Comparison of TF-IDF and Sentence-Transformer NLP Methods for a Perfume Recommendation System Based on User Descriptions with Streamlit Visualization

Kevin De Rafael Rio Aryanto[1*], Faulinda Ely Nastiti [2], Ridwan Dwi Irawan[3]

[1,2,3] *Universitas Duta Bangsa Surakarta, Jl. Bhayangkara No. 55 Tipes, Serengan, Surakarta, Central Java 57154, Indonesia*

Keywords
Content-Based Filtering; Natural Language Processing; Perfume Recommendation; Sentence-Transformer; TF-IDF.

Correspondence Email:
220103096@mhs.udb.ac.id

Abstract
The global perfume industry has grown considerably in recent years, yet helping consumers identify products that genuinely align with their personal taste remains a non-trivial problem, especially when preferences are communicated through unstructured, free-form natural language rather than standardized scent terminology. This study addresses that gap by evaluating and comparing two Natural Language Processing (NLP) text-representation methods, Term Frequency-Inverse Document Frequency (TF-IDF) and Sentence-Transformer, within a cross-lingual content-based filtering perfume recommendation system. The novelty lies in the cross-lingual setup, where unstructured Indonesian-language queries are matched directly against an English fragrance dataset of 42,115 records without any external translation module. A Streamlit web interface was deployed to collect data from 30 respondents, who submitted 180 graded relevance judgments for Top-3 recommendations evaluated through Precision@3 and Normalized Discounted Cumulative Gain (nDCG@3). System stability was assessed by partitioning the corpus into a training set of 33,692 records and a test set of 8,423 records. Sentence-Transformer, using the pre-trained paraphrase-multilingual-MiniLM-L12-v2 model, outperformed TF-IDF on all metrics, achieving a mean Precision@3 of 0.867 against 0.833 and a mean nDCG@3 of 0.912 against 0.850, with substantially lower score degradation across data partitions. These findings confirm that semantic embedding methods provide superior ranking quality and cross-lingual robustness, offering a scalable and translation-free solution applicable to multilingual product recommendation systems in commercial settings.

1. Introduction

The global perfume market has been growing at a pace that few consumer sectors can match. Market projections consistently point to continued revenue expansion through 2030 [1] , and within the broader beauty landscape, the fragrance category has emerged as one of the fastest-growing segments in prestige retail [2] . As the product catalog widens, finding a scent that genuinely resonates with individual taste becomes harder, not easier. Most consumers do not shop by technical specifications. They reach for

descriptions that feel natural to them, things like "a light, fresh scent for hot weather" or "something soft and sweet but not overwhelming." Static filter-based systems, which dominate most existing recommendation platforms, are simply not built to handle that kind of open-ended input. That mismatch between how people think about fragrance and how systems process queries is the problem this research sets out to address.

Prior attempts at solving this have followed two broad paths. The first relied on structured multicriteria methods. Tobing et al. [3] designed a men's perfume recommendation system using Analytic Hierarchy Process (AHP) combined with TOPSIS, ranking products by price and longevity attributes. These approaches are defensible in structured decision contexts but become rigid the moment a user wants to describe a mood rather than specify a criterion. The second path turned to Natural Language Processing. Agashe [4] demonstrated that NLP-driven recommendation pipelines can process free-text queries directly without predefined attribute mappings. Nurmuthia et al. [5] took this further by applying content-based filtering to perfume recommendation, measuring user-query similarity through Cosine and Jaccard metrics with promising results. TF-IDF paired with cosine similarity has since been adopted broadly, appearing in news article recommendation [6] , recipe retrieval [7] , tourism destination suggestion [8] , conversational AI systems [9] , and fragrance-specific applications [10] . Phalle and Bhushan [11] confirmed through a systematic comparison that content-based filtering outperforms collaborative filtering whenever user preference is expressed through descriptive language rather than interaction history.

Despite this progress, a fundamental limitation cuts across virtually all TF-IDF-based approaches. The method assigns weights to words, not meanings, which makes it dependent on exact lexical overlap between query and document. Birwadkar [12] showed that this surface-level matching breaks down systematically when semantically related texts differ in vocabulary, as is common with synonyms and paraphrases. In a cross-lingual setting, this weakness becomes far more severe. When a user submits a query in Indonesian and the fragrance database stores aroma profiles in English, TF-IDF often fails to find relevant matches simply because the same concept appears in two different languages. Bhangale et al. [13] provided empirical evidence that Transformer-based semantic methods consistently outperform conventional NLP approaches in multilingual matching tasks. Kim et al. [14] extended this to the perfume domain specifically, showing that Sentence-Transformer embeddings capture inter-sentence semantic relationships effectively in fragrance note estimation, a task structurally similar to the recommendation problem explored here.

