Disaster Tweet Classification Using DistilBERT

A Machine Learning NLP Project

By Ayush Gupta
Navya Kondaveeti

Project Overview

01

Objective

Classify disaster-related tweets using deep learning.

02

Tools Used:

Python, PyTorch, DistilBERT, Transformers, TQDM, Pandas.

03

Outcome

Predict whether a tweet is related to a disaster or not.

Dataset Information

Source

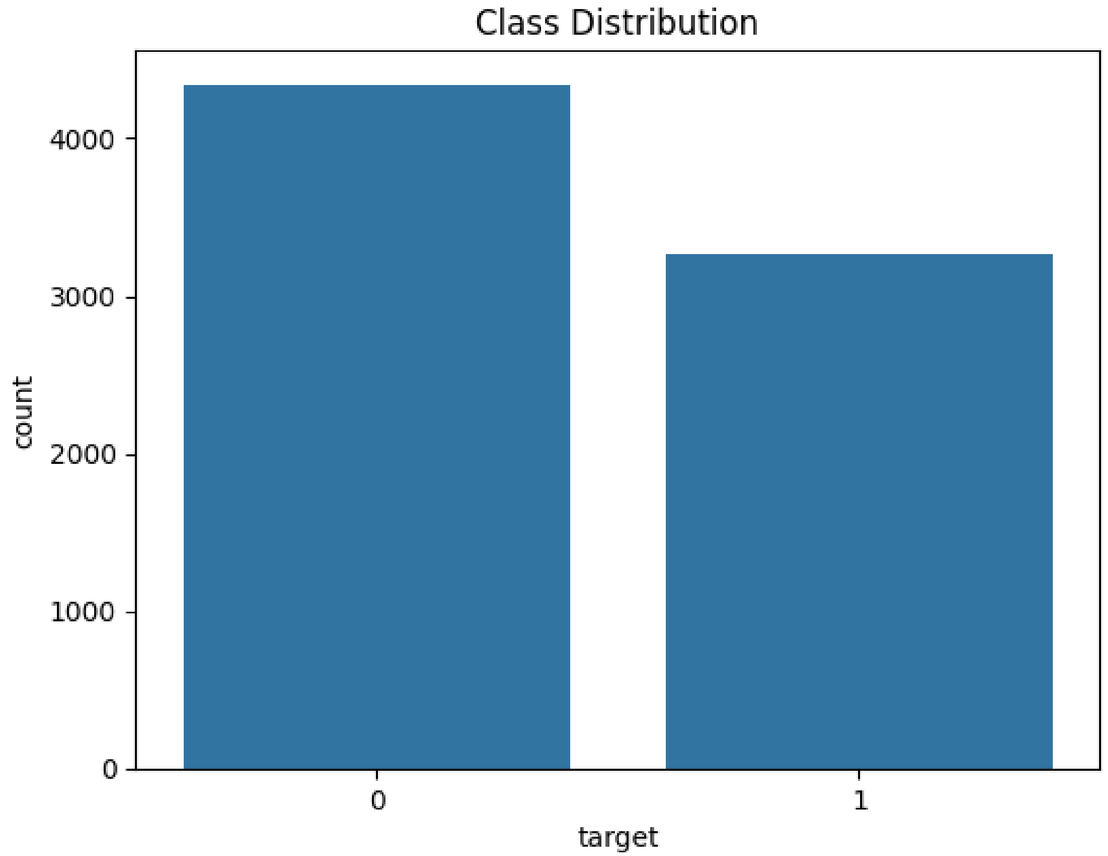
Kaggle Disaster Tweets Dataset

Classes

- **1 (Disaster):** Tweets about real disasters.
- **0 (Non-Disaster):** Other tweets.

Size

- **Train:** 7,613 tweets.
- **Test:** 3,263 tweets.



This chart shows the balance between disaster (1) and non-disaster (0) tweets.

