DeepGloVe: Extending GloVe Word Embeddings for Low-Resource African Languages

Anonymous ACL submission

Abstract

Learning high quality word embeddings for low resource African languages remains challenging due to limited text data, complex morphology, and few linguistic resources. Word embedding approaches, such as GloVe and Word2Vec, often struggle to encode semantic relationships under these conditions. In this paper, we propose DeepGloVe, a deep extension of the GloVe framework tailored for low-resource African languages. The architecture incorporates modifications to co-occurrence weighting, implementing a function that enhances the significance of infrequent yet informative co-occurrences, as well as the integration of a deep neural network architecture to learn richer semantic representations from global co-occurrence information. We evaluate the proposed model on two low resource African languages Sepedi (nso) and Setswana (tsn), using intrinsic semantic similarity benchmarks. The experimental result demonstrated that DeepGloVe methods consistently outperform strong baseline models, including W2V-SG, W2V-CBOW, and standard GloVe, in both quantitative and qualitative evaluation for both benchmark datasets. These findings indicate that deep representation learning improves semantic modeling for low resource African languages.

1 Introduction

Low-resource languages are languages that lack sufficient digitized texts, tools, and annotated datasets; due to these, many African languages are categorized as low-resource languages in the field of Natural Language Processing (NLP) (Ranathunga et al., 2023). This problem makes it harder to test NLP models thoroughly, especially for languages with few resources, and it creates a gap that makes NLP technologies less useful. This makes it impossible to use new technologies in areas like Q&A systems, neural machine translation, and large language models (Cho et al., 2014; Ranathunga et al., 2023; Devlin et al., 2019; Kwiatkowski et al., 2019; Patil et al., 2023) . To address these challenges in language technology, numerous issues require effective resolution. A fundamental aspect entails the representation of the symbolic characteristics of words, achieved through the application of word embedding techniques suitable for machine learning algorithms.

In modern NLP, word embeddings have emerged as a vital and fundamental resource; they enable models to encode syntactic and semantic information derived from their contextual surroundings within large text corpora. In recent years, word embeddings have been extensively employed in various NLP tasks, including sentiment analysis, named-entity recognition, and chatbots (Patil et al., 2023; Sezerer and Tekir, 2021) and showed great performance.

Word2Vec is one of the most popular word embedding methods that generates low-dimensional, real-valued vectors to represent word semantics (Mikolov et al., 2013b). These word vectors keep track of semantic features. This means that words with similar meanings can be found next to each other in the vector space. This method effectively incorporates semantic relationships among words, a capability not achievable by traditional methods, hence greatly improving the understanding of word meanings in a wide range of NLP applications. Nevertheless, Word2Vec does not work as well on low-resource languages, which do not have a lot of digital text data or linguistic resources, because it was made for high-resource languages.

Global vectors for word representation (GloVe) is another widely utilized word embedding algorithm designed for the generation of word embeddings (Pennington et al., 2014). It is a count-based model based on global statistical principles. This indicates that it takes into account the overall frequency of word co-occurrences, which results in

enhanced interpretability and improved efficiency in the domain of NLP. However, Glove's dependence on large, high-quality corpora creates a big problem, especially for low-resource African languages that often do not have enough digital text data and linguistic resources.

In summary, many of the existing African languages, including Sepedi and Setswana, exemplify this challenge, and the current research on African language word embeddings predominantly utilizes embedding models originally designed for high-resource languages. These approaches do not adequately address challenges such as significant data sparsity, complex morphology, and limited lexical coverage. Therefore, this work aims to address these limitations by extending the GloVe word embedding method to DeepGloVe, a method that incorporates linguistically informed word representations, thereby reducing data sparsity and mitigating morphological complexity. Specifically, we propose a weighted co-occurrence adjustment to mitigate data sparsity through learning from limited corpora, in conjunction with a deep learning-based factorization integration framework that leverages morphological structures to enhance embedding quality. Using Sepedi and Setswana languages with language codes nso and tsn, respectively, we demonstrate the potential of this approach to advance word representation learning for low-resource African languages.

The contributions of this paper are summarized as follows:

- We propose a modified GloVe formulation tailored for low-resource and morphologically complex African languages by modifying the co-occurrence weighting method to incorporate a weighting function that emphasizes infrequent yet informative co-occurrences.
- We develop a deep nonlinear extension of the GloVe framework by integrating a deep neural network architecture to enhance the semantic representation of morphologically rich words.
- We provide a comprehensive empirical evaluation of the proposed method against the baseline methods.

2 Related work

Word embeddings have established themselves as fundamental components of contemporary Natural Language Processing (NLP), allowing algorithms to encode the semantic and syntactic characteristics of words within dense continuous vector spaces. Early and influential embedding techniques, including Word2Vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017), have achieved significant success in numerous applications in high-resource languages such as English. These models depend on extensive corpora to acquire significant co-occurrence statistics and contextual patterns. However, their direct implementation in low-resource African languages continues to pose challenges owing to limited data availability, morphological complexity, and scarce linguistic resources.

