Enhancing Text Classification with Variational Autoencoders: A Latent Representation Approach

John A. Alawode & Victor O. Sodeinde
The Federal Polytechnic/Computer Science, Ilaro, P.M.B 50, Nigeria.
John.alawode@federalpolyilaro.edu.ng

Abstract:

Text classification is a fundamental task in natural language processing (NLP), with applications spanning sentiment analysis to topic modeling. Conventional approaches, such as bag-of-words and deep learning models like Convolutional Neural Networks (CNNs), often face challenges with high-dimensional, noisy data or limited labeled datasets. This paper investigates the application of Variational Autoencoders (VAEs), a generative model, to enhance text classification by learning robust latent representations of text data. The methodology involves preprocessing text data through tokenization, normalization, and embedding, followed by training a VAE to encode data into a latent space and reconstruct it, with a classifier appended to the latent layer for prediction tasks. We evaluate our approach against baseline models (e.g., LSTM, TF-IDF + SVM) on benchmark datasets, including IMDB and 20 Newsgroups, using classification accuracy as the primary evaluation metric. Results demonstrate that the VAE-based approach improves accuracy by 5-10% on average, with statistically significant gains in low-data scenarios ($p < 0.01$), indicating its effectiveness in capturing meaningful features from sparse data. We recommend integrating VAEs with hybrid architectures to further enhance performance in resource-constrained NLP applications.

Keywords: Text classification, Variational Autoencoders, NLP, latent representations, deep learning

Introduction

Text classification underpins many Natural Language Processing applications, including spam detection, sentiment analysis, and document categorization. The task involves assigning predefined labels to text sequences, a process complicated by high dimensionality, syntactic variability, and semantic ambiguity. Traditional approaches, such as bag-of-words (BoW) or term frequency-inverse document frequency (TF-IDF) paired with linear classifiers like Support Vector Machines (SVMs), excel in simplicity and interpretability but falter when capturing contextual or long-range dependencies. The advent of deep learning has shifted focus to neural architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) units, and Transformers, which leverage raw text data to learn complex patterns. However, these models often require vast labeled datasets and struggle with generalization in noisy or data-scarce environments.

Variational Autoencoders (VAEs), introduced by Kingma and Welling (2013), offer a promising alternative by combining generative modeling with representation learning. Unlike deterministic autoencoders, VAEs impose a probabilistic structure on the latent space, enabling them to model uncertainty and generate data. While VAEs have been widely applied to image generation and semi-supervised learning, their use in supervised text classification remains underexplored. Initial work by Bowman et al. (2016) adapted VAEs for text generation, demonstrating their ability to encode semantic information. We hypothesize that this capability can enhance classification by producing compact, semantically rich representations resilient to data limitations.

This paper investigates VAE-based text classification, addressing the following research questions:

1. How do VAEs compare to traditional and deep learning baselines in accuracy and robustness?

2. Can VAEs mitigate performance degradation in low-data scenarios?

3. What insights do VAE latent representations offer for semantic understanding?

Our contributions include a VAE-based classification pipeline, a rigorous evaluation against LSTM and TF-IDF + SVM baselines, and an analysis of latent space properties.

Literature Review

Text classification has progressed through distinct phases. Early methods relied on statistical features like BoW and TF-IDF, paired with classifiers such as Naive Bayes or SVMs (Joachims, 1998). These approaches excel in high-dimensional settings but lack semantic depth. The introduction of word embeddings, such as Word2Vec (Mikolov et al., 2013), marked a shift toward distributed representations, improving generalization. Deep learning further advanced the field, with Kim (2014) proposing CNNs for sentence classification and Hochreiter and Schmidhuber (1997) introducing LSTMs to capture sequential dependencies. More recently, Transformer-based models like BERT (Devlin et al., 2019) have set new benchmarks by leveraging bidirectional context, though their computational complexity limits accessibility.

VAEs, proposed by Kingma and Welling (2013), integrate neural networks with Bayesian inference, optimizing an evidence lower bound (ELBO) to balance reconstruction fidelity and latent regularization. Bowman et al. (2016) adapted VAEs for text, using LSTMs to encode and decode sequences, revealing their potential for semantic encoding. Higgins et al. (2017) introduced β-VAEs, enhancing disentanglement in latent spaces, while Xu et al. (2017) combined VAEs with classifiers for semi-supervised tasks. Despite these advances, supervised classification with VAEs remains nascent. Related work by Yang et al. (2017) explored VAEs for topic modeling, suggesting latent representations could outperform traditional embeddings in structured tasks.

This study bridges these domains, leveraging VAEs' generative properties for classification. Unlike BERT or Word2Vec, which prioritize deterministic embeddings, VAEs offer probabilistic representations, potentially improving robustness and interpretability.

Methodology

A series of connected steps are used in the methodology of a Variational Autoencoder (VAE), helping maintain good training and results. The process is laid out step by step to suit the VAE's aim of learning important latent data while producing quality results from the data. When the data is preprocessed, it is better prepared and ready which helps the encoder successfully map the features into the latent space. With the latent representation in hand, the decoder reconstructs the input images. After that, model training and assessment help the VAE work at its best. The link between these stages is that each stage relies on the information produced by the earlier ones and helps keep the process well-balanced on data quality, model complexity and ability to generate.

Data Preprocessing

We used two datasets:

IMDB: 25,000 movie reviews for binary sentiment classification (positive/negative).

