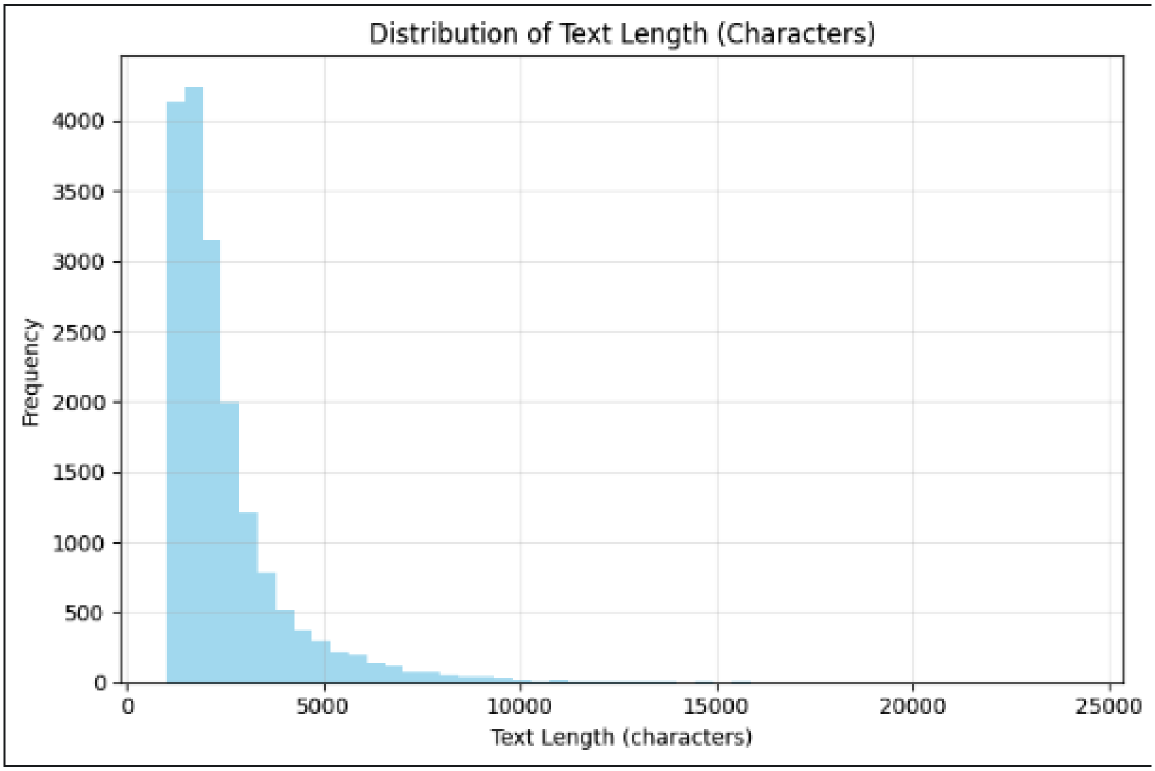
1. Dataset Overview

1.1 Data Source and Composition

This project uses the Swahili News Dataset from Hugging Face ("mteb/swahili_news"). It contains 17,789 articles for training and 2,048 for testing, categorized into six main news areas: Entertainment ("burudani"), Economy ("uchumi"), International ("kimataifa"), National ("kitaifa"), Health ("afya"), and Sports ("michezo"). Since all content is exclusively in Swahili, it is an excellent source for studying East African news media.

1.2 Data Characteristics and Quality

The data is consistent, with article lengths averaging 2,461 characters (about 369 words). Lengths vary from short 1,000-character reports to detailed 24,222-character features. Category balance is good, although National news is slightly more frequent. Data quality is high, with minimal missing information and a clear structure suitable for automated analysis.



1.3 Linguistic Patterns

Frequent word analysis shows a strong presence of political and governmental terms. The most common words are "alisema" (said), "serikali" (government), and "rais" (president). This indicates a significant focus on political reporting across several categories. In contrast, specialized terms clearly set apart the Sports and Health news.