Word2Vec generates word embeddings by predicting surrounding context words through either the Continuous Bag-of-Words (W2V-CBOW) or Skip-gram (W2V-SG) models (Mikolov et al., 2013a). Although effective in high-resource environments, Word2Vec considers words as indivisible units and requires extensive training data, thus restricting its efficacy for low resource African languages characterized by limited corpora and significant lexical diversity. Word2Vec for African languages continues to face challenges in accurately representing uncommon and morphologically complex words.

Global vectors for word representation (GloVe) (Pennington et al., 2014) is an unsupervised learning algorithm designed for the generation of word embeddings. GloVe is a count-based, global log-bilinear regression model that integrates two embedding techniques, namely global matrix factorization and local context window. The model is founded on the observation that the most suitable initial approach for word vector learning involves the ratios of co-occurrence probabilities rather than the probabilities themselves. In other words, the GloVe model is based on the premise that the ratios of co-occurrence probabilities between terms may encode certain relationships among them (Pennington et al., 2014). The model constructs a co-occurrence matrix X, where X_{ij} represents the number of times the word j appears in the context of the word i. Subsequently, the algorithm learns word vectors, ensuring their dot product closely reflects the logarithm of their co-occurrence probability. However, the GloVe formulation assumes access to extensive and balanced corpora in high-resource languages and demonstrates strong performance under such conditions, which are rarely met in low-resource African language datasets. Overall,

existing research on African language word embeddings primarily employs embedding models initially developed for high-resource languages, with limited adaptation to low-resource contexts. These methods do not sufficiently confront challenges such as severe data sparsity, complex morphology, and restricted lexical coverage. In contrast, the current study enhances an existing embedding model to more effectively support low-resource African languages by explicitly addressing these constraints, thereby producing more resilient and linguistically informed word representations.

3 Methodology

3.1 Standard GloVe Model

Our work extends the GloVe (Pennington et al., 2014) for low resource African languages. To extend the method, it is essential first to observe the problem formulation of GloVe.

Let V denote the vocabulary size and $X \in \mathbb{R}^{V \times V}$ the matrix of word-to-word co-occurrence matrix, where an entry X_{ij} counts the number of times word j appears in the context window of word i over the entire corpus. GloVe trains two embedding matrices, $W, \tilde{W} \in \mathbb{R}^{V \times d}$, and two bias vectors $b, \tilde{b} \in \mathbb{R}^{V}$, minimizing the weighted least-squares objective between the inner product of word and context vectors and the logarithm of their co-occurrence count using Equation 1.

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\vec{w}_i^\top \vec{\tilde{w}}_j + b_i + \tilde{b}_j - \log(X_{ij}) \right)^2 \tag{1}$$

where $\vec{w}_i$ and $\vec{\tilde{w}}_j$ are the ith and j^{th} rows of W and $\tilde{W}$, respectively, while b_i and b_j represent their respective bias parameters.. The weighting function $f(x)$ is employed to regulate the weight, preventing excessive escalation when word frequency becomes too high. After training, GloVe considers the final word representation as the average of the two embedding matrices, $v_w = (w_w + \tilde{w}_w)/2$ (Pennington et al., 2014).

$$f(x) = \begin{cases} \left(\frac{x}{x_{\max}}\right)^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

Common values are $x_{\max} = 100$ and $\alpha = 0.75$

Although GloVe produces strong representations for high-resource languages, its interaction term $w_i^\top \tilde{w}_j$ is strictly linear: the model cannot capture non-linear semantic relations, and its fixed weighting scheme may under-represent rare co-occurrences. Both issues are especially pronounced in morphologically rich, low-resource settings such as Sepedi (nso) and Setswana (tsn), where vocabulary coverage is limited and rare co-occurrences carry disproportionate linguistic signal. We address these shortcomings with the DeepGloVe architecture described below.

3.2 DeepGloVe

The challenges linked to low-resource African languages are not merely data-centric challenges; they are fundamentally algorithmic issues stemming from assumptions embedded in models tailored for high-resource, predominantly Indo-European languages. These existing embedding methods also rely on shallow co-occurrence statistics and cannot effectively model morphologically rich languages that reflect high-order semantic interactions.

This motivated us to recognize that research cannot merely involve scaling up existing high-resource models but must fundamentally reconsider how linguistic features are represented and processed at the embedding level. This concept is also corroborated by Arora's research on "internal structure" for African languages; the internal structure is more intricate and requires explicit, potentially non-linear, modeling within the embedding space (Arora et al., 2016).

3.2.1 Architecture

To address the issue, we utilize a deep learning framework to apply a non-linear transformation over the initial embedding matrix, enabling the model to capture high-order contextual dependencies found in morphologically rich languages, which shallow matrix factorization methods fail to capture. This representation encourages semantic similarity among agglutinatively related words and incorporates a non-linear representation on top of GloVe model. The proposed architecture is depicted in Figure 1.

