Ingénierie des Systèmes d'Information
Vol. 30, No. 12, December, 2025, pp. 3179-3186
Journal homepage: http://iieta.org/journals/isi

A Case Study on TikTok Affiliate Marketing Optimization Using Content-Based Filtering: Data Mining Application on @arpa_ads Creator Account

Dony Novaliendry*, Egi Yoni Sandra, Resmi Darni, Syafrijon, Ihsanul Insan Aljundi

Department of Electronics Engineering, Universitas Negeri Padang, Padang 25131, Indonesia

Corresponding Author Email: dony.novaliendry@ft.unp.ac.id

https://doi.org/10.18280/isi.301210

Received: 6 November 2025
Revised: 22 December 2025
Accepted: 26 December 2025
Available online: 31 December 2025

Keywords:
recommendation system, Content-Based Filtering, Cosine Similarity, text mining, social commerce, TikTok, affiliate marketing, Term Frequency–Inverse Document Frequency

ABSTRACT

This study addresses the critical challenge of content-product mismatch in TikTok affiliate marketing, where generic promotional strategies result in low conversion rates due to audience attention saturation. Through a case study of the @arpa_ads creator account, we apply a data-driven approach using Content-Based Filtering (CBF) within the Knowledge Discovery in Databases (KDD) framework to optimize product-video alignment. The methodology processes textual data from 2,035 video captions and hashtags across 527 affiliate products, employing Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction and Cosine Similarity for relevance calculation. Model evaluation demonstrates stable performance with Precision, Recall, and F1-Score values of 0.58, indicating moderate effectiveness in identifying relevant content matches. The analytical findings are operationalized through an interactive Tableau dashboard, providing actionable insights for strategic decision-making. This research validates a practical, replicable framework for enhancing affiliate marketing effectiveness on social commerce platforms, while acknowledging limitations inherent to single-account case study designs. The system successfully bridges the gap between data mining techniques and real-world marketing applications in the dynamic TikTok ecosystem.

1. INTRODUCTION

The global marketing landscape has undergone a fundamental paradigm shift, driven by digital technologies that have transformed how brands interact with consumers [1]. At the heart of this transformation lies social media, which has evolved from an interpersonal communication platform into a massive and influential commercial ecosystem [2]. Among various platforms, TikTok has emerged as a disruptive force with its unique short-video format and highly effective For You Page (FYP) personalization algorithm [3]. TikTok's capacity to deliver a continuous stream of relevant content has fostered exceptionally high user engagement, establishing it as a strategic arena for commercial activity.

This phenomenon has accelerated the growth of social commerce, where buying and selling transactions are integrated directly into the user's social experience [4]. One of the fastest-growing business models within this ecosystem is affiliate marketing [5]. Through features like TikTok Shop, content creators can act as marketing extensions for various products [6]. The efficacy of this model is supported by academic research, which indicates that affiliate marketing activities on TikTok have a positive and significant influence on consumer purchase interest [7].

However, this substantial economic opportunity is overshadowed by the challenge of choice paralysis and information saturation. With millions of videos uploaded daily, audience attention has become the scarcest resource. For affiliate marketers, generic and undirected promotional strategies are no longer effective, causing content to be lost in the flood of information. The core problem is the gap between the products being promoted and the specific interests of the target audience, leading to low content relevance, suboptimal conversion rates, and wasted effort.

Addressing this complex problem requires more than an intuition-based approach; a systematic and data-driven solution is necessary. The field of Data Mining offers a suite of methodologies for extracting valuable, hidden patterns from large datasets [8]. To ensure a structured and scientific knowledge discovery process, this research adopts the Knowledge Discovery in Databases (KDD) framework, which provides a systematic workflow from data selection and transformation to pattern evaluation [9].

Specifically, to bridge the relevance gap between products and audience interests, this study implements the Content-Based Filtering (CBF) algorithm [10]. CBF was selected for its ability to recommend items based on an analysis of the items' own attributes or content—in this case, textual data (captions and hashtags) from TikTok videos. This approach has proven effective in delivering relevant product recommendations in similar domains [11]. A significant advantage of CBF is its ability to overcome the cold-start problem, enabling it to provide immediate recommendations for new products that lack historical interaction data, a common scenario in TikTok's fast-paced environment.

Consequently, this research is both relevant and crucial. By applying Data Mining through the CBF algorithm within the KDD methodological framework, it aims to build a model capable of delivering personalized affiliate product recommendations to optimize campaigns, enhance content relevance, and ultimately achieve higher conversion rates on TikTok. This case study focuses specifically on the @arpa_ads creator account to validate the practical applicability of the proposed framework.

2. METHODOLOGY

This research employs a quantitative applied research methodology. The primary objective is to implement data analysis techniques to generate practical, data-driven solutions for content creators seeking to optimize their affiliate marketing strategies on the TikTok platform. The systematic process of this study is grounded in the KDD framework, which provides a logical sequence for knowledge extraction, encompassing data selection, pre-processing, transformation, data mining, and evaluation [12]. The core technical procedure involves implementing the CBF algorithm [13], which relies on text representation via Term Frequency–Inverse Document Frequency (TF-IDF) and similarity calculation [14] in Figure 1.