What remains absent from the literature is a direct, controlled comparison between TF-IDF and Sentence-Transformer in a cross-lingual perfume recommendation scenario, tested through real user judgments, and conducted without any external translation layer. This study fills that gap. The primary objective is to compare the recommendation quality of both methods using Precision@3 and Normalized Discounted Cumulative Gain (nDCG@3) as evaluation metrics, supplemented by a representation stability analysis that partitions the dataset into separate training and test sets to assess score consistency across data compositions. An interactive Streamlit web prototype [15] was deployed as both the demonstration interface and the instrument for collecting respondent ratings. The results show that Sentence-Transformer, using the pre-trained paraphrase-multilingual-MiniLM-L12-v2 model, outperforms TF-IDF on all evaluated metrics, and that its similarity score degrades far less when moving from training data to unseen test data. These findings carry practical implications for developers of multilingual product recommendation systems, offering evidence that semantic embedding methods deliver meaningful advantages in cross-lingual retrieval without the operational overhead of maintaining a translation module.

2. Research Methods

This study adopts a comparative experimental design that positions TF-IDF and Sentence-Transformer within an identical processing pipeline from start to finish. Structurally, the research follows a five-stage data mining framework covering data collection, preprocessing, modeling, system deployment alongside respondent data acquisition, and evaluation. One decision that shapes the entire flow is placing deployment in the middle rather than at the end. In most recommendation research, the system is built and then the evaluation happens afterward in a controlled setting. This study flips that logic. The deployed application

becomes the data collection environment itself, which means the relevance judgments respondents provide come from real interaction with a functioning system rather than from reviewing synthetic outputs on paper. Fig. 1 lays out the complete research pipeline.

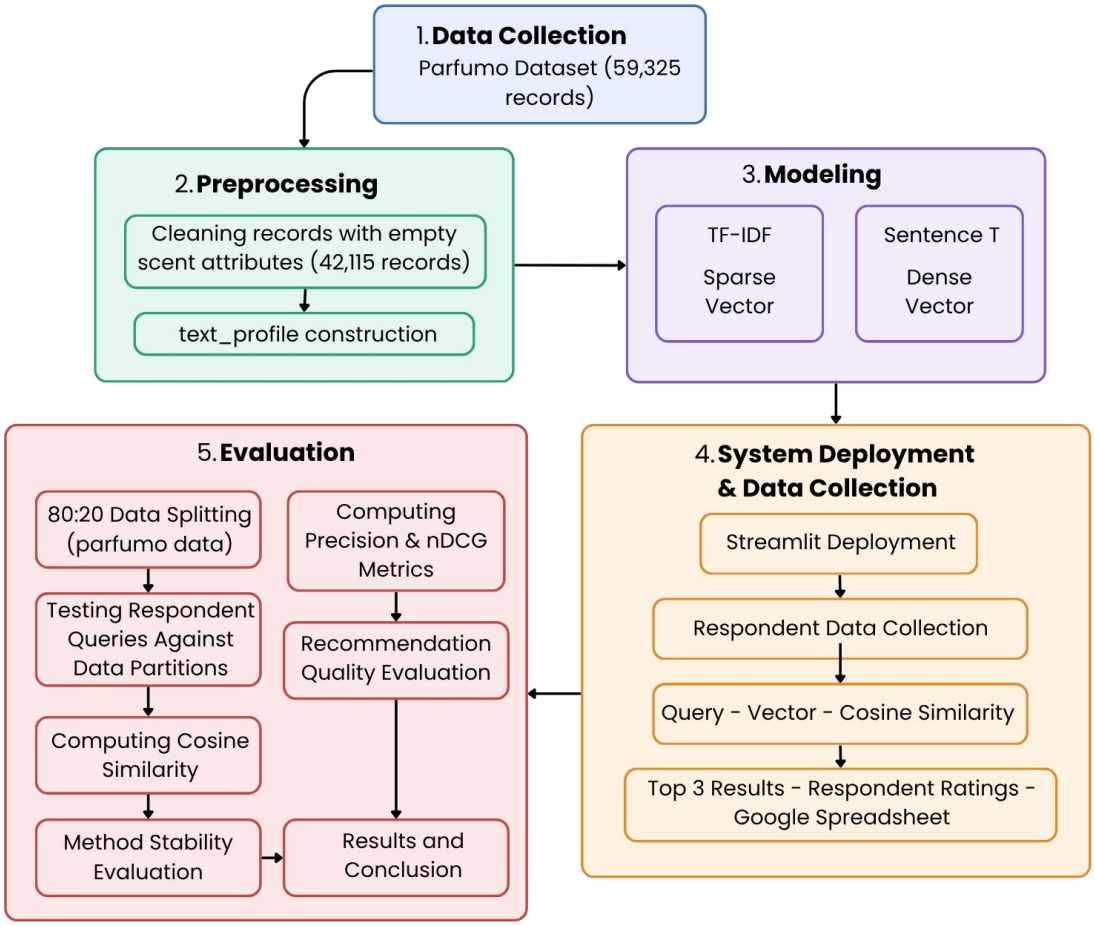

Fig. 1. Research System Workflow

2.1 Data Collection

Two distinct data sources feed this research. The first is a static fragrance corpus. The Parfumo Fragrance Dataset, retrieved from the TidyTuesday public repository at https://github.com/rfordatascience/tidytuesday/tree/main/data/2024/2024-12-10, provides 59,325 raw entries drawn from the Parfumo international fragrance community, a platform where users contribute aroma descriptions for commercial perfumes spanning multiple countries and price ranges. Six attributes were extracted for each entry: Name, Brand, Main Accords, Top Notes, Middle Notes, and Base Notes. These attributes together form the olfactory fingerprint of each product and serve as the raw textual material for the modeling stage.