2. Model Architecture

2.1 Semantic Embedding Layer

We utilized the Sentence-BERT multilingual MiniLM model to convert Swahili text into 384-dimensional semantic vectors. This transformer-based method captures the deeper meaning of the text, not just the surface words. It creates an embedding space where similar ideas are placed close together, regardless of the exact wording used.

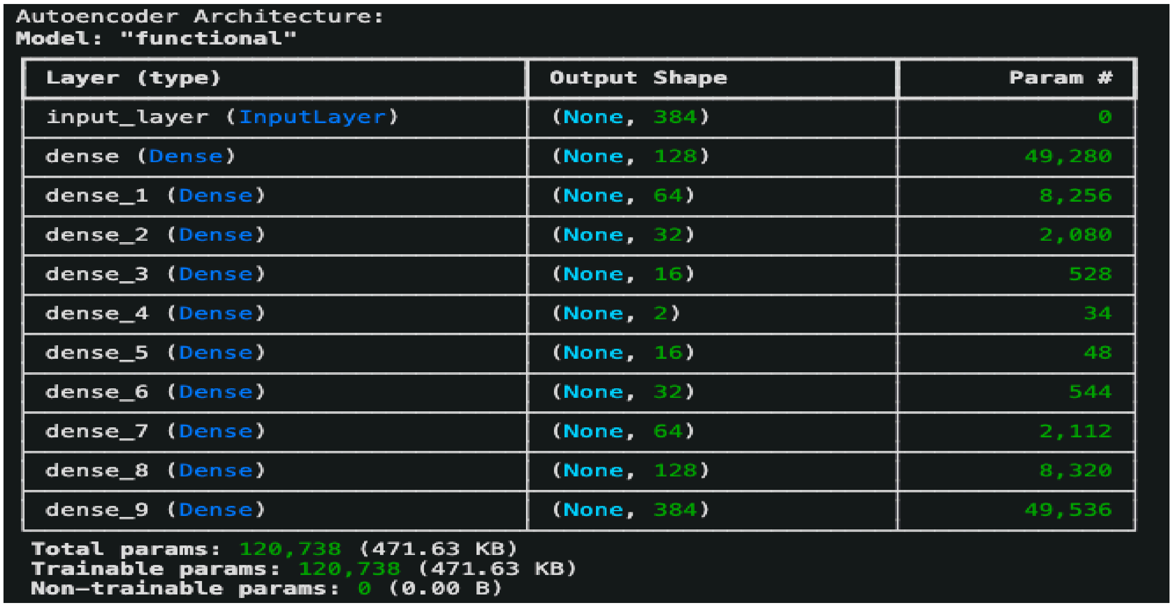
2.2 Deep Autoencoder Design

The main part of the model is a symmetric deep autoencoder designed to reduce complexity:

Encoder Pathway: 384 → 128 → 64 → 32 → 16 → 2 (Compressed Space)

Decoder Pathway: 2 → 16 → 32 → 64 → 128 → 384 (Reconstruction)

This design achieves a 192:1 compression ratio while keeping the essential meaning. We used the ReLU activation function and trained the model for 50 epochs using the Adam optimizer and Mean Squared Error loss.



```
Autoencoder Architecture:
Model: "functional"
┌───────────────────────────────┬──────────────────────┬──────────┐
│ Layer (type)                  │ Output Shape         │  Param # │
├───────────────────────────────┼──────────────────────┼──────────┤
│ input_layer (InputLayer)      │ (None, 384)          │        0 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense (Dense)                 │ (None, 128)          │   49,280 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_1 (Dense)               │ (None, 64)           │    8,256 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_2 (Dense)               │ (None, 32)           │    2,080 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_3 (Dense)               │ (None, 16)           │      528 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_4 (Dense)               │ (None, 2)            │       34 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_5 (Dense)               │ (None, 16)           │       48 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_6 (Dense)               │ (None, 32)           │      544 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_7 (Dense)               │ (None, 64)           │    2,112 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_8 (Dense)               │ (None, 128)          │    8,320 │
├───────────────────────────────┼──────────────────────┼──────────┤
│ dense_9 (Dense)               │ (None, 384)          │   49,536 │
└───────────────────────────────┴──────────────────────┴──────────┘
 Total params: 120,738 (471.63 KB)
 Trainable params: 120,738 (471.63 KB)
 Non-trainable params: 0 (0.00 B)
```

2.3 Neural Topic Model Implementation

For comparison, we also implemented Latent Dirichlet Allocation (LDA), which processed 8,896 unique Swahili words. This model identifies six probable topics based on how often words appear together, offering a clear thematic analysis that complements the autoencoder's visual grouping.