Data Preprocessing

01

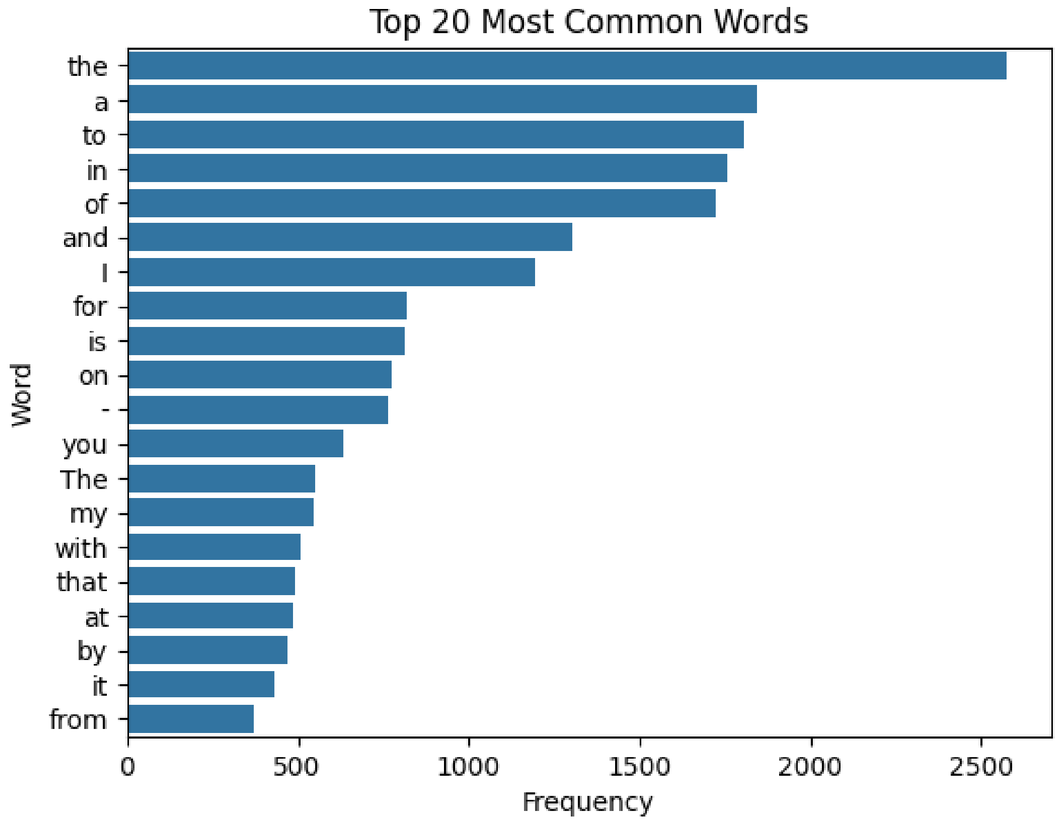
Convert text to lowercase.

02

Remove URLs, special characters, and extra spaces.

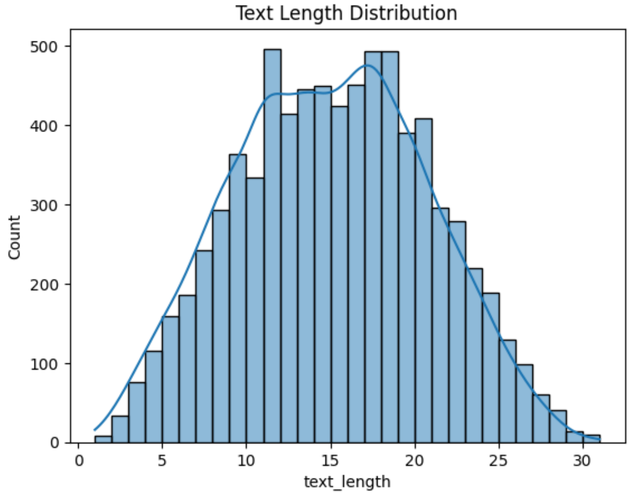
03

Use **DistilBERT tokenizer** to prepare data for the model.

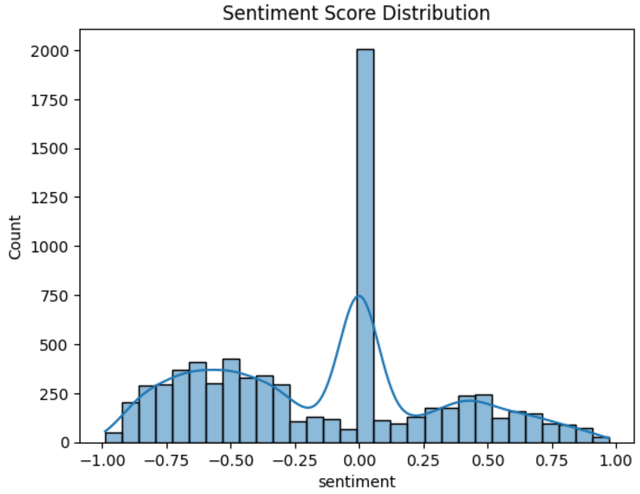


This Graph shows frequently used words in all tweets. Common words like *"the," "a," "to,"* and *"is"* are stopwords and have been be removed during preprocessing.

