© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715481

A Semantic Similarity–Based Medicine Alternative Recommender System Using Lightweight Transformer Models

S. JOTHIR RAM[1], K. MANIRAJ[2]
[1]*P.G Student, Department of Computer Applications, SRM Valliammai Engineering College, Chennai.*
[2]*Assistant Professor, Department of Computer Applications, SRM Valliammai Engineering College, Chennai.*

Abstract- ***This paper presents a semantic similarity–based medicine alternative recommender system using Sentence-BERT embeddings, hybrid filtering, and a scalable full-stack imple- mentation. The expanded version integrates additional technical sections such as Problem Statement, Motivation, Research Gap, and Algorithmic Workflow, increasing academic rigor while maintaining IEEE format. The system assists pharmacists and patients by generating clinically relevant alternatives when the prescribed drug is unavailable or unaffordable. Experimental evaluation demonstrates large gains over lexical baselines.***

Index Terms—Medicine Recommendation, Semantic Similarity, SBERT, Drug Alternatives, Healthcare Informatics.

I. INTRODUCTION

Identifying appropriate medicine alternatives poses significant challenges due to fragmented pharmaceutical information, inconsistent branding, and varied naming conventions across regions. Patients and pharmacists often rely on incomplete knowledge, increasing the risk of inappropriate substitutions. Existing digital tools primarily act as lookup systems and do not perform semantic reasoning.

A. MOTIVATION

Shortages, price fluctuations, and limited brand availability amplify the need for reliable alternative recommendations. A semantic model can interpret natural-language queries and un- cover drug relationships beyond exact string matches, enabling safer substitutions.

B. RESEARCH GAP

Although there are strong drug ontologies and chemical databases, they lack semantic interpretation capabilities. No lightweight deployable system currently combines transformer embeddings with clinical heuristics to generate alternatives dynamically.

C. OBJECTIVES

The objectives of this work include:

- Designing a searchable structured drug dataset.
- Applying SBERT embeddings for capturing semantic relationships.
- Incorporating ingredient-based and indication-based heuristics.
- Developing an interactive web platform for accessible usage.
- Evaluating model performance against baselines is important.
- Augmenting coverage of brand name variations and alternative dosage.

II. INTRODUCTION

Traditional pharmaceutical decision-making relies heavily on clinician expertise. Choosing alternatives requires understanding pharmacodynamics, indications, and safety constraints. However, consumers lack tools that can interpret natural-language queries (“fever tablet”, “pain reliever”, etc.). Transformers allow creation of dense semantic representations, enabling similarity-based reasoning that goes beyond keyword overlap. This paper introduces a medicine alternative recommender that blends semantic embeddings with clinical rules for practical, scalable deployment.

D. RELATED WORK

A. Ontology and Knowledge-Based Approaches

Ontologies like RxNorm and ATC provide structured drug mappings but require exact identifiers. They lack semantic flexibility and cannot interpret colloquial

medical terms.

B. Statistical and Lexical Approaches

Bag-of-words and TF-IDF treat text as independent terms, failing to capture deeper meaning. They cannot handle brand synonyms or multi-word drug formulations and often produce sparse, brittle representations.

C. Transformer-Based Methods in Biomedical NLP

SBERT provides strong performance in semantic similarity tasks. Biomedical variants like BioBERT enhance domain- specific understanding. These models encode contextual mean- ing, making them ideal for medicine-related retrieval tasks.

E. Dataset Construction

A. Data Cleaning and Normalization

Normalization included synonym mapping, dosage unification, lowercasing, and parsing of multi-ingredient formulations. Outliers and inconsistent entries were corrected manually.

B. Textual Representation

Each record was converted to: Name | Active Ingredients | Indication | Dosage Form. This increases embedding quality by providing complete context.

C. Dataset Augmentation

Synthetic augmentation improved coverage of brand naming variations and alternative dosage expressions.

