



Implementasi *Term Frequency - Inverse Document Frequency* dan *Cosine Similarity* untuk Analisis Kemiripan Deskripsi Produk Halal

Jennifer Velensia Santoti¹, Jennifer Jocelyn², Hafiz Irsyad³

^{1,2,3} Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang

¹jennifervelensiasantoti_2226250014@mhs.mdp.ac.id, ²jenniferjocelyn_2226250011@mhs.mdp.ac.id,
³hafizirsyad@mdp.ac.id*

ABSTRACT

In today's digital era, the clarity of product information has become a crucial aspect to support consumer decision-making during the purchasing process. This study focuses on the implementation of feature extraction from product descriptions using the TF-IDF (Term Frequency - Inverse Document Frequency) method and Cosine Similarity to predict confusing product descriptions. The research methodology includes several preprocessing steps such as tokenizing, stopword removal, filtering, removal of null and NaN data, as well as text feature extraction using TF-IDF and Cosine Similarity. Evaluation results show that the system successfully identifies halal products with a precision of 96%, recall of 98%, and an F1-score of 97%, indicating a good balance between precision and recall. For haram products, it achieves a precision of 98%, recall of 95%, and an F1-score of 97%. Overall, the system attains an accuracy of 97%. The evaluation results indicate that the model performs better at recognizing halal products, with a recall of 98%, compared to a recall of 95% for haram products. This suggests that the methods used are highly effective in predicting the clarity of product descriptions. The conclusion of this study confirms that the combination of TF-IDF and Cosine Similarity is effective in identifying ambiguities in product descriptions, thereby improving information transparency for consumers.

Keywords: *Cosine Similarity, feature extraction, halal products, haram products, TF-IDF*

ABSTRAK

Di era digital saat ini, kejelasan informasi produk telah menjadi aspek penting untuk mendukung keputusan konsumen dalam proses pembelian. Penelitian ini difokuskan pada implementasi ekstraksi fitur dari deskripsi produk menggunakan metode TF-IDF (*Term Frequency - Inverse Document Frequency*) dan *Cosine Similarity* untuk memprediksi deskripsi produk yang membingungkan. Metodologi penelitian ini meliputi beberapa tahap preprocessing, yang meliputi tokenizing, stopword removal, filtering, penghapusan data null dan data NaN, serta ekstraksi fitur teks menggunakan metode TF-IDF dan *Cosine Similarity*. Hasil evaluasi menunjukkan bahwa sistem berhasil mengenali produk halal dengan nilai precision sebesar 96%, recall sebesar 98%, dan F1-score sebesar 97%, yang mengindikasikan bahwa adanya keseimbangan yang baik antara precision dan recall. Untuk produk haram mencapai precision sebesar 98%, recall sebesar 95%, dan F1-score sebesar 97%. Secara keseluruhan, sistem berhasil mendapatkan nilai akurasi sebesar 97%. Hasil evaluasi menunjukkan bahwa model lebih baik dalam mengenali produk halal, dengan hasil recall sebesar 98%, sementara hasil recall produk haram sebesar 95%. Hal ini mengindikasikan bahwa metode yang digunakan sangat efektif dalam memprediksi kejelasan deskripsi produk. Kesimpulan dari penelitian ini menegaskan bahwa kombinasi TF-IDF dan *Cosine Similarity* efektif dalam mengidentifikasi ambiguitas deskripsi produk, sehingga dapat meningkatkan transparansi informasi bagi konsumen.

Kata kunci: *Cosine Similarity, ekstraksi fitur, produk halal, produk haram, TF-IDF*

1. PENDAHULUAN

Seiring dengan bertambahnya wawasan masyarakat mengenai produk bersertifikat halal, kejelasan dan konsistensi akan deskripsi pada produk halal di platform e-commerce merupakan hal yang penting. Pemilihan produk halal tidak hanya didasarkan pada label dan tampilannya saja, melainkan juga didasarkan pada deskripsi produk yang jelas dan terjamin akan sertifikasinya. Produk bersertifikasi halal merupakan produk yang digunakan oleh umat agama Islam karena menggunakan aturan agama Islam dalam pengolahannya [1]. Namun di tengah proses ini, masih ditemukan deskripsi produk halal yang kurang jelas dan membingungkan.

