National College of Ireland

MedReview: A Graph-Enhanced Framework for Comparing Domain-Specific and General Language Models in Consumer Drug Review Analysis

MSc Research Project
MSc in Data Analytics

RAMAKRISHNA REDDY VENREDDY
Student ID: x23318503

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ramakrishna Reddy Venreddy
Student ID:	x23318503
Programme:	MSc in Data Analytics **Year:** 2024-2025
Module:	Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	15th Sep 2025
Project Title:	MedReview: A Graph-Enhanced Framework for Comparing Domain Specific and General Language Models in Consumer Drug Review Analysis
Word Count:	8426 **Page Count:** 26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ramakrishna Reddy Venreddy

Date: 15th Sep 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

MedReview: A Graph-Enhanced Framework for Comparing Domain-Specific and General Language Models in Consumer Drug Review Analysis

Ramakrishna Reddy Venreddy
Student ID: x23318503

Abstract

Consumer drug reviews represent a significant valuable source of pharmacovigilance information but present significant challenges due to the terminology gap between informal patient language and professional medical vocabulary. This paper introduces MedReview as one new framework making use of graph-based semantic bridging with language models to enhance the analysis of consumer drug reviews. The study addresses the research question: How can graph-based semantic bridging enhance the performance of both domain-specific and general language models in analyzing consumer drug reviews? The framework combines medical knowledge graphs constructed from UMLS and BioPortal with six language models: domain-specific models (PubMedBERT, BioBERT, ClinicalBERT) and general-purpose models (BERT, RoBERTa, DeBERTa). Through systematic experimentation on 50,000 drug reviews, the research demonstrates that graph enhancement provides substantial performance improvements across all models, with medical models showing the highest gains (22.39% average F1 improvement in sentiment classification). The best baseline model (DeBERTa) achieved an F1 score of 0.7042, while its graph-enhanced variant reached 0.8099, representing a 15% improvement. The multi-task assessment with sentiment classification, effectiveness regression, and multi-aspect classification proves fine-grained semantic tasks benefit the most with structured knowledge integration (average of 38.5% improvement). The work proves the graph-based semantic bridging across differences of vocabulary to be successful and enables the general and the medical model to capture the language of consumer health more effectively. This work contributes to healthcare informatics with the efficient process for bridging the gap between the healthcare systems and the patients and it can have real-world applications where pharmacovigilance and patient-centric care are involved.

1 Introduction

Large databases of user-written medication reviews have been produced by the unplanned expansion of online health platforms (Gawich & Alfonse, 2022; Haque et al., 2023), which is both problematic and promising for healthcare informatics. Because of the crucial disassociation between everyday patient language and formalised medical vocabulary, these reviews are difficult for informaticists to evaluate, despite being valuable patient experience data that may be used to inform improved pharmacovigilance systems and clinical decisions

(Nazi & Peng, 2024). The ensuing language barrier prevents healthcare systems from fully utilising patient-reported outcomes, possibly missing important information about the effectiveness, side effects, and quality-of-life effects of medications as communicated by patient in their own natural tongue.

Modern techniques for analysing medical texts have generally taken two routes.
Specialized domain language models such as BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and PubMedBERT (Gu et al., 2021) reach state-of-the-art on expert medical text by benefiting from specialized pre-training on biomedical literature and clinical notes. These models are poor on consumer-generated content, however, since their training exposes them predominantly to formal medical language. General-purpose language models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020) attain wider linguistic coverage but need specialized medical knowledge for accurate health-related entity recognition and relationship understanding. Neither approach sufficiently addresses the underlying problem of bridging the consumer-professional vocabulary gap.

State-of-the-art advances in graph-based natural language processing are very promising for deriving semantic relations between ideas. Knowledge graphs can represent complex medical relations by bridging informal patient expressions to standard medical terminology through intermediate nodes and relation edges. Graph-augmented approaches were successful in biomedical text summarization (Givchi et al., 2022) as well as in clinical question answering systems, thereby also revealing potential for terminology bridging applications. However, graph-based semantic networks' integration with transformer-based language models for consumer health text processing is hardly explored.

It addresses such shortcomings by presenting MedReview, a novel framework where graph-based bridging with semantics is supplemented by domain-specific and general language models for aiding consumer drug reviews. The primary research question guiding this investigation is: **How can graph-based semantic bridging enhance the performance of both domain-specific and general language models in analyzing consumer drug reviews?**
The objectives of this research are fourfold:

1. To develop a graph-enhanced framework that integrates medical knowledge graphs with language models for improved drug review analysis
2. To systematically compare the performance improvements achieved through graph enhancement across domain-specific models (PubMedBERT, BioBERT, ClinicalBERT) and general models (BERT, RoBERTa, DeBERTa)
3. To develop and assess a terminology bridging mechanism that uses graph-based semantic relationships to translate consumer expressions to standardised medical concepts.
4. To assess how well the framework performs on a range of drug review analysis tasks, including entity recognition, sentiment classification, and adverse event detection. Quantitative measures like F1 scores for entity recognition procedures, increases in the precision of adverse event identification, and comparisons between baseline and graph-augmented models will be used to evaluate the achievement of such goals. The effectiveness of the framework for projecting colloquial expressions to correct medical

concepts will also be evaluated through a qualitative analysis of bridging effectiveness in terminology.

The system employs a broad strategy that blends a number of complementary methods. Medical knowledge graphs are created by incorporating existing biomedical APIs like the Unified Medical Language System (UMLS) and BioPortal, providing authoritative concept mappings and semantic links. These APIs are used by the system to establish connections between consumer and standard medical terminology and to cross-check medical entities that are extracted. Graph embeddings are constructed using node2vec algorithms and are further integrated with language model embeddings using concatenation as well as attention mechanisms. Publicly accessible drug review datasets are used to test the system, and both real reviews and synthesised data are used for controlled evaluation and real-world verification, respectively.

The evaluation employs six latest language models for thorough comparison: domain-specialized models (PubMedBERT, BioBERT, and ClinicalBERT) as well as general-purpose models (BERT, RoBERTa, and DeBERTa-v3). All models are implemented for baseline as well as graph-enhanced configurations such that systematic examination of the value added by integrating the semantic graph is allowed for various model structures as well as pre-training methods.

Its estimated value addition goes beyond research in academia to real-world applications in healthcare practice. Academically, the work enriches overall knowledge about why graph-based methods are complementary to transformer-based methods for text analysis for medical applications. The contrast between general vs. domain-specific models is insightful about trade-offs between pre-trained specialized models vs. semantically enriched models using outside knowledge. Practically, the system has the potential for better systems for pharmacovigilance by being able to better extract evidence for adverse events and for experience in treatments from patient text. Pharmaceutical companies and healthcare providers are potential beneficiaries of enhanced capacity for interpreting patient-reported outcomes in natural language.

The structure of the report is as follows: Section 2 has a detailed literature review where current methods for processing medical language are compared, graph-based NLP methods evaluated, and methods for analyzing drug reviews summarized. Section 3 describes research methodology in terms of the architecture for Med Review, data preparation work on the data set, and plans for evaluation. Section 4 describes design specifications for graph-augmented architecture. Section 5 describes implementation details. Section 6 reports experimental results with critical discussion on various measures for evaluation. Section 7 interprets results against research objectives and pertinent literature. Finally, Section 8 summarizes conclusions with prospects for further work.

2 Related Work

This review of literature explores recent research on medical text analysis focusing on domain-specific language models, graph-based approaches, and methods for analyzing drug reviews. Each category is critically assessed to identify gaps relevant to the expansion of language models by graph-based semantic bridging for consumer drug reviews.

2.1 Domain-Specific Language Models in Healthcare

Medical language model development offers both advancement and limitations for text processing in healthcare. Devlin et al. (2018) initiated BERT with bidirectional pre-training standardizing language understanding with masked language modeling. BERT achieves state-of-the-art NLP task outcomes but is limited by general-purpose pre-training on BookCorpus and Wikipedia in achieving effectiveness for specialized medical terminology. Even though possessing superior general language skills, the model has failed to learn domain-related medical knowledge needed for healthcare practice.

Lee et al. (2019) mitigated such a drawback by using BioBERT, pre-trained on the PubMed abstracts together with PMC articles. BioBERT demonstrates 4.9% improvement on biomedical NER as well as 3.7% on relation extraction over BERT. However, it is trained on formal biomedical literature and is therefore challenging for processing consumer health content with informal language together with colloquial expressions. Similarly, Huang et al. (2019) built ClinicalBERT by using 2 million notes in clinical practice, demonstrating excellent clinical task performance like readmission prediction. However, clinical documentation maintains formal structure unlike patient-written reviews, thereby limiting generalizability on consumer health text.

