Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 61(3): 227-245 (2024)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(61-3)842

Pakistan Academy of Sciences

Review Article

Advancements in Word Embeddings: A Comprehensive Survey and Analysis

Khushal Das[1*], Kamlish[2*], and Fazeel Abid[3]

[1]Department of Computer Science, Modelling, Electronics and Systems Engineering, University of Calabria, Rende, Itlay

[2]Department of Computer & Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan

[3]Department of Computer Science and Information Technology, University of Lahore, Lahore, Pakistan

Abstract: In recent years, the field of Natural Language Processing (NLP) has seen significant growth in the study of word representation, with word embeddings proving valuable for various NLP tasks by providing representations that encapsulate prior knowledge. We reviewed word embedding models, their applications, cross-lingual embeddings, model analyses, and techniques for model compression. We offered insights into the evolving landscape of word representations in NLP, focusing on the models and algorithms used to estimate word embeddings and their analysis strategies. To address this, we conducted a detailed examination and categorization of these evaluations and models, highlighting their significant strengths and weaknesses. We discussed a prevalent method of representing text data to capture semantics, emphasizing how different techniques can be effectively applied to interpret text data. Unlike traditional word representations, such as Word to Vector (word2vec), newer contextual embeddings, like Bidirectional Encoder Representations from Transformers (BERT) and Embeddings from Language Models (ELMo), have pushed the boundaries by capturing the use of words through diverse contexts and encoding information transfer across different languages. These embeddings leverage context to represent words, leading to innovative applications in various NLP tasks.

Keywords: Word Embeddings, Word Representations, NLP, Contextual Embeddings, BERT, ELMo, Word2Vec, Cross-Lingual Embeddings.

1. INTRODUCTION

Words are components of any speech belonging to meaning as well as significance. Further, characters of any written word never have a bit of significant sense by themselves, which shows that characters can't present a powerful sense of the written word individually. For instance, Book and Pen tend to be related to one another, but it is most unlikely that we will evaluate or come across the value of this relevancy by using only the characters of this pair of words [1]. Word embeddings are frequently described as models for strings that fulfil the essential function of providing meaningful representations for words or phrases. Moreover, word embeddings represent words within a continuous vector space, allowing for the modelling of well-defined relationships among them. In this space, words are arranged into vectors with familiar properties, establishing meaningful connections through geometric relationships [2-4].

The influence of word embeddings largely depends on their ability to capture natural language and geometric relationships. They allow expeditious end-to-end modules by modulating an exceptional real-life representation right into an unremitting space; because of this, they are

Received: November 2023; Revised: August 2024; Accepted: September 2024
* Corresponding Authors: Khushal Das <Khushal.das@dimes.unical.it>; Kamlish <kamlish@outlook.com>

well-liked in natural language processing (NLP) subjects: they will be easy to plug into deep learning modules [5]. Sentiment analysis stated that physical object recognition and many other everyday jobs surpassed their renowned matching part with these techniques. We are primarily concerned with word embeddings learned on the trained corpus. This group of representations is creating an effort to compile a complete plain text useful dataset into an unceasing vector representation with no professional familiarity. Thus, in this survey, we accept this claim that the logic of a word enormously relies on the words neighbouring it. Other word embedding practices enhance this supposition and usage of a language model to shape contextualized word representations, such as BERT [6]. Recent advancements in areas such as sign language translation [7], healthcare signal processing [8], and vehicle detection [9] further illustrate the versatility of deep learning models in enhancing communication and accuracy across diverse applications. Finally, a distinctive technique is to build vectors using familiarity or another basis of experienced knowledge; an example is the TransE approach, as described by Cano and Morisio [10]. The process of training and using word embeddings for a machine learning objective is illustrated in Figure 1 [11].

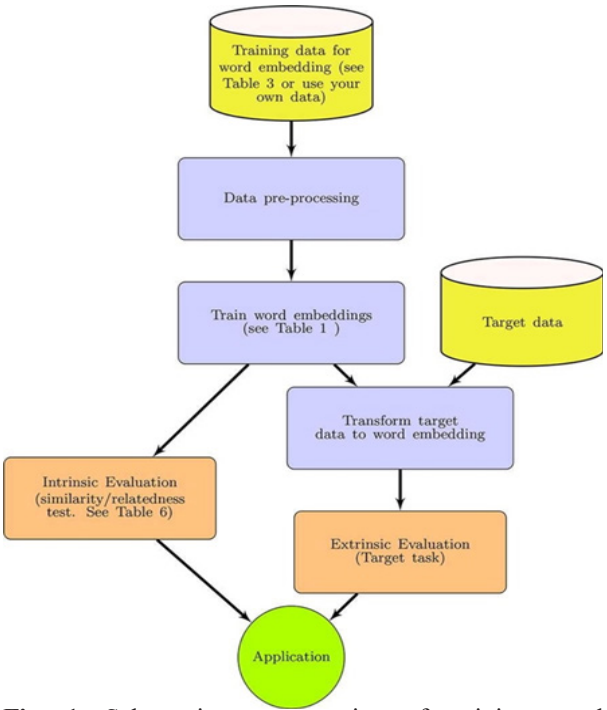

Fig. 1. Schematic representation of training word embeddings to use them for Machine learning objectives [11].

Previously, feature engineering NLP involved evolving vital mathematical functions to symbolize relevant sides of the text, such as the relation of pronouns to nouns. This method frequently required important domain information plus energy to find significant features [12]. Differently, word embeddings can be studied from the text's corpus and don't need any feature extraction or manual labelling; they are usually known in an unsupervised method [13]. So, we can say that word embeddings can be straightforwardly learned on whichever text data corpus. Word embedding is divided into two types: contextual and non-contextual word embeddings. The differentiation between these two types is that either the word embedding changes dynamically according to the context in which it appears or not.

Regardless of an excess of related works reachable on language models, word embeddings, and their advancements plus applications, no comprehensive survey collecting the detailed work done on word embeddings exists up to now. The current paper discusses the recent advances and innovations in word embeddings. In a study, Font and Costa-Jussà [14] employed a transfer translation architecture to examine incorporating two debiasing techniques using Global Vectors for Word Representation (GloVe) embeddings. The researchers put forth and assessed a scheme on the WMT English-Spanish benchmark task, observing improvements of up to one Bilingual Evaluation Understudy (BLEU) point. Regarding gender bias assessment, the researchers generated a collection of occupations and demonstrated that their system can mitigate inherent biases in the baseline system. Rezaeinia *et al*. [15] introduced Improved Word Vectors (IWV), a novel technique designed to enhance the accuracy of sentiment analysis by leveraging pre-trained word embeddings. Their methodology incorporated several approaches, including part-of-speech (POS) tagging, lexicon-based strategies, a word position algorithm, and Word2Vec or GloVe methods. Their plan's accuracy was validated using deep-learning models and benchmark datasets designed explicitly for sentiment analysis. The results of their experiment about sentiment analysis demonstrated the high effectiveness of IWV. Yao *et al*. [16] developed several intuitive evaluation methods for temporal word embeddings. Their quantitative and qualitative analyses indicate that their methodology consistently captures

evolutionary patterns. Furthermore, their approach steadily overtakes current state-of-the-art material embedding methods regarding semantic accuracy and structural quality. Zhao *et al*. [17] performed a series of intrinsic analyses, revealing several key findings. Firstly, ELMo, a language representation model, was observed to contain a significantly higher number of male objects than female objects. Secondly, the ELMo embeddings consistently incorporate gender-related information. Lastly, the encoding of gender data in ELMo was found to be uneven, with a noticeable disparity between male and female objects. Moreover, the researchers demonstrated that a prior system reliant on ELMo exhibits bias and identifies significant bias within the WinoBias dataset. Finally, the researchers examined two methodologies to mitigate gender bias and demonstrated the potential for eliminating the bias observed in the WinoBias dataset.