As depicted in the figure, the co-occurrence matrix is factorized into word (target) embeddings and context embeddings through the use of two parallel dense deep learning towers. This necessitates non-linear modeling within the embedding space, and each embedding goes through the dense layer in order to alleviate the complex interaction that occurs within the internal structure. The DeepGloVe method retains the embedding tables and bias vectors of standard GloVe but replaces the linear dot

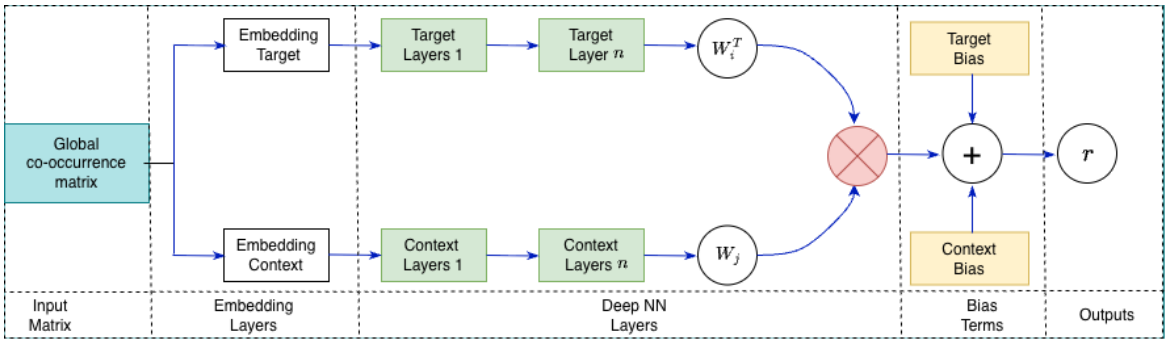


Figure 1: The proposed DeepGloVe architecture

product with a pair of parallel non-linear dense towers before computing the inner product. The problem formulation is given in as follows:

Let $\phi_u : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $\phi_v : \mathbb{R}^d \to \mathbb{R}^{d'}$ denote the target and context tower functions, respectively, each implemented as a feed-forward network with $\tanh$ activations:

$$\phi(x) = \tanh\Big(\tanh\Big(x\,W^{(1)} + b^{(1)}\Big)\,W^{(2)} + b^{(2)}\Big) \tag{3}$$

where $W^{(1)} \in \mathbb{R}^{d\times h}$, $W^{(2)} \in \mathbb{R}^{h\times d'}$, and the two towers have separate, independently learned parameter sets. In all experiments we use $d = 100$, $h = 256$, and $d' = 128$. The predicted log co-occurrence for the pair (i, j) is then

$$r_{ij} = \phi_u(w_i)^\top \phi_v(\tilde{w}_j) + b_i + \tilde{b}_j \tag{4}$$

Importantly, the bias terms b_i and $\tilde{b}_j$ are drawn directly from the embedding table biases rather than produced by the towers. This design choice preserves the word-level frequency offset that anchors the GloVe log-bilinear model while allowing the towers to learn a non-linear re-projection of the co-occurrence derived embedding space.

3.2.2 Adjusted Weighting Function

In standard GloVe, rare words are ignored, and they receive fewer gradient updates due to their low frequency, which also results in the data sparsity problem. This combined effect contributes to poor semantic discrimination. To address the challenges, we implement a weighting function that amplifies the influence of infrequent yet informative co-occurrences by modifying $f(x)$ as follows:

$$f_{\text{deep}}(x) = \begin{cases} g(x) + \lambda\left(1 - e^{-\beta x}\right) & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

where $\lambda = 0.8$ and $\beta = 0.1$ are fixed hyperparameters and $g(x) = 0.5\left(\frac{x}{x_{\max}}\right)^{\alpha}$.

In the first term of GloVe's weighting function, we downscale it by $\frac{1}{2}$ to ensure a balanced contribution to the mixture. The second term is a saturating exponential; for small values of x, a first order Taylor expansion approximate $\lambda(1 - e^{-\beta x}) \approx \lambda\beta x$, resulting in a linear increase at low counts. As $x \to x_{\max}$, both terms reach saturation, causing $f_{\text{deep}}(x) \to 1$. Compared to $f(x)$, the adjusted weighting function $f_{\text{deep}}(x)$ allocates greater weighting in the range $0 < x \ll x_{\max}$, while aligning at the upper limit, thereby preserving the original model's behavior for frequently occurring pairs. Unlike standard co-occurrence-based embeddings that ignore infrequent words due to large corpora in high-resource languages, this modification ensures the incorporation of notable infrequent co-occurrences by increasing the gradient contribution of rare tokens by preventing their collapse in the embedding spaces, while also reducing the influence of high-frequency tokens. This implementation effectively redistributes learning capacity for low-resource languages with limited digital data.

3.2.3 Post-Dense Embedding Extraction

One of the innovations in the DeepGloVe model is the extraction of the post dense word representation embeddings. After training, we pass every word through both parallel towers and average the outputs using Equation 6:

$$v_w^{\text{deep}} = \frac{\phi_u(w_w) + \phi_v(\tilde{w}_w)}{2} \in \mathbb{R}^{d'} \tag{6}$$

This mirrors the averaging operation performed in the GloVe method, but DeepGloVe operated it in a non-linear transformed space of dimension $d' = 128$, rather than the raw embedding space dimension we used $d = 100$.

