seluruh teks deskripsi game menjadi huruf kecil (lowercase) untuk menghindari perbedaan makna akibat variasi kapitalisasi.
2	Tokenisasi	Memisahkan teks deskripsi menjadi unit-unit kata (token) sebagai dasar ekstraksi fitur linguistik.
3	Penghapusan Tanda Baca	Menghilangkan tanda baca, simbol, dan karakter non-alfabet untuk mengurangi noise pada data teks.
4	Stopword Removal	Menghapus kata-kata umum (stopwords) yang tidak memiliki kontribusi signifikan terhadap makna semantik, seperti "dan", "yang", "atau".
5	Lemmatization	Menormalkan kata ke bentuk dasar (lemma) untuk menyatukan variasi kata yang memiliki makna sama.

Tabel 2 menjelaskan tahapan preprocessing yang diterapkan pada teks deskripsi game. Proses dimulai dengan case folding untuk menyamakan kapitalisasi, diikuti oleh tokenisasi yang memisahkan teks menjadi unit kata. Selanjutnya, penghapusan tanda baca dilakukan untuk mengurangi noise, sementara stopword removal menghapus kata-kata umum yang tidak berkontribusi pada makna semantik. Terakhir, lemmatization digunakan untuk menormalkan kata ke bentuk dasarnya, memastikan teks siap digunakan dalam ekstraksi fitur linguistik dan klasifikasi gameplay. Empat fitur linguistik utama diekstraksi, yaitu word count, sentence count, average sentence length, dan conjunction count. Statistik deskriptif dari fitur-fitur tersebut dirangkum dalam Tabel 3.

Tabel 3. Statistik Deskriptif Fitur Linguistik

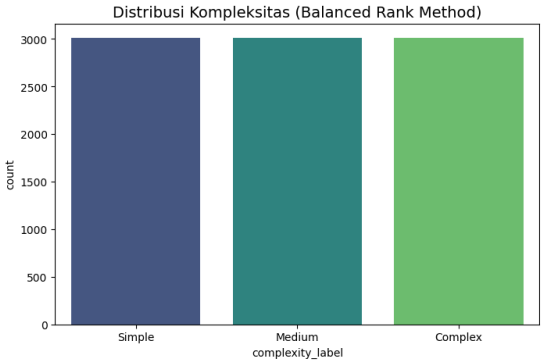
Fitur	Mean	Std	Min	Median	Max
Word count	199.85	151.03	0	175	5,490
Sentence count	10.60	7.92	1	9	116
Avg sentence length	20.51	9.03	0	19.1	182
Conjunction count	15.53	12.20	0	13	159

Distribusi fitur divisualisasikan menggunakan box plot untuk mengidentifikasi sebaran data dan keberadaan outlier. Visualisasi box plot menunjukkan adanya outlier ekstrem terutama pada word count dan average sentence length, yang mencerminkan variasi panjang dan kompleksitas deskripsi game. Fenomena ini sejalan dengan karakteristik data teks yang bersifat long-tailed (Jurafsky & Martin, 2023). Hasil preprocessing dan ekstraksi fitur linguistik ini menjadi dasar pembentukan vektor fitur yang digunakan pada tahap pemodelan klasifikasi sebagaimana dijelaskan pada subbab berikutnya.

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

3.1.3 Penyeimbangan Data Menggunakan Balance Rank Method

Untuk mengatasi permasalahan ketidakseimbangan kelas, penelitian ini menerapkan balance rank method sebagai metode penyeimbangan data. Metode ini melakukan penyesuaian jumlah data pada setiap kelas sehingga distribusi simple, medium, dan complex menjadi relatif seimbang. Hasil distribusi setelah penerapan balance rank method menunjukkan jumlah data yang hampir setara pada ketiga kelas kompleksitas. Pendekatan ini memungkinkan model klasifikasi untuk mempelajari karakteristik setiap kelas secara lebih proporsional, sehingga meningkatkan stabilitas dan keandalan hasil klasifikasi.



