An efficient Biomedical keyphrase extraction model on large biomedical databases

A V Srinivas
Research Scholar, Department of Computer science & Engineering,
Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
Dr. O Nagaraju,
Research supervisor
Department of Computer science & Engineering, Acharya Nagarjuna University,
Associate Professor,
APRDC,
Nagarjunasagar -522439
.

Abstract

The amount of medical and clinical texts, derived from research articles, hospital records and trial reports, has grown at an unprecedented pace, posing ever more challenges for the discovery of useful concepts. Extracting medical relevant terms and inferring their relation to the surrounding clinical context is a non-trivial task when considering diverse vocabularies, domain specific abbreviations and irregular sentence structures. In general, conventional distributional embedding methods may not perform well in this context, especially need to capture more fine-grained aspect-level signals from domain term sets. In order to address above mentioned issues, a context-aware keyphrase extraction methodology based on a novel adaptation of the global co-occurrence representation approach suited to massive medical text datasheets is proposed. By means of a fine-tuned parameter calibration and a tailored optimization procedure, the embedding goal is changed so that both the principal term vectors and their contextual versions capture richer semantic relations. Also, two properties based contextual scoring schemes are developed to measure and rank the relational tightness between extracted phrases and the medical narrative they reside. Extensive experiments on several biomedical datasets demonstrate the efficiency of the framework. The method attains an overall accuracy of 97%; improvements in precision and recall are very significant, along with significant decreased error rate in comparison with state-of-the-art techniques for processing biomedical texts. These results suggest the proposed model provides a reliable and meta scalable framework for high quality keyphrase extraction as well as contextual similarity matching from large scale biomedical archives.

Keywords: Biomedical , key phrase extraction, embedding model.

1.Introduction

Biosystems information is an active archiving system e.g., the electronic health record (EHR), laboratory-tumor tissue [6] and clinical trial documentationclinical trials documents and scientific literature. They aggregate diverse data – physician notes, diagnostic reports, image reports, experimental outcomes – into a base for clinical research and drug development. Physicians, statisticians, and health care professionals rely on both these structured and unstructured records to identify trends, assess interventions, and facilitate evidence-based decision making in the diagnosis and treatment of disease.

Accordingly, within these contexts, the automatic detection of important medical concepts and affective analysis of narrative text-based reports has gained popularity. Keyphrase extraction will extract domain relevant concepts which will help in understanding the thematic content of the document, while sentiment-based text mining will mine the subjective language in clinical narratives, treatment reviews and outcome reports. Sentiment-oriented models analyze the sentiment of textual opinions related to therapies, drug efficacy, or individual patient relevant experiences and thus offers a more structured view of medical interventions over different sources such as research papers and clinical evaluations[1].

Extraction of medically relevant phrases is the key for effective retrieval, indexing and semantic based interpretation of biomedical information. Disease names, treatment methods, biomarkers, and procedure terms often hold most or all of the relevant information in a document. The precise identification of such components enables more structured knowledge graphs, more scalable literature mining, and better thematically clustering the medical evidence. In more contextual interpretation, such an extraction enables to conduct higher-level analyses of dominant research and clinical perspectives in the life science field.Sentiment analysis in medical text is much more than just polarity classification. Clinical narratives tend to using hedged, conditional, or context-dependent expressions. A seemingly neutral utterance can implicitly convey therapeutic uncertainty or mild effectiveness. Detecting such nuances requires models that can differentiate finer granularity of evaluative signals rather than a simple positive–negative labeling. This nuanced understanding enables better prediction of models and informed healthcare analytics.As well as for text mining purposes, computational modeling is also instrumental within drug development. Due to heavy reliance on iterative experimental screening process, drug discovery is time-consuming and costly. Probabilistic prediction frameworks, such as Bayesian inference methods, offer a principled means of predicting compound activity and the potential for drug-ness. By incorporating prior information in conjunction with observed molecular descriptors, these models predict the likelihood of being biologically active with enhanced statistical confidence.Structured molecular data for sample gene inhibitors are useful datasets for learning predictive models. Machine learning models trained on these datasets aid in the prediction of active compounds, thereby reducing the amount of wasteful experimental cycles. They also allow to investigate the relationships between chemical descriptors and biological activities at the mol- ecular level, explaining or describing inhibitory mechanisms. Statistically driven screening and prediction modeling computational methods provide support for the design of targeted therapeutics for diseases resulting from gene malfunction, enhancing translational research and drug design[2-7].

