# Influencer Recommendation from Campaign Briefings: A Briefing-Centric Approach Using Sparse Representations and Density-Based Clustering

**Luiz Carlos Grandisoli, Gustavo Scalabrini Sampaio, Maria Amelia Eliseo**

[1]Graduate Program in Applied Computing, Universidade Presbiteriana Mackenzie
São Paulo, SP, Brazil

luiz.grandisoli@gmail.com

***Abstract.*** *This work evaluates a cold-start influencer recommendation pipeline using 110 synthetic campaign briefings and 1,100 synthetic influencer biographies in Brazilian Portuguese. The method compares TF-IDF and multilingual MiniLM embeddings under direct ranking and clustering-based retrieval. Results show that TF-IDF combined with DBSCAN achieves the best performance, reaching P@5 = 0.844 and NDCG@5 = 0.854 overall, and P@5 = 0.928 in the in-scope evaluation. The findings indicate that briefing-centric recommendation is feasible in cold-start settings, and that within this synthetic benchmark, where niche labels are defined by characteristic vocabulary, sparse lexical representations outperform dense embeddings for niche-level discrimination; the advantage persists under nested cross-validation ($p < 0.001$) but is specific to lexically marked niches rather than general.*

**Keywords:** Recommendation systems, natural language processing, influencer-based recommendation, text clustering, semantic similarity, content-based recommendation.

## 1. Introduction

Over the past decade, influencer marketing has expanded rapidly, leaving brands with a large and heterogeneous pool of creators to navigate [Sweet et al. 2019, Jha and Ray 2023]. Selection still relies on reach, engagement proxies, and manual inspection, procedures that are costly, difficult to scale, and insufficient when campaign effectiveness depends on semantic alignment rather than popularity [Gan et al. 2019, Wang et al. 2023]. Most existing approaches are creator-centric, modeling account content, audience attributes, or multimodal signals, but rarely treating the campaign briefing as the primary semantic input [Sweet et al. 2019, Jha and Ray 2023, Gan et al. 2019, Wang et al. 2023]. This leaves open a practical question: can the campaign briefing be used as the main retrieval input in cold-start influencer recommendation?

This paper formalizes influencer recommendation as a briefing-centric retrieval problem, where the campaign briefing acts as a semantic query over influencer profiles. The proposed pipeline combines text vectorization, unsupervised clustering, and cosine similarity ranking without prior interaction history, making it suitable for cold-start scenarios. We compare TF-IDF (Term Frequency-Inverse Document Frequency, a sparse lexical weighting method) and MiniLM-based contextual embeddings under two biography-length conditions and three clustering algorithms (K-Means, Agglomerative, DBSCAN),

reporting a reproducible evaluation protocol with tuned hyperparameters.[1] The study should be interpreted as a controlled benchmark for isolating representation and clustering effects, not as evidence of deployment-ready performance.

This work makes three contributions: (i) it formalizes influencer recommendation as a briefing-centric cold-start retrieval task; (ii) it introduces a controlled Portuguese synthetic benchmark to isolate the effects of textual richness, encoder choice, and clustering; and (iii) it shows that, in this setting, sparse lexical representations combined with density-based clustering outperform multilingual contextual embeddings for niche-level discrimination.

## 2. Related Work

Influencer recommendation has evolved from popularity-based selection to richer representations that combine content signals, audience attributes, and semantic analysis [Sweet et al. 2019, Jha and Ray 2023, Gan et al. 2019, Wang et al. 2023]. In parallel, the broader recommendation literature shows that content-based methods are effective when interaction history is unavailable [Aggarwal 2016], that hybrid strategies become advantageous when structured and unstructured data coexist [Zhang et al. 2019], and that cosine similarity over textual features remains a strong baseline [Stefanovic et al. 2019]. Regarding text encoding, Sentence-BERT-style models produce semantically meaningful sentence embeddings that can be compared using cosine similarity and applied to semantic search and clustering [Reimers and Gurevych 2019]. However, their advantage over sparse lexical representations is task- and domain-dependent, motivating the comparison with TF-IDF in this study.

Existing methods differ from this approach in at least one dimension: they are creator-centric, rely on interaction history unavailable in cold-start scenarios, or do not analyze the interaction between textual richness and encoder choice (Table 1). This article addresses this gap by treating the briefing as the semantic query and systematically comparing encoders, clustering strategies, and representation types.

