Concept2Vec: concept vector generation for biomedical literature using concept modelling

Suneetha Vazrala, Thayyaba Khatoon Mohammed
Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India

Article Info	ABSTRACT
	In the biomedical field due to tremendous medical research, every year the medical papers generated is very huge. In order to process these tremendous amounts of textual data, the information retrieval techniques are combined with machine learning methods like natural language processing (NLP) for finding the useful insights in data for NLP tasks. In this work, we have addressed the challenges involved in existing NLP techniques like Word2Vec, Doc2Vec, and Biosent2Vec and followed by the traditional classification of articles with Concept2Vec and random forest classification. Concept vector techniques are emerging methods in NLP and in information retrieval systems. Pretrained concept encoders are available in Machine learning, whereas no such exists in biomedical domain. In this paper a proposal on Concept2Vec trained from biomedical text is highlighted. Our work accurately emphasizes on the quality of semantics defined between the medical terms, which can help the medical practitioners in taking errorless decisions. In this work 20 thousand biomedical documents are considered for Concept2Vec embeddings and the proposed approach have shown best concept similarity in terms of relevant measures compared with existing approaches such as Word2Vec, global vectors (GloVe), fastText, and BioSentVec, bidirectional encoder representations from transformers (BERT), biomedical-BERT.

Corresponding Author:

Suneetha Vazrala
Department of Computer Science and Engineering, Malla Reddy University
Hyderabad, Telangana, India
Email: suneetha.vazrala@gmail.com

1. INTRODUCTION

In machine learning the latest advances in human replacement demands the need of a machine to understand the natural language. In this context language grams always played an important role. Language modelling has become the need of the hour in natural language processing by machines [1]. With the invention of language models, the machine is able to predict the next sequence of the text in sentence formation, the probability of the word sequences and topic sequences in recommendation system [2]. Language models are used in many of the machine learning applications such as sentence tagging, language translation, voice recognition, and question and answering systems, keyword search mechanisms, text processing (voice-text and text-voice) [3]. The language models vary depends on the machine application and generalised as two types.

Statistical machine language models: statistical WordNet defines the probabilistic relations between the words and the next sequence of the word [4]. It defines a chain rule method for generating the sequence. The probability of sequence of terms defined as shown in (1).

$$P\ (w1w2w3w4)\ = p\ (w1)\ p\ (w2|w1)\ p\ (w3|w1w2)\ p\ (w4|w1w2w3) \tag{1}$$

The co-occurrence probability is defined from Global Statistics as P (j/i) i.e., the number of times the word j occurs in the context of i [17]. Over the years the learning vector space word representations such as Word2Vec, fastText and GloVe has added a prominence over better-quality feature depiction in bagging the biomedical domain knowledge and also in natural language processing tasks [18]. Most of these techniques failed to capture the deep contextual relationships of the terms present in documents, so a novel context-based word embedding model was defined on one of the largest biomedical dataset PubMed, as the PubMed data set article statistics is mentioned in Table 1. To perform some of the applications in biomedicine like drug discovery, protein sequence analysis or concept search and question and answering a deep understanding of the biomedical concepts is required. The main challenge in word embeddings is that representation of words through word learning model. A composite embeddings and context aware embeddings are defined and compared with Bio-WordVec and the model outperformed [19].

Sentence similarity is an important task in medical text mining study and the traditional methods like bag of words may not define a topic model so accurately due to word ambiguity, so BioSentVec is defined with sentence embeddings learned from recurrent Pretrained sentence vectors [20]. Identification and mining of terms from large corpus is a laborious work to address this TermInformer project is proposed in this the system automatically mines the terms with the meaning form the biomedical documents and the system also reuses the existing generated word embeddings. This work compared the automatically mined term embeddings with the pretrained word embeddings and the result of the comparison showed that the auto mining of terms have shown better accuracy in defining semantic relationships [21].