In recent years, the era of significant data has brought about challenges related to information overload. Addressing these challenges, Du *et al*. [18] aimed to achieve precise and automatic categorization of Internet news edition data. Recognizing the limitations of single-topic and word embedding models, they planned a novel text representation method that combined Glove models, Word2VEC, LDA, and TF-IDF. Additionally, Suhartono *et al*. [19] introduced two CNN architectures that incorporated Glove and Word2Vec word embeddings to analyze sentiment in drug reviews, utilizing deep learning methods, for instance, BERT and RoBERTa. Haller *et al*. [20] provided a comprehensive taxonomy of ways in the field, spanning classical Machine Learning to Deep Learning approaches while emphasizing the need for adaptations in Deep Learning architectures for NLP to tackle evolving challenges in ASAG tasks. Gender bias in static word embeddings was scrutinized by Caliskan *et al*. [21], revealing preferences in semantic suggestions, word frequency, parts of speech, clustered concepts, word frequency, term, parts of speech, and word meaning dimensions. Meanwhile, Tang *et al*. [22] proposed an unsupervised method to learn Dynamic Contextual Word Embeddings (DCWEs) through time-adapting a pre-trained MLM using manual and automatic templates. Alnajjar *et al*. [23] contributed to the field by creating a sentiment analysis corpus for endangered languages and Finnish. The study conducted by Yen and Jeon [24] achieved significant accuracy improvements in embedding-matching A2W systems by generating multiple embeddings and incorporating pronunciation-based embeddings. Engler *et al*. [25] introduced SensePOLAR, offering word sense-aware interpretability for contextual word embeddings. Schiffers *et al*. [26] developed word embeddings tailored for the social sciences and compared them to general language models in a domain-specific context. Lastly, Zaland *et al*. [27] comprehensively evaluated existing word embedding algorithms on extrinsic classification tasks, shedding light on how these models encode word relations. The study by Worth [28] highlights that advancements such as Word2Vec, GloVe, ELMo, and BERT embeddings rely on the idea that a word's semantic meaning is shaped by its distributional properties within a text corpus. The study by Das and Kamlish [29] shows that knowledge about words' meaning helps make summaries better and more accurate. This research benefits tasks like finding information, organizing documents, and extracting knowledge. The new method also makes summarizing text easier and reduces the need for manual work. It shows how important it is to understand language when using automated tools. This method helps deal with the vast amount of text we have today. Abro *et al*. [30] combined Word2Vec and GloVe embeddings with a neural network to improve the model. We then tested its performance using different learning rates across ten developers. The results showed that when Convolution was combined with Word2Vec embeddings, the model tended to be more accurate on average during testing.

2. REPRESENTING TEXT WITH EMBEDDINGS

This section provides a concise overview of the different types of word embeddings. We conduct a detailed analysis of a text, focusing on its word sequence to explore the contextual relationships between the embeddings.

2.1. Representation of Word Embeddings

Numerous methodologies, such as Word2Vec, GloVe, and FastText, are commonly used to examine word embeddings, each employing distinct approaches for capturing semantic relationships in a corpus. For example, Word2Vec uses neural networks to predict word context. At the same

time, GloVe captures co-occurrence statistics of words in a large corpus, and FastText considers subword information, making it more effective for morphologically rich languages [31-33]. One common approach involves utilizing a one-hot encoding technique, which assigns a distinct index within a vocabulary dictionary to each word, creating a unique representation for every word in the corpus. A comment is demonstrated by a vector consisting of all zeros except for one within an appropriate context. The process of studying context-based prediction involves the utilization of word embedding strategies. These strategies enable mapping one-hot vectors to more compact representations, often with lower dimensions than vocabulary dimensions. These representations' components capture the language data's underlying symbolic meaning. The fundamental premise is that to achieve accurate word prediction, it is necessary to enhance and refine the representations of words through learning [11].

2.1.1. Word2vec

The Word-to-Vector (Word2vec) technique under consideration relies entirely on a predictive approach that can be implemented using the skip-gram (SG) and continuous bag-of-words (CBOW) models [31]. Small neural networks are used in Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) models to map words to specific points in a vector space. The distinction between these methods lies in whether the neural network endeavours to forecast a target term given its context (Continuous Bag of Words, CBOW) or vice versa. Two crucial factors determine the training of word2vec embeddings, as shown in Figure 2 [34]. Firstly, the embedding dimension is between fifty and five hundred through experimental methods. Secondly, the span of the context window refers to the number of words preceding and following the target word that is utilized as context for training the word embeddings. Additional significant hyperparameters are elaborated upon in the appendix section. The requirement for a more extensive training dataset is typically observed when training embeddings with more dimensions. It is crucial for each dimension to effectively capture a distinct aspect of meaning so that the embeddings possess the necessary capacity to differentiate between words. Paragraph2Vec and Doc2Vec are variants of the word2vec model

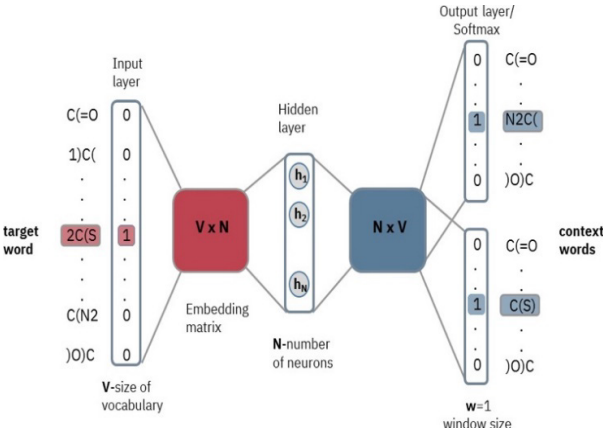

Fig. 2. Illustration of the Skip-Gram Architecture in the Word2Vec Algorithm [34].

designed to represent documents or paragraphs as vectors rather than individual words. There are two distinct types of doc2vec models: the PV-DM model, like the SG model of word2Vec, and the PV-DBOW model; both are used to distribute memory across paragraphs [35].

2.1.2. GloVe

Global Vectors for Word Representation (GloVe) model comprehends word embeddings proficiently using a word co-occurrence matrix rather than a word calculation job. A co-occurrence matrix is a VXV square matrix in which V indicates the vocabulary size. Every matrix element represents the frequency of occurrence of the specified vocabulary objects within a predetermined context window that spans the entire corpus. GloVe can comprehend vector embeddings, which facilitate the reduction of literal errors during the processing of co-occurrence statistics required by the model. Additionally, it considers the global co-occurrence statistics present in the preparation corpus. The model comprises multiple hyperparameters that must be vigilantly chosen, such as the dimension of the vector embedding and the size of the perspective window. The word vectors generated through the GloVe method exhibit epistemological equivalence to those obtained through word2vec. However, GloVe employs a count-based model as its foundation, in contrast to word2vec's predictive model [35]. GloVe, compared to word2vec, is known for its ability to capture longer-term dependencies due to its computation of statistics over more oversized context windows. However, it is essential to note that the order of these

dependencies needs to be preserved. Based on empirical observations, discernible advantage has yet to be identified for either GloVe or word2vec models. The overall reliability of these models is contingent upon various factors, such as the kind of data and the specific assessment job being measured. GloVe embeddings have proven highly effective in capturing semantic relationships in various natural language processing tasks, as demonstrated in recent studies across different application domains [36-38].

2.1.3. FastText

The FastText model extends word2vec and GloVe methods, incorporating a specific constraint. This model can lever novel, out-of-vocabulary (OOV) expressions by utilizing the word2vec skip-gram (SG) model, which includes inner subword data in character n-grams, representing sequences of adjacent characters. This approach entails constructing a vector representation for a word by considering the combination of its subword elements. This approach also enables the model to capture the structural and linguistic relationships between words and facilitates the creation of vectors for previously unanticipated words. Probabilistic FastText is a methodology used to combine FastText and Gaussian Mixture Models (GMMs). Howard and Ruder [39] did not provide any text to rewrite. The representation of each word is depicted as a Gaussian Mixture Model (GMM) consisting of n mechanisms, effectively capturing n distinct senses or meanings associated with the word. This representation can analyze the sub-word structure, distinguish between dissimilar word senses, and deliver improved representations of infrequent or hidden words. Recently, FastText has been widely applied in various fields of natural language processing, demonstrating its effectiveness in tasks such as sentiment analysis, cybersecurity, and machine learning applications [40-43].