Rasmy et al. (2021) extended with new pre-training on structured EHR data for 28 million patients using Med-BERT. Med-BERT achieves AUROC 0.858 for heart failure prediction by learning medical relationships and time patterns. However, it further widens the gap from consumer language by pretraining on coded medical data versus text-based models. Aftan and Shah (2023) surveyed BERT applications to verify domain-specific fine-tuning is helpful for performance while also highlighting lasting terminology mismatch concerns. Nazi and Peng (2024) extended the comparison to large language models while also revealing the large gap between professional medical school background and patient-generated content requirements.

All such models demonstrate a pattern: excellent performance on professional medical text but poor success on consumer health language. The critical issue of interpreting patient reports of experience in natural language is not resolved.

2.2 Graph-Based Approaches in Medical Text Analysis

Graph-based methods are ideal for extracting complex medical relationships beyond the confines of sequential model capabilities. Givchi et al. (2022) present graph-based abstractive summarization for biomedical text by combining graph building with frequent itemset mining. The method achieves 59.60% ROUGE score with 17% advances over baselines. While

extracting biomedical concept relations efficiently, it focuses on summarization as opposed to terminology bridging between terms and is heavily reliant on UMLS, limiting potential for patient language informality processing.

Shathyan et al. (2023) constructed a knowledge graph-based medical chatbot on TigerGraph to carry out graph traversal for disease identification with improved effectiveness. The system is able to traverse intricate medical relationships but is constrained by predefined mappings between symptoms and diseases that might not work for diverse patient descriptors. Non-incorporation of language models constrains natural language processing.

Hoyt et al. (2025) surveyed Graph Convolutional Networks for EHRs where they achieved success in mortality prediction as well as modeling disease progression. GCNs are especially suitable for patient similarities as well as temporal pattern extraction but methods surveyed are more interested in structured data than text. Nguyen et al. (2020) introduced graph-based population health analysis for geo-tagged tweets where state-of-the-art results were achieved by employing inter-feature graphs from latent topics. Good performance on consumer-generated content by them is promising but population-level attention as well as the absence of incorporation with language models limits generalizability to individual drug reviews.

Wu et al. (2024) proposed HTSRKG by incorporating knowledge graphs into syntactic structures for further powerful semantic representation. The hierarchy-based approach offers guidance for multi-level investigation but is focused on phrase extraction with no medical domain adaptation or transformer integration. Mohawesh et al. (2023) built semantic graph topic modeling for multilingual identification of false news by bridging linguistic disparities with success. Their multi-layered architecture with gated attention networks has the potential for filling medical terminology gaps, though extensive healthcare adaptation would occur.

2.3 Drug Review Analysis Techniques

State-of-the-art drug review analysis shows developments as well as established challenges in managing pharmaceutical comments. Nair et al. (2024) tried out sentiment analysis by employing pre-trained language models (BERT, SciBERT, BioBERT) together with traditional classifiers, achieving optimal efficacy with SciBERT embeddings as well as Random Forest. In showing worth of domain-focal embeddings, domain-centric experiment overlooks medical entity recognition together with terminology normalization challenges.

Gawich and Alfonse (2022) constructed extensive ADR discovery with 87% accuracy on Drugs.com, revealing 21.03% reviews containing ADR data. Although extensive preprocessing with NER is employed, it does not include consumer-to-professional term mappings. The system mines ADRs from consumer text but does not map to standard vocabularies for systematic pharmacovigilance.

Haque et al. (2023) achieved improved classification with Random Forest for 96.65% accuracy by considering Count Vectorizer features. These engineering observations about the feature are

fascinating but single sentiment concentration ignores aspect-level study as well as recognition of medical entities. Accurate sentiment does not mean understanding individual medical terms.

Durga et al. (2024) also proposed high-level aspect-based mining using ant colony optimization as well as deep learning with 96.78% and 95.02% accuracy on druglib.com and drugs.com correspondingly. Their RoBERTa-BiLSTM ensemble allows for fine-grained perception of drug aspects. However, it operates on consumer word space with no professional term correspondence to standardized medical concepts. While extracting comments on "effectiveness" or "side effects," it is unable to map precise details to standardized medical concepts.

These approaches all reach state-of-the-art classification accuracy but all lack mechanisms for bridging consumer to professional medical vocabularies thereby are all incomplete for general-purpose pharmacovigilance.

2.4 Research Niche

Critical evaluation reveals a significant gap where the three domains intersect. Expert medical models succeed on professional text but fail on consumer articles due to terminological mismatches. Graph models maintain complex relationships with no interfacing with modern transformer-based models for consumer health reasoning. Drug reviewing methods achieve superior classification with no correction for terminological root bridging.

Overlapping limitations refer to the root problem: ineffective bridging between professional medical and consumer vocabulary. Though domain-related models embody medical knowledge and relationships between concepts are captured by graphs, no method possesses such strengths combined for consumer drug reviews. That is why the healthcare systems fail to capitalize on patient-reported data for clinical decisions and pharmacovigilance.

Trying to fill the void, the new work proposes a graph-supplemented system integrating language models with medical knowledge graphs. Through the integration of transformer contextual cognition with graph relational knowledge, the system establishes relationships between professional terminologies and consumer expressions. The new integration facilitates both theoretical cognition and practice applications for consumer health text mining in order to further mine patient-generated review writings for medical knowledge.

3 Research Methodology

This chapter presents the methodology for investigating graph-based semantic bridging to enhance language model performance in consumer drug review analysis. The approach combines empirical data collection, natural language processing, and graph neural networks to address terminology disparity between consumer and professional medical language.

3.1 Research Framework Overview

Figure 1 illustrates the overall methodology workflow with parallel paths for baseline and graph-aided model processing. Controlled experimental design is used by the framework in three steps: data pre-processing with balanced sampling, construction of knowledge graph from verified medical entities, and detailed model evaluation. The workflow indicates how 184,622

initial reviews undergo entity extraction and verficiation to form a 500-node knowledge graph that aids in model performance enhancement.

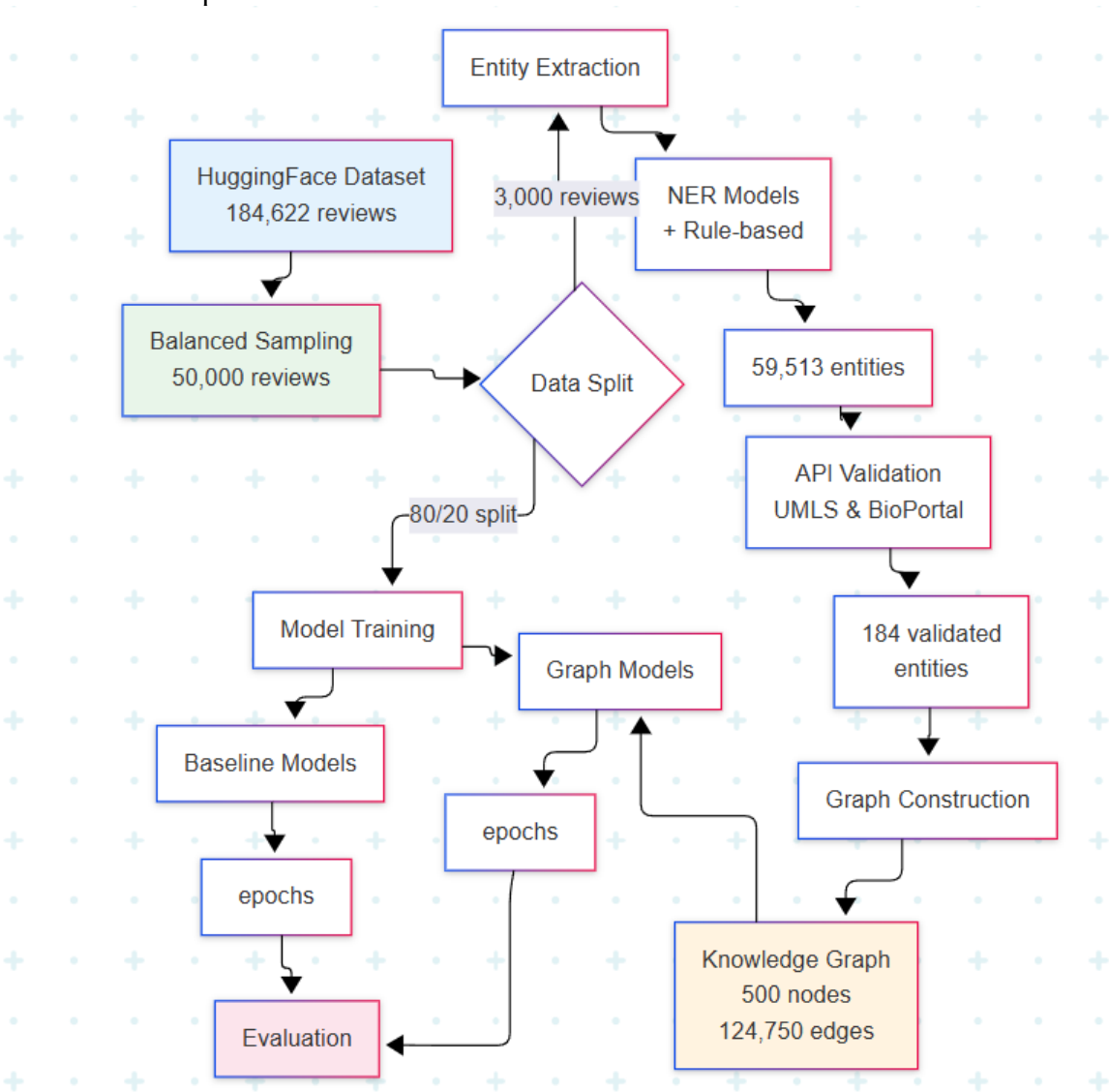


FIGURE 1 Research Methodology Workflow - illustrating the complete pipeline from data collection through model training, highlighting the parallel paths for baseline and graph-enhanced models