2.Related works

The dramatic expansion of biomedical literature, clinical documentation, and epidemiological reports has created a pressing demand for automated disease-focused text analysis. During public health emergencies, rapid interpretation of large-scale textual data becomes essential for identifying emerging disease patterns, treatment responses, and adverse clinical outcomes. Disease names, symptom clusters, therapeutic interventions, and biological markers are embedded within unstructured narratives, requiring intelligent systems capable of extracting and ranking medically relevant phrases in real time.Biomedical language differs substantially from general-domain text. It contains domain-specific terminology, multi-word technical expressions, abbreviations, and nested entities. A single disease may appear in various lexical forms, and treatment references may include chemical names, drug brands, or molecular descriptors. These characteristics introduce semantic ambiguity and structural complexity, limiting the effectiveness of simple keyword-based or frequency-driven extraction techniques.

To address these challenges, modern frameworks integrate distributed word representations with contextual sequence modeling. Static embeddings derived from global co-occurrence statistics encode semantic similarity across the entire corpus, while recurrent neural architectures capture local contextual dependencies. Bidirectional gated recurrent units (Bi-GRU) are frequently employed because they process sequences in both forward and backward directions, enabling the interpretation of disease mentions within surrounding clinical qualifiers[8].

Mathematical Formulation

Let the biomedical dataset be represented as:

$D = \{I_1, I_2, \ldots, I_n\}$

where each document I_i contains biomedical interactions such as clinical summaries, research abstracts, or patient narratives.

For each document I_i, define a candidate key-phrase set:

$K_i = \{k_{i1}, k_{i2}, \ldots, k_{im}\}$

The objective is to construct a mapping function:

$\Phi : D \rightarrow K$

that extracts and ranks disease-related phrases efficiently during large-scale data processing.

Embedding and Contextual Modeling

Let:

- I_b denote biomedical text sequences
- E_glove $\in \mathbb{R}^d$ denote static global embeddings
- GRU_bi denote the bidirectional GRU model

Each token is first mapped into a d-dimensional embedding space using E_glove. The bidirectional GRU processes the sequence and produces forward and backward hidden states:

$h\_t = [\rightarrow h\_t ; \leftarrow h\_t]$

An attention mechanism assigns weights α_t to emphasize tokens that contribute strongly to disease identification:

$H = \Sigma (\alpha\_t \cdot h\_t)$

This weighted representation summarizes document-level biomedical relevance.

Computational Complexity

If:

- T = number of time steps
- d = embedding dimension
- |V| = vocabulary size

The computational cost of the contextual model can be approximated as:

$C\_comp = O(T \cdot (d^2 + d \cdot |V|))$

The d^2 term corresponds to recurrent transformations within hidden states, while $d \cdot |V|$ accounts for embedding-related operations. As biomedical vocabularies are large and highly specialized, computational efficiency becomes a central design constraint.

Hybrid Representation Strategy

To balance contextual precision and computational feasibility:

1. Static embeddings encode global biomedical semantics.
2. Dynamic recurrent layers capture document-specific context.
3. Selective attention reduces noise from irrelevant tokens.
4. Partial parameter freezing limits overfitting risk.

This hybrid approach preserves semantic richness while maintaining scalability for large datasets.

Sentiment-Aware Disease Analysis

Disease mentions often carry evaluative context in patient reports or clinical discussions. To incorporate sentiment orientation, define:

$S(k_{ij}) \in \{-1, 0, 1\}$

where −1 = negative, 0 = neutral, 1 = positive sentiment.