The second data source is human judgment. Thirty adults participated as respondents through purposive convenience sampling, recruited based on their ability to express fragrance preferences in everyday natural language regardless of how much they knew about perfumery technically. No professional background in fragrance was required, and participants were not screened for specific demographic criteria beyond being adults. Through the system interface, each respondent completed a brief profile form that captured gender, age group, and self-rated familiarity with perfume products, ranging from occasional buyers to habitual collectors. This profile data characterizes the respondent pool and contributes to the transparency of the

study's human judgment layer. Across all 30 sessions, respondents submitted 27 unique queries total, the slight gap between participant count and query count arising from a small number of independently written descriptions that were textually identical once compared side by side.

2.2 Data Preprocessing

The raw dataset of 59,325 entries required cleaning before any modeling could take place. A thorough inspection found 17,210 records where all four scent-related columns, Main Accords, Top Notes, Middle Notes, and Base Notes, were completely empty. Entries in this condition were removed from the corpus. A perfume without any aroma data cannot be meaningfully recommended to anyone, and more practically, its corresponding vector would be all zeros, which would skew cosine similarity calculations in unpredictable ways. Removing these entries reduced the working corpus to 42,115 records.

Each surviving entry was then converted into a single unified text string called a text_profile, built by concatenating all six extracted attributes in a fixed sequential template. Both NLP methods compared in this study process one text input per document, so consolidating the attributes into one string is not just convenient but necessary. It also ensures that neither method operates with less information than the other, which matters for the fairness of the comparison. No external translation or language normalization was applied at any point in this stage. The corpus remained in English throughout, and Indonesian-language user queries would later be matched against it as-is. That deliberate choice is what makes the cross-lingual setting real rather than artificially resolved.

2.3 Modeling

The same 42,115 *text_profile* strings were fed into both methods under identical conditions, processed separately and independently so any difference in downstream performance can be traced back to the representation method rather than to any data asymmetry.

TF-IDF was built using English stop-word removal and a bigram range of (1,2). Extending to bigrams was a considered choice given the nature of fragrance language. Descriptors like *woody amber*, *french lavender*, and *sea breeze* are two-word units that mean something quite different from their individual components, and a unigram-only model would miss that. The resulting matrix was sparse with dimensions of 42,115 by 116,402, reflecting the large vocabulary that emerges from across tens of thousands of fragrance profiles.

Sentence-Transformer used the pre-trained paraphrase-multilingual-MiniLM-L12-v2 checkpoint, a compact model with demonstrated competitiveness across diverse semantic similarity tasks in large-scale embedding benchmarks [16] . All 42,115 profiles were encoded into a dense matrix of 42,115 by 384, with vectors normalized before storage. Saving the embeddings to disk meant they could be reused across both the deployment stage and the evaluation stage without recomputation, which kept the system response time acceptable for live respondent interaction.

2.4 System Deployment and Data Collection

Both vector representations were integrated into a Streamlit web application deployed to Streamlit Cloud, accessible at https://skripsi-parfum-app-xeztptfchcu26usvxkuuhw.streamlit.app/. The system was intentionally kept straightforward in its design. A respondent arrives at the interface, enters a free-form Indonesian-language description of the fragrance they are looking for, and the system returns the three highest-scoring perfumes from each method displayed side by side, six items in total per query. Matching happens by encoding the query using each method's pre-built model and computing cosine similarity against all 42,115 fragrance vectors. Cosine similarity was chosen specifically because it measures the angular distance between vectors rather than their magnitude, which makes it robust to variation in text length across queries and documents.

Content validity of the collected judgments comes from the self-referential structure of the rating task. Each respondent evaluated recommendations against a query they wrote themselves, which means the relevance standard being applied is the respondent's own mental model of what they wanted. There is no

external benchmark being imposed that might introduce construct mismatch. The person who typed the query is exactly the right person to judge whether the returned perfume matches it.

Rating reliability was supported through explicit anchor definitions displayed within the interface during the rating task. The three-point scale was not left open to interpretation. Score 0 was defined as not relevant, score 1 as somewhat relevant, and score 2 as highly relevant, with brief examples provided alongside each level. Providing concrete anchors reduces the ambiguity that would otherwise inflate variance across respondents who might naturally interpret a loose scale differently. All ratings were logged automatically to a shared Google Spreadsheet at the moment of submission, accessible at https://docs.google.com/spreadsheets/ d/1w_xcFAgyUzAjeloUHGsPXuAww59TQnA8/edit?usp=sharing, removing manual transcription as a potential error source. The complete dataset of 180 graded relevance judgments was collected through this process across all 30 respondents and 27 unique queries.

2.5 Evaluation

Evaluation ran along two parallel tracks designed to capture different aspects of method performance, and both tracks address a specific dimension of validity or reliability that the study needs to establish.