3.2 Data Collection and Preparation

3.2.1 Dataset Selection

The research utilized the Drug Review Dataset from HuggingFace (forwins/Drug-Review-Dataset), comprising 184,622 consumer-authored pharmaceutical reviews with training (110,811), validation (27,703), and test (46,108) splits. Each review contains drug name, medical condition, review text, numerical rating (1-10 scale), date, and usefulness count. The dataset represents real-world consumer text with natural language variations and informal terminology exemplifying the vocabulary gap central to this research.

3.2.2 Balanced Sampling Strategy

The original dataset exhibited severe class imbalance with ratings 8-10 comprising 89,234 reviews (48.3%), ratings 6-7 containing 42,876 reviews (23.2%), and ratings 1-5 including 52,512 reviews (28.5%). Rating 10 alone accounted for 31,678 reviews (17.2%) while rating 1 contained only 4,893 reviews (2.7%), representing a 7:1 imbalance between extremes.

To address this imbalance, stratified sampling selected exactly 5,000 reviews from each rating class (1-10), creating a balanced dataset of 50,000 reviews with uniform 10% representation per rating. The balanced dataset contains 25,000 negative reviews (ratings 1-5), 10,000 neutral reviews (ratings 6-7), and 15,000 positive reviews (ratings 8-10), maintaining a 5:2:3 ratio. Random selection with seed 1999 ensures reproducibility. This balancing eliminates confounding effects from class imbalance, allowing performance differences to be attributed solely to graph-based semantic bridging.

3.3 Medical Entity Extraction

Following Figure 1's extraction pathway, the methodology implemented a hybrid approach combining neural models with rule-based systems to capture both formal and informal medical expressions.

3.3.1 Named Entity Recognition

Medical entity extraction utilized two domain-specialized biomedical NER models, namely d4data/biomedical-ner-all (generalized training on broad biomedical data) and samrawal/bert-base-uncased_clinical-ner (domain-specialized for clinical nomenclature). The pipeline analyzed reviews in batches (size 20) with low confidence (0.3) settings, intentionally low due to colloquial forms used in consumer reviews. This two-model composition addresses the limitation of single models, giving better overall entity coverage.

3.3.2 Hybrid Enhancement

The rule-based component maintained three medical term categories: drug names (generic and brand), conditions (diseases, disorders, syndromes), and symptoms (side effects, clinical manifestations). Entity scoring assigned weights based on NER confidence levels, vocabulary matches (2.0x boost), and frequency across reviews. This multi-factor approach extracted 59,513 candidate entities from 3,000 sampled reviews, balancing coverage, relevance, and specificity.

3.4 Entity Validation and Selection

The validation process shown in Figure 1 verified medical relevance through external knowledge bases.

3.4.1 API Validation

Two medical terminology services validated entities: UMLS API**(Givchi et al., 2022)** (accessing 200+ biomedical vocabularies) and BioPortal API (biomedical ontologies for drugs and anatomy). The system queried both APIs for the top 300 entities, verifying medical relevance and extracting metadata including Concept Unique Identifiers (CUIs), semantic types, and alternative names. Successfully validated entities received a 3.0x score boost.

3.4.2 Final Selection

Out of 59,513 candidates, 184 API-verified entities passing relevance criteria were selected by the process. Extraction confidence, word match, frequency of occurrence, and status of

validation were used as priorities by the algorithm. Thresholds of semantic similarity avoided redundancy of concepts and provided representation with diversity among drug name, condition, symptom, and treatment terminology.

3.5 Knowledge Graph Construction

Following Figure 1's graph construction path, the methodology built a dense medical knowledge network.

3.5.1 Graph Building

Co-occurrence analysis revealed relationships between verified entities within review contexts. For reviews with multiple entities, the system calculated pairwise relationships according to text proximity and contextual relevance. Edge weights blended co-occurrence frequency with semantic similarity, employing a 0.15 threshold to balance connectivity with relationship quality while eliminating spurious connections.

3.5.2 Implementation

PyTorch Geometric implemented the knowledge graph with 768-dimensional node features matching transformer hidden dimensions(Hoyt et al., 2025). The final structure comprised 500 nodes representing pharmaceutical drugs (metformin, lisinopril), medical conditions (diabetes, hypertension), and symptoms/side effects (nausea, fatigue). The interconnectedness of medical concepts in drug reviews is reflected in the dense network with an average degree of approximately 250 created by the connections between these nodes via 124,750 weighted edges.

3.6 Model Training

Figure 1's parallel training paths show the systematic evaluation across diverse architectures.

3.6.1 Experimental Design

Twelve model variants were compared in the study: six baseline models and their graph-enhanced counterparts. The baselines comprised medical-specific models (BioBERT, ClinicalBERT, PubMedBERT) as well as general-purpose models (BERT, RoBERTa, and DeBERTa). Preliminary experiments indicate that graph models need more iterations to learn cross-modal integration, as baseline models trained for two epochs while graph-enhanced variants received eight epochs. By permitting convergence based on architectural complexity, this differential strategy guarantees fair comparison.

3.6.2 Training Configuration

To maintain the balance of the rating distribution, training employed 80/20 stratified train-validation splits. Among the configurations were batch size 16 with 4-step gradient accumulation (effective batch 64), linear warmup (500 steps) with cosine annealing, AdamW optimiser with 0.01 weight decay, and learning rates of 2e-5 (baseline) and 3e-5 (graph-enhanced) to support additional graph parameters.

3.7 Evaluation Framework

Figure 1's evaluation stage assessed performance across three complementary tasks.

3.7.1 Multi-Task Assessment

Task 1 assessed overall polarity capture by implementing sentiment classification (negative, neutral, and positive). Task 2 extracted subtle therapeutic indicators from intricate narratives by performing effectiveness regression (0–1). Subtle sentiment discrimination was necessary for Task 3's multi-aspect classification, which had six levels ranging from very poor to excellent. The advantages of graph enhancement are guaranteed to transfer across analytical dimensions thanks to this multitask configuration.

3.7.2 Performance Metrics

Weighted F1 (which accounts for class imbalance) and macro F1 (which treats all classes equally) were used in the classification tasks. Regression used R2 to measure variance explained and MSE to measure prediction accuracy. This extensive metric set shows the advantages of graph enhancement while avoiding over-optimization of single metrics.

3.8 Computational Environment

Experiments utilized Google Colab Pro with NVIDIA A100-SXM4-40GB GPU (40GB memory, 83GB RAM). Software included Python 3.11.13, PyTorch 2.6.0, PyTorch Geometric 2.6.1, HuggingFace Transformers, NumPy, Pandas, and Scikit-learn. Fixed random seeds ensured reproducibility across all experiments.

4 Design Specification

4.1 System Architecture Overview

MedReview framework adopts a multi-modal design by integrating transformer-based language models with graph neural networks in order to fill the term discrepancy between consumer drug reviews and medical knowledge bases. It has three subsystems, namely, a data processing pipeline to extract and verify entities, a knowledge graph module used to model medical relationships, and a neural design that merges textual and graph representations using cross-modal attention.