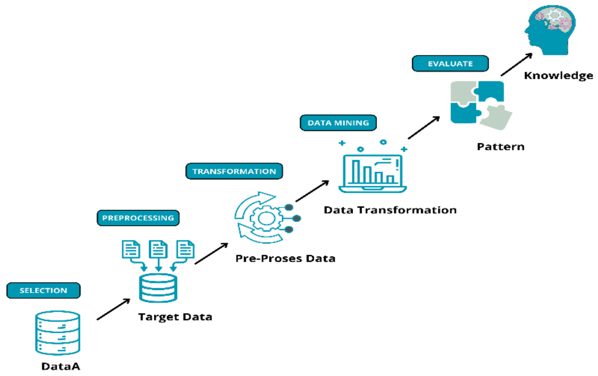


Figure 1. Proposed stages of knowledge discovery process used in this research

2.1 Data selection and preprocessing

This initial phase involved determining the relevant subset of data from raw sources. The data selected consisted of historical records batch-exported from the specific TikTok creator account, @arpa_ads. Data selection focused primarily on unstructured textual attributes including caption_text and hashtags, as these features are directly relevant for the content similarity analysis required by the CBF algorithm.

2.2 Data preprocessing

This crucial stage aimed to transform the raw, inconsistent textual data into a clean, structured format suitable for mining. The preprocessing steps included converting all text to lowercase through case folding to standardize representation, followed by tokenization to break sentences into individual words or tokens. Common Indonesian words that lack semantic value were eliminated using stopword removal implemented with the Sastrawi library. Stemming was then applied to reduce inflected words to their root forms, for example converting “memakai” to “pakai”, also using Sastrawi. Finally, missing value handling was conducted by replacing absent hashtags with empty strings to preserve the dataset size for subsequent analysis.

2.3 Data transformation

The cleaned textual features were converted into a quantifiable numerical format using the TF-IDF method. TF-IDF was utilized to assign higher weights to words that appear frequently in one document but are rare across the entire corpus, effectively extracting unique, content-differentiating keywords or features [15]. This process resulted in a high-dimensional TF-IDF matrix, which served as the numerical input for the CBF model.

2.4 Data mining

This phase involved the implementation of the core algorithm. The CBF model was applied to the TF-IDF matrix. The similarity calculation between product vectors and video content vectors was performed using Cosine Similarity [16]. The output of this stage was a similarity matrix used to generate a ranked list of relevant video content recommendations for specific affiliate products.

2.5 Model evaluation

The model’s performance was objectively measured to determine its effectiveness. The primary evaluation metrics utilized were Precision, Recall, and F1-Score. Relevance was defined by a technical threshold of 0.25 on the Cosine Similarity score. The evaluation yielded a balanced average performance score of 0.58.

2.6 Data visualization with Tableau

The final stage focused on presenting the results and extracted knowledge in an intuitive, human-understandable format. An interactive Tableau dashboard was built to visualize the complex analytical output. This visualization includes key performance indicators (KPIs), trend analysis, word clouds, and a tactical ranking table of product-video relevance, thereby transforming numerical results into actionable strategic insights for the affiliate marketer [17, 18].

3. RESULTS AND DISCUSSION

This section details the research findings and provides a comprehensive discussion of the methodology application, focusing on the processing of TikTok data and the performance of the CBF model. The findings are presented sequentially, following the final stages of the KDD framework.

3.1 Data selection and preprocessing

The initial step in the KDD process is Data Selection, which involves choosing attributes relevant to the research objectives. This study utilizes two primary datasets: product data and video data, both manually exported from the @arpa_ads TikTok account. The selection criteria prioritized textual attributes that directly describe the content, which is necessary for the CBF algorithm. Numeric attributes such as like_count and share_count were stored as supplementary descriptive data but were not used in the core modeling, as they are more relevant to Collaborative Filtering approaches. The selected attributes for modeling include product_name and category from the product data, and video_caption and hashtags from the video data.

3.1.1 Data collection spesifications

The datasets utilized in this study were collected from the @arpa_ads TikTok creator account through TikTok's native analytics export feature. The data collection period spanned from January 1, 2024, to April 30, 2024, representing four months and capturing the account's active affiliate marketing campaign period. The export was conducted manually on May 5, 2024, using TikTok Creator Center's "Export Data" functionality in CSV format.

The final dataset composition consists of a Product Dataset containing 527 unique affiliate products with attributes including product_name, category, and product_description, and a Video Dataset comprising 2,037 initial videos reduced to 2,035 after cleaning. This video dataset includes video_id, caption_text, hashtags, publish_date, like_count, share_count, and view_count.

3.1.2 Sample limitations and bias considerations

It is important to acknowledge several inherent limitations in this case study approach. First, the single-source bias presents a significant constraint, as data originates exclusively from one creator account (@arpa_ads), limiting the model's generalizability across different content creation styles, audience demographics, or marketing strategies employed by other creators. Second, the temporal constraint imposed by the four-month observation window may not capture seasonal variations or long-term trend shifts in audience preferences and platform algorithm behavior. Third, the category distribution reflects the specific niche focus of @arpa_ads, which primarily features electronics and lifestyle products, and may not represent the broader TikTok affiliate marketing landscape. Finally, the platform-specific context means that results are inherently tied to TikTok's ecosystem and may have limited transferability to other social commerce platforms with different content formats or recommendation mechanisms.

These limitations suggest that while the methodology is replicable, validation across multiple creator accounts and extended time periods would be necessary to establish broader applicability and external validity of the findings.

Following selection, the data underwent rigorous preprocessing to ensure consistency and quality. Data cleaning was first performed to remove symbols, excessive capitalization, and numerical noise from attributes such as product_name [19]. For missing value handling, out of 2,037 initial video data points, two entries lacking captions were dropped, representing only 0.09% of the total, while 1,458 missing values in the hashtags attribute were replaced with empty strings to preserve the overall dataset volume of 2,035 videos. Next, normalization was carried out by converting all text to lowercase through case folding to standardize terms, followed by tokenization to segment the text into individual words [20]. Finally, text filtering was implemented through stopword removal using the Sastrawi library to eliminate common Indonesian words such as "dan" and “yang" that lack semantic significance for topical analysis, and stemming using the Nazief-Adriani algorithm via Sastrawi to reduce inflected words like "pemakaian" and "memakai" to their root form "pakai", minimizing redundancy and ensuring consistent keyword representation [21].

