**Title:**

* Pendekatan Semantik Dalam Sistem Rekomendasi Produk UMKM Untuk Mengatasi Permasalahan *Cold Start*

**Abstract**

1. **Introduction**

Sistem rekomendasi telah menjadi komponen penting dalam berbagai platform digital modern, termasuk *e-commerce*, layanan media, hingga pariwisata dan kebudayaan. Sistem ini membantu pengguna dalam menyaring informasi dan menemukan produk atau layanan yang relevan berdasarkan preferensi pribadi, sehingga meningkatkan efisiensi interaksi dan kepuasan pengguna secara keseluruhan (Zhou et al., 2023).

Namun demikian, salah satu tantangan utama yang dihadapi oleh sistem rekomendasi adalah permasalahan *cold start*, yaitu kondisi ketika suatu item atau belum memiliki cukup data historis, seperti rating atau interaksi, sehingga sulit untuk diproses oleh algoritma rekomendasi tradisional (Ibrahim et al., 2023). Permasalahan ini menjadi sangat penting terutama pada produk lokal seperti kerajinan tradisional Bali, di mana produk-produk baru sering kali tidak mendapatkan eksposur yang memadai karena belum memiliki riwayat ulasan atau interaksi pengguna.

Kerajinan tradisional Bali sendiri merupakan warisan budaya bernilai tinggi, mencakup produk berbahan kayu, bambu, perak, dan kain, yang berasal dari berbagai daerah seperti Gianyar, Tabanan, dan Karangasem. Produk ini berfungsi sebagai hiasan, perlengkapan upacara, pakaian adat, hingga cinderamata. Namun, transformasi digital belum sepenuhnya mampu mengangkat potensi produk ini, terutama dalam konteks sistem rekomendasi yang masih bergantung pada riwayat data.

Sebagai solusi terhadap tantangan tersebut, penelitian ini mengusulkan sistem rekomendasi berbasis *metadata* dan *Natural Language Processing* (NLP) yang tidak bergantung pada riwayat interaksi pengguna. Dengan pendekatan ini, sistem membentuk representasi semantik produk melalui ekstraksi entitas menggunakan teknik *Named Entity Recognition* (NER) dan transformasi vektor menggunakan TF-IDF, sehingga mampu mengenali kemiripan semantik antar item berdasarkan atribut deskriptif seperti:

* Jenis bahan (misalnya kayu, bambu, perak, kain),
* Asal daerah (Gianyar, Tabanan, Karangasem),
* Fungsi (hiasan, upacara, pakaian, souvenir),
* Gaya visual (tradisional, kontemporer, modern Bali),
* Warna dominan, ukuran, dan harga.

Pendekatan ini sejalan dengan temuan Mishra et al. (2024), yang menunjukkan bahwa pemanfaatan konten deskriptif dan semantik melalui NLP dapat secara signifikan meningkatkan performa sistem rekomendasi dalam situasi data terbatas, seperti kasus cold start. Teknik embedding semantik berbasis TF-IDF memungkinkan sistem untuk mengukur kedekatan antar item dalam ruang vektor, yang kemudian digunakan dalam proses perhitungan *similarity*.

Hasil perhitungan kemiripan semantik antar produk kemudian digunakan untuk memprediksi kemungkinan ketertarikan pengguna terhadap produk baru. Dengan demikian, meskipun produk belum memiliki riwayat interaksi atau rating, sistem tetap dapat memberikan prediksi rekomendasi yang relevan berdasarkan profil semantik item tersebut. Pendekatan ini tidak hanya memperbaiki keterbatasan pada cold start, tetapi juga memberikan peluang lebih besar bagi produk lokal dan tradisional untuk mendapatkan visibilitas dalam platform digital.

1. **Related Work**

Permasalahan *cold start* telah menjadi salah satu tantangan utama dalam pengembangan sistem rekomendasi modern. Xiao et al. (2023) melalui pendekatan UPRec mengemukakan bahwa proses *pre-training* yang mengintegrasikan atribut pengguna secara eksplisit dapat memperkaya representasi pengguna dan mengurangi ketergantungan pada data historis yang terbatas, sehingga mampu mengatasi cold start secara efektif pada skenario *sequential recommendation*.

