Computers & Industrial Engineering 176 (2023) 108940

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

# Machine learning and ontology-based novel semantic document indexing for information retrieval

Anil Sharma [a,b,*], Suresh Kumar [c]

[a] University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Delhi 110078, India
[b] College of Computing Sciences & IT, Teerthanker Mahaveer University, Moradabad 244001, India
[c] Department of Computer Science and Engineering, Netaji Subhas University of Technology, East Campus, Delhi 110031, India

ARTICLE INFO

*Keywords:*
Document indexing
Concept extraction
Machine learning
Computer science ontology
Semantic web
Information retrieval
Natural language processing

ABSTRACT

The goal of information retrieval (IR) systems is to find the contents most closely related to the user's information needs from a pool of information. However, conventional IR methods neglect semantic descriptions of document contents and index documents based on the words that they include. When users and indexing systems use different terms to express the same subject, a vocabulary gap emerges. To overcome this limitation and to enhance the effectiveness of the IR systems, this paper introduced a novel hybrid semantic document indexing employing machine learning and domain ontology. The presented technique uses a skip-gram with negative sampling-based machine learning model and a domain ontology to determine the concepts for annotating unstructured documents. The proposed work also introduced multiple feature based novel concept ranking algorithm where statistical, semantic, and scientific named entity features of the concept were used to assign relevance weight to the annotations. The fuzzy analytical hierarchy process was used to derive the parameters of these feature weights. The final step is to rank the concepts according to their relevance to the document. Five benchmark publicly accessible datasets from the computer science domain were used in a series of experiments to validate the results of presented method. Experiment findings showed that the proposed method performs better than state-of-the-art techniques on these datasets, by improving average accuracy by 29%, while an improvement of 25% was recorded in F-measure. The improvement in average accuracy demonstrates that the performance of the proposed approach is better than the state-of-the-art methods in extracting document concepts accurately even when the same concept is referred to by distinct terms in the document and domain ontologies. The proposed system's ability to find similar concepts when the documents possess no concept from domain ontology is demonstrated by the improvement in F-measure, which is attributed to high recall rates of the proposed indexing scheme while maintaining high accuracy.

## 1. Introduction

Online resources have become the largest source of information and zillions of users are exploring these resources to satisfy their information needs. Information retrieval (IR) systems are designed to deliver results that are closely related to the user's information demand (Guo et al., 2020). The document indexing stage assigns keywords that describe the contents of the documents to make it easier to map a query to a set of documents during retrieval (Kumar, Singh, and De, 2012). To improve the effectiveness of information retrieval tasks, several solutions employing statistical, neural, and semantic techniques have been suggested in the literature (Upadhyay et al., 2020; Shanmuganathan et al., 2021; Anand and Kumar, 2022a). These techniques adhere to standard indexing procedures: The first step is pre-processing of documents, which includes lemmatization, tokenization, and elimination of undesirable features (stop words, symbols, punctuations and special characters,). The second step is to extract the promising key phrases based on a lexical and morphological pattern of terms. In the third step, multiple feature based relevance score was assigned to key phrases extracted from the text. Lastly selection and ranking of key phrases based on relevance score.

The documents are indexed in traditional IR methods based on the terms they contain rather than the concept that describe them. When users and domain experts use different terminology to express the same subject, a vocabulary gap exists. Semantically relevant contents without a lexical match between query and information contents are not

* Corresponding author.

https://doi.org/10.1016/j.cie.2022.108940

Available online 28 December 2022

included in the result by these IR systems. Term semantics and their relationships have been successfully captured using domain knowledge in document representation in an IR system (Anand and Kumar, 2022b). To formally express and make sense of domain knowledge, ontologies have long been employed in every area. While vocabulary gaps are no longer an issue thanks to ontology-based approaches for document indexing. However, a limitation of these methods is that they heavily rely on the scope and level of information of the original input ontology. Recent advances in neural network success in the field of information retrieval (IR) make them a viable solution for document indexing tasks (Shokat et al., 2020; Singh et al., 2021). The neural network techniques make use of the semantic relationships between words in contexts where they co-occur, but they ignore the significant semantic relational structures. Ontologies and semantic lexicons are examples of manually compiled knowledge bases that contain these semantic relationship structures.

The objectives of the proposed research are: first, to deal with the problem of vocabulary gap when different terms are used to represent the same concept in documents and domain ontology; and second, to provide concepts for indexing even when document terms don't have corresponding concepts in the domain ontology. The goal of this paper includes proposing a fine-grained semantic indexing of scientific literature beyond the descriptors of Computer Science Ontology at the semantic level of corresponding concepts so as to bring results close to those achieved by domain experts. As a solution to this issue, a novel machine learning and ontology-based semantic document indexing (MLOntoSDI) approach for information retrieval is proposed in this paper.

The proposed indexing approach is intended for unstructured text documents of domain-specific datasets, particularly scientific literature in the field of computer science. The datasets used is unique in the way that it consists of computer science domain-specific vocabulary features viz. methods, models, tools, system components, evaluation metrics, corpus, knowledge base, their relations, and coreferences. The proposed indexing method performed well on these datasets because the processing involves feature-dependent tasks; for example, identifying domain-specific named entities for part-of-speech tagging, and choosing stemming or lemmatization for getting the base or root form of domain-specific terms representing documents etc.

The MLOntoSDI technique uses word embeddings and computer science ontology (CSO) to identify the concept in a text. The proposed solution uses NLP techniques to extract the promising key phrases from document text and feeds them into the ML model. In order to find concepts, the skip-gram with negative sampling model supplements document vocabulary with semantically related terms and extracts corresponding concepts from a domain ontology. By performing a fuzzy match between document vocabulary and ontology concepts, the MLOntoSDI technique eliminates the problem of vocabulary gap. By using CSO, it extracts the concepts, their hyponyms and hypernyms. Based on the statistical and semantic features, augmented with named entity characteristics of the concepts, the relevance weights are assigned to them. For determining feature weighting parameters, the Fuzzy analytic hierarchy process (FAHP) is applied. The concepts are sorted based on their relevance weight.

The organization of this paper is as follows: the relevant literature is outlined in the section that follows. The motivation and objectives of the proposed scheme are explained in Section 3. The proposed semantic document indexing strategy is presented in Section 4. The evaluation is deliberated in Section 5. Section 6 presents findings and discussion. And the conclusion and recommendations are stated in Section 7.

## 2. Literature work

There have been numerous document indexing techniques presented in the literature, each with advantages and disadvantages (Prasanth and Gunasekaran, 2019; Hammache and Boughanem, 2021). These indexing techniques can be broadly divided into two categories: free vocabulary and controlled vocabulary-based methods. In a free vocabulary-based indexing scheme, any term composing a document can be used to represent the document. One concept may be represented by more than one term as there is no standardization of terms and control of synonyms and homographs. The usage of a free vocabulary leads to many problems, such as a vocabulary gap as different authors use different keywords to denote the same idea. A controlled vocabulary is a set of words and phrases that have been structured for the purpose of indexing information for browsing. It often describes a certain domain and uses the concept's preferred and alternative terminology. This research work utilized a controlled vocabulary by employing Computer Science Ontology (CSO) as an external knowledge source. The major differences between these two indexing schemes are presented in Table 1.

**Table 1**
Comparison between free vocabulary and controlled vocabulary based document indexing.

| Parameters | Document Indexing<br>Free Vocabulary based | Controlled Vocabulary based |
|---|---|---|
| Goal | To create document indexing using the terms composing a document | Link concept and it's alternative as well as preferred terms logically and hierarchically to create controlled vocabulary for document indexing |
| Representation of concept | One concept may be represented by more than one terms for indexing | One concept is represented with one term and its alternative and preferred terms for indexing |
| Standardization of terms | No standardization and control over terms (synonyms and homographs) | Predefined authorized alternative and preferred terms for a concept in the domain |
| External resource | Not used | Uses external resources and controlled vocabulary for indexing |
| Semantic features | Not available | Available |
| Selection of terms | Different authors use different terms to represent the same concept | Uniformity of terms used to represent a concept |
| Usage | Not much applied in knowledge organization systems | Used in subject headings, thesauri, taxonomies and other knowledge organization systems |
| Limitations | Vocabulary gap problem | Existence of knowledge resource of the domain is prerequisite |
| Applicability with respect to machine learning | Used extensively; but lack semantic relational structure between terms | Yet to be explored fully; support semantic relational structure between terms |

### 2.1. Free language based document indexing

These schemes employ strictly the document's keywords to identify a document. Correia et al. (Correia et al., 2022) presented a concept annotation scheme using NLP in the legal domain. The authors investigated a method for named entity recognition (NER) in the legal field in Portuguese. Two schemes based on bidirectional long-short term memory networks (BiLSTMs) and conditional random fields for NER in the legal domain. Singh et al. (Singh et al., 2022) proposed a classification approach using deep learning for medical images. The preprocessing of the image dataset involves the use of stationary discrete wavelet transformation, while a convolutional neural network is employed for the classification of processed image data.