**Table 1. Comparison with related studies on influencer recommendations.**

| Method | Main signal | Briefing as primary query? | Explicit cold-start setting? | Encoder comparison? |
|---|---|---|---|---|
| Sweet et al. [Sweet et al. 2019] | Instagram content | No | No | No |
| Jha and Ray [Jha and Ray 2023] | YouTube titles (NLP) | Partial | No | No |
| Gan et al. [Gan et al. 2019] | Visual + text posts | No | No | No |
| Wang et al. [Wang et al. 2023] | Multi-perspective account representation | No | No | No |
| This work | Briefing text | Yes | Yes | Yes |

[1]The synthetic dataset [Grandisoli et al. 2026a] and the reproduction code [Grandisoli et al. 2026b] are publicly available on Zenodo.

## 3. Proposed Method

### 3.1. Problem Definition

Let $B$ denote a campaign briefing and $I = \{i_1, i_2, \ldots, i_n\}$ a set of influencer profiles. Each briefing and influencer profile includes structured attributes and textual descriptions. The task is to define and evaluate a ranking function $f(B, I)$ that orders influencers by estimated compatibility with the briefing. Compatibility is modeled through semantic similarity between the briefing and the influencer biography, optionally combined with structured attributes, following a content-based recommendation approach in which auxiliary features support the recommendation process [Aggarwal 2016]. Since no brand-creator interaction history is available, the problem is treated as a cold-start retrieval scenario. The workflow is shown in Figure 1.

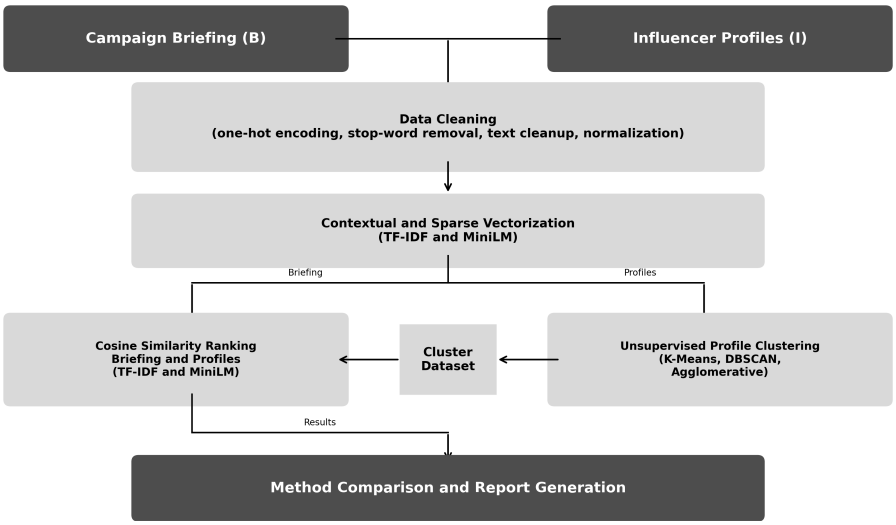



**Figure 1. Briefing-centric recommendation pipeline.**

### 3.2. Synthetic Data Set

The experiments use a synthetic dataset with 110 campaign briefings and 1,100 influencer profiles across eight domains: self-care, wellness, fitness, longevity, mental health, sports nutrition, healthy eating, and weight management. Profiles include followers, engagement rate, content type, tone (communication voice, e.g., educational or motivational), and biography, while briefings include objective, product category, target audience, tone, and description. The briefings focus on five in-scope niches (longevity, self-care, mental health, fitness, and wellness), representing campaigns that cover only part of the available creator pool. The dataset was generated with a large language model (OpenAI ChatGPT) under a documented per-niche prompt protocol: each profile fixes the niche and samples a tone of voice, content type, target audience, and structured attributes (followers, engagement, posting frequency, age, and geography), yielding varied combinations within each niche, while enriched biographies inject niche-specific vocabulary; briefings follow an eight-field marketing template (context, objective, audience, creative territory, influencer profile, deliverables, key message, and proof) instantiated per niche. The protocol controlled niche, tone, and semantic coherence, ensured reproducibility, avoided real personal

data, reflected a Brazilian Portuguese non-English retrieval context, and passed automated consistency checks. Labels were used only for benchmark construction and evaluation, not as input features during ranking.