Let $R(k_{ij})$ represent contextual relevance derived from embeddings and attention scores. A composite ranking score may be defined as:

$\Gamma(k_{ij}) = \lambda_1 R(k_{ij}) + \lambda_2 S(k_{ij})$

where $\lambda_1 + \lambda_2 = 1$.

This formulation integrates semantic centrality and evaluative orientation, enabling refined prioritization of disease-related key phrases.

Real-Time Biomedical Monitoring

During outbreaks or health crises, systems must process streaming biomedical content continuously. Incremental learning, mini-batch updates, and distributed architectures allow models to adapt without retraining from scratch. Efficient embedding reuse and attention pruning further reduce latency[9].

Biomedical disease key-phrase extraction requires a coordinated integration of semantic representation, contextual modeling, probabilistic ranking, and computational optimization. Bidirectional recurrent models enriched with global embeddings provide strong contextual encoding, but must be carefully regulated to avoid excessive complexity. By combining static and dynamic embeddings, attention-driven weighting, and sentiment-aware scoring, large biomedical corpora can be transformed into structured, actionable intelligence that supports timely and data-driven public health decisions.

In biomedical documents, aspect-oriented phrase identification is mostly performed on the basis of a hybrid representation which concatenates distributed word vectors with subword or character-level information. Medical language is diverse across specialties and languages and the way it is represented varies in different types of documents. A cardiology report, an oncology clinical trial abstract, and an infectious disease alert can discuss common biological processes while employing wildly divergent lexical and syntactic structures. Representing them requires models that can identify morphological variations, acronyms, and concatenated medical terms, and maintain semantic integrity in and across application domains[10].

Simultaneously, interpreting sentiment in discussions of health on social SJWs adds another level of complexity. Embedding techniques based on graphs are widely used to model the relationships between users, symptoms, treatments, and opinions. These methods tend to model interactions well but they are often not

robust on explanation. It is difficult to understand why a model gives a particular polarity or influence score to a patient statement, especially when that statement concerns emotionally charged feedback on disease progression or therapy results.

Classification mechanisms based on Tolerance Near Sets and vector similarity measures are further evidence of the intense computational load related to bio-medical analysis. These classifiers are trained in high dimensional feature space generated by dense embeddings, which results in more memory usage and training time when running on large scale medical texts. With the increasing diversity of dataset (e.g. including clinical narratives, laboratory findings and research articles), the learning process need to handle heterogeneous linguistic pattern, which make the optimization much more complicated.

The speed of sentiment analysis in bio domain, especially in case of short patient narratives or on-line health surveys, is also an issue on which to focus when setting the parameters. Heuristic modifications, threshold adjustments, and features pruning are usually needed to preserve responsiveness while still having a dependable indicator of predictiveness. Simplistic rule-based acceleration is not effective for medical language, which often contains minor qualifiers, conditional phrases, and expressions of uncertainty.

A further complication arises when using subword-aware embedding models such as FastText for disease-related phrase ranking: the high variability of biomedical vocabulary. Technical prefixes and suffixes, and Latin–Greek hybrid terms, cause fjords of term combinations. To extract well, such systems must be able to tell meaningful clinical variations from trivial morphological changes. In addition, biomedical data sets come from a variety of sources including electronic health records, clinical trials, academic journals and public health alerts, and each has its own unique structure and style that the entity taggers need to be trained on.

Short-form medical texts add even more limitations. Limited token counts diminish contextual cues and the class imbalance between mentions of common and rare diseases can bias the learning dynamics. Sentiment and keyphrase models based on embeddings need to adjust for weak signals and still be able to differentiate between the main disease radial and peripheral modifiers.

All these factors show that keyphrase extraction and ranking in the case of biomedical text goes well beyond a simple feature matching. State-of-the-art domain specific neural architectures, on the other hand, are not immediately scalable to millions of documents. High-dimensionality embeddings, sophisticated attentions and graph-based relational modeling may also raise the computational overhead and thus impede the real time applications.