Hassani et al. (Hassani et al., 2022) presented an indexing scheme for scientific videos in the education domain. The authors identified keywords from features of the text extracted from video frames and audio signals. The proposed approach was tested on datasets from English and

Portuguese language. The authors further applied Friedman's non-parametric statistical test for further analysis of the results. Garg and Sharma (Garg and Sharma, 2022) introduced a linguistic feature-based model to identify misinformation on social networking sites. The authors applied various linguistic features for the extraction of important features for the embedding of news content. Further, a word embedding based machine learning model was used for fact checking purposes.

Guillon et al. (Guillon et al., 2021) proposed a case based reasoning framework to help customers by providing the information needed in a bidding process. The framework utilized a knowledge base for storing, categorizing, and reasoning the information associated with the bidding process. Spolaor et al. (Spolaôr et al., 2021) introduced an indexing scheme for video retrieval using speech or text commands. The authors applied speech to text algorithm to generate annotations and index videos for retrieval. Aman et al. (Aman et al., 2021) proposed latent semantic indexing and a clustering-based key phrase extraction approach. The authors employed frequency-based features for assigning weights for the ranking key phrases.

Ullah, Khusro, and Ahmad (Ullah et al., 2021) presented an IR approach for book retrieval using metadata from social networking sites, bibliographic features, and semantic analysis. The authors applied a metadata based indexing and re-ranking algorithm for book retrieval. Chinnasamy and Deepalakashmi (Chinnasamy and Deepalakshmi, 2022) proposed a secure framework to access electronic health data. The authors applied RSA algorithm based key generation for encryption of health records and the Blowfish encryption technique for encryption of generated key values. The framework used steganography for the exchange of keys, and encrypted health care information is retrieved using keyword search. Abasi et al. (Abasi et al., 2021) presented k-means clustering and multi-verse optimizer-based concept extraction from text documents. The authors also used an ensemble based concept identification algorithm to extract key phrases from documents related to scientific publications.

Bhunia et al. (Bhunia et al., 2020) presented a segmentation based word spotting model for resource scarce Indic languages. The authors employed hidden Markov model based segmentation of text line images with character filler techniques. Bordoloi et al. (Bordoloi et al., 2020) presented a supervised algorithm for key phrase extraction from web resources by combining graph-based technique and vector space model. The authors used a statistical supervised weighting scheme for data labeling and term similarity for incorporating mutual information.

Sharma, Gupta, and Juneja (Sharma et al., 2021) developed a key-phrase recognition and document indexing system based on natural language processing. The unsupervised key-phrase extraction method used by this model is based on term frequency, term vectorization, ontology, and phrase-based features. Euclidean distance was used for proposed features grouping similar key phrases, while cosine distance was used for word embedding vectors. Based on these two scores, the ordering of key keywords is also determined. For extracting spoken documents, Gupta and Yadav (Gupta and Yadav, 2021) presented using an IR method based on hidden Markov models and deep learning. Audio files were converted to text scripts using the Kaldi toolkit, which is based on deep learning. The authors developed wavelet tree-based text document indexing. Documents are ranked based on cosine similarity.

A general framework for indexing and retrieving multimedia on smartphones utilizing machine learning (ML) and artificial intelligence is introduced by Wagenpfeil et al. (Wagenpfeil et al., 2021). To start, they conducted a multimedia feature vector graph using semantic analysis of the medium. Refined feature weights are used to construct semantic indexing. The authors of this proposal processed queries using natural languages and represented them using SPARQL.

### 2.2. Controlled language based document indexing

These schemes use keywords composing a document as well as concepts inferred from domain ontology to represent the documents. For the medical domain, an indexing system based on description logic and Best Match 25 (BM25) model for unstructured documents was presented by Boukhari and Omri (Boukhari and Omri, 2020). The query phrase and the medical thesaurus were roughly matched by the VSM. The proposed technique was presented with enhanced domain knowledge and inference capacity owing to description logic. In the end, less important concepts were eliminated.

Cross-language IR was presented by Rahimi, Montazeralghaem, and Shakery (Rahimi, Montazeralghaem, and Shakery, 2020) based on translation knowledge. In order to calculate the frequency and discrimination values of the query phrases, the authors used the aggregation function. The axiomatic analysis of restrictions and the hierarchical computation of discrimination values served as the foundation for document ranking. Jiang (Jiang, 2020) has presented an IR model based on three different external sources. The presented idea used information from WordNet, Wikipedia, and domain ontology to model a keyword's score in a document. The proposal introduced a semantic network based semantic similarity for matching query terms with the document set. A strategy for information retrieval that integrates external knowledge sources and topic modelling was presented by Subramaniam et al. (Subramaniam et al., 2021). The authors used fuzzy c-means clustering to create document clusters and a modified firefly approach to identify document features. The relevant documents are retrieved using Latent Dirichlet Allocation from appropriate clusters.

Nentidis et al. (Nentidis et al., 2020) proposed a concept annotation method in the biomedical domain using MeSH descriptors. The authors also proposed semantic indexing based on the classification and the supervised predictive technique. The authors tested their approach on two diseases and found that this approach provided better concept annotation than heuristic based state-of-the-art annotation methods. Vidal et al. (Vidal et al., 2014) presented a semantic description generator for documents in the education domain using an ontology. The authors applied graph structure based on domain ontology to extract semantic annotation and augmented contextual description using linked data. Li et al. (Li et al., 2020) introduced a bag of concepts model to represent documents using external knowledge sources. The authors also performed clustering for concept sense disambiguation.

Liu et al. (Liu et al., 2022) presented a crawler framework based on web text and the structure of the web links, to optimize weight assignment to hyperlinks and other web resources. Further, the framework involves an ontology and formal concept analysis for concept extraction and accumulation of related web resources. Bertola and Patti (Bertola and Patti, 2016) presented sentiment analysis based on visitor's comment tags for artwork on social networking sites. The authors created an ontology based categorization of opinions. This model provides interoperability using linked open data platforms to test semantic applications for the Italian language dataset.

Li et al. (Li et al., 2022) introduced a fuzzy theory based knowledge representation and reasoning approach for knowledge graphs. The authors applied a fuzzy extension of inference rules for knowledge graphs to discover more knowledge from external sources. Dourado, Pedronette, and da Silva Torres (Dourado, Pedronette, and da Silva Torres, 2019) proposed an unsupervised technique for rank aggregation from multiple rankers. The authors applied graph based approach for merging and unification of ranks and to find inter-relationship information among them. Further, a graph based contextual semantic similarity metric for retrieving results was also put forward.

Dhayne et al. (Dhayne et al., 2021) introduced a framework for the integration of clinical trial data and patient healthcare data to create a knowledge base for medical research. The proposal used SNOMED-CT as an external source, machine learning and NLP methods to extract concepts from clinical and patient data. Further, these concepts from two datasets were mapped by using neural word embeddings. Lee, Wang, and Trappey (Lee, Wang, and Trappey, 2015) presented an ontology based approach to handling and resolving consumer grievances with a case-based reasoning technique. The approach includes creating a

