MAGE: Multi-Head Attention Guided Embeddings for Low Resource Sentiment Classification

V. Vashisht[1][0009−0002−3466−4676], S. Singh[1][0009−0004−1652−595X], M. Konduskar[1][0009−0009−3203−6976], J.S. Walia[1][0000−0002−9255−5446], and V. Marivate[2][0000−0002−6731−6276]

[1] School of Computer Science and Engineering, Vellore Institute of Technology
[2] University of Pretoria

Abstract. Due to the lack of quality data for low-resource Bantu languages, significant challenges are presented in text classification and other practical implementations. In this paper, we introduce an high-performing model combining Language-Independent Data Augmentation (LiDA) with multi-head attention based weighted representations to selectively enhance critical features within the embedding space and improve text classification performance. This integration allows us to create stable data augmentation strategies that are accurate across various linguistic contexts, ensuring that our model can handle the unique syntactic and semantic features of Bantu languages. This approach not only addresses the data scarcity issue but also sets a foundation for future research in low-resource language processing and classification tasks.

1 Introduction

Text classification is one of the most widely explored tasks in Natural Language Processing (NLP) due to its diverse applications, including spam detection, sentiment analysis, and topic modeling. Despite the impressive advancement achieved through deep learning, these methods rely heavily on large amounts of labeled data, posing a challenge for low-resource languages [13, 14]. African languages in general exemplify this challenge, as the scarcity of annotated datasets limits the development of accurate text classification models [1, 2]. Data augmentation has emerged as a promising solution for addressing data scarcity by generating synthetic data from original datasets [7]. Traditional augmentation techniques, including synonym replacement, sentence back-translation, and generative models, rely heavily on language-specific resources such as pre-trained word embeddings, language models, or linguistic databases, such as WordNet [4, 11, 15, 24]. This language dependence makes these approaches less accurate for underrepresented languages, such as Bantu languages, which lack these linguistic resources [27]. To overcome these limitations, Language-Independent Data Augmentation (LiDA) was introduced [22], operates on sentence embeddings (embedding space) rather

than surface text at the word/token level. LiDA transforms sentence embeddings to generate synthetic data, bypassing the need for language-specific resources. Building upon this foundation, we propose MAGE (Multi-Head Attention Guided Embeddings), a framework designed to enhance text classification performance for low-resource languages. MAGE extends the LiDA framework by introducing significant innovations to the embedding and augmentation process. Specifically, it replaces the traditional Denoising Autoencoder with the Variational Autoencoder (VAE) to enable more expressive and diverse synthetic embeddings. Additionally, MAGE incorporates a novel multi-head attention mechanism that selectively emphasizes salient features in the embeddings. This focus on multi-head attention improves the model's capacity to capture critical syntactic and semantic nuances, making it particularly accurate for low-resource languages. Using the AfriSenti SemEval dataset [12], a collection of tweets annotated with positive, negative, and neutral sentiments for Kinyarwanda, Swahili, and Xitsonga, we evaluate the performance of MAGE in sentiment classification. Our results demonstrate that MAGE outperforms baseline approaches in low-resource settings. Moreover, comparative analyses highlight the advantages of MAGE over self-attention-based models, further establishing its value as a stable framework for addressing the challenges posed by data scarcity in low-resource languages.

Bantu languages pose particular challenges for natural language processing due to their linguistic complexity. They exhibit rich morphology, large noun class systems, and agglutinative structures, where multiple morphemes are combined within a single word. In addition, frequent code-switching and orthographic variation further complicate text processing. These properties make token-level augmentation methods less accurate, since word boundaries and surface forms often fail to capture the underlying linguistic structure. This motivates our shift toward embedding-level augmentation, which can better generalize across morphological and orthographic variation.

This work not only addresses the pressing issue of data scarcity in Bantu languages but also provides a scalable and adaptable framework for extending text classification capabilities to other low-resource language families. Through the introduction of MAGE, we set the stage for future research in low-resource language processing and establish a pathway to improve the inclusivity and generalizability of NLP technologies.

2 Related Work

2.1 Data Augmentation Techniques

In recent years, data augmentation techniques have gained significant attention, especially for low-resource languages, due to the scarcity of properly annotated datasets and general lack of resources. Prior studies on data augmentation span several dimensions, and we group them here into three categories: lexical-level, contextual-level, and embedding-level augmentations.

Lexical-level augmentations One of the earliest and most widely used techniques is back-translation. Sennrich et al. [21] leverage monolingual target language data for textual-based augmentation using back-translation to enhance model performance, though at the cost of requiring an additional pretrained Neural Machine Translation (NMT) model. Lample et al. [8] propose a related method that relies solely on monolingual corpora by mapping sentences from two languages into a shared latent space via a shared encoder–decoder architecture. While these approaches reduce dependency on parallel corpora, they remain limited by the availability of monolingual data. Another influential lexical-level approach is Easy Data Augmentation (EDA) by Wei and Zou [24], which applies synonym replacement, random insertion, random swap, and random deletion. Despite its simplicity and reliance only on a synonym dictionary such as WordNet [11], EDA significantly improves model performance even when training on small datasets.

