International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com
Volume 3 Issue 8 , August 2026
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

HARNESSING MESH ONTOLOGY FOR SEMANTIC CLASSIFICATION IN MEDICAL TEXTS

S. Narmatha[1], Dr. V. Maniraj[2]
[1] *Research Scholar, Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli), Poondi, Thanjavur, Tamilnadu, India.*
[2] *Research Advisor, PG and Research Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli)., Poondi, Thanjavur, Tamilnadu, India..*

Article Info

Article History:

Published: 6 August 2026

Publication Issue:
Volume 3, Issue 8
August-2026

Page Number:
57-65

Corresponding Author:
S. Narmatha

Abstract:

This research investigates the difficulties associated with employing domain-specific ontologies for the classification of online textual content. The study specifically examines the use of the Medical Subject Headings (MeSH) ontology to enhance the classification of medical documents. A novel content representation model is introduced based on the insights obtained from this investigation. The proposed method is evaluated against conventional stem-based text representations using two well-known data mining algorithms, namely C4.5 and K-Nearest Neighbors (KNN). Experimental results demonstrate that the proposed approach consistently outperforms traditional methods. By integrating semantic concepts and hierarchical relationships, such as hypernyms, from domain ontologies, the quality of feature representations for document classification is significantly improved. Validation on the Ohsumed biomedical dataset shows that the proposed model achieves up to a 30% increase in classification accuracy compared to standard stem-based techniques.

Keywords: K-Nearest Neighbours, semantic classification

1. Introduction

The widespread availability of information on the internet has greatly facilitated content sharing and access; however, identifying and retrieving information that is both relevant and valuable remains a major challenge. As a result, considerable research efforts are directed toward enhancing automated information retrieval methods. These advances are also critical to the development of the Semantic Web, which relies on novel mechanisms for structuring data and enriching it with semantic annotations, typically through metadata. In this context, ontologies—formal representations of knowledge within a specific domain—serve as fundamental tools. By defining concepts and the relationships among them, ontologies enable a structured and semantically meaningful organization of information.

This study investigates the role of a conceptual, ontology-based framework in supporting the automatic classification of medical documents, which constitutes the main objective of this work. The remainder of the paper is structured as follows. Section 2 surveys existing research related to medical document classification and ontology-based methods. Section 3 presents the proposed approach, detailing its architecture, methodology, and distinguishing characteristics. Section 4 describes the Ohsumed dataset and the MeSH medical ontology and explains how they are used to evaluate the proposed model against a conventional bag-of-stems representation. Section 5 discusses the experimental setup and provides an in-depth analysis of the results. Finally, Section 6 summarizes the findings and outlines potential directions for future research.

2. Related Work

Although notable progress has been made in document classification research, a large number of existing approaches continue to depend on the traditional bag-of-words representation, which is increasingly regarded as insufficient. One of its main drawbacks is the lack of semantic awareness, as it treats terms independently and ignores relationships between them. To overcome these limitations, recent studies have shifted toward concept-based representations of textual data, with ontologies gaining recognition as an effective alternative.

In particular, the work of Amine demonstrates the benefits of ontology utilization—such as WordNet—for enhancing document clustering performance. More recently, G has shown that ontology-based classification techniques are both feasible and efficient. Their research illustrates the use of multilingual ontologies to classify bilingual websites and proposes the construction of domain-specific ontologies by combining linguistic analysis with statistical approaches.

These studies highlight the importance of semantically enriched and precise content classification, especially in ontology-driven systems. For such systems to function effectively, it is essential to develop and optimize robust algorithms capable of automatically classifying online documents using domain-specific ontologies. This requirement becomes even more critical in contexts where predefined knowledge bases or trained learning models are not available.

3. A Method for Conceptual Representation

Effective document classification requires a representation model that balances computational efficiency with the preservation of informative features. While the bag-of-words model remains widely used, its inherent limitations have encouraged researchers to investigate more sophisticated alternatives. In this work, we propose a concept-based representation framework that both enhances semantic richness and reduces feature vector dimensionality, leading to improved classification performance.

To enable a fair comparison with conventional approaches, our method is implemented through two main stages:

1. Conceptual Transformation

In the first stage, selected textual terms are mapped to domain-specific concepts using an appropriate matching and disambiguation mechanism. This process enriches the representation by incorporating contextual and semantic information, allowing the model to capture meaning beyond simple term frequency.