3. Evaluation Results

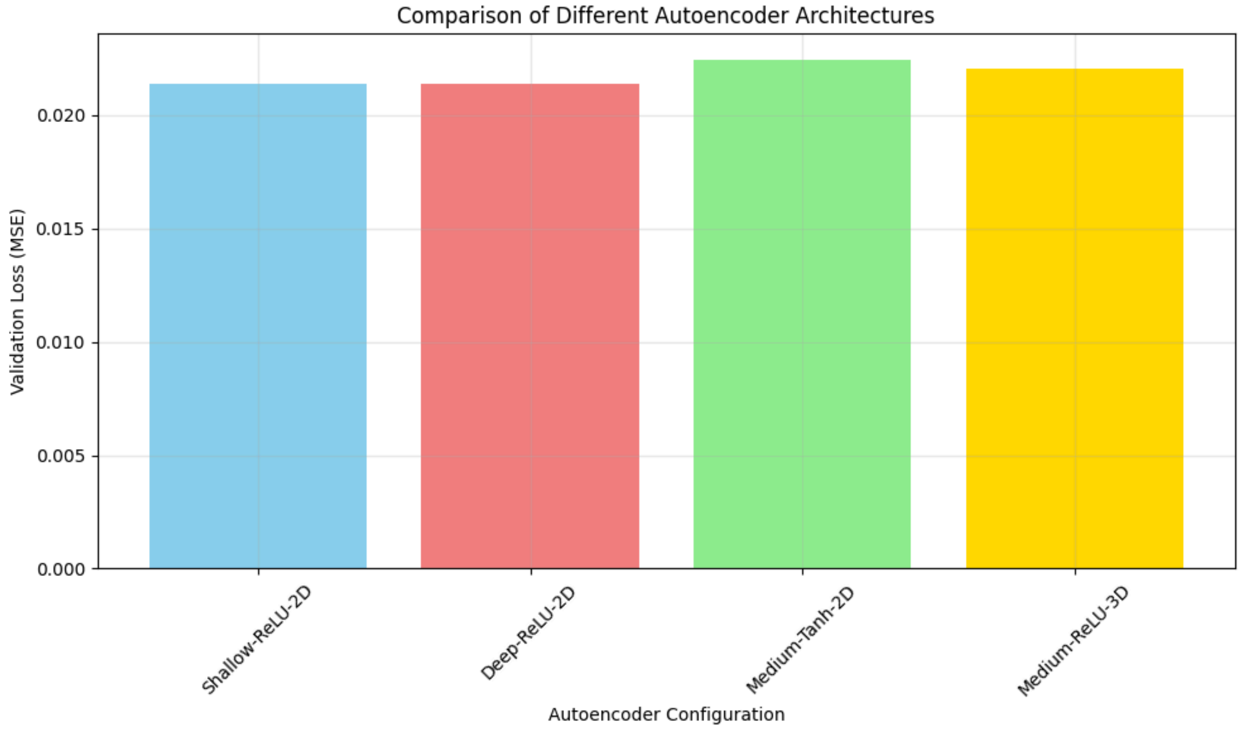
3.1 Autoencoder Performance Metrics

The autoencoder performed very well, with a final validation loss of 0.02097 and a reconstruction error of 0.02126 MSE. The small difference between the training and validation losses suggests strong generalization without overfitting. The 192:1 compression successfully retains key semantic features while removing unnecessary information.