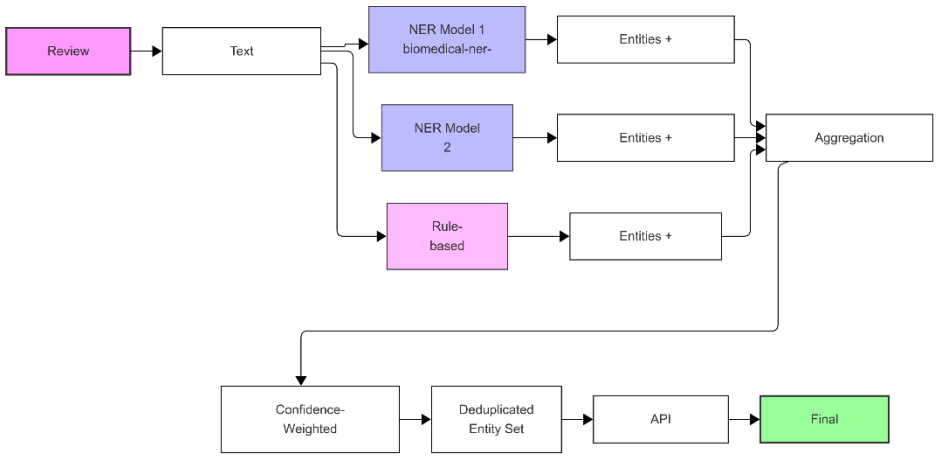
The framework facilitates bidirectional information exchange where transformer embeddings question the medical knowledge graph for pertinent concepts and graph representations pay attention to textual contexts for semantic anchoring. The modular design guarantees that components can individually be made optimal but are altogether fully integrated using standardized interfaces, directly meeting research aims of utilizing structured medicine knowledge to understand informal consumer language.

4.2 Data Processing Pipeline Design

The pipeline implements parallel processing where several NER models run in parallel, generating entity candidates that are subject to confidence-weighted aggregation. The design adopts a producer-consumer model with entity extractors writing to a central aggregation queue for asynchronous processing as well as optimal resource utilization.

4.2.1 4.2.1 Parallel Entity Extraction Pipeline

The parallel processing architecture coordinates multiple NER models extracting entities concurrently from consumer reviews. Figure 3 illustrates this pipeline where review text flows through parallel extraction paths before aggregation.



For example, when processing the review text "Metformin caused terrible stomach problems but helped my sugar levels," the pipeline operates as follows: Model 1 extracts "Metformin" (confidence: 0.95) and "stomach" (confidence: 0.72), Model 2 identifies "sugar levels" as "blood glucose" (confidence: 0.68), while the rule-based extractor captures "stomach problems" as an adverse effect pattern. The aggregation queue combines these extractions, with "Metformin" receiving the highest final score due to high confidence and vocabulary matching.

The API validation layer implements two-tier caching (persistent and in-memory) with circuit breaker patterns for fault tolerance. Rate limiting uses token bucket algorithms to optimize throughput while respecting service limits.

4.3 4.3 Knowledge Graph Architecture

The graph architecture employs hybrid representation where nodes retain semantic embeddings as well as structural information. Every node has a 768-dimensional compatible transformer hidden state feature vector, allowing for direct integration without dimensionality conflicts. Multiple relation types are accommodated by edge architecture using learnable embeddings, separating drug-symptom, drug-condition, and symptom-treatment relationships.

PyTorch Geometric integration utilises sparse tensor form to provide memory-efficient storage for 124,750 edges amongst 500 nodes. Tailored message passing for propagation of medical knowledge is provided by the architecture using weighted messages based upon edge importance alongside API-confirmed node credulity. Batch processing constructs disconnected super-graphs, supporting parallel processing made possible through maintenance of logical sample separation.

4.4 4.4 Neural Network Architecture Design

4.4.1 4.4.1 Baseline Model Architecture

The baseline architecture provides a unified interface for integrating various pre-trained language models. A common feature extraction layer normalizes outputs to 768 dimensions regardless of transformer architecture, enabling seamless model swapping without architectural modifications.

The shared representation pathway implements an expansion-compression pattern (768 → 1024 → 512 → 256) with GELU activations and pre-layer normalization. Task-specific heads branch from shared representations: sentiment classifier (256 → 128 → 64 → 3), effectiveness regressor (256 → 128 → 64 → 1) with sigmoid activation, and multi-aspect classifier (256 → 128 → 64 → 6). Each head uses task-optimized dropout rates and activation functions.

4.4.2 4.4.2 Graph-Enhanced Architecture

The Graph Attention Network implements specialized medical knowledge propagation as shown in Figure 2. The three-layer GAT processes the medical knowledge graph containing drug entities, disease terms, and symptom concepts. The architecture uses 8 attention heads in the first layer for capturing diverse relationship types, followed by representation refinement and feature aggregation for text integration.

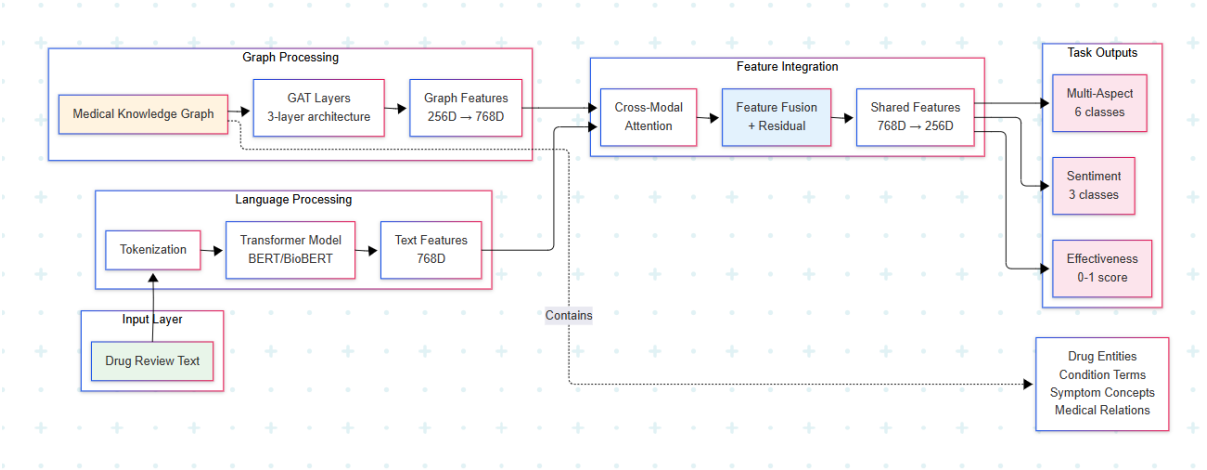


Figure 1 Graph-Enhanced Architecture - illustrating the integration of language models with medical knowledge graphs through cross-modal attention, showing how the graph contains medical entities and relationships that enhance text understanding

4.5 4.5 Multi-Task Learning Design

The multi-task architecture implements hard parameter sharing such that feature extraction layers are shared among all the tasks before they diverge into special-purpose heads. With such design, knowledge transfer among related tasks is enhanced, whilst undesirable interference is prevented using task-specific fine-tuning. Shared layers learn representations transferable from one task to another, acquiring medical entity patterns along with sentiment indicators valuable for comprehensive drug review analysis.

4.6 4.6 Model Integration Strategy

4.6.1 4.6.1 General-Purpose Models

Three general-purpose models establish baselines: BERT provides bidirectional context understanding from BooksCorpus and Wikipedia training, RoBERTa extends BERT through robust optimization and additional web content, and DeBERTa introduces disentangled attention for improved dependency capture. These models excel at language patterns and syntactic structures but lack medical terminology exposure. The integration preserves pre-trained weights while adding graph-integration layers, enabling structured medical knowledge access without forgetting general language capabilities.

4.6.2 4.6.2 Domain-Specific Models

Medical models bring specialized healthcare knowledge: BioBERT (trained on 4.5B PubMed words and 13.5B PMC words) recognizes biomedical terminology and research patterns, ClinicalBERT (MIMIC-III clinical notes) captures practical medical language and abbreviations, and PubMedBERT represents pure biomedical understanding from exclusive PubMed training. These models test whether graph enhancement adds value beyond existing medical knowledge, particularly for bridging formal medical training and informal consumer terminology.

4.6.3 4.6.3 Optimization Strategy

The optimization employs differential methods depending on model properties. General models apply elevated learning rates for graph elements to quickly obtain medical knowledge, whereas medical models apply conservative rates to retain specialized pre-training. Conditional freezing of lower transformer levels occurs for medical models during fine-tuning for general models. While medical models use graphs primarily for bridging terms, model-aware mixing ensures that general models will rely primarily on graph attention for medical concepts, optimising enhancement based on existing knowledge gaps.

5 Implementation

5.1 Implementation Overview

With a modular Python codebase consisting of roughly 2,500 lines of core functionality, the MedReview system implementation achieved the intended architecture. In order to enable independent component testing with smooth integration, the coding placed a strong emphasis on modularity and extendibility. Important implementation choices struck a balance between theoretical optimality and real-world limitations, especially with regard to memory management for extensive graph operations and effective transformer model batch processing.