2. Hypernym-Based Enhancement

The conceptual representation is further refined by introducing hypernyms, which represent higher-level conceptual categories. This enhancement emphasizes broader semantic relationships and contributes to a more informative and robust feature vector for document classification.

Preprocessing Stage

Prior to applying the proposed conceptual representation model, the textual data undergoes a comprehensive preprocessing phase. This step involves removing irrelevant elements such as HTML tags, images, and other non-textual content. The preprocessing process aims to reduce noise, eliminate redundant or semantically equivalent terms, and retain only the most meaningful information required for accurate classification.

Key preprocessing steps include:

1. Stop Word Removal

Stop words are commonly occurring terms that carry minimal semantic value, such as *"the," "and,"* and *"of."* Eliminating these terms helps reduce the dimensionality of the dataset and improves processing efficiency without sacrificing meaningful content. Typically, stop words are identified using predefined, language-specific lists.

2. Stemming

Stemming is the process of reducing words to their base or root forms in order to consolidate different morphological variations of the same term into a single feature. Originally proposed by Porter in 1980 for English, stemming enables classification systems to recognize words such as *"walk," "walking,"* and *"walker"* as representations of the same concept. Without this step, each variation would be treated as a distinct feature, potentially lowering classification performance. The Porter stemming algorithm operates by analyzing word patterns to extract common roots and has since been adapted for various languages, often with guidance from linguistic experts.

Figure 1 provides an example of a text after being cleaned and stemmed, highlighting how the refined representation improves clarity and reduces redundancy in the classification process.

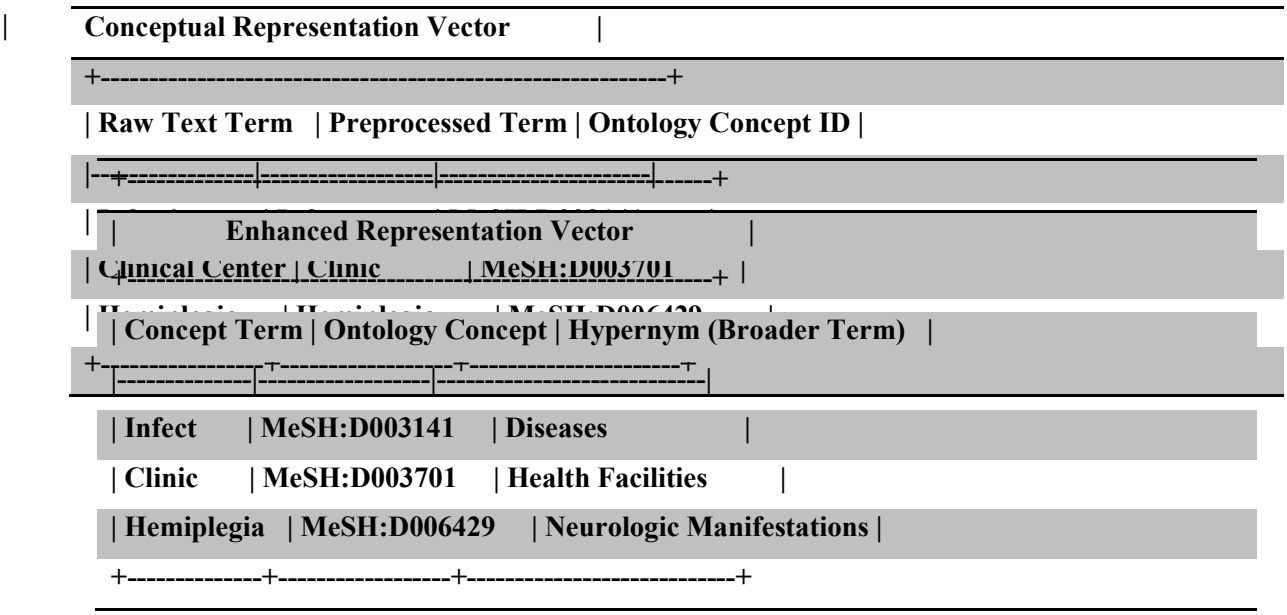


Figure 1. This is an example of an Ohsumed benchmark paragraph representation vector.

