International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Semantic Marginal Autoencoder Model for the Word Embedding Technique for the Marginal Denoising in the Different Languages

**Dr. Deepak Kumar[1], L. Vertivendan[2], K. Velmurugan[3], Dr. Kumarasamy M[4], Ms. Dhanashree Toradmalle[5], Khan Vajid Nabilal[6]**

**Submitted:** 04/11/2022 **Accepted:** 01/02/2023

**Abstract:** The words are comprised of the smaller elements for the practical evaluation of the languages for the election of effective sematic. The conventional semantic technique subjected to the challenges associated with the incorporation of the different feature variables for the computation. However, the word embedding technique is complex due to the presence of the difference in the language features. This paper aimed to develop as an effective semantic model integrated with the Auto Encoder model. The proposed model is termed as Sematic Marginal Auto Encoder (SMarginalAE) for the different language sequences. The proposed model comprises of the Marginal features with the neighborhood estimation of the features. The proposed SMarginalAE achieves the neighborhood accuracy of 92.45% and the pair-wise accuracy is estimated as the 88.94%. The comparative analysis emphasised that the suggested SMarginalAE framework achieves the ~3% enhanced efficiency than the conventional techniques.

***Keywords:*** *Semantics, Word Embedding, Marginal Estimation, Neighbourhood Estimation, Accuracy*

## 1. Introduction

The smallest units of spoken or written language that have a clear, functional meaning are called words. The definition of a word in a language is its semantics [1]. A comprehensive examination of the ways in which people use language to express thoughts and emotions is required for a proper natural language semantics. In order to extract "word meaning" from a natural language's numerical representation, multiple techniques have been developed by researchers in natural language processing (NLP). These models may perform parsing, word sense disambiguation, and machine translation, among other NLP tasks [2]. One such method that determines the vector representation of a word's meaning based on its context is neural word embedding [3]. These vectors convert the semantic space, also known as semantic relative positions, between words. Words with comparable contexts will have vectors that are closer to one another in such a space. These fixed-dimensional vectors lack a clear definition of what each of its dimensions means. The main goal of this thesis is to use neural word embeddings to investigate the semantic space of word vectors.

In recent years, exceptional success rates have been recorded in a variety of NLP downstream jobs employing these neural word embeddings [4]. Nevertheless, the methodology to estimate the word embedding's efficiency in preserving the semantics of the word remains an open problem. Given that monolingual word embeddings have been the beneficiary of extensive research in diverse aspects like application-oriented and evaluation strategies, a part of the work presented in this dissertation is devoted to the development of methods for evaluation of bilingual word embeddings for the English-Tamil language pairs. The cost of procuring and cumulating these literatures' proprietary content for linguistic tools development is very high. Another difficulty is the enormous morphological variation between languages; in agglutinative languages like Tamil, formulated by the morphological affixes for the computation of the gender, mood, tense, person, voice and

[1] *Assistant Professor, Department of Humanities and Social Sciences, Maulana Azad National Institute of Technology, Bhopal, India.*
*drdkabhinav@gmail.com*
*0000-0002-7312-8561*
[2] *Assistant Professor, Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.*
*vetrivendan.l@galgotiasuniversity.edu.in,*
*0000-0001-5499-7019*
[3] *Assistant Professor, Department of English, Anurag University, Telangana, India.*
*ommvel23@gmail.com*
*0000-0003-4067-6745*
[4] *Assistant Professor, Department of Computer Science, College of Engineering and Technology, Wollega University, Nekemte, Oromia Region, Ethiopia.*
*drmksamy115@gmail.com*
[5] *Associate Professor, Department of Computer Engineering, K J Somaiya Institute of Technology, India.*
*dhanashree.t@somaiya.edu*
[6] *Associate Professor, Computer Engineering, KJ College of Engineering and Management Research, Maharashtra, India.*
*kvajid12@gmail.com*
*0000-0002-0999-9776*