**Table 2**
Comparative analysis of reviewed document indexing frameworks.

| Framework | Basic Approach | Tools/Techniques used | Key features | Limitations |
|---|---|---|---|---|
| Rahimi, Montazeralghaem, and Shakery, 2020 | Free vocabulary | Probabilistic structured query, language modeling, axiomatic analysis for translation models in IR | Aggregation function for the frequency of a given query term, estimation of discrimination value function for cross language IR | Effects on various factors on document ranking by modeling of constraints to be explored further |
| Wagenpfeil et al., 2021 | Free vocabulary | Machine learning, feature vectors, social media contents, graph codes | Optimum utilization of phone hardware | Performance depends upon annotations of multimedia contents |
| Gupta and Yadav, 2021 | Free vocabulary | Hidden Markov model, deep neural networks, Gaussian mixture model, TF-IDF | Wavelet tree indexing, needs less memory with less processing time | Not validated for big data processing environment |
| Correia et al., 2022 | Free vocabulary | NLP, conditional random field, BiLSTM | Largest corpus with Legal NER in Portuguese | Not fully automated annotations |
| Singh, Pannu and Malhi, 2021 | Free vocabulary | CNN, feature extraction, image classification | CNN based medical image classification | Lack multi model ML features |
| Hassani, Ershadi and Mohebi, 2022 | Free vocabulary | Unsupervised keyword extraction, statistical feature of audio signals and video frames, | Language independent algorithm | Complexity increases significantly with increase in number of features |
| Garg and Sharma, 2022 | Free vocabulary | Linguistic features, TF-IDF, count vectorizer, hash vectorizer | Machine learning and word embedding based feature extraction | Lack of semantic features |
| Guillon et al., 2021 | Free vocabulary | Knowledge base (KB) | Easy to access KB for bidding information | Redundant rules and codependent modules |
| Boukhari and Omri, 2020 | Controlled vocabulary | Description Logics, VSM, Medical Subject Headings (MeSH) | Identify morphological variations of terms | Inaccurate weight assignment for new biomedical terms |
| Jiang, 2020 | Controlled vocabulary | Wikipedia, WordNet, domain ontology, labeled dynamic semantic network | Wikipedia based semantics of new terms that are not included in ontology and WordNet | Performance compared only with Lucene (keyword based) |
| Subramaniam et al., 2021 | Controlled vocabulary | Modified firefly algorithm, ontology, fuzzy c-mean clustering, LDA | Hybrid method combining Latent Dirichlet Allocation and domain ontology | Limited by the degree of coverage of concepts in the domain ontology |
| Nentidis et al., 2020 | Controlled vocabulary | Latent and semantic feature, TF-IDF, multi label classification, MeSH, UMLS | modeling indexing into multi label classification | Weak supervision method employed to train annotator |
| Bertola and Patti, 2016 | Controlled vocabulary | Ontology sentiment analysis, NLP, Plutchik's circumplex model, linked open data | Ontology based sentiment analysis using user's feedback tags | System is designed specifically for Italian language |
| Li et al., 2022 | Controlled vocabulary | Knowledge Graph, fuzzy relations, RDF triples, fuzzy reasoning | Fuzzy semantic information representation and reasoning using Knowledge Graph | Computational complexity of the proposed system |
| Dhayne et al., 2021 | Controlled vocabulary | NLP, Neural networks, SNOMED-CT ontology, Vector Space Model | Conversion of unstructured electronic medical records to queryable data with semantic reasoning | Missing field errors |
| Li et al., 2020 | Controlled vocabulary | Domain ontology, concept score inverse soft document frequency | Bag of concept based representation of documents with semantic knowledge | Hierarchical relations between concepts were not considered |
| Liu et al., 2022 | Controlled vocabulary | Pareto optimal solution, Hyperlink analysis, domain ontology, formal concept analysis | Domain ontology and hyperlink structure based web crawler | Premature convergence of crawler |
| Vidal et al., 2014 | Controlled vocabulary | Linked data, domain ontology, content based graph filtering | Semantic annotation of unstructured documents, each relevant term is connected to sub graph of ontology | Filtering linked data is expensive |
| Dourado, Pedronette, and da Silva Torres, 2019 | Controlled vocabulary | Rank aggregation, fusion graph, ad hoc retrieval | Minimum common sub graph based similarity score, no hyperparameter is required | Issues with vector representation based on fusion graph |
| Lee, Wang, and Trappey, 2015 | Controlled vocabulary | Web ontology language, case based reasoning, KB | Construction of customer complaint ontology | Limited in coverage of complaint cases |
| Sharma, Gupta and Juneja, 2021 | Controlled vocabulary | Wikipedia, TF-IDF, POS tagging based TF-IDF, word embeddings | Multiple unsupervised feature based keyword extraction and indexing for unstructured documents | Issue in keywords with stop words and domain specific keywords |
| Cimiano and Völker, 2005 | Controlled vocabulary | Text2Onto, Statistical features, TF-IDF | Machine readable extraction of concepts | Relevance weight assignment issue with concept with multiple occurrences |
| Kang, Haghighi, and Burstein, 2014 | Controlled vocabulary | CFinder, Statistical and semantic features, POS tagging, modified TF-IDF | Unsupervised method, Domain ontology based key concept extraction | Relevance weight assignment issue with concept with multiple occurrences |

knowledge base by developing an ontology to represent, index and find customer complaints. A comparative analysis of reviewed literature is presented in Table 2.

## 3. Motivation and objectives

The proposed approach used hybrid method employing neural word embeddings and external sources to learn document annotation from unstructured text documents from scientific literature. The neural word embedding model employs a huge corpus to discover semantic properties among co-occurring words but ignores the semantic relationship structure between them. The use of domain ontology as an external source was intended to address this issue in the concept extraction task. Our comprehension of the relationships between two words is improved by these semantic relational structures found in the domain ontology. Our approach depends on the approximate matching of document terms to the concept in ontology, which enables the successful recognition of morphological variations of the word.

The following are the key contributions of our proposed approach.:

(1) This research work presented a novel hybrid approach to integrate two complementary approaches - machine learning and domain ontology for semantic document indexing.

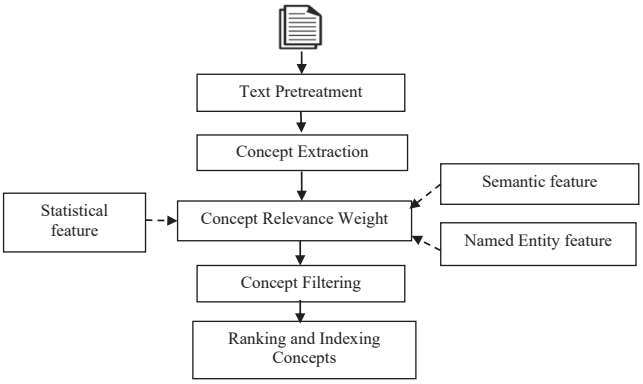


**Fig. 1.** Semantic document indexing based on neural word embeddings and domain ontology.

(2) This research work also proposed multiple features based novel concept ranking algorithm for unstructured text documents.
(3) When multiple terms are used to refer to the same concept in documents or domain ontology, then also presented technique extract such concepts accurately. Additionally, when the document terms have few or no associated concepts in the domain ontology, even then it offers a significant number of semantically related concepts.
(4) The proposed method used the word2vec model trained on more than four million scientific documents from computer science publications for the word embeddings task.
(5) The proposed method is analyzed by comparing its results with two state-of-the-art models on five benchmark datasets in the domain of information retrieval.

## 4. Machine learning and ontology-based novel semantic document indexing

This paper introduced machine learning and ontology-based novel hybrid semantic document indexing (MLOntoSDI) technique to address the information retrieval issues. Machine learning based semantics lacks hierarchical relational structures and results in less relevant concepts to be extracted for the indexing. While domain ontology is limited by degree of coverage of concepts and fails to extract sufficient concepts when the keyword is not covered in the ontology. The benefit of hybrid approach is that it possesses both corpus-based semantic features of machine learning as well as hierarchical relationship structures among concepts present in ontology for document indexing. The presented method used a neural word embedding based Skip-gram with negative sampling (SGNS) model (Mikolov et al., 2013) and a comprehensive standard ontology named computer science ontology (CSO) (Salatino et al., 2018) as an external resource to extract the promising features from the unstructured text.

The SGNS is a shallow neural network-based word embedding model which is highly efficient and lightweight. The pre-treatment of documents is the initial stage in the proposed indexing scheme. The concept extraction task is included in the second stage. The MLOntoSDI technique employs a fuzzy string match between document terms and concepts from external resources to account for morphological variances. Statistical, as well as semantic features along with domain-specific named entity features, were used to provide relevance scores to promising concepts. Due to fuzzy matches, the concept extraction stage may identify several concepts for each document. In the concept selection stage, the less significant concepts retrieved in the concept extraction step are eliminated and ranked according to relevance weight. The stages of the MLOntoSDI approach are shown in Fig. 1.

### 4.1. Pretreatment of text

The pretreatment stage for documents entails changing text to lower case, creating tokens, and eliminating unwanted features including stop words, HTML tags, punctuation, symbols, and special characters. The most frequent words that appear in text regularly and have no relevance in text representation are stop words. A parser based stemming technique and part-of-speech (POS) tagging were applied for text preprocessing (Loper and Bird, 2002). Some of the challenges faced during the document pre-processing involve:

- Choosing the right process (stemming or lemmatization) for getting the base form of domain-specific terms composing text documents.
- Developing lexical syntactic pattern for Identification of promising *n*-grams from documents that could be used as preferred terms or phrases in concept extraction stage.
- Recognizing variations of document keywords by considering fuzziness and vagueness in natural language text, without compromising accuracy.
- Identification of domain-specific named entities for part-of-speech tagging.

