Style Extraction on Text Embeddings using VAE and Parallel Dataset

Anonymous ACL submission

Abstract

This study investigates the stylistic differences among various Bible translations using a Variational Autoencoder (VAE) model. By embedding textual data into high-dimensional vectors, the study aims to detect and analyze stylistic variations between translations, with a specific focus on distinguishing the American Standard Version (ASV) from other translations. The results demonstrate that each translation exhibits a unique stylistic distribution, which can be effectively identified using the VAE model. These findings suggest that the VAE model is proficient in capturing and differentiating textual styles, although it is primarily optimized for distinguishing a single style. The study highlights the model's potential for broader applications in AI-based text generation and stylistic analysis, while also acknowledging the need for further model refinement to address the complexity of multi-dimensional stylistic relationships. Future research could extend this methodology to other text domains, offering deeper insights into the stylistic features embedded within various types of textual data.

1 Introduction

Language, in both speech and writing, consists of two essential components: content and style. Broadly speaking, content refers to what is being expressed, while style pertains to how it is expressed. Specifically, style encompasses the variability of linguistic forms in actual language use (Babatunji, 2024). Historically, style in language has been studied within the field of Stylistics, a branch of applied linguistics that examines writing styles in literary criticism as well as tone and accent in discourse analysis.

When people hear the term "style," however, they often associate it with visual aesthetics in images rather than linguistic expression. Indeed, with the advancement of Generative AI, research on style has predominantly focused on images, particularly on style transfer techniques introduced by Gatys et al. (2015). This foundational work led to significant progress in computer vision, fostering a deep exploration of style transfer in visual contexts.

Although research on linguistic style in language models has been less explored compared to studies in the visual domain, language still provides a valuable area for stylistic analysis. In particular, linguistic style transfer has shown significant practical applications in real-world contexts, such as AI-driven text generation systems like Character AI, highlighting its growing importance in everyday use.

While style transfer techniques in images have made significant strides, their direct application to linguistic contexts encounter practical limitations. To overcome these challenges, this study adopts an alternative approach by quantifying style using embeddings. A central question we address is whether textual style can be represented as a measurable entity. Style, inherently subjective and complex, has been challenging to formalize, but we draw inspiration from the concept of word embeddings, which allow for semantic operations such as "king - man + woman = queen." By extending this concept to sentence embeddings, we aim to establish a robust framework for analyzing linguistic properties. Specifically, our research examines whether manipulated embeddings can effectively capture and quantify stylistic and semantic relationships in textual data. Furthermore, we investigate how these relationships emerge when texts with similar styles are modeled as belonging to the same probability distribution.

Although texts with similar styles may share underlying distributions, the exact differences between these distributions are not well understood. To address this, we employ Variational Autoencoders (VAEs) to normalize and regularize stylistic distributions, enabling a clearer analysis of whether

these distributions are distinctly separated. This approach allows us to quantify and compare stylistic variations in a mathematically robust manner.

The significance of this study lies in its ability to extract and quantify "style" from textual data as a measurable distribution. This quantification of style provides a foundation for practical applications, such as using these stylistic representations in generative models like GANs to enhance AI-driven content creation. By enabling more precise stylistic modeling, this research opens new possibilities for both theoretical exploration and applied advancements in text generation and stylistic analysis.

2 Related Works

2.1 Styles in Natural Languages

In linguistics, style has been seen as the unique way individuals or groups engage in conversation, conveying politeness or formality, and able to be controlled and adjusted to suit the intended social context, as Labov (1997) discusses. Additionally, style is the set of linguistic features such as tone, punctuation, word choice, and syntactic structure, playing a key role in stylistics and sentiment analysis (Pang and Lee, 2008).

Subsequently, Shen et al. (2017) considered style as the specific manner in which ideas are expressed in text, distinguishable from the content. In style transfer tasks, style is represented by personal style, formality, politeness, offensiveness, genre, and sentiment (Toshevska and Gievska, 2022). Current studies of styles focus on computational models for style transfer; Cross-Alignment with non-parallel text (Shen et al., 2017), Retrieve-and-Edit approach (Li et al., 2018), Unsupervised style transfer (Prabhumoye et al., 2018), Generative probabilistic model (He et al., 2020), and Reinforcement Learning for style transfer (Gong et al., 2019).