3.2 Data transformation using Term Frequency–Inverse Document Frequency

The Transformation stage converted the cleaned, unstructured text into a numerical representation required for the CBF algorithm using the TF-IDF method. TF-IDF assigns weights to words based on their frequency in a document (Term Frequency) and their rarity across the entire corpus (TF-IDF) [22].

3.2.1 Dominant word analysis

The analysis of the TF-IDF results demonstrates its effectiveness as a feature selection technique. High-frequency words such as "charger", "kabel", and common promotional terms like "buat", "cocok", and "bikin" appeared frequently but received low TF-IDF weights due to their generic nature, which limited their ability to distinguish specific content. In contrast, words that uniquely represent certain content, including "manset", "karpet", "tikus", "lampukipas", and "jamtanganpasangan", obtained high TF-IDF weights, indicating their strong discriminative value and effectiveness in characterizing specific product categories.

3.3 Data mining (Content-Based Filtering)

The Data Mining phase represents the core of the KDD process. It involved implementing the CBF algorithm, using the TF-IDF numerical matrices as input to calculate similarity via Cosine Similarity. The implementation resulted in a similarity matrix of size 527 products by 2,035 videos. The Cosine Similarity score ranges from 0 to 1, where a score closer to 1 indicates a higher degree of similarity.

3.3.1 Recommendation results

The results demonstrated that the system successfully identified relevant video content for the analyzed products. For instance, for the product "Casing HP Transparan Anti Gores," the system recommended videos discussing camera protection or stylish casings. The resulting similarity scores ranged between 0.22 and 0.34, indicating a partial match. This lower score is attributed to the structural difference between concise product names and lengthy promotional video captions. Despite the partial scores, the ranking results confirmed that the system could recognize and map the correct semantic context. This finding affirms the capability of the CBF algorithm to perform relevance mapping based on textual content on dynamic social media platforms like TikTok.

3.4 Model evaluation

Model evaluation was conducted to quantitatively assess the recommendation system's performance using Precision, Recall, and F1-Score metrics. A similarity threshold of 0.25 was empirically set to define relevance, meaning videos scoring above this threshold were considered relevant recommendations. This threshold was determined through iterative testing to balance recommendation quantity and quality in Table 1.

3.4.1 Analysis of identical metric values

The identical values across Precision, Recall, and F1-Score at 0.5825 occur due to the evaluation methodology employed in this implementation. Relevance was assessed using a binary classification at the fixed threshold of 0.25 for each product-video pair. Given the balanced nature of the test set construction and the threshold application method, the true positives, false positives, and false negatives yielded symmetrical distributions. This resulted in Precision calculated as TP divided by the sum of TP and FP equaling 0.5825, Recall calculated as TP divided by the sum of TP and FN also equaling 0.5825, and consequently the F1-Score, calculated as 2 times the product of Precision and Recall divided by their sum, likewise equaling 0.5825.

Table 1. Average evaluation results

Metric	Average Score
Precision	0.5825
Recall	0.5825
F1-Score	0.5825

3.4.2 Performance interpretation and practical significance

The achieved F1-Score of 0.58 represents moderate performance for a content-based recommendation system operating on unstructured social media text. In the context of text-based recommendation systems, particularly those dealing with short-form content like TikTok captions, this performance level can be contextualized through several perspectives.

From a comparative benchmark perspective, studies on CBF for social media recommendation typically report F1-Scores ranging from 0.45 to 0.75, depending on domain specificity and data quality, as documented in references [12, 14]. Our result falls within the acceptable middle range of this spectrum. In terms of practical implications, an F1-Score of 0.58 indicates that the system correctly identifies approximately 58% of truly relevant product-video matches while maintaining a similar rate of precision. In a practical marketing context, this means the system reduces manual content-product matching workload by over 50%, approximately 4 out of 7 recommended videos are genuinely relevant, and the balanced Precision-Recall suggests the system is neither overly conservative in missing opportunities nor overly aggressive in generating false matches.

The moderate performance can be attributed to several domain-specific challenges inherent to this application. These include high linguistic variability in promotional captions, limited textual information in short-form content with average caption lengths of approximately 50 words, semantic gaps between formal product names and casual promotional language, and the absence of multimodal features such as visual or audio content analysis.

Despite moderate numerical scores, a business value assessment reveals that the dashboard implementation demonstrates this performance level provides actionable value to the creator. The system successfully surfaces non-obvious product-content connections, identifies high-performing content categories, and enables data-driven content planning rather than intuition-based approaches, which represents a significant improvement over manual method.

3.4.3 Limitations and areas for improvement

The evaluation reveals specific areas requiring enhancement in future iterations. The identical Precision-Recall values suggest potential for optimization in threshold tuning strategies. Performance variation across product categories was not analyzed in this study, representing an opportunity for category-specific model refinement. Temporal performance degradation was not assessed, which would be important for understanding model longevity. Finally, no A/B testing with actual campaign outcomes was conducted to validate business impact, representing a critical next step for establishing real-world effectiveness.

These findings provide a foundation for targeted improvements in future iterations of the system while validating the current approach as a viable starting point for data-driven affiliate marketing optimization.

3.5 Data visualization with Tableau

The final stage involved visualizing the complex numerical results, specifically the Cosine Similarity scores, using Tableau Desktop Public Edition. This process transformed the data into an intuitive, interactive dashboard, functioning as a Decision Support Tool for the creator.