Permasalahan ini dapat menimbulkan keraguan bagi konsumen sekaligus berpotensi melanggar ketentuan syariah. Sebagian besar deskripsi produk di e-commerce Indonesia masih menggunakan terminologi kehalalan yang ambigu. Misalnya seperti frasa “bebas zat haram” sering muncul tanpa disertai rincian komposisi yang jelas, sehingga menimbulkan multi tafsir. Fenomena ini diperkuat oleh hasil analisis sentimen yang menunjukkan bahwa konsumen mengeluhkan minimnya transparansi informasi halal pada platform daring [2]. Deskripsi yang ambigu ini tidak hanya merugikan konsumen domestik, tetapi juga berdampak pada persepsi internasional terhadap kredibilitas sertifikasi halal Indonesia [3]. Padahal, sertifikasi halal ini tidak hanya digunakan untuk memastikan kepatuhan proses produksi terhadap aturan syariah agama, tetapi juga memberi jaminan akan kualitas produk, seperti kebersihan, tingkat higienis, serta sikap etis [1].

Kondisi ini semakin diperburuk oleh masuknya produk impor tanpa sertifikasi resmi, sehingga menimbulkan keraguan bagi masyarakat muslim. Keberadaan produk semacam ini memperlihatkan lemahnya pengawasan dan pengelolaan data kehalalan di Indonesia. Tidak hanya itu, ada banyak pelaku usaha lokal yang masih cenderung mengabaikan aspek kehalalan karena belum menyadari pentingnya sertifikasi halal pada produknya [4]. Fenomena ini diperkuat oleh data yang menunjukkan ketidaksesuaian antara label produk dan isi konten deskripsi yang dilaporkan dalam analisis deskriptif.

Saat ini, pemahaman dan kepatuhan pelaku usaha kecil terhadap regulasi halal masih sangatlah rendah, hanya 25% dari mereka yang mengetahui dan memahami tentang regulasi tersebut [5]. Hal ini menunjukkan bahwa masih perlunya edukasi dan sosialisasi yang lebih intensif mengenai pentingnya sertifikasi halal, terutama bagi pelaku usaha kecil yang seringkali menjadi penyedia utama produk makanan dan minuman di masyarakat. Rendahnya literasi halal ini turut berkontribusi pada maraknya deskripsi produk yang ambigu di platform e-commerce.

Untuk mengatasi permasalahan tersebut, metode analisis tekstual berbasis machine learning seperti TF-IDF dan Cosine Similarity telah banyak digunakan pada domain lain, khususnya untuk mengukur tingkat kemiripan antar dokumen. Berdasarkan studi sebelumnya, Metode TF-IDF terbukti efektif dalam mengekstraksi fitur penting dari suatu teks [6], sedangkan *Cosine similarity* mampu mengukur tingkat kemiripan dokumen dengan akurasi hingga 93,6% [7]. Gabungan kedua metode ini berhasil menilai tingkat kesesuaian antara kutipan dan sumber referensinya, dan menunjukkan hasil yang cukup akurat dalam mengidentifikasi keterkaitan antar teks [8]. Keberhasilan TF-IDF dan *Cosine similarity* ini juga diperkuat oleh penelitian lain [9] yang membandingkan berbagai metode untuk mengukur kemiripan pada data fatwa MUI. Hasil dari penelitian menunjukkan bahwa kombinasi TF-IDF dan Cosine similarity menghasilkan F1-Score tertinggi sebesar 0.299, hasil ini berhasil mengungguli metode lain seperti Jaccard, Dice, Euclidean, dan IndoBert.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan sebuah sistem analisis kemiripan pada deskripsi produk halal dengan menggunakan metode TF-IDF dan Cosine Similarity. Sistem ini diharapkan dapat secara otomatis mendeteksi ambiguitas dalam deskripsi dan membantu regulator dalam memantau kepatuhan terhadap prinsip-prinsip halal. Dengan menganalisis deskripsi produk yang didapat dari platform Kaggle, sistem ini diharapkan mampu meminimalkan kesalahan interpretasi konsumen sekaligus menjadi alat bantu bagi regulator dalam memantau kepatuhan syariah.