The primary track used Precision@3 and Normalized Discounted Cumulative Gain (nDCG@3) to measure the quality of recommendations as experienced by respondents. Both are recognized standards in information retrieval and recommender systems research [17] [18] [19], which provides metric validity, the assurance that the instruments being used actually measure what they claim to measure in this domain. Precision@3 counts the proportion of the three returned items that a respondent judged relevant at any level, treating scores of 1 and 2 as relevant and score 0 as not. The nDCG@3 metric weighs both relevance and position, rewarding systems that rank the most relevant items first and penalizing those that bury them lower, using the graded scores as the weight at each rank position. Together these two metrics cover retrieval accuracy from a binary perspective and ranking quality from a graded one, which makes them complementary rather than redundant.

The secondary track addressed the reliability of each method's representational output through a partition stability test. The fragrance corpus of 42,115 records was divided into a training partition of 33,692 records and a held-out test partition of 8,423 records. Respondent queries were applied to both partitions independently, and the mean cosine similarity score of the top-three results was recorded for each. The gap between a method's training-partition score and its test-partition score quantifies how much its output degrades when the candidate pool changes in size and composition. A method that holds its scores steady across both partitions is demonstrably reliable in producing consistent representations, and a large degradation suggests the method leans heavily on vocabulary overlap that shrinks as the corpus shrinks. Neither model was retrained for this test. The pre-built vectors were simply queried against a reduced subset of the original corpus, which keeps the analysis focused on score consistency rather than on overfitting or generalization in a machine learning sense.

3. Results and Discussion

This section reports findings from each stage of the research pipeline in the order they were executed, then moves into a comparative discussion that interprets those findings against the existing literature.

3.1 Data Collection Results

The Parfumo Fragrance Dataset yielded 59,325 raw entries upon initial extraction. Table 1 presents a representative sample of the collected records, illustrating the range of aroma attributes available across the dataset.

Table 1. Sample of Collected Parfumo Dataset Records

ID	Name	Brand	Main Accords	Top Notes	Middle Notes	Base Notes
12	I am a Dandelion	CB I Hate Perfume	Sweet, Gourmand, Green	Meringue, Magnolia	Parsley, Clove	Green vanilla, White oud
14	On A Clear Day	CB I Hate Perfume	Smoky, Sweet, Spicy	Cardamom, Pink pepper	Juniper, Vanilla	Cypress, Musk
26	À L'Apogée de Vert	Le Ré Noir	Spicy, Green, Woody	Basil, Galbanum	Frankincense, Cedar	Cistus, Labdanum
…	…	…	…	…	…	…
59.325	Eros Flame	Versace	Fresh, Spicy, Warm	Black pepper, Mandarin	Geranium, Rosemary	Vanilla, Patchouli

(Source: Parfumo Fragrance Dataset, 2024)

As shown in Table 1, each entry contains multilayered olfactory descriptors distributed across Main Accords, Top Notes, Middle Notes, and Base Notes. The diversity of fragrance vocabulary across these attributes, spanning floral, woody, fresh, spicy, and gourmand families, is what makes this dataset well-suited for testing text-based representation methods. It is also what makes the cross-lingual challenge real. All of those descriptors are in English, and every query submitted by respondents would be in Indonesian.

3.2 Preprocessing Results

Cleaning the dataset was a necessary and not entirely trivial step. Inspection identified 17,210 entries where all four scent attribute columns were empty, meaning there was no aroma information of any kind attached to those records. Table 2 illustrates examples of such incomplete entries.

Table 2. Example Records with Empty Aroma Information

Name	Brand	Main Accords	Top Notes	Middle Notes	Base Notes
Tabac Écarlate	Le Ré Noir	-	-	-	-
Wet Stone	CB I Hate Perfume	-	-	-	-
Chocolate Box	CB I Hate Perfume	-	-	-	-
My Birthday Cake	CB I Hate Perfume	-	-	-	-
Beautiful Launderette	CB I Hate Perfume	-	-	-	-

(Source: Parfumo Fragrance Dataset, 2024)

As Table 2 confirms, these entries carry product names and brands but contribute nothing usable for aroma-based matching. Removing them brought the final workable corpus to 42,115 records, a reduction of roughly 29 percent from the raw total. Each surviving record was then consolidated into a single text_profile string. Table 3 demonstrates what this concatenation produces in practice.

Table 3. Example of text_profile Formation

ID	text_profile
1	Name: Tidal Pool Brand: CB I Hate Perfume Main accords: Top notes: Bergamot Middle notes: French lavender Base notes: Musk, Foulness

(Source: Research data processing, 2026)

Table 3 shows how six separate attributes collapse into one continuous text representation. That unified string is what both methods receive as input, ensuring they operate under exactly the same informational conditions.

3.3 Modeling Results

The modeling stage produced two fundamentally different types of vector representations from the same 42,115 *text_profile* inputs. TF-IDF generated a sparse matrix of 42,115 by 116,402 dimensions, where dimensionality reflects the total unique vocabulary across the corpus. Sentence-Transformer produced a dense embedding matrix of 42,115 by 384 dimensions, fixed regardless of vocabulary size. Figure 2 visualizes the structural contrast between these two output types.