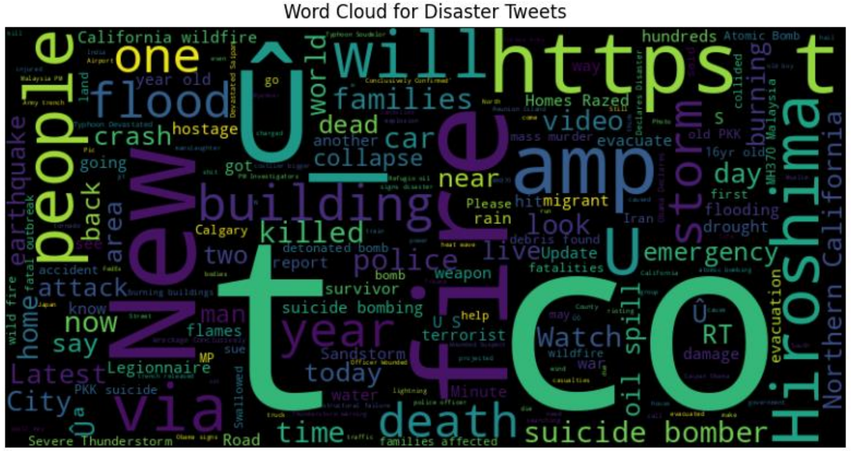
Data Preprocessing & EDA



This shows the number of words per tweet. Most tweets contain 10-20 words, so padding/truncation at 128 tokens is sufficient.



Shows the range of sentiment scores in tweets. Most tweets lean towards neutral sentiment, but disaster-related tweets tend to be negative.



This word cloud highlights the most frequent words in disaster tweets. Words like *"fire," "flood," "earthquake,"* and *"evacuation"* are common, supporting the model's learning.

Model Architecture

Base Model: *Pre-trained DistilBERT*.

01 Used a **classification head** on top of DistilBERT.

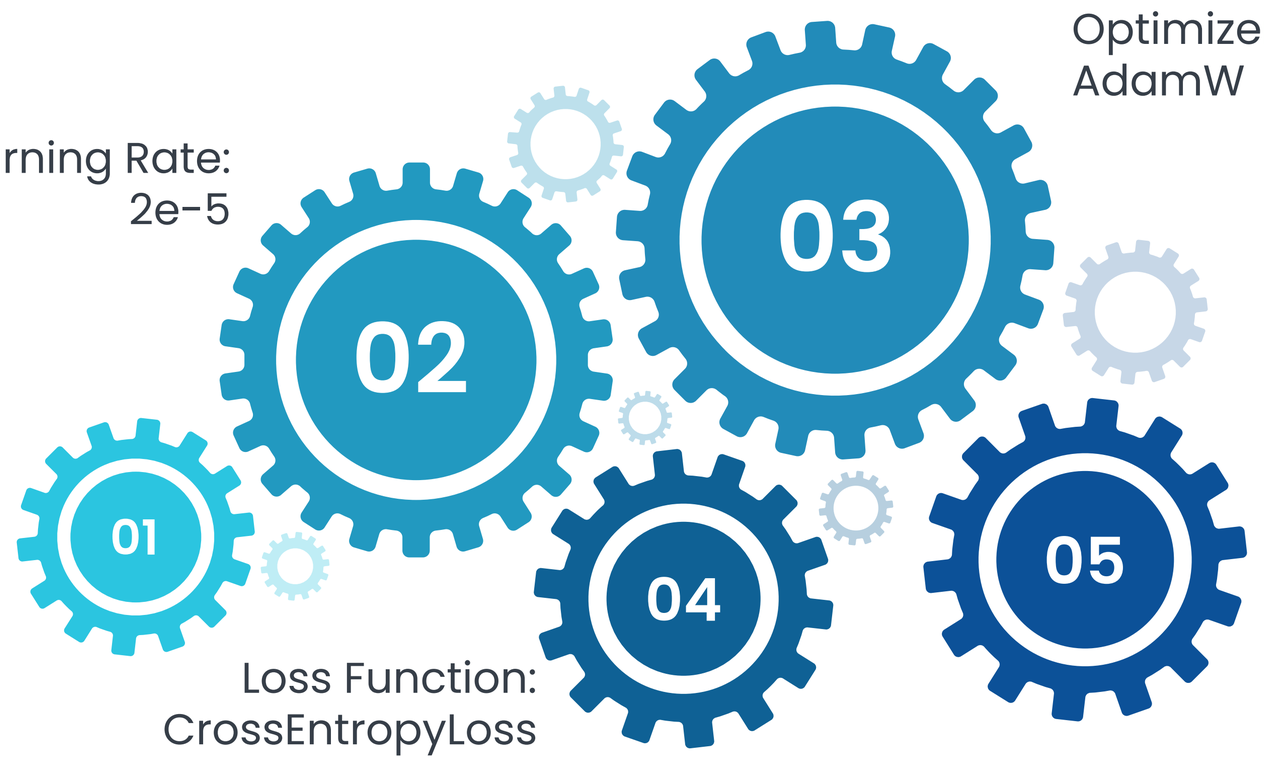
02 Input: Tokenized text.