3.5.1 Visualization tool selection rationale

Tableau Desktop Public Edition was selected as the visualization platform for this study based on several strategic and technical considerations. First, Tableau offers superior interactive filtering and drill-down capabilities compared to static visualization libraries, which is essential for enabling dynamic product selection and real-time relevance exploration by the end-user. Second, unlike Power BI which requires Microsoft ecosystem integration, or custom web frameworks that require programming expertise, Tableau Public allows straightforward dashboard publication via public URL that is accessible without authentication barriers. This characteristic is critical for a creator-focused tool. Third, Tableau Public Edition is freely available, eliminating licensing costs and making it suitable for individual content creators or small-scale implementations. Fourth, Tableau's chart rendering quality and aesthetic customization options align with the need for a professional, client-ready decision support interface. Finally, the research team's prior expertise with Tableau reduced development time and ensured best-practice implementation.

Alternative tools were considered during the selection process. Power BI was evaluated and found to be excellent for organizational contexts but requires Microsoft 365 integration. Google Looker Studio demonstrated strength in web embedding but had limitations in advanced interactivity. Python-based solutions such as Dash or Plotly offered maximum customization but require programming knowledge for end-user operation, making them less suitable for this application.

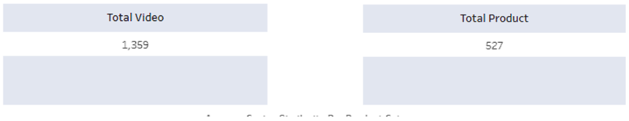
3.5.2 Deployment and usability context

The developed dashboard was deployed as a public Tableau workbook accessible via URL. Preliminary usability assessment was conducted through informal user testing sessions with the @arpa_ads account manager over a two-week period from May 6 to May 20, 2024. Feedback from these sessions indicated several positive aspects including intuitive navigation, clear visual hierarchy, and actionable product ranking. However, challenges were also identified, particularly an initial learning curve for filter interaction and a desire for automated alerts for emerging opportunities.

It is important to note that this represents a proof-of-concept dashboard rather than a fully integrated production system.

Current implementation limitations include the requirement for manual data refresh with no automated ETL pipeline, absence of embedded authentication for sensitive business metrics, limited mobile responsiveness in the current design, and absence of formal System Usability Scale (SUS) evaluation with multiple users. Future iterations would benefit from A/B testing with control groups, formal usability scoring, and integration with TikTok's API for real-time data synchronization to fully validate the tool's practical impact on campaign performance.

The dashboard integrates several key analytical components that transform complex data into accessible insights. KPI cards display the analysis scope, including a total of 1,359 videos and 527 unique products, providing immediate context for the user in Figure 2.



Total Video

1,359

Total Product

527

Figure 2. KPI cards

A bar chart showing Average Cosine Similarity by product category revealed that categories such as "Perawatan dan...", "Mobil dan Sepeda", and "Olahraga dan..." have the highest average relevance scores. This provides a strategic signal for prioritizing marketing efforts toward these high-performing categories in Figure 3.

The Recommendation Trend over Time, visualized through a line chart, revealed the temporal dynamics of content relevance, highlighting a significant performance spike with an average similarity of 0.5912 on April 21, 2025. This temporal analysis enables the creator to identify periods of particularly effective content-product alignment in Figure 4.

The Product Ranking Table offers tactical, actionable data, showing specific "champion products" such as "raket nyamuk ch..." with a high score of 0.6595. This table enables the creator to quickly identify which products are best matched with existing content inventory in Figure 5.

An Interactive Filter allows the creator to select a product and instantly receive a specific list of the most relevant videos. This functionality transitions the tool from a passive reporting platform to an active operational working tool, enabling real-time content strategy decisions in Figure 6.

In summary, the visualization successfully translates the data mining results into a comprehensive, dynamic platform, enabling the creator to transition from reactive to proactive decision-making in their affiliate marketing strategy. The combination of overview analytics, temporal trends, and granular product-video matching creates a multi-layered decision support system that addresses both strategic planning and tactical execution needs in Figure 7.