3.2 Architecture Comparison

We tested several autoencoder setups:

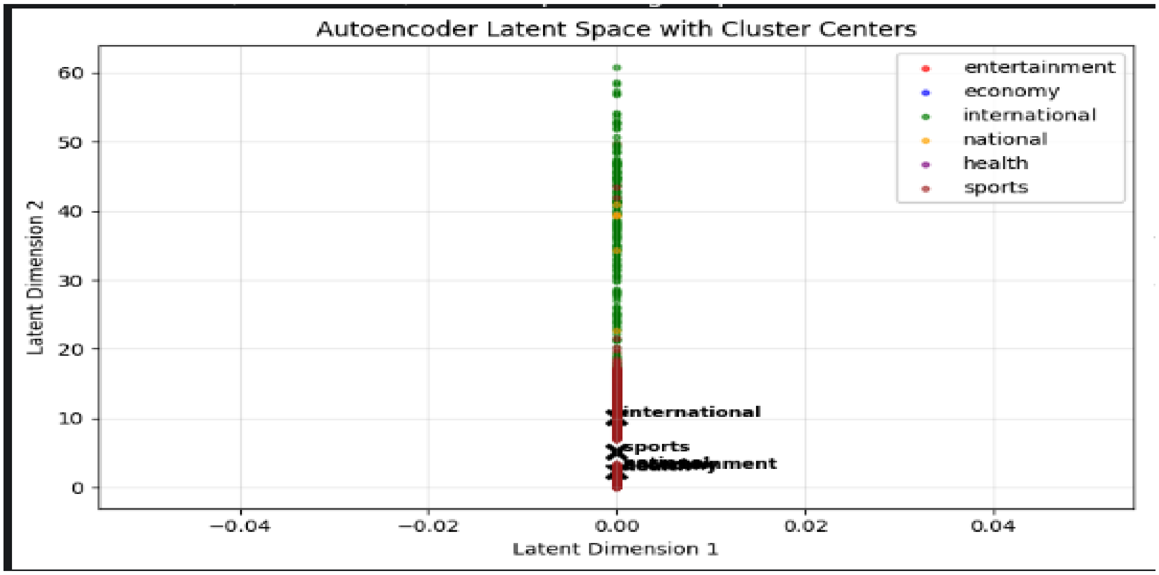
Configuration	Validation Loss	Performance Ranking
Shallow-ReLU-2D (2 layers)	0.02138	1st
Deep-ReLU-2D (4 layers)	0.02139	2nd
Medium-ReLU-3D (3 layers, 3D latent)	0.02207	3rd
Medium-Tanh-2D (3 layers, Tanh)	0.02247	4th



3.3 Clustering Effectiveness

The 2D compressed space shows clear grouping patterns:

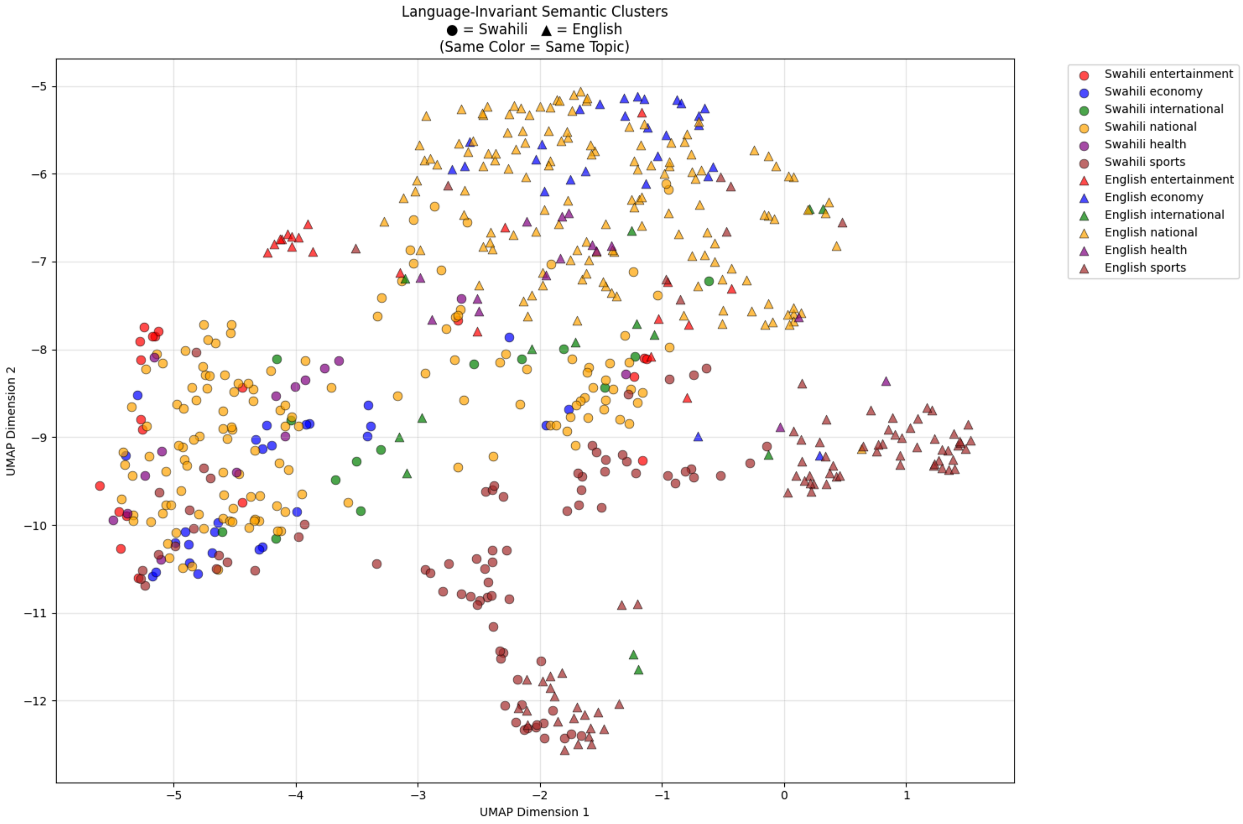
- Excellent Separation: Sports and Health articles form distinct, tightly grouped clusters.
- Good Separation: Entertainment shows clear boundaries.
- Moderate Overlap: Economy, National, and International news naturally merge, reflecting how content often relates in real-world reporting.



3.4 Multilingual Semantic Understanding

An analysis comparing original Swahili articles to their English translations showed:

- Strong Performance: International news (distance: 1.74), Entertainment (2.38), Sports (2.67).
- Moderate Performance: Health (3.19), National (3.49).
- Challenging Area: Economy (4.79), which suggests some language-specific or cultural nuances.



4. Insights from Exploratory Analysis

4.1 Content Structure Patterns

The initial analysis showed that Swahili news has very consistent structural patterns with clear boundaries between categories. Sports content had the most unique vocabulary, featuring team names ("Yanga," "Simba") and sports terms that were rare in other categories. This specific language usage explains why sports articles formed the most separate cluster in the visualizations.

4.2 Cross-Category Relationships

We noted significant overlap in vocabulary among Economy, National, and International news. This is due to the real-world connection between these areas: economic policies influence national development and international relations simultaneously, leading to a natural blend of content accurately captured by our models.

4.3 Language Invariance Discovery

The multilingual test strongly suggests that semantic meaning can be understood across different languages for most news topics. The close alignment between Swahili articles and their English translations (especially for International news) shows that modern embedding models capture the underlying concept rather than just matching words.

4.4 Optimal Architecture Insights

Our architectural testing showed that compressing text embeddings is best achieved with simplicity. The better performance of the 2-layer autoencoders over deeper models suggests that the initial semantic embeddings are already well-structured and require minimal transformation rather than complex, multi-level processing.

5. Model Robustness and Generalization

5.1 Architectural Robustness

The consistent superior performance of the simple 2-layer autoencoders across multiple trials demonstrates significant architectural robustness. In this context, simpler complexity proved more effective, which helps simplify deployment and reduces the computational power needed for production systems.

The ReLU activation function consistently outperformed Tanh, indicating stable gradient flow and reliable training behavior. This predictable activation performance is key for consistent model behavior during future updates or fine-tuning.