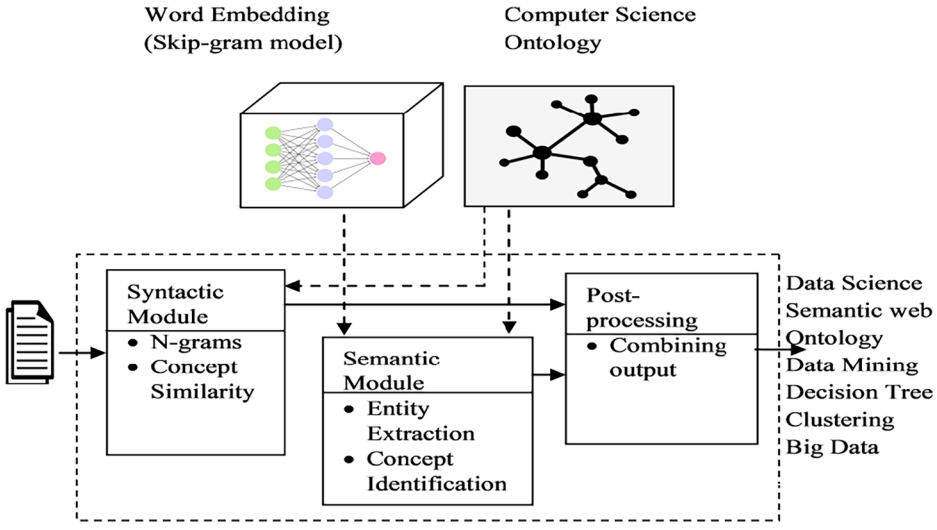


**Fig. 2.** Word Embeddings and domain ontology-based concept extraction (Salatino et al., 2019).

### 4.2. Concept extraction

This step identifies candidate concepts from unstructured text documents. The accuracy of the ontology concepts extracted from the document text dictates the effectiveness of the document indexing approach (Jiang and Tan, 2010). Fig. 2 shows the steps of the concept extraction process of the proposed indexing approach based on Salatino et al. (Salatino et al., 2019). This process includes syntactic and semantic feature based concept extraction. The syntactic module finds concepts that are explicitly referenced in a document. The syntactic module takes the pre-processed text and extracts the domain ontology concepts that are explicitly present in the text. To identify *n*-grams that my result into candidate concepts, lexical syntactic patterns of key phrases are identified by semantic module. Using neural word embedding based skip-gram with negative sampling (SGNS), words that are semantically related to these *n*-grams are extracted. The concept and its hypernyms are obtained by mapping these semantically related *n*-grams onto an ontology. Finally, the outcomes from the syntactic and semantic modules are integrated to get a comprehensive list of document concepts.

#### 4.2.1. Syntactic feature based concept extraction

The document text is first subjected to NLP pre-processing algorithms, and *n*-grams are produced. These *n*-grams from the texts are mapped to ontology concepts in the syntactic feature based concept extraction. The fuzzy similarity between *n*-grams and ontology concepts is measured using *Levenshtein* Similarity. From the list of extracted concepts, ontology concepts with similarity equal to or greater than a threshold (*minRel*) are considered. Algo. 1 shows pseudo-code for syntactic feature based concept extraction.

Algorithm 1
Semantic feature based document annotation.

**Input:** Domain ontology *Onto*, document *d*, minimum similarity threshold *minRel*
**Output:** Set of concepts *Koncept_list*

```
for each document d do
    perform preprocessing(d)
    nGrams = generate nGrams(d)        /* generating unigram, bigram, trigram */
end for
for each nGram in nGrams do
    mapping nGram to onto_concepts concepts in domain ontology
                                       /* concept mapping from ontology */
    if similarity(nGram, onto_concepts, 'levenshtein_similarity') ≥ minRel then
        add onto_concepts to Koncept_list
    end if
end for
return Koncept_list
```

#### 4.2.2. Semantic feature based concept extraction

To uncover concepts that are conceptually associated to documents but may not be explicitly addressed in them, semantic feature based concept extraction is applied. A specific lexical syntactic pattern of *n-grams*, such as a noun or adjective preceded by one or multiple nouns, is identified for concept identification. To determine the semantic similarity between document *n-grams* and ontology items, the SGNS-based word embedding was applied. For avoiding combinations of *n-grams* that may result in false positive concepts, the document *n-grams* with a specific lexical syntactic pattern are considered to identify candidate concepts. Eq (1) is used to discover these patterns using grammar.

$$< ADJ.* > * < NOUN.* > + \quad (1)$$

where adjectives are indicated by ADJ and nouns by NOUN. These learnt key phrases are converted to unigrams, bigrams, and trigrams and having cosine similarity equal to or better than 0.7 were used to extract their semantically related terms from the SGNS model. In order to identify a similar word in the SGNS model, the gap between bigram and trigram token words is substituted with the symbol "underscore." The average of the embedded vectors for all the words in the SGNS model's vocabulary is used when *n*-grams are not present in the vocabulary. Next, concepts and their hypernyms are extracted by mapping all *n*-grams and associated terms to concepts in the domain ontology (Salatino et al., 2018). Algo. 2 demonstrates pseudo code of Semantic feature based concept extraction. When multiple *n*-grams refer to the same ontology concept or when the same *n*-gram appears multiple times in a document's content, an ontology concept may be mentioned more than once. So, concepts are given weights based on their general relevance to the document. Eq (3) is used to determine a concept's significance to a document.

The value of *minRel* was empirically set to 0.94. When mapping document keywords to ontology concepts, multiple variations of ontology concepts, plurals, and hyphens between words may be handled by an optimal value of *minRel*. For example, variations between the terms "hierarchical clustering" and "hierarchical-clustering", and "hierarchical clustering" and "hierarchical clusterings". The $minRel = 1$ considers only the exact match between document keywords and ontology concepts. A value of *minRel* slightly lower than one allows partial match by considering more vagueness or fuzziness in concept matching.

### 4.3. Concept weight assignment

In this step, a concept's relevance to a document is represented by assigning its relevance weight based on its statistical, semantic, and scientific named entity properties. To determine the statistical feature weight of a concept, the Okapi BM25 approach, a variation of the term frequency inverse document frequency (TFIDF) was applied (Jiménez et al., 2018).

Algorithm 2
Semantic feature based document annotation.

**Input:** Domain ontology *Onto*, document *d*, Skip gram model *SGNS*, minimum similarity threshold *minRel*
**Output:** Set of concepts *Koncept_list*

```
for each document d do
    POS_tags = assign POS_tags(d)                    // part-of-speech tagging
    tokens = tokenization(POS_tags, 'lexical_syntactic_pattern')
end for
for each token in tokens do
    nGrams = generate nGrams(token, 1, 3)
    for each nGram in nGrams do
        keyphrase = join(nGram, '_')
        if keyphrase found_in SGNS then
            related_terms = SGNS.similar(keyphrase, topN = 15, min_cosine_sim = 0.7)
        else
        word_embedd = [ ]
        for each related_term in nGram do
            word_embedd.append(SGNS[related_term])
        end for
        new_embed = mean(word_embedd)
        related_terms = SGNS.similar(new_embed, topN = 15, min_cosine_sim = 0.7)
        end if
        for each term in related_terms do
            mapping term to onto_concepts concepts in the domain ontology
            relevance = Similarity(term, onto_concept, 'levenshtein_similarity')
            if relevance ≥ minRel then
                append onto_concept to Koncept_list
                append Super_Concept_Of(onto_concept) to Koncept_list
            end if
        end for
    end for
end for
return Koncept_list
```

The Okapi BM25 is a successful approach for document representation, but it was unable to capture the semantic context of the document's terms. By deriving semantic information from text, ontology-based approaches address the semantic context of a document (Kumar et al., 2013). The idea of NER has also proved successful in improving the effectiveness of IR systems (Goyal, Gupta, and Kumar, 2018). To obtain the final concept weight, the weights from these three characteristics were linearly integrated. For a concept $k$ the relevance score *Relwt* in corpus $C$ is calculated as follows:

$$Relwt(k, C) = \sum_{d \in C} Relwt(k, d) \tag{2}$$

A concept's relative importance *Relwt* in document $d$ is specified as follows:

$$Relwt(k, d) = \delta RelStat(k, d) + \rho RelSem(k, d) + \lambda RelSNEwt(k) \tag{3}$$

where $\delta$, $\rho$, and $\lambda$ are the parameters for feature weights, and these parameters depend on the problem in consideration. The method to calculate these parameters is illustrated in Section 5. The expression $RelStat(k, d)$ denotes statistical feature weight of a concept $k$ in document $d$ as follows:

$$RelStat(k, d) = BM25_{Score}(k, d) \tag{4}$$

where $BM25_{Score}(k, d)$ (Jiménez et al. 2018) denotes weight of concept $k$ in document $d$ as follows:

$$BM25_{Score}(k, d) = cFreq^* \left[ \frac{\log\left(\frac{D-dk+0.5}{dk+0.5}\right)}{kFreq + \psi^*\left(\left(1-\tau\right) + \tau \frac{DocLen}{AvgDocLen}\right)} \right] \tag{5}$$

with *cFreq* is the number of times concept $k$ appears in doc $d$ and D is the number of documents in dataset. *dk* is the number of documents having concept $k$. *DocLen* represents length of documents (number of concepts). *AvgDocLen* denotes average length of documents. The parameters $\psi$ and $\tau$ are normalization factors and the values of these parameters was set by experimentation as $\psi \in [1.2, 2]$ and $\tau = 0.75$. The expression $RelSem(k, d)$ represents semantic feature weight of a concept $k$ in document $d$ as follows (Salatino et al., 2019):

$$RelSem(k, d) = (\#KOnto)^*(\#KnGramOnto) \tag{6}$$

where the term $\#KOnto$ is the number of times concept $k$ is discovered in the CSO and $\#KnGramOnto$ denotes number of distinct $n$-grams mapping concept $k$ corresponding to document $d$. The term $RelSNE(k)$ represents scientific named entity feature weight of a concept $k$ and computed based on whether the concept $k$ is present as a named entity in the domain or not (Luan et al., 2018). A score of 0.5 is given if concept $k$ represents a named entity in the considered domain; otherwise, a score

of 0.5 is given.

$$RelSNE(k) = \begin{cases} 0.5; k \in SNE\_List \\ 0; Otherwise \end{cases} \quad (7)$$

### 4.4. Concept selection

In the process of extracting concepts from a text, many concepts that are only vaguely associated to the document may be found. The weight is given to the concept in order to determine how relevant it is to the document and is computed by Eq. (3). In this process, a concept's weight is set to the highest value if it is specifically mentioned in the document. Finally, the algorithm ranks the extracted concepts using their relevance weight and the top n concepts are returned for each document. To create the final document index, concepts are sorted based on their relevance weight. Algo. 3 demonstrated the steps of the proposed indexing approach.