number [5]. The magnitude of the Tamil words are higher than the English languages with the embedding technique to withstand the NLP languages integrated with the machine learning paradigmss. The machine learning paradigm with the language comprises of the multilingual scenario for embedding the same space vector. The trained multiple are computed in the vector form with the independently generated mapping function in the single vector known as bilingual embedding. However, the conventional bilingual embedding demands for the higher parallel resources for the comparable corpora in the aligned words and sentences subsets. In both languages the both words are embedded with the computation of the comparability of the real semantics [6]. The monolingual tasks are embedded with the language specific semantics in the vector form. For example, The English statement "A tiny river rushes into the sea" uses the Tamil word "Oodu" as its counterpart to the English word "run". However, Like in the previous line, "run" and "Oodu" both have several meanings that are incompatible. Common semantics are captured through bilingual embedding, while variances are rejected., an inevitable trade-off in the interest of a common vector space. The proposed SMarginalAE involved in the computation of the words in the languages with consideration of features.

## 2. Related Works

Even before the VSM attained its notoriety, the first evaluation of word embedding was conducted. A comprehensive review of duties that could be handled by VSM was put out in [7]. These challenges are therefore seen as proof of the effectiveness of word embedding. The initial comparisons of the outcomes of the various VSMs were published in the same year by [8]. Following the development of the Word2Vec tool in 2013, [9] suggested a word analogy task, a cutting-edge method for assessing word embedding. A thorough overview of methods for measuring word embedding is given in [10]. Extrinsic and intrinsic evaluation are the two main groups of contemporary approaches that were systematised in [11]..

When translating text from one language to another using statistical models and a big corpus, this process is known as phrase-based statistical machine translation [12]. The translation and language prototype elements are combined in the log-linear PB-SMT paradigm. The best resource for multilingual learning is the phrase table, which in its simplest form has four probabilities values. The co-occurrence of phrase pairs in the training corpus is shown by these scores, not their semantic connection. Therefore, a number of methods were put out for adding semantic knowledge to the phrase-table scores. In order to increase SMT's effectiveness, recent efforts have used distributional representations of words and phrases. This approach is predicated on the idea that words and phrases with comparable distributive representations exist across languages. The foundations for the Neural Probabilistic Language Model were developed via the work of [13], who focused on continuous space language models. In these models, n-gram probabilities are calculated using a continuous word representation rather than a conventional discrete representation (NPLM). Similar words are represented by NPLM adjacent to one another in a continuous area. Its basic design, which consists of two hidden layers underneath the top SoftMax layer and acts on the distributed representation of words, is straightforward. NPLM learns a distributed representation as well as a distributive n-gram probability across words, however it is limited to working with a small vocabulary size. For a powerful vocabulary application like the MT system, the vocabulary size restriction makes NPLM challenging. The NPLM version created in [14] is superior to the earlier LMs. With a broad vocabulary and a huge corpus, this new NPLM can be trained more quickly and is integrated into an SMT system decoder. A neural network addition called the Structured Output Layer (SOUL) uses a class-based language model to increase MT efficiency with less training data. A recurrent neural network provides the foundation for the integrated language and translation model. Depending upon that unbounded context of both the source and target words, it predicts the target words.

## 3. Proposed Model

In the bilingual/monolingual model, the language for the website can be crawled with the extraction of the main text in the designed web pages. Work embedding techniques are comprised of the incorporation of the raw text with the inclusion of hyperlinks, headlines, images, publications, and so on. To perform the word embedding technique this paper proposed a SMarginalAE model. The SMarginalAE can be normalized and cleaned with the elimination of quotations, currency, brackets, and punctuation with the incorporation of digits and characters. On the website, the bilingual words in the website are processed parallel manner with the assigned delimiter sentence alignment. The proposed SMarginalAE model are corrected manually with the assigned sentence with the splitting of more than 30 words to achieve consistency in data. The sentences in the websites are processed and computed based on the linguistic pattern with variation in the token length for the average corpus tokens per sentence in the functional words. The process involved in the proposed SMarginalAE model are presented in figure 1.