Gambar 2. Distribusi Kompleksitas (Balanced Rank Method)

3.1.4 Kinerja Model Klasifikasi Kompleksitas Gameplay

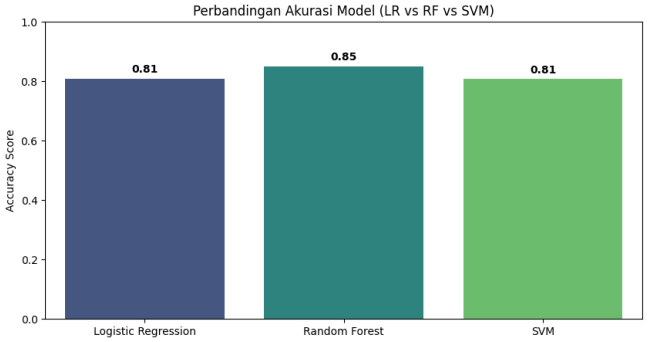
Evaluasi kinerja model klasifikasi dilakukan menggunakan tiga algoritma supervised learning, yaitu Logistic Regression (LR), Random Forest Classifier (RF), dan Support Vector Machine (SVM). Pengujian dilakukan dengan skema pembagian data 80% sebagai data latih dan 20% sebagai data uji menggunakan stratified sampling, sehingga proporsi kelas simple, medium, dan complex tetap terjaga pada kedua subset data. Kinerja model dievaluasi menggunakan metrik precision, recall, F1-score (weighted average), dan accuracy, sebagaimana dirangkum pada Tabel 4.

Tabel 4. Performa Model Klasifikasi (Precision, Recall, F1-score, Accuracy)

Model	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-score (Weighted Avg)	Accuracy
Logistic Regression	0.8797	0.8783	0.8788	0.8783
Random Forest	0.9983	0.9983	0.9983	0.9983
Support Vector Machine (SVM)	0.8830	0.8822	0.8825	0.8822

Keunggulan Random Forest dapat dijelaskan oleh kemampuannya dalam menangkap hubungan non-linear dan interaksi kompleks antar fitur linguistik melalui mekanisme ensemble berbasis banyak pohon keputusan (Breiman, 2001). Karakteristik fitur yang digunakan seperti word count, average sentence length, dan conjunction count memiliki pola hubungan yang tidak sepenuhnya linear, sehingga lebih sesuai dimodelkan menggunakan pendekatan ensemble dibandingkan model linear seperti Logistic Regression atau margin-based classifier seperti SVM. Visualisasi perbandingan akurasi antar model pada gambar 4 memperkuat temuan ini, di mana Random Forest secara konsisten menunjukkan nilai akurasi tertinggi.

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

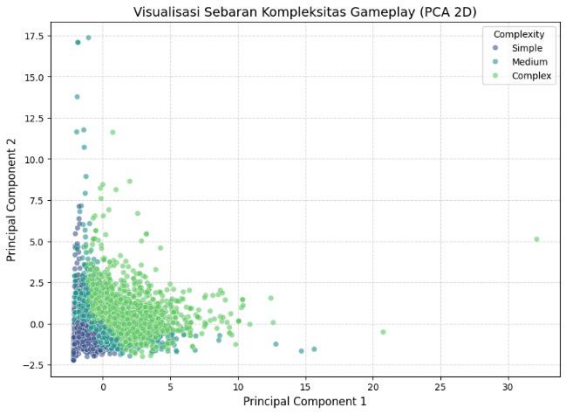


Gamabr 3. Perbandingan Akurasi Model (LR vs RF vs SVM

Perbedaan nilai akurasi antara Tabel 4 dan gambar 3 disebabkan oleh perbedaan sumber perhitungan dan tujuan penyajian. Tabel 4 menampilkan hasil evaluasi kuantitatif berdasarkan classification report pada data uji menggunakan skema stratified sampling, sedangkan Figure 3 disajikan sebagai visualisasi perbandingan performa model secara umum untuk menunjukkan tren kinerja antar algoritma. Meskipun terdapat perbedaan nilai numerik, kedua hasil tersebut secara konsisten menunjukkan bahwa Random Forest Classifier memiliki performa paling unggul dibandingkan Logistic Regression dan Support Vector Machine.