Algorithm 3
Neural word embedding and ontology based semantic document indexing.

**Input:** Domain ontology *Onto*, dataset *docSet*, minimum relevance threshold *minRel*
**Output:** Semantic document index *DocIndex*

```
for each document d in docSet do
    extracted_koncept_list = generate syntactic feature based document annotations
                                                              //using Algo. (1)
    extracted_koncept_list + = generate semantic feature based document annotations
                                                              //using Algo. (2)
    for each extracted_koncept in extracted_koncept_list do
        calculate statistical feature weight                  //using Eq (4)
        calculate semantic feature weight                     //using Eq. (5)
        calculate scientific named entity feature weight      //using Eq. (6)
        calculate RelevanceScore                              //using Eq. (3)
        if RelevanceScore(extracted_koncept, d) ≥ minRel then
            add extracted_koncept to extracted_koncept_list
        end if
    end for
    DocIndex = sort(extracted_koncept_list, RelevanceScore)
end for
return DocIndex
```

## 5. Evaluation

The MLOntoSDI indexing scheme employs a skip-gram with negative sampling (SGNS) model for word embedding. The SGNS model allows us to only modify a small percentage of the weights, rather than all of them for each training sample. The SGNS model's execution time is affected by the size of the training corpus and grows linearly as more data is employed to train the model. The vocabulary size $v$ is not a significant determinant. The proposed indexing scheme's time complexity is $O(n * \log(v))$, where $n$ is the size of the entire corpus and $v$ is the number of unique words in the vocabulary as the implementation of the proposed indexing uses a binary search over the vocabulary size to achieve the sampling of negative examples.

### 5.1. Evaluation benchmark

A collection of research articles in the domain of computer science and library science were chosen for the empirical evaluation of the MLOntoSDI method, and its performance was assessed using standard evaluation measures in the field of IR. Five benchmark datasets namely CACM, CISI, LISA (University of Glasgow, 2020), KDD (Caragea et al., 2014), and Inspec (Hulth, 2003) were used in evaluation process.

The Communications of the ACM (CACM) journal's 3204 scholarly articles, together with their titles, abstracts, and author information, are included in the CACM dataset. 1,460 scholarly papers in the field of library science make up the CISI dataset compiled by the University of Glasgow's (UofG) Centre for Inventions and Scientific Information (CISI). The UofG IR Group compiled the LISA dataset, which includes 6004 documents that comprise abstracts of library and information science. A collection of 834 knowledge discovery and data mining abstracts from the ACM conference make up the KDD dataset, while 500 scholarly articles with abstract from the computer & control and IT domain are included in the Inspec dataset. A list of related terms assigned by domain experts is provided with each dataset. Table 3 throws light on the description of each dataset.

### 5.2. Evaluation measures

The performance of any information retrieval system is measured using standard evaluation metrics. Five evaluation metrics, namely precision, precision at $x$, recall, mean average precision and F-measure are employed for comparative analysis of the MLOntoSDI with state-of-the-art models. The percentage of concepts returned that are actually relevant to the document is known as precision (P). The percentage of relevant concept returned by an indexing technique is known as the recall (R). F-measure (Fm), which combines accuracy and recall into a single metric, is the harmonic mean of these two metrics. The gold standard used to assess the proposed system is the set of annotated concepts with which a domain expert annotated the documents.

Let *#RelCon* denotes the total count of relevant concepts fetched and *#RtrConc* represents total count of retrieved concepts by the proposed system, while *#TotRelCon* represents the total number of relevant concepts assigned to a document by the gold standard. Precision at $x$ ($P@x$, $x = 10$, 20, and 50) relates to the accuracy of the top $x$ concepts returned by the system and investigated by the user. The proposed system performances are measured as follows as shown in Eq. (8), (9), (10) and (11) for precision, precision at $x$, recall and F-measures respectively.

**Table 3**
Description of benchmark for evaluation.

| | Dataset | | | | |
|---|---|---|---|---|---|
| | **CACM** | **CISI** | **LISA** | **KDD** | **Inspec** |
| Document Type | Computer Science | Library Science | Library and Information Science | Computer Science | Computers and Controls, IT |
| No. of Documents | 3204 | 1460 | 6004 | 834 | 500 |
| No. of Words | 368,460 | 166,440 | 360,240 | 83,230 | 60,110 |

**Table 4**
Tuning hyper parameters for the word embedding model.

| **Hyper-parameter** | Embedding size | Context window size | Negative sampling | Maximum iteration | Minimum count cut-off |
|---|---|---|---|---|---|
| **Value** | 128 | 10 | 5 | 5 | 10 |

$$P = \frac{\#RelCon}{\#RtrCon} \tag{8}$$

$$P@x = \frac{\#RelConatrankx}{\#RtrConatrankx} \tag{9}$$

$$R = \frac{\#RelCon}{\#TotRelCon} \tag{10}$$

$$Fm = 2^{*}\frac{P^{*}R}{P+R} \tag{11}$$

Mean Average Precision (*MeanAvgP*) being a stable metric encapsulates the system's effectiveness in terms of precision, recall and ranking. Average precision for a single user query is calculated by averaging the precision values for the top x concepts after each relevant concepts has been retrieved, and further averaged again by the total documents count. *P@x* is the precision of rank-ordered concepts chosen at rank *x*, *Rel@x* is either 1 or 0 depending on whether the chosen concept at rank x is part of the gold standard or not. While *#RelCon* denotes the number of relevant concepts returned for document doc, α is the maximum number of returned concepts evaluated by the gold standard, and #DocSet is the number of documents in dataset.

$$MeanAvgP = \frac{1}{\#RelCon}\sum_{l=1}^{\#DocSet}\frac{1}{\min(\#RelCon,\alpha)}\sum_{x=1}^{\alpha}(P@x.Rel@x) \tag{12}$$

### 5.3. Hyperparameter computation for word embedding model

The authors pre-processed the training dataset from Microsoft Academic Graph by replacing *n*-grams mapping to the concept in domain ontology with underscores symbol (for example, "document clustering" became "document_clustering" and "neural word embeddings" became "neural_word_embeddings") (Mikolov et al., 2013). The quality of output from a machine learning model is significantly influenced by hyper-parameter tuning. The SGNS model's hyper-parameters were empirically set in the proposed work as shown in Table 4.

### 5.4. Parameters computation for assigning concept weight

The analytic hierarchy process (AHP) was considered for calculating the parameter's weight for concept relevance due to the suitability for modelling the nature of problem. In order to deal with the uncertainty involved in the problem under consideration, we also explored the fuzzy analytic hierarchy process (FAHP) for the same and found that it is more effective for calculating parameters weights in the proposed work. By applying FAHP, average precision and recall values were improved by 2.65 % and 1.85 % respectively as compared to the results on same metrics when parameters weights were calculated using AHP.

The FAHP is a relative measurement-based multiple-criteria decision analysis (MCDA) approach for comparing alternative solutions (Xu and Liao, 2013). In Eq. (13), a hierarchical structure of weights is established, and the values of the matrix's parameters are determined by a domain expert based on the significance of the various criteria used to calculate the concept weights. In this study, scientific named entity features account for 1/5 of statistical features whereas semantic features account for 1/3 of statistical features of the concept. Eq. (13) and (14) displays the pairwise comparison matrix for the AHP and fuzzy AHP respectively.

$$\begin{array}{c} \\ \delta \\ \rho \\ \lambda \end{array}\begin{array}{c} \begin{array}{ccc} \delta & \rho & \lambda \end{array} \\ \begin{pmatrix} 1 & 1/3 & 1/5 \\ 1/3 & 1 & 1/5 \\ 1/5 & 1/3 & 1 \end{pmatrix} \end{array} \tag{13}$$

$$\begin{array}{c} \\ \delta \\ \rho \\ \lambda \end{array}\begin{array}{c} \begin{array}{ccc} \delta & \rho & \lambda \end{array} \\ \begin{pmatrix} (1,1,1) & (2,3,4) & (4,5,6) \\ (1/4,1/3,1/5) & (1,1,1) & (2,3,4) \\ (1/6,1/5,1/4) & (1/4,1/3,1/2) & (1,1,1) \end{pmatrix} \end{array} \tag{14}$$

Using Eq. (13) and (14), the values of δ, ρ and λ computed using AHP and FAHP are demonstrated in Table 5 and 6 respectively.