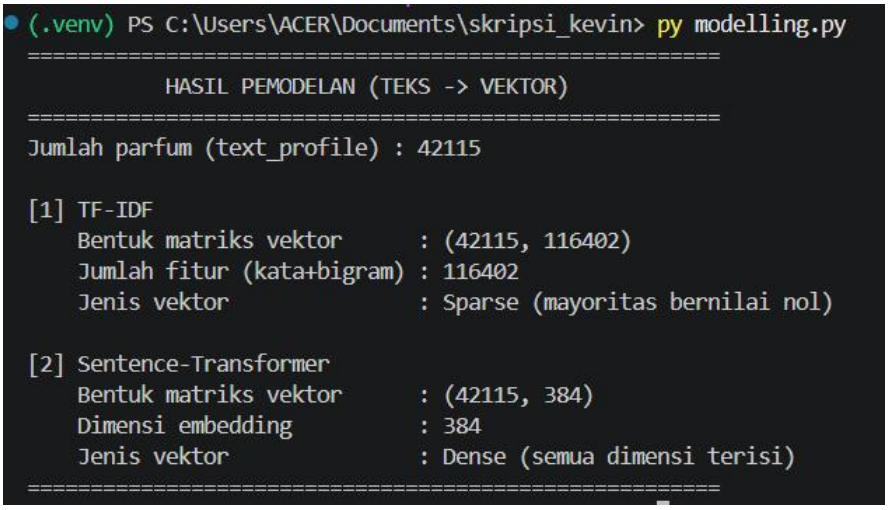

Fig. 2. Vector Representation Output of Both Methods

As shown in Fig. 2, the sparse representation contains vast stretches of zero values because most vocabulary terms appear in only a fraction of all documents. The dense embedding, by contrast, encodes meaning into every dimension, resulting in a compact representation that carries contextual information rather than just word occurrence statistics. This architectural difference is not incidental. It is the core reason why the two methods behave so differently when confronted with cross-lingual queries in the evaluation stage.

3.4 Deployment and Data Collection Results

Figure 3 shows the live system interface as respondents encountered it through Streamlit Cloud.

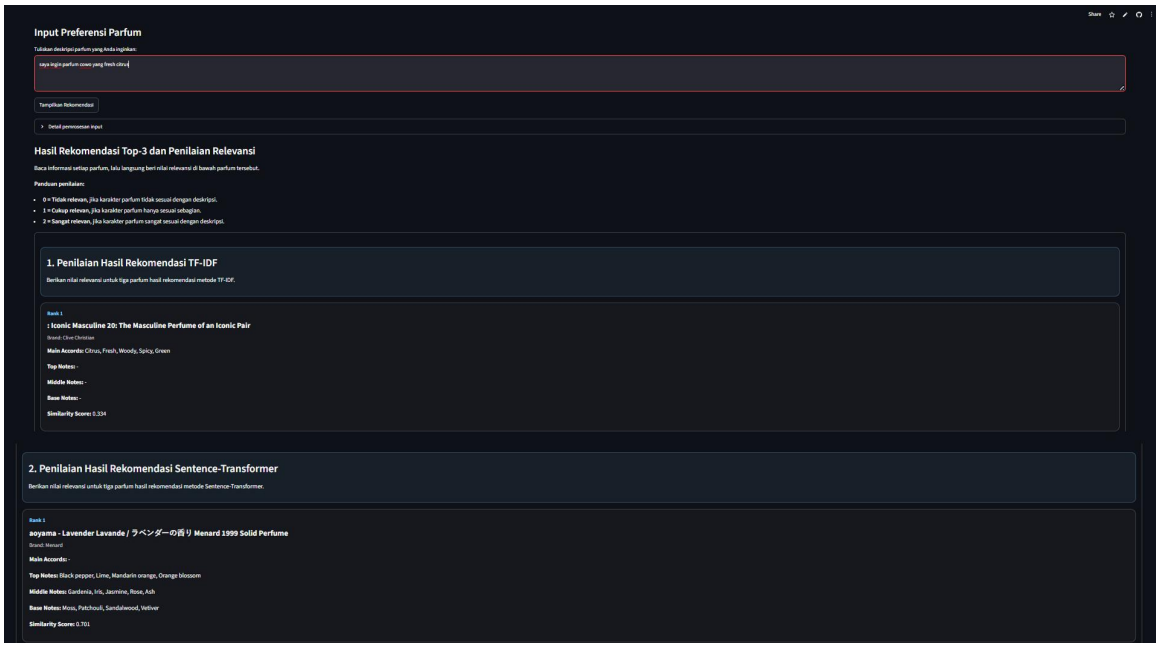

Fig. 3. Perfume Recommendation System Interface on Streamlit

As Fig. 3 illustrates, the interface presents both methods' top-three results side by side, allowing respondents to compare outputs before assigning relevance scores. The deployment phase collected 180 graded relevance judgments from 30 respondents across 27 unique Indonesian-language queries. The distribution of scores fell as follows: 96 judgments received a score of 2, indicating high relevance; 57 received a score of 1, indicating partial relevance; and 27 received a score of 0, indicating no relevance. This distribution, where roughly half of all returned items were rated highly relevant, suggests the system as a whole performed reasonably well, though the difference between methods becomes visible once the data is broken down by method in the next stage.