Contextual-level augmentations To overcome the limitations of predefined dictionaries, Kobayashi [7] proposed contextual augmentation, which leverages bidirectional language models to generate substitute words based on surrounding context. This produces more semantically appropriate alternatives and outperforms lexical-level methods like EDA, especially in low-resource settings. Generative adversarial models have also been applied in this space: Yu et al. [25] introduced SeqGAN, which combines reinforcement learning with GANs to generate discrete sequences. SeqGAN has been shown to improve fluency and diversity in sequence generation tasks such as NLP and music generation. Related adversarial approaches include Jia and Liang [5], who designed adversarial examples for reading comprehension, demonstrating performance drops even in state-of-the-art models on SQuAD [19]. Beyond augmentation, Raffel et al. [17] explored transfer learning for low-resource languages via sentence-level alignment and multilingual embeddings, and Li et al. [9] investigated synthetic data generation using large language models in zero- and few-shot settings. These contextual-level techniques extend augmentation beyond surface-level manipulation and exploit broader semantic and generative modeling.

Embedding-level augmentations More recent work explores perturbations directly in the representation space. Chen et al. [3] proposed TMix, which interpolates hidden representations of text samples, combined with entropy minimization and consistency regularization, to improve generalization in resource-limited settings. Such mixup-style strategies avoid the brittleness of token-level methods and are better suited to morphologically complex languages. Building on this line of work, we investigate embedding-level augmentation tailored to low-resource Bantu languages, where surface token manipulations often fail due to agglutination and orthographic variation.

2.2 Data Augmentation for Low Resource Corpora Text Classification

For downstream NLP tasks in low-resource settings, various augmentation methods have been proposed. Rahamim et al. [18] introduced TAU-DR, which employs soft prompts while keeping the language model frozen, reconstructing hidden representations into synthetic sentences. This improves multi-class classification without requiring additional model training. Thangaraj et al. [23] investigated cross-lingual transfer in African languages, benchmarking forgetting metrics, though without applying augmentation. Karimi et al. [6] proposed AEDA, which introduces punctuation marks into sentences as a lightweight augmentation, preserving semantic consistency and outperforming EDA in low-resource settings. Litake et al. [10] developed IndiText Boost, a framework designed for underrepresented Indian languages, which combines EDA and back-translation to outperform more complex LLM-based methods on classification tasks. Prompt-based augmentation has also been explored: Sahu et al. [20] used pretrained LLMs like GPT-2 [16] to generate synthetic intent-classification data, though reliability issues with LLMs sometimes degrade data quality. Zhao et al. [26] introduced EPiDA, an augmentation framework combining conditional entropy minimization with relative entropy maximization, balancing diversity and quality. EPiDA consistently outperforms earlier techniques across text classification tasks, highlighting its applicability for low-resource settings.

2.3 LiDA - Language Independent Data Augmentation

Sujana and Kao [22] proposed LiDA, a language-independent augmentation method for text classification. Instead of generating new sentences, LiDA perturbs sentence-level embeddings trained with multilingual SBERT, resulting in consistent gains of 2-3% in LSTM-based classification. Its language independence stems from the multilingual dataset used to train the underlying embeddings, making it broadly applicable to low-resource contexts.

3 Methodology

Given the morphological richness and agglutinative nature of Bantu languages, token-level augmentation risks introducing noise rather than diversity. For instance, splitting or replacing tokens without accounting for noun class agreement can distort meaning. Similarly, code-switching and orthographic inconsistencies challenge augmentation methods that assume stable token inventories. To address these issues, we adopt an embedding-level approach that operates on distributed representations rather than surface tokens, allowing us to capture linguistic variation more robustly.

Hence, further refining the LiDA architecture, we propose a multi-head attention-based mechanism to quantitatively highlight and weight the individual embeddings to emphasize the important contributions of the LiDA architecture for the text-classification goal.

3.1 Dataset

The dataset referred to is the AfriSenti SemEval Shared Task - 12 dataset by Muhammad et al. [12] based on tweet sentiment analysis. As the study focuses on the Bantu language family, the datasets of the following 3 Bantu languages were chosen - Kinyarwanda, Xitsonga and Swahili having 7940 tweet-label pairs in the combined training set and 1482 tweet-label pairs in the combined test set. We observe a skewness in the data towards Kinyarwanda due to Kinyarwanda having the highest data points at 5155 tweets, with Swahili being the second highest at 3009 tweets and Xitsonga having the least data at 1258 tweets.

Field	Description
ID	Alpha-Numeric Serial Numbers
Tweet	Tweet Content
Label	Tweet Label

Table 1. Field descriptions of the dataset.

The dataset consists of 3 main fields, namely ID, Tweet, and Label. On preprocessing the data, the ID field was dropped, and the tweets accordingly preprocessed: lowercasing and removing punctuations, hyperlinks, and emojis. The labels in the label field as mentioned in table 1 were given as Negative, Neutral and Positive to indicate the tweet sentiment which was label-encoded as 0,1,2 respectively for further processing.