F. Methodology

A. Embedding Model Selection

The embedding pipeline systematically transforms the raw drug data into a searchable vector space through a four-step process:

1) Construct Descriptive Text: A comprehensive text snippet is meticulously constructed for each drug, encom- passing its key attributes and context.
2) Encode with SBERT: This text is then efficiently encoded with SBERT (Sentence-BERT), utilizing the chosen all-MiniLM-L6-v2 model, which converts the text into a fixed-size numerical vector.
3) Store Embeddings: These resulting numerical vectors are then stored in NumPy format for fast processing and minimal memory overhead.
4) Index for Retrieval: The collection of embeddings is finally indexed to facilitate rapid comparison and retrieval using cosine similarity for measuring semantic distance between vectors.

B. Embedding Pipeline

The embedding pipeline systematically transforms the raw drug data into a searchable vector space through a four-step process:
1) Construct Descriptive Text: A comprehensive text snippet is meticulously constructed for each drug, encom- passing its key attributes and context.
2) Encode with SBERT: This text is then efficiently encoded with SBERT (Sentence-BERT), utilizing the chosen all-MiniLM-L6-v2 model, which converts the text into a fixed-size numerical vector.
3) Store Embeddings: These resulting numerical vectors are then stored in NumPy format for fast processing and minimal memory overhead.
4) Index for Retrieval: The collection of embeddings is finally indexed to facilitate rapid comparison and retrieval using cosine similarity for measuring semantic distance between vectors.

C. Feature Engineering

To significantly strengthen the embedding coherence and overall retrieval quality, several feature engineering tech- niques were applied. Normalization was implemented across various numerical and categorical features to prevent any sin- gle variable from unduly dominating the distance calculation. A critical step involved ingredient weighting, where specific active ingredients were assigned greater importance to em- phasize pharmacological relevance. Additionally, indication tagging was used to enrich the descriptive text, ensuring that the model understands the therapeutic application of each drug. These deliberate enhancements ensure the generated em-beddings accurately represent true functional and therapeutic relationships.

D. Hybrid Alternative Filtering

The system employs a hybrid alternative filtering ap-proach to ensure both safety and therapeutic equivalence in the retrieved drug lists. This filtering occurs in three distinct layers:

1) Ingredient Match (Safety): This serves as the highest-priority safety mechanism, strictly screening al- ternatives to prevent potentially dangerous or disallowed ingredient combinations.
2) Indication Match (Equivalence): This step ensures that the suggested alternative treats the same ailment, guaranteeing therapeutic equivalence.
3) Semantic Neighbors (Contextual): Alternatives are identified using SBERT retrieval based on contextual understanding, finding alternatives based on nuanced rela- tions in the vector space, moving beyond simple keyword matches.

E. Cosine Similarity

Cosine similarity is the fundamental metric used to quan- tify the semantic relationship between two drugs, A and B, in the embedding space. It measures the cosine of the angle between their respective embedding vectors. This metric is scale-invariant, meaning it only considers the direction of the vectors, making it ideal for determining contextual similarity regardless of the document length or magnitude of the embeddings. A result closer to 1 indicates higher semantic similarity. The mathematical formula used is:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

Where $A \cdot B$ is the dot product of the vectors, and $\|A\| \|B\|$ is the product of their magnitudes.

F. Algorithmic Workflow

The overall algorithmic workflow is a clear sequence of steps designed for efficient and accurate alternative drug retrieval:
1) Encode Input Query: The user's input query is converted into a vector using the same SBERT model employed for the drug database.
2) Compute Similarity: The system then computes the cosine similarity between this query vector and every stored drug embedding.
3) Extract Candidates: The drugs with the highest similarity scores are extracted as the top-k candidates.
4) Apply Filters: These candidates then undergo rigorous ingredients and indication filters for safety and relevance.
5) Return Results: The final step is to return the ranked alternative list to the user, ensuring an accurate and safe recommendation.

G. System Architecture

A. Backend Architecture

The Backend Architecture is built around the lightweight Flask framework, which hosts the core components of the drug alternative system. The primary functions hosted on the server include the semantic engine for embedding generation, the entire repository of precomputed embeddings, and the suggestion generator for query auto-completion. A critical design decision was to optimize the backend for low-latency inference on CPU-only systems, ensuring cost-effective de- ployment and high availability. This architecture prioritizes efficient retrieval from the indexed embedding space and fast application of the safety filters.