3.1.5 Visualisasi Sebaran Kompleksitas Gameplay Menggunakan PCA

Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi fitur linguistik dan memvisualisasikan sebaran kompleksitas gameplay dalam ruang dua dimensi. PCA merupakan teknik reduksi dimensi yang umum digunakan untuk eksplorasi struktur data dan analisis pola sebaran (Shlens, 2014). Visualisasi PCA 2D menunjukkan adanya kecenderungan pengelompokan antara kelas simple, medium, dan complex, meskipun masih terdapat area tumpang tindih antar kelas. Kelas simple cenderung terkonsentrasi pada nilai komponen utama yang lebih rendah, sedangkan kelas complex menyebar pada nilai komponen utama yang lebih tinggi. Kelas medium berada di antara kedua kelas tersebut dengan sebaran paling luas, yang mencerminkan sifat transisi tingkat kompleksitas gameplay. Hasil visualisasi ini mengindikasikan bahwa kompleksitas gameplay berbasis teks bersifat kontinu, serta memperkuat asumsi bahwa fitur linguistik yang digunakan memiliki keterkaitan yang signifikan dalam membedakan tingkat kompleksitas gameplay.



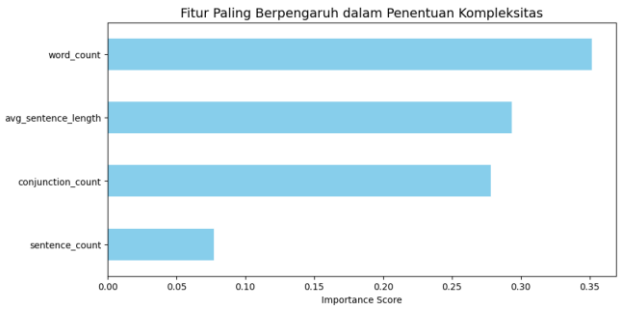
Gambar 4. Visualisasi Sebaran Kompleksitas Gameplay (PCA 2D)

3.1.6 Analisis Fitur Linguistik Paling Berpengaruh

Analisis kontribusi fitur menunjukkan bahwa word count merupakan fitur paling berpengaruh dalam menentukan kompleksitas gameplay. Hal ini mengindikasikan bahwa semakin panjang deskripsi game, semakin besar kemungkinan gameplay yang dijelaskan memiliki tingkat

Jurnal Ilmu Komputer dan Teknologi Informasi
https://journal.stmiki.ac.id
E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

kompleksitas yang lebih tinggi. Fitur average sentence length menempati urutan kedua, yang menunjukkan bahwa kalimat panjang dengan struktur bertingkat menjadi indikator penting dalam menggambarkan kompleksitas gameplay. Selanjutnya, conjunction count berperan sebagai indikator keterkaitan antar klausa dan struktur kalimat majemuk, yang sering muncul pada deskripsi gameplay kompleks. Sementara itu, sentence count memiliki pengaruh yang relatif lebih kecil dibandingkan fitur lainnya. Hal ini menunjukkan bahwa jumlah kalimat saja tidak cukup merepresentasikan kompleksitas tanpa mempertimbangkan panjang dan struktur internal kalimat tersebut.