Statistics	Sepedi	Setswana
Number of Sentences	451412	462049
Number of Words	8.3M	7.1M
Unique Words	133K	118K
Avg. WpS	18.44	15.37

Table 1: This table presents the statistics of the dataset after GlotLID, including average number of words per sentence (Avg.WpS).

4 Experiments

4.1 Experimental Datasets

Training data is essential for creating robust and accurate models, and high-quality, clean training data can elevate the model to a high quality. This study utilized popular data resources such as National Centre for Human Language Technology (NCHLT) (Eiselen and Puttkammer, 2014) and Vuk'zenzele (Lastrucci et al., 2023) to curate monolingual datasets for the two languages: nso and tsn. However, these datasets contain unwanted artifacts such as other languages, URLs, and other internet artifacts (e.g Emojis). To remedy this, we used the publicly available GlotLID tool (Kargaran et al., 2023) to filter out content that does not belong to the target language, following (Sindane and Marivate, 2024). From this, we build a clean corpora for nso and tsn with statistics as shown in Table 1.

4.2 Evaluation Benchmarks

In this study, we evaluated the DeepGloVe alongside the baselines, standard Glove, W2V-SG, and W2V-CBOW embeddings using similarity tasks on the WordSim353, and SimLex999 datasets for the selected languages (Makgatho et al., 2021). The WordSim353 dataset is the translated version of the WordSim353 benchmark translated into nso and tsn languages, where the translated pairs have been assigned similarity ratings by human annotators to measure semantic relatedness. In this case, a high score would mean high relatedness, while a low score means no semantic relationship. The SimLex999 on the other hand, restricts similarity to pure semantic relatedness, such as synonymous words. Both datasets have been extensively used to measure the quality of monolingual embeddings representations, which we also explore extensively in this study. The objective here is to assess the quality of an embedding space. A high-quality embedding is expected to encode semantic structure such that words with similar meanings are represented by proximate vectors, while dissimilar words are positioned farther apart in the vector space. Each evaluation was conducted employing Spearman's correlation coefficient to determine similarity measures. The Spearman correlation coefficient ρ is a widely employed measure for evaluating the effectiveness of word embeddings. It measures the monotonic relationship between similarity scores produced by the model and human evaluations, provides valuable insights into the closeness of vectors in high dimensions, and can be calculated between two embedding vectors as shown in Equation 7 below: We publicly release all the word embeddings to support research for low resourse African languages.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{7}$$

where d_i is the difference between the ranks of each pair of values and n is the number of pairs.

4.3 Experimental Setup

In the experiment, we generated embeddings of the dimension 100 as found to be a good trade-off between efficiency and computational cost for low-resourced languages (Makgatho et al., 2021). For GloVe we experimented with hyperparameters: window size of 10, minimum count of 5, 25 training epochs, and a learning rate of 0.05, while most were left as default according to the original paper's recommendations (Pennington et al., 2014). For Word2Vec we explored hyperparameters: we used both skipgram and CBOW as the model, a window size of 5, 25 training epochs, a minimum count of 5, a learning rate of 0.05, and defaults otherwise (Bojanowski et al., 2017).

The DeepGloVe model is trained by minimizing the objective function in Equation 8, which is structurally similar to the standard GloVe objective function in Equation 1, but with a richer prediction function of Equation 4 and the modified function in Equation 5. The optimization was performed using the Adam optimizer (Kingma and Ba, 2014), employing a learning rate of 0.001, using mini batches of 8192. We trained the model for 25 epochs. We utilized two hidden layers for the deep neural network, with in the hidden layer neurons specified as $\{256, 128\}$, and the embedding dimension is set be 100. To obtain these parameters, we conducted hyperparameter tuning for the DeepGloVe model manually, evaluating embedding dimensions of $50, 100, 200, 300$, with the optimal dimension

being 100. For learning rate, we tested values of $0.01, 0.001, 0.0001$, and 0.001 as the best for the Adam optimizer. While we implemented two hidden layers, increasing their number yielded minimal improvement rather than increasing the number of parameters. For the two layers, we found 256 and 128 neurons performed best. The proposed DeepGloVe and all baseline models were implemented in Python using TensorFlow, and trained using NVIDIA GeForce RTX 4060 GPU laptop. The Spearman result for the DeepGloVe was obtained from a five-fold cross validation, with the average reported as the representative result.

$$J_{\text{deep}} = \sum_{i,j=1}^{V} f_{\text{deep}}(X_{ij})(r_{ij} - \log X_{ij})^2 \quad (8)$$

5 Results and Discussion

5.1 Quantitative Results

The intrinsic experimental results in the WordSim353 and SimLex999 word similarity benchmarks shown in Table 2 demonstrate that the proposed DeepGloVe method consistently and significantly outperforms all baseline embedding models, in Sepedi(nso) and Setswana(tsn) languages. Performance gains are notable in semantic similarity tasks, indicating that the proposed DeepGloVe model learns richer semantic representations than the baselines. As depicted in Figure 2, DeepGlove beats all baseline methods. For instance, when we observe the relative improvement gain over the GloVe method using the WordSim353 dataset, DeepGloVe achieved a 24.94% improvement for nso and 63.93% improvement for SimLex999.