3.5 Evaluation Results

Before presenting the comparative figures, it is worth grounding the metrics formally. Precision@K is defined as the proportion of relevant items among the top K returned results, where relevance is treated as binary (scores of 1 and 2 count as relevant, 0 does not), as expressed in Equation 1.

$$Precision@K = (\Sigma_{i=1..K}\, rel_i) / K \qquad (1)$$

where rel_i equals 1 if the item at rank i is judged relevant and 0 otherwise. Normalized Discounted Cumulative Gain uses the graded relevance scores directly, weighting higher-ranked positions more heavily through a logarithmic discount, computed via Equations 2 and 3.

$$DCG@K = \Sigma_{i=1..K}\, (2^{rel_i} - 1) / log_2(i + 1) \qquad (2)$$

$$nDCG@K = DCG@K / IDCG@K \qquad (3)$$

where IDCG@K is the maximum possible DCG@K under the ideal ranking of items from most to least relevant. These formulations follow the conventions established in information retrieval literature [17] [18] [19].

The comparative results across both metrics are summarized in Table 4.

Table 4. Comparative Performance of Both Methods

Method	Mean Precision@3	Mean nDCG@3
TF-IDF	0,833	0,850

Sentence-Transformer	0,867	0,912

(Source: Research data processing, 2026)

Table 4 shows that Sentence-Transformer outperformed TF-IDF on both metrics. On Precision@3, the margin was modest: 0.867 versus 0.833, a difference of 0.034. On nDCG@3, the gap widened considerably: 0.912 versus 0.850, a difference of 0.062. That divergence between the two metrics carries an important interpretive implication and will be addressed in the discussion.

3.6 Representation Stability Results

The stability test partitioned the 42,115-record fragrance corpus into a training partition of 33,692 records and a held-out test partition of 8,423 records. Each of the 27 respondent queries was applied to both partitions independently using each method's pre-built vector representations. Table 5 reports the resulting mean cosine similarity scores and the degradation between partitions.

Table 5. Cosine Similarity Score Consistency Across Data Partitions

Method	Training Partition (80%)	Test Partition (20%)	Score Degradation
TF-IDF	0,1992	0,1209	0,0783
Sentence-Transformer	0,6038	0,5820	0,0218

(Source: Research data processing, 2026)

As Table 5 shows, the absolute cosine similarity values are not comparable between methods because they originate from different vector spaces, sparse for TF-IDF and dense for Sentence-Transformer. The meaningful comparison is the degradation column. TF-IDF's score dropped by 0.0783 when moving from the full training partition to the smaller test set. Sentence-Transformer's score dropped by only 0.0218 across the same transition. That is a roughly fourfold difference in sensitivity to corpus size variation, which points to a substantial difference in representational robustness.

3.7 Discussion

The findings reported above converge on two conclusions that deserve careful interpretation rather than simple acceptance at face value.

The first and most direct conclusion is that Sentence-Transformer delivered stronger recommendation quality than TF-IDF in this cross-lingual setting. But the way it was stronger is worth examining. The Precision@3 gap (0.034) was relatively narrow, meaning both methods returned roughly similar proportions of relevant items. The nDCG@3 gap (0.062) was nearly twice as large, which tells a different story. This means TF-IDF often found relevant perfumes in its top-three, but it did not consistently put the most relevant ones first. Sentence-Transformer did both things better, but its primary advantage was in ranking quality. That distinction matters practically. A recommendation system that retrieves relevant items but buries them at position three while an irrelevant item sits at position one will frustrate users even if, technically, something relevant did appear. The nDCG metric captures exactly that kind of failure, and this is where the semantic approach showed its clearest edge.

This pattern is consistent with what Birwadkar [12] documented about the structural limits of lexical methods: when vocabulary does not overlap between query and document, the entire retrieval mechanism weakens. In a cross-lingual context, that overlap failure is systematic rather than occasional. An Indonesian query like "aroma segar dan ringan" shares essentially zero lexical content with English descriptors like "fresh", "light", or "green", even though those words mean the same thing. TF-IDF has no mechanism to bridge that gap. Sentence-Transformer, trained on multilingual data, maps both inputs into the same semantic space, which is precisely why Bhangale et al. [13] found Transformer-based methods to dominate cross-lingual matching tasks and why Kim et al. [14] observed strong performance from sentence embeddings specifically in the fragrance domain.

Where this study's findings differ from prior work is in what was being compared. Nurmuthia et al. [5] held the representation method constant and varied the similarity metric, comparing Cosine against Jaccard Similarity. This study did the opposite: it held Cosine Similarity constant and varied the representation method. That inversion is what makes the contribution distinct. The results here do not simply confirm that content-based filtering works for perfume recommendation, which Nurmuthia et al. already established. They establish specifically that how text is represented matters as much as how similarity is computed, and that in a cross-lingual setup, the representational choice is arguably the more consequential of the two decisions.