2.2 Evaluating Style Transfer in Text

Evaluation metrics are vital for text style transfer as they provide precise, quantitative assessments of how effectively the generated text adheres to the target stylistic attributes while preserving semantic integrity. However, evaluation can be challenging due to the subjective nature of style. It typically involves automatic evaluation and human evaluation (Jin et al., 2022).

2.2.1 Automatic Evaluation

The automatic evaluation measures how well the meaning of the original sentence was preserved in the output (generated sentence). The following metrics are commonly used:

- BLEU: Measures n-gram precision between generated text and references (Papineni et al., 2002).
- ROUGE: Assesses overlap of n-grams, focusing on recall to evaluate content coverage (Lin, 2004).
- METEOR: Evaluates translation quality using precision, recall, stemming, and synonymy (Banerjee and Lavie, 2005).
- BERTScore: Utilizes BERT embeddings to measure semantic similarity between generated and reference texts (Zhang et al., 2020).

2.2.2 Human Evaluation

Human judges assess how well the generated text adheres to the desired style and maintains semantic integrity. Yamshchikov et al. (2021) delineates the distinctions between human evaluation and automatic methods, illustrating how human assessment captures nuanced stylistic and semantic subtleties. However, it is costly and lacks the consistency, objectivity, and scalability provided by automatic evaluation methods. Additionally, both methods are limited by their reliance on reference texts, which may not fully capture the breadth of acceptable outputs or the creative potential of the generated text.

2.3 Anomaly Detection and VAE

Anomaly detection has evolved through various methodologies to address the challenge of identifying outliers across different domains (Schölkopf et al., 1999; Liu et al., 2008). The advent of deep learning introduced Autoencoders (Hinton and Salakhutdinov, 2006), which makes it possible to detect anomalies in high-dimensional data by analyzing reconstruction errors. Further advancements have been made with Variational Autoencoders (VAE), which leverage both probabilistic modeling and latent space representations (Kingma and Welling, 2022). We will employ a VAE model, trained on high-dimensional embedding vectors representing a single stylistic attribute, to identify anomalies by capturing deviations in stylistic characteristics.

3 Methodology

3.1 Data Collection and Preprocessing

This study utilizes biblical data collected from *Bible SuperSearch* (bib), a platform operating under the GNU GPL open source license. Ten different versions were initially considered: KJV, NET, ASV, ASVS, Coverdale, Geneva, KJV_Strongs, Bishops, Tyndale, and WEB. However, Bishops, Tyndale, and WEB were excluded due to insufficient parallel data. The remaining versions were selected for their linguistic diversity and historical backgrounds to enhance the depth of our style classification study.

The biblical texts are publicly available under the GNU GPL license, allowing free use for research purposes. Our study adhered to these guidelines without altering the original texts. In the preprocessing phase, we extracted the biblical data in JSON format and encoded all text files using UTF-8 to handle special characters. The initial data quality was high, minimizing the need for extensive text cleaning.

3.2 Embedding and Model Training

We employed OpenAI's text-embedding-3-small model to embed each biblical sentence into 1536-dimensional vectors. This model was chosen for its balance between performance and computational efficiency, making it suitable for our research. These high-dimensional vectors capture the nuanced language style of the sentences, providing foundational data for style-based classification.

3.3 Style Extraction

Text embedding is assumed to include both content and style, as represented by the following equation:

$$\text{text\_embedding} = \text{style\_embedding} + \text{content\_embedding}$$

Under this assumption, text embedding can be seen as simultaneously containing both the content and stylistic features of the text. In this study, we utilized this assumption to perform an analysis based on Bible data. The Bible data consists of the same verse expressed in multiple translations in a parallel structure, where the content remains the same, but the style varies. This characteristic of Bible data justifies the assumption that each translation's content embedding is identical. That is, the differences between the translations are primarily due to style, allowing for style analysis to be conducted. The core assumption of this study is that the difference in text embeddings between translations reflects the difference in styles. This can be expressed mathematically as follows:

$$\text{KJV\_embedding} - \text{Other\_embedding} = \text{KJV\_style\_embedding} - \text{Other\_style\_embedding}$$

Through this relationship, we calculated the difference between the two text embeddings and, based on this, measured the difference in style between the translations. Specifically, the goal of the study was to analyze the text embedding differences between KJV (King James Version) and other translations (e.g., ASV (American Standard Version)) to quantify the stylistic features. To do this, we calculated the difference between embeddings, represented as 1536-dimensional vectors, and used Variational Autoencoder (VAE) as a tool to analyze the distribution of these vectors.