3.2 Architecture

The previous architectures and frameworks discussed have focused primarily on widely studied languages such as English, Indonesian, Chinese, French, and others. Although these languages benefit from extensive resources and established linguistic frameworks, our work diverges by addressing African Bantu languages, which are linguistically distinct and underrepresented in computational research. Bantu languages exhibit unique structural and morphological characteristics, requiring specialized approaches that go beyond the methodologies applied to more commonly studied languages. Hence, the embedding models and the architectural complexities in the components of previous frameworks do not conform to the requirements of the Bantu languages.

Taking LiDA as our base framework, we propose our modified architecture in 3.2 creating a stable architecture that caters to the demands of Bantu languages.

LiDA Architecture The LiDA architecture (*figure 1*) makes use of the multilingual SBERT (Sentence-BERT) model making the architecture *language independent* in essence. The embeddings so generated are passed through three functions - linear transformation, autoencoder model, denoising autoencoder model - before being concatenated with the original embeddings and henceforth classified using LSTM and BERT classifiers.

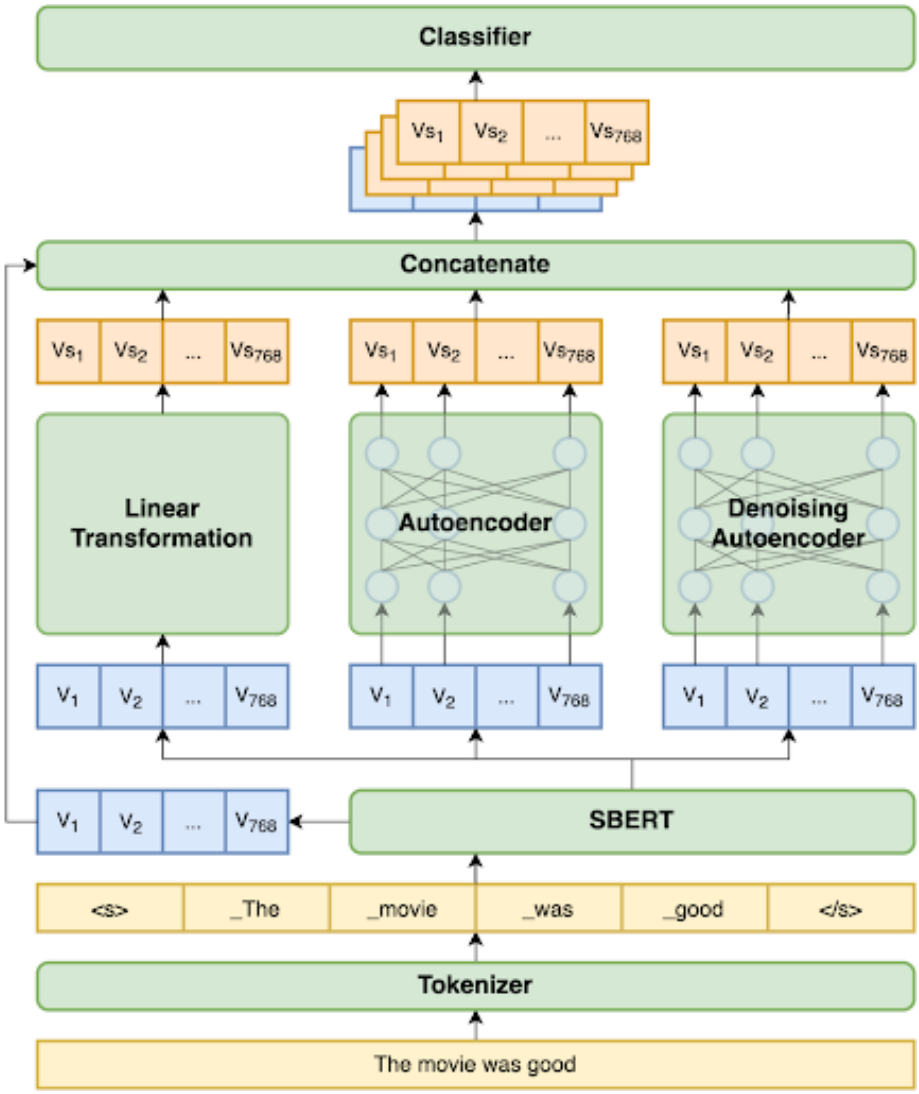

Fig. 1. LiDA Framework Reproduced from [22] [22].

Proposed Architecture Figure 2 shows our modification to the original LiDA architecture by changing the embedding model to AfriBERTa, replacing the denoising autoencoder with a variational autoencoder and addressing the concatenation of complex low-resource languages such as the Bantu language family by the introduction of weighted concatenation using multi-head attention. The tweets are passed through the AfriBERTa model, the choice of which is discussed in the sub-section 3.3. The model outputs a 768-dimensional representation of the text which is passed through the aforementioned transformation functions.

The Linear Transformation Layer introduces controlled variability into the input embeddings by applying a randomized shift, enhancing the robustness and generalizability of the representations. For each embedding, a random noise vector r is sampled uniformly within a range $[r_{\min}, r_{\max}]$ and added to the original embedding, resulting in a transformed embedding $e' = e+r$. This operation is performed independently for all embeddings in the training and testing datasets. The parameters $r_{\min}$ and $r_{\max}$ can be adjusted to control the magnitude of perturbation, ensuring that the embeddings retain their original semantic structure while introducing sufficient variability to aid learning.