In WordSim353, the baseline models achieve a relatively weaker correlation in both languages. For Sepedi, all the baseline models obtain a smaller Spearman correlation, and all results are not statistically significant($p > 0.05$). In contrast, DeepGlove achieves a significantly higher correlation with values of $p < 0.05$. A similar pattern is observed for Setswana, where the proposed method outperforms all the predictive and count based baseline models. Interestingly, DeepGloVe is the only model that consistently resulted in strong and statistically reliable improvement across both low resource African languages.

As depicted in Table 2 and Figures 6 and 7 (See Appendix A), the confidence interval (CI) further supports the robustness of DeepGloVe's improvements. For W2V-SG, W2V-CBOW, and GloVe on WordSim353, the intervals overlap with zero for WordSim353 dataset, indicating weak semantic associations. However, for DeepGloVe, the interval is positive for both benchmark datasets, showing that the learned representation captures meaningful semantic structure beyond random correlation effects. This is crucial in low resource African languages, where limited co occurrence can result in unreliable embeddings. This performance improvement is also supported on the SimLex999 dataset, which is widely considered a more challenging benchmark as it assesses true semantic similarity. For both nso and tsn benchmarks, DeepGloVe achieves a higher correlation by significantly outperforming all baseline models. These performance improvements suggest that DeepGloVe is more adept at capturing nuanced semantic relationships, which are challenging in morphologically rich African languages.

Importantly, we observe a significant performance difference between DeepGloVe and standard GloVe across both datasets. Given that DeepGloVe extends the original GloVe framework with deeper non linear transformation, these findings imply that the traditional matrix factorization may not be sufficient for capturing the intricate semantic pattern found in low resource African language settings. The deeper framework likely allows the model to better explore high order co occurrence patterns.

Interestingly, DeepGloVe and all baseline models exhibit identical, relatively high coverage rates (see Table 2) across both benchmarks: $64-68\%$for WordSim353 and 81% for SimLex999, suggesting that the vocabulary learned by DeepGloVe is adequately representative even with limited data. This consistency in convergence ensures that the observed performance improvement attributes are due to the methodological enhancement done on the embedding learning method rather than different vocabulary sizes. These semantically meaningful embedding enhancements are especially significant within low-resource Sepedi and Setswana language environments, where data scarcity and intricate linguistic structures present substantial challenges.

As observed in Figures 6 and 7 (see Appendix A), the 5-fold cross validation results further reinforce the fact that DeepGloVe generates more stable semantic representations compared to the baseline models. In both benchmarks, it consis-

Dataset	Language	Model	ρ	CI Low	CI High	p-value	n	Coverage
WordSim-353	nso	W2V-SG	0.1091	-0.0407	0.2338	0.1018	226	0.6402
	nso	W2V-CBOW	-0.0200	-0.1643	0.1191	0.7650	226	0.6402
	nso	GloVe	0.1194	-0.0160	0.2493	0.0733	226	0.6402
	nso	**DeepGloVe**	**0.1492**	0.0100	0.2823	0.0248	226	0.6402
	tsn	W2V-SG	0.0767	-0.0520	0.2008	0.2368	240	0.6799
	tsn	W2V-CBOW	0.0696	-0.0603	0.2007	0.2831	240	0.6799
	tsn	GloVe	0.0603	-0.0790	0.1951	0.3522	240	0.6799
	tsn	**DeepGloVe**	**0.1971**	0.0620	0.3191	0.0022	240	0.6799
SimLex-999	nso	W2V-SG	0.2045	0.1324	0.2766	0.0000	817	0.8178
	nso	W2V-CBOW	0.1470	0.0750	0.2159	0.0000	817	0.8178
	nso	GloVe	0.1944	0.1193	0.2647	0.0000	817	0.8178
	nso	**DeepGloVe**	**0.3185**	0.2478	0.3838	0.0000	817	0.8178
	tsn	W2V-SG	0.1960	0.1282	0.2683	0.0000	827	0.8278
	tsn	W2V-CBOW	0.0736	-0.0035	0.1421	0.0343	827	0.8278
	tsn	GloVe	0.1527	0.0800	0.2259	0.0000	827	0.8278
	tsn	**DeepGloVe**	**0.2571**	0.1895	0.3208	0.0000	827	0.8278

Table 2: Spearman's ρ correlation results on WordSim-353 and SimLex-999 datasets for nso and tsn languages.

tently achieves the highest ρ while showing significantly lower variance across folds. This pattern is evident for both nso and tsn, referring that the improvements are not limited to a specific dateset, but applicable across different languages. The reduced variance shown in DeepGloVe is a significant finding for low resource languages. the strong consistency across folds indicates that the proposed architecture allows the extraction of more semantic patterns. This contradicts the baseline models, which seem to be more vulnerable to noise. From this, we can infer that the improvement made by DeepGloVe is not due to random variation, showing its superior semantic learning, and generalization compared to baseline models. The consistency across the two languages indicates that the proposed architecture is applicable across related low resource languages.