This illustration appears to represent how a piece of text or a phrase (such as "Infect Clinic Hemiplegia") is converted into a **representation vector**—a structured numerical format. This vector serves to encapsulate the **semantic content and relationships** among the words, enabling computational systems to interpret, classify, or analyze the information effectively. By mapping each term to its conceptual meaning—often using a domain-specific ontology like MeSH—the vector becomes a **semantic abstraction** of the original text, rather than a mere collection of raw words. This transformation allows machine learning algorithms to perform tasks like classification or clustering with greater accuracy and context awareness.

THE MAPPING OF TERMS TO CONCEPTS

The connection between words and their underlying ideas is illustrated in Figure 2, which serves as an example of how individual terms are linked to specific concepts. This process of mapping is fundamental in natural language processing because it allows the **system to move** beyond mere words and capture the abstract meanings and relationships they represent. By associating terms with their corresponding concepts, we can better understand and organize the semantic content of text. While Figure 2 likely provides a visual representation of this mapping process, without the actual figure, a more detailed description is not possible.

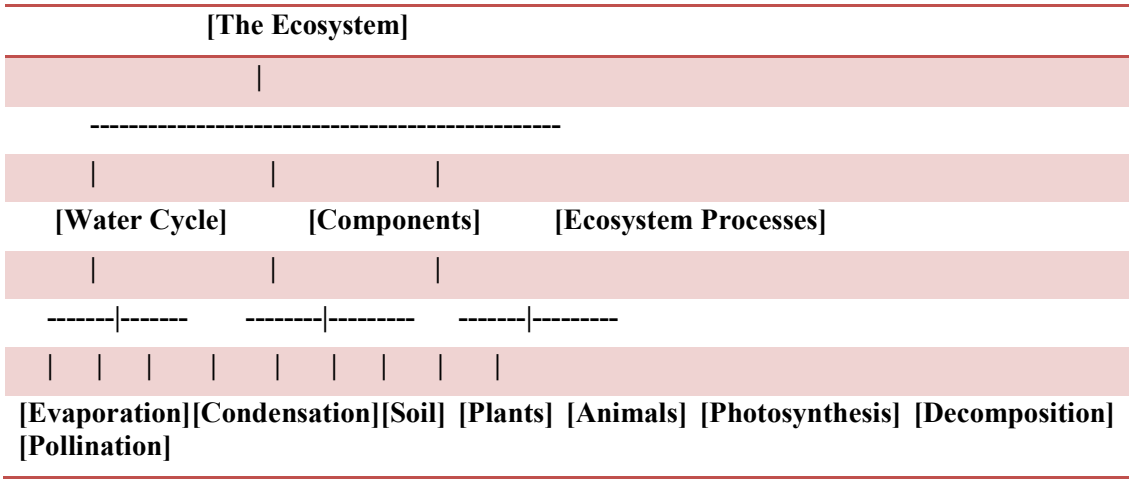


Figure 2. Illustration of the correlation between terms and concepts

- Water Cycle includes Evaporation, Condensation, Precipitation

- Components include Soil, Plants, Animals

- Ecosystem Processes include Photosynthesis, Decomposition, Pollination.

The ontology functions as a bridge linking specific terms to the concepts they represent. For instance, the words "appendicitis" (with a frequency of 2) and "appendiceal" (frequency of 1) both correspond to the same underlying concept of appendicitis. Their individual term frequencies are then combined to form the total frequency for that concept. At this stage, we can identify three theoretical strategies for incorporating or replacing terms with concepts:

- **ReplacingWordswithConcepts:**
  This strategy is similar to the first approach but uses a concept vector that integrates all terms found in MeSH (Medical Subject Headings), instead of simply duplicating them in the new representation. The resulting term vector will only retain terms that do not appear in MeSH.

- **Concept-OnlyVector:**
  In contrast, this method completely excludes all terms, even those listed in MeSH, creating a representation composed solely of concepts.

- **AddressingAmbiguity:**
  Since language is inherently ambiguous, assigning exact meanings to words can be challenging. A single term may have multiple interpretations depending on context. Word Sense Disambiguation (WSD)—the task of identifying the correct meaning—is one of the most complex problems in artificial intelligence. Because advanced WSD techniques can be impractical or difficult to apply, we focus on two simpler methods instead. This approach assumes that every document contains central themes or core ideas, which naturally surface and aid in revealing more relevant topics, even as the number of dimensions grows. Concept frequencies are computed accordingly to reflect this understanding.