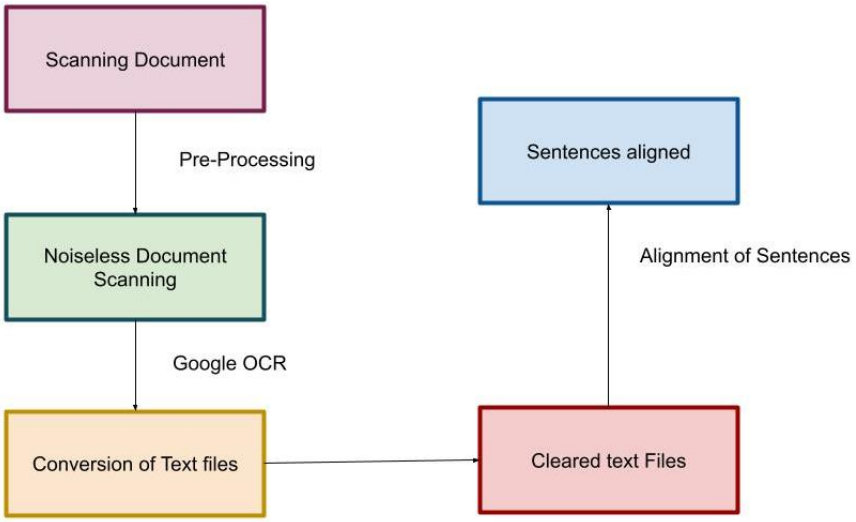



**Fig. 1:** Process in SMarginalAE

The proposed SMarginalAE sequences are mapped with the pre-trained source embedding (X) with the mapping function for the target embedding (Y). The proposed model SMarginalAE incorporates the word embedding in English and Tamil form. The examination is based on the processing of the word embedding methodology with the consideration of the sentences as presented in table 1.

**Table 1:** Sentences for Analysis

| | English Sentences | Tamil Sentences |
|---|---|---|
| Monolingual | 465734 | 573493 |
| Bilingual | 55783 | 56723 |

The word embedding is applied using the targeted contemporary vector sequence training algorithm for the limited corpus target. The word embedding is processed based on the pre-trained billion-word corpus text for the different online sources. Through the incorporation of the machine learning-based autoencoder is stated as M with the mapping function is defined as X to Y. However, with the information in the bilingual the features are mapped with the X and Y in the information dictionary for the created dataset D. The trained model uses the transfer function F for the generation of the vector target in the source word. In figure 2 the word embedding with the autoencoder model is presented.

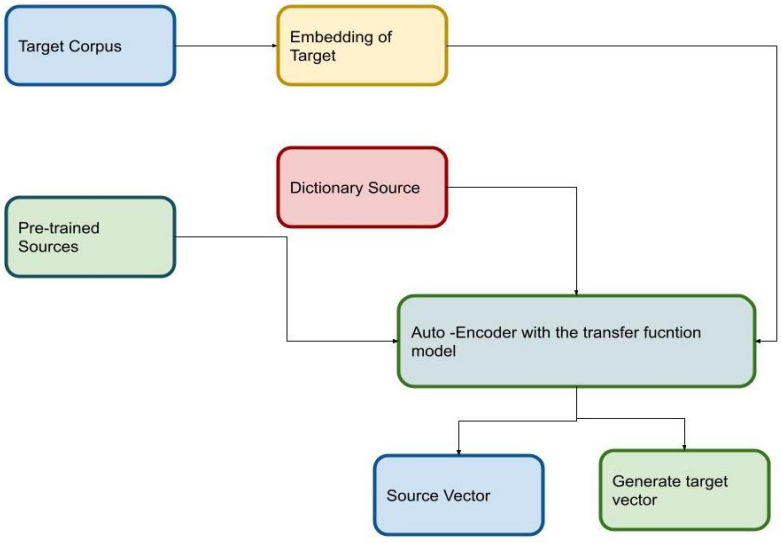



**Fig. 2:** Flow Chart of Transfer Learning

The data sources considered for the examination of the proposed SMarginalAE for the estimation is presented in table 2.