Pendekatan serupa juga ditinjau secara komprehensif oleh Zhou et al. (2023), yang mengelompokkan metode-metode berbasis deep learning seperti BERT4Rec dan *Graph Neural Network* (GNN) sebagai solusi mutakhir untuk masalah sparsitas dan cold start, melalui pembentukan embedding pengguna dan item yang lebih dalam dan kontekstual.

Dalam ranah berbasis konten dan emosi, Jiang et al. (2020) mengembangkan sistem rekomendasi dengan menggabungkan Word2Vec dan LSTM untuk analisis sentimen, yang menunjukkan bahwa ulasan pendek sekalipun dapat dimanfaatkan untuk membentuk pemodelan emosi pengguna dalam sistem rekomendasi sosial. Pendekatan berbasis NLP juga diadopsi oleh Mishra et al. (2024), yang menggabungkan klasifikasi ulasan menggunakan Random Forest dan model filtrasi berbasis *collaborative filtering*. Penelitian tersebut membuktikan bahwa kualitas ulasan dan analisis teks mampu memperbaiki akurasi rekomendasi meskipun pada kondisi cold start.

Di sisi lain, pendekatan *content-based filtering* yang mengandalkan metadata produk telah banyak dikembangkan untuk domain yang menghadapi keterbatasan interaksi pengguna. Negara et al. (2023) membangun sistem rekomendasi untuk NFT marketplace dengan hanya mengandalkan fitur nama dan deskripsi item berbasis TF-IDF dan cosine similarity, yang terbukti mampu memberikan hasil relevan bahkan tanpa rating pengguna sama sekali.

Pendekatan berbasis metadata juga diusulkan oleh Lestari et al. (2024) dalam konteks marketplace UMKM. Mereka menggabungkan teknik *clustering* dan *imputation* untuk mengisi kekosongan data pada produk baru, dan menunjukkan bahwa kombinasi ini mampu meningkatkan akurasi prediksi secara signifikan hingga 100% pada algoritma KNN dan Naïve Bayes.

Penelitian ini mengambil inspirasi dari pendekatan-pendekatan di atas, dengan kontribusi utama pada pemanfaatan metadata dan deskripsi semantik dari produk budaya lokal, yaitu kerajinan tradisional Bali. Tidak seperti domain *e-commerce* konvensional, produk-produk kerajinan Bali memiliki deskripsi yang kaya akan informasi simbolik dan kultural. Dengan mengimplementasikan teknik *Named Entity Recognition* (NER) dan representasi embedding, penelitian ini bertujuan membangun sistem rekomendasi yang mampu mengatasi cold start secara semantik berbasis konten, serta mempertahankan kekayaan budaya dalam konteks digital.

1. **The Proposed Model**

Model yang diusulkan merupakan sistem estimasi rating berbasis konten dengan pendekatan semantik yang dirancang khusus untuk produk kerajinan tradisional Bali. Sistem ini bertujuan mengatasi permasalahan *cold start* dengan tidak mengandalkan data interaksi pengguna, melainkan pada kekayaan informasi deskriptif (metadata) yang tersedia dari produk.

Sistem ini terdiri dari beberapa komponen utama:

* Ekstraksi Metadata

Sistem dimulai dengan mengekstraksi atribut penting dari produk kerajinan, seperti bahan (kayu, bambu, perak, kain), asal daerah (Gianyar, Tabanan, Karangasem), fungsi (hiasan, upacara, pakaian, souvenir), serta atribut lain seperti warna dominan, ukuran, dan harga. Informasi ini mencerminkan karakteristik unik dari kerajinan tradisional Bali yang menjadi dasar untuk proses semantik.

* *Named Entity Recognition* (NER)

Untuk meningkatkan pemahaman terhadap konteks deskripsi produk, sistem menggunakan teknik NLP berupa *Named Entity Recognition* (NER). Teknik ini memungkinkan sistem mengidentifikasi entitas penting dalam teks deskriptif produk, seperti nama daerah, bahan, dan kegunaan, yang sering kali tidak ditulis dalam struktur metadata formal.

* Representasi Vektor (TF-IDF)

Metadata dan entitas hasil ekstraksi kemudian dikonversi menjadi representasi vektor menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF). Representasi ini bertujuan untuk menangkap bobot pentingnya kata-kata dalam konteks semantik antar produk. TF-IDF memungkinkan sistem memahami kesamaan makna meskipun struktur kata berbeda, sebagaimana dijelaskan dalam pendekatan NLP modern (Mishra et al., 2024)

* *Similarity Computation* (*Cosine Similarity*)

Setelah data produk direpresentasikan dalam bentuk vektor, sistem menghitung *cosine similarity* untuk mengukur derajat kemiripan antara produk baru dan produk populer. Dengan metode ini, sistem dapat mengidentifikasi produk-produk yang memiliki kedekatan semantik tinggi meskipun belum memiliki interaksi dari pengguna.