The VAE is an unsupervised learning method that models the distribution of data in a latent space. In this study, we aimed to utilize the VAE to classify the embedding differences between translations and detect stylistic differences through anomaly detection. By compressing the input data and reconstructing it, VAE retains the important features while learning the distribution, allowing for the modeling of stylistic differences between translations.

During the training process of the VAE, we used the distribution differences between *KJV_embedding* and *ASV_embedding*. The VAE learned the difference between KJV and ASV embeddings in the latent space and then measured the similarity between the reconstructed distribution and the original distribution. We computed the L2-norm in this reconstruction process to quantitatively evaluate the stylistic similarity or difference between the translations. This allowed us to analyze the stylistic differences between KJV and ASV, as well as conduct comparative analyses with other translations.

In conclusion, this study evaluated the stylistic differences between Bible translations using VAE for anomaly detection. Through this process, we effectively quantified the stylistic similarities and differences between various translations. Based on the VAE model, trained on the difference be-

SYMBOL	DESCRIPTION
$\mathbf{k}^{(i)}$	EMBEDDING OF KJV, $i = 1, \cdots, N$
$\mathbf{a}^{(i)}$	EMBEDDING OF ASV
$\mathbf{y}_j^{(i)}$	EMBEDDING OF OTHER BIBLES, $j = 1, \cdots, 5$
$\mathbf{x}^{(i)}$	KJV_STYLE_EMBEDDING − ASV_STYLE_EMBEDDING
$\mathbb{R}^d$	d-DIMENSIONAL INPUT SPACE
$\mathbb{R}^p$	p-DIMENSIONAL FEATURE SPACE ($p < d$)
$\psi : \mathbb{R}^d \to \mathbb{R}^p$	ENCODER OF VAE
$\theta : \mathbb{R}^p \to \mathbb{R}^d$	DECODER OF VAE

Table 1: Notation used throughout this article.

tween *KJV_embedding* and *ASV_embedding*, we similarly analyzed the stylistic differences between other translations. This methodology enabled sophisticated text analysis that went beyond merely examining content features to include stylistic features. Thus, we provided new insights into how stylistic differences manifest within the embedding space.

3.4 Model Architecture and Training Details

The VAE model used in this study has an input dimension of 1536, and both encoder and decoder use fully connected (FC) layers. The size of each hidden layer follows a geometric sequence from the input dimension of 1536 to the final feature dimension (rounded to the nearest integer). Batch normalization is applied to all layers except the final output layers of both the encoder and decoder. The activation function used is Leaky ReLU (α=1e-2) except for the final output layer of the encoder and decoder. The final output layer of the decoder uses a Sigmoid-based activation function to ensure that the output distribution lies within the range [-1,1].

The hyperparameters are as follows: 6 values for the number of hidden layers (ranging from 1 to 6) and 6 values for the feature dimension (ranging from 2^3 to 2^8), resulting in 36 total combinations.

We split 13,823 sentence vectors into training and test sets with a 9:1 ratio, using KJV-ASV differences as training data. The model employs fully connected layers with batch normalization and Leaky ReLU activation, and is trained using the Adam optimizer and MSE loss function. A schematic of the model structure is provided in Figure 1.

Additionally, The model was trained using the Adam optimizer with a learning rate (lr) of 0.001. The batch size was set to 100, and training was conducted over 500 epochs. The input data was processed on a device configured to use CUDA. The input dimension of the sentence embeddings was 1536, and the fully connected layer dimensions were defined as [617, 247, 100, 40]. The latent space was set to a feature dimension of 16.