2. METODE PENELITIAN

Pada penelitian ini, terdapat dua teknik utama yang digunakan dalam menganalisis kemiripan deskripsi produk halal, yaitu meliputi teknik TF-IDF (*Term Frequency - Inverse Document Frequency*) dan *Cosine Similarity*. TF-IDF merupakan sebuah teknik yang memberikan nilai lebih pada kata-kata yang jarang ditemukan namun memiliki makna penting dalam dokumen tertentu. TF-IDF mampu mengurangi pengaruh kata-kata umum yang sering muncul di banyak dokumen, seperti kata sambung atau kata fungsi. Sementara itu, *Cosine Similarity* merupakan metode yang digunakan untuk menilai tingkat kemiripan antara dokumen dengan membandingkan vektor query dan vektor dokumen [10]. Penelitian ini dilaksanakan melalui enam langkah utama yang sebagaimana digambarkan pada Gambar 1.



Gambar 1. Tahapan metode penelitian

Tahapan pertama dalam penelitian ini dimulai dengan mengidentifikasi masalah. Fokus utama penelitian ini mengacu pada isu ketidakonsistenan deskripsi produk halal yang sering ditemukan di berbagai platform *e-commerce*. Setelah diidentifikasi, dilakukan analisis lebih mendalam terhadap teks deskripsi produk serta label pada produk.

Setelah masalah teridentifikasi, langkah selanjutnya adalah mengumpulkan data yang diperlukan untuk mendukung proses analisis. Pengumpulan data merupakan langkah penting yang bertujuan untuk memperoleh informasi yang diperlukan guna mencapai tujuan penelitian [11]. Dalam penelitian ini, data yang digunakan berupa daftar deskripsi produk bahan makanan dan minuman yang telah memiliki sertifikasi label halal. Data tersebut diperoleh dari platform Kaggle dan mencakup sebanyak 39773 data deskripsi produk makanan dan minuman dengan sertifikasi halal. Namun, pada proses ini hanya digunakan sebanyak 20.000 data, hal ini dikarenakan memori perangkat pemrosesan data tidak dapat menangani dataset dalam ukuran yang terlalu besar.

Setelah data berhasil dikumpulkan, langkah selanjutnya adalah melakukan preprocessing pada data. Tahap ini dilakukan untuk membersihkan unsur yang kurang relevan pada data sebelum dianalisis, seperti tanda baca, huruf kapital, serta kata-kata umum (*stopwords*) dalam deskripsi produk. Proses preprocessing data ini terdiri dari beberapa tahapan. Tahap pertama dimulai dari *tokenizing*, yang digunakan untuk memisahkan teks deskripsi menjadi satuan kata [12]. Hasil akhir dari tahap ini berupa serangkaian token yang akan digunakan di tahap berikutnya. Setelah itu, dilakukan *stopword removal*, yakni proses penghilangan kata umum yang tidak memiliki makna yang penting dalam proses klasifikasi dan analisis, misalnya seperti kata yang, atau, ke, dan sebagainya [13]. Tahap selanjutnya adalah *filtering*, yang dilakukan untuk memeriksa bahwa semua text deskripsi produk hanya terdiri dari huruf alfabet. Token yang memenuhi kriteria akan disimpan, sementara token yang tidak memenuhi kriteria akan dihapus. Terakhir, data deskripsi yang bersifat *null* atau bernilai kosong dan data NaN (*Not a Number*) akan dihapus dalam tahap penghapusan data *null* dan data NaN. Hal ini dilakukan agar tidak mempengaruhi hasil analisis produk bersertifikasi halal.

Setelah data dibersihkan melalui tahap preprocessing, Tahap selanjutnya adalah mengolah data menggunakan metode TF-IDF. Metode ini digunakan untuk memberikan nilai bobot pada setiap kata dalam deskripsi produk, sehingga kata yang unik dan informatif akan memiliki nilai bobot yang lebih tinggi. TF (*Terms Frequency*) digunakan untuk menghitung frekuensi munculnya kata tertentu pada suatu dokumen [14]. Rumus seberapa sering sebuah kata i muncul dalam dokumen j , dibandingkan dengan kata yang paling sering muncul di dokumen tersebut. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 1.

$$TF(i, j) = \frac{freq(i, j)}{maks_k f_{kj}} \quad (1)$$

Keterangan :

$TF(i, j)$ = frekuensi munculnya kata tertentu pada suatu dokumen.

$freq(i, j)$ = jumlah kemunculan kata i di dokumen j .

$maks_k f_{kj}$ = frekuensi kata yang paling sering muncul di dokumen j .