2.1.4. ELMo

Embedding from Language Model (ELMo) is a language representation model emphasising words through character-level and word-level embeddings. Instead of employing a stable embedding for every word, ELMo evaluates the entire sentence and assigns each word an embedding [44]. The embeddings are constructed utilizing a trained bidirectional recurrent neural network (RNN) for a particular task. The embedding's bidirectional architecture is based on both preceding and subsequent words. One significant innovation of ELMo is the incorporation of task-specific weighting coefficients for the embeddings. It allows the model to be trained on one objective or task and then applied to a different task, effectively combining shared information while focusing on specific semantic aspects. Integrating ELMo word embeddings with deep learning multimodal transformers has shown promising results in enhancing image description tasks, as demonstrated in recent research by Cheng *et al*. [45]. In a recent study by Rong *et al*. [46], the advancements in multimodal deep learning, particularly integrating ELMo word embeddings with transformers, have significantly improved image description capabilities.

2.1.5. CoVe

Contextualized Word Vectors (CoVe) use a profound Long Short-Term Memory (LSTM) encoder derived from a cognitive sequence-to-sequence model specializing in machine translation [47]. This method is cast to provide word vectors with context. CoVe word embeddings are a mechanism for processing the entire input sequence. From an architectural standpoint, this model is characterized by its simplicity and lack of logic. The initial step is deleting the dual-layer, one-way LSTM encoder from the machine translation (MT) model. The process encrypts the static pre-trained GloVe embeddings used as context vectors. These context vectors are appended to the GloVe embeddings and provided as input to subsequent NLP tasks. CoVe has improved in numerous NLP charges, including sentiment analysis, question classification, entailment, and question answering [48]. Furthermore, it has brought attention to the dynamic manifestation of language.

2.1.6. ULMFit

The technique known as universal language model fine-tuning (ULMFit) was initially presented by Luong *et al*. [49]. This approach frequently employs language modelling, specifically utilizing LSTM networks to leverage extensive untagged statistics effectively. Precisely, the ULMFit model consists of three distinct phases:

- To acquire knowledge about linguistic

characteristics, a language model undergoes training using a substantial corpus of commonly used language.

- Subsequently, the model is refined by training it on a specific corpus of job-related texts, allowing it to understand job-specific language patterns better.
- Lastly, the model undergoes an advanced fine-tuning process, incorporating objective classification of job-related entities.

Moreover, the previously proposed two effective strategies, namely slanted triangular studying rates and discriminative fine-tuning, to enhance the fine-tuning process of objective-domain language models. Each neural network layer possesses distinct significance, with higher layers capturing semantic information and lower layers representing syntactic information. Hence, it is imperative to consider the unshared learning rate as the primary indicator of prejudicial fine-tuning. Ideally, the model can retain the knowledge acquired from the standard domain data while developing the latest features specific to the target domain. Additionally, the algorithm must exhibit rapid convergence towards an appropriate initial state during training, gradually refining the parameters. In pursuit of this objective, the proposal suggests using sloped triangular learning rates, characterized by a gradual initial increase followed by a subsequent linear decay. One main advantage of using a small learning rate is effectively preserving data in the pre-trained parameter. To address the issue of catastrophic forgetting, researchers proposed a gradual unfreezing approach known as the step-by-step unfreezing mechanism. This involves unfreezing the pre-trained model, starting from the last layer and progressing gradually. Additionally, three new fine-tuning tactics were introduced, which have gained popularity in subsequent research. As a result of these advancements, ULMFit has demonstrated superior performance to the present state-of-the-art models across seven text classification tasks.

2.1.7. XLNet

XLNet (eXtra Long Network) is a generalized autoregressive pretraining model for language understanding. It extends the Transformer-XL model and improves upon BERT by leveraging a permutation-based autoregressive approach to model word sequences. It allows XLNet to capture bidirectional context while maintaining the advantages of autoregressive models [50]. BERT is an autoencoding pre-training method connected to the latest autoregressive (AR) techniques used to calculate a text corpus's probability distribution using autoregressive models such as GPT and ELMo. The primary goal of BERT is to rebuild the unique data from corrupted input. Given that the compactness approximation is excluded from the objective of the BERT model, it can readily leverage bidirectional contexts for reconstruction purposes. Furthermore, this analysis aims to provide a comprehensive examination of the observed advantages of this method compared to the most recent augmented reality (AR) techniques. Nevertheless, using Artificial representations, such as [MASK], has indicated a discrepancy between the pre-training and fine-tuning processes, leading toward a need for more consistency between the two. Moreover, the BERT model adheres to the notion that the predicted tokens are independent, potentially compromising its ability to effectively capture long-range, high-order dependencies prevalent in natural language. A pre-trained non-specialized autoencoder (AE) strategy relies on transformer-xl to address such issues. This approach leverages bidirectional contexts to enhance the predictability of various factorization orders and surpasses the limitations of BERT through autoregressive preparation. Moreover, XLNet employs the permutation language modelling objective, combining the compensations of autoencoding and autoregressive approaches while mitigating their limitations [51].

2.1.8. BERT

The language representation model Bidirectional Encoder Representations from Transformers (BERT), developed by Devlin *et al*. [52], is designed to understand language by looking at both the left and right context of words in a sentence at the same time, making it more effective than earlier models [39, 44] that only considered one direction at a time. In the study, Devlin *et al*. [52] proposed a model founded upon a multilayer bi-directional transformer-encoder that serves as a contextualized word representation model. Unlike traditional sequential recurrence, this model employs parallel tending layers within the transformer neural network. The present model has undergone pre-training on two unsupervised tasks: The proposed

approach involves utilizing a covered language model, wherein approximately 15% of the tokens are unsystematically replaced with a unique "[MASK]" token. The standard is then proficient to predict the masked tickets. Additionally, a subsequent sentence prediction (NSP) task is employed, wherein the model is presented with a set of sentences and trained to identify whether the second sentence logically follows the first. This second task aims to gather more information on enduring or practical aspects. BERT is trained on a corpus of a book and text paragraphs sourced from the English language Wikipedia. The corpus contains approximately eight hundred words. Two sizes are available for pre-trained BERT models, namely BERT-base and BERT-large. BERT can be employed by directly using the pre-trained model on unannotated data or fine-tuning it on task-specific data. The pre-existing anonymized model and accompanying code for fine-tuning can be accessed through online platforms. The user's text needs more information to be rewritten academically. Numerous domain-specific iterations of BERT have undergone training or fine-tuning on text specific to a particular domain. Some examples of these iterations include:

• **BioBERT** is a modification of the BERT model specifically adapted for biomedical script analysis [53]. Its architecture has been modified and pre-trained using a large corpus of PubMed descriptions and PMC full-text snippets. The system is optimized for biomedical text mining tasks, including question answering, entity identification NER, and relation extraction.

• **ClinicalBERT** is proficient in clinical text from the publicly available mimic-iii database, which contains about 2 million clinical notes [54]. The model was introduced to the following types of messages:

i. Clinical BERT
ii. Clinical bioBERT
iii. Discharge summary BERT
iv. Discharge summary bioBERT

• **SciBERT** proficiently uses an arbitrary sample of 1.14 million semantic scholar full-text papers. SciBERT is a model that undergoes unsupervised training on various scientific publications from various fields. This pre-training helps boost its effectiveness in handling scientific NLP tasks [55]. There are four forms of Seibert:

i. The Cased (Both uppercase and lowercase vocabulary).
ii. The Uncased (Only lowercase vocabulary).
iii. Those models which are using BaseVocab.
iv. Those models are models using SciVocab and are trained from scratch.

2.1.9. MorphoRNN

Using word n-grams enables more efficient exploitation of the complex internal semantics. However, English is characterized by numerous meaningful affixes, including prefixes, roots, and suffixes. In a study, Sennrich *et al*. [56] introduced the concept of morphology to progress the learning process of word embeddings. A representation of the subword is obtained by training the fix with RNN. The embedding of the parent word is determined by considering all morphemes except those discussed by Xu and Liu [57], who focus on morphological aspects. RNN models linguistic units on a morpheme level instead of a word level. In their analysis, scholars consider the morpheme the fundamental natural language unit, conveying a unique vector to each morpheme for classification purposes. The embeddings of morphological texts are derived from the embeddings of their basic morphemes. An additional parent word embedding is derived by combining a stem and affix embedding.