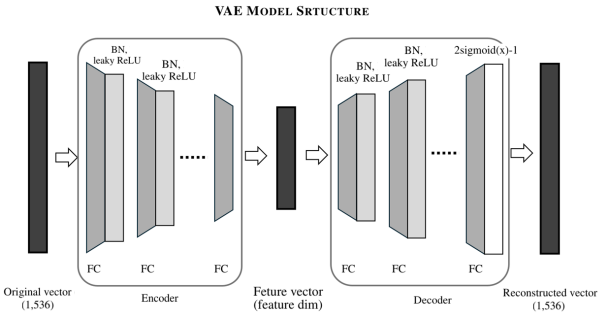

Figure 1: A schematic illustration of the VAE model. The encoder receives a 1,536-dimensional original (sentence embedding) vector as input and outputs a feature vector of the feature dimension. The decoder takes the feature vector of the feature dimension as input and outputs a 1,536-dimensional reconstructed vector.

3.5 Evaluation Metrics

According to our hypothesis, the KJV-ASV vector is expected to contain information related to the style of ASV, with KJV as the reference point. If a VAE with a sufficiently small feature dimension can effectively reconstruct this vector, it suggests that the VAE is leveraging specific stylistic features during the encoding-decoding process. On the other hand, if data not included in the model's training process are reconstructed through the VAE, the reconstruction quality is expected to be poor compared to the original. Based on this characteristic, we aim to perform anomaly detection using the VAE.

We aim to verify whether the VAE, trained using KJV-ASV vectors, has effectively learned the unique style of ASV. To do so, the trained VAE will be applied to six Bible translations (ASV, NET, ASVS, Coverdale, Geneva, and KJV Strongs), and we will examine if the model successfully distinguishes ASV's unique style compared to other translations. For the test dataset (not used during model training), ASV will serve as the normal data, and the other five translations (NET, ASVS, Coverdale, Geneva, and KJV Strongs) will serve

as anomaly data, consisting of sentence embedding vectors corresponding to the same Bible verses as in the test dataset. To remove the context of KJV during ASV training, the VAE was trained on the differences between the sentence vectors of ASV and KJV (KJV-ASV). Similarly, the anomaly data from the other Bible translations will be processed by subtracting the corresponding KJV sentence vectors, following the same procedure.

Among the 36 hyperparameter sets, the model that most clearly differentiates the reconstruction L2 error distribution between the training data and the anomalies will be considered the most effective in detecting the unique style of ASV. We will evaluate how well the original data and anomaly data are distinguished using Fisher's Linear Discriminant (FLD). FLD increases as the squared difference between the means of the two distributions becomes larger, and the sum of their variances becomes smaller. The formula for FLD S is as follows:

$$S = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

where μ_1 and μ_2 are the means of the original data and anomaly data distributions, respectively, and σ_1 and σ_2 are the variances of the original data and anomaly data distributions, respectively. This metric will help quantify how well the model separates the original data from anomalies based on reconstruction errors.

4 Results

4.1 Training Convergence and Loss Analysis

For all 36 hyperparameter combinations, both the training loss and test loss decreased and eventually converged, indicating that the models successfully learned from the data and reached a stable state in terms of reconstruction error. Detailed loss curves and analysis are provided in Figure 2.

4.2 L2 Error Distribution and FLD Analysis

The L2 error distribution for each model is presented in Figure 3. The minimum Fisher's Linear Discriminant (FLD) between the L2 norm distributions of the reconstructed sentence vectors from the trained dataset (ASV) and the anomaly datasets (NET, ASVS, Coverdale, Geneva, KJV Strongs) across the 36 models is shown in Figure 4.

The minimum FLD is more important than the maximum FLD for determining the separa-

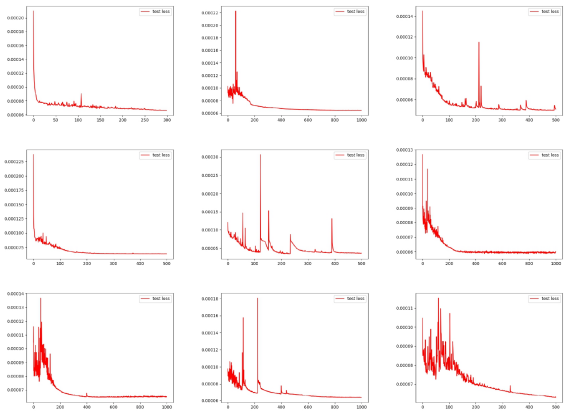

Figure 2: Test set loss during training. The x-axis represents the number of epochs, and the y-axis represents the mean error. The hyperparameters of each model are as follows: starting from left the 1st, 2nd, and 3rd columns represent feature dimensions of 8, 64, and 256, respectively, and the starting from top 1st, 2nd, and 3rd rows represent 1, 3, and 6 hidden layers, respectively.