## 6. Results and discussions

The MLOntoSDI method is evaluated for the document indexing task on five benchmark datasets. The proposed method was compared with statistical and external knowledge-based two state-of-art methods namely Text2Onto (Cimiano and Völker, 2005) and CFinder (Kang, Haghighi, and Burstein, 2014). For comparative analysis, three versions of the proposed indexing scheme based on the types of features employed for the concept extraction task were taken into consideration. The first variation of the proposed approach SyntacticMLOnto solely used the syntactic features and the second version SemanticMLOnto only utilized the semantic features, while the proposed method MLOntoSDI utilized both the syntactic and semantic features for concept extraction.

The precision and recall values for comparative algorithms on benchmark datasets are shown in Figs. 3 and 4. On these datasets, the MLOntoSDI method yields greater precision values than CFinder and Text2Onto. On the KDD, CACM, Inspec, LISA, and CISI datasets, the

**Table 5**
Analytic Hierarchy Process based parameters computation for assigning concept weight.

| **Parameter** | $\delta$ | $\rho$ | $\lambda$ |
|---|---|---|---|
| **Value** | 0.6369 | 0. 2582 | 0.1047 |

**Table 6**
Fuzzy Analytic Hierarchy Process based parameters computation for assigning concept weight.

| **Parameter** | $\delta$ | $\rho$ | $\lambda$ |
|---|---|---|---|
| **Value** | 0.6295 | 0. 2632 | 0.1073 |

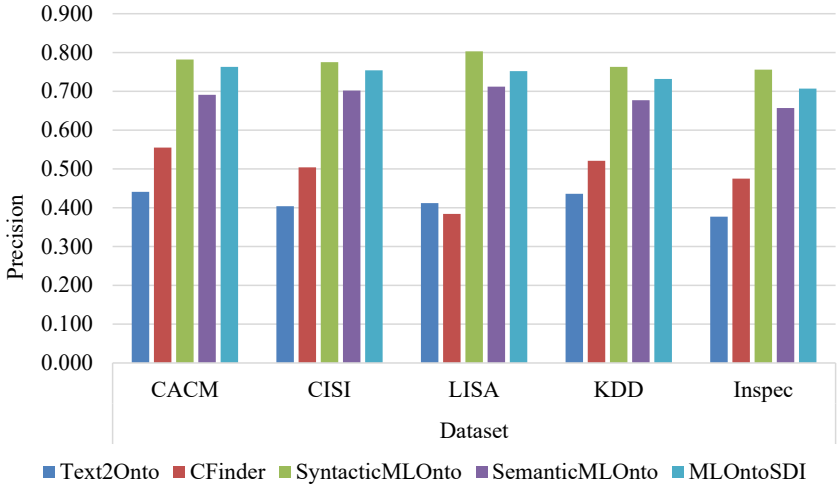


**Fig. 3.** Results of comparison on precision metric.

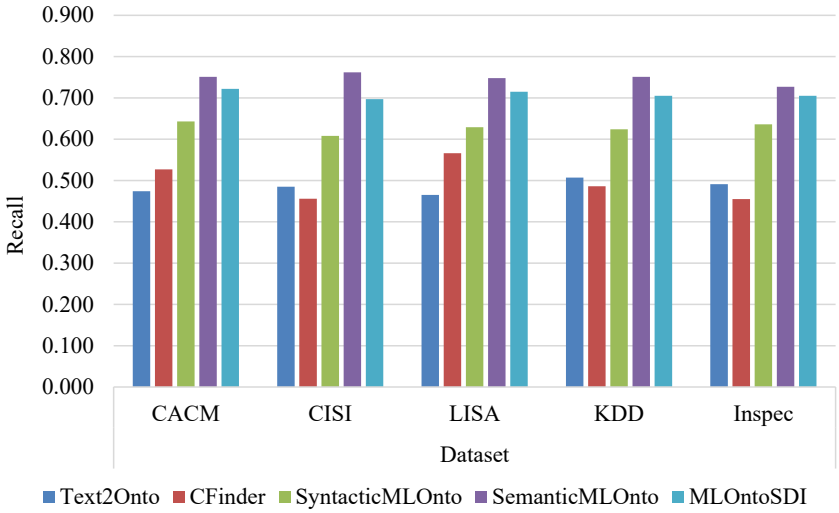


**Fig. 4.** Results of comparison on recall metric.

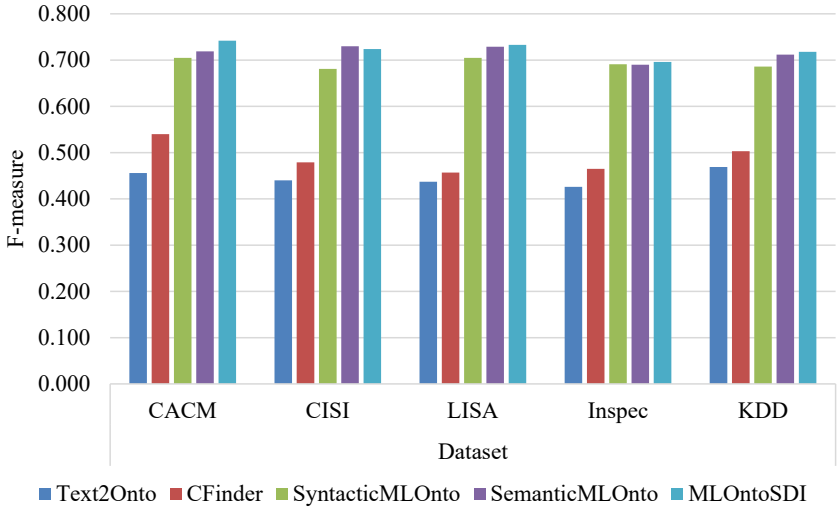


**Fig. 5.** Results of comparison on F-measure metric.

**Table 7**
Performance analysis on Precision@k.

| Metric | Text2Onto | CFinder | SyntacticMLOnto | SemanticMLOnto | MLOntoSDI |
|---|---|---|---|---|---|
| | **CACM** | | | | |
| P@20 | 0.45 | 0.55 | 0.80 | 0.75 | 0.77 |
| P@50 | 0.40 | 0.46 | 0.72 | 0.68 | 0.74 |
| P@100 | 0.38 | 0.44 | 0.70 | 0.64 | 0.70 |
| | **CISI** | | | | |
| P@20 | 0.45 | 0.55 | 0.75 | 0.71 | 0.74 |
| P@50 | 0.38 | 0.44 | 0.74 | 0.66 | 0.72 |
| P@100 | 0.37 | 0.42 | 0.70 | 0.65 | 0.68 |
| | **LISA** | | | | |
| P@20 | 0.45 | 0.60 | 0.80 | 0.70 | 0.70 |
| P@50 | 0.40 | 0.45 | 0.70 | 0.65 | 0.72 |
| P@100 | 0.38 | 0.43 | 0.71 | 0.66 | 0.66 |
| | **KDD** | | | | |
| P@20 | 0.47 | 0.57 | 0.81 | 0.76 | 0.82 |
| P@50 | 0.42 | 0.48 | 0.74 | 0.70 | 0.74 |
| P@100 | 0.38 | 0.45 | 0.70 | 0.68 | 0.69 |
| | **Inspec** | | | | |
| P@20 | 0.42 | 0.55 | 0.84 | 0.72 | 0.72 |
| P@50 | 0.38 | 0.45 | 0.70 | 0.69 | 0.69 |
| P@100 | 0.35 | 0.42 | 0.68 | 0.66 | 0.68 |

MLOntoSDI technique increased the precision roughly by 33 %, 32 %, 29 %, 34 % and 35 % percent as compared to CFinder and Text2Onto. On the KDD, CACM, Inspec, LISA, and CISI datasets, the MLOntoSDI technique outscored the CFinder and Text2Onto approaches on the recall by around 21 %, 24 %, 19 %, 25 % and 21 % respectively. On each of the five benchmark datasets, the F-measure for various approaches is compared in Fig. 5. On KDD, CACM, Inspec, LISA, and CISI datasets, respectively, the MLOntoSDI technique improved the F-measure by a margin of 27 %, 28 %, 24 %, 29 %, and 29 % as compared to CFinder and Text2Onto.

Due to a syntactic match between domain ontology concepts and document terms, SyntacticMLOnto achieved highest precision among all compared methods. Using this technique, it is possible to extract the concepts that are highly relevant and explicitly contained in the document. The accuracy of the MLOntoSDI is higher than that of the SemanticMLOnto but lower than that of the SyntacticMLOnto. It may be concluded that SemanticMLOnto extracts concepts that are semantically implied but not directly refenced in the document, however, SemanticMLOnto may produce false positives. Though it trailed SemanticMLOnto in recall value, MLOntoSDI showed somewhat higher performance than SyntacticMLOnto. Due to the inclusion of several concepts that are semantically similar, SemanticMLOnto underperformed SyntacticMLOnto on precision metric, although outperformed on recall and F-measure.

Results from Figs. 3–5 demonstrate that on all five datasets, the proposed technique outperformed Text2Onto and CFinder in terms of average accuracy measure. This performance is attributed to the proposed model's concept weight assignment step, which maintains the relevant concepts on higher rank, whereas CFinder and Text2Onto lacked to do so when a concept exists more than once in a document.

