# Implementasi Klasifikasi *Naïve Bayes* dan Pemodelan Topik dengan *Latent Dirichlet Allocation* untuk Data Ulasan *Video Game* Lokal Pada *Platform Steam*

**Yusra Sakti Wardhana [1*], Ayundyah Kesumawati[1]**

[1]Universitas Islam Indonesia, Jl. Kaliurang Km 14.5, Kabupaten Sleman, 55584, dan Indonesia
*Corresponding author: yusra.wardhana@students.uii.ac.id

**P-ISSN:** 2986-4178
**E-ISSN:** 2988-4004

**Riwayat Artikel**
Dikirim: 29 Maret 2023
Direvisi: 08 Agustus 2023
Diterima: 27 November 2023

**ABSTRAK**

Industri *video game global* terus meroket dengan partisipasi pemain yang meningkat setiap tahun. *Steam*, sebagai *platform* distribusi digital terbesar yang dikembangkan oleh *Valve Corporation*, memungkinkan pengguna memberikan ulasan terbuka tentang *video game*. Prediksi sentimen yang akurat dalam ulasan *online* dapat meningkatkan peluang keuntungan bagi pengembang *video game*. Penelitian ini fokus pada klasifikasi dan pemodelan topik ulasan *video game* lokal di *Steam* menggunakan metode *Naïve Bayes* dan *Latent Dirichlet Allocation*. Sebelum penyeimbangan data, tingkat akurasi klasifikasi *Naïve Bayes* mencapai 86%, dan setelah penyeimbangan data, turun menjadi 81%. Untuk pemodelan topik, penelitian ini mendapatkan 5 topik dengan nilai probabilitas 0.38807 untuk ulasan "*Recommended*" yaitu topik 1 membahas fitur *video game* stardew valley, topik 2 membahas *puzzle* dengan tema *fantasy*, topik 3 membahas *visual art* dan *soundtrack*, topik 4 membahas *update patch* disertai *puzzle* dengan musik, dan topik 5 membahas karakter dan *gameplay*, sedangkan 3 topik dengan nilai probabilitas 0.28095 untuk ulasan "*Not Recommended*" yaitu topik 1 membahas masalah *bug gameplay*, topik 2 membahas *bug boss battle*, dan topik 3 membahas masalah performa *video game*.

**Kata Kunci:** *Video Game*, Klasifikasi *Naïve Bayes*, Pemodelan Topik, *Latent Dirichlet Allocation* (LDA).

***ABSTRACT***

*The global video game industry continues to skyrocket with player participation increasing every year. Steam, as the largest digital distribution platform developed by Valve Corporation, allows users to provide open reviews of video games. Accurate sentiment prediction in online reviews can increase profit opportunities for video game developers. This research focuses on the classification and topic modeling of local video game reviews on Steam using Naïve Bayes and Latent Dirichlet Allocation methods. Before data balancing, the Naïve Bayes classification accuracy rate reached 86%, and after data balancing, it dropped to 81%. For topic modeling, this study obtained 5 topics with a probability value of 0.38807 for "Recommended" reviews, namely topic 1 discussing stardew*

*valley video game features, topic 2 discussing puzzles with fantasy themes, topic 3 discussing visual art and soundtracks, topic 4 discussing patch updates accompanied by puzzles with music, and topic 5 discussing characters and gameplay, while 3 topics with a probability value of 0.28095 for "Not Recommended" reviews, namely topic 1 discussing gameplay bug problems, topic 2 discussing boss battle bugs, and topic 3 discussing video game performance problems.*

***Keywords:*** *Video Game*, Klasifikasi *Naïve Bayes*, *Topic Modeling, Latent Dirichlet Allocation* (LDA)

## 1. Pendahuluan

*Ernst and Young* melaporkan kemunculan industri ekonomi kreatif, yang menghasilkan $2.250 miliar pada tahun 2015, menyumbang 3% dari PDB global dan menciptakan 29,5 juta lapangan kerja baru pada tahun 2013 [1]. Di sisi lain, menurut Badan Ekonomi Kreatif Indonesia, sektor ekonomi kreatif di Indonesia mencapai Rp 1.153,4 triliun pada 2019, yang setara dengan 7,28% dari PDB nasional pada tahun tersebut [2]. Di Indonesia, ekonomi kreatif terdiri dari 17 subsektor, salah satunya adalah subsektor "Aplikasi dan *Game Developer*". Subsektor ini telah diidentifikasi memiliki potensi bisnis yang tinggi untuk ekonomi global. Pada tahun 2020, sub-sektor ini menyumbang Rp 24,88 triliun terhadap PDB nasional, atau sekitar 2,19%, dan berada di urutan ketujuh di antara sub-sektor ekonomi kreatif lainnya dalam hal tingkat pertumbuhan PDB [3]. Valuasi pasar video game global untuk subsektor ini diperkirakan mencapai USD 184,4 juta pada tahun 2022 [4].