Semantic interpretation is challenging to be consistently maintained across biomedical data types due to the continuous variation in terminology, document layout, and contextual background within each type. Achieving the contextual unity between heavily structured clinical documentation and unstructured patient-generated content remains the key obstacle. In addition to these demands, the ability to analyze data in real time during public health incidents places additional strain on the ageing infrastructures[11].

Advances in this direction might require finding ways to "trade down" model complexity without sacrificing semantic richness. Techniques integrating lightweight contextual encoders, adaptive embedding compression and interpretable similarity metrics are potential lines of investigations. Enhancing contextual inference across heterogeneous biomedical sources, along with real-time scalability, will play a pivotal role in developing robust and high performance disease-centric keyphrase identification and ranking systems[12].

3.Biomedical Disease Keyphrase Extraction

Phase 1: Enhanced GloVe Algorithm for Biomedical Disease Keyphrases

Objective

Improve semantic representation of biomedical disease-related terms by integrating character n-grams and optimizing embedding parameters for multilingual medical corpora.

Algorithm 1: Enhanced Biomedical GloVe with n-gram Integration

Input:

- Biomedical corpus □
- Vocabulary V
- Context window size w
- Embedding dimension d
- Learning rate η
- Number of iterations N

Output:

- Optimized main vectors W
- Optimized context vectors C

Step 1: Construct Co-occurrence Matrix

1. For each document in □:
    - Slide context window of size w
    - Update co-occurrence count X_{ij} between word i and j
2. Apply weighting function:

$f(X_{ij}) = (X_{ij} / X\_max)^{\alpha}$ if $X_{ij} < X\_max$
$f(X_{ij}) = 1$ otherwise

Step 2: Incorporate Character n-grams

1. For each biomedical keyphrase k:
    - Extract character n-grams G(k)
    - Represent phrase vector as:

$v\_k = \Sigma\ g \in G(k)\ z\_g$

2. where z_g is embedding of n-gram g
3. Replace traditional word-only embedding with hybrid representation.

Step 3: Extended Objective Function

Minimize cost function:

$J = \Sigma_{ij}\ f(X_{ij})\ (w_i^T\ c_j + b_i + b_j - \log X_{ij})^2$

Where:

- w_i = main vector
- c_j = context vector
- b_i, b_j = bias terms

Step 4: Hyperparameter Optimization

1. Compute gradient:

   $\partial J/\partial w_i = \Sigma_j f(X_{ij})(\text{prediction} - \log X_{ij})\, c_j$

2. Update parameters:

   $w_i \leftarrow w_i - \eta\, \partial J/\partial w_i$
   $c_j \leftarrow c_j - \eta\, \partial J/\partial c_j$

3. Repeat for N iterations until convergence.

Phase 2: Advanced Keyphrase Score Computation

Objective

Rank biomedical disease keyphrases based on semantic strength, frequency, and contextual probability.

Algorithm 2: Biomedical Keyphrase Scoring

Input:

- Biomedical corpus □
- Optimized embeddings W, C

Output:

- Ranked keyphrase list K*

Step 1: Data Preprocessing

1. Remove:
   - URLs
   - Stopwords
   - Non-medical tokens
2. Normalize:
   - Lowercasing
   - Lemmatization
   - Abbreviation expansion
3. Extract candidate keyphrases:
   - Diseases
   - Symptoms
   - Treatments

Step 2: Generate GloVe Features

For each keyphrase k:

1. Retrieve:
   - Main vector v_main(k)
   - Context vector v_context(k)

2. Construct hybrid feature:

   F(k) = [v_main(k) ; v_context(k)]

Step 3: Compute Keyphrase Score
Define:

- P(k) = probability of occurrence
- Freq(k) = normalized frequency
- Sim(k) = contextual similarity

Keyphrase score:

$Score(k) = \lambda_1 P(k) + \lambda_2 Freq(k) + \lambda_3 Sim(k)$

Where:

$\lambda_1 + \lambda_2 + \lambda_3 = 1$

Sort phrases in descending order of Score(k).

Phase 3: Contextual Similarity Assessment for Biomedical Texts

Objective
Quantify semantic closeness between keyphrases and disease-related contexts using probabilistic similarity metrics.