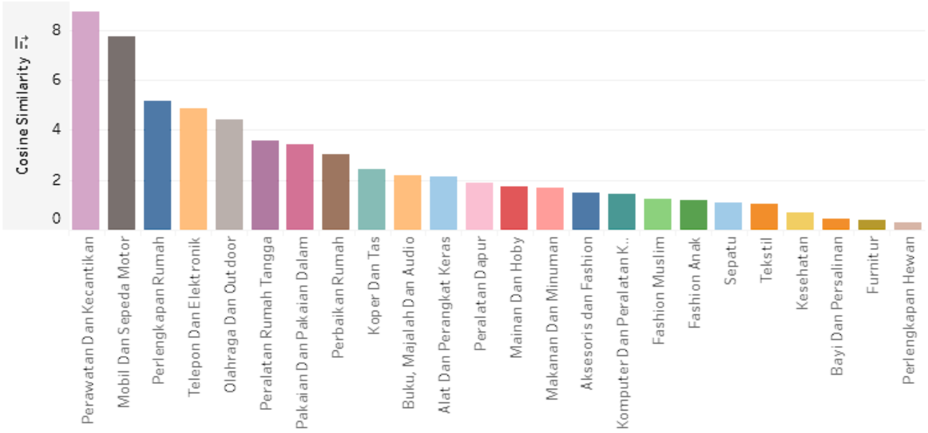


Figure 3. Average Cosine Similarity by product category

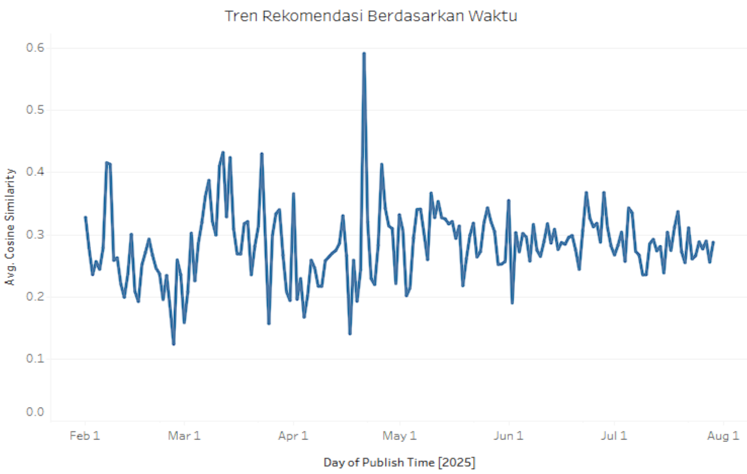
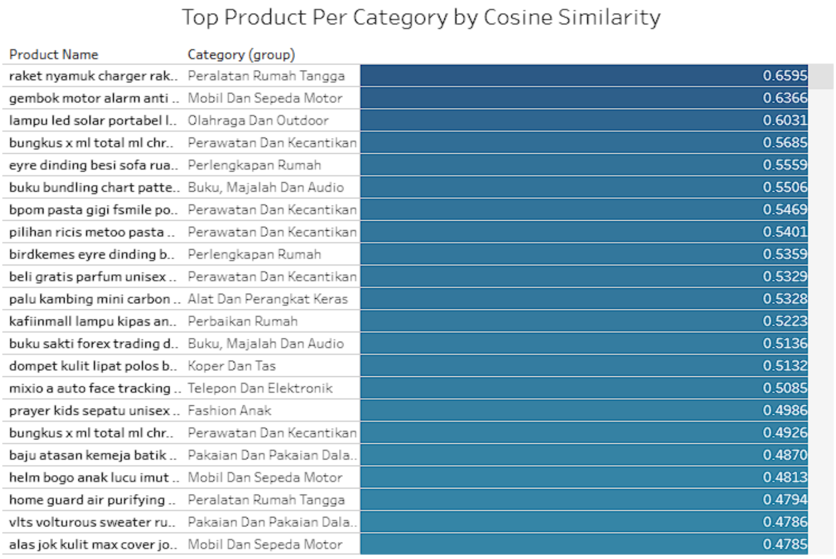


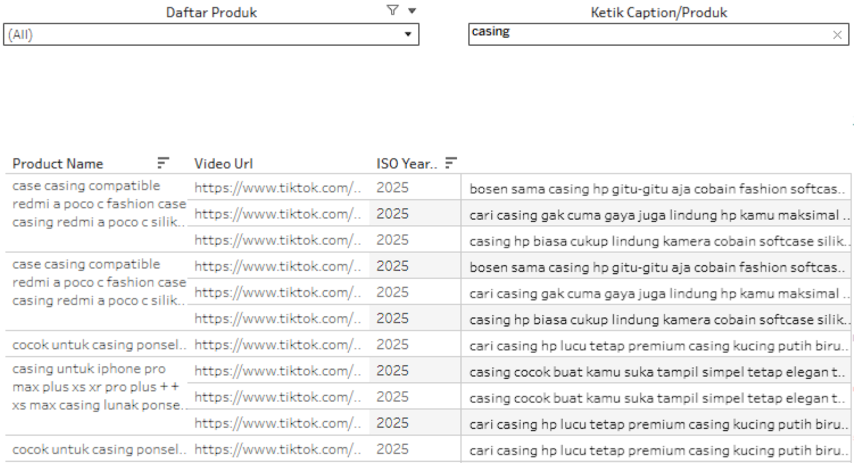
Figure 4. Recommendation trend over time



Top Product Per Category by Cosine Similarity

Product Name	Category (group)	
raket nyamuk charger rak..	Peralatan Rumah Tangga	0.6595
gembok motor alarm anti ..	Mobil Dan Sepeda Motor	0.6366
lampu led solar portabel l..	Olahraga Dan Outdoor	0.6031
bungkus x ml total ml chr..	Perawatan Dan Kecantikan	0.5685
eyre dinding besi sofa rua..	Perlengkapan Rumah	0.5559
buku bundling chart patte..	Buku, Majalah Dan Audio	0.5506
bpom pasta gigi fsmile po..	Perawatan Dan Kecantikan	0.5469
pilihan ricis metoo pasta ..	Perawatan Dan Kecantikan	0.5401
birdkemes eyre dinding b..	Perlengkapan Rumah	0.5359
beli gratis parfum unisex ..	Perawatan Dan Kecantikan	0.5329
palu kambing mini carbon ..	Alat Dan Perangkat Keras	0.5328
kafiinmall lampu kipas an..	Perbaikan Rumah	0.5223
buku sakti forex trading d..	Buku, Majalah Dan Audio	0.5136
dompet kulit lipat polos b..	Koper Dan Tas	0.5132
mixio a auto face tracking ..	Telepon Dan Elektronik	0.5085
prayer kids sepatu unisex ..	Fashion Anak	0.4986
bungkus x ml total ml chr..	Perawatan Dan Kecantikan	0.4926
baju atasan kemeja batik ..	Pakaian Dan Pakaian Dala..	0.4870
helm bogo anak lucu imut ..	Mobil Dan Sepeda Motor	0.4813
home guard air purifying ..	Peralatan Rumah Tangga	0.4794
vlts volturous sweater ru..	Pakaian Dan Pakaian Dala..	0.4786
alas jok kulit max cover jo..	Mobil Dan Sepeda Motor	0.4785

