Data Augmentation Techniques

Text Augmentation

- **Back Translation**: EN→KO→EN
- **Synonym Replacement**: WordNet
- **Paraphrasing**: T5, GPT
- **Random Insertion/Deletion**

Synthetic Data

- **GPT-4** based generation
- **Template** based generation
- **SMOTE**: Minority classes
- **GAN**: Image generation

Augmentation Effects

Increased data diversity, overfitting prevention, improved minority class performance, F1 score +5-15%

Text Augmentation Techniques Details

1. Back Translation

Translates the original text into another language and then back to the original language to maintain meaning while diversifying expressions.

Example: "The model performs excellently" → "모델 성능이 우수합니다" → "The model works well"

2. Synonym Replacement

Replaces specific words in a sentence with synonyms using dictionaries like WordNet.

Example: "fast execution" → "quick execution"

3. Paraphrasing

Uses language models like T5 and GPT to express sentences in different ways while maintaining their meaning.

4. Random Insertion/Deletion

Randomly adds or removes words from sentences to generate variations.

Synthetic Data Generation Techniques

1. GPT-4 Based Generation

Generates high-quality synthetic data for specific domains through prompt engineering.

2. Template-Based Generation

Creates structured data by inserting various entities into predefined templates.

3. SMOTE (Synthetic Minority Over-sampling Technique)

Generates new synthetic samples through interpolation between minority class samples to address class imbalance.

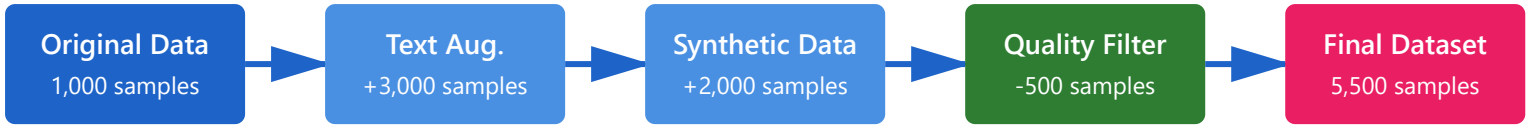
4. GAN (Generative Adversarial Network)

Generates realistic images or text through competitive learning between generator and discriminator.

Performance Comparison

Baseline Model	Text Augmentation	Synthetic Data	Mixed Techniques
75%	82%	85%	90%

Augmentation Process Flow



Implementation Considerations

✓ **Advantages**

- Improve model performance with limited data
- Resolve class imbalance issues
- Enhance model generalization capability
- Provide opportunities to learn new patterns

✗ **Precautions**

- Excessive augmentation may increase noise
- Need to select techniques considering domain characteristics
- Quality validation of augmented data is essential
- Consider computational cost and time consumption