Dalam perhitungan TF, nilai suatu kata dihitung dengan membagi jumlah kemunculan suatu kata dengan total keseluruhan kata pada suatu dokumen. Hal ini menunjukkan kata yang sering muncul dalam deskripsi akan mendapatkan nilai TF yang lebih tinggi daripada kata yang jarang muncul [15]. Selanjutnya, IDF (Inverse Document Frequency) digunakan untuk memberikan bobot yang lebih tinggi pada kata yang jarang muncul di dokumen karena dianggap penting dalam menentukan topik pada suatu dokumen [15]. Hal ini berarti kata yang sering muncul akan memiliki nilai idf yang lebih kecil dibandingkan kata yang jarang muncul. Hal ini dikarenakan kata yang sering muncul akan dianggap kurang informatif [15]. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 2.

$$IDF(i) = \log \frac{N}{n(i)} \quad (2)$$

Keterangan :

$IDF(i)$ = keunikan kata i di seluruh kumpulan dokumen.

N = total jumlah dokumen dalam kumpulan data.

$n(i)$ = jumlah dokumen yang mengandung kata i.

Berikut merupakan rumus TF-IDF pada kata dalam suatu dokumen. Perhitungan TF-IDF dilakukan dengan mengkalikan hasil TF dengan hasil IDF. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 3.

$$TF - IDF(i, j) = TF(i, j) \times IDF(i) \quad (3)$$

Keterangan :

TF = *Terms Frequency*

$TF(i, j)$ = frekuensi munculnya kata tertentu pada suatu dokumen.

$IDF(i)$ = keunikan kata i di seluruh kumpulan dokumen.

Setelah bobot dari setiap kata dihitung menggunakan TF-IDF, tahap selanjutnya adalah menghitung tingkat kemiripan teks deskripsi produk makanan dan minuman dengan menggunakan *Cosine Similarity*. Metode ini sangat membantu dalam proses pengelompokan maupun identifikasi pada produk yang memiliki deskripsi serupa. *Cosine similarity* dinilai sangat cocok untuk mengukur tingkat kemiripan dokumen yang memiliki banyak kata, karena metode ini tidak memperhatikan besarnya nilai vektor, melainkan hanya memperhatikan arah vektor [16]. Berikut merupakan rumus perhitungan *Cosine Similarity*. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 4.

$$\cos \theta = \frac{A \cdot B}{|A| |B|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} * \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

Keterangan :

$\cos \theta$ = *Cosine similarity*

A = Vektor pertama (misalnya representasi dokumen 1)

B = Vektor kedua (misalnya representasi dokumen 2)

A_i = Komponen ke-i dari vektor A

B_i = Komponen ke-i dari vektor B

Nilai kemiripan yang diperoleh dari tahap *Cosine Similarity* kemudian dianalisis lebih lanjut pada tahap evaluasi. Tahap ini dilakukan untuk mengevaluasi model dan menilai tingkat konsistensi informasi mengenai status halal pada deskripsi produk yang mirip satu sama lain. Terdapat beberapa metrik yang digunakan untuk mengevaluasi model. Metrik pertama adalah Presisi (*Precision*), yang digunakan untuk mengukur tingkat akurasi model dalam mengklasifikasikan produk ke dalam status halal ataupun haram. Berikut merupakan rumus perhitungan presisi [17], [18]. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Keterangan :

Precision = Presisi: mengukur tingkat akurasi model

TP = *True Positive*: jumlah prediksi positif yang benar

FP = *False Positive*: jumlah prediksi positif yang salah

Metrik selanjutnya adalah *Recall*, yang digunakan untuk mengukur tingkat keberhasilan model dalam mengidentifikasi produk halal dari keseluruhan produk berstatus halal. Semakin tinggi nilai recall, maka semakin baik pula kemampuan model dalam mendeteksi seluruh produk, hal ini mengurangi resiko adanya produk halal yang tidak terdeteksi. Perhitungan metrik ini dilakukan dengan membandingkan jumlah produk yang diprediksi sebagai halal dengan total jumlah produk halal sesungguhnya. Berikut merupakan rumus perhitungan *Recall* [17], [18]. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Keterangan :