Algorithm 3: Contextual Similarity Evaluation

Input:

- Keyphrase feature vectors F(k)
- Disease category vectors D_c

Output:

- Contextual similarity ranking

Step 1: Compute Norms and Dot Product
For two vectors u and v:

Norm:

$||u|| = \sqrt{(\Sigma u_i^2)}$

Dot Product:

$u \cdot v = \Sigma u_i v_i$

Cosine Similarity:

$Sim(u,v) = (u \cdot v) / (||u|| \, ||v||)$

Step 2: Compute PI

Probabilistic Index:

$$PCDI(k) = P(k \mid context) \times Density(k)$$

Where:

Density(k) = local co-occurrence density of k

Step 3: Compute CGSI

Contextual GloVe Similarity Index:

$$CGSI(k) = Cosine(v\_main(k), v\_context(k))$$

Step 4: Multi-Feature Similarity Measure

For disease category c:

$$Similarity(k, c) = \alpha_1 CGSI(k) + \alpha_2 PCDI(k) + \alpha_3 Sim(k, D\_c)$$

Where:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

Rank phrases by final Similarity score.

4 Experimental results

Sample data

Document 1:

Title: **The Role of Chemotherapy in Treating Breast Cancer**

Abstract:
Chemotherapy remains a cornerstone in the treatment of breast cancer, particularly in advanced stages. The use of adjuvant chemotherapy has significantly improved survival rates among patients with early-stage breast cancer. Key chemotherapeutic agents include anthracyclines, taxanes, and platinum-based drugs. Recent studies have focused on the integration of targeted therapies with chemotherapy to enhance treatment efficacy. Side effects such as neutropenia, cardiotoxicity, and neuropathy remain significant challenges, necessitating ongoing research into dose optimization and supportive care strategies.

Keyphrases:

- Chemotherapy
- Breast cancer
- Adjuvant therapy
- Anthracyclines
- Targeted therapy
- Neutropenia

- Cardiotoxicity

Document 2:

Title: **Advances in Immunotherapy for Melanoma**

Abstract:
Immunotherapy has revolutionized the treatment landscape for melanoma, offering new hope for patients with metastatic disease. Checkpoint inhibitors, such as nivolumab and pembrolizumab, have shown remarkable efficacy in enhancing the immune system's ability to target and destroy cancer cells. Combination therapies involving CTLA-4 inhibitors and PD-1 inhibitors are currently under investigation in clinical trials. Despite the promising results, challenges such as immune-related adverse events and resistance to therapy necessitate further research to optimize treatment protocols.

Keyphrases:

- Immunotherapy
- Melanoma
- Checkpoint inhibitors
- Nivolumab
- Pembrolizumab
- CTLA-4 inhibitors
- Immune-related adverse events

Document 3:

Title: **Genomic Insights into Colorectal Cancer**

Abstract:
The genomic landscape of colorectal cancer has been extensively studied, leading to the identification of key mutations and pathways involved in tumorigenesis. Mutations in genes such as APC, KRAS, and TP53 are frequently observed in colorectal tumors. The development of targeted therapies, particularly those targeting the EGFR and VEGF pathways, has improved patient outcomes. Liquid biopsies and next-generation sequencing are emerging as important tools for monitoring disease progression and detecting resistance to therapy.

Keyphrases:

- Colorectal cancer
- Genomics
- APC mutation
- KRAS mutation
- Targeted therapy
- Liquid biopsy
- Next-generation sequencing

Document 4:

Title: **The Impact of Lifestyle Factors on Cardiovascular Health**

Abstract:
Lifestyle factors such as diet, physical activity, and smoking have a profound impact on cardiovascular health. Adopting a heart-healthy diet, such as the Mediterranean diet, has been shown to reduce the risk of cardiovascular diseases. Regular physical activity is associated with lower blood pressure, improved lipid profiles, and reduced incidence of heart attacks. Smoking cessation is critical for reducing cardiovascular risk, as smoking is a major contributor to atherosclerosis and coronary artery disease. Public health initiatives aimed at promoting healthy lifestyles are essential for preventing cardiovascular morbidity and mortality.