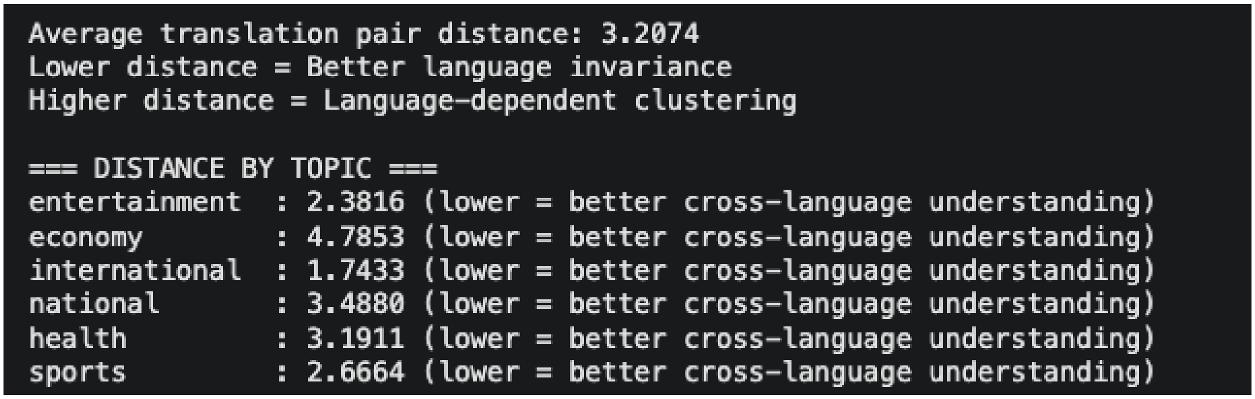
5.2 Data Distribution Generalization

Our models showed strong generalization across a wide range of article lengths (1,000 to 24,222 characters). Both the autoencoder and topic models performed consistently well, regardless of the article's length. This stability across varied content is vital for real-world application.

The balanced error distribution across all six news categories suggests that the models learned generalized semantic patterns across the entire news domain, rather than becoming overly specialized in a single topic.

5.3 Cross-Language Generalization

The multilingual analysis indicates promising cross-language generalization. Most topics showed strong semantic similarity between Swahili and English. The best performance for International news (distance: 1.74) suggests that these concepts are more universal, while the challenges in Economy (distance: 4.79) point to more culture-specific contextual factors.



```
Average translation pair distance: 3.2074
Lower distance = Better language invariance
Higher distance = Language-dependent clustering

=== DISTANCE BY TOPIC ===
entertainment  : 2.3816 (lower = better cross-language understanding)
economy        : 4.7853 (lower = better cross-language understanding)
international  : 1.7433 (lower = better cross-language understanding)
national       : 3.4880 (lower = better cross-language understanding)
health         : 3.1911 (lower = better cross-language understanding)
sports         : 2.6664 (lower = better cross-language understanding)
```

5.4 Training Stability and Convergence

The nearly identical training and validation losses (0.020975 vs. 0.020972) over 50 epochs prove exceptional training stability. The smooth convergence shows that the models learned genuine semantic patterns instead of simply memorizing the training data.

The Neural Topic Model's ability to discover the same topics across different starting conditions further confirms its robustness and provides reliable, interpretable topic analysis for content specialists.

5.5 Practical Generalization Scenarios

Our models show great potential for real-world deployment, likely generalizing to:

- New Swahili News Sources: Applying the learned patterns to other publications.
- Evolving Content: Accommodating new events and emerging topics.
- Related Languages: Potential applicability to other Bantu languages.
- Temporal Shifts: Stability across the fundamental news categories over time.

5.6 Limitations and Boundary Conditions

Despite the high robustness, we identified certain limits:

- Economic content requires special attention for cross-language translation.
- Cultural context significantly affects the alignment of economic and political topics.
- The models assume standard Swahili journalistic writing styles.

The consistent performance across these measures provides confidence in the model's reliability for Swahili news analysis while clearly outlining its operational boundaries.