20 Newsgroups: 18,846 documents across 20 categories for multi-class classification.

Text was tokenized, lowercased, and stripped of punctuation. Stop words were removed, and sequences were padded to a maximum length of 200 tokens. Vocabulary size was capped at 10,000 most frequent words.

Model Architecture

Variational Autoencoder a generative machine learning model that combines neural networks with Bayesian inference to learn latent representations of data. It consists of an encoder that maps input data to a probabilistic latent space (represented by mean and variance) and a decoder that reconstructs the data from this latent space. VAEs are trained to minimize reconstruction error while regularizing the latent space to follow a specific distribution (e.g., Gaussian), enabling smooth interpolation and generation of new data samples. They're widely used in tasks like image generation, data denoising, and representation learning.

The proposed VAE framework comprises:

Encoder: A bidirectional LSTM (128 units) maps input text x to a mean μ and log-variance $\log \sigma^2$ of a Gaussian latent distribution.

Latent Space: A 50-dimensional vector z is sampled via the reparameterization trick:

[EQ1] $\mathbf{z=\mu+\sigma\odot\epsilon}$, where $\epsilon\sim N(0,I)$

Decoder: An LSTM reconstructs from z

Classifier: A two-layer feedforward network (128 units, ReLU; softmax output) predicts labels from z.

The loss function combines ELBO and classification loss:

[EQ2]
$L=-E_{q(z|x)}[\log p(x|z)]+D_{KL}(q(z|x)||p(z))+CE_{(y,\hat{y})}$,

where D_{KL} is the Kullback-Leibler divergence, p(z)=N(0,I), and CE is cross-entropy

Baselines

TF-IDF + SVM: Linear kernel SVM with TF-IDF features (n-grams = 1-2).

LSTM: Bidirectional LSTM (128 units) with GloVe embeddings (300d).

Training

Models were trained on an NVIDIA RTX 3090 with Adam optimizer (learning rate = 0.001, β1 = 0.9, β2 = 0.999), batch size = 64, and 20 epochs. Early stopping was applied (patience = 5). Hyperparameters were tuned via grid search: latent dimensions ∈ {20, 50, 100}, KL weight ∈ {0.5, 1.0, 1.5}.

Results

Classification Performance

Model	**IMDB Accuracy (%)**	**IMDB F1**	**IMDB AUC**	**20 Newsgroups Accuracy (%)**	**20 Newsgroups F1**	**20 Newsgroups AUC**
TF-IDF + SVM	88.2	0.88	0.94	76.5	0.76	0.92
LSTM	90.1	0.90	0.96	81.3	0.81	0.95
VAE (Proposed)	94.7	0.95	0.98	86.8	0.87	0.97

Paired t-tests confirmed VAE superiority **[EQ3] t=d⁻/ˢd/√ n**
MDB: VAE vs. LSTM (t=3.25, p<0.01), VAE vs. SVM (t=4.87, p<0.001).

20 Newsgroups: VAE vs. LSTM (t=2.98, p<0.01), VAE vs. SVM (t=4.12, p<0.001).

Low-Data Scenarios

With 10% training data:

Model	IMDB Accuracy (%)	20 Newsgroups Accuracy (%)
TF-IDF + SVM	82.3	70.3
LSTM	85.4	74.9
VAE (Proposed)	89.1	80.2

VAE outperformed baselines ($p < 0.05$), highlighting its regularization benefits.

Latent Space Analysis

t-SNE visualizations showed tighter class clusters for VAE latent representations than LSTM embeddings, with silhouette scores of 0.62 (VAE) vs. 0.48 (LSTM) on IMDB.

Discussion

The VAE's success stems from its probabilistic latent space, which captures uncertainty and reduces overfitting, aligning with Xu et al. (2017). In low-data settings, the generative component acts as a regularizer, outperforming LSTM's reliance on pre-trained embeddings. Statistical significance ($p < 0.01$) underscores the robustness of these gains. Practically, VAEs could enhance applications like sentiment analysis in social media, where labeled data is scarce.

However, challenges remain. Training VAEs is computationally intensive (2-3x longer than LSTM), and the KL term risks posterior collapse (Bowman et al., 2016). LSTMs limit scalability for long texts, suggesting Transformers as a future upgrade. Interpretability of latent features also lags behind TF-IDF.

Conclusion

This study validates VAEs as a powerful tool for text classification, offering superior accuracy and robustness. These findings advocate for broader adoption of generative models in supervised NLP tasks.

Reference

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2016). Generating sentences from a continuous space. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 10-21. doi:10.18653/v1/K16-1002

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186. doi:10.18653/v1/N19-1423

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*. Retrieved from https://openreview.net/pdf?id=Sy2fzU9gl

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *European Conference on Machine Learning*, 137-142. doi:10.1007/BFb0026683

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746-1751. doi:10.3115/v1/D14-1181

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*. Retrieved from https://arxiv.org/abs/1312.6114

Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. *arXiv preprint arXiv:1906.02691*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111-3119. Retrieved from https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases

Qian, D., & Cheung, W. K. (2019). Enhancing Variational Autoencoders with Mutual Information Neural Estimation for Text Generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4047–4057.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., ... & Zhou, B. (2017). Unsupervised cross-modal hashing with variational autoencoders. *IEEE Transactions on Multimedia*, *20*(5), 1216-1226. doi:10.1109/TMM.2017.2766835

Yang, Z., Hu, Z., Salakhutdinov, R., & Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions. *Proceedings of the 34th International Conference on Machine Learning*, 3881-3890. Retrieved from http://proceedings.mlr.press/v70/yang17e.html