Autoencoder is a key part of the augmentation process, designed to refine and diversify input embeddings by learning compressed representations while retaining essential features. This is accomplished through an encoder-decoder architecture that reduces the input embedding dimensions to a latent space and reconstructs them back to the original size. This introduces subtle variations while preserving essential semantic features, enhancing the diversity of augmented data. The model has been slightly enhanced from the original model used in Sujana and Kao ([22]) with Leaky ReLU activations, Batch Normalization - to ensure stable training and mitigate vanishing gradients - and Dropout layers to help with regularization, improve generalization and reduce overfitting for the Bantu language family. The encoder consists of sequential linear layers that progressively reduce the embedding size from the original *768* dimensions to *32* dimensions in the latent space with a learning rate of 0.001 for stabilized learning. The decoder mirrors the encoder structure, gradually increasing the dimensionality from the latent space back to the original embedding size of *768*, with the final layer applying a sigmoid activation for bounded output. This structure ensures that the embeddings are refined through compression and reconstruction, creating more diverse and stable representations.

The Variational Autoencoder (VAE) proves to be integral to our data augmentation process, providing a more flexible and expressive approach to embedding refinement compared to the original denoising autoencoder. Unlike traditional autoencoders, VAEs model the input data as probabilistic distributions rather than deterministic mappings, allowing for more varied and stable synthetic embeddings. This probabilistic framework facilitates the generation of diverse augmented data points, which is crucial in low-resource language tasks like those involving Bantu languages. The VAE learns a distribution over the latent space, enabling the generation of new samples by sampling from the learned distribution, enhancing the diversity of the training data and helping the model

generalize better. The VAE architecture consists of an encoder and a decoder. The encoder first maps the input embeddings from their original input dimension of *768* to a latent space of dimension *256*. The size of the intermediate layers,i.e., hidden dimension is set to 512, balancing capacity and complexity. Batch Normalization, ReLU activations, and Dropout (set to 0.2) are applied to improve stability, avoid overfitting, and ensure stable learning. The encoder then produces two outputs: the mean μ and log-variance $\log(\sigma^2)$ of the latent distribution. The reparameterization trick is applied to sample from this distribution, allowing gradients to propagate through the sampling process and enabling accurate training. Just as in the autoencoder, the decoder mirrors the encoder, first mapping the latent representation of latent dimension back to the hidden space, and then expanding to the original input dimension. The final output of the decoder is a probabilistic reconstruction of the original input. Also we are not proposing that VAE will always outperform the original DAE (Denoising Autoencoder) configuration mentioned in the LiDA architecture, but it may prove to be comparable or even superior in some scenarios, and both configurations will provide us with the desired end results when used along with attention mechanisms. Finally, the embeddings we get through these transformations are weighted and concatenated with the original embeddings using a multi-head attention mechanism section 3.4.

3.3 AfriBERTa as the Embedding Model

To evaluate the effectiveness of multilingual transformer embeddings for low-resource sentiment classification, we fine-tuned several candidate models—mBERT, XLM-R, AfriBERTa, and BantuBERTa—on three languages (Kinyarwanda, Swahili, and Tsonga) as well as on a combined dataset. Each model was trained and evaluated under identical experimental settings, and we report the classification metrics averaged over runs for accuracy, precision, recall, and F1-score.

From Table 2, AfriBERTa demonstrates superior overall performance across nearly all metrics and datasets. For Kinyarwanda, AfriBERTa surpasses the next-best model (BantuBERTa) by over 1% in accuracy, while for Swahili and Tsonga, it maintains the highest precision and balanced recall–F1 trade-off. On the combined multilingual dataset, AfriBERTa yields the best macro-average scores (0.5695 accuracy, 0.5803 F1), showing its adaptability to cross-lingual data and justifying its selection as the embedding model for all subsequent MAGE experiments.

3.4 MAGE

The multi-head attention component is designed to enhance the model's ability to focus on important embeddings by assigning distinct weights to different embeddings using multiple attention heads. This method allows for the dynamic selection of which embeddings have more influence on the final classification decision, effectively highlighting the critical features. The num_heads was set to *4* ensuring that multiple perspectives of the embeddings are captured simultaneously. The use of attention mechanism stemmed from the observation that