$$cf(d,c) = tf\{d, t \in T \mid c \in (ref_c(t))\}$$

The key idea behind this approach is to focus on the most common meanings of a word. Essentially, the ontology generates a ranked list of concepts, giving higher priority to the more frequently used interpretations while downplaying the less common ones. This is a standard practice in most ontologies. Concept frequencies are then calculated based on this prioritized ordering as follows:

$$cf(d,c)=tf\{d,t \in T|\ first(ref_c(t)) = c\}$$

4. **IncorporatingHypernyms:**
   Grasping the relationships between concepts is crucial for accurately capturing the main themes within texts. Simply replacing terms with their corresponding concepts, without considering how these concepts relate to one another, may not significantly improve performance—and in some cases, might even be less effective than using the original terms. This observation is consistent with findings in recent research. To overcome this limitation, we combine the frequency of each concept in the text with the frequencies of its hyponyms, thereby leveraging the hierarchical (hypernym–hyponym) relationships among concepts. The next step involves updating the frequency values in the concept vector accordingly, where H(c) denotes the set of hyponyms linked to the concept c.

$$cf'(d,c) = \sum_{b \in H(c)} cf(d,b)$$

5. **SelectingandReducingDescriptors:**
   The mapping process is performed for every document in the training dataset. The concepts extracted from the ontology serve as descriptors that form the vector representation of each document. A concept's frequency within a particular category of the training corpus reflects its importance. The aim of this selection step is to identify descriptors that best distinguish each category from the others. Assigning weights to terms highlights their relevance within a specific category. Simply counting how often a term appears in a category is a basic approach, but it fails to consider how distinctive the term is compared to other categories. The more widely used and effective method is TF-IDF (Term Frequency-Inverse Document Frequency), which provides better weighting by balancing term frequency with its uniqueness across categories. This method is applied within the vector space model as follows:

«Frequency Term »* « Inverse document frequency »

Term Frequency (TF) refers to how many times a particular term appears within a given category. **Inverse Document Frequency (IDF)** is calculated by dividing the total number of categories by the number of categories that include the term under evaluation. This helps measure the term's importance by reducing the weight of commonly occurring terms across many categories. The formula is expressed as:

$$\mathbf{FTIDF(Ci, Wj)} = TF(Ci, wj) * log \frac{\mathbf{nbr_{category}}}{\mathbf{DF(wj)}}$$

The process of dimensionality reduction involves selecting a subset of features from the full set that better differentiates between categories. Controlling the number of features in the vector space is important for two main reasons. First, the complexity of a learning algorithm depends on both the number of features and the number of training samples. By reducing the number of indexed terms, the efficiency of the learning process can be improved. While having more features might seem beneficial for accuracy, it is not always helpful if many features are irrelevant or redundant.

6. Dimensionality Reduction and Feature Selection:
Although additional features can provide more information, overwhelming the learning algorithm with too many,

including irrelevant ones, can degrade performance. Dimensionality reduction addresses this issue by focusing on terms that strongly correlate with specific categories, as these are more useful for distinguishing among classes. During this step, terms with low χ^2 (chi-square) values are removed. The χ^2 test is a supervised statistical method that evaluates not only the frequency of terms within categories but also how terms interact with categories overall. To apply this technique, each category is identified, and then the top K features that best describe that category in comparison to others are selected**.**

COMPUTING X2 USING MATHEMATICS

Here's a simplified mathematical formula to compute $\chi 2$:

$$X^2(D, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) * (N_{11}N_{00} - N_{10} + N_{01})^2}{(N_{11} + N_{01}) * (N_{11} + N_{10}) * (N_{10} + N_{00}) * (N_{01} + N_{00})}$$

Where:

- $\chi 2$ (chi-squared) represents the association between a term and a category

Ntc is the number of documents for category c, with value 0.1 (this represents a small fraction of the total to ensure the calculation accounts for sparse data)

N11 (n_0_1) is the total number of documents that contain the term and are classified as belonging to category c.

7. Understanding the Components of χ^2 Calculation:

In the context of using the χ^2 (chi-square) method for feature selection, it is important to understand the counts related to whether terms appear in documents within specific categories:

- **N10:** The number of documents that include the term but do **not** belong to the target category.
- **N01:** The number of documents that belong to the target category but do **not** contain the term.
- **N00:** The number of documents that neither contain the term nor belong to the category.