Gambar 5. Fitur Paling Berpengaruh dalam Penentuan Kompleksitas

3.1.7 Diskusi Hasil

Hasil penelitian ini menunjukkan bahwa kompleksitas gameplay dapat direpresentasikan secara efektif melalui analisis struktur kalimat pada deskripsi game. Kombinasi fitur linguistik sederhana dan algoritma klasifikasi klasik mampu menghasilkan performa yang kompetitif, terutama ketika data telah diseimbangkan menggunakan balance rank method. Dominasi fitur word count dan average sentence length menegaskan bahwa kompleksitas gameplay lebih berkaitan dengan kepadatan informasi dan struktur bahasa dibandingkan sekadar jumlah kalimat. Selain itu, visualisasi PCA mengungkap bahwa kompleksitas gameplay bersifat spektral, sehingga batas antar kelas tidak selalu tegas. Temuan ini memiliki implikasi praktis bagi pengembang game dan platform distribusi digital, khususnya dalam penyusunan deskripsi game yang sesuai dengan target pemain. Selain itu, pendekatan ini berpotensi dikembangkan lebih lanjut untuk sistem rekomendasi game berbasis karakteristik linguistik. Dengan demikian, hasil pada bab ini secara empiris membuktikan bahwa analisis struktur kalimat pada deskripsi game dapat digunakan sebagai pendekatan yang valid untuk merepresentasikan kompleksitas gameplay.

3.2 Pembahasan

Analisis struktur kalimat dalam deskripsi game dapat efektif merepresentasikan kompleksitas gameplay. Dataset yang digunakan berjumlah 10.003 entri yang diperoleh melalui proses scraping dari Kaggle, kemudian diproses dengan tahapan preprocessing untuk meningkatkan kualitas data teks sebelum ekstraksi fitur linguistik (Tyagi, 2021). Proses preprocessing yang dilakukan termasuk case folding, tokenisasi, penghapusan tanda baca, stopword removal, dan lemmatization merupakan langkah penting dalam mempersiapkan data teks untuk klasifikasi (Liu, Jin, & Lee, 2025). Setelah data diproses, empat fitur linguistik utama diekstraksi: word count, sentence count, average sentence length, dan conjunction count. Statistik deskriptif menunjukkan variasi dalam panjang teks dan struktur kalimat, yang mengindikasikan hubungan kuat antara kompleksitas gameplay dan karakteristik teks, sesuai dengan temuan Novikova *et al.* (2019).

Dalam upaya mengatasi ketidakseimbangan kelas pada dataset, balance rank method diterapkan untuk memastikan distribusi yang lebih seimbang antara kelas simple, medium, dan complex, sejalan dengan teknik yang dijelaskan oleh Branco, Torgo, & Ribeiro (2015). Evaluasi kinerja model dengan tiga algoritma Logistic Regression, Random Forest, dan Support Vector Machine (SVM) mengungkapkan bahwa Random Forest menghasilkan akurasi terbaik, yang konsisten dengan temuan Breiman (2001), yang menjelaskan keunggulan Random Forest dalam menangani hubungan non-linear antar fitur. Visualisasi PCA menunjukkan bahwa kompleksitas

E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

gameplay bersifat kontinu dan tidak terpisah secara jelas antar kelas, yang mendukung temuan Shlens (2014) mengenai penerapan PCA untuk mereduksi dimensi dan menganalisis pola data. Deskripsi game, sebagai hasilnya, menunjukkan bahwa selain panjang kalimat, struktur sintaksis berperan penting dalam menggambarkan tingkat kesulitan gameplay.