03 Output: Binary classification (disaster or non-disaster).

```python
# Prepare DataLoaders
BATCH_SIZE = 32  # Increase batch size for faster training
train_dataset = TextDataset(X_train_texts, y_train)
train_loader = DataLoader(train_dataset, batch_size=BATCH_SIZE, shuffle=True)

test_dataset = TextDataset(X_test_texts)
test_loader = DataLoader(test_dataset, batch_size=BATCH_SIZE, shuffle=False)

# Load model with built-in classification head
model = DistilBertForSequenceClassification.from_pretrained("distilbert-base-uncased", num_labels=2).to(device)

# Freeze first few layers to speed up training
for param in model.distilbert.embeddings.parameters():
    param.requires_grad = False
for layer in model.distilbert.transformer.layer[:2]:  # Freeze first 2 layers
    for param in layer.parameters():
        param.requires_grad = False

# Define loss function and optimizer
criterion = nn.CrossEntropyLoss()
optimizer = optim.AdamW(model.parameters(), lr=2e-5)

# Learning rate scheduler
num_training_steps = len(train_loader) * 3  # 3 epochs
lr_scheduler = get_scheduler("linear", optimizer=optimizer, num_warmup_steps=0, num_training_steps=num_training_steps)

# Mixed precision training for faster execution
scaler = torch.cuda.amp.GradScaler()

# Training loop with optimizations
EPOCHS = 3
model.train()
```

Training Configuration

Hyperparameters and Epochs



Batch Size: 32

Learning Rate:
2e-5

Loss Function:
CrossEntropyLoss

Optimizer:
AdamW

Mixed Precision
Training for
efficiency.

Number of Epochs = 3

Model Optimization

Techniques Used

01

Freezing Layers

First 2 layers of DistilBERT are frozen to speed up training.

02

Gradient Scaling

Mixed precision training using AMP.