Key Characteristics of the χ^2 Method:

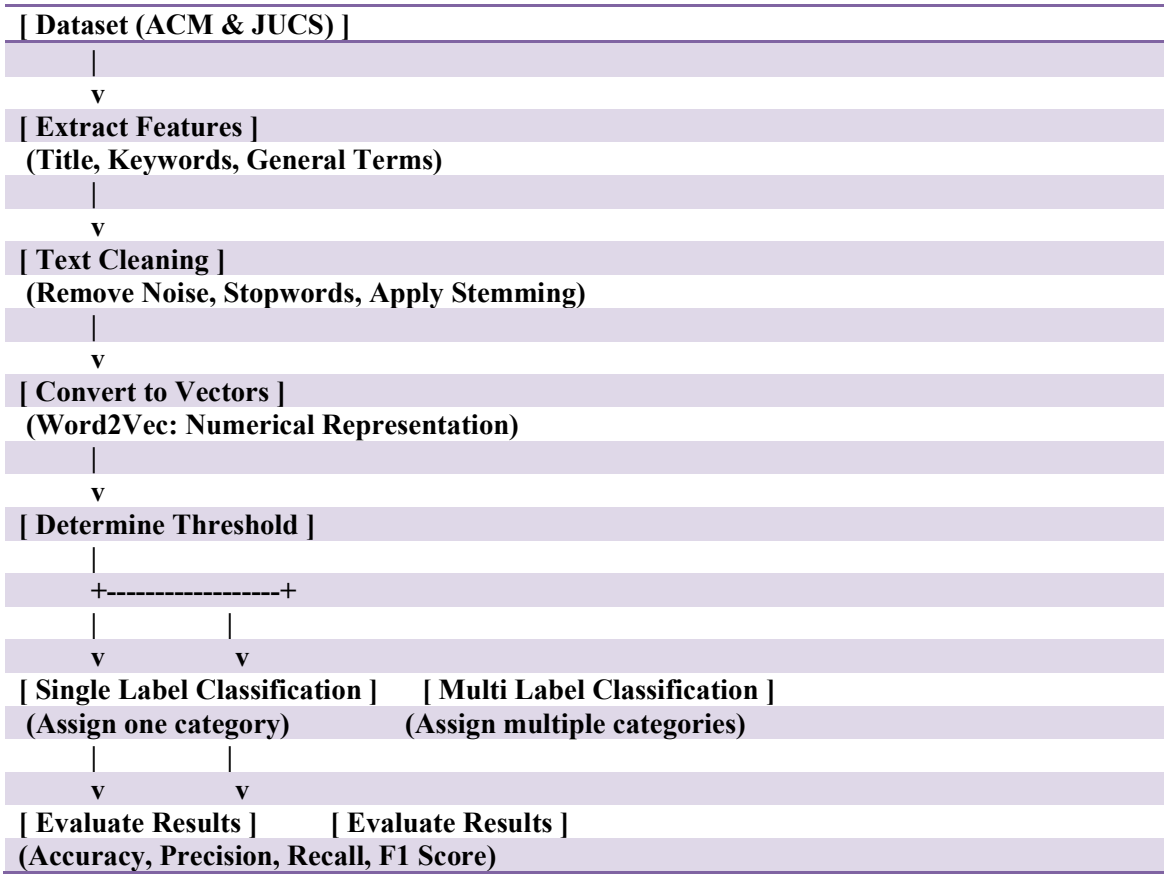
- **Supervised Technique:** The χ^2 approach is supervised because it depends on labeled data, utilizing category information to guide the selection of relevant features.
- **Multivariate Analysis:** This method is multivariate as it evaluates not only individual term-category relationships but also considers the interaction between multiple features and categories.
- **Focus on Relationships:** The primary goal is to investigate how features relate and interact with categories, helping to identify the most meaningful terms for classification.
- **Linear Computational Complexity:** Despite its thoroughness, the χ^2 method scales efficiently, with calculations growing linearly relative to the size of the dataset, making it suitable for large-scale applications.

In summary, χ^2 leverages labeled category data to analyze feature-category interactions effectively while maintaining computational efficiency, making it a valuable tool for feature selection in classification.

Building and Evaluating a Text Classification Model:

Once documents are labeled and have undergone preprocessing and conceptual representation, the next step is to build a machine learning model using the concept vector matrix. This step follows conventional supervised classification workflows. For this study, we will assess the performance of two popular classifiers: **C4.5** and **K-Nearest Neighbors (KNN)**. Due to their common use, we will concentrate on the conceptual understanding rather than detailed algorithmic explanations.