The dataset is nearly balanced across eight niches: longevity (138), fitness (138), wellness (138), mental health (138), self-care (137), healthy eating (137), weight management (137), and sports nutrition (137). These niches reflect major wellness segments, including health, fitness, mindfulness, appearance, nutrition, longevity, and weight management [Callaghan et al. 2024, Global Wellness Institute 2024].

Two biography-length conditions are evaluated: short (12–21 words, avg. 16), approximating Instagram's 150-character constraint; and enriched (110–127 words, avg. 117), including niche-specific vocabulary such as "biohacking," "psychoeducation," and "meal prep." Enriched biographies provide more lexical evidence for sparse retrieval, consistent with vocabulary-growth assumptions in information retrieval [Manning et al. 2008]. At the same time, Sentence-BERT-style models produce dense sentence embeddings suitable for semantic similarity and clustering [Reimers and Gurevych 2019]. This condition therefore evaluates whether biography richness improves both sparse and dense retrieval.

The relevance ground truth is defined over five in-scope niches: longevity, self-care, mental health, fitness, and wellness, with 100 briefings distributed across these niches. Ten additional briefings target fashion (5) and technology (5), two niches absent from the influencer pool. These out-of-scope briefings provide a conservative stress-test condition in which no relevant profile exists under the binary relevance definition. The remaining three niches, sports nutrition, healthy eating, and weight management, are present in the influencer pool but have no corresponding briefings, acting as distractors that the system must bypass. Results are reported in two ways: overall performance with all 110 briefings and in-scope performance with 100 briefings, excluding the 10 out-of-scope briefings.

**Lexical overlap by construction.** Because briefings and biographies share the same niche-controlled prompt protocol, the niche label is partly recoverable from surface vocabulary, favouring lexical methods by construction. Over the five in-scope niches, the content-vocabulary overlap of a briefing with same-niche biographies exceeds that with other niches (0.228 vs. 0.183; diagonal lift $1.24\times$, robust to tokenisation), and in all five the correct niche is the one of maximal overlap. Part of the niche signal is thus available to a bag-of-words model by design, which tempers the TF-IDF advantage reported below.

### 3.3. Data Preparation Workflow

The pipeline performs five operations: dataset validation (1,100 profiles, 8 niches), one-hot encoding of categorical variables, MinMaxScaler normalization of numeric variables, text vectorization (TF-IDF or MiniLM-based sentence embeddings), and concatenation into a unified feature space. The TF-IDF weight is defined as:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \log\left(\frac{N}{\mathrm{df}(t)}\right), \tag{1}$$

where $\mathrm{tf}(t, d)$ is the frequency of term $t$ in document $d$, $N$ is the corpus size,

**Table 2. Summary of the synthetic dataset.**

| Element | Value |
|---|---|
| Campaign briefings | 110 (100 in-scope + 10 out-of-scope) |
| Briefing niches | 7 (5 in-scope and 2 out-of-scope) |
| Briefing subject areas | Mental health, wellness, longevity, fitness, self-care, technology (out-of-scope), and fashion (out-of-scope) |
| Influencer profiles | 1,100 |
| Influencer relevant niches | 5 (longevity, self-care, mental health, fitness, wellness) |
| Influencer distracting niches | 3 (sports nutrition, weight management, healthy eating) |
| Short bio | 16 words on average |
| Enriched bio | 117 words on average |
| Relevance criterion | Binary thematic compatibility (briefing niche = influencer niche) |

and $\mathrm{df}(t)$ is the number of documents containing $t$ [Stefanovic et al. 2019]. TF-IDF is adopted as a sparse baseline given its interpretability, while MiniLM-based contextual embeddings test whether dense representations offer measurable gains under richer profile descriptions. The compatibility between campaign briefings and influencer biographies is calculated using cosine similarity. Cosine similarity measures angular proximity independently of vector magnitude, making it suitable for texts of variable length [Stefanovic et al. 2019].