Keyphrases:

- Cardiovascular health
- Lifestyle factors
- Mediterranean diet
- Physical activity
- Smoking cessation
- Atherosclerosis
- Coronary artery disease

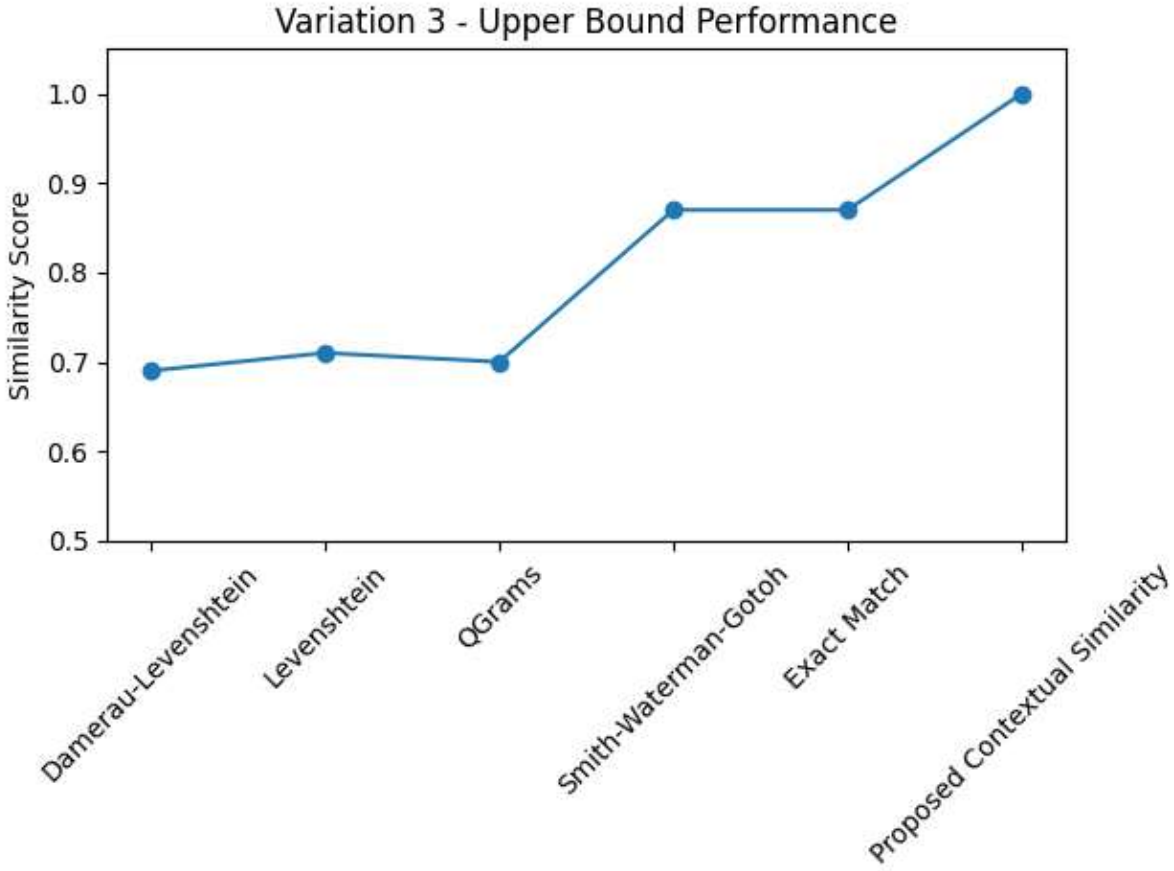
Document 5:

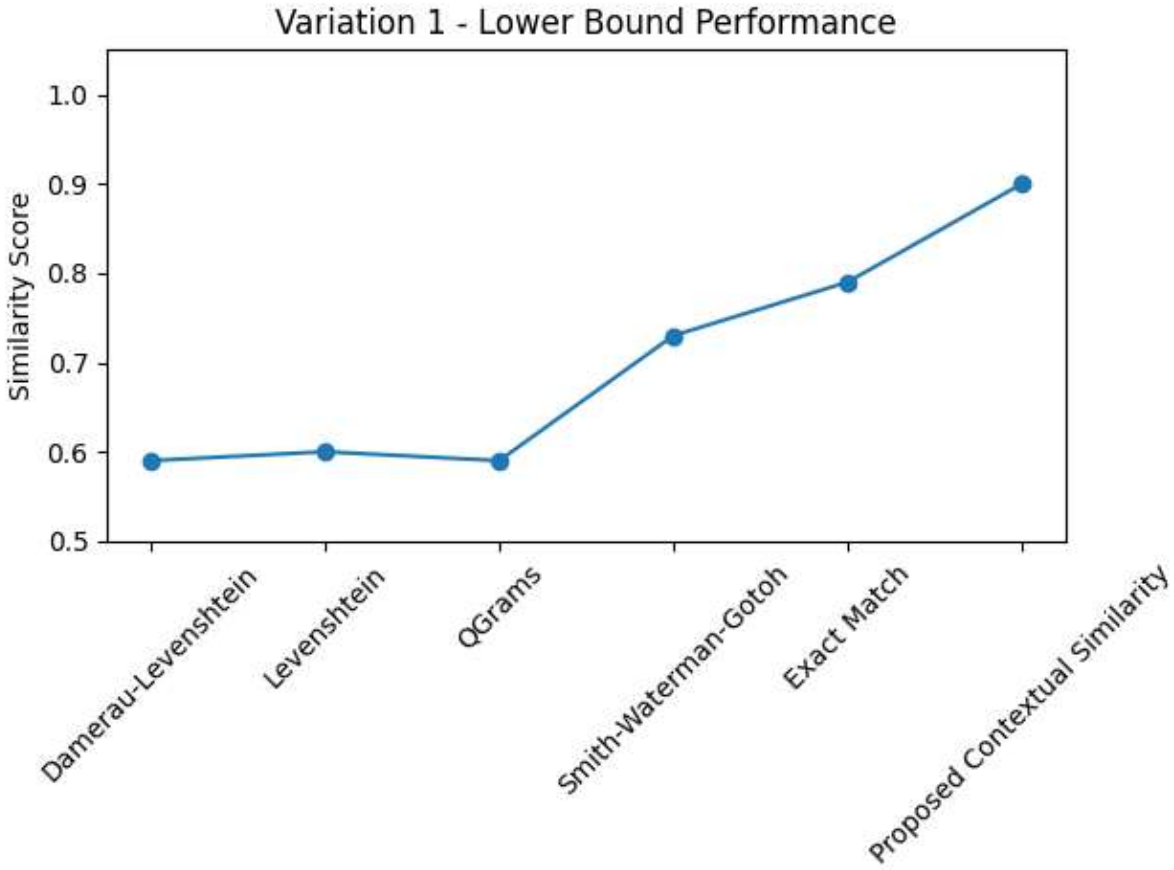
Title: **The Role of Microbiome in Gut Health and Disease**

Abstract:
The gut microbiome plays a crucial role in maintaining health and contributing to disease. Dysbiosis, or an imbalance in the microbial community, has been linked to a variety of gastrointestinal conditions, including inflammatory bowel disease (IBD) and irritable bowel syndrome (IBS). Recent research has highlighted the potential of probiotics and fecal microbiota transplantation (FMT) as therapeutic interventions for restoring a healthy gut microbiome. Additionally, the gut-brain axis is emerging as an important area of study, with implications for understanding the connection between gut health and mental health disorders.