* Prediksi Rating (*Rating Predicted*)

Berdasarkan nilai kemiripan tersebut, sistem memperkirakan rating untuk produk baru dengan merujuk pada produk populer yang mirip secara semantik. Prediksi ini memungkinkan sistem memberikan rekomendasi yang relevan, meskipun tanpa data interaksi historis, sehingga meningkatkan eksposur produk kerajinan baru.

1. **Model Analysis**
2. **Item Classification**

Sebagaimana dijelaskan pada bagian sebelumnya, sistem yang diusulkan bertujuan untuk memperkirakan rating awal terhadap produk-produk baru yang belum memiliki riwayat interaksi pengguna. Untuk mencapai tujuan tersebut, langkah awal yang dilakukan adalah mengelompokkan produk baru ke dalam kategori semantik tertentu berdasarkan data deskriptif (metadata) yang tersedia. Proses ini disebut sebagai *item classification*, dan dilakukan dengan menerapkan algoritma klasifikasi berbasis pembelajaran mesin terhadap fitur metadata yang telah dikonversi dalam bentuk vektor.

Dalam sistem ini, klasifikasi item digunakan untuk menentukan kelas semantik dari suatu produk, seperti *hiasan kayu tradisional dari Gianyar*, *aksesoris perak kontemporer dari Karangasem*, atau *souvenir bambu modern dari Tabanan*. Pengelompokan ini penting agar produk baru dapat dicocokkan dengan kelompok produk populer yang serupa secara semantik, yang nantinya menjadi referensi dalam estimasi rating.

Untuk mendukung klasifikasi multikelas, digunakan pendekatan *one-vs-all* (OvA) (Milgram, Cheriet, & Sabourin, 2006), yang mengubah permasalahan multikelas menjadi sejumlah model klasifikasi biner. Dalam pendekatan ini, setiap model bertugas membedakan satu kelas semantik terhadap seluruh kelas lainnya. Proses pelatihan dilakukan dengan data produk historis yang telah diberi label kelas berdasarkan analisis domain (bahan, fungsi, daerah, gaya), sementara prediksi dilakukan terhadap item baru berdasarkan vektor fitur hasil proses embedding seperti TF-IDF.

Prediksi kelas semantik untuk item baru ditentukan dengan cara memilih model klasifikasi yang menghasilkan skor kepercayaan tertinggi terhadap input , yaitu:

Di mana:

* adalah representasi vektor dari fitur metadata produk ,
* adalah skor klasifikasi untuk kelas ke- ,
* adalah jumlah total kelas semantik.

Untuk keperluan klasifikasi, sistem menggunakan dua algoritma populer: C4.5 (Kotsiantis, 2007) dan Naive Bayes (Zhang, 2004), yang telah terbukti efisien dalam menangani data deskriptif dengan jumlah fitur terbatas.

**Algoritma 1. One-vs-All untuk Klasifikasi Item**

Input:

* : Algoritma pelatihan (misal: C4.5 atau Naive Bayes)
* : Dataset fitur (metadata vektor)
* : Label kelas semantik dari item pelatihan

Output:

* Model klasifikasi biner ​ untuk setiap kelas semantik

Pseudocode:

|  |
| --- |
| Begin  For each class k ∈ {1, 2, ..., K} do  For each item i in training set do  If y\_i = k then  y′\_i ← 1  Else  y′\_i ← 0  End for  Train classifier f\_k ← L(X, y′)  End for  End |

1. **Item similarity**

Setelah produk baru berhasil diklasifikasikan ke dalam kategori semantik tertentu, sistem melanjutkan ke tahap identifikasi produk-produk referensi yang serupa. Tujuan utama dari tahap ini adalah menghitung tingkat kemiripan semantik antara produk baru dengan produk yang telah memiliki riwayat rating, untuk digunakan sebagai dasar dalam proses estimasi rating awal.