**Table 2:** Embedded word Sequences

| Data Sources | Data Target | Number of Dictionary words | Attribute for processing | Algorithm |
|---|---|---|---|---|
| Wikipedia | Wikipedia (ta) | 17842 | Comparable Form | FastText |
| GloVe | Monolingual Tamil (ta) | 14783 | Monolingual Form | GloVe |
| Google | Tamil (ta) | 12782 | Monolingual Form | Word2Vec |
| Tam (en) | cEnTam (ta) | 5893 | Parallel Process | BilBOWA |

In order for the proposed SMarginalAE to be taken into consideration by the evaluation metrics to achieve the word embedding for pair-wise usability for the semantic words. The semantic features of the words comprise the cosine global neighborhood estimation for the isolated words with relative topology for the target word embedding.

### 3.1 Linear Mapping

The proposed SMarginalAE model incorporates linear mapping features for the elementary vector spaces. With the English embedding vector, the projects are embedded in the Tamil space. The vector is represented in the 300 dimensions in the vector with the English word presented in T ,atrix. The features in the vector form for 300 × 1 as represented in equation (1)

$$Tx = y \tag{1}$$

In above equation (1) the variable $x$ denoted as the embedding space for the English language and spanning tree for the embedded space in Tamil is represented as $y$. In the Moore Penrose inverse is presented in equation (2) and (3) with computation of variables $X^{+} = (X^{T}X)^{-1}X^{T}$ for the matrix operator transfer function

$$TX = Y \tag{2}$$

$$T = YX^{+} \tag{3}$$

However, accurate features are computed with the variables $Y$ and $X$ in the appropriate column for the unrelated column features in the embedding space. With the incorporation of the AutoEncoder model deep learning network features are estimated with the fully connected network for the non-linear vector function value as , $f$ such that $f(x) = y$ where the vector of English is defined as $X$ and the tamil features are represented as $y$. The training sequences vector is computed based on the reduction of the loss function in the target vector. The loss function comprises of the cosine function computed between the variables with the word embedding techniques. The proximity of the generated target is represented as Tˆ with the embedded target value presented in equation (4)

$$Cos.prox = 1 - \cos(\hat{T}, T) \tag{4}$$

The pipeling architecture of the autoencoder mode for the word embedding techniques are presented in the figure 3 with the hyper parameter estimation.

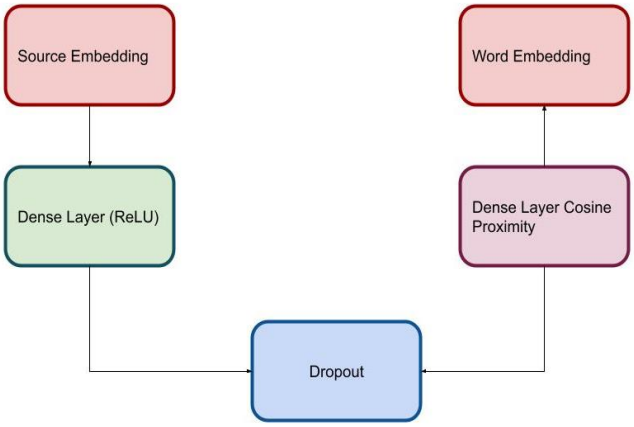



**Fig. 3:** Hyper Parameter word embedding technique

With the computation of the hyper parameters the features in the embedded word sequences are evaluated with the semantic collected word-pair evaluation in the computation of the linguistics relations. The sample word-pair are computed based on the cosine feature with pairwise operation in the equation (5). The pipeline architecture for the SMarginalAE model computed for the accuracy operation with consideration of the cosine distance in the vector a and b

$$\cos(a, b) = \frac{a^{T}.b}{\|a\|.\|b\|} \tag{5}$$

The computation of the pairwise cosine accuracy is computed based on the measured linguistics relation with the embedding, accuracy estimation with the computation of the word embedding topology.