### 3.4. Profile Grouping

Before ranking, influencer profiles are organized using unsupervised clustering to structure the search space into semantically coherent groups. Three clustering families were selected to structure the retrieval space from complementary perspectives [Fuchs and Hopken 2022]: K-Means, as a partition-based method that forms compact centroid-based groups; Agglomerative clustering, to test whether hierarchical bottom-up organization better preserves semantic proximity across adjacent niches; and DBSCAN, as a density-based alternative that identifies clusters of arbitrary shape and isolates noisy profiles without requiring a predefined number of groups [Cretulescu et al. 2011].

K-Means is an unsupervised algorithm that partitions data into $k$ clusters by assigning observations to the nearest centroid and minimizing the within-cluster sum of squares, usually based on Euclidean distance [Fuchs and Hopken 2022]. Agglomerative clustering uses a bottom-up merging process with Ward and complete linkage criteria. Ward forms compact clusters by minimizing within-cluster variance, while complete linkage enforces stricter separation through maximum pairwise distance [Cretulescu et al. 2011]. DBSCAN groups observations located in dense regions and identifies sparse observations as noise or outliers. Unlike K-Means, it does not require defining the number of clusters in advance and can detect clusters with irregular shapes [Cretulescu et al. 2011].

Final ranking scores are computed as:

$$s(B, i) = \cos(\mathbf{v}_B, \mathbf{v}_i) + \beta \cdot \mathbb{I}\{c(i) = c_B\}. \quad (2)$$

The cluster-boost parameter $\beta = 0.10$ was fixed across all configurations; for the best configuration, P@5 varies by at most 0.008 over $\beta \in [0, 0.30]$ and plateaus for $\beta \geq 0.10$, so results are insensitive to this choice. The campaign briefing is used as a semantic query in the influencer space. Each influencer receives a cosine similarity score relative to the briefing, plus cluster boosting where applicable. The system returns a ranked list of the top $K$ creators, ordered by final score. Hyperparameters are selected through grid search over the clustering family, number of clusters, linkage, vectorizer size, n-gram range, feature weights, $\varepsilon$, and minimum samples.

### 3.5. MiniLM-based Multilingual Sentence Encoder

A complementary condition replaces TF-IDF with the off-the-shelf `paraphrase-multilingual-MiniLM-L12-v2` model, used zero-shot without domain or Portuguese-specific fine-tuning and chosen as a lightweight multilingual baseline compatible with Portuguese; it therefore measures the contribution of a generic dense encoder rather than the best achievable performance of contextual embeddings. Embeddings are unit-normalized before cosine similarity calculation. In this condition, Euclidean distance is used as the DBSCAN metric, given that dense embeddings are not sparse. All other pipeline components remain identical, allowing direct comparison with TF-IDF.

## 4. Experimental Setup

### 4.1. Assessment Protocol

Recommendation quality is evaluated in a cold-start scenario with no interaction history: an influencer is considered relevant when their niche matches the target niche specified in the campaign briefing. Since no relevant profiles exist for out-of-scope briefings under the binary relevance definition, P@5 necessarily equals zero for these cases. They are therefore included as a conservative stress test of the retrieval setting, not as an evaluation of abstention or out-of-distribution detection. The evaluation uses $K = 5$, as the system is intended for shortlist generation. P@5 measures the proportion of relevant influencers in the top five positions, while NDCG@5 complements this by rewarding rankings in which relevant results appear earlier. Results are reported in two registers: (i) overall performance ($n = 110$), including out-of-scope briefings as a conservative estimate; and (ii) in-scope performance ($n = 100$), isolating behavior when the target niche is represented in the pool.

### 4.2. Experimental Conditions

Ten experimental conditions were evaluated through grid search, prioritizing P@5 and NDCG@5. The random baseline used uniform ranking over 100 runs (P@5 = 0.114; NDCG@5 = 0.114). The structured-only condition used follower count and engagement rate. TF-IDF conditions included direct cosine ranking and clustering with K-Means ($k = 8$), Agglomerative ($k = 8$, Ward), and DBSCAN ($\varepsilon = 0.30$, `min_samples`=3, cosine metric). MiniLM conditions included direct ranking, K-Means (hybrid, $w_{\text{text}} =$

$0.8$, $w_{\text{num}} = 0.2$, $k = 6$), Agglomerative ($k = 6$, complete linkage), and DBSCAN ($\varepsilon = 0.30$, `min_samples`=8, Euclidean). Short and enriched biographies were evaluated with text-only and hybrid representations, and ablations compared structured-only, direct, and clustered configurations.