Recall = mengukur tingkat keberhasilan model

TP = *True Positive*: jumlah prediksi positif yang benar

FN = *False Negative*: jumlah kasus positif yang tidak terdeteksi

Metrik selanjutnya adalah *F1-Score*, yang digunakan untuk menggabungkan nilai *Precision* dengan *Recall* sebagai satu kesatuan. Hal ini dilakukan untuk memberikan gambaran yang lebih menyeluruh mengenai cara kerja model. *F1-Score* akan bernilai tinggi apabila nilai *Precision* dan *Recall* juga tinggi. Berikut merupakan rumus perhitungan *F1-Score* [18]. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 7.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (7)$$

Keterangan :

Precision = Presisi: mengukur tingkat akurasi model

Recall = mengukur tingkat keberhasilan model

Selanjutnya, metrik terakhir adalah *Accuracy*, yang digunakan untuk mengukur frekuensi bernilai benar pada prediksi model, baik dalam memprediksi status halal maupun haram, dari keseluruhan prediksi yang dilakukan oleh model. Perhitungan metrik ini dilakukan dengan membagi jumlah prediksi benar dengan total keseluruhan prediksi yang dilakukan oleh model. Perhitungan metrik ini dilakukan dengan membagi jumlah prediksi benar dengan total keseluruhan prediksi yang telah dilakukan. Berikut merupakan rumus yang digunakan untuk menghitung *Accuracy* [17]. Rumus ditulis menggunakan *equation editor* dengan nomor indeks seperti rumus 8.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Keterangan :

TP = *True Positive*: jumlah prediksi positif yang benar

TN = *True Negative*: jumlah prediksi negatif yang benar

FP = *False Positive*: jumlah prediksi positif yang salah

FN = *False Negative*: jumlah kasus positif yang tidak terdeteksi

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini dirancang untuk memuat informasi mengenai deskripsi produk dan label halal haram. Sebelum digunakan, data di dalam dataset ini disaring untuk memastikan deskripsi produk yang digunakan dalam proses pelatihan. Tahapan *preprocessing* mencakup langkah-langkah seperti *Tokenizing*, *Stopword Removal*, *Filtering*, dan menghapus data *null* dan data *NaN*. tahap ini membantu menjaga konsistensi dan kebersihan data sehingga model tidak terpengaruh oleh *noise*. Tabel 1 menunjukkan data sebelum proses *preprocessing* dilakukan, sedangkan Tabel 2 menampilkan data setelah melalui proses *preprocessing*.

Tabel 1. Dataset Deskripsi Produk

No	Text	Label
1	vegetable oil	Halal
2	beef stock contains less than of mirepoix carr...	Halal
3	clam stock potatoes clams cream vegetable oil...	Haram
4	water cream broccoli celery vegetable oil corn...	Haram
5	chicken stock contains less than of yeast extr...	Halal

Tabel 2. Hasil Preprocessing Dataset Deskripsi Produk

No	Text	Label
1	[vegetable, oil]	Halal
2	[beef, stock, contains, less, mirepoix, carrot..]	Halal
3	[clam, stock, potatoes, clams, cream, vegetabl...]	Haram
4	[water, cream, broccoli, celery, vegetable, oi..]	Haram
5	[chicken, stock, contains, less, yeast, extrac...]	Halal

Setelah tahap *preprocessing*, data hasil dari preprocessing akan diubah menjadi representasi numerik menggunakan menggunakan TF-IDF. Hal ini dilakukan agar *Cosine Similarity* dapat menghitung kesamaan produk dan sekaligus untuk melatih model. Lalu hasil dari proses akan ditampilkan dalam bentuk teks yang ditampilkan pada Tabel 3.

Tabel 3. Hasil TF-IDF dalam Bentuk Teks

Coords	Values
(0, 3320)	0.298574755562751
(0, 3977)	0.3329245667109259
(0, 2338)	0.14517015104852817
(0, 6136)	0.09744708792721599
(0, 4962)	0.07838018186993589
(0, 3064)	0.2025523713164477
(0, 31)	0.1041594821897661
(0, 5279)	0.25719757511384506
(0, 3505)	0.33953204800476427
(0, 1253)	0.14441398730107517
(0, 4800)	0.34582892301213355
(0, 505)	0.2328858638110653
(0, 4517)	0.579220440455422
(1, 4346)	0.9665684697365662
(2, 2960)	0.27804354865371583
(2, 6136)	0.0830885840015987
(2, 31)	0.08738587411456064
(2, 505)	0.1984520679514699
(2, 875)	0.236124248266321
(2, 2666)	0.17500131888162396
(2, 2286)	0.1747027420496759
(2, 1371)	0.1001808842682626
(2, 5664)	0.1151294879540834
(2, 3765)	0.0921822481986531
(2, 324)	0.13501457878944248
...	
(19998, 6236)	0.2531645676659984
(19998, 5033)	0.2547279791471803
(19999, 2960)	0.08942826751111105
(19999, 4507)	0.2772009050455627