## 4. Results and Discussion

The effectiveness of the suggested SMarginalAE is comparatively examined with the existing techniques such

as Linear Mapping (LM), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) - and a bilingual model, BilBOWA for the comparative examination. The experimental evaluation is performed according to the viewpoint of the pair-wise cosine accuracy (P.Accuracy) and neighbourhood accuracy (N.accuracy) with the proposed SMarginalAE as presented in the table 3.

**Table 3:** Comparison of Accuracy

| Transfer Functions | N.Accuracy (%) | P.Accuracy (%) |
|---|---|---|
| LMW | 76.23 | 28.91 |
| LMG | 79.34 | 45.68 |
| LMF | 81.34 | 49.83 |
| MLPW | 84.76 | 53.27 |
| MLPG | 83.42 | 55.73 |
| MLPF | 85.93 | 59.84 |
| CNNW | 83.68 | 62.64 |
| CNNG | 88.36 | 66.83 |
| CNNF | 86.72 | 72.37 |
| SMarginalAE | 92.45 | 88.94 |

The comparative examination of the proposed SMarginalAE with the existing model is presented in figure 4 and 5. The figure 6 provides the word embedding technique implemented for the tamil and English languages are presented.

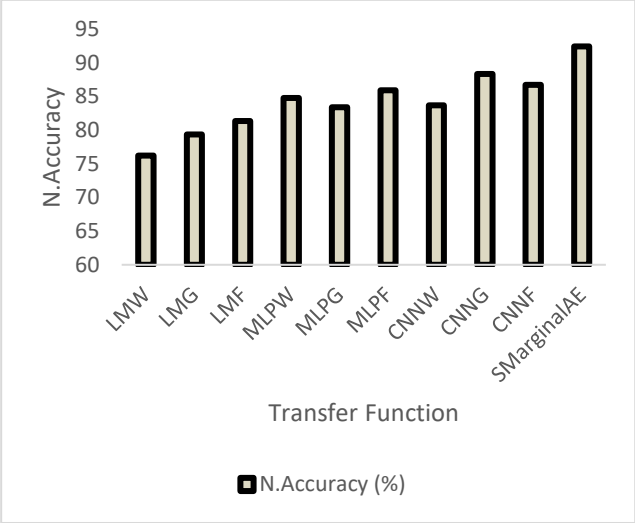



**Fig. 4:** Comparison of N.Accuracy

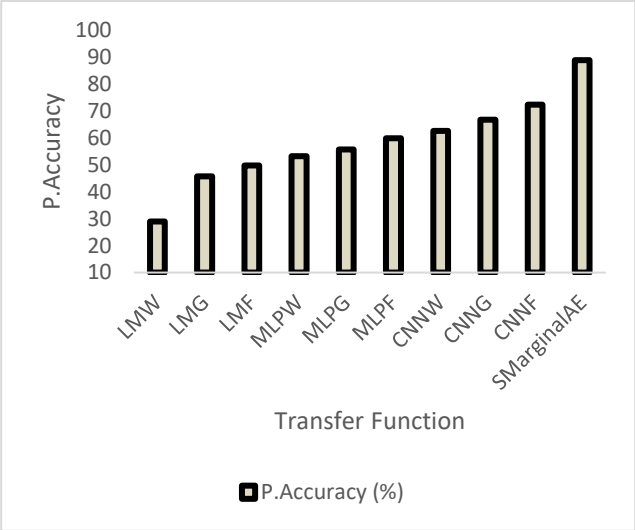
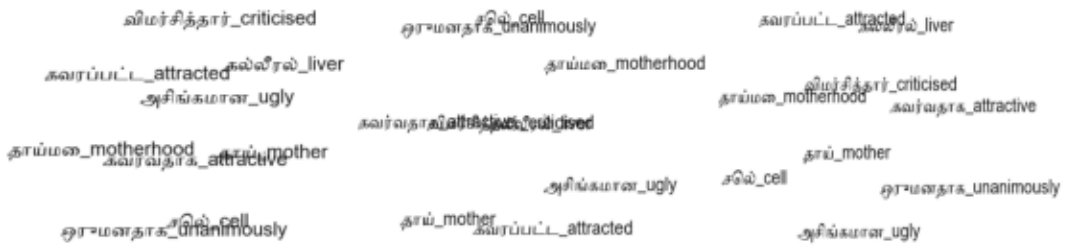