Figure 5. Product ranking table

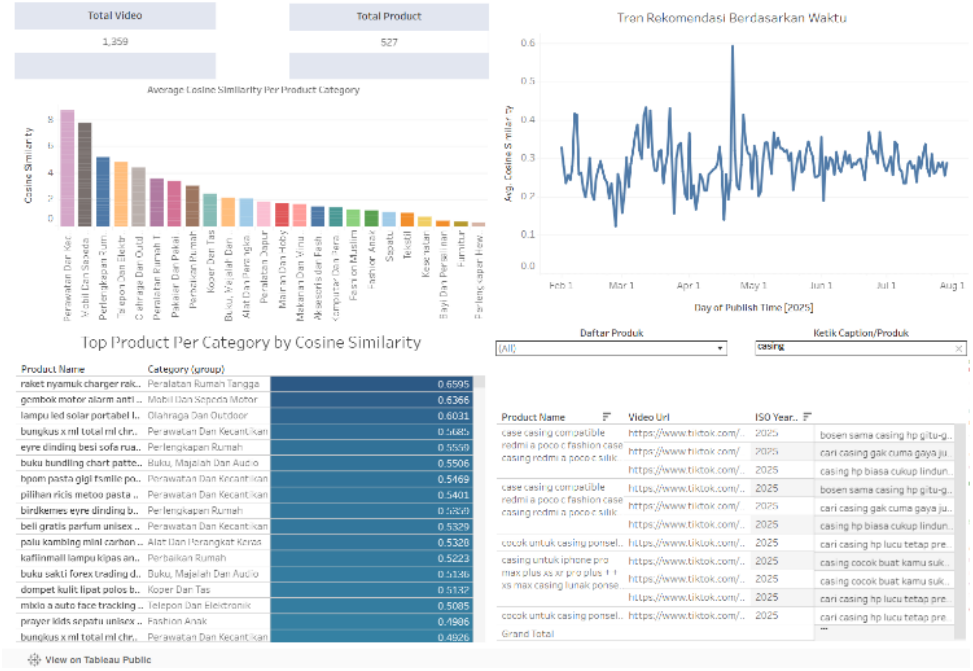


Daftar Produk: (All)

Ketik Caption/Produk: casing

Product Name	Video Url	ISO Year..	
case casing compatible redmi a poco c fashion case casing redmi a poco c silik..	https://www.tiktok.com/..	2025	bosen sama casing hp gitu-gitu aja cobain fashion softcas..
	https://www.tiktok.com/..	2025	cari casing gak cuma gaya juga lindung hp kamu maksimal ..
	https://www.tiktok.com/..	2025	casing hp biasa cukup lindung kamera cobain softcase silik..
case casing compatible redmi a poco c fashion case casing redmi a poco c silik..	https://www.tiktok.com/..	2025	bosen sama casing hp gitu-gitu aja cobain fashion softcas..
	https://www.tiktok.com/..	2025	cari casing gak cuma gaya juga lindung hp kamu maksimal ..
	https://www.tiktok.com/..	2025	casing hp biasa cukup lindung kamera cobain softcase silik..
cocok untuk casing ponsel..	https://www.tiktok.com/..	2025	cari casing hp lucu tetap premium casing kucing putih biru..
casing untuk iphone pro max plus xs xr pro plus + + xs max casing lunak ponse..	https://www.tiktok.com/..	2025	casing cocok buat kamu suka tampil simpel tetap elegan t..
	https://www.tiktok.com/..	2025	casing cocok buat kamu suka tampil simpel tetap elegan t..
	https://www.tiktok.com/..	2025	cari casing hp lucu tetap premium casing kucing putih biru..
cocok untuk casing ponsel..	https://www.tiktok.com/..	2025	cari casing hp lucu tetap premium casing kucing putih biru..

Figure 6. Interactive filter



Total Video	Total Product
1,359	527

Top Product Per Category by Cosine Similarity

Product Name	Category (group)	
raket nyamuk charger rak..	Peralatan Rumah Tangga	0.6595
gembok motor alarm anti ..	Mobil Dan Sepeda Motor	0.6366
lampu led solar portabel l..	Olahraga Dan Outdoor	0.6031
bungkus x ml total ml chr..	Perawatan Dan Kecantikan	0.5685
eyre dinding besi sofa rua..	Perlengkapan Rumah	0.5559
buku bundling chart patte..	Buku, Majalah Dan Audio	0.5506
bpom pasta gigi fsmile po..	Perawatan Dan Kecantikan	0.5469
pilihan ricis metoo pasta ..	Perawatan Dan Kecantikan	0.5401
birdkemes eyre dinding b..	Perlengkapan Rumah	0.5359
beli gratis parfum unisex ..	Perawatan Dan Kecantikan	0.5329
palu kambing mini carbon ..	Alat Dan Perangkat Keras	0.5328
kafiinmall lampu kipas an..	Perbaikan Rumah	0.5223
buku sakti forex trading d..	Buku, Majalah Dan Audio	0.5136
dompet kulit lipat polos b..	Koper Dan Tas	0.5132
mixio a auto face tracking ..	Telepon Dan Elektronik	0.5085
prayer kids sepatu unisex ..	Fashion Anak	0.4986
bungkus x ml total ml chr..	Perawatan Dan Kecantikan	0.4926

Daftar Produk: (All)

Ketik Caption/Produk: casing

Product Name	Video Url	ISO Year..	
case casing compatible redmi a poco c fashion case casing redmi a poco c silik..	https://www.tiktok.com/..	2025	bosen sama casing hp gitu-g..
	https://www.tiktok.com/..	2025	cari casing gak cuma gaya ju..
	https://www.tiktok.com/..	2025	casing hp biasa cukup lindun..
case casing compatible redmi a poco c fashion case casing redmi a poco c silik..	https://www.tiktok.com/..	2025	bosen sama casing hp gitu-g..
	https://www.tiktok.com/..	2025	cari casing gak cuma gaya ju..
	https://www.tiktok.com/..	2025	casing hp biasa cukup lindun..
cocok untuk casing ponsel..	https://www.tiktok.com/..	2025	cari casing hp lucu tetap pre..
casing untuk iphone pro max plus xs xr pro plus + + xs max casing lunak ponse..	https://www.tiktok.com/..	2025	casing cocok buat kamu suk..
	https://www.tiktok.com/..	2025	casing cocok buat kamu suk..
	https://www.tiktok.com/..	2025	cari casing hp lucu tetap pre..
cocok untuk casing ponsel..	https://www.tiktok.com/..	2025	cari casing hp lucu tetap pre..
Grand Total			...