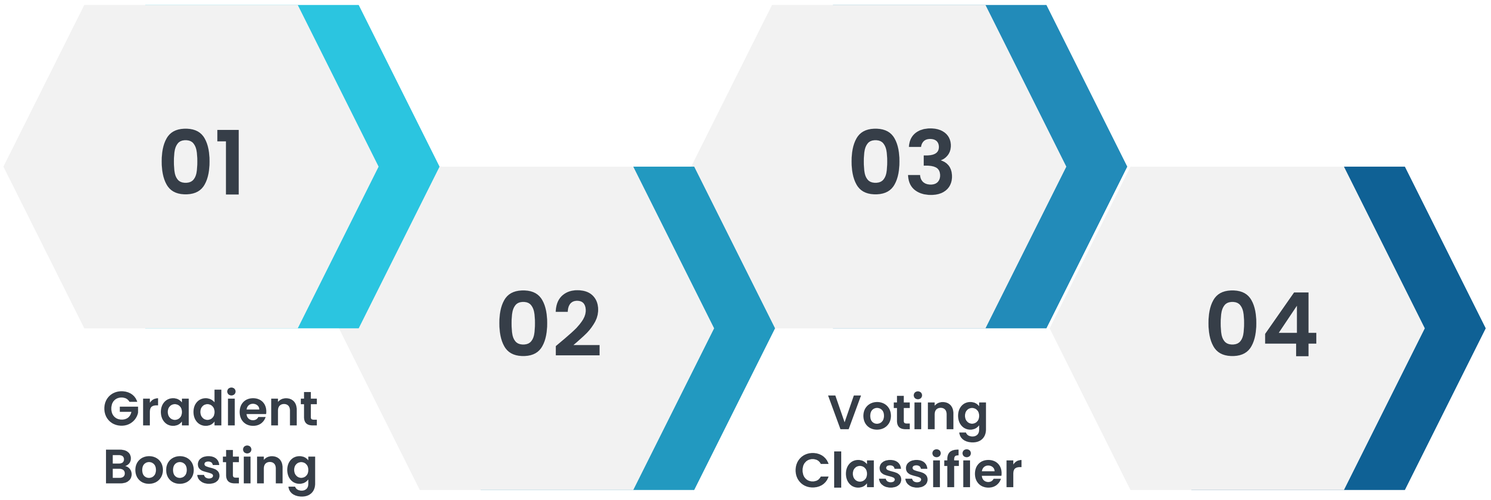
03

Learning Rate Scheduler

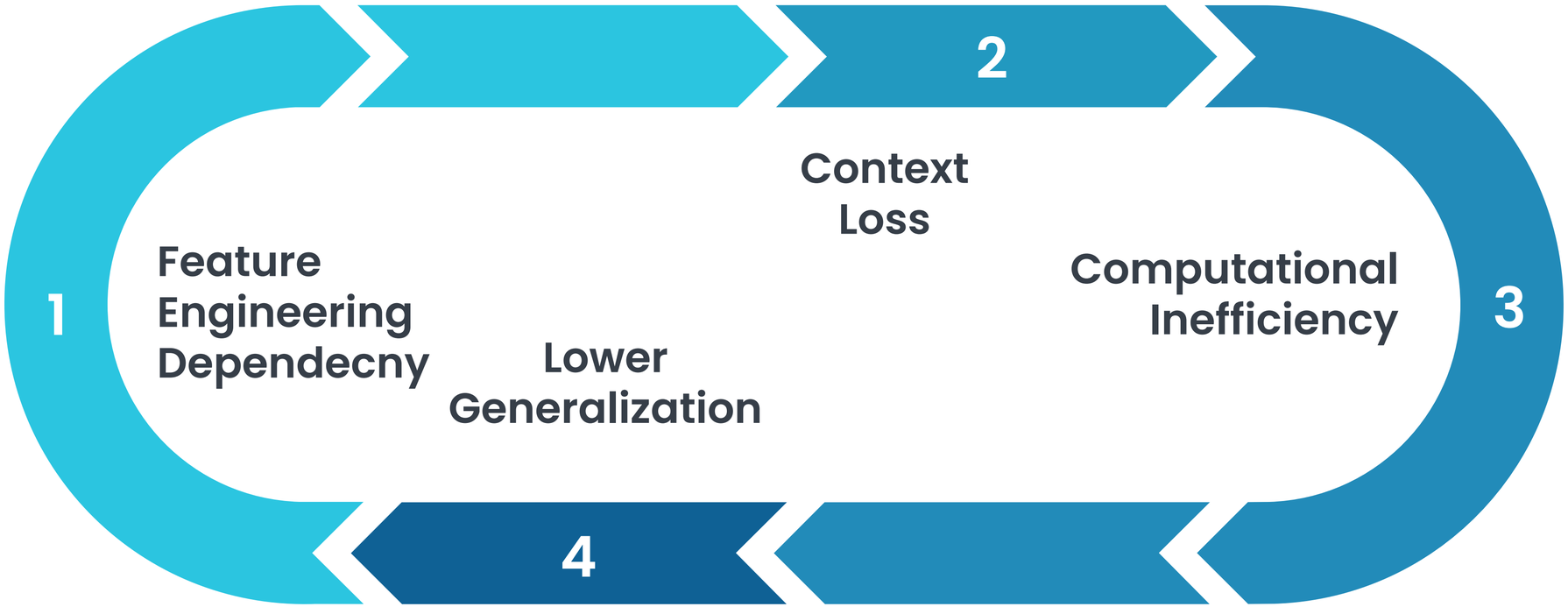
Linear decay for better generalization.

Approach 1(TF-IDF + Ensemble Model)

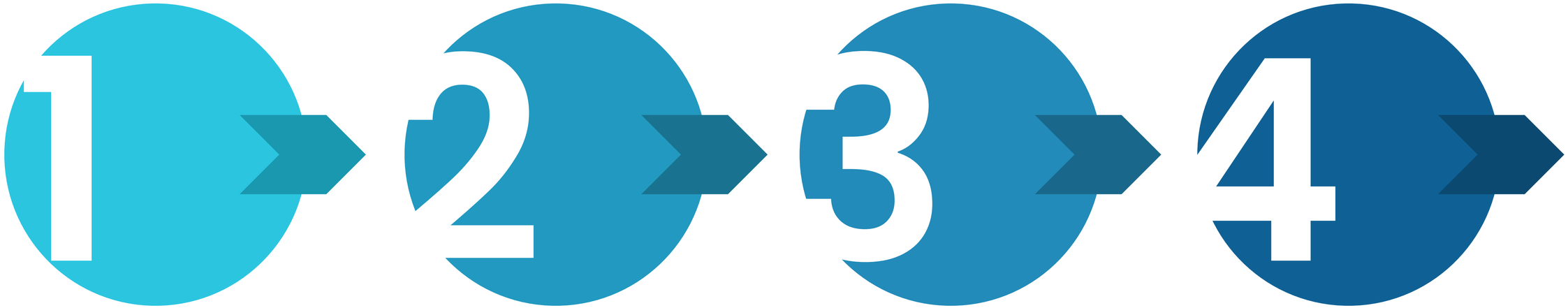
Text preprocessing includes stopword removal, lowercasing, and removing URLs, numbers, and punctuation, followed by TF-IDF vectorization to convert text into numerical features. An ensemble model (e.g., Random Forest or Gradient Boosting) is then trained on these features for classification.