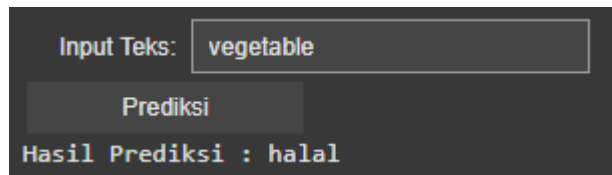
(19999, 4944)	0.0778647518419653
(19999, 5447)	0.14764244261231721
(19999, 2171)	0.06859201651467263
(19999, 4835)	0.09843889275213421

Setelah proses pembobotan menggunakan TF-IDF, langkah berikutnya adalah menerapkan Cosine Similarity untuk mengukur tingkat kesamaan antar deskripsi teks produk berdasarkan representasi dari hasil pembobotan tersebut. Kemudian, hasil tersebut akan disajikan dalam bentuk teks yang ditampilkan pada Tabel 4.

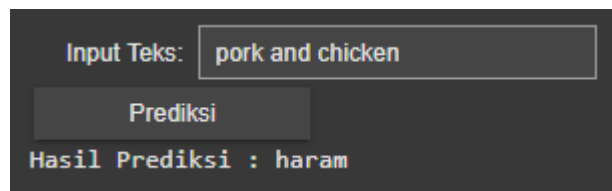
Tabel 4. Hasil Cosine Similarity dalam Bentuk Teks

1.	0.	0.06355361	...	0.030931	0.03352875	0
0.	1.	0.	...	0.	0	0.02475374
0.06355361	0.	1.	...	0.04028357	0.05364976	0
.....						
0.030931	0.	0.04028357	...	1.	0.12575809	0.0584797
0.03352875	0.	0.05364976	...	0.12575809	1.	0.09721327
0.	0.02475374	0.	...	0.0584797	0.09721327	1.

Setelah melalui seluruh tahapan tersebut, dibuatlah sebuah implementasi ekstraksi fitur mendeteksi status kehalalan produk berdasarkan deskripsi produk yang diberikan. Sistem ini dibangun menggunakan *widget* yang tersedia di *Google Collab*, memungkinkan pengguna untuk memasukkan deskripsi produk yang ingin diperiksa. Selanjutnya, dilakukan uji prediksi apakah produk tersebut tergolong halal atau haram. Tampilan uji coba diperlihatkan pada Gambar 5 dan Gambar 6.