**Fig. 5:** Comparison of P.Accuracy




(a) Original Embedding (b) Auto Encoder Embedding (c)SMarginalAE Embedding

**Fig. 6**

The computation of the accuracy value is cross-validated with the set of 300 words with the 50 epochs for the created pair-wise data feature in the intrinsic bilingual embedding techniques as presented in table 4.

**Table 4:** Comparison of Accuracy

| Language | Transfer Function | Model | P. Accuracy (%) | N. Accuracy (%) |
|---|---|---|---|---|
| Hindi | Auto Encoder | Eng-Tam Matrix | 11,34 | 88.45 |
| | | Re-computed Matrix | 19.34 | 90.46 |
| | SMarginal AE | Re-computed Matrix | 76.95 | 75.78 |
| | | Re-trained Network | 83.02 | 84.56 |
| Tamil | Auto Encoder | Eng-Tam Matrix | 43.67 | 84.78 |
| | | Re-computed Matrix | 78.82 | 94.72 |
| | SMarginal AE | Eng-Tam Network | 83.56 | 88.72 |
| | | Re-trained Network | 90.56 | 92.91 |

The different features in the neighbourhood estimation is evaluated under the POS categories for the classification under four categories such as Noun, Verb, Adverb and Adjectives. The measured cosine neighboorhood for the each category is presented in table 5.

**Table 5:** Tamil Embedding Marginal Estimation

| Category | Embedding Models | | |
|---|---|---|---|
| | **Word2Vec** | **GloVe** | **FastText** |
| Nouns | 83.45 | 94.56 | 88.56 |
| Verbs | 84.78 | 92.34 | 91.57 |
| Adjectives | 82.56 | 90.84 | 86.78 |
| Adverbs | 86.82 | 92.56 | 90.78 |

The experimental evaluation of the presented SMarginalAE is determined by utilising the computation of the marginal features as in table 6. In figure 7 embedding model for the estimation of the different features based on different categories are presented.

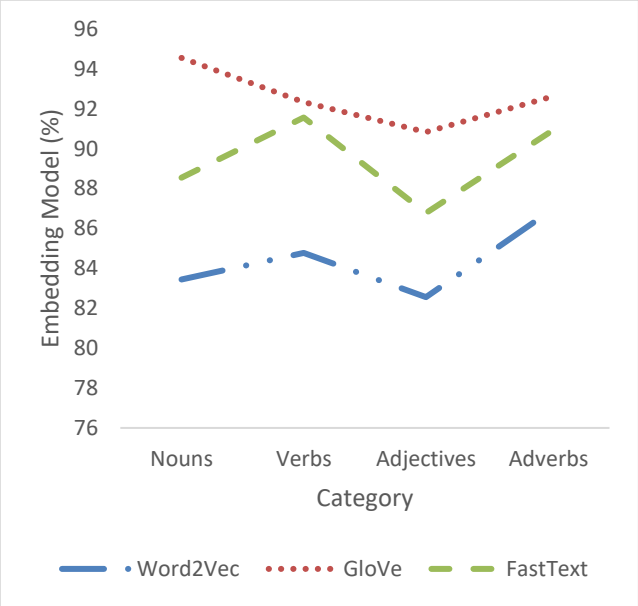



**Fig. 7:** Comparison of Embedding Models

**Table 6:** Word Embedding with the AutoEncoder

| **Text size** | **Summary Size** | | **Different Sentences** |
|---|---|---|---|
| | Original Text | Transfer Learning with Auto Encoder | |
| 20 | 2 | 2 | 0 |
| 40 | 6 | 6 | 0 |
| 60 | 12 | 12 | 0 |
| 80 | 16 | 16 | 0 |

In the table 6 the different sentences are estimated based on captured sematics with the word embedding for the proposed AutoEncoder is presented. In the table 7 word embedding for the proposed SMarginalAE model is presented.