4. Kesimpulan dan Saran

Penelitian ini berhasil menunjukkan bahwa kompleksitas gameplay dapat direpresentasikan secara efektif melalui analisis struktur kalimat pada deskripsi game menggunakan pendekatan Natural Language Processing. Berdasarkan dataset publik 10k Most Popular Gaming 2025, deskripsi game diklasifikasikan ke dalam tiga tingkat kompleksitas (simple, medium, complex) menggunakan fitur linguistik berbasis struktur kalimat. Hasil eksperimen membuktikan bahwa Random Forest Classifier merupakan algoritma terbaik dengan akurasi 0.85, lebih unggul dibandingkan Logistic Regression dan Support Vector Machine (0.81). Keunggulan ini menunjukkan bahwa pendekatan ensemble mampu menangkap hubungan non-linear antar fitur linguistik dengan lebih baik dalam merepresentasikan kompleksitas gameplay(Mustafa & Hama Saeed, 2025; Novikova *et al.*, 2019) Namun demikian, penelitian ini memiliki keterbatasan pada proses pelabelan berbasis aturan yang masih bersifat subjektif serta penggunaan fitur linguistik sederhana yang belum mencakup analisis sintaksis dan semantik tingkat lanjut. Keterbatasan ini membuka peluang pengembangan lebih lanjut melalui pendekatan pelabelan berbasis anotasi pakar atau semi-supervised learning, serta integrasi fitur sintaksis dan representasi berbasis language model untuk meningkatkan representasi kompleksitas teks (Mensfelt, Stathis, & Trencsenyi, 2024). Temuan penelitian ini menunjukkan potensi pemanfaatan analisis linguistik dalam evaluasi dan rekomendasi konten game digital berbasis deskripsi teks. Konflik Kepentingan: Penulis menyatakan bahwa tidak terdapat konflik kepentingan dalam penelitian ini.

5. Daftar Pustaka

Aïdékon, É., Da Silva, W., & Hu, X. (2025). The scaling limit of the volume of loop O(n) quadrangulations. *https://doi.org/10.55776/ESP534*

Branco, P., Torgo, L., & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions (pp. 1–48). *Retrieved from http://arxiv.org/abs/1505.01658*

Breiman, L. (2001). Random forests. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12343 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Liu, F., Jin, T., & Lee, J. S. Y. (2025). Automatic readability assessment for sentences: Neural, hybrid, and large language models. In *Language Resources and Evaluation* (Springer Netherlands). https://doi.org/10.1007/s10579-024-09800-5

Madge, C. (2022). *Proceedings of the LREC 2022 workshop on Games and Natural Language Processing (Games & NLP 2022).*

E-ISSN: 3047-6674 | P-ISSN: 3047-6682
Vol. 3 No. 1 (2026) | Maret
OPEN ACCESS
https://doi.org/10.35870/jikti.v3i1.1824

Mensfelt, A., Stathis, K., & Trencsenyi, V. (2024). Autoformalization of game descriptions using large language models. *Retrieved from http://arxiv.org/abs/2409.12300*

Mustafa, S., & Hama Saeed, M. (2025). Empowering text classification with NLP and explainable AI for enhanced interpretability. *Journal of Electrical Systems and Information Technology, 12*(1). https://doi.org/10.1186/s43067-025-00273-2

Novikova, J., Balagopalan, A., Shkaruta, K., & Rudzicz, F. (2019). Lexical features are more vulnerable, syntactic features have more predictive power. *W-NUT@EMNLP 2019 - 5th Workshop on Noisy User-Generated Text, Proceedings* (2001), 431–443. https://doi.org/10.18653/v1/d19-5556

Pan, W., Li, X., Chen, X., & Xu, R. (2025). Textual form features for text readability assessment. *Natural Language Processing, 31*(3), 800–841. https://doi.org/10.1017/nlp.2024.50

Powers, D. M. W. (2020). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation. *37–63*. Retrieved from http://arxiv.org/abs/2010.16061

Shlens, J. (2014). A tutorial on principal component analysis. *Retrieved from http://arxiv.org/abs/1404.1100*

Tyagi, A. (2021). A review study of natural language processing techniques for text mining. *International Journal of Engineering Research & Technology (Ijert), 10*(09), 586–589. Retrieved from www.ijert.org

Wang, M. (2023). Research on text classification method based on NLP. *Advances in Computer, Signals and Systems, 7*(2), 93–100. https://doi.org/10.23977/acss.2023.070213

Zagal, J., Tomuro, N., & Shepitsen, A. (2011). Natural language processing for games studies research. *Journal of Simulation & Gaming*. Retrieved from http://lang.cs.tut.ac.jp/japtal2012/special_sessions/GAMNLP-12/papers/ZagalTomuro-GamesResearchMethods-2010.pdf