Generation of biomedical word embeddings even it's emerging with unstructured data recently many embeddings' generations on structured data proposed. Term embedding approaches on structured data doesn't have a systematic method for computing the quality of these word similarity embeddings and performance of these embeddings in NLP chores. A structure comprising three dissimilar tasks associated with the aspects of ontological concepts: i) the categorization, ii) the hierarchal, and iii) the relational features is proposed and in different aspect the metrics for evaluating the quality of the task is also suggested [22].

In biomedical text natural language processing (BioNLP) applications and research still the statistical language modeling is a primary and essential step even there is continuous improvement in contextual based word embeddings using neural methods. Several neural architectures are proposed to meet the specific BioNLP task objectives and downstream applications. Biomedical natural language processing related tasks are continuously showing much improvement in adapting the neural network and architectures and latest advances of the networks and pretrained transformers using long short-term memory (LSTM), recurrent neural networks (RNN) techniques and bidirectional encoder and decoder techniques [23].

Existing word embedding methods Word2Vec and GloVe model have limitations in terms word vector representation where each word is represented as vector and internal sub structure of the word is ignored. More over the existing methods concentrate on generating embeddings from one large corpus ignoring the domain knowledge. If a topics model suggested can apprehend the sub term information and if can generate the insights of the words then it will greatly benefit the BioNLP applications. For each word generated the sub word information or heading information is captured with the concept of meta thesaurus these generated sub headings exploited as medical subject headings (MeSH). With this domain specific medical headings knowledge the medical word embeddings can qualitative in terms of BioNLP application performance [24].

PMCVec is an unsupervised technique used to generate the phrase embeddings from the PubMed abstracts. Phrase embeddings further explore the possibility of N-gram structures of the word embeddings whereas the existing models only suggested the unigram representation of the words. PMCVec proposed multi step model; i) biomedical article pre-processing, ii) using chunking approach generating phrases from this medical articles, and iii) filtering the generated phrases and ranking and re-ranking iv) creating distributed representation of phrases by phrase tagging [25].

Unified medical language (UML) system is defined from medical subheadings concept modeling. The phrases in this meta thesaurus model are combined to generate paraphrases of the unified medical language concepts. The paraphrases are created in three phases of this UML concept modeling i) feature collection, ii) feature selection, and iii) UML concept modeling [12].

Word embeddings are trained from four diverse repositories like biomedical publications, electronic health records, news articles and Wikipedia and these embeddings are evaluated against quantitative and qualitative approach and the medical subheadings considered from three subjects' disorder, symptoms, and drug. Intrinsic and extrinsic quantitative evaluation methods are used to evaluate the quality of the generated embeddings from the three categories. In intrinsic evaluation the word embeddings are checked with respect to medical semantics i.e., measuring the semantic similarity between the medical terms [13]. Sentence embeddings and the other word embeddings techniques in the biomedical articles domain require a unified methodology for assessment, so medical sentence evaluation system is proposed for ten different classification tasks. The novel MedSentEvalis proposed for BioNLP problems [14].

3. DOCUMENT REPRESENTATIONS METHOD

Word representation is depending on the distributional proposition. The terms belonging to same distributions share similar meanings. To represent the medical document as medical term model many language representation models proposed [15].

The word vector representation consists of chain rule probability distribution of words and their co-occurrence. Many approaches have been proposed on different corpuses and for biomedical domain also among these proposed models many have been implemented in various application like: i) keyword concept search, ii) question and answering system, and iii) medical recommendations [15].

The proposed models on biomedical domain in the early studies are bag-of-words model to represent the document words and their co-occurrence as vectors. To represent a bag of words (BoW) model for medical articles a high dimensional matrix model is used and many methods have been proposed for normalizing this representations like latent semantic indexing (LSA), and many more representation in neural network model like (PCA) principal component analysis and random indexing (RI) for reducing the high dimensional sparse matrix to low dimensional [15].

Word vectors have been upgraded to embeddings using neural networks over the years with the different models like global vectors neural model [4], fastText model [3], and Word2Vec model [3]. Embeddings from language model [15], bidirectional encoding representations from transformers (BERT) [15]. Medical word embeddings learned from biomedical copra as part of neural network language models have been used for performing specific tasks such as protein sequence identification, gene bank copra analysis, and gene structure analysis.