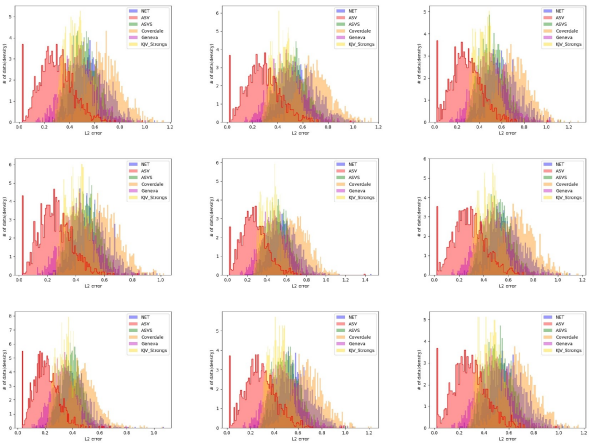

Figure 3: L2 error distribution on ASV, NET, ASVS, Coverdale, Geneva, and KJV Strongs. The x-axis represents the L2 error between the original and reconstructed sentence vector, and the y-axis represents the distribution density. The hyperparameters of each model are as follows: starting from left the 1st, 2nd, and 3rd columns represent feature dimensions of 8, 64, and 256, respectively, and starting from top the 1st, 2nd, and 3rd rows represent 1, 3, and 6 hidden layers, respectively.

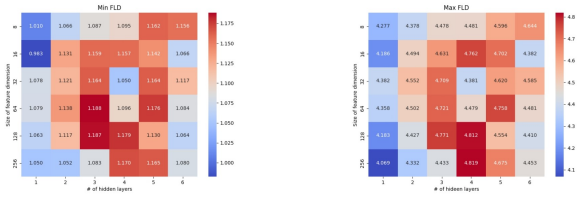

Figure 4: (Left) Minimum and (Right) Maximum of FLD between ASV and other 5 anomaly datasets (NET, ASVS, Coverdale, Geneva, and KJV Strongs). A higher minimum FLD indicates better differentiation between ASV and anomaly L2 error distributions.

tion between normal and anomaly data. A high minimum FLD represents the model that has the most differentiation between the ASV original and the anomaly reconstructions, indicating the best-performing model in terms of distinguishing between the original and anomalous styles based on the L2 norm distribution. Figure 4 shows that the minimum FLD is maximized in models with 3 hidden layers and a feature dimension size between 32 and 128. Models with too small or too large hidden layers and feature dimensions tend to perform poorly in anomaly differentiation.

Across the 36 models, the anomaly dataset that produced the minimum FLD most frequently was Geneva, appearing 31 times, followed by KJV Strongs, which appeared 5 times. This suggests that the L2 error distribution of the Geneva dataset was generally the closest to that of ASV, making it the hardest to distinguish from ASV. Conversely, the anomaly dataset that consistently produced the maximum FLD in all 36 models was Coverdale, indicating that it was the easiest to distinguish from ASV based on the L2 error distribution. This result highlights the distinctiveness of Coverdale's style compared to ASV, while Geneva's style appears more similar.

4.3 Impact of Context Subtraction on VAE Performance

Training the VAE without subtracting context parallel sentence (KJV) vectors demonstrated that both the training loss and test loss decreased and converged, indicating successful learning. However, as shown in Figure 5, the mean L2 error across all distributions was higher compared to the models trained with parallel sentence subtraction.

When comparing Figures 6 and 4, the Fisher's Linear Discriminant (FLD) for the no-subtraction case (from context parallel sentence vectors) is sig-