Beberapa penelitian sebelumnya yang digunakan untuk referensi. Penelitian pertama ialah jurnal oleh Khofifah [5]. Dari hasil penelitian menunjukkan bahwa 2 dari 5 pantai mendapat *review* negatif yaitu pantai Cibendo mendapat 0,550 dan pantai Tanjung Baru 0,650. dan 3 pantai lainnya mendapat lebih banyak *review* positif. Penelitian kedua yaitu jurnal oleh Ramadhan et al [6] dari Universitas Telkom. Dari hasil penelitian menunjukkan bahwa sentimen positif dominan sebesar 69.8% dengan akurasi algoritma 75.45%. Penelitian ini juga menunjukkan bahwa "*story*", "*character*", "*music*", dan "*art*" termasuk istilah yang sering muncul di antara topik-topik dominan. Penelitian ketiga yaitu jurnal oleh Rashif et al [7] dari Institut Teknologi Sepuluh Nopember. Dari hasil penelitian menunjukkan bahwa lima topik teratas antara lain tentang kondisi dan dampak pandemi saat ini, himbauan untuk menjaga jarak agar Kesehatan tetap terjaga, perkembangan penyebaran Covid-19 yang ada di Indonesia, vaksinasi yang terjadi di beberapa wilayah di Indonesia, dan cara menghadapi Covid-19.

Pada penelitian ini dilakukan klasifikasi ulasan dan pemodelan topik untuk data ulasan *video game* lokal pada *platform Steam* dengan menggunakan metode klasifikasi *Naïve Bayes* dan *Latent Dirichlet Allocation*. Penelitian ini memiliki perbedaan dari penelitian sebelumnya dengan fokus pada *video game* lokal. Tujuan penelitian ini adalah mengidentifikasi emosi dan tema yang dominan dalam ulasan *video game*. Hal ini menjadi penting karena membantu pengembang memahami preferensi pemain, mendapatkan masukan konstruktif, dan memastikan pengembangan produk yang lebih baik sesuai dengan kebutuhan pasar. Oleh karena itu, peneliti membuat penelitian yang berjudul "Implementasi Klasifikasi *Naïve Bayes* dan Pemodelan Topik dengan *Latent Dirichlet Allocation* untuk Data Ulasan *Video Game* Lokal Pada *Platform Steam*".

## 2. Metodologi Penelitian

### 2.1. Klasifikasi *Naïve Bayes*

Klasifikasi *Naive Bayes* adalah model statistik yang didasarkan pada teorema Bayes. Teorema Bayes menyatakan bahwa ada atau tidaknya suatu karakteristik dalam suatu kelas tidak bergantung pada ada atau tidaknya karakteristik lainnya. Klasifikasi *Naive Bayes* adalah algoritma *supervised learning* yang dapat dilatih secara efisien sebagai model probabilistik. Estimasi parameter model *Naive Bayes* dapat dilakukan dengan menggunakan metode maksimum *likelihood*. Model *Naive Bayes* dapat digunakan tanpa probabilitas Bayesian atau metode Bayesian lainnya [8]. Tahapan-tahapan algoritma klasifikasi *Naïve Bayes* sebagai berikut [9].

1. Hitung probabilitas bersyarat/*likelihood*:
$$P(x_i|C) = \frac{P(x)P(C|x)}{P(x_1)P(C|x_1) + P(x_2)P(C|x_2) + \cdots + P(x_n)P(C|x_n)} \quad (1)$$
dengan $C$ = kelas; $x$ = vektor dari nilai atribut *n*
2. Hitung *probabilitas prior* untuk tiap kelas:
$$P(C) = \frac{N_j}{N} \quad (2)$$
dengan $N_j$ = jumlah dokumen pada suatu kelas; $N$ = jumlah total dokumen
3. Hitung probabilitas *posterior* dengan rumus:
$$P(C|x) = \frac{P(x|c)P(C)}{P(x)} \quad (3)$$