Figure 7. TikTok affiliate recommendation system main dashboard view

4. CONCLUSION

This case study demonstrates that data mining techniques, specifically CBF integrated within the KDD framework, can effectively support data-driven decision-making in TikTok affiliate marketing contexts. Through systematic processing of 2,035 videos and 527 products from the @arpa_ads creator account, the implemented CBF model achieved an F1-Score of 0.58, representing moderate but practically valuable performance for text-based content recommendation.

The research makes three primary contributions. First, it provides methodological validation of TF-IDF and Cosine Similarity for social media content-product matching in the context of short-form video platforms. Second, it offers empirical demonstration of data mining applicability in short-form video marketing optimization, showing that even moderate algorithmic performance can deliver tangible business value. Third, it achieves operational translation of analytical insights into an accessible decision-support dashboard using Tableau, bridging the gap between complex algorithms and practical usability.

However, this study's findings must be interpreted within the constraints of its case study design. The single-creator data source, four-month observation window, and purely textual analysis approach limit the generalizability of results across different creator profiles, longer time periods, and the inherently multimodal nature of video content. The moderate F1-Score indicates room for substantial improvement, particularly through integration of multimodal features including visual and audio analysis, as well as collaborative filtering approaches that leverage cross-creator engagement patterns.

Despite these limitations, the validated framework provides a replicable foundation for content creators seeking to transition from intuition-based to evidence-based affiliate marketing strategies in competitive social commerce environments. The proof-of-concept demonstrates that systematic data mining approaches can deliver measurable value even with limited scope and resources.

5. FUTURE WORK

Based on the specific limitations identified in this study, we propose the following targeted research directions to advance both the theoretical understanding and practical implementation of data-driven affiliate marketing optimization.

A multi-creator validation study should be conducted by replicating the methodology across 10 to 15 diverse TikTok creator accounts spanning different niches including fashion, technology, beauty, and food. This would assess model transferability and identify category-specific optimization requirements, addressing the current limitation of single-source bias.

Multimodal feature integration represents a critical enhancement opportunity. The current text-only approach should be extended by incorporating computer vision analysis of video thumbnails and visual content, audio feature extraction from video soundtracks, and temporal engagement patterns such as view duration and drop-off rates. This addresses the fundamental limitation that TikTok content is inherently visual and auditory, not purely textual, and that ignoring these modalities likely constrains model performance.

Development of a hybrid recommendation architecture should combine the current CBF approach for cold-start scenarios with Collaborative Filtering using cross-creator engagement data, and context-aware filtering incorporating time-of-day, trending hashtags, and seasonal factors. This hybrid system should be evaluated using advanced metrics like NDCG@K and Mean Average Precision to capture ranking quality beyond simple classification accuracy.

A longitudinal performance analysis over a 12-month period would assess model performance degradation over time, develop automated retraining pipelines, and identify optimal model refresh intervals. This would address the temporal constraint limitation and provide insights into the dynamic nature of platform algorithms and user preferences.

Causal impact evaluation through controlled experimentation is essential for validating real-world effectiveness. A study should be designed where a treatment group of creators uses the recommendation system while a control group follows standard practices, with measurement of actual business outcomes including conversion rates, commission revenue, and return on investment. This would move beyond correlation to establish causal effectiveness, addressing the current lack of A/B testing validation.

An API-integrated production system should transform the proof-of-concept into a production-grade web application featuring real-time TikTok API integration for automated data ingestion, responsive mobile interface for on-the-go content planning, push notification system for emerging product-content opportunities, and an embedded A/B testing framework for continuous improvement. This addresses the current manual refresh and limited deployment limitations.

Finally, explainable AI enhancement through implementation of SHAP (SHapley Additive exPlanations) or LIME would provide interpretable recommendations, showing creators why specific videos match certain products. This would enhance trust and actionability by making the recommendation logic transparent and educational for users.

These directions address the core limitations of single-source data, text-only analysis, and lack of real-world validation, while building upon the validated foundational methodology established in this study. Together, they chart a path toward a more robust, generalizable, and practically impactful affiliate marketing optimization system for social commerce platforms.

ACKNOWLEDGMENT

The authors would like to thank Universitas Negeri Padang for financial support toward the APC of this article, funded by the EQUITY Kemdiktisaintek Program supported by LPDP, under contract number 4310/B3/DT.03.08/2025 and 2692/UN35/KS/2025.

REFERENCES

[1] Sutriawan, Muljono, Khairunnisa, Alamin, Z., Lorosae, T.A., Ramadhan, S. (2024). Improving performance sentiment movie review classification using hybrid feature TFIDF, N-Gram, information gain and support vector machine. Mathematical Modelling of Engineering Problems, 11(2): 375-384.

https://doi.org/10.18280/mmep.110209

[2] Pebrianti, D., Ahmad, D., Bayuaji, L., Wijayanti, L., Mulyadi, M. (2024). Using Content-Based Filtering and apriori for recommendation systems in a smart shopping system. Indonesian Journal of Computing, Engineering, and Design (IJoCED), 6(1): 58-70. https://doi.org/10.35806/ijoced.v6i1.393

[3] Cao, Z.Q. (2020). Classification of digital teaching resources based on data mining. Ingénierie des Systèmes d'Information, 25(4): 521-526. https://doi.org/10.18280/isi.250416