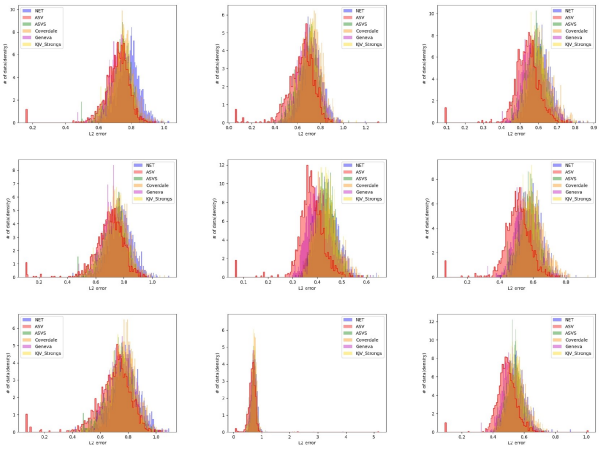

Figure 5: L2 error distribution on ASV, NET, ASVS, Coverdale, Geneva, and KJV Strongs, without parallel sentence (KJV) subtraction. The x-axis represents the L2 error between the original and reconstructed sentence vector, and the y-axis represents the distribution density. The hyperparameters of each model are as follows: starting from left the 1st, 2nd, and 3rd columns represent feature dimensions of 8, 64, and 256, respectively, and the starting from top 1st, 2nd, and 3rd rows represent 1, 3, and 6 hidden layers, respectively.

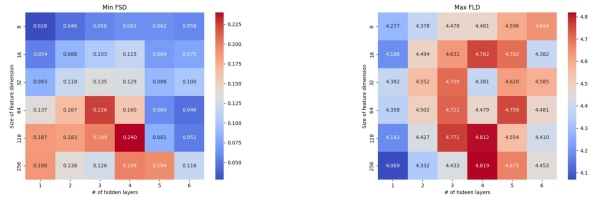

Figure 6: (Left) Minimum and (Right) Maximum of FLD between ASV and other 5 anomaly datasets (NET, ASVS, Coverdale, Geneva, and KJV Strongs), without parallel sentence (KJV) subtraction. A higher minimum FLD indicates better differentiation between ASV and anomaly L2 error distributions.

nificantly lower than for the subtracted case. Specifically, the mean of the minimum FLD across the 36 models in the subtracted case is 1.111, while the mean for the no-subtraction case is 0.116, making the FLD approximately 9.6 times lower without subtraction.

Furthermore, the highest maximum FLD in the no-subtraction case (1.000) is nearly the same as the lowest minimum FLD in the subtracted case (0.983). This stark difference in FLD highlights that when trained without subtracting the context parallel sentence vectors, the VAE's ability to distinguish anomalies from normal (trained domain) data is significantly diminished. This result reinforces the idea that the subtraction of context helps the VAE better capture stylistic differences, leading to clearer separation between ASV and other translations.

5 Discussion

This study extracted the styles of various Bible translations and utilized a Variational Autoencoder (VAE) model to analyze how these styles differ, particularly in comparison to the American Standard Version (ASV). The results revealed that the styles of each Bible translation followed a normal distribution, and these distributions could be clearly distinguished from that of the ASV. This indicates that there are stylistic differences between the ASV and other translations, and that these differences can be effectively detected using the VAE model.

After optimizing the VAE model's hyperparameters, the process of distinguishing between the ASV and other translation styles resulted in a Type 1 error of 8.7% and a Type 2 error of 6.7%, with a total error rate of 15.3%. Conversely, the model achieved an accuracy of 84.7%, demonstrating its ability to effectively differentiate styles. This level of accuracy suggests that the model can clearly recognize the distribution of a specific style and use it as a basis to distinguish between the styles of different translations.

However, the VAE model was optimized for distinguishing a single style. While it was useful for detecting differences between a specific translation style and the ASV, it had limitations when it came to distinguishing multiple styles simultaneously or understanding the relationships between complex, multi-dimensional styles. These limitations stem from the structural characteristics of the VAE, which compresses the data's features during learning, making it inherently challenging to fully capture the complex characteristics of the data. Therefore, to distinguish multiple styles simultaneously, it may be necessary to use other models or train the VAE model in a more sophisticated manner.

MODEL	ACCURACY	TYPE I ERROR	TYPE II ERROR
MODEL 1	83.5%	9.8%	6.7%
MODEL 2	82.9%	10.1%	7.0%
MODEL 3	83.4%	9.8%	6.8%
AVERAGE	**83.3%**	**9.9%**	**6.8%**

Table 2: Accuracy & Error Rates of Models 1, 2, and 3 on Anomaly Detection

The ability to extract a specific style suggests that the style's characteristics can be quantified and represented as a probability distribution. This means that AI can utilize this quantified style representation to generate text that adheres to a specific style. For example, in text generation tasks where a particular writing style or tone is required, a 'style metric' could be used as a numerical and comparable indicator to assess and ensure that the generated text conforms to the desired style.