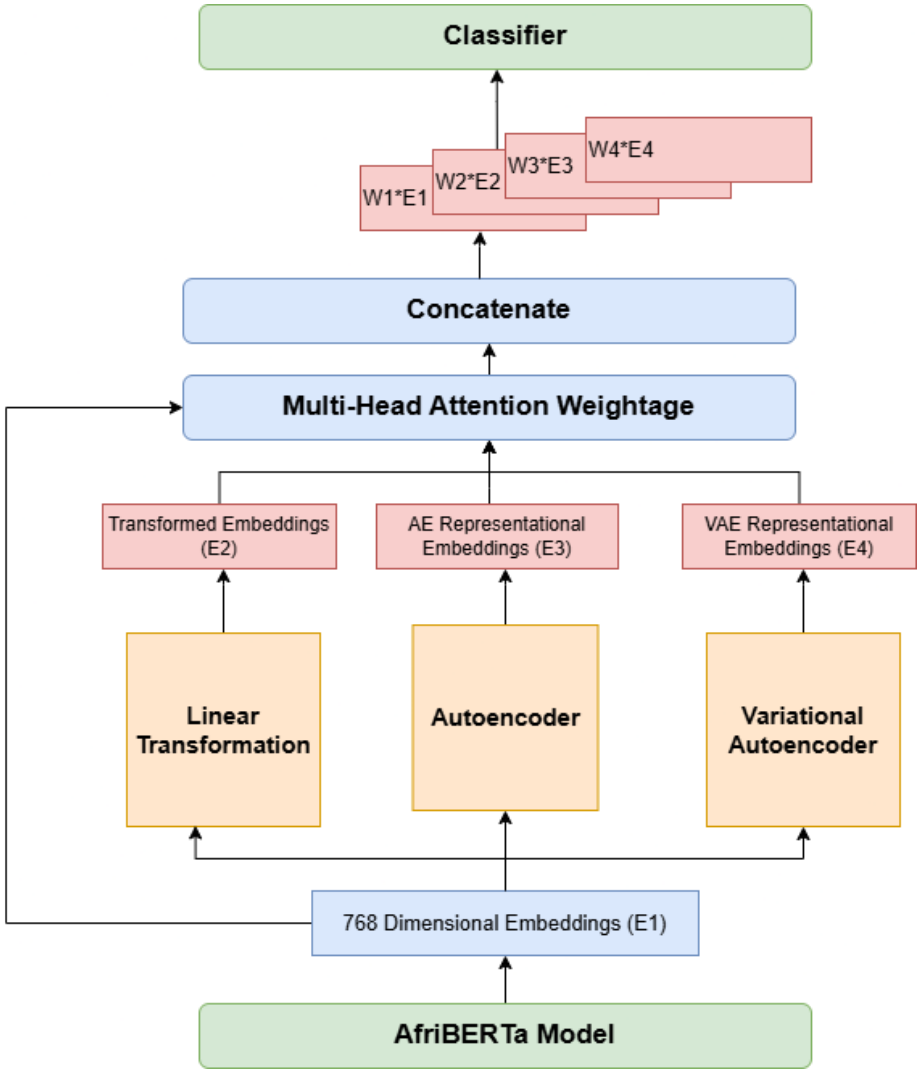

Fig. 2. Modified LiDA - MAGE Framework

Table 2. Language-wise classification performance of candidate embedding models. Metrics are averaged across experimental runs. AfriBERTa achieves consistently strong results across languages, outperforming others on the combined dataset.

Language	Model	Accuracy	Precision	Recall	F1
4*Kinyarwanda	mBERT	0.5296	0.5344	0.5196	0.5229
	XLM-R	0.3808	0.1269	0.3333	0.1838
	AfriBERTa	**0.6408**	**0.6430**	**0.6391**	**0.6407**
	BantuBERTa	0.6306	0.6346	0.6285	0.6309
4*Swahili	mBERT	0.5739	0.5193	0.4070	0.4052
	XLM-R	0.5960	0.4205	0.3405	0.2666
	AfriBERTa	**0.6004**	**0.5427**	**0.4806**	**0.4966**
	BantuBERTa	0.5960	0.5122	0.4869	0.4967
4*Tsonga	mBERT	0.5320	0.3569	0.4085	0.3692
	XLM-R	0.4729	0.1576	0.3333	0.2140
	AfriBERTa	**0.5418**	**0.4892**	**0.4123**	**0.3772**
	BantuBERTa	0.5172	0.4100	0.4099	0.3844
4*Combined Dataset	mBERT	0.4832	0.4640	0.4552	0.4250
	XLM-R	0.5207	0.5282	0.5497	0.5228
	AfriBERTa	**0.5695**	**0.5785**	**0.5600**	**0.5803**
	BantuBERTa	0.5266	0.5209	0.5113	0.5130

manually weighting embeddings improved classification results were observed, and attention provides a learnable way to optimize this process. In this approach, the embeddings are attended to using multi-head attention, where each head independently processes the embeddings and captures different aspects of the feature space. The context vectors, which are trainable, guide the attention mechanism to focus on the most relevant embeddings. The outputs from each attention head are then concatenated, and an aggregation is performed by summing the resulting vectors, capturing the most important features across different heads.

3.5 Classification

We employed two architectures for classification, namely, LSTM and logistic regression and the results were evaluated using accuracy, precision, recall and F1 score. The LSTM classifier was used with an input dimension of *768*, hidden dimension of *128*, and a single layer. The model was trained with a learning rate of *0.001* using CrossEntropyLoss. An early stopping mechanism with a patience of *3* epochs was employed, along with the StepLR scheduler to adjust the learning rate.

The second classifier used was logistic regression, which serves as a lightweight yet accurate baseline for classification. The model was trained with a maximum of 1000 iterations using the LBFGS(Limited-memory Broyden-Fletcher-Goldfarb-Shanno) solver for optimization. Since logistic regression is a simple

linear model, it provides a useful comparison against the LSTM's sequential feature extraction capabilities. By analyzing both models, we aim to assess the impact of complex sequential modeling versus traditional linear classification on our dataset.