Hyperparameter selection was conducted in a preliminary experimental stage before the final comparison among recommendation configurations. This stage evaluated alternative values for text vectorization, representation weighting, clustering structure, distance metrics, and DBSCAN density parameters. For each algorithmic family, the configuration selected for the final evaluation was the one that maximized top-$k$ recommendation quality, prioritizing P@5 and NDCG@5, while also recording internal clustering indicators such as silhouette score, Davies-Bouldin index, Calinski-Harabasz score, number of clusters, and noise ratio when applicable. As a result, configurations within the same encoder family may use different vectorization hyperparameters, since each was tuned independently to maximize retrieval quality rather than to equalize input representations.

Because selection and final comparison initially used the same briefings, we validated the main comparison with niche-stratified nested five-fold cross-validation over the 100 in-scope briefings (hyperparameters tuned on training folds, metrics on held-out folds, biography index fixed, clustering restricted to noise $\leq 20\%$). TF-IDF reaches held-out P@5 $= 0.910 \pm 0.035$ against $0.750 \pm 0.075$ for MiniLM ($p < 0.001$), recovering the reported TF-IDF + DBSCAN configuration in most folds; the reported figures are thus not a tuning artefact.

A set of research questions was formulated to guide the empirical design, experimental comparisons, and interpretation of results, focusing on whether semantic ranking improves influencer selection based on campaign briefings.

### 4.3. Research Questions

RQ1: Does briefing-centered semantic ranking outperform random selection?

RQ2: To what extent does clustering improve ranking quality under different levels of textual richness?

RQ3: Does hybrid representation surpass text-only and structured-only representation?

RQ4: Under what conditions do the MiniLM-based multilingual sentence embeddings outperform TF-IDF, and to what extent is this effect mediated by biography richness?

## 5. Results and Discussion

### 5.1. Clustering Quality

The grid-search results show that cluster quality and recommendation quality do not always move in lockstep. In enriched biographies, TF-IDF K-Means and TF-IDF Agglomerative clustering obtain strong internal clustering scores, but their retrieval performance remains slightly below the best TF-IDF + DBSCAN configuration. The strongest retrieval configuration is TF-IDF + DBSCAN ($\varepsilon = 0.30$; `min_samples`=3; cosine metric; 17

clusters; silhouette = 0.317; Davies-Bouldin = 1.548), achieving P@5 = 0.844 overall and P@5 = 0.928 in-scope. This indicates that TF-IDF representations create sharper cluster boundaries in this benchmark, resulting in better search-space structuring for niche-level retrieval.

### 5.2. Recommendation Results

Table 3 presents the final P@5 and NDCG@5 values for the selected hyperparameter configurations. The table reports both the overall evaluation with 110 briefings and the in-scope evaluation with 100 briefings, excluding the ten intentionally out-of-scope briefings related to technology and fashion.

**Table 3. Final recommendation results after hyperparameter selection. Overall metrics include 110 briefings; in-scope metrics include 100 briefings.**

| Method | Key configuration | Overall P@5 / NDCG@5 | In-scope P@5 / NDCG@5 |
|---|---|---|---|
| Random baseline | 100 runs | 0.114 / 0.114 | – |
| Structured only | followers + engagement | 0.096 / 0.100 | 0.106 / 0.110 |
| TF-IDF direct | 3k features; ngram=(1,2) | 0.776 / 0.774 | 0.854 / 0.851 |
| TF-IDF + K-Means | $k = 8$ | 0.836 / 0.848 | 0.920 / 0.933 |
| TF-IDF + Agglomerative | $k = 8$; Ward | 0.836 / 0.848 | 0.920 / 0.933 |
| TF-IDF + DBSCAN | eps=.30; min samples=3; cosine | 0.844 / 0.854 | 0.928 / 0.939 |
| MiniLM direct | text only; MiniLM-L12-v2 | 0.665 / 0.666 | 0.732 / 0.732 |
| MiniLM + K-Means | hybrid; $k = 6$ | 0.640 / 0.643 | 0.704 / 0.708 |
| MiniLM + Agglomerative | $k = 6$; hybrid or text only | 0.635 / 0.642 | 0.678 / 0.686 |
| MiniLM + DBSCAN | hybrid; eps=.30; min samples=8 | 0.616 / 0.623 | 0.678 / 0.686 |