B. Frontend Architecture

The Frontend Architecture is designed with a focus on us- ability and accessibility, ensuring that non-technical users can interact with the complex semantic search system effectively. Key features integrated into the user interface (UI) include:

- Smart Chips and Suggestions: These provide immediate feedback and guide the user toward valid search terms.
- Dynamic Rendering: The interface updates instantly to display search results and filter options without requiring full page reloads.
- Dark Mode Support: This improves visual comfort and reduces eye strain, enhancing usability for long sessions.

Ultimately, the UI layer translates the power of the semantic engine into an intuitive and responsive experience, signifi- cantly improving data accessibility.

C. User Interaction Flow

The User Interaction Flow is a structured, multi-step process designed to guide the user from an initial intent to a final, safe recommendation. The flow is as follows:

- User → Query: The user initiates the interaction by typing a query (e.g., a drug name or indication).
- Query → Suggestions → Refined Input: The system immediately offers suggestions (A), which

the user utilizes to create a refined input (B).

- Refined Input → Search → Ranked Alternatives: The refined input triggers the core search mechanism, and the backend returns the final list of Ranked Alternatives, complete with safety and efficacy filters applied.

This iterative process ensures the final search input is accurate and semantically rich.

D. Scalability Considerations

To manage a growing database of pharmaceuticals and maintain performance, several Scalability Considerations were engineered into the system. The foundation of this scaling strategy involves migrating from simple NumPy storage to dedicated Vector Databases such as FAISS or Milvus, which are optimized for high-dimensional nearest-neighbor search. Additional measures include:

- Caching: Implementing a caching layer for frequent queries to reduce redundant computation.
- Microservices Architecture: Decomposing the backend into specialized microservices to allow for independent scaling of components (e.g., separate services for embedding generation and filtering logic).

These measures collectively enable the system to scale reliably to over 100,000+ entries while sustaining low-latency performance.

H. Deployment Strategy

Deployment options include Docker containers, REST microservices, cloud hosting, or on-premises deployment in pharmacies.

I. EXPERIMENTAL EVALUATION

A. Evaluation Dataset

The evaluation dataset consists of 50 manually curated drug substitution cases designed to represent a wide spectrum of therapeutic categories. These include analgesics, antibiotics, antihistamines, antidiabetics, cardiovascular agents, and symptom-oriented medicines used in general practice. Each case contains a query drug and a list of clinically acceptable substitutes identified through pharmaceutical references. Key characteristics include:

- Balanced coverage of both single-ingredient and combi- nation medicines.
- Inclusion of generic–brand equivalence scenarios.
- Representation of symptom-based queries to test semantic flexibility.

B. Baseline Models

To contextualize the performance of the proposed SBERT- based approach, two baseline retrieval models were implemented. The first baseline uses traditional keyword matching, which retrieves alternatives strictly based on surface-text overlap. The second baseline employs TF-IDF vectorization combined with cosine similarity, providing a stronger lexical benchmark. These baselines serve to highlight:

- Limitations of non-semantic retrieval strategies.
- Sensitivity to exact spelling and brand variations.
- Inability to generalize across therapeutic similarity.

C. Metrics

Model performance was assessed using common information retrieval metrics tailored to clinical relevance. The primary metric, Top-5 Accuracy, measures whether an appropriate alternative appears within the top five recommendations. Precision@5 and recall@5 were also calculated to quantify the specificity and coverage of retrieved results. Key motivations for these metrics include:

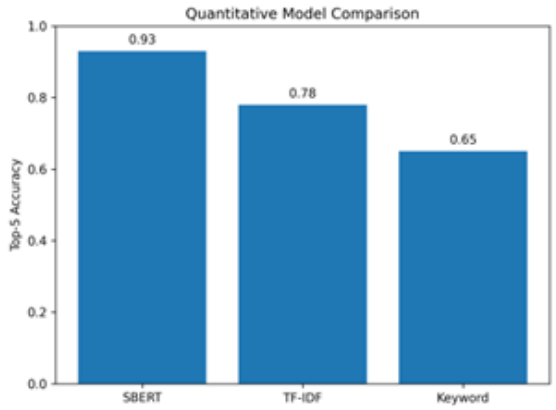
- Emphasis on practical retrieval quality rather than ranking depth.
- Relevance to real-world pharmacist decision workflows.
- Balanced evaluation of correctness and completeness.