In the BoW neural model as shown in Figure 1, the similar words co-occurrence is processed in different layers of the neural network and a language model is developed with word sequence generation as shown in Figure 2. The input documents are vectored based on the Word2Vec model and the input layer is fed with N-input Word2Vecs generated from MED Corpus and the sequence similarity is processed in the hidden layers and language is constructed from generated sequent words in the dense layers and the output layer is sequence od terms generated which will help in taking crucial medical decisions like medicine recommendation, decease prediction. The input Word2Vec consists of the vectors with frequency of the word in medical document and the weights are assigned to these terms calculated from frequency and the inverse term frequency based on these weights the cosine similarity of the words calculated.
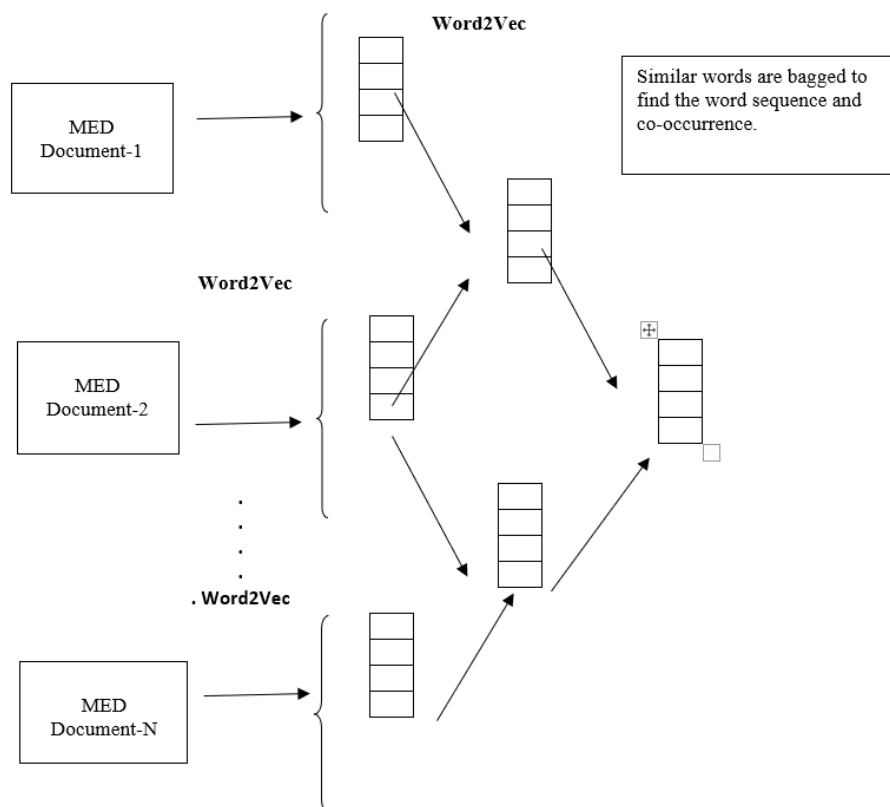

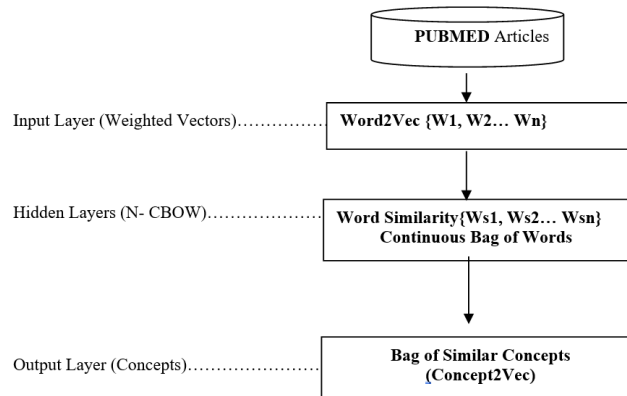Figure 1. Bag of words generation from medical documents

Figure 2. Word sequence generation using recurrent neural network

4.   PROPOSED CONCEPT MODELING METHOD (CONCEPT2VEC)

Term similarity task is extended with concept similarity in this paper and the complexity in sparse matrix is reduced using concept embeddings generation from medical corpus. Concepts can be generated from the N-grams constructed from the medical documents. These N-grams probability of co-occurrence is calculated and the values are vectorized as concept model 2-gram, 3-gram and so on N-gram.