**Table 7:** Word Embedding with SMarginalAE

| **Text size** | **Summary Size** | | **Different Sentences** |
|---|---|---|---|
| | Original Text | Transfer Learning with SMarginalAE | |
| 20 | 2 | 2 | 1 |
| 40 | 6 | 6 | 8 |
| 60 | 12 | 12 | 14 |
| 80 | 16 | 16 | 16 |

The summary variable for the different text features are computed for the varying text size with the transfer learning in the proposed SMarginalAE. The comparison study

indicates the suggested SMarginalAE model achieves the superior performance to the standard method.

## 5. Conclusion

Word embedding techniques with the semantic features are estimated based on the word embedding techniques for the consideration of the Tamil and English languages. The generated feature variables are computed based on the sematic based AutoEncoder model with the marginal feature estimation. The proposed SMarginalAE perform the pairwise estimation of the features. The proposed model achieves the higher 92.45% for the neighbood estimation which is significantly minimal than the existing technique. The proposed model computed as sentence as the value of 1, 8, 14 and 16 for the varying text data.

## Reference


[1] Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application*, 267-281.

[2] Neelakandan, S., Arun, A., Bhukya, R. R., Hardas, B. M., Kumar, T., & Ashok, M. (2022). An automated word embedding with parameter tuned model for web crawling. *Intelligent Automation & Soft Computing*, *32*(3), 1617-1632.

[3] Roman, M., Shahid, A., Khan, S., Koubaa, A., & Yu, L. (2021). Citation intent classification using word embedding. *Ieee Access*, *9*, 9982-9995.

[4] Srinivasan, S., Ravi, V., Alazab, M., Ketha, S., Al-Zoubi, A. M., & Kotti Padannayil, S. (2021). Spam emails detection based on distributed word embedding with deep learning. In *Machine intelligence and big data analytics for cybersecurity applications* (pp. 161-189). Springer, Cham.

[5] Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, *8*(4), 881-893.

[6] Singh, K. N., Devi, S. D., Devi, H. M., & Mahanta, A. K. (2022). A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights*, *2*(1), 100061.

[7] Habib, M., Faris, M., Alomari, A., & Faris, H. (2021). AltibbiVec: A Word Embedding Model for Medical and Health Applications in the Arabic Language. *IEEE Access*, *9*, 133875-133888.

[8] Li, S., Pan, R., Luo, H., Liu, X., & Zhao, G. (2021). Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling. *Knowledge-Based Systems*, *218*, 106827.

[9] Rani, R., & Lobiyal, D. K. (2021). A weighted word embedding based approach for extractive text summarization. *Expert Systems with Applications*, *186*, 115867.

[10] Fesseha, A., Xiong, S., Emiru, E. D., Diallo, M., & Dahou, A. (2021). Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Information*, *12*(2), 52.

[11] Kumhar, S. H., Kirmani, M. M., Sheetlani, J., & Hassan, M. (2021). Word embedding generation for Urdu language using Word2vec model. *Materials Today: Proceedings*.

[12] Wu, F., Yang, R., Zhang, C., & Zhang, L. (2021). A deep learning framework combined with word embedding to identify DNA replication origins. *Scientific reports*, *11*(1), 1-19.

[13] Du, Y., Fang, Q., & Nguyen, D. (2021). Assessing the reliability of word embedding gender bias measures. *arXiv preprint arXiv:2109.04732*.

[14] Zuo, Y., Li, C., Lin, H., & Wu, J. (2021). Topic modeling of short texts: A pseudo-document view with word embedding enhancement. *IEEE Transactions on Knowledge and Data Engineering*.