Concept weights in Text2Onto were determined based on TF-IDF and a concept that appears several times in a text and is related to the domain could be given less weight because of the way TF-IDF work. Concept weights in CFinder were determined using a modified TF-IDF. In this technique, during the concept supplementation phase, the adjectives are taken out of the concepts, creating repetitious concepts. As a result, giving these concepts the appropriate weight results in a concept ranking issue. The proposed technique, which learns word semantics using the SGNS neural word embeddings and ontology, outperformed Text2Onto and CFinder on recall metric.

Table 7 indicates that the proposed approach scored 0.77, 0.74, and 0.70 on the CACM dataset at P@20, P@50, and P@100 respectively. On the CISI dataset, the proposed approach achieved values of 0.74, 0.72, and 0.68 at P@20, P@50, and P@100, respectively. On LISA dataset, the proposed approach attained a score of 0.70, 0.72, and 0.66 at P@20, P@50, and P@100, respectively. On the KDD dataset, the suggested approach shows scores of 0.82, 0.74, and 0.69 at P@20, P@50, and P@100, respectively. The proposed approach attained a score of 0.72, 0.69 and 0.68 at P@20, P@50, and P@100 respectively on the Inspec dataset. It is included from Table 7, that the MLOntoSDI approach outperformed the Text2Onto on P@k metric approximately by 32.66 %, 32.66, 31.33 %, 28.33 % and 31.33 % on KDD, CACM, Inspec, LISA, and CISI datasets respectively. While the proposed approach improved P@k score by 25 %, 25.33 %, 22.33 %, 20 % and 24.33 % compared to CFinder on KDD, CACM, Inspec, LISA, and CISI datasets respectively. The outcomes demonstrate that the proposed technique outperformed CFinder and Text2Onto on the P@k measure across all five datasets.

Table 8 shows that the proposed approach showed a score of 0.77, 0.83, 0.75, 0.72 and 0.77 on MeanAvgP metric for KDD, CACM, Inspec, LISA, and CISI datasets respectively. Further, the comparative analysis on MeanAvgP metric showed that the proposed approach outperformed Text2Onto with a margin of 32 %, 27 %, 28 %, 25 % and 32 % on KDD, CACM, Inspec, LISA, and CISI datasets respectively. While the proposed approach improved MeanAvgP score by of 18 %, 19 %, 15 %, 17 % and 18 % compared to CFinder on KDD, CACM, Inspec, LISA, and CISI datasets respectively. The findings demonstrate that on all five datasets, the proposed technique outperformed CFinder and Text2Onto in terms of MeanAvgP performance.

**Table 8**
Results analysis on mean average precision.

| Dataset | Model | | | | |
|---|---|---|---|---|---|
| | Text2Onto | CFinder | SyntacticMLOnto | SemanticMLOnto | MLOntoSDI |
| CACM | 0.56 | 0.64 | 0.78 | 0.73 | 0.83 |
| CISI | 0.45 | 0.59 | 0.67 | 0.70 | 0.77 |
| LISA | 0.47 | 0.55 | 0.60 | 0.68 | 0.72 |
| KDD | 0.45 | 0.59 | 0.67 | 0.72 | 0.77 |
| Inspec | 0.47 | 0.60 | 0.65 | 0.71 | 0.75 |

The empirical evaluation revealed that the proposed approach performed better than CFinder and Text2Onto with 29 % improvement in average accuracy and F-measure was improved by 25 % on all five standard benchmark datasets related to scholarly publications in the computer science field. The improvement in average accuracy showed that the proposed indexing approach outperformed state-of-the-art techniques in extracting document concepts even when the same concept was referred by distinct words in the document and domain ontology. The proposed system's ability to find similar concepts when the documents possess no concept from domain ontology is demonstrated by the improvement in F-measure, which is attributed to the proposed scheme's high recall rates while maintaining high accuracy.

The central idea in the proposed approach is concept extraction which is mapping between document *n*-grams and ontology concepts. This mapping algorithm has linear time complexity. The processing time of the proposed algorithm depends on the size of the key phrase also called token and grows linearly with the number of *n*-grams in the key phrase. There are no additional requirements for concept extraction in the proposed indexing method. (how linear time complexity is impacting our proposed scheme).

The key phase in the query processing is query concept extraction using ontology. In the proposed approach, there is a mapping between query *n*-grams and ontology concepts. This mapping algorithm has linear time complexity. The processing time of this algorithm depends on the size of the query and grows linearly with the number of *n*-grams in the user query. Moreover, there are no additional requirements for query processing in the proposed indexing method.

## 7. Conclusions and future work

This research work presented a novel hybrid semantic indexing method for unstructured text documents by combining machine learning with domain ontology. The proposed approach's ability to recognize concepts that are semantically connected to document text was improved by using the machine learning-based skip-gram model. A fuzzy match between document words and ontology concepts is used to incorporate the morphological variations of concepts. The proposed method assigned weights to concepts by syntactic, semantic, and scientific named entity features and ranks concept based on relevance to the document. The empirical evaluation revealed that the proposed method performed better than the state-of-the-art CFinder and Text2Onto with 29 % on average accuracy and also improved F-measure by a margin of 25 % on all five standard benchmark datasets related to scholarly papers in the field of computer science. The only essential requirements for the proposed method are availability of a comprehensive domain ontology and neural word embedding model trained on domain corpus.

In further development, performance of the proposed indexing may be improved by tuning the hyperparameters of the skip-gram with negative sampling model using optimization techniques. For the goal of concept extraction, knowledge graph embedding may also be investigated. Further, the ontology from other domain may be applied to test the presented indexing scheme.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgement

The authors acknowledge that this research work has not received any funding.

## References

Abasi, A. K., Khader, A. T., Al-Betar, M. A., Naim, S., Alyasseri, Z. A. A., & Makhadmeh, S. N. (2021). An ensemble topic extraction approach based on optimization clusters using hybrid multi-verse optimizer for scientific publications. *Journal of Ambient Intelligence and Humanized Computing, 12*(2), 2765–2801. https://doi.org/10.1007/s12652-020-02439-4

Aman, M., Abdulkadir, S. J., Aziz, I. A., Alhussian, H., & Ullah, I. (2021). KP-Rank: A semantic-based unsupervised approach for keyphrase extraction from text data. *Multimedia Tools and Applications, 80*(8), 12469–12506. https://doi.org/10.1007/s11042-020-10215-x

Anand, S. K., & Kumar, S. (2022a). Uncertainty analysis in ontology-based knowledge representation. *New Generation Computing, 40*(1), 339–376. https://doi.org/10.1007/s00354-022-00162-6

Anand, S. K., & Kumar, S. (2022b). Experimental comparisons of clustering approaches for data representation. *ACM Computing Surveys, 55*(3), 1–33. https://doi.org/10.1145/3490384

Bertola, F., & Patti, V. (2016). Ontology-based affective models to organize artworks in the social semantic web. *Information Processing & Management, 52*(1), 139–162. https://doi.org/10.1016/j.ipm.2015.10.003

Bhunia, A. K., Roy, P. P., Sain, A., & Pal, U. (2020). Zone-based keyword spotting in Bangla and Devanagari documents. *Multimedia Tools and Applications, 79*(37), 27365–27389. https://doi.org/10.1007/s11042-019-08442-y

Bordoloi, M., Chatterjee, P. C., Biswas, S. K., & Purkayastha, B. (2020). Keyword extraction using supervised cumulative TextRank. *Multimedia Tools and Applications, 79*(41), 31467–31496. https://doi.org/10.1007/s11042-020-09335-1

Boukhari, K., & Omri, M. N. (2020). DL-VSM based document indexing approach for information retrieval. *Journal of Ambient Intelligence and Humanized Computing, 1–12*. https://doi.org/10.1007/s12652-020-01684-x

Caragea, C., Bulgarov, F., Godea, A., & Gollapalli, S. D. (2014, October). Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1435-1446). http://doi.org/10.3115/v1/D14-1150.

Chinnasamy, P., & Deepalakshmi, P. (2022). HCAC-EHR: Hybrid cryptographic access control for secure EHR retrieval in healthcare cloud. *Journal of Ambient Intelligence and Humanized Computing, 13*(2), 1001–1019. https://doi.org/10.1007/s12652-021-02942-2

Cimiano, P., & Völker, J. (2005). A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)* (pp. 227-238). 10.1007/11428817_21.