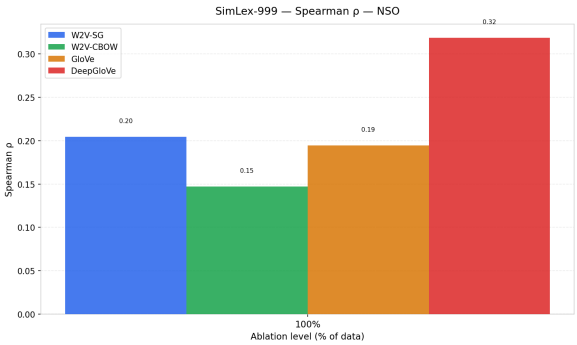
5.2 Qualitative Results

To further confirm the effectiveness of the proposed DeepGloVe model, we performed qualitative analysis, evaluating the embedded word embeddings via visualizations produced by t-SNE (Van der Maaten and Hinton, 2008). These are dimensionality reduction techniques that map high-dimensional word vectors into a two-dimensional space. Figures 3, 4, and 5 (see Appendix A) show the semantic organization of the tsn and nso embeddings learned by all the baselines and proposed models. The visualization from the DeepGloVe reveals relatively high semantically coherent clusters as compared to the baselines, where words belonging to similar categories are grouped closely together in the embedding space. This indicates that the proposed method captures meaningful semantic relationships between linguistically related terms.

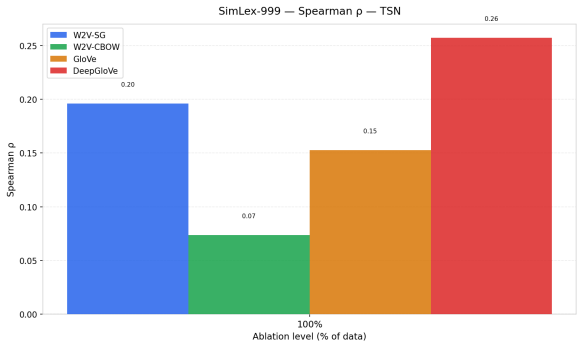
The t-SNE visualization also highlights that related words are clustered together and unrelated categories are distanced in the embedding space. This also demonstrates that the method captures structured semantic manifolds. In contrast, all the baseline embedding models produce less distinguishable structures, with more dispersed and overlapping clusters. This limited separation shows weaker semantic separation, aligning with quantitative evaluation where DeepGloVe consistently outperformed all baseline models in correlation score and variance.

The achievement observed in DeepGloVe results from the deeper nonlinear transformation advancement made on the GloVe framework, allowing the model to capture high order semantic dependencies that shallow methods struggle to represent. This ability is essential for low resource African languages, where a sparse training corpus and rich morphology complicate representation learning.

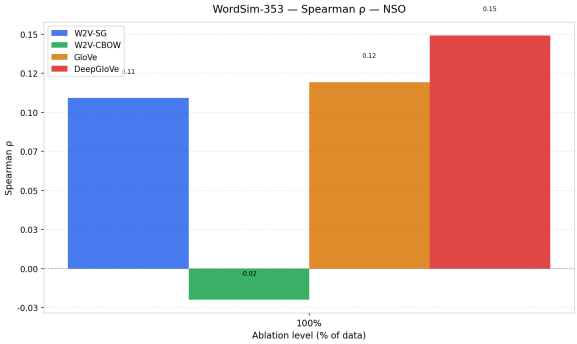
Overall, visualization analysis offers qualitative evidence that the proposed method learns a more semantically coherent embedding space compared to baseline models for African languages. These qualitative findings supplement the quantitative as-



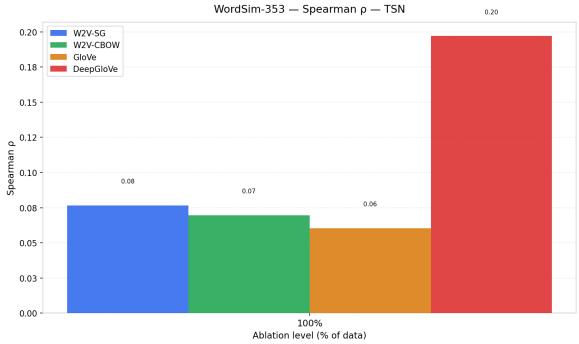
(a) SimLex-999 Spearman ρ results for NSO.



(b) SimLex-999 Spearman ρ results for TSN.



(c) WordSim-353 Spearman ρ results for NSO.



(d) WordSim-353 Spearman ρ results for TSN.

Figure 2: Spearman ρ correlation comparison across embedding models for SimLex-999 and WordSim-353 datasets.

sessments and further validate the efficacy of the proposed approach in capturing semantic relationships, especially in scenarios involving limited data and intricate linguistic frameworks for the designed languages.