### 2.2. *Latent Dirichlet Allocation* (LDA)

*Latent Dirichlet Allocation* adalah sebuah model untuk topik yang bersifat probabilistik dan generatif. Dalam model ini, setiap dokumen dianggap sebagai campuran acak dari topik laten dan setiap topik direpresentasikan sebagai distribusi atas sekumpulan kata yang tetap. Tujuan utama dari LDA adalah untuk menemukan struktur topik laten yang mendasari dari data yang diamati. Dalam LDA, kata-kata dari sebuah dokumen dianggap sebagai data yang diamati. Untuk setiap dokumen dalam korpus, kata-kata dihasilkan. Kata-kata tersebut merupakan bagian dari kosakata, yang diindeks oleh $\{1, \ldots, V\}$, membentuk sebuah urutan dari N kata $W = (W_1, W_2, \ldots, W_n)$, dan korpus terdiri dari kumpulan M dokumen yang direpresentasikan sebagai $D = (W_1, W_2, \ldots, W_n)$ [10].

$$P(W, Z, \theta, \varphi|\alpha, \beta) = P(\varphi|\beta)P(\theta|\alpha)P(Z|\theta)P(W|\varphi_k) \quad (4)$$

dengan jumlah topik dilambangkan dengan K, sedangkan M merupakan banyaknya dokumen yang ada dalam korpus. N adalah banyaknya kata dalam dokumen tertentu, dan $\alpha$ merupakan probabilitas pada distribusi topik per dokumen. Sementara itu, $\beta$ merupakan probabilitas pada distribusi kata per topik. Distribusi untuk dokumen *d* disebut dengan $\theta$, sedangkan $\varphi$ adalah distribusi kata untuk topik *k*. Z merupakan topik untuk kata yang ke-*n* dari suatu dokumen, dan W merupakan kata khusus yang terdapat dalam korpus.

### 2.3. *Collapsed Gibbs Sampling* (CGS)

Algoritma *Collapsed Gibbs Sampling* adalah sebuah metode untuk melakukan *Latent Dirichlet Allocation* yang diperkenalkan oleh Griffiths dan Stevyers. Tidak seperti pendekatan lainnya, CGS hanya mempertimbangkan variabel tugas laten Z, mengabaikan parameter $\varphi$ dan $\theta$. Hasilnya, CGS merupakan algoritma *Markov Chain Monte Carlo* (MCMC) yang jauh lebih sederhana, tetapi juga lebih efektif dan lebih cepat daripada algoritma pengambilan sampel *Gibbs* yang naif. Untuk membuat pengambilan sampel lebih efisien, algoritma CGS menggunakan beberapa matriks hitungan. *Gibbs sampling* adalah sebuah teknik yang digunakan untuk mengestimasi $P(Z|W)$. Dimulai dengan menginisialisasi topik secara acak untuk setiap kata dan kemudian mengambil sampel Z

untuk semua kata dari distribusi kata dengan topik dari semua dokumen dan topik dengan dokumen. Terakhir, teknik ini menetapkan topik yang disampel untuk setiap kata [11].

$$\theta_{dw,k} = \frac{n^{w}_{w,k+\beta}}{n_{w,k} + W\beta} \frac{n^{dw}_{w,k+\alpha}}{n^{dw}_{w+K\alpha}} \tag{5}$$

dengan $\theta_{dw,k} = P(Z|W)$ atau probabilitas kata terhadap topik; $n^{w}_{w,k}$ = banyak kata *w* di *assign* ke topik di setiap dokumen; $\beta$ = *dirichlet* parameter atas distribusi kata terhadap topik di *corpus;* $n^{dw}_{w,k}$ = banyak topik $k$ di *assign* ke dokumen $d$; $\alpha$ = *dirichlet* parameter atas distribusi topik terhadap dokumen; $n_{w,k}$= banyak kata selain *w* di *assign* ke topik $k$ di setiap dokumen; $W$ = jumlah variasi kata di dalam *corpus;* $n^{dw}_{w}$ = banyak topik selain *k* di *assign* ke $d$.