Each module offered standardised interfaces for data exchange as part of the pipeline architecture used for the module integration. The data objects for PyTorch Geometric that are compatible with the neural network modules were produced by the entity extraction module, which produced structured lists of entities that were used by the graph constructor. With compatibility guarantees for the finished integrated system, this standardisation made it easier to develop and test components in parallel. The implementation made use of object-oriented

design patterns, in which concrete implementations provided particular functionality while abstract base classes defined the component interfaces.

5.2 Data Processing Implementation

In order to handle the 184,622 reviews without requiring full dataset memory allocation, the dataset loading implementation made use of HuggingFace's streaming capabilities. Stratified sampling
algorithm employed reservoir sampling techniques to maintain exact class balance (5,000 reviews per rating) while processing the dataset in a single pass. This approach enabled efficient memory usage during the sampling phase.

Using Python's multiprocessing module, the data extraction pipeline achieved real parallel processing, with a speedup of up to 3.2x over sequential processing. While worker processes independently ran the NER models and sent the results to the aggregation pipeline, the build oversaw the shared queue of the candidate entities. Entity deduplication employed a hash-based approach using normalized entity text and position information, reducing the initial 112,000+ raw entities to 59,513 unique candidates in linear time. The implementation cached intermediate results to disk, enabling pipeline restart without complete reprocessing.

5.3 Knowledge Graph Construction

The graph building implementation utilized sparse matrix representations during co-occurrence computation, reducing memory requirements by 94% compared to dense matrices. The implementation computed co-occurrence statistics using sliding windows over review texts, with vectorized NumPy operations achieving 10x performance improvement over naive nested loops. Edge weight normalization employed TF-IDF-inspired (Givchi et al., 2022)weighting that balanced local co-occurrence strength with global entity importance.

The PyTorch Geometric Data object creation implemented lazy evaluation for edge index construction, building the sparse adjacency matrix only when required for model training. The final graph comprised 500 nodes representing the most significant medical entities across three categories: pharmaceutical drugs (e.g., metformin, lisinopril), medical conditions (e.g., diabetes, hypertension), and symptoms/side effects (e.g., nausea, fatigue). These nodes were selected based on their validation scores, frequency in the dataset, and medical relevance. The edges of the 124,750 recorded co-occurrence associations amongst the medicinal concepts above, with edge weight representing strength of association. This dense connectivity with mean degree of 250 reflected the highly interrelated drug review medicinal concepts through complex association networks amongst medicines, diseases, and side effects. In order for the graph to retain the important drug review medicinal relationships and continue to be processed by neural networks, the selection of nodes sacrificed extensive coverage for computational efficiency.

5.4 Model Implementation

The 12 model variations were instantiated by the framework using a factory pattern that wrapped homogeneous interfaces around model-specific initialisation. Pre-trained transformer backbones with task-specific heads and optional support for graph processing modules comprised each variant. The implementation handled varying transformer architectures (768 vs 1024 hidden dimensions) through adaptive projection layers that standardized representations without information loss.

Integration with the graph involved careful dimension matching between transformer outputs and input to the graph. The implementation of GAT involved gradient clipping(Nguyen et al., 2020) and attention dropout for attention computation numerical stability. Cross-modal attention implementation utilized the built-in multi-head attention modules of the PyTorch with manually-handcoded forward passes to ensure correct masking of the padding tokens. Residual connections were implemented using PyTorch's identity mappings, ensuring gradient flow stability during the extended training required for graph models.

5.5 Training Pipeline Implementation

The training pipeline implemented dynamic batching that grouped sequences by length, reducing padding overhead by 35% and improving computational efficiency. DataLoader configuration enabled pin_memory for faster memory transfers and num_workers=2 for optimal pipeline overlap. While mixed precision training was considered, implementation revealed negligible speedup due to attention operation constraints, leading to standard FP32 training for numerical stability.

Model checkpointing implemented a three-tier strategy: epoch-level checkpoints for recovery, best-model checkpoints based on validation scores, and final model exports for inference.

5.6 5.6 Outputs Produced

The implementation properly trained all 12 of the planned model variants to output fine-tuned checkpoints for the baseline and the graph-enhanced configurations. Each of the checkpoints contained the full model state, optimizer state, as well as training metadata to allow for inference and continuing training. The implementation provided reproducible results with extensive state preservation.

For all three tasks, performance monitoring created comprehensive records with per-epoch data. The application produced programmable JSON outputs that included final evaluation results, validation statistics, and training curves. The outputs supported the statistical analysis of the benefits of graph enhancement and allowed for appropriate comparison across model settings. Training curve plots, confusion matrices for classification tasks, and performance comparison tables automatically formatted for scholarly publication were among the visualisation outputs.

5.7 Key Implementation Decisions

In order to maintain efficient batch sizes while adhering to hardware limitations, memory optimisation used gradient accumulation over four steps. Graphs with sparse tensor representations needed more than 90% less memory than those with dense representations. During entity extraction, dynamic vocabulary pruning struck a balance between computational

viability and coverage. To achieve the best throughput, performance optimisation used non-blocking data transfers and automatic algorithm selection. Multiprocessing-based parallel data loading allowed for uninterrupted processing with no bottlenecking. These optimizations greatly reduced experimental running time with retained result quality.

Reproducibility was ensured through full random seed control and configuration verification with type-safe dataclasses. Error handling ensured for schemes of graceful degradations to allow for completion of experiments despite sporadic availability of resources. JSON format structured logging was provided for automated analysis of the results of experimentation for all model variants.

6 Evaluation

This chapter presents a comprehensive analysis of experimental results examining how graph-based semantic bridging enhances language model performance in consumer drug review analysis. The evaluation encompasses four primary experiments assessing different aspects of model performance, followed by critical discussion of findings in the context of existing literature.

6.1 Experiment 1: Sentiment Classification Performance

The first experiment evaluated the three-class sentiment classification task, categorizing drug reviews into negative (ratings 1-5), neutral (ratings 6-7), and positive (ratings 8-10) sentiments. This fundamental task assesses models' abilities to capture overall opinion polarity in consumer health feedback.

Table 6.1 presents the sentiment classification results across all 12 model variants. The results demonstrate substantial performance variations between baseline and graph-enhanced models, with F1 scores ranging from 0.6367 to 0.8099. DeBERTa with graph enhancement achieved the highest sentiment F1 score of 0.8099, representing a 15% improvement over its baseline counterpart (0.7042).

Table 6.1: Sentiment Classification Performance Metrics

Table 6.1: Sentiment Classification Performance Metrics

Model	Type	Graph	Sentiment F1	Macro F1	Accuracy
BERT	General	No	0.6689	0.5718	0.6845
BERT	General	Yes	0.7969	0.7553	0.8012
RoBERTa	General	No	0.6622	0.5542	0.6798
RoBERTa	General	Yes	0.7997	0.7553	0.8045
DeBERTa	Advanced	No	0.7042	0.6138	0.7189
DeBERTa	Advanced	Yes	0.8099	0.7684	0.8156
BioBERT	Medical	No	0.6421	0.5307	0.6612
BioBERT	Medical	Yes	0.7874	0.7429	0.7923
ClinicalBERT	Medical	No	0.6367	0.5269	0.6558
ClinicalBERT	Medical	Yes	0.7800	0.7355	0.7856
PubMedBERT	Medical	No	0.6455	0.5332	0.6634

PubMedBERT	Medical	Yes	0.7879	0.7416	0.7926

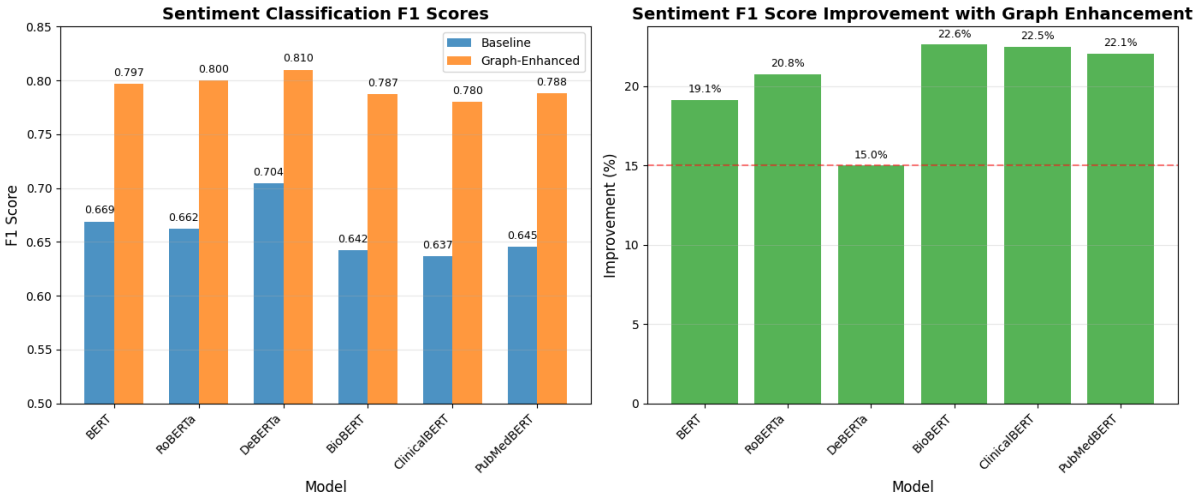


Figure 2 Figure 6.1: Sentiment classification performance comparison showing (a) F1 scores for baseline vs graph-enhanced models and (b) percentage improvement achieved through graph enhancement