Limitations Of Approach 1



DistilBERT better than Approach 1



The first approach (TF-IDF + ML models) provided a structured pipeline but lacked deep contextual understanding.

The final approach (DistilBERT fine-tuning) significantly improved performance by leveraging pre-trained NLP models.

Transformers are now the preferred choice for text classification due to their ability to learn contextual representations.

Better generalization, less feature engineering, optimized GPU usage, and higher Kaggle score.

Approach 2(Naïve Bayes)

Convert text into numerical features using TF-IDF bigrams and classify tweets with a Multinomial Naïve Bayes model.

Text Pre-Processing

Clean and preprocess the text data

Feature Extraction

Use TF-IDF vectorization to convert the

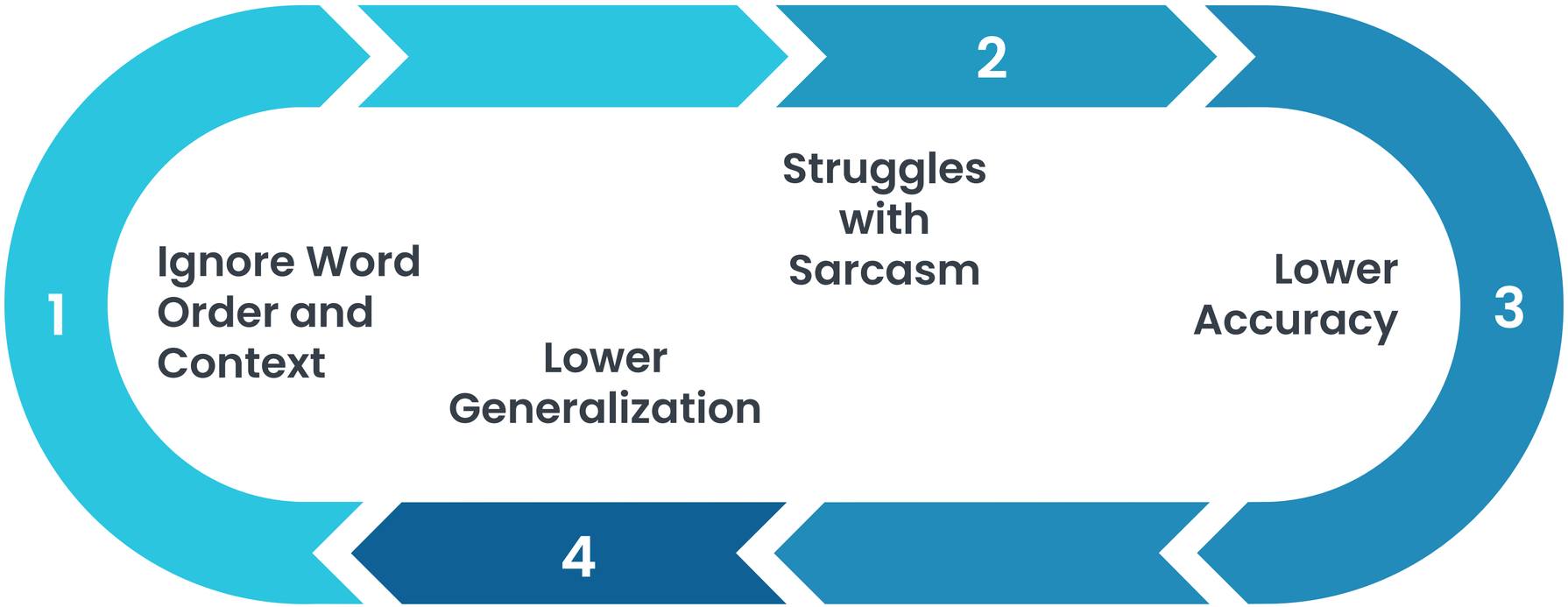
Model Training

Combined both models to improve prediction stability.