Short biographies produced lower retrieval across all families (Table 4; best: TF-IDF + Agglomerative, P@5 = 0.709 overall and 0.780 in-scope), confirming that biography richness is decisive; TF-IDF outperforms MiniLM at every configuration in this condition as well. In the in-scope evaluation of the best configuration (TF-IDF + DBSCAN), 77 of 100 briefings achieved perfect P@5 = 1.0 and none scored 0.0. Performance varies by niche, from near-perfect for mental health, fitness, wellness, and longevity to self-care at 4/20 perfect cases (mean P@5 = 0.73), reinforcing the failure mode of Section 5.5: niches with distinctive vocabulary are well captured, while diffuse niches like self-care suffer lexical overlap with adjacent categories.

All TF-IDF clustering variants significantly outperform TF-IDF direct ranking, whereas clustering does not improve MiniLM, which trails TF-IDF throughout (Table 3).

Statistical significance was assessed using the Wilcoxon signed-rank paired test on P@5 scores per briefing, $n = 100$ in-scope briefings, two-tailed, $\alpha = 0.05$. The main

**Table 4. Short-biography results, tuned per family. Overall metrics include 110 briefings; in-scope metrics include 100. Random and structured-only baselines are unchanged from Table 3.**

| Method | Key configuration | Overall P@5 / NDCG@5 | In-scope P@5 / NDCG@5 |
|---|---|---|---|
| TF-IDF direct | 3k features; ngram=(1,2) | 0.644 / 0.647 | 0.708 / 0.711 |
| TF-IDF + K-Means | $k = 6$; 3k; ngram=(1,2) | 0.682 / 0.686 | 0.750 / 0.755 |
| TF-IDF + Agglomerative | $k = 3$; Ward | 0.709 / 0.711 | 0.780 / 0.783 |
| TF-IDF + DBSCAN | eps=.50; min samples=4; cosine | 0.680 / 0.683 | 0.748 / 0.752 |
| MiniLM direct | text only | 0.496 / 0.499 | 0.546 / 0.549 |
| MiniLM + K-Means | text only; $k = 4$ | 0.575 / 0.572 | 0.632 / 0.629 |
| MiniLM + Agglomerative | hybrid; $k = 10$; complete | 0.589 / 0.589 | 0.648 / 0.648 |
| MiniLM + DBSCAN | hybrid; eps=.50; min samples=2 | 0.555 / 0.556 | 0.610 / 0.611 |

comparisons are:

- TF-IDF + DBSCAN vs. MiniLM direct: $\Delta$P@5 = +0.196, $p = 0.000005$;
- TF-IDF + DBSCAN vs. TF-IDF direct: $\Delta$P@5 = +0.074, $p = 0.003$;
- TF-IDF + K-Means vs. MiniLM + K-Means: $\Delta$P@5 = +0.196, $p = 0.000011$;
- MiniLM direct vs. MiniLM + K-Means: $\Delta$P@5 = +0.026, $p = 0.132$;
- TF-IDF + DBSCAN vs. TF-IDF + K-Means: $\Delta$P@5 = +0.008, $p = 0.545$.

To account for multiple pairwise comparisons, Holm-Bonferroni correction was applied. The main significant differences involving TF-IDF + DBSCAN remained significant after correction. Relative to the random baseline (P@5 = 0.114), the best configuration achieves a $7.4\times$ improvement.

### 5.3. Ablation Study

The ablation isolates four factors. Text richness: enriched biographies far exceed short ones (Tables 3–4). Encoder: direct TF-IDF (P@5 = 0.776) beats direct MiniLM (0.665) by 11.1 points. Clustering: DBSCAN raises TF-IDF to 0.844 ($p = 0.003$) but does not help MiniLM (0.616–0.640 vs. 0.665). Structured features: followers and engagement alone (0.096) fall below the random baseline. The strongest configuration therefore combines sparse lexical representations with density-based clustering; interpretation follows in Section 5.6.

### 5.4. Answering the Research Questions

RQ1 is answered affirmatively. All semantic methods outperform the random baseline (P@5 = 0.114), with the best configuration, TF-IDF + DBSCAN, achieving P@5 = 0.844 overall and P@5 = 0.928 in the in-scope evaluation, representing a $7.4\times$ improvement.