4 Results

In this section, we present the classification performance of various embedding configurations and attention mechanisms, evaluated using standard metrics such as accuracy, precision, recall, and F1 score. We further extended our experiments through multiple independent runs with shuffled datasets to assess the robustness and consistency of our proposed approach.

4.1 Effect of DAE and VAE

We first compare the performance of the original embeddings using the proposed VAE vs DAE configurations using LSTM and Logistic Regression without weighted attention concatenation mechanism.

Metric	Original	With DAE	With VAE
Accuracy	0.5680	0.5739	0.5769
Precision	0.5589	0.5659	0.5681
Recall	0.5554	0.5697	0.5562
F1 Score	0.5566	0.5672	0.5589

Table 3. Comparison between DAE and VAE configurations (LSTM).

Observing table 3, we see a consistent improvement in all metrics over the original results for both DAE and VAE configurations using the LSTM classifier.

Metric	Original	With DAE	With VAE
Accuracy	0.5710	0.5729	0.5735
Precision	0.5636	0.5672	0.5659
Recall	0.5596	0.5624	0.5621
F1 Score	0.5592	0.5619	0.5618

Table 4. Comparison between DAE and VAE configurations (Logistic Regression).

Similarly, for Logistic Regression in table 4, we observe for the VAE configuration a 0.34%, 0.37%, 0.50%, and 0.48% improvement in accuracy, precision, recall and F1 score respectively over the original classification results. The

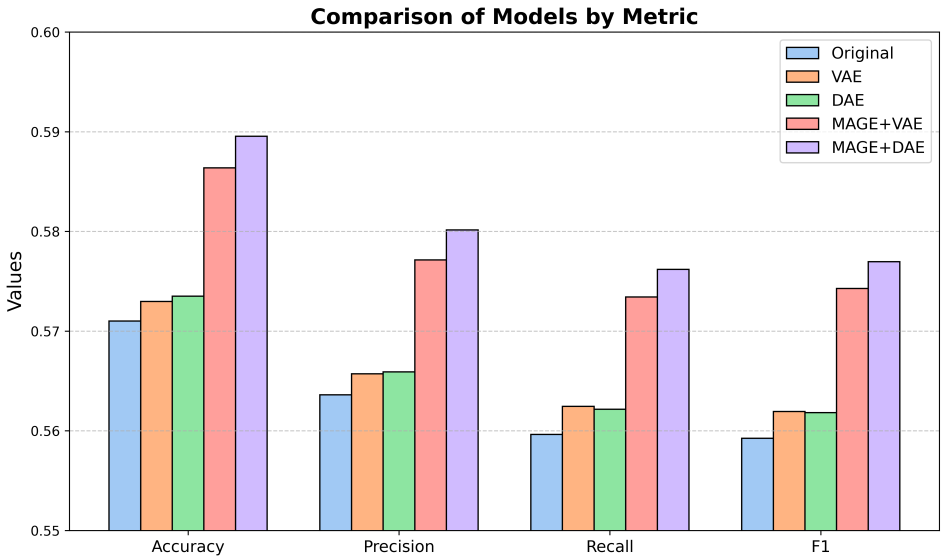

Fig. 3. Comparative Results

DAE configuration shows comparable improvements of 0.43%, 0.40%, 0.45%, and 0.46% for the same metrics. Both LSTM and Logistic Regression provide similar relative trends, suggesting that either autoencoder configuration can be effectively leveraged for feature enhancement.

4.2 Integrating Multi-Head Attention

To further test the robustness of our proposed attention-guided framework, we extended the experiments by introducing a Multi-Head attention mechanism. Each configuration was benchmarked by shuffling the dataset four times and running five independent training iterations per shuffle for both the LSTM and Logistic Regression classifiers. This approach allows us to evaluate performance stability and generalization beyond a single train-test split.

Figure 4 visualizes the average accuracy achieved across runs when comparing models with and without Multi-Head attention using both DAE and VAE configurations. The substantial performance gains are evident: **MAGE+DAE** achieves the highest accuracy improvement of 3.64% over the baseline, closely followed by **MAGE+VAE** at 2.51%. The attention mechanism further enhances both configurations, yielding an additional 0.4% and 1.21% increase over their respective non-attention counterparts.

A similar multi-run evaluation was conducted for the Logistic Regression classifier, where the averaged-out results (visualized in figure 3) exhibit consistent improvements across all metrics. To quantify these trends, we computed the overall mean and standard deviation (mean $\pm$ std) across runs for both LSTM

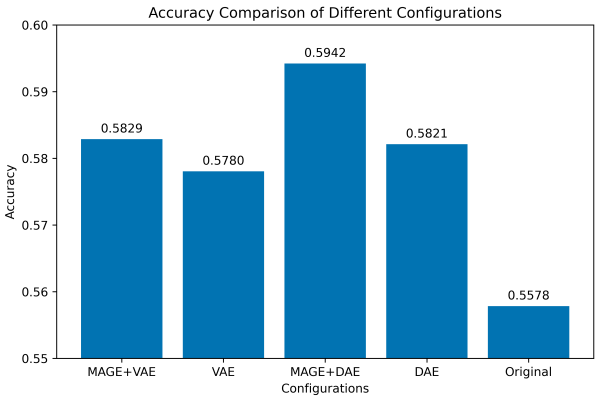

Fig. 4. Comparative Results

Table 5. Aggregated performance (mean±std) for Logistic Regression models averaged across 5 dataset shuffles and 5 independent runs.