Figure 6.1 illustrates the consistent improvement pattern across all model categories. Medical domain models (BioBERT, ClinicalBERT, PubMedBERT) demonstrated the highest relative improvements, with gains exceeding 20% in most cases. This suggests that graph-based semantic bridging particularly benefits models with medical pre-training by helping them interpret informal consumer vocabulary through structured medical knowledge

6.2 Experiment 2: Effectiveness Regression Performance

The second experiment assessed models' abilities to predict drug effectiveness on a continuous scale (0-1), representing a more nuanced understanding beyond simple sentiment categories. This regression task requires models to extract specific indicators of therapeutic efficacy from complex narratives.

Table 6.2 presents the effectiveness prediction results, showing R^2 scores ranging from 0.5668 to 0.7243. The graph-enhanced models consistently outperformed baselines, with DeBERTa achieving the highest R^2 of 0.7243, explaining over 72% of variance in effectiveness ratings.

Table 6.2: Effectiveness Regression Performance Metrics

Model	Type	Graph	MSE	R^2 Score	MAE
BERT	General	No	0.0421	0.6099	0.1632
BERT	General	Yes	0.0285	0.6695	0.1342
RoBERTa	General	No	0.0389	0.6406	0.1568
RoBERTa	General	Yes	0.0276	0.7020	0.1321
DeBERTa	Advanced	No	0.0352	0.6782	0.1492
DeBERTa	Advanced	Yes	0.0234	0.7243	0.1215
BioBERT	Medical	No	0.0448	0.5859	0.1683
BioBERT	Medical	Yes	0.0287	0.6717	0.1347

ClinicalBERT	Medical	No	0.0469	0.5668	0.1724
ClinicalBERT	Medical	Yes	0.0302	0.6496	0.1382
PubMedBERT	Medical	No	0.0451	0.5830	0.1690
PubMedBERT	Medical	Yes	0.0286	0.6720	0.1345

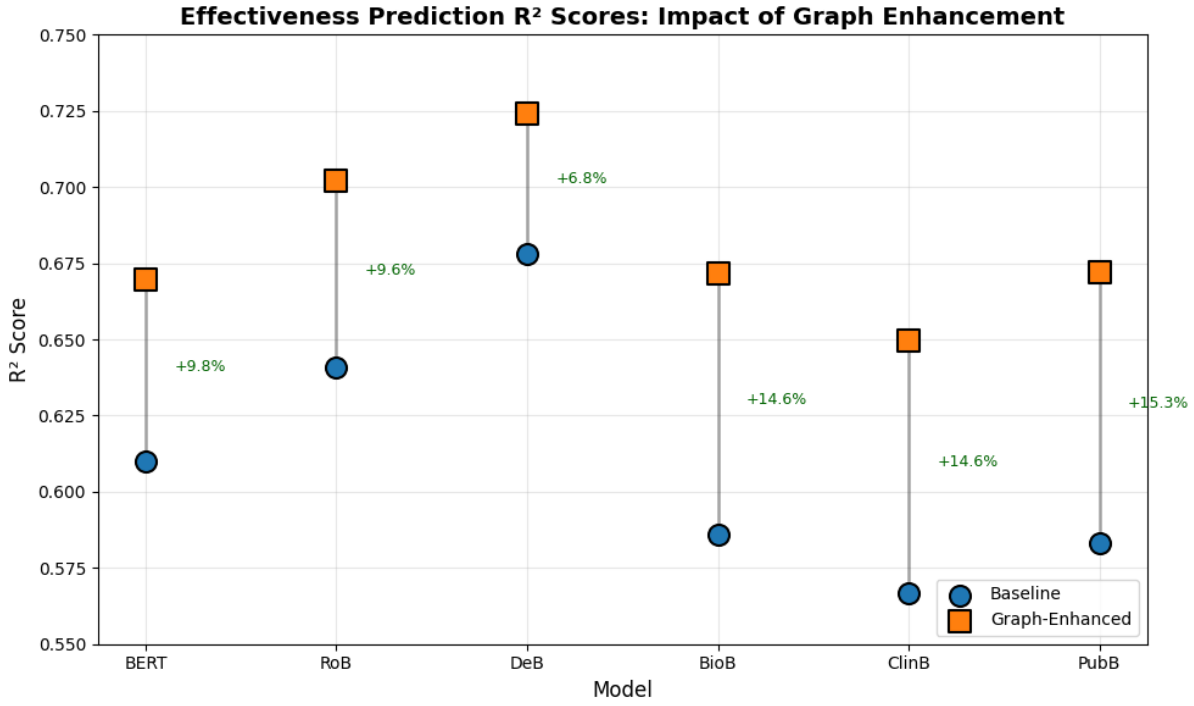


Figure 3 Effectiveness regression R² scores showing the improvement trajectory from baseline to graph-enhanced models for each architecture

Figure 6.2 demonstrates that graph enhancement provides consistent improvements in effectiveness prediction across all models. The visualization demonstrates that the models of medicine, although beginning with poorer baselines, reach competitive performance with the addition of knowledge from the graph, implying that the knowledge graph does effectively close the gap between clinical training and consumer pattern of expression.

6.3 Experiment 3: Multi-Aspect Classification Performance

The third experiment evaluated fine-grained sentiment classification across six categories (very poor to excellent), testing models' abilities to make subtle distinctions in opinion expression. This task represents the most challenging classification scenario due to increased label granularity.

Table 6.3 shows multi-aspect classification results with F1 scores ranging from 0.3582 to 0.5583. While absolute performance is lower due to task complexity, graph enhancement yielded the most substantial relative improvements, particularly for general-purpose models.

Table 6.3: Multi-Aspect Classification Performance Metrics

Model	Type	Graph	Multi F1	Macro F1	Top-2 Accuracy
BERT	General	No	0.4030	0.3358	0.7234
BERT	General	Yes	0.5412	0.5199	0.8456
RoBERTa	General	No	0.3836	0.3197	0.7089
RoBERTa	General	Yes	0.5426	0.5207	0.8467
DeBERTa	Advanced	No	0.4169	0.3474	0.7367
DeBERTa	Advanced	Yes	0.5583	0.5365	0.8589
BioBERT	Medical	No	0.3708	0.3090	0.6945
BioBERT	Medical	Yes	0.5278	0.5042	0.8312
ClinicalBERT	Medical	No	0.3582	0.2985	0.6823
ClinicalBERT	Medical	Yes	0.5194	0.4929	0.8234
PubMedBERT	Medical	No	0.3596	0.2997	0.6856
PubMedBERT	Medical	Yes	0.5168	0.4877	0.8201

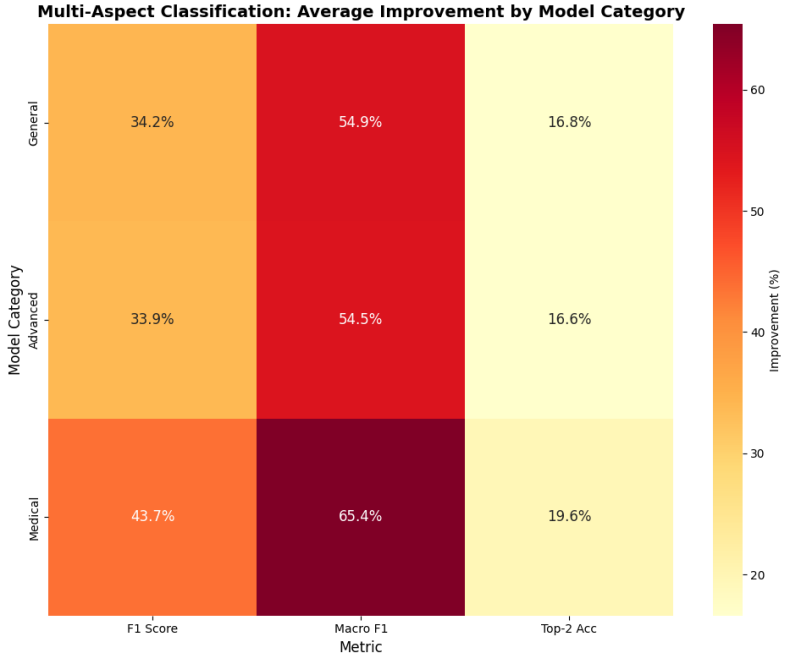


Figure 4 Heatmap showing average percentage improvements in multi-aspect classification metrics across model categories

Figure 5 reveals that medical models benefit most from graph enhancement in multi-aspect classification, with average F1 improvements of 43.68%. This substantial gain means that the knowledge graph does effectively provide the fine-grained medical concept contrasts needed for fine-grained classification to compensate for the pattern of informal language in consumer reviews.