The N-gram is constructed from the N words of medical document as:

$$Document - 1 = \{W1, W2 \dots Wn\} \dots\dots\dots vector\ of\ words$$

$$N - Gram_{Doc-1} = P\ (W1|\ W2 \dots Wn\}, P\ (W2|\ W3 \dots Wn) \dots P\ (Wn|\ W1 \dots Wn-1)$$

a concept hierarchy is generated, using this hierarchal model the next sequence of sentence is predicted with a 98% accuracy compare with other existing models. In the generation of this concept hierarchy a term frequency measure is a basic for generating the key terms. The key terms generated importance is calculated the weight function calculated from inverse document frequency [16].

$$Term - Weight\ (tw) = TF/IDF$$

These weighted terms and their similarity will generate the set of similar terms and that can be called as a concept [16].

$$Concept = (tw1, tw2 \dots twn)$$

$$C11 = P\ (tw11, tw12 \dots tw1i)$$

$$N - Gram_{Doc-1} = Document - 1 =$$
$$(C11, C12 \dots C1m), (C11, C12 \dots C1n), (C11, C12 \dots C1o) \dots\dots\dots. (C11, C12 \dots C1z)$$

$$N - Gram_{Doc-2} = Document - 2 =$$
$$(C21, C22 \dots C2m), (C21, C22 \dots C2n), (C21, C22 \dots C2o) \dots\dots\dots. (C21, C22 \dots C2z)$$

$$N - Gram_{Doc-J} = Document - J =$$
$$(CJ1, CJ2 \dots CJm), (CJ1, CJ2 \dots CJn), (CJ1, CJ2 \dots CJo) \dots\dots\dots. (CJ1, CJ2 \dots CJz)$$

Each N-gram generated from the medical documents can be considered to be a set of concepts and using concept similarity measures the correlation between the concepts calculated to generate the bag of concepts (BoC). BoC is the set of vectors generated from with different concept from PubMed documents. Using recurrent neural model, a concept language model can be generated from the concept hierarchy generation. From a medical document different set of correlated concepts generated with a probability of the N-Grams generated from the same documents and using different dense layers of the neural network a concept hierarchy model is generated which can predict the next sequence using the current sequence as shown in Figure 3.
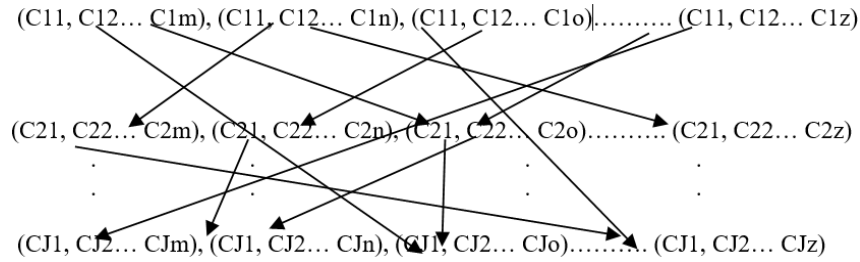
(C11, C12… C1m), (C11, C12… C1n), (C11, C12… C1o)……… (C11, C12… C1z)

(C21, C22… C2m), (C21, C22… C2n), (C21, C22… C2o)……… (C21, C22… C2z)

(CJ1, CJ2… CJm), (CJ1, CJ2… CJn), (CJ1, CJ2… CJo)……. (CJ1, CJ2… CJz)