Model	Accuracy	Precision	Recall	F1 Score
MAGE+VAE	0.5829 ± 0.0078	0.5773 ± 0.0065	0.5745 ± 0.0069	0.5752 ± 0.0070
VAE	0.5780 ± 0.0060	0.5628 ± 0.0054	0.5624 ± 0.0057	0.5618 ± 0.0058
MAGE+DAE	**0.5942 ± 0.0102**	**0.5839 ± 0.0081**	**0.5803 ± 0.0073**	**0.5805 ± 0.0076**
DAE	0.5821 ± 0.0040	0.5631 ± 0.0046	0.5612 ± 0.0041	0.5613 ± 0.0042
Original	0.5578 ± 0.0132	0.5629 ± 0.0109	0.5602 ± 0.0107	0.5604 ± 0.0109

and Logistic Regression models, aggregated across accuracy, precision, recall, and F1.

Table 6. Aggregated performance (mean ± std) for LSTM models averaged across 4 dataset shuffles and 5 independent runs.

Model	Accuracy	Precision	Recall	F1 Score
MAGE+VAE	0.5864 ± 0.0069	0.5802 ± 0.0062	0.5791 ± 0.0065	0.5805 ± 0.0066
VAE	0.5795 ± 0.0055	0.5734 ± 0.0052	0.5706 ± 0.0056	0.5720 ± 0.0053
MAGE+DAE	**0.5973 ± 0.0088**	**0.5861 ± 0.0074**	**0.5835 ± 0.0071**	**0.5852 ± 0.0072**
DAE	0.5842 ± 0.0048	0.5778 ± 0.0046	0.5751 ± 0.0047	0.5762 ± 0.0048
Original	0.5591 ± 0.0125	0.5613 ± 0.0111	0.5585 ± 0.0108	0.5588 ± 0.0110

From these results, we observe that MAGE consistently improves performance across all four evaluation metrics for both classifiers. The performance gains are particularly pronounced in F1 score and precision, highlighting MAGE's ability to better capture sentiment nuances. **MAGE+DAE** once again achieves

the best balance between accuracy and consistency, outperforming the baseline and single-autoencoder variants with low standard deviations across runs.

Overall, these experiments demonstrate that integrating Multi-Head Attention within the MAGE framework consistently improves classification accuracy, precision, recall, and F1 score across both LSTM and Logistic Regression models, validating its robustness and reproducibility across independent runs and shuffled datasets.

5 Conclusion

We thus present an innovative approach to embedding refinement and classification for the Bantu language family by integrating embedding-level transformations and high-performing attention mechanisms. Through systematic experimentation, we demonstrated that denoising and variational autoencoders enhance the quality of embeddings by refining their structure while preserving semantic integrity. A key contribution of this study was the introduction of a multi-head attention mechanism. The attention mechanism dynamically assigns different weights to various embeddings, enabling the model to focus on the most relevant features. This approach allowed us to capture crucial aspects of the embedding space more effectively. The results from our experiments clearly indicate that multi-head attention significantly boosted classification performance across various metrics. This validates the hypothesis that refining embeddings and emphasizing important features through attention enhances classification performance. The framework proposed in this study provides a scalable and stable solution, particularly for low-resource languages where linguistic diversity and data scarcity pose unique challenges.

6 Future Works

While this study demonstrates significant improvements in embedding refinement and classification through a novel architecture and attention mechanisms, it is limited to three Bantu languages. Testing on a broader range of Bantu and other low-resource languages is needed to assess the generalizability of the approach. Additionally, expanding the variety of pre-trained embeddings and exploring novel attention mechanisms, such as hierarchical or adaptive attention, could further enhance performance and applicability.

7 Limitations

While our proposed approach demonstrates notable improvements in embedding refinement and classification, it has several limitations. First, our experiments were conducted on a limited set of three Bantu languages, which restricts the generalizability of our findings to other Bantu and low-resource languages. Second, the dataset exhibits an imbalance, with Kinyarwanda comprising the majority of data points. This skewness may introduce biases in model learning and

affect the performance across languages. Third, the dataset size may be insufficient for training complex components like the Denoising Autoencoder, Variational Autoencoder, and the standard Autoencoder as they require a large and diverse dataset to learn meaningful latent representations effectively. The limited training data could lead to suboptimal embeddings, affecting downstream classification performance. Fourth, while embedding-level transformations using denoising and variational autoencoders refine embedding structures, their impact on preserving linguistic nuances requires further investigation. Finally, the computational complexity of our approach, particularly the attention mechanism, may pose challenges for real-time applications in resource-constrained environments. Addressing these limitations in future research will be essential for broader adoption and scalability.