2.1.10. MWE

Multi-Word Expressions (MWEs) are fixed or semi-fixed expressions that consist of multiple words but function as a single semantic unit [58]. Using word vector models to incorporate prior knowledge is a commonly employed technique for improving performance. The general practice represents a word's suffix, root, and prefix as separate tokens. The objective of MWE is to use a stylized approach in conveying the combined significance of a suffix and a prefix [57]. The model has been constructed based on the hypothesis that all meanings of morphemes in a token have equal help to the given structure of tokens, denoted as w = {w1, w2, ..., wn}. We aim to determine the meanings for each morpheme, characterized as mi, for wi (where i ranges from 1 to n). The term "mi" can be conceptually divided into three distinct components, namely "pi," "ri," and "si." These components represent the prefix denoting a collection or set, the root indicating a

group or set in addition to the suffix signifying a set of "wi." Hence, in cases where we serve as the contextual basis for wj, the altered representation of wi.

Furthermore, numerous studies focus on news at the sub-word (SW) level. In their research, Sennrich *et al*. [56] introduced the Byte Pair Encoding (BPE) approach, which involves merging frequently occurring neighbouring components within the subword to enhance word representation. Additionally, the authors demonstrated that their method outperformed alternative strategies in the milieu of neural machine translation tasks. Ustun *et al*. [59] introduced an enhanced logarithmic bilinear model (LBL) and emphasized its role in assigning morpheme labels. We have compared all pre-trained models in Table 1. The research demonstrated that this approach resulted in word embeddings that effectively preserved morphological interactions. Bian *et al*. [60] combined graphical and textual representations to enhance the effectiveness of word embeddings, demonstrating improvements through experiments involving word analogy, uniformity, and completion tasks. Their method employs a forward LSTM model to capture the prefix and root of a word and a reverse LSTM model to acquire the suffix and root, focusing on character-level information within a word. Cao and Rei [61] introduced a char2vec model using a Bidirectional Long Short-Term Memory (Bi-LSTM) network to generate embeddings for fictional representations, succeeding in morpheme boundary recovery and syntactic analogy tasks. Regarding morpheme boundary recovery, the researchers demonstrated that their morphological exploration was like that of specialized morphological investigators. Additionally, their research performed well in answering syntactic analogies. Kim *et al*. [62] introduced a novel approach that utilizes Convolutional Neural Networks (CNN) with max pooling for word embeddings. They also demonstrated that their proposed model could reduce the number of variables while enhancing performance.

2.2. Visualization of Word Embeddings

The newer word embedding techniques represent words in high-dimensional vector spaces, which allows them to learn subtle semantic relationships between words. The disadvantage of high-dimensional embeddings is that they are difficult to interpret. In most cases, such embeddings must be projected in a 2D or 3D space to facilitate critical analysis and interpretation. Another popular method of word-embedding visualization is t-SNE, which projects the embeddings into a lower dimension while trying to preserve their local structure. It has effectively shown semantic clustering, making word-embedding models more

Table 1. Comparison of pre-trained models.

Method	Architecture	Encoder	Decoder	Objective
Word2Vec	NN	No	No	Skip-gram and CBOW
GloVe	Matrix factorization	No	No	Global word-word co-occurrence Statistics
FastText	NN	No	No	Skip-gram and CBOW with sub-word information
ELMo	LSTM	Yes	Yes	Language modelling
CoVe	LSTM	Yes	No	Language modelling and word prediction
UMLFit	NN	No	No	Unsupervised machine learning of word embedding
XLNet	Transformer	Yes	No	Masked language modelling and next-sentence prediction
BERT	Transformer	Yes	No	Masked language modelling and next-sentence prediction
Morpho-RNN	NN	Yes	No	Language modelling with morphological information
MWE	NN	Yes	No	Language modelling with multi-word expression information

interpretable since related words congregate in low-dimensional spaces [63]. The effectiveness of the visualization method in many studies conducted on the performance of various word embedding algorithms, such as t-SNE. For example, t-SNE has been useful in visualizations to show clear clusters of word embeddings emanating from models such as Word2Vec and FastText [64]. These visualizations display groups of semantically similar words, offering insights into the quality of representations learned by recent research conducted by Bandyopadhyay *et al.* [65] through embeddings used in natural language understanding tasks.

These visualization techniques have also helped compare word embeddings from a classic model and those taken out of more advanced deep learning architectures, such as transformers. The research by Robinson and Pierce-Hoffman [66] have shown the importance of visualising the contextualized embeddings that transformer models create, such as BERT, where t-SNE and PCA have been used to contrast semantic similarities and differences amongst word contexts. Such a method is useful in understanding how modern embedding techniques handle word polysemy and context-dependent meanings, giving the embeddings more interpretability for downstream applications. Similarly, word embedding visualizations have discovered significant findings in health on the relationship of medical terms with their contexts. Therefore, recent efforts leveraged visualization methods to analyze embeddings derived from electronic health record data that exposed meaningful semantic relationships, helping in predictive modelling and decision-making processes [67]. The clustering patterns emerging from such visualizations have played an instrumental role in enhancing the interpretability of medical embeddings, particularly for identifying semantically related diagnostic phrases or treatment options.

Recent developments in embedding visualization have included MDS and UMAP, which are increasingly used with t-SNE for improved interpretability and scalability. These methods have showcased more intuitive visualizations of word embedding spaces, especially when dealing with large datasets or intricate models. Besides clustering words, they can also expose outliers, anomalies, and rare occurrences of words in embedding spaces, which gives further detailed insight into how different models represent such rare terms and contexts [68]. Word embedding visualization is an essential tool in embedding model analysis and interpretation, helping researchers further understand relationships and structures within data. Visualization techniques such as t-SNE, PCA, and UMAP remain vital in handling the quality of embeddings, especially in their evolution with sophisticated deep learning architectures. These visualizations allow intuition to more easily understand how embeddings encode semantic information of critical importance in developing natural language processing and beyond.

3. HISTORY OF PRE-TRAINED MODELS

Pre-training has consistently been regarded as a highly effective methodology for acquiring knowledge about the variables within deep neural networks, which are refined through fine-tuning processes for downstream tasks [69]. The year 2006 witnessed a significant advancement in deep learning, as it saw the resurgence of the acquisitive layer-wise unsupervised pre-training technique, which was subsequently combined with supervised fine-tuning [70]. In computer vision, it is a common habit to initially train models on the extensive ImageNet dataset, followed by fine-tuning on smaller datasets for specific tasks. This approach exhibits notable advantages compared to an unplanned initialization, as the model acquires comprehensive image features that can be leveraged across various vision-related tasks. In NLP, it has been demonstrated that Pre-trained Models (PTMs) trained on extensive corpora are helpful for multiple downstream NLP tasks, ranging from basic word embeddings to complex deep neural models.

3.1. Abbreviations and Acronyms

The practice of representative words as fixed-length vectors has a longstanding historical background. The concept of "modern" word embedding was first introduced in the ground-breaking research of neural network language models (NNLM). Collobert *et al*. [71] demonstrated that pre-trained word embeddings on unlabelled data can significantly enhance performance in NLP tasks. To tackle the issue of computational complexity, the researchers opted to train word embeddings using a pairwise top-ranking

job rather than relying on language modelling. This study represents the initial endeavour to acquire universal word embeddings that can be utilized for various tasks using unannotated data. According to the findings presented by Mikolov *et al*. [72], it has been revealed that deep neural networks do not yield significant benefits in developing effective word embeddings. Skip-gram (SG) and Continuous bag-of-words (CBOW) models are two shallow Architectures proposed by the authors. Despite their simplicity, these methods can acquire high-quality word embeddings that capture the underlying syntactic and semantic similarities between words. Word2Vec is widely recognized as one of the most standard NLP model implementations. It facilitates the use of pre-trained word embeddings for multiple NLP tasks. Moreover, GloVe [32] is a popular model for gaining pre-trained word embeddings. These embeddings are derived from global word-word co-occurrence facts extracted from a corpus of considerable size. While pre-trained word embeddings are valuable in NLP charges, they often lack context sensitivity and are primarily trained using shallow models. When employed in a subsequent project, the entirety of the classic must still be acquired anew. Numerous researchers also endeavour to acquire embeddings of textual elements such as reading materials, sentences, or reports during the concurrent time frame. Examples of these efforts include the utilization of paragraph vectors [73], skip-thought vectors [74], and context2vec [75]. In contrast to their contemporary counterparts, these rudimentary sentence embedding models aim to transform entered sentences into a vector representation of stable dimensions instead of generating contextual words for individual tokens.