RQ2 is answered affirmatively for TF-IDF and negatively for MiniLM. TF-IDF improves from P@5 = 0.776 to 0.844 with DBSCAN clustering ($p = 0.003$). However, clustering does not improve retrieval based on MiniLM: MiniLM direct (P@5 = 0.665) outperforms all MiniLM clustering variants (P@5 = 0.616–0.640), although without statistical significance. This asymmetric effect suggests that clustering benefits depend on the geometry of the representation space.

RQ3 is answered negatively. Structured-only retrieval (P@5 = 0.096) performs below the random baseline. The best configurations are text-only (TF-IDF DBSCAN) or text-dominant. A constrained hybrid MiniLM K-Means configuration (P@5 = 0.640) does not outperform text-only MiniLM direct (P@5 = 0.665), so structured attributes do not constitute a beneficial component.

RQ4 is answered negatively. In this benchmark, TF-IDF outperforms MiniLM across all evaluated configurations. Direct TF-IDF (P@5 = 0.776) exceeds direct MiniLM (P@5 = 0.665), and TF-IDF + DBSCAN (P@5 = 0.844) exceeds the best MiniLM configuration (P@5 = 0.665), both with $p < 0.001$. This result indicates that contextual embeddings do not provide an advantage when niche discrimination relies on specific lexical markers that are well captured by sparse representations.

### 5.5. Qualitative Examples

Two examples illustrate the behavior of the best configuration, TF-IDF + DBSCAN on enriched biographies. Success case: a fitness-for-beginners briefing (Briefing 21) retrieves five fitness influencers in the top-5 (P@5 = 1.0); all contain niche-specific markers such as "preparador físico," "corrida," "treino," and "força" aligned with the briefing, showing how the method captures discriminative vocabulary within the target niche.

Failure case: a self-care briefing (Briefing 85) retrieves three correct profiles but also a wellness mentor and a mental-health speaker (P@5 = 0.60); both contain expressions such as "qualidade de vida," "hábitos," and "burnout" that overlap lexically with self-care, and their similarity scores (0.392–0.393) are close to the correct ones (0.396–0.425). This is the main failure mode, sparse representations overvaluing shared vocabulary across adjacent wellness niches, which future work could address through graded relevance or niche-aware re-ranking.

### 5.6. Discussion

The results support three methodological findings. First, the main performance gain comes from combining sparse lexical representations with density-based clustering, rather than from replacing TF-IDF with contextual embeddings. TF-IDF + DBSCAN achieves P@5 = 0.844, significantly outperforming all MiniLM configurations ($p < 0.001$). This suggests that, for niche discrimination in Portuguese wellness domains, specific lexical markers ("biohacking," "psychoeducation," "meal prep") carry stronger discriminative signal than contextual embeddings, which tend to map semantically adjacent niches into overlapping embedding regions, suggesting that lexical signal strength is platform and domain dependent. This contrasts with [Gan et al. 2019], whose text-only features were nearly uninformative on Instagram (AUC = 0.531), and is consistent with the broader IR finding that off-the-shelf dense encoders often trail sparse lexical baselines in zero-shot, out-of-domain retrieval [Thakur et al. 2021].

Second, clustering benefits depend on representation geometry: DBSCAN significantly improves TF-IDF ($p = 0.003$) but not MiniLM, whose more homogeneous space does not benefit from spatial structuring. This parallels [Wang et al. 2023], where complementary sources are needed when single-perspective representations leave niche boundaries implicit.

Third, structured features remain weak as primary signals: followers and engagement alone fall below the random baseline (P@5 = 0.096) and hybrids do not beat text-only, so popularity should be treated as a post-retrieval business filter, consistent with [Gan et al. 2019, Jha and Ray 2023].

## 6. Conclusion

This study investigated a briefing-centric recommendation pipeline for influencer selection under cold-start conditions, comparing sparse and contextual text representations with multiple clustering strategies using enriched biographies. The results show that TF-IDF combined with DBSCAN clustering achieves the strongest performance, reaching P@5 = 0.844 and NDCG@5 = 0.854 overall, and P@5 = 0.928 in the in-scope evaluation with 100 briefings. TF-IDF significantly outperforms MiniLM-based contextual embeddings across all configurations ($p < 0.001$), and clustering significantly improves TF-IDF retrieval ($p = 0.003$) but does not improve ranking based on MiniLM. Structured attributes such as followers and engagement are weak as standalone signals and should be treated as post-retrieval business constraints rather than primary semantic features.