## 2.4. *Topic Coherence*

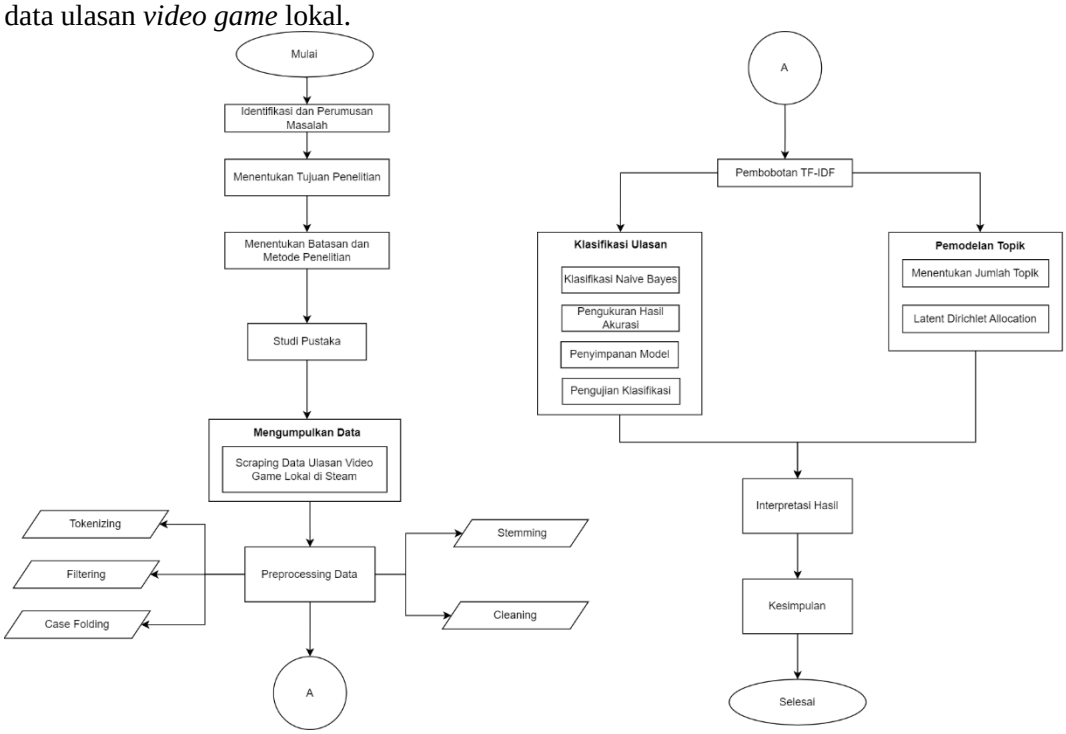
*Topic coherence* adalah matriks yang digunakan untuk mengevaluasi kualitas pemodelan topik. Matriks ini mengevaluasi sebuah topik dengan mengukur seberapa mirip secara semantik kata-kata dengan skor tertinggi dalam topik tersebut [12]. Perhitungan untuk nilai *coherence* diberikan pada persamaan (6) [13].

$$coherence\ (V) = \sum_{(v_i, v_j)\epsilon v} score(v_i, v_j, \epsilon) \tag{6}$$

dengan $V$ = himpunan kata yang menjelaskan topik; $\epsilon$ = faktor pemulusan yang menjamin bahwa skor mengembalikan bilangan asli. Kemudian skor dihitung dengan matriks UMass dengan persamaan (7).

$$score\ (v_i, v_j, \epsilon) = log \frac{D(v_i, v_j) + \epsilon}{D(v_j)} \tag{7}$$

dengan $(v_i, v_j)$ = menghitung jumlah dokumen yang mengandung kata *i* dan *j*; $D(v_j)$ = menghitung jumlah dokumen yang mengandung *j*.Data yang digunakan dalam penelitian ini yaitu data sekunder yang diambil dari metode text mining pada *platform Steam* yaitu data ulasan *video game* lokal.




**Gambar 1.** Diagram Alir Penelitian

## 3. Hasil dan Pembahasan

### 3.1. TF-IDF (*Term Frequency – Inverse Document Frequency*)

Data dilakukan pembobotan TF-IDF untuk menentukan signifikansi sebuah kata dalam koleksi dokumen yang dianalisis. Nilai TF-IDF sebuah kata dalam dokumen dihitung dengan mengalikan dua matriks. *Term Frequency* (TF), yang merupakan frekuensi kata dalam dokumen, dan *Inverse Document Frequency* (IDF), yang merupakan kebalikan dari frekuensi kata di seluruh dokumen.