3.2. Second-generation PTMs: Pre-trained Contextual Encoders

Given that NLP endeavours extend outside the scope of individual words, it is customary to pre-train neural encoders at the sentence level or higher. The vectors produced by neural encoders, commonly called contextual word embeddings, modify the semantic representation of texts based on their surrounding perspective. The primary successful example of PTM for NLP was introduced by Dai and Le [76]. The authors employ a language model LM or a system autoencoder to digitize extended short-term memory networks LSTMs. They observe that pre-training can enhance the guidance process and improve the inductive reasoning capabilities of LSTMs in various text classification tasks. Liu *et al*. [77] conducted pre-training of a shared LSTM encoder using a language model LM and subsequently fine-tuned it within the multi-task learning MTL framework. The authors observed that incorporating pre-training and fine-tuning techniques can significantly enhance the enactment of MTL in various text classification tasks. Please provide more information or specify what you want me to rewrite academically. It has been observed that the performance of seq2seq models can be meaningfully enriched using unattended pre-training [78]. The encoder and decoder weights are initialized using pre-trained weights from two language models. Subsequently, these weights are fine-tuned using tagged data.

Furthermore, the contextual encoder can be pre-trained with a language model (LM). In a study, McCann *et al*. [47] utilized a pre-trained deep LSTM encoder derived from an attentional sequence-to-sequence model employed in machine translation (MT). The performance of various common NLP objectives can be enhanced by utilizing the context vectors (CoVe) generated by the pre-trained encoder. The contemporary post-translational modifications (PTMs) have advanced significantly compared to their precursor PTMs. They now possess enhanced proficiency in handling extensive corpora, employ more robust and intricate architectures such as transformers, and engage in novel pre-training tasks.

The model proposed by Peters *et al.* [44] is a pre-trained two-layer LSTM encoder incorporating a bidirectional language model (BiLM). This BiLM comprises both a forward language model and a reversed language model. The contextual representations generated by the Pre-trained BiLM, ELMo (embeddings from language models), have yielded significant improvements across diverse NLP tasks. Please provide more information or specify what you want me to rewrite academically. The word meaning was determined using contextual string embeddings pre-trained with a character-level language model [79]. However, these two pre-trained models (PTMs) are commonly employed as feature extractors to generate contextual word embeddings. These embeddings are then utilized as input for the primary model in downstream tasks.

The variables in question are held constant, while the previous model's remaining variables continue to be trained from the beginning. The ULMFit (universal language model fine-tuning) approach, as described by Howard and Ruder in [39], aimed to fine-tune a pre-trained language model (LM) for the task of text classification (TC). This method achieved state-of-the-art results on six commonly used TC datasets. The ULM-fit methodology comprises three distinct phases: (i) initial training of the language model (LM) using common-domain data; (ii) subsequent fine-tuning of the LM using target-specific data; and (iii) further fine-tuning of the LM on the specific target task. ULMFit additionally incorporates several valuable techniques for fine-tuning, namely prejudicial fine-tuning, sloped triangular learning rates, and step-by-step releasing.

Lately, there has been a growing recognition of the significant capabilities of deep pre-trained models (PTMs) in acquiring universal language representations. Prominent examples include OpenAI GPT (generative pre-training) and BERT (bidirectional encoder representation from the transformer). Additionally, there is a rising trend of introducing a greater variety of self-supervised tasks to enhance the acquaintance acquisition of these pre-trained models from vast text corpora. Following the emergence of ULMFit and BERT, fine-tuning has emerged as the prevailing method for adapting pre-trained models to suit downstream tasks.

4. CROSS-LINGUAL WORD EMBEDDINGS

When observing the presence of various languages, it becomes evident that approximately seven thousand languages are currently in use. However, it is essential to note that only a limited number of languages possess abundant human-interpreted resources. The task at hand involves acquiring cross-lingual shift learning of word embeddings. We employ a model trained on languages with great linguistic resources to accomplish this. This model then maps the input embeddings of languages with limited resources onto a joint semantic space. These embeddings are commonly referred to as cross-lingual word embeddings [80]. Founded on the classification of monolingual embeddings, cross-lingual embedding learning approaches can be categorized as dynamic or static. The static method has received considerable attention in recent studies, while numerous studies are currently investigating the active process. Additionally, these approaches can be categorized into offline and online categories based on the training objective. In general, online strategies aim to optimize cross-lingual and monolingual objectives simultaneously. Conversely, offline methods involve utilizing pre-trained monolingual word embeddings from different languages as involvement and mapping them into a shared semantic space [81]. As a survey by Ruder *et al*. [80] noted, most cross-lingual word embedding models are optimized using similar objective functions, and differences in performance often stem from data requirements rather than architecture.

4.1. Static Cross-lingual Word Embeddings

When examining still embeddings, it is observed that specific methods involve learning language models for both objective and source languages. These methods then collectively enhance their respective objectives by utilizing cross-lingual goals. An approach was proposed by Klementiev *et al*. [82] to acquire bilingual word embeddings and word alignments simultaneously. Subsequently, using the monolingual skip-gram model, the researchers endeavoured to develop proficiency in bilingual embeddings, encompassing both sentence and word-level alignments. The model proposed by Lample and Conneau [83] aims to acquire bilingual embeddings that enhance the semantic coherence of sentence pairs with a specific orientation.

Guan *et al*. [84] proposed a methodology for leveraging document-aligned similar corpora to acquire bilingual embeddings. The absorption of two aligned documents made this into a pseudo-bilingual paper, which was then used to train a skip-gram model. Offline methods involve learning a projection that facilitates the transformation of the source language's vector space to the target language's vector space. The acquisition of such a matrix can be achieved through a supervised approach, wherein the objective is to minimize the squared Euclidean distance, also known as the Mean Squared Error (MSE), between the target word embedding of a translated word and the converted source word embedding. The matrix can typically be acquired by replacing the mean squared error with a max-margin hinge loss or by employing

singular value decomposition. Bengio *et al*. [85] have introduced an alternative approach for aligning word embeddings in target and source languages. This method utilizes canonical correlation analysis (CCA) to project the embeddings onto a shared space. The researchers discovered that incorporating cross-lingual embeddings into dependency analysis and comprehensive supplementary features such as lexical characteristics and word clusters yielded significant performance improvements. The authors extended their research efforts and incorporated nonlinearity into the mapping procedure.

In addition to supervised approaches to cross-lingual embedding learning, unsupervised methods have also resigned promising outcomes. The initial step involved constructing a bilingual dictionary using adversarial learning techniques, as described by Radford *et al*. [86]. Subsequently, bilingual embeddings were generated, along with a modification approach. In another study, Peters *et al*. [44] introduced a similar framework that adopts a two-step approach for acquiring multilingual embeddings. Notably, this framework considers the interdependencies among numerous languages, a factor that previous research needs to consider. To address the challenges associated with uncertainty in acquiring cross-lingual embeddings for reserved language sets, Wang *et al*. [87] introduced a resilient framework. This framework enables learning a common multilingual embedding space by iteratively incorporating additional languages into the existing space.

4.2. Dynamic Cross-lingual Word Embeddings

Many researchers have explored and shared their findings and studies on dynamic word embeddings with cross-lingual transfer, drawing inspiration from the significant advancements made in active word embeddings for monolingual applications. In a study, one of the online approaches examined by Akbik *et al*. [79] focuses on ELMo, a model that heavily on which it heavily relies. This approach aims to create a polyglot model that captures character-level information from multilingual data to generate relative representations. Lample and Conneau [83] primarily centred on BERT and its objectives, explicitly examining the utilization of cross-lingual supervision from parallel data to investigate cross-lingual language models (XMLS). This approach yielded highly favourable results on various cross-lingual tasks, establishing a new benchmark in the field. Subsequently, the researchers demonstrated that large-scale pre-trained multilingual language models significantly improved evaluating cross-lingual transfer tasks. It highlights the potential of multilingual modelling, excluding compromising the evaluation of individual language-specific outcomes.