The approach taken in this study opens up the possibility of expanding the research to other parallel text datasets. By applying this methodology to other text domains, researchers can study the stylistic differences and their implications within each domain. For example, the approach could be extended to analyze the styles of different translations of literary works, legal document translations, or works by various authors.

We have demonstrated that the VAE model can distinguish between the original and anomaly data using the reconstruction L2 error. To measure the overall accuracy, False Positive Rate (FPR), and False Negative Rate (FNR) of the model, we created an Accuracy Test Dataset using data not included in the training set. This dataset consisted of 1,000 samples, with 50% of the samples being from ASV and the remaining 50% from five anomaly datasets (NET, ASVS, Coverdale, Geneva, and KJV Strongs).

The binary classification results showed that the lowest overall error rate was achieved when the threshold was set at mean + 0.8 std. The average overall error rate across the three models was 16.8%. The relatively high FNR, particularly with Geneva, suggests that modern English Bible trans-

VERSE	TRANSLATION
GEN 1:1	**ASV:** IN THE BEGINNING GOD CREATED THE HEAVENS AND THE EARTH. **GENEVA:** IN THE BEGINNING GOD CREATED THE HEAUEN AND THE EARTH. **COVERDALE:** IN YE BEGYNNYNGE GOD CREATED HEAUEN & EARTH:
MAT 1:1	**ASV:** THE BOOK OF THE GENERATION OF JESUS CHRIST, THE SON OF DAVID, THE SON OF ABRAHAM. **GENEVA:** THE BOOK OF THE GENERATION OF JESUS CHRIST THE SON OF DAVID, THE SON OF ABRAHAM. **COVERDALE:** THIS IS THE BOKE OF THE GENERACION OF IESUS CHRIST YE SONNE OF DAUID, THE SONNE OF ABRAHAM.
JOH 3:16	**ASV:** FOR GOD SO LOVED THE WORLD, THAT HE GAVE HIS ONLY BEGOTTEN SON, THAT WHOSOEVER BELIEVETH ON HIM SHOULD NOT PERISH, BUT HAVE ETERNAL LIFE. **GENEVA:** FOR GOD SO LOVETH THE WORLD, THAT HE HATH GIVEN HIS ONLY BEGOTTEN SON, THAT WHOSOEVER BELIEVETH IN HIM, SHOULD NOT PERISH, BUT HAVE EVERLASTING LIFE. **COVERDALE:** FOR GOD SO LOUED THE WORLDE, THAT HE GAUE HIS ONELY SONNE, THAT WHO SO EUER BELEUETH IN HI, SHULDE NOT PERISHE, BUT HAUE EUERLASTINGE LIFE.

Table 3: Original Sentences of 3 Different Versions: ASV, Geneva, Coverdale

lations inherently do not exhibit distinct stylistic differences.

The results of the anomaly detection using the VAE in this study also show trends similar to what would be expected when humans classify ASV and other Bible versions. In this study, the L2 error distributions of ASV and Geneva had a significant overlap, making it difficult to classify them with a low error rate using a specific threshold. In contrast, the L2 error distribution of Coverdale barely overlapped with ASV, and the FLD was the highest across all models. More typically, 67.8% (ASV 34.6%, KJV Strongs 33.2%) of type 2 error in anomaly detection is from Geneva and KJV Strongs.

Table 3 illustrates the textual differences between three versions of the Bible (ASV, Geneva, Coverdale), which could influence the VAE's ability to distinguish anomalies. The relatively low accuracy of anomaly detection using the VAE in this study may be attributed to the subtle stylistic differences between the texts. This implies that using sentences with clearer stylistic differences and more varied contexts in future experiments could result in better accuracy.

6 Conclusion

This study has successfully demonstrated the application of a Variational Autoencoder (VAE) model to analyze and distinguish the stylistic differences among various Bible translations, with a particular focus on the American Standard Version (ASV). By embedding textual data into high-dimensional vectors and applying anomaly detection techniques, the study identified unique stylistic distributions for each translation, showcasing the model's capability to differentiate between these styles with a notable accuracy rate of 84.7%. The findings confirm that the VAE model can effectively capture and differentiate textual styles, though it is primarily optimized for distinguishing a single style.