**Tabel 1** TF-IDF

| Token | TF | | | | | DF | D/DF | IDF | TF-IDF |
|---|---|---|---|---|---|---|---|---|---|
| | D0 | D1 | D2 | D3 | D4 | | | | |
| although | | | 1 | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| appeal | | | 1 | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| aspect | | | 1 | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| atmosphere | | | | | 1 | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| awesomely | | | 1 | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| best | | | 1 | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| category | | 1 | | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| story | | | | | 1 | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| survival | | | 1 | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| work | | 1 | | | | 1 | 5/1=5 | Ln(5)+1=2,609 | 2,609 |
| write | 2 | | | | | 2 | 5/2=2,5 | Ln(2,5)+2=1,916 | 3,833 |

Pada **Tabel 1** dilakukan perhitungan D/DF terlebih dahulu untuk dapat mencari nilai IDF. Contohnya untuk hasil IDF dari kata "*write*" dihitung dengan D = 2,5 yang artinya banyaknya dokumen yaitu 5 dibagi dengan DF yaitu 2 sehingga didapatkan nilai IDF yaitu 1,916. Lalu untuk nilai TF-IDF dihitung dari hasil perkalian IDF dengan D/DF sehingga menghasilkan TF-IDF sebesar 3,833.

### 3.2. Klasifkasi *Naive Bayes*

Penelitian ini menggunakan metode *confusion matrix* untuk mengevaluasi keakuratan klasifikasi dalam proses evaluasi. Metode ini merupakan alat evaluasi yang penting dalam pembelajaran mesin dan biasanya digunakan untuk situasi yang melibatkan dua atau lebih kategori [14]. Setiap elemen dalam matriks merepresentasikan jumlah contoh uji untuk kelas aktual, yang ditampilkan sebagai baris, sedangkan kelas yang diprediksi direpresentasikan dalam kolom. Terlihat pada **Tabel 2** merupakan hasil *confusion matrix* dari prediksi dua kelas sentimen yaitu kelas "*Recommended"* dan "*Not Recommended"* dengan nilai akurasi, *precision*, dan *recall*. Sebagai ilustrasi, dalam ulasan "*This game has amazing features and captivating puzzles*" *Naive Bayes* akan menghitung probabilitas kata-kata seperti "*amazing*" dan "*captivating*" muncul dalam kategori "*Recommended*" dan dengan demikian, memberikan prediksi klasifikasi sebagai "*Recommended*" dengan probabilitas tertinggi.

**Tabel 2** Hasil Confusion Matrix

| Prediksi | Aktual | | Precision |
|---|---|---|---|
| | Recommended | Not Recommended | |
| Recommended | 2106 | 341 | 86% |
| Not Recommended | 0 | 0 | 0% |
| Recall | 100% | 100% | |
| Akurasi | | | |
| 86% | | | |

Berdasarkan **Tabel 2**, dengan menggunakan metode klasifikasi *Naïve Bayes* diperoleh hasil prediksi bahwa pada kelas "*Recommended*" terdapat 2.106 ulasan yang

diklasifikasikan dengan benar dari total jumlah ulasan "*Recommended*", dan tidak terdapat kesalahan prediksi yang dilakukan dalam prediksi ulasan sebagai "*Not Recommended*". Dengan demikian, nilai *precision* untuk kelas "*Recommended*" adalah 86%. Di sisi lain, untuk kelas "*Not Recommended*" tidak terdapat ulasan diklasifikasikan dengan benar sebagai "*Not Recommended*" dari total jumlah ulasan, dan terdapat 341 ulasan kesalahan prediksi yang dilakukan dalam mengkategorikan ulasan sebagai "*Recommended*". Oleh karena itu, nilai *precision* untuk kelas "*Not Recommended* " adalah 0%. Nilai *confusion matrix* menunjukkan tingkat akurasi sebesar 86%, yang berarti bahwa dari 2.447 ulasan yang diuji, 2.106 ulasan diklasifikasikan dengan benar oleh model klasifikasi *Naïve Bayes*. Hasil klasifikasi memiliki nilai 0 terjadi karena data tidak sama banyak atau proporsional sehingga gagal dalam memprediksi. Selanjutnya dilakukan tahapan untuk menyamakan data "*Recommended*" dan "*Not Recommended"*.