Figure 3. Example ConceptNet model

5.   EXPERIMENTATION RESULTS

In this paper we have used the medical documents of PubMed abstracts of NCBI and considered based on the concept search "protein- protein interaction". PubMed central have interface to collect biomedical documents using e-utilities of national library of medicine and we have also collected 20k medical abstracts using the same utilities with the keyword "protein-protein interaction". Top 20 thousand documents are considered for the research, in this model the collected abstracts are processed in different layers of recurrent neural model for generating the Concept2Vecs. The process followed in this model as shown in Figure 4 is: i) abstract pre-processing, ii) BoC generation, and iii) concept modeling (Concept2Vec).
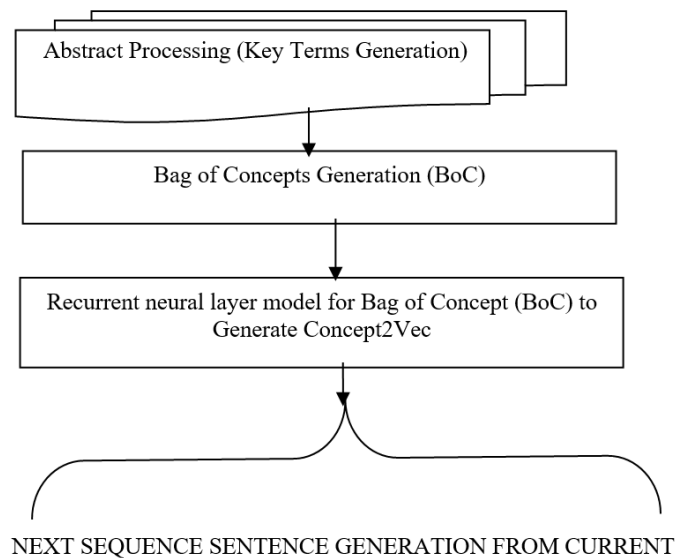

Figure 4. Concept 2Vec generation proposed model

Extracted documents are pre-processed with stop word removal and stemming methods for generating tokens of each document. Each token importance is calculated with the weight measure and all-important tokens of the documents are identified and indexed in matrix and the matrix size is reduced with the dimensionality reduction methods. The important terms are combined as a set then the probability of co-occurrence is calculated and indexed as vector which is named as concept vector. These generated concept vectors are processed in different layers of neural network as multi-layer perceptron (MLP) model. This MLP model is helps in identifying the sentence sequence.

The MLP model with BOC is represented as:

$$MLP\ (BOC) = Layer_N(SoftMax\ (Ln-1\ (RELU\ (Ln-2\ (.....(RELU\ (Concept2Vec)))))\ (5)$$

the next generated sequence is evaluated with the accuracy measures such as precision, recall and f-measure against different aspect of embeddings. Different combination of embeddings and similarity measures results evaluated are listed as follows shown in Table 2 [5].

Table 2. Different training set and the similarity measure results on two embedding models

Data set	Similarity measure	Word2Vec embeddings model	Concept2Vec embeddings model
5k Tokenset	Cosine	0.66	0.80
10k Tokenset	Cosine	0.67	0.85
15k Tokenset	Cosine	0.67	0.89
20k Tokenset	Cosine	0.69	0.98

Different model combinations are experimented and generated results are compared and listed as shown in Table 3. The evaluation results of above represent the different method combination for generating the sentence sequence from history of concept hierarchy. The experiments results reveal that using Word2Vec model the next word generation isn't accurate enough and the existing unsupervised models are less accurate and complex compared to the Concept2Vec model.

Table 3. Embeddings evaluation against different combination methods

Method	Precision	Recall	F1 measure
Word2Vec	0.648	0.628	0.637
Word2Vec+RNN	0.725	0.757	0.766
Concept2Vec	0.834	0.864	0.852
Concept2Vec+RNN	0.986	0.972	0.984

6. CONCLUSION

In this research, we have addressed the problem of matrix sparsity and defined a Concept2Vec model. Word distributions and representations increases the matrix sparsity number of dimensions will increase based on the single word distribution. So, we have defined the concepts from set of words and using these concepts a Concept2Vec model is defined for reducing number of dimensions. The Concept2Vec network is processed recurrently in different layers of the network to predict the next sequence of words and this model has outperformed every other method combination of the existing models.