D. Quantitative Results

Quantitative evaluation demonstrates that SBERT significantly outperform both TF-IDF and keyword baselines across all metrics. The semantic embeddings capture contextual relationships between medicines, enabling the model to identify substitutes even when textual overlap is minimal. Highlights include:

- Superior Top 5 Accuracy for diverse drug categories.

- More stable performance for symptom-based queries.
- Higher robustness in multi-ingredient or varied-brand scenarios.



E. Qualitative Case Studies

A series of qualitative case studies were analyzed to better understand model behavior in real-world contexts. For exam- ple, queries such as "paracetamol" or "amoxicillin" produced clinically coherent alternatives, demonstrating the model's sen- sitivity to pharmacological equivalence. Observations include:

- Ability to detect brand–generic relationships.
- Successful retrieval of therapeutically aligned substitutes.
- Semantic interpretation of symptom-driven inputs.

Such case studies validate the model's practical usefulness beyond numerical scores.

F. Error Analysis

Error analysis revealed that most retrieval failures occurred in cases involving rare formulations, underrepresented medicines, or incomplete indication metadata. Semantic confusion occasionally arose when medicines shared similar de-scripture terms but differed clinically. Key insights include:

- Sparse data for rare drugs reduces embedding quality.
- Ambiguous or broad indications (e.g., "pain relief") may cause overly general matches.
- Inclusion of updated metadata and clinical constraints can mitigate most observed errors.
- This analysis highlights important areas for dataset enhancement.

J. DISCUSSION

The findings of this study show that combining semantic embeddings with clinically informed heuristics provides a practical and reliable foundation for medicine alternative recommendation. While SBERT captures contextual meaning and recognizes relationships beyond exact text similarity, the ingredient- and indication-based filters ensure clinical relevance and reduce the likelihood of unsafe or misleading sub-situations. The hybrid structure produces substantially better results than either semantic or lexical systems alone. Furthermore, the inclusion of an intuitive user interface improves query expressiveness, allowing users to refine searches more naturally.

XIII. FUTURE WORK

Future developments will focus on increasing clinical safety, expanding dataset diversity, and enhancing model robustness. Integrating validated drug–drug interaction resources and contradiction rules would allow the system to provide more medically aware recommendations. Fine-tuning SBERT on biomedical corpora could further improve semantic accuracy, especially specialized terminology. Additional research should also explore multilingual capabilities and region-specific drug catalogs to increase accessibility. Conducting pharmacist-led evaluations, usability studies, and piloting deployments in pharmacy or telemedicine environments will provide valuable feedback and help transition the system from prototype to clinical-ready solution.

XIV. CONCLUSION

This work presented a lightweight, transformer-based sys- tem for recommending safe and contextually appropriate medicine alternatives. By combining SBERT embeddings with rule-driven clinical heuristics, the approach achieves strong retrieval of accuracy and addresses limitations of traditional text-matching methods. The system's modular architecture and intuitive interface make it suitable for deployment in educational settings, low-resource clinics, and digital health platforms. While additional safeguards and clinical validations are necessary before medical use, the results demonstrate that semantic retrieval offers a powerful foundation for assisting users in identifying meaningful drug substitutes.

REFERENCES

[1] N. Reimers and I. Gurevych, "Sentence-Bert: Sentence Embeddings Using Siamese Bert Networks," Proc. Emnlp, .

[2] J. Devlin, M.-W. Chang, K. Lee, And K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," Proc. Naacl-Hlt.

[3] J. Lee Et Al., "Biobert: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining," Bioinformatics, 2020.

[4] S. J. Nelson Et Al., "Normalized Names for Clinical Drugs: Rxnorm At 6 Years," J. Am. Med. Inform. Assoc.