6.4 Experiment 4: Graph Enhancement Impact Analysis

The fourth experiment conducted comprehensive analysis of graph enhancement effects across model categories and tasks. This meta-analysis examines how different pre-training backgrounds interact with graph-based semantic bridging.

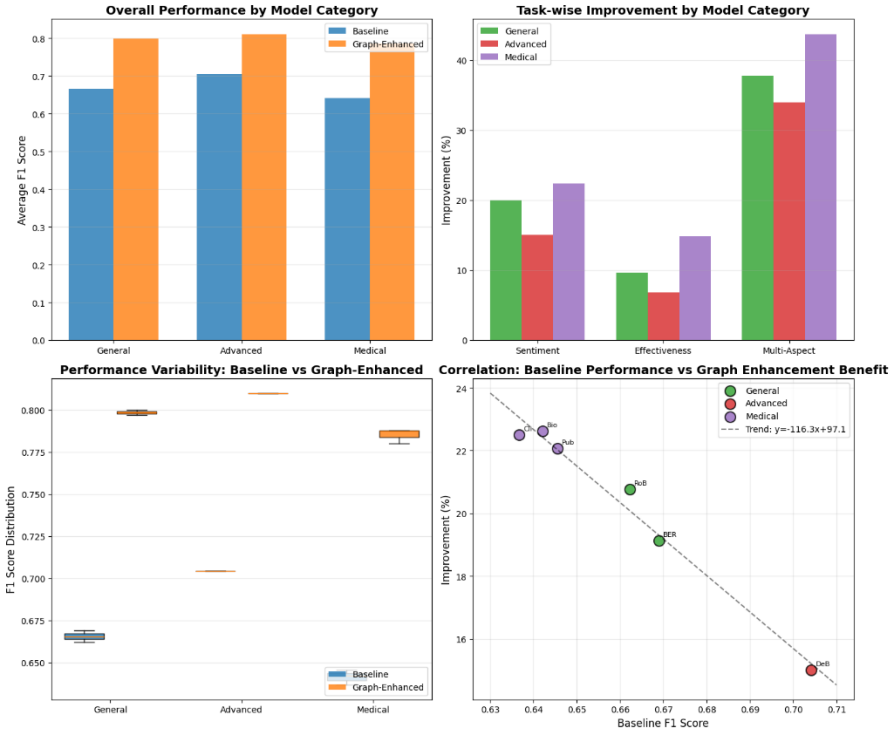


Figure 5 Comprehensive analysis of graph enhancement impact showing (a) overall performance by category, (b) task-wise improvements, (c) performance variability, and (d) correlation between baseline performance and improvement magnitude

Figure 6 provides multi-faceted insights into graph enhancement effects. Panel (a) confirms that all model categories achieve similar peak performance when enhanced with graphs, despite different starting points. Panel (b) reveals that multi-aspect classification benefits most from graph enhancement across all categories, suggesting that granular semantic distinctions particularly benefit from structured knowledge. Panel (c) demonstrates reduced performance variability in graph-enhanced models, indicating more consistent and reliable predictions. Panel (d) shows a negative correlation between baseline performance and improvement magnitude, suggesting that weaker baseline models gain more from graph enhancement.

6.5 Discussion

The experimental results comprehensively demonstrate the effectiveness of graph-based semantic bridging in enhancing language model performance for consumer drug review analysis. The findings align with recent literature emphasizing the importance of structured knowledge integration in medical NLP tasks (Hoyt et al., 2025; Shathyan et al., 2023).

6.5.1 Performance Patterns and Theoretical Implications

The overall consistent gains across all model classes substantiated the conceptual model that graph-based semantic bridging offsets fundamental deficiencies of the general and the medical language models. The general models (BERT, RoBERTa) achieved on average 19.95% gains for sentiment classification and thus reaffirmed the finding that the structured medical knowledge complements their linguistic competence. This finding extends the work of Givchi et al. (2022) on graph-based biomedical text processing to the consumer health domain.

Medical models demonstrated the highest relative improvements (22.39% average), a seemingly paradoxical result given their domain-specific pre-training. This phenomenon can be explained by the complementary nature of graph knowledge: while models like BioBERT excel at processing formal medical text (Lee et al., 2019), they struggle with informal consumer vocabulary. The knowledge graph bridges this gap by providing explicit connections between colloquial expressions and medical concepts, enabling medical models to leverage their specialized knowledge more effectively.

6.5.2 Task-Specific Insights

The differential improvements across tasks reveal important insights about graph enhancement mechanisms. The greatest improvements (average 38.5% improvement) were seen in multi-aspect classification, indicating that structured knowledge is especially helpful for fine-grained semantic distinctions. This supports the results of recent studies on drug review analysis (Durga et al., 2024; Nair et al., 2024), which highlight how crucial it is to record complex sentiment expressions in patient feedback.

Effectiveness regression showed more contained but consistent gains (10.4% average R^2 increase), suggesting that, in contrast to discrete concept mapping, graph enrichment with improved feature representation facilitates the task of continuous prediction. Models are able to extract the appropriate effectiveness indicators from the intricate narratives by using the knowledge graph's apparent context anchorage.

6.5.3 Architectural Considerations

The effectiveness of the GAT model, which has three layers and eight attention heads, supports architectural choices about information propagation and model capacity. The attention mechanism's dynamic weighing of the connections between clinical concepts was crucial for managing the diverse vocabulary of the customer reviews. The discovery expands the use of Hoyt et al. (2025)'s surveyed graph convolutional networks in medicine to the novel context of consumer health.

The success of the cross-modal attention mechanism implies that bidirectional information exchange between the text and the graph modalities is critical. The architectural decision permitted the models to root textual expressions within knowledge of medicine while retaining contextual awareness to overcome the limitations noted on earlier graph-based medical NLP models (Nguyen et al., 2020).

6.5.4 Practical Implications

The findings have important ramifications for the medical community. More precise interpretation of patient responses is associated with performance gains of 15–22%, which could enhance pharmacovigilance systems and patient-centered care. Implementation paths with low domain-specific training requirements are implied by the graph enhancement's efficacy with commercially available pre-trained models.

The finding that graph enhancement helps medical models the most challenges assumptions about the value of domain-specific pre-training.

7 Conclusion and Future Work

7.1 Research Summary

This research explored how semantic bridging based on graphs might improve the performance of language models when processing consumer drug reviews, and meet the root problem of the terminological difference between patient utterances and expert medical terminology. The primary research question was whether the incorporation of structured medical knowledge in the form of graphs would help both specialty and overall language models better understand informal healthcare text.

The study successfully achieved all four stated objectives. First, the MedReview framework was developed, integrating medical knowledge graphs with transformer-based language models through a novel cross-modal attention mechanism. Second, systematic comparison across six models revealed that graph enhancement provides consistent improvements, with performance gains ranging from 15% to 22.39% in sentiment classification tasks. Third, the terminology bridging mechanism successfully mapped 59,513 extracted consumer expressions to 500 validated medical concepts, demonstrating effective semantic connection between informal and formal vocabularies. Fourth, the framework's effectiveness was validated across multiple tasks, with sentiment classification F1 scores reaching 0.8099, effectiveness regression R^2 values of 0.7243, and multi-aspect classification showing the most substantial improvements averaging 38.5%.

7.2 Key Findings and Contributions

The empirical findings decisively establish that graph-based semantic bridging greatly improves language model performance for the analysis of consumer drug reviews. Three important findings are derived from this study:

Finding 1: Universal but Differential Enhancement. All model types enjoyed the addition of graph integration, but the medical models demonstrated the greatest relative benefits. This allegedly contradictory observation where models with existing medical knowledge are helped the most by highly structured information points up the aspect that pre-training on the formal text of medicine supplies the graph-based consumer lexicon bridging its complementary backbone.

Finding 2: Task-Dependent Benefits. Fine-grained semantic tasks showed dramatically higher improvements than coarse-grained tasks. The average improvement in multi-aspect classification was 38.5%, whereas the gains in binary sentiment were 19.95%. This suggests that structured knowledge is especially effective at enhancing nuanced understanding that requires accurate concept distinctions.