The stability analysis adds a second dimension to this argument. A method that degrades substantially when the candidate pool shrinks is one whose output is partly an artifact of corpus size rather than a genuine reflection of semantic alignment. TF-IDF's degradation of 0.0783 versus Sentence-Transformer's 0.0218 suggests that a significant portion of TF-IDF's retrieval behavior depends on incidental vocabulary co-occurrence patterns that disappear as the corpus thins. Dense embeddings carry meaning in a way that is less sensitive to those incidental co-occurrences. This observation aligns with what Sarode and Javaji [20] and Javaji and Sarode [21] demonstrated in sequential recommendation contexts: embedding-based representations hold their quality advantage not just in average-case performance but under varied data conditions.

Taken together, these findings speak to both research objectives stated in the introduction. The first objective, comparing recommendation quality, is answered by Table 4: Sentence-Transformer outperforms TF-IDF, and the margin is more pronounced in ranking quality than in retrieval accuracy. The second objective, assessing representational stability, is answered by Table 5: semantic embeddings degrade far less across corpus variations than lexical frequency vectors. The Streamlit prototype demonstrated that both findings hold under live interaction conditions, not just controlled batch evaluation, which strengthens the practical validity of the conclusions.

That said, this study does not claim Sentence-Transformer is categorically superior in all recommendation contexts. Both methods achieved Precision@3 values above 0.83, which indicates that even TF-IDF can perform reasonably well in this domain given the structured nature of fragrance vocabulary. The difference becomes most visible under the specific pressure of cross-lingual retrieval and ranking quality, which is exactly the pressure this study was designed to test.

4. Conclusion

What three research questions set out to establish, this study answered with measurable results. The central question was whether the choice of text representation method changes the quality of recommendations when queries arrive in Indonesian and the product corpus is entirely in English. It does, and the way it changes things reveals something more nuanced than a simple win-loss verdict.

Sentence-Transformer, using the pre-trained paraphrase-multilingual-MiniLM-L12-v2 checkpoint, outperformed TF-IDF on both evaluation metrics drawn from 180 graded relevance judgments across 30 respondents. On Precision@3 the margin was 0.867 against 0.833, a difference narrow enough that TF-IDF could not be called a failure. On nDCG@3 the margin was 0.912 against 0.850, nearly double the width of the Precision gap. That divergence is the most telling finding in the entire study. Both methods retrieved relevant perfumes at a broadly similar rate, but TF-IDF consistently misranked them, placing items with lower relevance scores above items that respondents cared about more. For any real user, that ranking error is what degrades the experience. The Streamlit prototype confirmed this under live interaction conditions, validating that the measured difference is not an artifact of controlled evaluation.

The stability analysis extended the finding further. When the fragrance corpus was divided into a training partition of 33,692 records and a test partition of 8,423 records, TF-IDF's mean cosine similarity degraded by 0.0783 across partitions. Sentence-Transformer degraded by only 0.0218, roughly one-quarter of that amount. Dense multilingual embeddings carry semantic meaning in a way that is structurally insensitive to

how large or small the candidate pool is. Lexical frequency vectors do not have that property, because their scores depend on vocabulary overlap and vocabulary thins as the corpus shrinks.

The contribution of this work to the field is specific rather than sweeping. No prior study had placed TF-IDF and Sentence-Transformer in a controlled head-to-head comparison within a cross-lingual Indonesian-to-English fragrance retrieval setting, without any translation module and evaluated through real user judgments. The results confirm that the Transformer-based semantic advantage documented in general multilingual matching tasks by Bhangale et al. [13] and in fragrance-specific note estimation by Kim et al. [14] extends to the full recommendation pipeline. For developers building multilingual product recommendation systems under resource constraints, this offers a concrete and evidence-backed case for adopting compact multilingual embedding models over classical frequency-based approaches, even without domain-specific fine-tuning.

Three directions remain open for follow-up research. The most impactful would be fine-tuning the multilingual embedding model on an Indonesian fragrance description corpus, since the checkpoint used here was trained on general-domain multilingual data and never exposed to perfume-specific vocabulary in either language. Domain adaptation would likely sharpen both metrics. Expanding the respondent pool well beyond 30 participants, and diversifying across age groups, purchase behaviors, and familiarity with perfumery, would also strengthen the generalizability of the evaluation results. Finally, a hybrid architecture that fuses TF-IDF's lexical precision with Sentence-Transformer's semantic breadth is a promising unexplored direction, potentially offering retrieval behavior that neither method achieves independently.

5. References

[1] Mandalika, "Tren Market Parfum Tahun 2025," 2025. [Online]. Available: https://mandalikaperfume.co.id/tren-market-parfum-tahun-2025/

[2] L. Jensen, "US Beauty Industry Sales Grow for the Fourth Consecutive Year, Circana Reports," 2025. [Online]. Available: https://www.circana.com/post/us-beauty-industry-sales-grow

[3] F. A. T. Tobing *et al.*, "Men's Perfume Recommendation System Using Analytic Hierarchy Process (AHP) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) Method," *International Journal of Science, Technology and Management*, vol. 4, no. 6, pp. 1552–1559, 2023,

[4] N. S. Agashe, "Product Recommender Chat Bot," *International Journal of Engineering Research and Technology (IJERT)*, vol. 10, no. 6, 2021, [Online]. Available: https://www.ijert.org/product-recommender-chat-bot

[5] L. D. Nurmuthia, R. Imtiyaz, and N. T. M. Sagala, "Personalized Perfume Recommendations Based on User Descriptions Using Cosine and Jaccard Similarity," *Procedia Comput. Sci.*, vol. 269, pp. 1097–1106, 2025, doi: 10.1016/j.procs.2025.09.051.