Proses ini melibatkan perbandingan atribut metadata yang telah dikonversi menjadi representasi vektor, menggunakan teknik embedding seperti TF-IDF. Diberikan sebuah produk baru, dan himpunan produk populer yang telah memiliki rating historis, sistem menghitung skor kemiripan untuk setiap berdasarkan kombinasi atribut semantik seperti bahan, fungsi, asal daerah, gaya visual, dan atribut numerik seperti ukuran dan harga.

Skor kemiripan dihitung dengan rumus:

dengan:

* S: nilai kesamaan atribut ke- antara produk dan ,
* : bobot pentingnya atribut ke-,
* : jumlah total atribut yang digunakan dalam perhitungan.

**Algoritma 2. Estimasi Kemiripan Antar Produk**

Input:

* ​: Himpunan produk baru
* : Himpunan produk populer (memiliki rating historis)
* : Daftar atribut metadata
* : Bobot tiap atribut

Output:

* ): Himpunan produk tetangga paling mirip untuk setiap produk baru

Pseudocode:

|  |
| --- |
| Begin  For each product p ∈ P\_baru do  NG(p) ← ∅  For each product q ∈ Q do  total\_weight ← 0  sim\_score ← 0  For each attribute dj ∈ D do  SFj ← Similarity(p\_j, q\_j)  sim\_score ← sim\_score + (SFj × wj)  total\_weight ← total\_weight + wj  End for  final\_similarity ← sim\_score / total\_weight  If final\_similarity ≥ threshold then  NG(p) ← NG(p) ∪ {q}  End if  End for  End for  End |

1. **Ratings prediction**

Tahap akhir dalam proses rekomendasi adalah estimasi nilai rating awal untuk produk baru berdasarkan kemiripan semantik terhadap produk-produk populer yang telah memiliki rating historis. Mengingat tidak tersedianya data interaksi pada kondisi cold start, sistem menggunakan pendekatan content-based filtering untuk menghitung prediksi berdasarkan kemiripan antar item yang telah diperoleh pada tahap sebelumnya.

Prediksi dilakukan dengan menggunakan prinsip *weighted average*, di mana nilai rating produk baru dihitung sebagai rata-rata tertimbang dari rating produk-produk tetangga berdasarkan skor kemiripan semantik. Semakin tinggi kemiripan suatu produk referensi terhadap produk baru, semakin besar kontribusinya terhadap estimasi rating.

Rumus prediksi dirumuskan sebagai berikut:

dengan:

* adalah rating prediksi untuk produk baru p,
* ​ adalah rating historis dari produk referensi q,
* adalah skor kemiripan semantik antara produk p dan q,
* ) adalah himpunan produk mirip untuk item p.

1. **Performance Assessment**

Untuk menilai efektivitas pendekatan semantik yang diusulkan dalam menangani permasalahan cold start, dilakukan serangkaian eksperimen yang dirancang untuk mengukur akurasi sistem dalam memprediksi rating awal produk UMKM yang belum memiliki riwayat interaksi pengguna. Evaluasi difokuskan pada tiga aspek utama: akurasi prediksi (menggunakan MAE dan RMSE), sensitivitas terhadap jumlah data referensi, dan pengaruh variasi bobot atribut metadata terhadap performa sistem.

Eksperimen dilakukan menggunakan dataset produk kerajinan tradisional Bali yang disusun secara semi-sintetik. Setiap item dalam dataset dianotasi dengan atribut metadata seperti bahan, fungsi, daerah asal, gaya visual, ukuran, dan harga. Sebagian item dilengkapi dengan rating historis, sementara sebagian lainnya ditandai sebagai produk baru tanpa rating, yang mensimulasikan kondisi cold start.

1. **Evaluation metrics**

Dua metrik utama digunakan dalam pengukuran akurasi sistem:

1. **Mean Absolute Error (MAE)**

MAE mengukur rata-rata selisih absolut antara rating yang diprediksi dengan rating aktual dari produk:

1. **Root Mean Square Error (RMSE)**

RMSE mengukur akar kuadrat dari selisih rata-rata antara prediksi dan rating aktual:

Kedua metrik digunakan untuk memberikan gambaran yang lebih menyeluruh: MAE memberikan rata-rata kesalahan absolut, sedangkan RMSE memberikan penalti lebih besar pada kesalahan prediksi yang tinggi.

1. **Experimental evaluation**
2. **Experimental scenarios**
3. **Conclusions**

**References**