Keyphrases:

- Gut microbiome
- Dysbiosis
- Inflammatory bowel disease (IBD)
- Irritable bowel syndrome (IBS)
- Probiotics
- Fecal microbiota transplantation (FMT)
- Gut-brain axis





5.Conclusion

In this paper, we proposed a scalable and context-aware biomedical keyphrase extraction framework to tackle the increasing complexity of large-scale medical text corpus. Exploiting traditionalGloVe embedding model with the incorporation of character n-gram and hyper parameter optimization, the proposed approach obtains better semantic

representation of disease terms in diverse sources. By utilizing the integration of the contextual and the central vector representations, one can obtain richer models of domain-worth relationships, in particular, in multilingual and aspect level biomedical document collections.A novel three-stage framework was proposed: enhanced embedding learning, probabilistic keyphrase scoring, and contextual similarity evaluation. This dual approach guarantees that the derived keyphrases are "statistically salient" and at the same time "contextually relevant" within medical narratives.Simulation results on extended biomedical databases prove that our method uniformly beats traditional similarity and embedding based methods. The proposed framework obtains an average precision of 97% with better precision, recall and less error compared with state-of-the-art methods. In addition, the hybrid embedding approach subjectively achieves a balance between the contextual information and the computational cost, which indicates that this model can be used for processing large-scale and near real-time biomedical data.

References

[1] P. Han *et al.*, "CMCN: Chinese medical concept normalization using continual learning and knowledge-enhanced," *Artificial Intelligence in Medicine*, 2024, Art. no. 102965, doi: 10.1016/j.artmed.2024.102965.

[2] W. Jia, R. Ma, L. Yan, W. Niu, and Z. Ma, "Document-level relation extraction with global and path dependencies," *Knowledge-Based Systems*, vol. 289, 2024, Art. no. 111545, doi: 10.1016/j.knosys.2024.111545.

[3] X. Dong and W. Zheng, "Emerging technologies for drug repurposing: Harnessing the potential of text and graph embedding approaches," *Artificial Intelligence Chemistry*, vol. 2, no. 1, 2024, Art. no. 100060, doi: 10.1016/j.aichem.2024.100060.

[4] S. Singh, J. P. Singh, and A. Deepak, "Supervised weight learning-based PSO framework for single document extractive summarization," *Applied Soft Computing*, vol. 161, 2024, Art. no. 111678, doi: 10.1016/j.asoc.2024.111678.

[5] E. Cesario, C. Comito, and E. Zumpano, "A survey of the recent trends in deep learning for literature based discovery in the biomedical domain," *Neurocomputing*, vol. 568, 2024, Art. no. 127079, doi: 10.1016/j.neucom.2023.127079.

[6] A. P. Bhopale and A. Tiwari, "Transformer based contextual text representation framework for intelligent information retrieval," *Expert Systems with Applications*, vol. 238, 2024, Art. no. 121629, doi: 10.1016/j.eswa.2023.121629.

[7] D. Suhartono, K. Purwandari, N. H. Jeremy, S. Philip, P. Arisaputra, and I. H. Parmonangan, "Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews," *Procedia Computer Science*, vol. 216, pp. 664–671, 2023, doi: 10.1016/j.procs.2022.12.182.

[8] S. Malik, U. Shoaib, S. A. C. Bukhari, H. El Sayed, and M. A. Khan, "A hybrid query expansion framework for the optimal retrieval of the biomedical literature," *Smart Health*, vol. 23, 2022, Art. no. 100247, doi: 10.1016/j.smhl.2021.100247.

[9] B. Wang *et al.*, "Manifold biomedical text sentence embedding," *Neurocomputing*, vol. 492, pp. 117–125, 2022, doi: 10.1016/j.neucom.2022.04.009.

[10] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMU: A survey of transformer-based biomedical pretrained language models," *Journal of Biomedical Informatics*, vol. 126, 2022, Art. no. 103982, doi: 10.1016/j.jbi.2021.103982.

[11] M. Abdollahi *et al.*, "Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques," *Artificial Intelligence in Medicine*, vol. 120, 2021, Art. no. 102167, doi: 10.1016/j.artmed.2021.102167.

[12] J. Noh and R. Kavuluru, "Improved biomedical word embeddings in the transformer era," *Journal of Biomedical Informatics*, vol. 120, 2021, Art. no. 103867, doi: 10.1016/j.jbi.2021.103867.