After training, the model can classify new documents by converting them into concept vectors through the same preprocessing pipeline and feeding these vectors into the trained classifier.



4. ASSESSMENT OF THE METHODOLOGY

THE OHSUMED DATASET

In 2000, we employed the OHSUMED dataset as a component of the TREC9 Task-Filtering framework. This dataset comprises 270 medical articles published or summarized between 1987 and 1991. Each document contains six sections: the title (.T), summary (.W), author (.A), source (.S), and publication date (.P). Additionally, the text features MeSH-indexed concepts (.M).

Evaluation Approach:

In this section, we present experimental outcomes using the **F-measure**, a metric that balances recall and precision by calculating their harmonic mean:

In text classification, **precision** and **recall** are standard metrics used to assess the performance of algorithms for a given category. They are defined as follows:

$$F = \frac{2 * recall * precision}{recall + precision}$$

- **Precision** measures the proportion of correctly identified positive instances out of all instances labeled positive:

- **Recall** evaluates how many actual positive instances were correctly detected:

$$Precision = \frac{True\ positive}{True\ positive + false\ positive}$$

$$recall = \frac{True\ Positive}{True\ positive + false\ negative}$$

Our experiments leverage the **MeSH Ontology**, a comprehensive biomedical thesaurus developed by the National Library of Medicine (NLM), USA. MeSH has been a cornerstone resource since its first public release in 1954 and has grown significantly, containing over 25,000 descriptors as of 2010. These descriptors are organized hierarchically across sixteen categories (e.g., anatomical terms in Category A, diseases in Category C), with a maximum depth of eleven levels, facilitating detailed and structured medical indexing.

To validate our method, we applied it to the **Ohsumed** database and tested two data mining algorithms: **C4.5**, a decision tree classifier, and **K-Nearest Neighbors (KNN)**. Both are widely recognized for their effectiveness in document classification. The performance was compared to stem-based text representation, across eight Ohsumed categories (see Table 1).

Our results indicate that ontology-based representations outperform stem-based methods by a significant margin—achieving approximately a 30% improvement in classification accuracy, a noteworthy advancement for theoretical research.

Moreover, enriching the representation vectors by integrating hypernyms and related concepts further boosted performance. The strong results also highlight the effectiveness of the KNN algorithm, particularly in combination with the **CHI 2** feature reduction technique. The synergy between KNN and CHI 2 may explain the improved outcomes, suggesting avenues for future exploration to refine classification and representation strategies further.

5. Conclusion and Future Directions

The main goal of our study was to enhance document classification by leveraging domain-specific ontology. We focused on the medical domain due to its importance and growing interest within the data mining field. Our method was tested on a standard benchmark dataset using two widely-used classifiers, KNN and C4.5, each evaluated over three iterations. We began by using the traditional stems-based representation as a baseline for comparison, followed by a detailed analysis of the results. Then, we incorporated key concepts and hypernym relationships from the MeSH ontology into the document representation.

The experimental outcomes confirm the strength of our approach, revealing an impressive 30% improvement in classification performance—surpassing our initial expectations. This clearly demonstrates that ontology-driven, subject-based document annotation is an effective and robust technique.

Looking forward, several avenues for future research are promising. In particular, exploring hyperonymy further could enable us to expand beyond a single hierarchical level of the MeSH ontology, potentially improving classification

accuracy and enabling better generalization. Another potential direction is applying this theoretical framework to a multilingual MeSH ontology, which would facilitate classification across diverse languages and broaden the scope of our methodology to handle multilingual medical literature.

References

1.Hassan, Sahar, Franck Hétroy, and Olivier Palombi. "Ontology-guided mesh segmentation." *FOCUS K3D Conference on Semantic 3D Media and Content*. 2010.

2.Osborne, John D., et al. "Interpreting microarray results with gene ontology and MeSH." *Microarray Data Analysis: Methods and Applications* (2007): 223-241.

3.Trieschnigg, Dolf, et al. "MeSH Up: effective MeSH text classification for improved document retrieval." *Bioinformatics* 25.11 (2009): 1412-1418.

4.Elberrichi, Zakaria, Belaggoun Amel, and Taibi Malika. "Medical Documents Classification Based on the Domain Ontology MeSH." *arXiv preprint arXiv:1207.0446* (2012).

5.Yoo, Illhoi, and Xiaohua Hu. "Biomedical ontology mesh improves document clustering qualify on medline articles: A comparison study." *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, 2006.