[6] A. A. Huda, R. Fajarudin, and A. Hadinegoro, "Sistem Rekomendasi Content-Based Filtering Menggunakan TF-IDF Vector Similarity untuk Rekomendasi Artikel Berita," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, pp. 1679–1686, 2022, doi: 10.47065/bits.v4i3.2511.

[7] R. Priskila, N. N. K. Sari, and P. B. A. A. Putra, "Implementasi Content-Based Filtering Menggunakan TF-IDF dan Cosine Similarity untuk Sistem Rekomendasi Resep Masakan," *Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika*, vol. 18, no. 1, 2024, doi: 10.47111/jti.v18i1.12543.

[8] S. Oyadila, D. Abdullah, and A. Razi, "Implementasi Content-Based Filtering dengan TF-IDF dan Cosine Similarity untuk Sistem Rekomendasi Destinasi Wisata di Aceh Tengah," *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 10, no. 2, pp. 1329–1339, 2025, doi: 10.36341/rabit.v10i2.6532.

[9] T. R. Masuzzahra, K. Umam, H. Mustofa, and M. R. Handayani, "HANA: An AI Chatbot for Islamic Jurisprudence on Menstruation Using SBERT and TF-IDF," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, pp. 1013–1024, 2025, doi: 10.30871/jaic.v9i3.9449.

[10] H. A. Malaiga and S. D. Sancoko, "Development of the Scentify Application as a Web and Mobile-Based Perfume Recommendation System Using Content-Based Filtering Based on User Preferences," *J-INTECH (Journal of Information and Technology)*, vol. 13, no. 2, pp. 354–364, 2025, doi: 10.32664/j-intech.v13i02.2143.

[11] M. T. S. Phalle and P. S. Bhushan, "Content Based Filtering and Collaborative Filtering: A Comparative Study," *Journal of Advanced Zoology*, vol. 45, pp. 96–100, 2024, doi: 10.53555/jaz.v45is4.4158.

[12] R. Birwadkar, "A Hybrid Plagiarism Detection Framework Using Lexical and Semantic Similarity with Lightweight Sentence Transformers," *Journal of Computer and Forensic Sciences*, vol. 4, no. 2, pp. 3–16, 2025, doi: 10.5937/jcfs4-64419.

[13] C. S. Bhangale, P. A. Gabhane, and A. K. Madasamy, "Fine-Tuned Sentence Transformer for Multilingual Job Title Matching," in *CEUR Workshop Proceedings*, 2025, pp. 4411–4420.

[14] J. Kim, K. Oh, and B. S. Oh, "An NLP-Based Perfume Note Estimation Based on Descriptive Sentences," *Applied Sciences*, vol. 14, no. 20, p. 9293, Oct. 2024, doi: 10.3390/app14209293.

[15] Streamlit, "Streamlit Documentation," 2024.

[16] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive Text Embedding Benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Dubrovnik, Croatia, May 2023, pp. 2014–2037. doi: 10.48550/arXiv.2210.07316.

[17] J. U. A. Fauzi and A. N. Rohman, "Recommendation System Yogyakarta Tourism Using TF-IDF and Cosine Similarity Methods with Word Normalizer," *Journal of Applied Informatics and Computing (JAIC)*, vol. 10, no. 1, pp. 935–945, 2026, doi: 10.30871/jaic.v10i1.11751.

[18] M. Farhan, M. R. Andreawan, R. Antonius, and N. D. Ulhaq, "Evaluasi dan Pengembangan Sistem Rekomendasi Game Berbasis Content-Based Filtering dengan TF-IDF dan Cosine Similarity," *BINER: Jurnal Ilmu Komputer, Teknik dan Multimedia*, vol. 2, no. 4, pp. 400–408, 2024,

[19] R. Al Rasyid and D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian pada Dataset Destinasi Wisata," *Jurnal Teknologi Informasi dan Komunikasi (JTIK)*, vol. 8, no. 1, pp. 170–178, 2024, doi: 10.35870/jtik.v8i1.1416.

[20] K. Sarode and S. R. Javaji, "Multi-BERT for Embeddings for Recommendation System," 2023. [Online]. Available: https://arxiv.org/abs/2308.13050

[21] S. R. Javaji and K. Sarode, "Enriching Sequential Recommendations with Contextual Auxiliary Information Using Sentence-BERT," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 12, 2024,

Acknowledgment

The authors express sincere gratitude to all thirty respondents who voluntarily participated in the user study, contributed their time to evaluate the recommendation outputs, and provided the graded relevance judgments that form the empirical backbone of this research. Their willingness to engage with the live system made the evaluation meaningful in a way that no synthetic benchmark could replicate. This research was carried out as part of the undergraduate thesis program at the Faculty of Computer Science, Universitas Duta Bangsa Surakarta. The authors also acknowledge the use of Claude AI (Anthropic) as an assistive tool in the English translation and language editing process of this manuscript.