From a practical perspective, the study shows that briefing-centric influencer recommendation is feasible even without historical interaction logs, if creator profiles contain sufficiently informative textual descriptions and the representation space is organized through tuned density-based clustering. From a methodological perspective, sparse lexical representations outperform contextual embeddings for niche discrimination in this benchmark, and the advantage persists under nested cross-validation ($p < 0.001$). Because the synthetic niche labels are themselves defined by characteristic vocabulary (Section 3.2), part of this advantage reflects recovery of generation-time keywords rather than a general superiority of sparse over dense encoders; it is best read as evidence that sparse representations excel when niche boundaries are lexically marked.

This work is limited by its reliance on synthetic data, the absence of behavioral or multimodal signals, and a purely binary relevance definition; in particular, generating both briefings and biographies with a single LLM embeds a measurable lexical overlap (Section 3.2) that favours sparse retrieval. Future research should validate with real campaign data, diversify generation across multiple LLMs, investigate graded relevance, and test whether the TF-IDF advantage persists with less specialized vocabulary and against fine-tuned or Portuguese-specific dense encoders.

## Acknowledgments

ChatGPT (OpenAI) was used to generate the synthetic dataset and support language revision and editorial refinement under a documented instruction protocol. No generative AI model is listed as author. The authors assume full responsibility for the manuscript, including originality, correctness, and verification of results, in accordance with the conference policy on generative AI use.

## References


Sweet, T., Rothwell, A. and Luo, X. (2019). Machine learning techniques for brand-influencer matchmaking on the Instagram social network. arXiv preprint arXiv:1906.05026.

Jha, A. K. and Ray, S. (2023). Influencer recommendation system: choosing the right influencer using a network analysis approach. Marketing Intelligence & Planning, 41(3):369–384.

Gan, T., Wang, S., Liu, M., Song, X., Yao, Y. and Nie, L. (2019). Seeking micro-influencers for brand promotion. In Proceedings of the 27th ACM International Conference on Multimedia, pages 1595–1603. ACM.

Wang, S., Gan, T., Liu, Y., Wu, J., Cheng, Y. and Nie, L. (2023). Micro-influencer recommendation by multi-perspective account representation learning. IEEE Transactions on Multimedia, 25:7463–7475.

Zhang, S., Yao, L., Sun, A. and Tay, Y. (2019). Deep learning based recommender system: a survey and new perspectives. ACM Computing Surveys, 52(1):1–38.

Stefanovic, P., Kurasova, O. and Strimaitis, R. (2019). The N-Grams based text similarity detection approach using self-organizing maps and similarity measures. Applied Sciences, 9(9):1870.

Cretulescu, R., Morariu, D. and Vintan, L. (2011). Clustering text documents: an overview. Acta Universitatis Cibiniensis, Technical Series.

Callaghan, S., Doner, H., Medalsy, J., Pione, A. and Teichner, W. (2024). The trends defining the global wellness market in 2024. McKinsey & Company.

Global Wellness Institute (2024). Global wellness economy monitor 2024. Global Wellness Institute.

Fuchs, M. and Hopken, W. (2022). Clustering: hierarchical, k-Means, DBSCAN. In Egger, R., editor, Applied Data Science in Tourism, pages 129–130. Springer.

Manning, C. D., Raghavan, P. and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Aggarwal, C. C. (2016). Recommender Systems: The Textbook. Springer.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP, pages 3982–3992. Association for Computational Linguistics.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I. (2021). BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Proceedings of NeurIPS Datasets and Benchmarks.

Grandisoli, L. C., Scalabrini Sampaio, G. and Eliseo, M. A. (2026a). A synthetic Brazilian Portuguese benchmark for briefing-centric influencer recommendation [Data set]. Zenodo. `https://doi.org/10.5281/zenodo.22047229`.

Grandisoli, L. C., Scalabrini Sampaio, G. and Eliseo, M. A. (2026b). Reproduction code for influencer recommendation from campaign briefings [Software]. Zenodo. `https://doi.org/10.5281/zenodo.22048849`.