Gambar 5. Hasil Prediksi Produk Halal

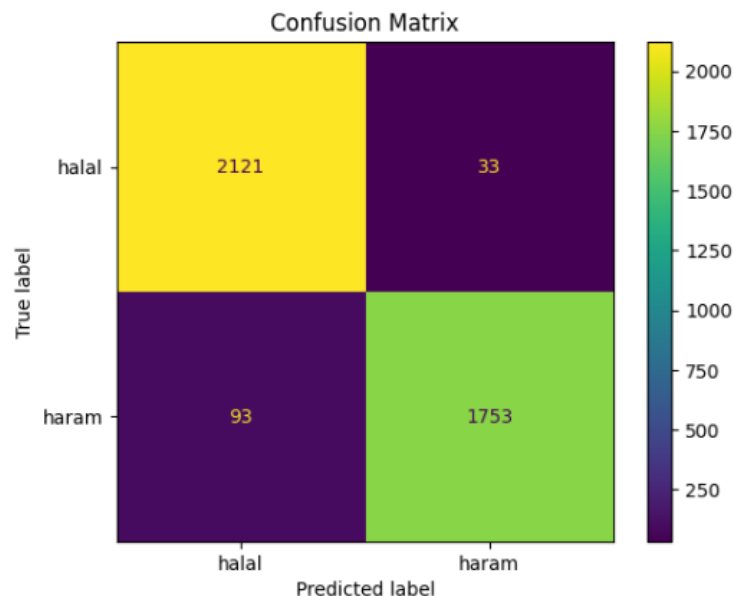


Gambar 6. Hasil Prediksi Produk Haram

Proses evaluasi dilakukan dengan menggunakan metrik *precision*, *recall*, *F1-score*, dan *accuracy*. Sebelum evaluasi, data dibagi menjadi dua bagian, yaitu 80% untuk data pelatihan dan 20% untuk data pengujian. Model yang digunakan dalam evaluasi adalah *Logistic Regression*. Hasil evaluasi menunjukkan bahwa untuk kategori halal, *precision* mencapai 0.96, yang berarti 96% dari prediksi halal adalah benar, sementara *recall* sebesar 0.98 menunjukkan bahwa model berhasil mendeteksi 98% dari total produk halal. *F1-score* untuk kategori halal tercatat sebesar 0.97, mengindikasikan keseimbangan yang baik antara *precision* dan *recall*.

Untuk kategori haram, *precision* tercatat sebesar 0.98, yang menunjukkan 98% dari prediksi haram adalah benar, namun *recall* sebesar 0.95 menunjukkan model hanya mendeteksi 95% dari total produk haram. *F1-score* untuk kategori haram berada pada angka 0.97. Secara keseluruhan, *accuracy* model mencapai 97%, yang berarti model mampu membuat prediksi yang benar untuk 97% dari data pengujian. Evaluasi juga dilakukan dengan *Confusion Matrix* menunjukkan bahwa model lebih baik dalam mengenali produk halal dibandingkan produk haram, dengan

recall yang lebih tinggi untuk kategori halal (98%) dibandingkan haram (95%). Hal ini menunjukkan bahwa model lebih sering meloloskan produk haram sebagai halal dibandingkan sebaliknya. Hasil evaluasi *Confusion Matrix* terlihat pada Gambar 7.



Gambar 7. Hasil Evaluasi *Confusion Matrix*

4. KESIMPULAN

Penelitian ini berhasil implementasi ekstraksi fitur dari deskripsi produk halal dan haram dengan memanfaatkan metode TF-IDF dan *Cosine Similarity* untuk menganalisis deskripsi produk secara efektif. Penggunaan TF-IDF sebagai representasi fitur teks, dipadukan dengan *Cosine Similarity* sebagai ukuran kesamaan, terbukti meningkatkan performa klasifikasi secara signifikan. Berdasarkan hasil evaluasi (*precision*, *recall*, *F1-score*, dan *accuracy*), sistem ini menunjukkan tingkat akurasi yang tinggi, sehingga mampu memprediksi status produk dengan lebih andal. Sistem ini diharapkan menjadi langkah awal menuju transparansi informasi halal di platform e-commerce, sekaligus memperkuat kredibilitas Indonesia sebagai negara dengan sertifikasi halal yang terpercaya.

Berdasarkan penelitian yang telah dilakukan terhadap deskripsi produk halal, terdapat beberapa hal yang dapat ditingkatkan dan dikembangkan lebih lanjut. Salah satunya adalah perlunya menggunakan komputasi yang lebih memadai serta memory yang besar agar metode *Cosine Similarity* dapat memproses data dalam jumlah banyak. Dalam penelitian ini, jumlah dataset yang digunakan harus dibatasi agar dapat diproses dengan baik oleh metode *Cosine Similarity*. Ketika dataset tidak dibatasi, sistem komputasi selalu mengalami gangguan error sehingga muncul pemberitahuan berupa “*session crashed*”, yang mengindikasikan terjadinya keterbatasan pada memory maupun sistem komputasi yang digunakan. Untuk mengatasi permasalahan tersebut, riset selanjutnya dapat memanfaatkan metode distribusi data atau pemrosesan paralel berbasis cloud agar mampu menangani volume data yang lebih besar secara efisien.