Untuk mengatasi jumlah sampel yang tidak proporsional antar kelas dalam data, penyeimbangan data diterapkan. Teknik ini digunakan untuk mengurangi dampak *imbalance* data terhadap kinerja model *machine learning* dan hasil prediksinya. Ketika adanya *imbalance class*, model dapat menjadi bias terhadap kelas mayoritas dan mengabaikan kelas minoritas, yang mengarah pada prediksi yang tidak akurat. Pada penelitian ini, SMOTE (*Synthetic Minority Oversampling Technique*) digunakan untuk menyeimbangkan data dengan meningkatkan jumlah sampel dengan atribut sentimen minoritas. Karena data dengan kelas "*Recommended*" berjumlah lebih banyak daripada data dengan kelas "*Not Recommended*", maka SMOTE digunakan untuk menyeimbangkan kedua kelas tersebut dengan cara menambahkan sejumlah data dengan kelas "*Not Recommended*" secara acak untuk menyamai jumlah kelas "*Recommended*". Terlihat pada **Tabel 3** merupakan hasil *confusion matrix* setelah dilakukan SMOTE dari prediksi dua kelas sentimen yaitu kelas "*Recommended*" dan "*Not Recommended*" dengan nilai akurasi, *precision*, dan *recall.* Meskipun data belum sepenuhnya seimbang, model telah mampu memberikan hasil klasifikasi yang baik.

**Tabel 3** Hasil Confusion Matrix Setelah SMOTE

| Prediksi | Aktual | | Precision |
|---|---|---|---|
| | Recommended | Not Recommended | |
| Recommended | 1693 | 49 | 97% |
| Not Recommended | 411 | 294 | 42% |
| Recall | 80% | 86% | |
| Akurasi | | | |
| 81% | | | |

Berdasarkan **Tabel 3**, dengan menggunakan metode klasifikasi *Naïve Bayes* dengan SMOTE diperoleh hasil prediksi bahwa pada kelas "*Recommended*", terdapat 1693 ulasan yang diklasifikasikan dengan benar dari total jumlah ulasan "*Recommended*", dan terdapat 411 ulasan kesalahan prediksi yang dilakukan dalam mengkategorikan ulasan sebagai "*Not Recommended*". Dengan demikian, nilai *precision* untuk kelas "*Recommended* " adalah 97%. Di sisi lain, untuk kelas "*Not Recommended*" terdapat 294 ulasan diklasifikasikan dengan benar sebagai "*Not Recommended*" dari total jumlah ulasan, dan terdapat 49 ulasan kesalahan prediksi yang dilakukan dalam mengkategorikan ulasan sebagai "*Recommended*". Oleh karena itu, nilai *precision* untuk kelas "*Not Recommended*" adalah 42%. Nilai *confusion matrix* menunjukkan tingkat akurasi sebesar 81%, yang berarti bahwa dari 2.447 ulasan yang diuji, 1.987 ulasan diklasifikasikan dengan benar oleh model klasifikasi *Naïve Bayes.*

### 3.3. Pemodelan Topik dengan *Latent Dirichlet Allocation*

Metode LDA dalam pemodelan topik melibatkan pengklasifikasian atau pengorganisasian teks ke dalam dokumen dan kata-kata per topik. Pada penelitian ini, pemodelan topik dilakukan pada ulasan "*Recommended*" *video game* lokal dari *platform Steam*, untuk mengkategorikannya ke dalam beberapa topik dan menentukan topik yang paling sering muncul terkait *video game*. **Tabel 4** menampilkan *coherence value* ulasan "*Recommended*".

**Tabel 4** *Coherence Value* Ulasan *Recommended*

| *Num Value* | Coherence Value | *Num Value* | Coherence Value |
|---|---|---|---|
| 1 | 0.3228 | 6 | 0.35582 |
| 2 | 0.31777 | 7 | 0.35728 |
| 3 | 0.31208 | 8 | 0.35804 |
| 4 | 0.38275 | 9 | 0.33804 |
| 5 | 0.38807 | 10 | 0.36896 |

Pada **Tabel 4**, pemodelan topik dilakukan dengan 5 topik yang memiliki probabilitas tertinggi yaitu 0. 38807, dan akan digunakan sebagai acuan. Angka-angka dan kata-kata tersebut merupakan hasil dari pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation* (LDA). Angka di sebelah kata menunjukkan bobot relatif dari masing-masing kata dalam membentuk topik. Pemahaman dari hasil ini memberikan kesimpulan tentang topik.