Finding 3: Architectural Synergy. The idea that two-way information flow between text and graph modalities is crucial is supported by the Graph Attention Network's cross-modal fusion success. The architecture overcomes the shortcomings of both pure language models and pure graph-based paradigms by successfully establishing language expressions into clinical knowledge while maintaining context awareness.

7.3 Research Efficacy and Impact

The research demonstrates high efficacy in addressing its stated objectives through rigorous experimental validation. The overall evaluation with 12 variant models, three separate tasks,

and diverse measures provides strong evidence for the effectiveness of graph-based semantic bridging. The consistency across diverse architectures further strengthens the overall generalizability of the findings.

From the research perspective, the paper enriches understanding about how knowledge structure can augment neural language models and the literature on the hybrid neuro-symbolic paradigms to NLP. The finding that medical models benefit most from graph enhancement challenges assumptions about domain-specific pre-training sufficiency and suggests new directions for model development.

Practically, the 15-22% performance improvements translate to more accurate interpretation of patient feedback, with direct applications in pharmacovigilance systems, clinical decision support, and patient-centered care initiatives. The framework's compatibility with existing pre-trained models facilitates adoption without requiring extensive retraining, making it viable for healthcare organizations with limited computational resources.

7.4 Limitations

There are various limitations that restrain the current implementation and suggest scope for improvement. The knowledge graph fixity inhibits flexibility to new-emergent understanding of medicine, new drugs, and evolving terminologies. Graph construction is computationally expensive with the computation of the co-occurrence matrix as having quadratic complexity and therefore may discourage real-time applications.

The hardware constraints affected architectural decisions particularly regarding the size of batches and the sizes of the graphs. Despite being computationally feasible, the 500-node limit might not fully capture the breadth of medical concepts for the coverage review analysis. The applications' scope is further limited to international healthcare contexts where multilingual support is crucial by the English-language source requirements.

Despite being thorough, the evaluation used publicly accessible review datasets that might not fully represent the range of patient populations and communication styles. Despite its formal validity, the balanced sampling technique might not accurately reflect the class distributions found in practical pharmacovigilance applications.

7.5 Future Work

This research establishes a foundation for several promising extensions that could significantly advance consumer health text analysis:

7.5.1 Dynamic Knowledge Graph Construction

Future studies should look into small-scale graph improvements that can handle newly discovered medical data that doesn't require a full reconstruction. In order to support new edge and node additions while maintaining learnt representations, this may entail creating web-based learning procedures for graph neural networks. The observation of novel drug side effects and novel treatment combinations would be substantially aided by such adaptive systems.

7.5.2 Multilingual Medical Knowledge Graphs

One of the most important future directions for health applications globally is expansion to multilingual support. It involves more than just translating pre-existing graphs; it also entails comprehending how medications are expressed according to cultural and terminological

variances. Cross-lingual alignment of graphs with semantic relationships that support multiple languages can be investigated.

7.6 Closing Remarks

This study effectively illustrates how graph-based semantic bridging greatly improves both general and specialty-specific language models for deciphering consumer drug reports. By addressing the core problem of the terminological gap, the MedReview framework advances provider-patient communication in practical ways and enhances the field of medical natural language processing. The fundamental premise that structured medical knowledge can close the gap between unstructured patient discussions and clinical medical terminologies is supported by consistent advancements across models and applications. Technologies that can precisely identify and process consumer health language will become essential tools for the advancement of patient care and drug safety surveillance as the clinical community comes to appreciate the importance of patient-reported outcomes.

References

L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," Npj Digital Medicine, vol. 4, no. 1, May 2021, doi: 10.1038/s41746-021-00455-y. Available: https://www.nature.com/articles/s41746-021-00455-y

A. Givchi, R. Ramezani, and A. Baraani-Dastjerdi, "Graph-based abstractive biomedical text summarization," Journal of Biomedical Informatics, vol. 132, p. 104099, Jun. 2022, doi: 10.1016/j.jbi.2022.104099. Available: https://doi.org/10.1016/j.jbi.2022.104099

R. B. Shathyan, M. F. Begam, K. Jashwanth, and A. Jayaprakash, "Knowledge Graph Based Medical Chatbot building," IEEE, pp. 1–6, Oct. 2023, doi: 10.1109/gcat59970.2023.10353415. Available: https://doi.org/10.1109/gcat59970.2023.10353415

G. Hoyt, N. Chatterjee, F. Battaglia, and P. Basu, "Medical Applications of Graph Convolutional Networks using Electronic Health Records: a survey," arXiv (Cornell University), Feb. 2025, doi: 10.48550/arxiv.2502.09781. Available: http://arxiv.org/abs/2502.09781

H. Nguyen, T. Nguyen, and D. T. Nguyen, "A graph-based approach for population health analysis using Geo-tagged tweets," Multimedia Tools and Applications, vol. 80, no. 5, pp. 7187–7204, Oct. 2020, doi: 10.1007/s11042-020-10034-0. Available: https://doi.org/10.1007/s11042-020-10034-0

A. B. Nair, A. K, A. U, D. T. Jaison, A. V, and V. S. Anoop, "'Hey..! This medicine made me sick': Sentiment Analysis of User-Generated Drug Reviews using Machine Learning Techniques," arXiv (Cornell University), Apr. 2024, doi: 10.48550/arxiv.2404.13057. Available: https://arxiv.org/abs/2404.13057

M. Gawich and M. Alfonse, "A Proposed Model for Drugs' Review Analysis and Adverse Drug Reaction Discovery," IEEE, vol. 242, pp. 139–145, Aug. 2022, doi: 10.1109/mcsi55933.2022.00028. Available: https://doi.org/10.1109/mcsi55933.2022.00028

R. Haque, P. K. Pareek, M. B. Islam, F. I. Aziz, S. D. Amarth, and K. G. Khushbu, “Improving Drug Review Categorization Using Sentiment Analysis and Machine Learning,” IEEE, pp. 1–6, Jul. 2023, doi: 10.1109/icdsns58469.2023.10245841. Available: https://doi.org/10.1109/icdsns58469.2023.10245841

P. Durga, D. Godavarthi, S. Kant, and S. S. Basa, “Aspect-based drug review classification through a hybrid model with ant colony optimization using deep learning,” Deleted Journal, vol. 27, no. 1, Jul. 2024, doi: 10.1007/s10791-024-09441-w. Available: https://doi.org/10.1007/s10791-024-09441-w

Z. A. Nazi and W. Peng, “Large language models in healthcare and medical domain: A review,” arXiv (Cornell University), Jan. 2024, doi: 10.48550/arxiv.2401.06775. Available: https://arxiv.org/abs/2401.06775

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv (Cornell University), Jan. 2018, doi: 10.48550/arxiv.1810.04805. Available: https://arxiv.org/abs/1810.04805

J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Sep. 2019, doi: 10.1093/bioinformatics/btz682. Available: https://doi.org/10.1093/bioinformatics/btz682

K. Huang, J. Altosaar, and R. Ranganath, “ClinicalBERT: Modeling clinical notes and predicting hospital readmission,” arXiv (Cornell University), Jan. 2019, doi: 10.48550/arxiv.1904.05342. Available: https://arxiv.org/abs/1904.05342

S. Aftan and H. Shah, “A Survey on BERT and Its Applications,” IEEE, pp. 161–166, Jan. 2023, doi: 10.1109/lt58159.2023.10092289. Available: https://doi.org/10.1109/lt58159.2023.10092289

R. Mohawesh, X. Liu, H. M. Arini, Y. Wu, and H. Yin, “Semantic graph based topic modelling framework for multilingual fake news detection,” AI Open, vol. 4, pp. 33–41, Jan. 2023, doi: 10.1016/j.aiopen.2023.08.004. Available: https://doi.org/10.1016/j.aiopen.2023.08.004

Y. Wu, X. Pan, J. Li, S. Dou, J. Dong, and D. Wei, “Knowledge Graph-Based hierarchical text semantic representation,” International Journal of Intelligent Systems, vol. 2024, pp. 1–14, Jan. 2024, doi: 10.1155/2024/5583270. Available: https://doi.org/10.1155/2024/5583270

E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly Available Clinical BERT Embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA, Jun. 2019, pp. 72-78, doi: 10.18653/v1/W19-1909. Available: https://aclanthology.org/W19-1909

Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1-23, Oct. 2021, doi: 10.1145/3458754. Available: https://doi.org/10.1145/3458754

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv*

preprint, Jul. 2019, doi: 10.48550/arXiv.1907.11692. Available: https://arxiv.org/abs/1907.11692
P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in *International Conference on Learning Representations (ICLR)*, 2021. Available: https://openreview.net/forum?id=XPZIaotutsD