In contrast, offline methodologies have employed linear projection to generate contextualized pre-trained embeddings [60]. The approach used in our study involved utilizing averaged contextualized embeddings as a reference point for individual words and acquiring knowledge of the shift matrix within the reference space. Wang *et al*. [87] introduced a method for directly acquiring this transformation within the given context, preserving word sense in cross-lingual dynamic embeddings. McCann *et al*. [47] evaluated current methods for dynamic cross-lingual embeddings and demonstrated their significant potential in enhancing cross-lingual dependency parsing. Additionally, they have shown that online methodologies exhibit superior encoding of cross-lingual lexical correspondence compared to offline techniques.

4.3. Multilingual Word Embeddings

In addition to the practice of transferring embedding models from resource-rich to low-resource languages through a plan, there have also been efforts to train embedding models in multiple languages simultaneously. In their study, Bengio *et al*. [85] introduced a novel language model called multi-BERT. This model was trained on a collection of mono-lingual Wikipedia corpora from a total of 104 languages. Notably, the model exhibited exceptional performance in zero-shot cross-lingual model shifts. The researchers demonstrated through a diverse set of investigative experiments that the multi-BERT model possesses the ability to seamlessly transition between languages, even in the absence of any explicit lexical cues. It is achieved by effectively capturing and understanding multilingual contexts.

In addition, Wag *et al*. [88] investigated the multi-BERT model's generalisation ability. They devised an alternative method for transferring lexical information from a monolingual model

to new languages. The outcome challenges the prevailing notion that multi-BERT exhibits strong generalization capabilities due to its utilization of a shared sub-word vocabulary and simultaneous training across multiple languages. In contrast, it was suggested that the monolingual representations should acquire abstract concepts that can be applied to various languages.

5. EMBEDDING FOR OUT-OF-VOCABULARY WORDS

The word2vec model is known for its simplicity and efficiency in learning semantic representations of words from large data files [89]. However, it has limitations in learning embeddings for OOV texts. OOV words can be categorized into terms not in the open vocabulary and words not encountered in the current corpus [90]. OOV texts can be broadly classified into three forms. (i) The dynamic lexicon, precisely online terminology, is developing continuously. (ii) Proper names refer to specific entities such as geographical locations, organizations, individuals, automated expressions, and temporal references. (iii) Investigating the Terminology of Research Fields and Professional Titles. In academic discourse, various terminologies encompass elements, such as the titles of literary works or newly created artistic pieces, including documentaries or novels. In most instances, expanding one's vocabulary is the optimal approach. In addition to linguistic processing, there is an expansion of vocabulary, enabling us to delve into the realm of OOV words, particularly those that fall within the extensive range of less favoured components. Words that are often unfamiliar are typically ignored, removed, or replaced with an 'unknown' tag (UNK), which is an insufficient solution. Addressing the challenges posed by OOV words is crucial. Recently, various neural network-based models, such as FastText, MorphoRNN, and MWE, have been developed to tackle this issue effectively.

6. DATASETS AND EVALUATION FRAMEWORKS

The measurement assessment system for current word embeddings can be categorized based on intrinsic and extrinsic evaluation. However, these evaluation approaches have faced extensive criticism in the existing literature. In this fragment, we present a brief overview of the two types of evaluations discussed in the previous quarter and direct readers to recent research studies for a thorough explanation and analysis.

6.1. Intrinsic Evaluation

Intrinsic evaluations establish relationships between words by assessing their syntactic or semantic properties, relying on artificial assessments as a basis. Through careful observation of the methods employed to acquire these assessments, it is possible to categorize such approaches into two distinct types: absolute and relative intrinsic evaluation. In the initial category, the individual reviews are gathered before victimization, serving as a reference point for word embedding methodologies. In intrinsic evaluation, the comparative approach uses accessors to assess word embeddings candidly grounded on their performance in an exact word relation objective or charge [91]. Due to its independence from human involvement or interaction, the absolute form of intrinsic evaluation is frequently employed alongside comparative inherent evaluation. In the following section, we will briefly introduce several well-known assessment techniques. The method used for assessing semantic similarity, known as similarity checking, is widely utilized due to its effectiveness in determining the relationship between word distances in human heuristic judgments and embedding space. The test sets commonly employed in current research include WordSim-353 [92], Mammals, Entities, Natural kinds (MEN) [93], and SimVerb-3500 [94].

The word analogy technique has gained significant recognition due to its integration with the well-known CBOW and Skip-gram representations. In this context, the embeddings of three words, w, x, and y, are employed to forecast the word z. The objective is to identify z in a manner that maintains the exact relationship between w and x as y and z. As an illustration, let us consider the scenario where w represents Pakistan, b represents Islamabad, and c represents India. In this case, d would correspond to Delhi. Prominent examples of trial sets of this nature include the WordRep, Microsoft Research Syntactic Analogies Dataset, and Google Analogy [95]. The synonym detection technique assesses the capacity of embeddings to accurately identify the most similar word to a given word from a pool of candidates. When considering a specific goal

word, such as "levied," one must select among options such as "imposed" (correct), "believed," "requested," and "correlated." The datasets that could be utilized in this methodology encompass the Test of English as a Foreign Language (TOEFL), English as a Second Language (ESL), and Reading and Writing for Academic Purposes (RDWP) [96].

The word embedding space in the concept categorization technique is evaluated through clustering. This task categorizes a set of specific terms into distinct subsets. For instance, the words "goat" and "dog" will be classified under the mammal category, while "oranges" and "grapes" will be categorized as fruits [97]. The identification of verb-noun pairs in textual data is facilitated by utilizing a technique known as the Sectional Preference method. Commonly used word embeddings can identify verb-noun pairs in which the noun is the subject or object of the verb. For example, the noun "humanity" is often used as the subject instead of the object of the verb "serve." Greenberg, Sayeed and Danberg (GDS) [98], and Ulrike and Pado (UP) [99] are commonly employed lexical sets.

6.2. Extrinsic Evaluation

Word embeddings are used as feedback for downstream tasks and to measure the influence of these tasks using specific metrics in extrinsic evaluations. Word embeddings have demonstrated significant applicability across various functions in the NLP domain. These embeddings can be utilized for multiple parts, as perceiving all such tasks as non-essential assessments is theoretically possible. One category of downstream tasks within this field encompasses language modelling, named entity recognition, POS tagging, chunking, machine reading comprehension, sentiment analysis, semantic role labelling, dependency parsing, machine translation, and natural language inference [100]. The assumption inherent in these non-essential evaluations is that word embeddings that yield positive results in one task will also deliver positive results in other studies. This assumption has been extensively explored and analyzed in the existing literature. Empirical observations have provided evidence that distinct NLP tasks prefer specific embeddings. Therefore, although extrinsic evaluations can help compare embeddings about a particular mission or objective, they are not mentioned as metrics for the overall review of word embeddings' excellence.

7. CONCLUSIONS

In this review, we reflect upon the evolution and impact of word embeddings within the domain of NLP. Word embeddings have formed a crucial basis for carrying out manifold tasks in NLP and have revolutionized the way text is represented; hence, semantic understanding has proficiently been achieved. Through such a detailed critical analysis, we have showcased their relative strengths and limitations while considering a host of various NLP tasks, including but not limited to sentiment analyses and machine translations. The paper has critically discussed the evolution of word embeddings from static to contextual, from the traditional Word2Vec and GloVe models to more advanced BERT and ELMo models.

Such a comparison highlights the advantages of contextualized embeddings well, which even pushed the limit of word representation further by incorporating dynamic context and improving performance on downstream tasks. However, just like any other machine learning model, embedding biases, capturing long-range dependencies, and inefficiency with out-of-vocabulary words remain critical points of concern that continue to drive research and innovation.

We also reviewed some applications of cross-lingual embedding and why multilingual models contribute to more effective language transfer and alignment across diverse languages. The emergent techniques for handling OOV words, dynamically changing embeddings, and domain-specific models have opened new vistas for applying NLP, further showcasing versatility and scalability.

Overall, word embeddings have given an effective way of encoding semantic information that furthers NLP. However, much research still needs to be carried out regarding critical challenges, such as bias, generalization toward poor-resource languages, and handling linguistic complexities. Ongoing developments of more sophisticated models and hybrid approaches are bound to shape the future of NLP, enabling highly accurate and meaningful language understanding in and out of general and specialist contexts.

8. CONFLICT OF INTEREST

The authors declare no conflict of interest.