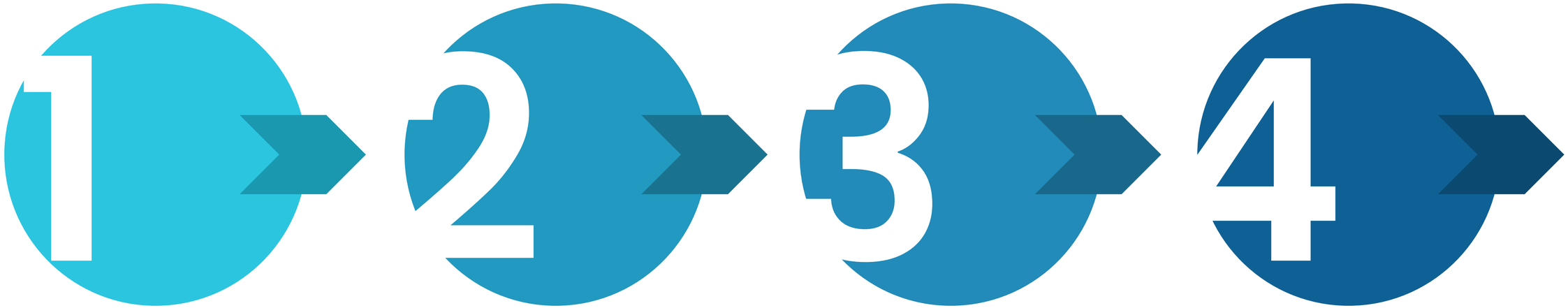
Prediction and Evaluation

Use the trained Naïve Bayes model to classify

Limitations Of Approach 2



DistilBERT better than Approach 2



While Naïve Bayes + TF-IDF is an interpretable and quick solution, it lacks context awareness.

DistilBERT effectively understands tweet semantics, making it the superior approach.

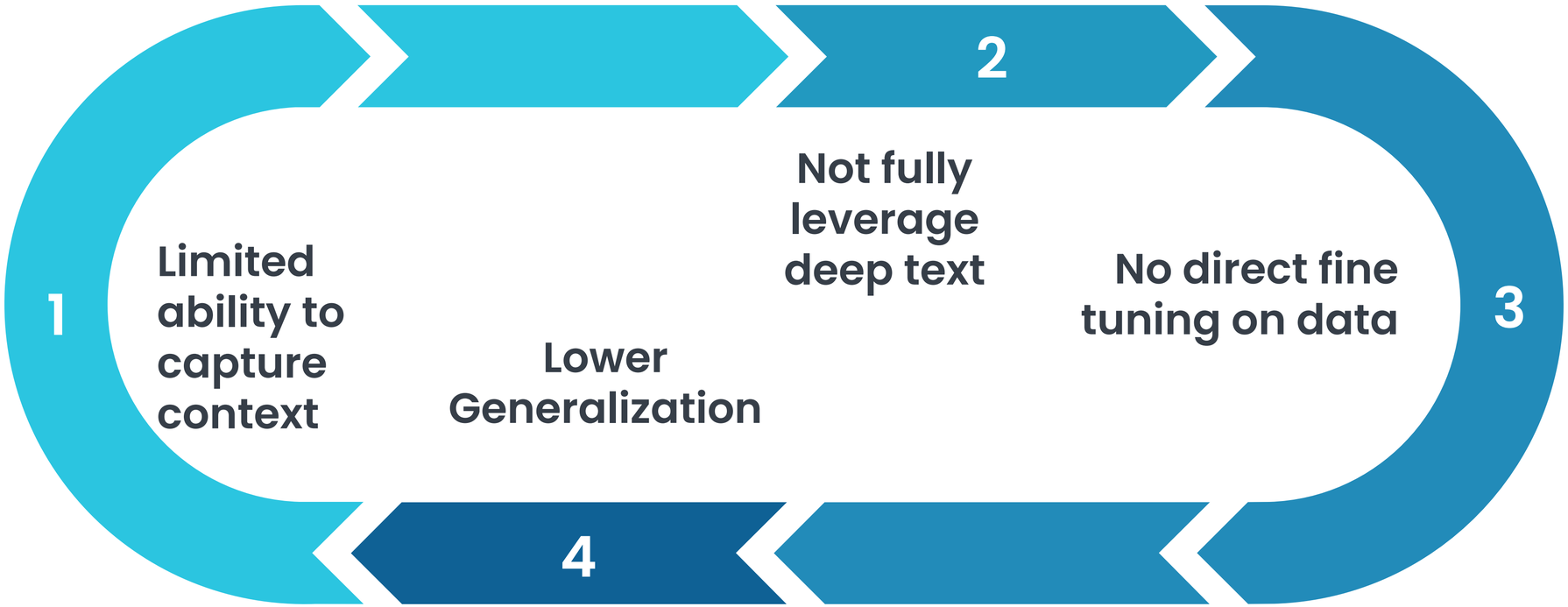
Pre-trained transformers significantly outperform traditional ML methods in NLP tasks.