REFERENCES

[1] L. Snidaro, G. Ferrin, and G. L. Foresti, "Distributional memory explainable word embeddings in continuous space," in *FUSION 2019 - 22nd International Conference on Information Fusion*, Jul. 2019, pp. 1–7, doi: 10.23919/fusion43075.2019.9011324.
[2] L. Zhao, J. Wang, L. Cheng, and C. Wang, "OntoSem: an ontology semantic representation methodology for biomedical domain," in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, Dec. 2020, pp. 523–527, doi: 10.1109/BIBM49941.2020.9313128.
[3] T. Huanling, Z. Hui, W. Hongmin, Z. Han, and M. Xueli, "Representation of semantic word embeddings based on SLDA and Word2Vec model," *Chinese Journal of Electronics*, vol. 02, pp. 1–9, May 2021, doi: 10.23919/cje.2021.00.113.
[4] H. Usselmann, R. Ahmad, and D. Siemon, "A personality mining system for German twitter posts with global vectors word embedding," *IEEE Access*, vol. 9, pp. 165576–165610, 2021, doi: 10.1109/ACCESS.2021.3130937.
[5] A. Singh and W. Jin, "On using composite word embeddings to improve biomedical term similarity," in *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, Oct. 2020, pp. 281–287, doi: 10.1109/BIBE50027.2020.00053.
[6] Q. Chen, Y. Peng, and Z. Lu, "BioSentVec: creating sentence embeddings for biomedical texts," in *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, Jun. 2019, pp. 1–5, doi: 10.1109/ICHI.2019.8904728.
[7] P. Tiwari, S. Uprety, S. Dehdashti, and M. S. Hossain, "TermInformer: unsupervised term mining and analysis in biomedical literature," *Neural Computing and Applications*, Sep. 2020, doi: 10.1007/s00521-020-05335-2.
[8] F. Alshargi, S. Shekarpour, T. Soru, and A. Sheth, "Metrics for evaluating quality of embeddings for ontological concepts," *CEUR Workshop Proceedings*, vol. 2350, Mar. 2019.
[9] V. Kumar, D. R. Recupero, D. Riboni, and R. Helaoui, "Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes," *IEEE Access*, vol. 9, pp. 7107–7126, 2021, doi: 10.1109/ACCESS.2020.3043221.
[10] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," *Scientific Data*, vol. 6, no. 1, p. 52, May 2019, doi: 10.1038/s41597-019-0055-0.
[11] Z. Gero and J. Ho, "PMCVec: distributed phrase representation for biomedical text processing," *Journal of Biomedical Informatics: X*, vol. 3, p. 100047, 2019, doi: 10.1016/j.yjbinx.2019.100047.
[12] W. Ning *et al.*, "Feature extraction for phenotyping from semantic and knowledge resources," *Journal of Biomedical Informatics*, vol. 91, p. 103122, Mar. 2019, doi: 10.1016/j.jbi.2019.103122.
[13] Y. Zhang, F. Tiryaki, M. Jiang, and H. Xu, "Parsing clinical text: How good are the state-of-the-art deep learning based parsers?," in *Proceedings - 2018 IEEE International Conference on Healthcare Informatics Workshops, ICHI-W 2018*, Jun. 2018, pp. 80–81, doi: 10.1109/ICHI-W.2018.00029.
[14] N. S. Tawfik and M. R. Spruit, "Evaluating sentence representations for biomedical text: methods and experimental results," *Journal of Biomedical Informatics*, vol. 104, p. 103396, Apr. 2020, doi: 10.1016/j.jbi.2020.103396.
[15] B. Chiu and S. Baker, "Word embeddings for biomedical natural language processing: a survey," *Language and Linguistics Compass*, vol. 14, no. 12, Dec. 2020, doi: 10.1111/lnc3.12402.
[16] T. Khatoon, A. Govardhan, and D. Sujatha, "Improving document relevant accuracy by distinguish doc2query matching mechanisms on biomedical literature," in *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering*, Jan. 2020, pp. 727–732, doi: 10.1109/Confluence47617.2020.9058299.