**Tabel 5** Hasil Model Topik Ulasan *Recommended*

| Topik | Model | Kesimpulan |
|---|---|---|
| 1 | 0.009*"stardew" + 0.008*"valley" + 0.006*"moon" + 0.006*"quest" + 0.006*"fun" + 0.005*"time" + 0.004*"feel" + 0.004*"cute" + 0.004*"character" + 0.004*"combat" | Fitur *video game* stardew valley |
| 2 | 0.009*"stardew" + 0.008*"valley" + 0.006*"moon" + 0.006*"quest" + 0.006*"fun" + 0.005*"time" + 0.004*"feel" + 0.004*"cute" + 0.004*"character" + 0.004*"combat" | *Puzzle* dengan tema *fantasy* |
| 3 | 0.037*"fun" + 0.027*"cute" + 0.015*"relax" + 0.010*"lovely" + 0.009*"music" + 0.009*"art" + 0.008*"beautiful" + 0.007*"super" + 0.007*"simple" + 0.006*"animal" | *Visual art* dan *soundtrack* |
| 4 | 0.018*"beautiful" + 0.013*"awesome" + 0.009*"art" + 0.009*"puzzle" + 0.007*" *patch* addict" + 0.006*"short" + 0.005*"music" + 0.005*"sweet" + 0.005*"grind" + 0.005*"beautifully" | *Update* disertai *puzzle* dengan musik |
| 5 | 0.096*"good" + 0.029*"wife" + 0.022*"cat" + 0.021*"amaze" + 0.018*"cry" + 0.012*"release" + 0.011*"meow" + 0.008*"grind" + 0.007*"harvest" + 0.007*"moon" | Karakter dan *gameplay* |

Pada **Tabel 5** didapatkan kesimpulan bahwa topik 1 membahas tentang fitur *video game* stardew valley, topik 2 membahas tentang *puzzle* dengan tema *fantasy*, topik 3 membahas tentang *visual art* dan *soundtrack*, topik 4 tentang membahas *update patch* disertai *puzzle* dengan musik, dan topik 5 tentang membahas karakter dan *gameplay*. Selanjutnya, pemodelan topik dilakukan pada ulasan "*Not Recommended*" *video game* lokal dari *platform Steam*, untuk mengkategorikannya ke dalam beberapa topik dan menentukan topik yang paling sering muncul terkait *video game*. **Tabel 6** menampilkan *coherence value* ulasan "*Not Recommended*".

**Tabel 6** *Coherence Value* Ulasan *Not Recommended*

| *Num Value* | Coherence Value | *Num Value* | Coherence Value |
|---|---|---|---|
| 1 | 0.26333 | 6 | 0.22441 |
| 2 | 0.24879 | 7 | 0.24007 |
| 3 | 0.28095 | 8 | 0.23875 |
| 4 | 0.26039 | 9 | 0.24684 |
| 5 | 0.23158 | 10 | 0.24086 |

Pada **Tabel 6**, pemodelan topik dilakukan dengan 3 topik yang memiliki probabilitas tertinggi yaitu 0. 28095, dan akan digunakan sebagai acuan.

**Tabel 7** Hasil Model Topik Ulasan *Not Recommended*

| Topik | Model | Kesimpulan |
|---|---|---|
| 1 | 0.003*"bug" + 0.003*"gameplay" + 0.003*"control" + 0.003*"boring" + 0.003*"ghost" + 0.002*"refund" + 0.002*"puzzle" + 0.002*"graphic" + 0.002*"point" + 0.002*"review"" | Masalah *bug gameplay* |
| 2 | 0.004*"bug" + 0.003*"bos" + 0.003*"buy" + 0.003*"horror" + 0.003*"load" + 0.002*"die" + 0.002*"fight" + 0.002*"ghost" + 0.002*"mechanic" + 0.002*"act" | Masalah *bug boss battle* |
| 3 | 0.003*"card" + 0.003*"horror" + 0.003*"combat" + 0.003*"lack" + 0.003*"buy" + 0.002*"gameplay" + 0.002*"hard" + 0.002*"bug" + 0.002*"dreadout" + 0.002*"money" | Masalah performa *video game* |

Pada **Tabel 7** didapatkan kesimpulan bahwa topik 1 membahas tentang masalah *bug gameplay*, topik 2 membahas tentang *bug boss battle*, dan topik 3 tentang membahas masalah performa *video game*.