Correia, F. A., Almeida, A. A., Nunes, J. L., Santos, K. G., Hartmann, I. A., Silva, F. A., et al. (2022). Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court. *Information Processing & Management, 59*(1), Article 102794. https://doi.org/10.1016/j.ipm.2021.102794

Dhayne, H., Kilany, R., Haque, R., & Taher, Y. (2021). EMR2vec: Bridging the gap between patient data and clinical trial. *Computers & Industrial Engineering, 156*, Article 107236. https://doi.org/10.1016/j.cie.2021.107236

Dourado, I. C., Pedronette, D. C. G., & da Silva Torres, R. (2019). Unsupervised graph-based rank aggregation for improved retrieval. *Information Processing & Management, 56*(4), 1260–1279. https://doi.org/10.1016/j.ipm.2019.03.008

Garg, S., & Sharma, D. K. (2022). Linguistic features based framework for automatic fake news detection. *Computers & Industrial Engineering, 108432*. https://doi.org/10.1016/j.cie.2022.108432

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review, 29*, 21–43. https://doi.org/10.1016/j.cosrev.2018.06.001

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., et al. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management, 57*(6), Article 102067. https://doi.org/10.1016/j.ipm.2020.102282

Guillon, D., Villeneuve, E., Merlo, C., Vareilles, É., & Aldanondo, M. (2021). ISIEM: A methodology to deploy a knowledge-based system to support bidding process. *Computers & Industrial Engineering, 161*, Article 107638. https://doi.org/10.1016/j.cie.2021.107638

Gupta, A., & Yadav, D. (2021). A novel approach to perform context-based automatic spoken document retrieval of political speeches based on wavelet tree indexing. *Multimedia Tools and Applications, 80*(14), 22209–22229. https://doi.org/10.1007/s11042-021-10800-8

Hammache, A., & Boughanem, M. (2021). Term position-based language model for information retrieval. *Journal of the Association for Information Science and Technology, 72*(5), 627–642. https://doi.org/10.1002/asi.24431

Hassani, H., Ershadi, M. J., & Mohebi, A. (2022). LVTIA: A new method for keyphrase extraction from scientific video lectures. *Information Processing & Management, 59*(2), Article 102802. https://doi.org/10.1016/j.ipm.2021.102802

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 216-223). 10.3115/1119355.1119383.

Jiang, X., & Tan, A. H. (2010). CRCTOL: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology, 61*(1), 150–168. https://doi.org/10.1002/asi.21231

Jiang, Y. (2020). Semantically-enhanced information retrieval using multiple knowledge sources. *Cluster Computing, 23*(4), 2925–2944. https://doi.org/10.1007/s10586-020-03057-7

Jiménez, S., Cucerzan, S., González, F. A., Gelbukh, A. F., & Dueñas, G. (2018). BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies. *Journal of Intelligent & Fuzzy Systems, 34*(5), 2887–2899.

Kang, Y. B., Haghighi, P. D., & Burstein, F. (2014). CFinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications, 41*(9), 4494–4504. https://doi.org/10.1016/j.eswa.2014.01.006

Kumar, S., Singh, M., & De, A. (2012, December). OWL-based ontology indexing and retrieving algorithms for Semantic Search Engine. In 2012 7th International Conference on Computing and Convergence Technology (ICCCT) (pp. 1135-1140). IEEE.

Kumar, S., Kumar, N., Singh, M., & De, A. (2013). A Rule-based approach for extraction of link-context from anchor-text structure. In Intelligent Informatics (pp. 261-271). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32063-7_28.

Lee, C. H., Wang, Y. H., & Trappey, A. J. (2015). Ontology-based reasoning for the intelligent handling of customer complaints. *Computers & Industrial Engineering, 84*, 144–155. https://doi.org/10.1016/j.cie.2014.11.019

Li, P., Mao, K., Xu, Y., Li, Q., & Zhang, J. (2020). Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. *Knowledge-Based Systems, 193*, Article 105436. https://doi.org/10.1016/j.knosys.2019.105436

Liu, J., Li, X., Zhang, Q., & Zhong, G. (2022). A novel focused crawler combining Web space evolution and domain ontology. *Knowledge-Based Systems, 243*, Article 108495. https://doi.org/10.1016/j.knosys.2022.108495

Li, P., Wang, X., Liang, H., Zhang, S., Zhang, Y., Jiang, Y., et al. (2022). A fuzzy semantic representation and reasoning model for multiple associative predicates in knowledge graph. *Information Sciences, 599*, 208–230. https://doi.org/10.1016/j.ins.2022.03.079

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. In *Proc. ETMTNLP*. (pp. 63–70), Pennsylvania, USA. 10.3115/1118108.1118117.

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proc. EMNLP*. (pp. 3219–3232). Belgium. http://doi.org/10.18653/v1/D18-1360.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. (pp. 3111–3119).

Nentidis, A., Krithara, A., Tsoumakas, G., & Paliouras, G. (2020). Beyond MeSH: Fine-grained semantic indexing of biomedical literature based on weak supervision. *Information Processing & Management, 57*(5), Article 102282. https://doi.org/10.1016/j.ipm.2020.102282

Prasanth, T., & Gunasekaran, M. (2019). Effective big data retrieval using deep learning modified neural networks. *Mobile Networks and Applications, 24*(1), 282–294. https://doi.org/10.1007/s11036-018-1204-y

Rahimi, R., Montazeralghaem, A., & Shakery, A. (2020). An axiomatic approach to corpus-based cross-language information retrieval. *Information Retrieval Journal, 23*(3), 191–215. https://doi.org/10.1007/s10791-020-09372-2

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., & Motta, E. (2018, October). The computer science ontology: a large-scale taxonomy of research areas. In *International Semantic Web Conference* (pp. 187-205). Springer, Cham. https://doi.org/10.1007/978-3-030-00668-6_12.

Salatino, A. A., Osborne, F., Thanapalasingam, T., & Motta, E. (2019, September). The CSO classifier: Ontology-driven detection of research topics in scholarly articles. In *International Conference on Theory and Practice of Digital Libraries* (pp. 296-311). Springer, Cham. https://doi.org/10.1007/978-3-030-30760-8_26.

Shanmuganathan, V., Yesudhas, H. R., Madasamy, K., Alaboudi, A. A., Luhach, A. K., & Jhanjhi, N. Z. (2021). AI Based Forecasting of Influenza Patterns from Twitter Information Using Random Forest Algorithm. *Human-centric Computing and Information Sciences*, *11*:*33*, 1-14. 10.22967/HCIS.2021.11.033.

Shokat, S., Riaz, R., Rizvi, S. S., Abbasi, A. M., Abbasi, A. A., & Kwon, S. J. (2020). Deep learning scheme for character prediction with position-free touch screen-based Braille input method. *Human-centric Computing and Information Sciences, 10*(1), 1–24. https://doi.org/10.1186/s13673-020-00246-6

Sharma, S., Gupta, V., & Juneja, M. (2021). Diverse feature set based Keyphrase extraction and indexing techniques. *Multimedia Tools and Applications, 80*(3), 4111–4142. https://doi.org/10.1007/s11042-020-09423-2

Singh, S. K., Cha, J., Kim, T. W., & Park, J. H. (2021). Machine learning based distributed big data analysis framework for next generation web in IoT. *Computer Science and Information Systems, 18*(2), 597–618. https://doi.org/10.2298/CSIS200330012S

Singh, A., Pannu, H. S., & Malhi, A. (2022). Explainable information retrieval using deep learning for medical images. *Computer Science and Information Systems, 19*(1), 277–307. https://doi.org/10.2298/CSIS201030049S

Spolaôr, N., Lee, H. D., Takaki, W. S. R., Ensina, L. A., Parmezan, A. R. S., Oliva, J. T., et al. (2021). A video indexing and retrieval computational prototype based on transcribed speech. *Multimedia Tools and Applications, 80*(25), 33971–34017. https://doi.org/10.1007/s11042-021-11401-1

Subramaniam, M., Kathirvel, A., Sabitha, E., & Basha, H. A. (2021). Modified firefly algorithm and fuzzy C-mean clustering based semantic information retrieval. *Journal of Web Engineering, 20*(1), 33–52. https://doi.org/10.13052/jwe1540-9589.2012

Ullah, I., Khusro, S., & Ahmad, I. (2021). Improving social book search using structure semantics, bibliographic descriptions and social metadata. *Multimedia Tools and Applications, 80*(4), 5131–5172. https://doi.org/10.1007/s11042-020-09811-8

University of Glasgow (UofG), (2020). Information Retrieval Test Collections. Available at http://ir.dcs.gla.ac.uk/resources/test_collections/. Accessed (15.08.2021).

Upadhyay, P., Bedathur, S., Chakraborty, T., & Ramanath, M. (2020, April). Aspect-based academic search using domain-specific KB. In *European Conference on Information Retrieval* (pp. 418-424). Springer, Cham. 10.1007/978-3-030-45442-5_52.

Vidal, J. C., Lama, M., Otero-García, E., & Bugarín, A. (2014). Graph-based semantic annotation for enriching educational content with linked data. *Knowledge-Based Systems,, 55*, 29–42. https://doi.org/10.1016/j.knosys.2013.10.007

Wagenpfeil, S., Engel, F., Kevitt, P. M., & Hemmje, M. (2021). Ai-based semantic multimedia indexing and retrieval for social media on smartphones. *Information*, *12*(1), 43:1-30.

Xu, Z., & Liao, H. (2013). Intuitionistic fuzzy analytic hierarchy process. *IEEE Transactions on Fuzzy Systems,, 22*(4), 749–761. https://doi.org/10.1109/TFUZZ.2013.2272585