[17]   Z. Zhang, F. Han, H. Zhang, T. Aoki, and K. Ogasawara, "Examining the effect of the ratio of biomedical domain to general domain data in corpus in biomedical literature mining," *Applied Sciences (Switzerland)*, vol. 12, no. 1, p. 154, Dec. 2022, doi: 10.3390/app12010154.

[18]   M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014, doi: 10.1109/TMM.2014.2306655.

[19]   B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, "How to train goodword embeddings for biomedical nlp," in *BioNLP 2016 - Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 2016, pp. 166–174, doi: 10.18653/v1/w16-2922.

[20]   A. Forestiero and G. Papuzzo, "Natural language processing approach for distributed health data management," in *Proceedings - 2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2020*, Mar. 2020, pp. 360–363, doi: 10.1109/PDP50117.2020.00061.

[21]   D. A. Koutsomitropoulos and A. D. Andriopoulos, "Thesaurus-based word embeddings for automated biomedical literature classification," *Neural Computing and Applications*, vol. 34, no. 2, pp. 937–950, Jan. 2022, doi: 10.1007/s00521-021-06053-z.

[22]   S. Althubaiti, Ş. Kafkas, M. Abdelhakim, and R. Hoehndorf, "Combining lexical and context features for automatic ontology extension," *Journal of Biomedical Semantics*, vol. 11, no. 1, p. 1, Dec. 2020, doi: 10.1186/s13326-019-0218-0.

[23]   V. I. Ilie, C. O. Truica, E. S. Apostol, and A. Paschke, "Context-aware misinformation detection: a benchmark of deep learning architectures using word embeddings," *IEEE Access*, vol. 9, pp. 162122–162146, 2021, doi: 10.1109/ACCESS.2021.3132502.

[24]   M. Alawad and G. Tourassi, "Computationally efficient learning of quality controlled word embeddings for natural language processing," in *Proceedings of IEEE Computer Society Annual Symposium on VLSI, ISVLSI*, Jul. 2019, vol. 2019-July, pp. 134–139, doi: 10.1109/ISVLSI.2019.00033.

[25]   S. Robinson, "Conceptual modelling for simulation: progress and grand challenges," *Journal of Simulation*, vol. 14, no. 1, pp. 1–20, Jan. 2020, doi: 10.1080/17477778.2019.1604466.

BIOGRAPHIES OF AUTHORS

Suneetha Vazrala 🆔 🅖 SC ⚫ research scholar in Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, Telangana, India. She received her B.Tech. Degree from Mekapati Rajamohan Reddy Institute of Technology and Science, Affiliated with JNTU Hyderabad, Telangana, India in 2008, and her M.Tech. degree from the Holy Mary Institute of Technology and science JNTU Hyderabad, Telangana, India in 2010 under the Computer Science and Engineering Department. She published two Patents and 6 research papers in international journals. She can be contacted at email: suneetha.vazrala@gmail.com.

Dr. Thayyaba Khatoon Mohammed 🆔 🅖 SC ⚫ Ph.D. (JNTU Hyderabad), M.Tech. (JNTU Hyderabad), B.Tech. (Kakatiya University, Warangal). Area of interest: data mining, machine learning, bioinformatics, artificial intelligence, pattern recognition and neural networks, cloud computing, cyber security. She has published three patents and 20 above research papers in international journals. She have more than 14 years experience in teaching computer science and engineering subjects. She is serving as reviewer for various Scopus indexed, science citation indexed and Taylor and Francis Journals. She has organized various symposiums, workshops, faculty development programs and conferences at national and international levels. She have administrative experience of 7 years as Head of the Department. She was the chairman of board of studies in an autonomous Engineering College. She was expertise in NBA, NAAC, JNTUH, AICTE and UGC accreditation process. She received funds from DRDO and DST for organizing conferences. She can be contacted at email: thayyaba.khatoon16@gmail.com.