[4] Christina, Sanjaya, F., Cahyaningtyas, V.T., Edbert, I.S., Suhartono, D. (2023). Distance metric analysis in recommendation system using Content-Based Filtering method. In 2023 6th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 47-52. https://doi.org/10.1109/ICOIACT59844.2023.10455814

[5] Rolanda, V., Gunawan, T.S., Wanayumini, W. (2023). Content-Based Filtering recommendation system using categories search engine. International Journal of Research in Vocational Studies (IJRVOCAS), 2(4): 120-125. https://doi.org/10.53893/ijrvocas.v2i4.177

[6] Shree, P., Suvvari, S. (2024). Parallel memory-based collaborative filtering for distributed big data environments. International Journal of Computational Methods and Experimental Measurements, 12(3): 217-225. https://doi.org/10.18280/ijcmem.120303

[7] Chen, C., Hao, L.F., Bai, B., Zhang, G.J. (2025). Knowledge discovery from database: MRI radiomic features to assess recurrence risk in high-grade meningiomas. BMC Medical Imaging, 25: 14. https://doi.org/10.1186/s12880-024-01483-2

[8] Koç, B. (2023). The role of user interactions in social media on recommendation algorithms: Evaluation of Tiktok's personalization practices from user's perspective. ResearchGate. https://doi.org/10.13140/RG.2.2.34692.71040

[9] Doğan, M., Chelery Komath, M.A., Sayilir, Ö. (2025). Credit rating prediction with ESG data using data mining methods. Future Business Journal, 11: 79. https://doi.org/10.1186/s43093-025-00490-1

[10] Januzaj, Y., Luma, A. (2022). Cosine similarity – A computing approach to match similarity between higher education programs and job market demands based on maximum number of common words. International Journal of Emerging Technologies in Learning (iJET), 17(12): 258-268. https://doi.org/10.3991/ijet.v17i12.30375

[11] Sintia, S., Defit, S., Nurcahyo, G.W. (2021). Product codefication accuracy with Cosine Similarity and weighted term frequency and inverse document frequency (TF-IDF). Journal of Applied Engineering and Technological Science (JAETS), 2(2): 62-69. https://doi.org/10.37385/jaets.v2i2.210

[12] Al Sabri, M.A., Zubair, S., Alnuhait, H.A. (2025). Improved prediction on recommendation system by creating a new model that employs Mahout collaborative filtering with Content-Based Filtering based on genetic algorithm methods. Discover Artificial Intelligence, 6: 20. https://doi.org/10.1007/s44163-025-00678-y

[13] Soedarsono, D.K., Mohamad, B., Adamu, A.A., Aline Pradita, K. (2020). Managing digital marketing communication of coffee shop using instagram. International Journal of Interactive Mobile Technologies (iJIM), 14(5): 108-118. https://doi.org/10.3991/ijim.v14i05.13351

[14] Yuensuk, T., Limpinan, P., Nuankaew, W., Nuankaew, P. (2022). Information systems for cultural tourism management using text analytics and data mining techniques. International Journal of Interactive Mobile Technologies (iJIM), 16(9): 146-163. https://doi.org/10.3991/ijim.v16i09.30439

[15] Ouadoud, M., Chkouri, M.Y., Nejjari, A. (2018). LeaderTICE: A platforms recommendation system based on a comparative and evaluative study of free e-learning platforms. International Journal of Online Engineering (iJOE), 14(1): 132-161. https://doi.org/10.3991/ijoe.v14i01.7865

[16] Mukhopadhyay, S., Kumar, A., Parashar, D., Singh, M. (2024). Enhanced music recommendation systems: A comparative study of Content-Based Filtering and K-Means clustering approaches. Revue d'Intelligence Artificielle, 38(1): 365-376. https://doi.org/10.18280/ria.340138

[17] Hakim, A.F., Irawan, Y., Setiawan, R.R. (2025). Implementation of the Crisp-Dm methodology and naive bayes algorithm on a raw material requirement prediction system to reduce food waste (Case study: Adamsafee Bakery, Resto, & Cafe). Jurnal Teknologi Informasi dan Pendidikan, 18(2): 968-983. https://doi.org/10.24036/jtip.v18i2.990

[18] Wahyudi, T., Silfia, T. (2022). Implementation of data mining using K-Means clustering method to determine sales strategy in S&R Baby Store. Journal of Applied Engineering and Technological Science (JAETS), 4(1): 93-103. https://doi.org/10.37385/jaets.v4i1.913

[19] Ramesh, M., Elangovan, V.R. (2025). Optimizing e-learning journey: Fusing collaborative and Content-Based Filtering in hybrid recommender systems. OPSEARCH. https://doi.org/10.1007/s12597-025-01057-y

[20] Novaliendry, D., Permana, A., Dwiyani, N., Ardi, N., Yang, C.H., Saragih, F.M. (2024). Development of a semantic text classification mobile application using TensorFlow Lite and Firebase ML Kit. Journal Européen des Systèmes Automatisés, 57(6): 1603-1611. https://doi.org/10.18280/jesa.570607

[21] Novaliendry, D., Wibowo, T., Ardi, N., Evi, T., Admojo, D. (2023). Optimizing patient medical records grouping through data mining and K-Means clustering algorithm: A case study at RSUD mohammad natsir solok. International Journal of Online Engineering (iJOE), 19(12): 144-155. https://doi.org/10.3991/ijoe.v19i12.42147

[22] Novaliendry, D., Ramah, S., Badaruddin, M., Utiarahman, S.A., Husain, H., Zain, R.H., Darwin, W., Erdisna. (2025). Implementation of the Analytical Hierarchy Process (AHP) method for choosing majors in the web-based new student admission information system (PPDB). Salud, Ciencia y Tecnología, 5: 1887. https://doi.org/10.56294/saludcyt20251887