References

1. Adelani, D.I.: Natural Language Processing for African Languages. Dissertation, Universität des Saarlandes (2022). https://doi.org/10.22028/D291-40305, http://dx.doi.org/10.22028/D291-40305
2. Amol, C.J., Chimoto, E.A., Gesicho, R.D., Gitau, A.M., Etori, N.A., Kinyanjui, C., Ndung'u, S., Moruye, L., Ooko, S.O., Kitonga, K., Muhia, B., Gitau, C., Ndolo, A., Wanzare, L.D.A., Kahira, A.N., Tombe, R.: State of nlp in kenya: A survey (2024), https://arxiv.org/abs/2410.09948
3. Chen, J., Yang, Z., Yang, D.: Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification (2020), https://arxiv.org/abs/2004.12239
4. Jahan, M.S., Beddiar, D.R., Oussalah, M., Mohamed, M.: Data expansion using wordnet-based semantic expansion and word disambiguation for cyberbullying detection. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). pp. 1761–1770. European Language Resources Association (ELRA), Marseille, France (June 20–25 2022), https://aclanthology.org/2022.lrec-1.187.pdf
5. Jia, R., Liang, P.: Adversarial examples for evaluating reading comprehension systems (2017), https://arxiv.org/abs/1707.07328
6. Karimi, A., Rossi, L., Prati, A.: AEDA: An easier data augmentation technique for text classification. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2748–2754. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.234, https://aclanthology.org/2021.findings-emnlp.234/
7. Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations (2018), https://arxiv.org/abs/1805.06201
8. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only (2018), https://arxiv.org/abs/1711.00043
9. Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations (2023), https://arxiv.org/abs/2310.07849
10. Litake, O., Yagnik, N., Labhsetwar, S.R.: Inditext boost: Text augmentation for low resource india languages. ArXiv **abs/2401.13085** (2024), https://api.semanticscholar.org/CorpusID:267200269

11. Miller, G.A.: WordNet: A lexical database for English. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), https://aclanthology.org/H94-1111/
12. Muhammad, S.H., Abdulmumin, I., Yimam, S.M., Adelani, D.I., Ahmad, I.S., Ousidhoum, N., Ayele, A.A., Mohammad, S.M., Beloucif, M., Ruder, S.: SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). Association for Computational Linguistics (2023)
13. Nie, E., Liang, S., Schmid, H., Schütze, H.: Cross-lingual retrieval augmented prompt for low-resource languages. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 8320–8340. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.528, https://aclanthology.org/2023.findings-acl.528/
14. Ogueji, K., Zhu, Y., Lin, J.: Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In: Ataman, D., Birch, A., Conneau, A., Firat, O., Ruder, S., Sahin, G.G. (eds.) Proceedings of the 1st Workshop on Multilingual Representation Learning. pp. 116–126. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.mrl-1.11, https://aclanthology.org/2021.mrl-1.11/
15. Perçin, S., Galassi, A., Lagioia, F., Ruggeri, F., Santin, P., Sartor, G., Torroni, P.: Combining WordNet and word embeddings in data augmentation for legal texts. In: Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., Preoțiuc-Pietro, D. (eds.) Proceedings of the Natural Legal Language Processing Workshop 2022. pp. 47–52. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022). https://doi.org/10.18653/v1/2022.nllp-1.4, https://aclanthology.org/2022.nllp-1.4/
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023), https://arxiv.org/abs/1910.10683
18. Rahamim, A., Uziel, G., Goldbraich, E., Anaby Tavor, A.: Text augmentation using dataset reconstruction for low-resource classification. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 7389–7402. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.466, https://aclanthology.org/2023.findings-acl.466/
19. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1264, https://aclanthology.org/D16-1264
20. Sahu, G., Rodriguez, P., Laradji, I., Atighehchian, P., Vazquez, D., Bahdanau, D.: Data augmentation for intent classification with off-the-shelf large language models. In: Liu, B., Papangelis, A., Ultes, S., Rastogi, A., Chen, Y.N., Spithourakis, G., Nouri, E., Shi, W. (eds.) Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 47–57. Association for Computational Linguis-

tics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.nlp4convai-1.5, https://aclanthology.org/2022.nlp4convai-1.5/
21. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data (2016), https://arxiv.org/abs/1511.06709
22. Sujana, Y., Kao, H.Y.: Lida: Language-independent data augmentation for text classification. IEEE Access **PP**, 1–1 (01 2023). https://doi.org/10.1109/ACCESS.2023.3234019
23. Thangaraj, H., Chenat, A., Walia, J.S., Marivate, V.: Cross-lingual transfer of multilingual models on low resource african languages (2024), https://arxiv.org/abs/2409.10965
24. Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks (2019), https://arxiv.org/abs/1901.11196
25. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: Sequence generative adversarial nets with policy gradient (2017), https://arxiv.org/abs/1609.05473
26. Zhao, M., Zhang, L., Xu, Y., Ding, J., Guan, J., Zhou, S.: EPiDA: An easy plug-in data augmentation framework for high performance text classification. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4742–4752. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.349, https://aclanthology.org/2022.naacl-main.349/
27. Şahin, G.G.: To augment or not to augment? a comparative study on text augmentation techniques for low-resource nlp. Computational Linguistics **48**(1), 5–42 (04 2022). https://doi.org/10.1162/coli_a_00425, https://doi.org/10.1162/coli_a_00425