In summary, the superior performance of the proposed model is principally attributable to the modification of the co-occurrence weighting function for low-resource languages such as nso and tsn. In contrast to the standard GloVe formulation, which is developed for high-resource languages and employs a static heuristic weighting approach, the proposed weighting function is engineered to more effectively manage sparse co-occurrence statistics. In corpora with limited resources, the infrequent occurrence of numerous word pairings can cause baseline models to either excessively penalize or completely ignore these informative yet uncommon co-occurrences. Therefore, the proposed DeepGloVe uses a modified weighting formula to maintain significant semantic information from low-frequency word co-occurrences that might otherwise be lost. This improves the quality of the generated word representations. Adding deep factorization techniques to the embedding learning process makes the proposed DeepGloVe model better at finding semantic relationships and also lets the model find more complex and non-linear relationships between words. This is especially important for African languages that have several different forms and structures.

6 Conclusion

This paper introduces an extension of the GloVe model designed for low-resource, morphologically intricate languages, with a particular emphasis on agglutinative languages, exemplified by the nso and tsn languages. Employing a data-sparsity-aware weighting function that enhances the significance of infrequent yet informative co-occurrences and the integration of a deep neural network architecture to improve the semantic representation of morphologically rich words to derive more robust and semantically meaningful embeddings. We evaluated the method using the WordSim353, and SimLex999 benchmark datasets, employing Spearman's correlation coefficient. The results show that the proposed DeepGloVe method demonstrated state-of-the-art representation quality compared to the baseline models, particularly for low-resource African languages such as nso and tsn.

Limitations

One limitation of the DeepGloVe model was that it was trained on a relatively small number of corpora, which may not fully capture the domain variation of the language. It remains a static embedding rather than a fully contextual one, as it is based on GloVe. Additionally, the model's evaluation protocol is restricted to intrinsic tasks, which assess semantic structure but do not necessarily evaluate performance in downstream tasks. As a result, the practical effectiveness of the proposed model remains uncertain. Finally, the scarcity of standardized datasets, like analogy datasets for nso and tsn, constrains the scope of the evaluation. Future work will involve expanding the methodology to additional languages and extending the method to contextual embedding.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Roald Eiselen and Martin J Puttkammer. 2014. Developing text resources for ten south african languages. In *LREC*, pages 3698–3703.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Richard Lastrucci, Jenalea Rajab, Matimba Shingange, Daniel Njini, and Vukosi Marivate. 2023. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 18–25.

Mack Makgatho, Vukosi Marivate, Tshephisho Sefara, and Valencia Wagner. 2021. Training cross-lingual embeddings for setswana and sepedi. *arXiv preprint arXiv:2111.06230*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A survey of text representation and embedding techniques in nlp. *IEEe Access*, 11:36120–36146.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Erhan Sezerer and Selma Tekir. 2021. A survey on neural word embeddings. *arXiv preprint arXiv:2110.01804*.

Thapelo Andrew Sindane and Vukosi Marivate. 2024. From n-grams to pre-trained multilingual models for language identification. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 229–239.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

A Appendix

Model Size and Cost The number of parameters for the Deep-GloVe (TSN) language model includes 8,365,022 parameters for embedding and 117,504 parameters for the dense tower. This represents an increase of 117,504 parameters compared to the baseline due to the introduction of two towers, which is not a significant increase. In terms of computational cost, the model took approximately 51 seconds for 25 iterations, while preparing the co-occurrence matrix required around 61 seconds, resulting in a total time of approximately 1 hour. Regarding AI, we used AI-based grammar checking tools to improve writing clarity.

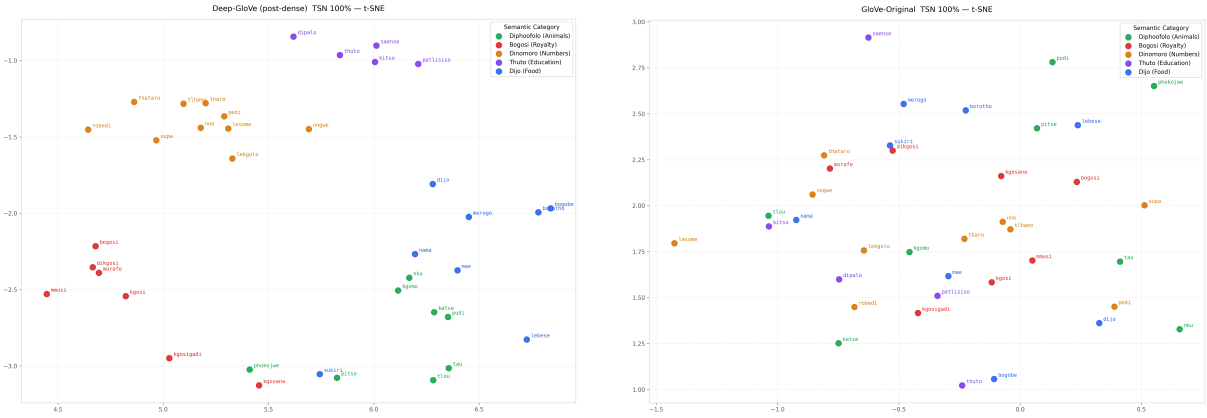


Figure 3: DeepGloVe and GloVe t-SNE visualization for tsn.