9. REFERENCES

1. C. Liu and K.K.H. Chung. The relationships between paired associate learning and Chinese word writing in kindergarten children. *Reading and Writing* 34(8): 2127-2148 (2021).
2. P. Aceves and J.A. Evans. Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science* 35(3): 788-814 (2024).
3. A. Berenguer, J.-N. Maz'on, and D. Tom'as. Word embeddings for retrieving tabular data from research publications. *Machine Learning* 113(4): 2227-2248 (2024).
4. F. Incitti, F. Urli, and L. Snidaro. Beyond word embeddings: A survey. *Information Fusion* 89: 418-436 (2023).
5. M. Toshevska. The Ability of Word Embeddings to Capture Word Similarities. *International Journal on Natural Language Computing (IJNLC)* 9(3): 25-42 (2020).
6. M.-C. Hung, P.-H. Hung, X.-J. Kuang, and S.-K. Lin. Intelligent portfolio construction via news sentiment analysis. *International Review of Economics and Finance* 89: 605-617 (2024).
7. K. Das, F. Abid, J. Rasheed, Kamlish, T. Asuroglu, S. Alsubai, and S. Soomro. Enhancing Communication Accessibility: UrSL-CNN Approach to Urdu Sign Language Translation for Hearing-Impaired Individuals. *CMES-Computer Modeling in Engineering and Sciences* 141(1): 689-711 (2024).
8. B. Lal, R. Gravina, F. Spagnolo, and P. Corsonello. Compressed sensing approach for physiological signals: A review. *IEEE Sensors Journal* 23(6): 5513-5534 (2023).
9. A. Baloch, T.D. Memon, F. Memon, B. Lal, V. Viyas, and T. Jan. Hardware synthesis and performance analysis of intelligent transportation using canny edge detection algorithm. *International Journal of Engineering and Manufacturing* 11(4): 22-32 (2021).
10. E. Çano and M. Morisio. Word Embeddings for Sentiment Analysis: A Comprehensive Empirical Survey. *Preprint ArXiv* 1: 1902.00753 (2019).
11. F.K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics* 100: 100057 (2019).
12. A. Agarwal, B. Agarwal, and P. Harjule. Understanding the Role of Feature Engineering in Fake News Detection. In: *Soft Computing: Theories and Applications: Proceedings of SoCTA 2021*, *Singapore* pp. 769-789 (2022).
13. R.A. Stein, P.A. Jaques, and J.F. Valiati. An analysis of hierarchical text classification using word embeddings. *Information Sciences* 471: 216-232 (2017).
14. J.E. Font and M.R. Costa-Jussà. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. *Preprint ArXiv* 2: 1901.03116 (2019).
15. S.M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems With Applications* 117: 139–147 (2019).
16. Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. Dynamic Word Embeddings for Evolving Semantic Discovery. *Preprint ArXiv* 2: 1703-00607 (2018).
17. J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender Bias in Contextualized Word Embeddings. *Preprint ArXiv* 1: 1904.0331 (2019).
18. Q. Du, N. Li, W. Liu, D. Sun, S. Yang, and F. Yue. A Topic Recognition Method of News Text Based on Word Embedding Enhancement. *Computational Intelligence and Neuroscience* 2022(1): 4582480 (2022).
19. D. Suhartono, K. Purwandari, N.H. Jeremy, S. Philip, P. Arisaputra, and I.H. Parmonangan. Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews. *Procedia Computer Science* 216: 664-671 (2023).
20. S. Haller, A. Aldea, C. Seifert, and N. Strisciuglio. Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers. *Preprint ArXiv* 1: 2204.03503 (2022).
21. A. Çalışkan, P.P. Ajay, T. Charlesworth, R. Wolfe, and M.R. Banaji. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency Syntax, and Semantics. *Preprint ArXiv* 1: 2206.03390 (2022).
22. X. Tang, Y. Zhou, and D. Bollegala. Learning Dynamic Contextualised Word Embeddings via Template-based Temporal Adaptation. *Preprint ArXiv* 3: 2208-10734 (2023).
23. K. Alnajjar, M. Hämäläinen, and J. Rueter. Sentiment Analysis Using Aligned Word

Embeddings for Uralic Languages. *Preprint ArXiv* 1: 2305.15380 (2023).
24. H. Yen and W. Jeon. Improvements to Embedding-Matching Acoustic-to-Word ASR Using Multiple-Hypothesis Pronunciation-Based Embeddings. *Preprint ArXiv* 2: 2210.16726 (2023).
25. J. Engler, S. Sikdar, M. Lutz, and M. Strohmaier. SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings. *Preprint ArXiv* 1: 2301.04704 (2023).
26. R. Schiffers, D. Kern, and D. Hienert. Evaluation of Word Embeddings for the Social Sciences. *Preprint ArXiv* 1: 2302.06174 (2023).
27. O. Zaland, M. Abulaish, and M. Fazil. A Comprehensive Empirical Evaluation of Existing Word Embedding Approaches. *Preprint ArXiv* 2: 2303.07196 (2024).
28. P.J. Worth. Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science* 13(1): 1-21 (2023).
29. K. Das and Kamlish. Enhancing Automated Text Summarization: A Survey and Novel Method with Semantic Information for Domain-Specific Summaries. *Journal of Computing & Biomedical Informatics* 5(2): 102-113 (2023).
30. Z.F. Abro, S.U. Rehman, K. Das, and A. Goswami. An Analysis of Deep Neural Network for Recommending Developers to Fix Reported Bugs. *Europan Journal of Science and Technology* (24): 375-379 (2021).
31. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Preprint ArXiv* 3: 1301.3781 (2013).
32. J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar* (25-29 October, 2014) pp. 1532-1543 (2014).
33. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135-146 (2017).
34. H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli. Exploring Chemical Space Using Natural Language Processing Methodologies for Drug Discovery. *Preprint ArXiv* 1: 2002-06053 (2020).
35. F. Torregrossa, R. Allesiardo, V. Claveau, N. Kooli, and G. Gravier. A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics* 11(2): 85-103 (2021).
36. G. Curto, M.F.J. Acosta, F. Comim, and B. Garcia-Zapirain. Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI and Society* 39(2): 617-632 (2024).
37. P. Rakshit and A. Sarkar. A supervised deep learning-based sentiment analysis by the implementation of Word2Vec and GloVe Embedding techniques. *Multimedia Tools and Applications* pp. 1-34 (2024).
38. M. Greeshma and P. Simon. Bidirectional Gated Recurrent Unit with Glove Embedding and Attention Mechanism for Movie Review Classification. *Procedia Computer Science* 233: 528-536 (2024).
39. J. Howard and S. Ruder. Universal Language Model Fine-Tuning for Text Classification. *Preprint ArXiv* 5: 1801.06146 (2018).
40. A. Faruq, M. Lestandy, and A. Nugraha. Analyzing Reddit Data: Hybrid Model for Depression Sentiment using FastText Embedding. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* 8(2): 288-297 (2024).
41. S.S. Alqahtani. Security bug reports classification using fasttext. *International Journal of Information Security* 23(2): 1343-1358 (2024).
42. D. Voskergian, R. Jayousi, and M. Yousef. Enhanced TextNetTopics for Text Classification Using the GSM Approach with Filtered fastText-Based LDA Topics and RF-Based Topic Scoring: fasTNT. *Applied Science* 14(19): 8914 (2024).
43. N.A. Nasution, E.B. Nababan, and H. Mawengkang. Comparing LSTM Algorithm with Word Embedding: FastText and Word2Vec in Bahasa Batak-English Translation. *12th International Conference on Information and Communication Technology (ICoICT)*, *Bandung, Indonesia* (7 Aug - 8 Aug, 2024) pp. 306-313 (2024).
44. M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *New Orleans, Louisiana* (1-6 June, 2018) pp. 2227-2237 (2018).
45. X. Cheng, T. Mei, Y. Zi, Q. Wang, Z. Gao, and H. Yang. Algorithm Research of ELMo Word Embedding and Deep Learning Multimodal Transformer in Image Description. *Preprint ArXiv*: 2408.06357 (2024).
46. L. Rong, Y. Ding, M. Wang, A.E. Saddik, and M.S.

Hossain. A Multi-Modal ELMo Model for Image Sentiment Recognition of Consumer Data. *IEEE Transactions on Consumer Electronics* 7(1): 3697-3708 (2024).