## 4. Kesimpulan

Berdasarkan hasil analisis menggunakan metode *Naïve Bayes*, terlihat bahwa model dapat mengklasifikasikan ulasan *video game* lokal dengan tingkat akurasi sebesar 86%, khususnya dalam mengidentifikasi kategori "*Recommended*", tetapi tidak mampu mengidentifikasi kategori "*Not Recommended*". Meskipun terjadi penurunan akurasi menjadi 81% setelah penerapan SMOTE, model mampu dalam mengklasifikasikan data untuk kedua kategori, "*Recommended*" dan "*Not Recommended*", dengan 1.982 data yang benar diklasifikasikan. Selain itu, melalui analisis pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA), ditemukan 5 topik utama dengan pada ulasan "*Recommended*" dengan probabilitas yaitu 0.38807. Topik 1 membahas fitur *video game* stardew valley, topik 2 membahas *puzzle* dengan tema *fantasy*, topik 3 membahas *visual art* dan *soundtrack*, topik 4 membahas *update patch* disertai *puzzle* dengan musik, dan topik 5 membahas karakter dan *gameplay*. Lalu, 3 topik pada ulasan "*Not Recommended*" dengan probabilitas yaitu 0.28095. Topik 1 membahas masalah *bug gameplay*, topik 2 membahas *bug boss battle*, dan topik 3 membahas masalah performa *video game*. Hasil ini menunjukkan kemampuan model dalam mengidentifikasi dan memodelkan topik-topik utama yang mendominasi dalam ulasan *video game* lokal, memberikan wawasan yang berharga terkait preferensi dan tema yang muncul dari perspektif pemain.

## 5. Daftar Pustaka

[1] Ernst and Young, Cultural times The first global map of cultural and creative industries, EY, 2015.

[2] Kementerian Pariwisata dan Ekonomi Kreatif RI, Cetak Biru Ekonomi Kreatif Indonesia Menuju 2025, Kementerian Pariwisata dan Ekonomi Kreatif, 2014.

[3] D. Novianty and D. Prastya, "Menparekraf: Game dan Aplikasi Sumbang Rp 24,88 Triliun PDB Indonesia," 17 Agustus 2021. [Online]. Available: https://www.suara.com/tekno/2021/08/17/114950/menparekraf-game-dan-aplikasi-sumbang-rp-2488-triliun-pdb-indonesia?page=1.

[4] S. Totilo, "Report: Video game revenue shrinks in 2022, snapping growth streak," 16 November 2022. [Online]. Available: https://www.axios.com/2022/11/15/global-video-game-market-decline-2022.

[5] W. Khofifah, D. N. Rahayu and A. M. Yusuf, "Analisis Sentimen Menggunakan Naive Bayes untuk Melihat Review Masyarakat Terhadap Tempat Wisata Pantai di Kabupaten Karawang pada Ulasan Google Maps," *Jurnal Interkom: Jurnal Publikasi Ilmiah Bidang Tenologi Informasi dan Komunikasi*, pp. 28-38, 2022.

[6] S. R. Ramadhan, M. Y. Febrianta and S. Widiyanesti, "Analisis Ulasan Indie Video Game Lokal pada Steam Menggunakan Analisis Sentimen dan Pemodelan Topik Berbasis Latent Dirichlet Allocation," *Journal of Animation & Games Studies*, pp. 117-144, 2021.

[7] F. Rashif, G. I. P. Nirvana, M. A. Noor and N. A. Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Cuitan Akun Bot Twitter bertagar #Covid-19," *Cogito Smart Journal*, pp. 170-181, 2021.

[8] L. Wilianto, T. H. Pudhiantoro and F. R. Umbara, "Analisis Sentimen Terhadap Tempat Wisata dari Komentar Pengunjung dengan Menggunakan Metode Naive Bayes Classifier Studi Kasus Jawa Barat," *Prosiding SNATIF*, pp. 439-448, 2017.

[9] D. A. Muthia, "Opinion Mining Pada Review Produk Kecantikan Menggunakan Algoritma Naive Bayes," *Jurnal Sistem Iinformasi STMIK Antar Bangsa*, pp. 21-26, 2018.

[10] M. Habibi, A. Priadana, A. B. Saputra and P. W. Cahyo, "Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA)," in *Conference: International Conference on Health and Medical Sciences (AHMS 2020)*, 2021.

[11] U. T. Setijohatmo, S. Rachmat, T. Susilawati and Y. Rahman, "Analisa Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik," *Prosiding The 11th Industrial Research Workshop and National Seminar* , pp. 402-408, 2020.

[12] F. Tang, "Beginner's Guide to LDA Topic Modelling with R," 14 July 2019. [Online]. Available: https://towardsdatascience.com/beginners-guide-to-lda-topic-modelling-with-r-e57a5a8e7a25.

[13] K. Stevens, P. Kegelmeyer, D. Andrzejewski and D. Buttler, "Exploring topic coherence over many models and many topics," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952-961, 2012.

[14] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.