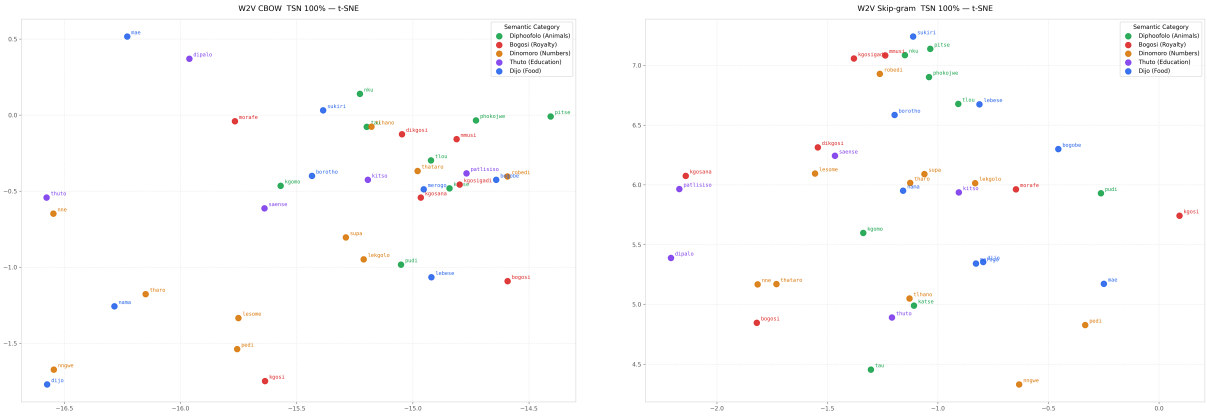


Figure 4: W2V-SG and W2V-CBOW t-SNE visualization for tsn.

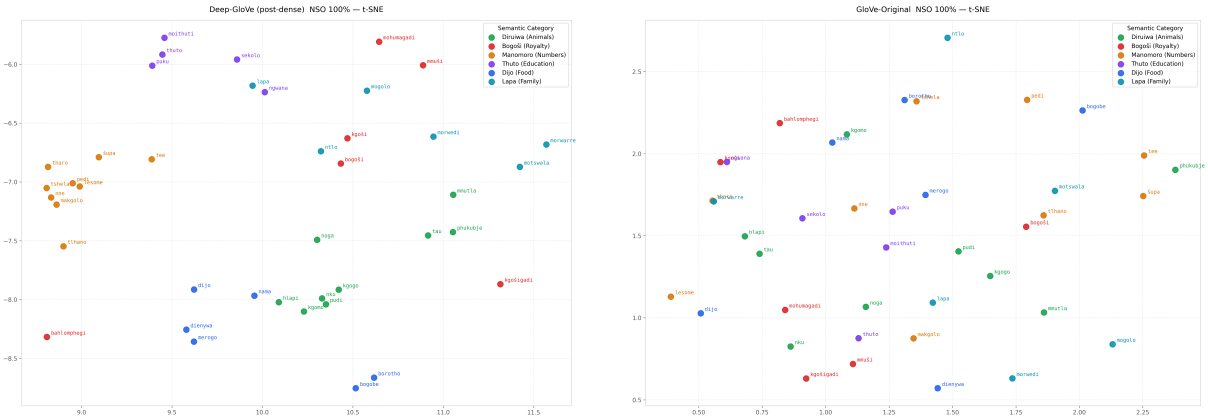


Figure 5: DeepGloVe and GloVe t-SNE visualization for nso.

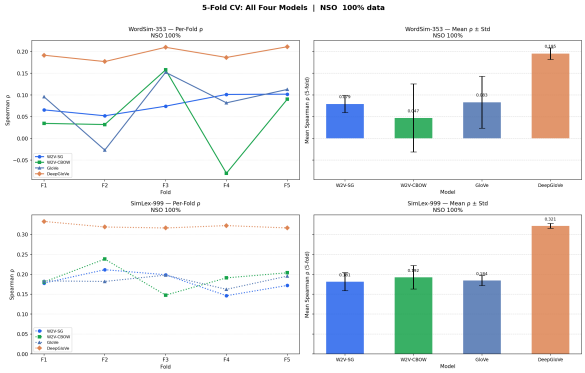


Figure 6: Spearman correlation with a 95% confidence interval and error bar across embedding models for the language nso, using both WordSim-353 and SimLex-999 datasets.

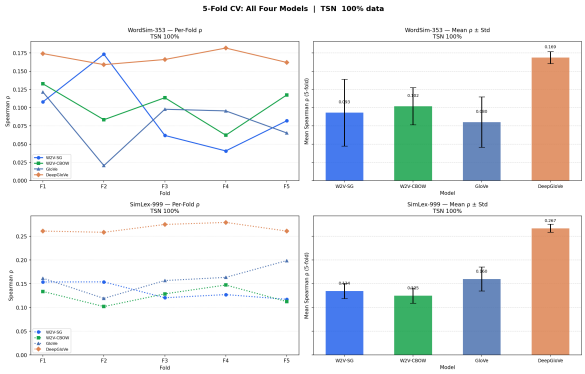


Figure 7: Spearman correlation with a 95% confidence interval and error bar across embedding models for the language tsn, using both WordSim-353 and SimLex-999 datasets.