DAFTAR PUSTAKA

- [1] R. Maulizah and S. Sugianto, “Pentingnya Produk Halal di Indonesia: Analisis Kesadaran Konsumen, Tantangan dan Peluang,” *El-Suffah J. Stud. Islam*, vol. 1, no. 2, pp. 129–147, 2024, doi: 10.70742/suffah.v1i2.49.
- [2] B. A. H. Hakim, A. S. Mujahidah, and A. S. Rusydiana, “Sentiment Analysis on Halal Certification,” *Harmoni*, vol. 21, no. 1, pp. 78–93, 2022, doi: 10.32488/harmoni.v21i1.609.
- [3] B. Brahimi, M. Touahria, and A. Tari, “Improving sentiment analysis in Arabic: A combined approach,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 10, pp. 1242–1250, 2021, doi: 10.1016/j.jksuci.2019.07.011.
- [4] Z. E. Fitri and A. Jumiono, “Sertifikasi Halal Produk Olahan Pangan,” *J. Pangan Halal*, vol. 3, no. 2, pp. 1–7, 2021, doi: 10.30997/jiph.v3i2.9676.
- [5] N. U. Khasanah, M. E. Safira, W. K. S. Agung, M. Chotib, S. Bin Lahuri, and E. P. Nimasari, “Regulation

- of halal and healthy products for small-scaled businesses as consumer protection,” *Open Access Maced. J. Med. Sci.*, vol. 9, no. E, pp. 749–753, 2021, doi: 10.3889/oamjms.2021.6643.
- [6] M. Liang and T. Niu, “Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs,” *Procedia Comput. Sci.*, vol. 208, pp. 460–470, 2022, doi: 10.1016/j.procs.2022.10.064.
 - [7] A. Widiyanto, E. Pebriyanto, F. Fitriyanti, and M. Marna, “Document Similarity using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity,” *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 4, no. 2, pp. 149–153, 2024, doi: 10.20895/dinda.v4i2.1589.
 - [8] U. Mardatillah, W. B. Zulfikar, A. R. Rialdy, I. Taufik, and W. Uriawan, “Citation Analysis on Scientific Articles Using Cosine Similarity,” *Proceeding 2021 7th Int. Conf. Wirel. Telemat. (ICWT 2021)*, 2021, doi: 10.1109/ICWT52862.2021.9678402.
 - [9] M. F. Syafiudin, G. Adiatmaja, and B. Hidayaturrohmah, “Calculation of Similarity between MUI Fatwas: A Comparison of Word Extraction and String-Matching Algorithms,” *Halal Res.*, vol. 5, no. 1, pp. 1–13, 2025, doi: 10.12962/j22759970.v5i1.1226.
 - [10] I. G. Anugrah, “Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi,” *Build Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 275–284, 2021, doi: 10.47065/bits.v3i3.1058.
 - [11] L. Suryani and K. Edy, “Pengembangan Aplikasi “Lost & Found” Berbasis Android dengan Menggunakan Metode Term Frequency – Inverse Document Frequency (TF-IDF) dan Cosine Similarity,” *Electro Lucueat*, vol. 6, no. 2, pp. 190–204, 2020, doi: 10.32531/jelekn.v6i2.232.
 - [12] Nofiyani and Wulandari, “Implementasi Electronic Data Processing untuk Meningkatkan Efektifitas dan Efisiensi Pada Text Mining,” *J. Media Inform. Budidarma*, vol. 6, no. 3, pp. 1621–1629, 2022, doi: 10.30865/mib.v6i3.4332.
 - [13] R. Al Rasyid and D. H. U. Ningsih, “Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata,” *J. Teknol. Inf. dan Komun.*, vol. 8, no. 1, pp. 170–178, 2024, doi: 10.35870/jtik.v8i1.1416.
 - [14] R. Wati, S. Ernawati, and H. Rachmi, “Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH,” *J. Manaj. Inform.*, vol. 13, no. 1, pp. 84–93, 2023, doi: 10.34010/jamika.v13i1.9424.
 - [15] S. Herimanto, K. Samosir, and F. Ginting, “A Comparative Analysis of Content-Based Filtering and TF-IDF Approaches for Enhancing Sports Recommendation Systems,” *Innov. Res. Informatics*, vol. 6, no. 2, pp. 90–97, 2024, doi: 10.37058/innovatics.v6i2.12404.
 - [16] S. Amaliah, M. Nusrang, and A. Aswi, “Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng,” *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variensiunm31.
 - [17] M. J. Mokarrama, S. Khatun, and M. S. Arefin, “A content-based recommender system for choosing universities,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 4, pp. 2128–2142, 2020, doi: 10.3906/elk-1911-37.
 - [18] M. Das, S. Kamalanathan, and P. Alphonse, “A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset,” *CEUR Workshop Proc.*, vol. 2870, pp. 98–107, 2021, doi: 10.48550/arXiv.2308.04037.