The implications of this research extend beyond academic inquiry, offering significant potential applications in the field of AI-driven text generation. The ability to extract and measure specific stylistic features opens up possibilities for generating texts with targeted stylistic attributes, which can be invaluable in automated writing tools, personalized content creation, and stylistic analysis of literary works. Moreover, the methodology employed in this study can be adapted to other text domains, providing a framework for analyzing stylistic differences across various types of textual data, including literary translations, legal documents, and author-specific writing styles.

In conclusion, while this study provides a solid foundation for the analysis of textual styles using VAE, it also sets the stage for future research to explore more sophisticated models and methodologies. By expanding this approach to other text domains and enhancing the model's capabilities, future work can continue to deepen our understanding of textual styles and their applications in AI and beyond.

7 Limitations

This study demonstrates the potential of Variational Autoencoder (VAE) models for extracting and analyzing stylistic differences in text embeddings. However, several limitations exist that need to be addressed in future research:

Single-Style Optimization The VAE model employed in this study is primarily optimized for distinguishing a single style, limiting its capability to simultaneously differentiate multiple styles or analyze complex, multi-dimensional stylistic relationships often present in natural language data. This focus on single-style optimization stems from the model's training, which is designed to determine the presence or absence of a single style. Addressing this limitation may involve incorporating methodologies that explicitly classify and separate different types within the latent space.

Challenges with Subtle Style Differences For certain translations, such as ASV and Geneva, the model struggles due to significant overlap in their L2 error distributions. This indicates a limitation in distinguishing texts with subtle stylistic differences, where stylistic nuances may not be adequately captured.

Dependence on Context Subtraction The model's effectiveness relies heavily on subtracting context-parallel sentences (e.g., KJV embeddings). Without this preprocessing step, its ability to differentiate between normal data and anomalies is significantly reduced, demonstrating a strong dependence on this technique.

Incorporating large language models (LLMs) such as ChatGPT could potentially enhance this process. These models are expected to facilitate the generation of parallel datasets through methods like translation or simplifying text into "kindergarten English." Such approaches could improve the preprocessing pipeline and enhance the model's ability to separate anomalies from normal data.

However, these methods serve as interim solutions, as they do not fully preserve the original content. To achieve robust and effective preprocessing, approaches that can completely retain the content must be developed and applied.

High False Negative Rate (FNR) The study reports a relatively high false negative rate, particularly for Geneva translations. This suggests that certain stylistic differences in modern English Bible translations are too subtle for the model to reliably detect.

Error Rates Despite achieving a commendable accuracy of 84.7%, the model exhibits significant error rates, including a Type I error rate of 8.7% and a Type II error rate of 6.7%. These error rates could hinder its applicability in tasks requiring higher precision.

Need for Advanced Techniques Addressing these limitations may require integrating more advanced machine learning techniques, such as models capable of handling multi-dimensional stylistic relationships or other unsupervised learning approaches. Further refinement of the VAE architecture could also enhance its performance. Currently, the study relies on extracting style through embedding subtraction. While the linearity of word embeddings is well-documented, the linearity of sentence embeddings has only been assumed. Exploring alternative methodologies for style extraction could provide significant improvements.

In summary, while this study provides a solid foundation for analyzing stylistic differences using VAE, these limitations highlight the need for future research to explore more robust and generalizable methods.

References

Bible SuperSearch. https://www.biblesupersearch.com/Bible. Accessed: 2024-08-22.

Adepoju Babatunji. 2024. *LINGUISTIC STYLISTICS*, pages 61–80. Publisher Name.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A Neural Algorithm of Artistic Style. *arXiv preprint arXiv:1508.06576*.

Hongyu Gong, Suma Bhat, Lingfei Wu, Jinjun Xiong, and Wen mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. *Preprint*, arXiv:1903.10671.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *Preprint*, arXiv:2002.03912.

G.E. Hinton and R.R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313:504–7.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text

style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Diederik P Kingma and Max Welling. 2022. Auto-encoding variational bayes. *Preprint*, arXiv:1312.6114.

William Labov. 1997. *The Social Stratification of (r) in New York City Department Stores*, pages 168–178. Macmillan Education UK, London.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Martina Toshevska and Sonja Gievska. 2022. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.