47. B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in Translation: Contextualized Word Vectors. *Advances in Neural Information Processing Systems* 30: 1-12 (2017).
48. W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal* 4: 100026 (2023).
49. T. Luong, R. Socher, and C.D. Manning. Better Word Representations with Recursive Neural Networks for Morphology. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, *Sofia, Bulgaria* (8-9 August, 2013) pp. 104–113 (2013).
50. Z. Yang. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Preprint ArXiv* 2: 1906.08237 (2019).
51. A.F. Adoma, N.-M. Henry, and W. Chen. Comparative analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. *17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, *IEEE* pp. 117-121 (2020).
52. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint ArXiv* 2: 1810.04805 (2018).
53. K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint ArXiv* 3: 1904.05342 (2019).
54. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4): 1234-1240 (2019).
55. I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. *Preprint ArXiv* 3: 1903.10676 (2019).
56. R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. *Preprint ArXiv* 5: 1508.07909 (2016).
57. Y. Xu and J. Liu. Implicitly Incorporating Morphological Information into Word Embedding. *Preprint ArXiv* 3: 1701.02481 (2017).
58. T. Baldwin and S.N. Kim. Multiword expressions. In: Handbook of Natural Language Processing. 2nd Edition. N. Indurkhya and F.J. Damerau (Eds.). *Taylor and Frances Group* pp. 267-292 (2010).
59. A. Üstün, M. Kurfalı, and B. Can. Characters or Morphemes: How to Represent Words?. *Proceedings of the 3rd Workshop on Representation Learning for NLP, Melbourne, Australia* (20 July 2018) pp. 144-153 (2018).
60. J. Bian, B. Gao, and T.-Y. Liu. Knowledge-powered deep Learning for word embedding. In: Machine Learning and Knowledge Discovery in Databases. T. Calders, F. Esposito, E. Hüllermeier, and R. Meo (eds). *Springer, Berlin, Heidelberg* pp. 132–148 (2014).
61. K. Cao and M. Rei. A joint Model for Word Embedding and Word Morphology. *Preprint ArXiv* 1: 1606.02601 (2016).
62. Y. Kim, Y. Jernite, D. Sontag, and A.M. Rush. Character-Aware Neural Language Models. *Preprint ArXiv* 4: 1508.06615 (2016).
63. D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F.B. Viegas, and M. Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *Preprint ArXiv* 1: 1611.05469 (2016).
64. S. Liu, P.-T. Bremer, J.J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, and V. Pascucci. Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics* 24(1): 553-562 (2018).
65. S. Bandyopadhyay, J. Xu, N. Pawar, and D. Touretzky. Interactive visualizations of word embeddings for k-12 students. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(11): 12713-12720 (2022).
66. I. Robinson and E. Pierce-Hoffman. Tree-SNE: Hierarchical clustering and visualization using t-SNE. *Preprint ArXiv* 1: 2002.05687 (2020).
67. N. Oubenali, S. Messaoud, A. Filiot, A. Lamer, and P. Andrey. Visualization of medical concepts represented using word embeddings: a scoping review. *BMC Medical Informatics and Decision Making* 22(1): 83 (2022).
68. M. Gniewkowski and T. Walkowiak. Assessment of document similarity visualization methods. In: Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science. Z. Vetulani, P. Paroubek, and M. Kubis (Eds.). *Springer, Cham* pp. 348-363 (2019).
69. X. Han, Z. Zhang, N. Ding, Y. Gu, and *et al*. Pre-trained models: Past, present and future. *AI Open*

2: 225-250 (2021).

70. D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, and P. Vincent. Why Does Unsupervised Pre-training Help Deep Learning?. *Journal of Machine Learning Research* 1: 625-660 (2010).
71. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural Language Processing (almost) from Scratch. *Preprint ArXiv* 1: 1103.0398 (2011).
72. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *Preprint ArXiv* 1: 1310-4556 (2013).
73. Q.V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. *Preprint ArXiv* 2: 1405-4053 (2014).
74. J. Pilault, J. Park, and C. Pal. On the impressive performance of randomly weighted encoders in summarization tasks. *Preprint ArXiv* 1: 2002.09084 (2020).
75. O. Melamud, J. Goldberger, and I. Dagan. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin Germany* (7-12 August 2016) pp. 51-61 (2016).
76. A.M. Dai and Q.V. Le. Semi-supervised Sequence Learning. *Preprint ArXiv* 1: 1511-01432 (2015).
77. P. Liu, X. Qiu, and X. Huang. Recurrent Neural Network for Text Classification with Multi-Task Learning. *Preprint ArXiv* 1: 1605-05101 (2016).
78. P. Ramachandran, P.J. Liu, and Q.V. Le. Unsupervised Pretraining for Sequence to Sequence Learning. *Preprint ArXiv* 2: 1611.02683 (2018).
79. A. Akbik, D. Blythe, and R. Vollgraf. Contextual String Embeddings for Sequence Labeling. *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA* (20-26 August 2018) pp. 1638-1649 (2018).
80. S. Ruder, I. Vulic, and A. ogaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65: 569-631 (2019).
81. S. Conia and R. Navigli. Conception: Multilingually-Enhanced, Human-Readable Concept Vector Representations. *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (online)* (8-13 December 2020) pp. 3268-3284 (2020).
82. A. Klementiev, I. Titov, and B. Bhattarai. Inducing Crosslingual Distributed Representations of Words. *Proceedings of COLING 2012, Mumbai, India* (December 2012) pp. 1459-1474 (2012).
83. G. Lample and A. Conneau. Cross-lingual Language Model Pretraining. *Preprint ArXiv* 1: 1901.07291 (2019).
84. J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics* 8: 93-108 (2020).
85. Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural Probabilistic language model. *Journal of Machine Learning Research* 3: 1137-1155 (2003).
86. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1(8): 9 (2019).
87. W. Wang, B. Bi, M. Yan, C. Wu, Z. Bao, J. Xia, L. Peng, and L. Si. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. *Preprint ArXiv* 3: 1908.04577 (2019).
88. S. Wang, W. Zhou, and C. Jiang. A survey of word embeddings based on deep learning. *Computing* 102: 717–740 (2020).
89. L. Ma and Y. Zhang. Using Word2Vec to process big text data. *2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA* (29 October-01 November 2015) pp. 2895-2897 (2015).
90. O. Kwon, D. Kim, S.-R. Lee, J. Choi, and S. Lee. Handling Out-of-Vocabulary Problem in Hangeul Word Embeddings. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, *Online* (2021) pp. 3213-3221 (2021).
91. A. Gladkova and A. Drozd. Intrinsic evaluations of word embeddings: What can we do better?. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Berlin, Germany* (12 August 2016) pp. 36-42 (2016).
92. E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A Study on Similarity and Relatedness Using Distributional and Wordnet-based Approaches. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistic*, *Boulder, Colorado* (June 2009) pp. 19-27 (2009).
93. E. Bruni, N.-K. Tran, and M. Baroni. Multimodal Distributional Semantics. *Journal of Artificial*

Intelligence Research 49: 1-47 (2014).

94. D. Gerz, I. Vulic, F. Hill, R. Reichart, and A. Korhonen. Simverb-3500: A Large-Scale Evaluation Set of Verb Similarity. *Preprint ArXiv* 4:1608.00869 (2016).
95. B. Gao, J. Bian, and T.-Y. Liu. Wordrep: A Benchmark for Research on Learning Word Representations. *Preprint ArXiv* 1: 1407.1640 (2014).
96. S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. Methods and Datasets for the Evaluation of Semantic Measures. *Semantic Similarity from Natural Language and Ontology Analysis*: 131-157 (2015).
97. B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo. Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing* 8: e19 (2019).
98. C. Greenberg, V. Demberg, and A. Sayeed. Verb polysemy and frequency effects in thematic fit modeling. *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, *Denver, Colorado* (4 June 2015) pp. 48-57 (2015).
99. S. Pado and M. Lapata. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics* 33(2): 161-199 (2007).
100. F. Nooralahzadeh, L. Ovrelid, and J.T. onning. Evaluation of Domain-specific Word Embeddings using Knowledge Resources. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, *Miyazaki, Japan* (7-12 May2018) (2018).