Deep contextual understanding and pre-trained knowledge enable better generalization. Unlike TF-IDF, it avoids overfitting and treats words in context.

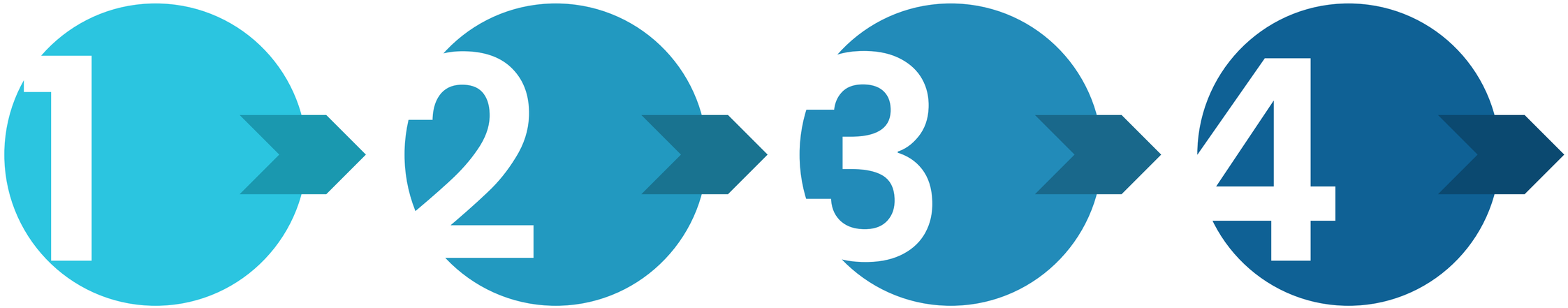
Approach 3(BERT with Logistic Regression)

Using only embeddings for training speeds up the process and reduces computational costs. The simpler model is also easier to interpret compared to full fine-tuning.

Limitations Of Approach 3



DistilBERT better than Approach 3



Fine-tuning **DistilBERT** provided a **significant boost in accuracy**

Precomputed BERT embeddings with logistic regression was **a good baseline** but lacked adaptability

Optimizing **training efficiency** (e.g., freezing layers, mixed precision) helped balance cost vs performance

End-to-end fine-tuning on the disaster tweet dataset captures language nuances and improves performance, resulting in higher accuracy and a better Kaggle leaderboard ranking.

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.

Initiation
Write your paragraph here

Planning
Write your paragraph here

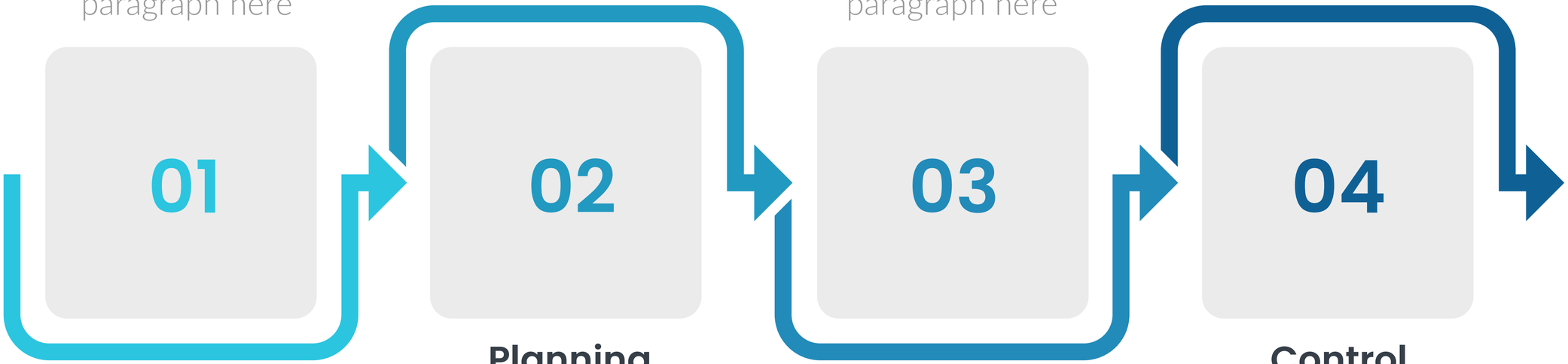
Execution
Write your paragraph here

Control
Write your paragraph here

Closure
Write your paragraph here

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.



Initiation
Write your paragraph here

Planning
Write your paragraph here

Execution
Write your paragraph here

Control
Write your paragraph here

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.

Initiation
Write your paragraph here

Planning
Write your paragraph here

Execution
Write your paragraph here

Control
Write your paragraph here

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.

Initiation
Write your paragraph here

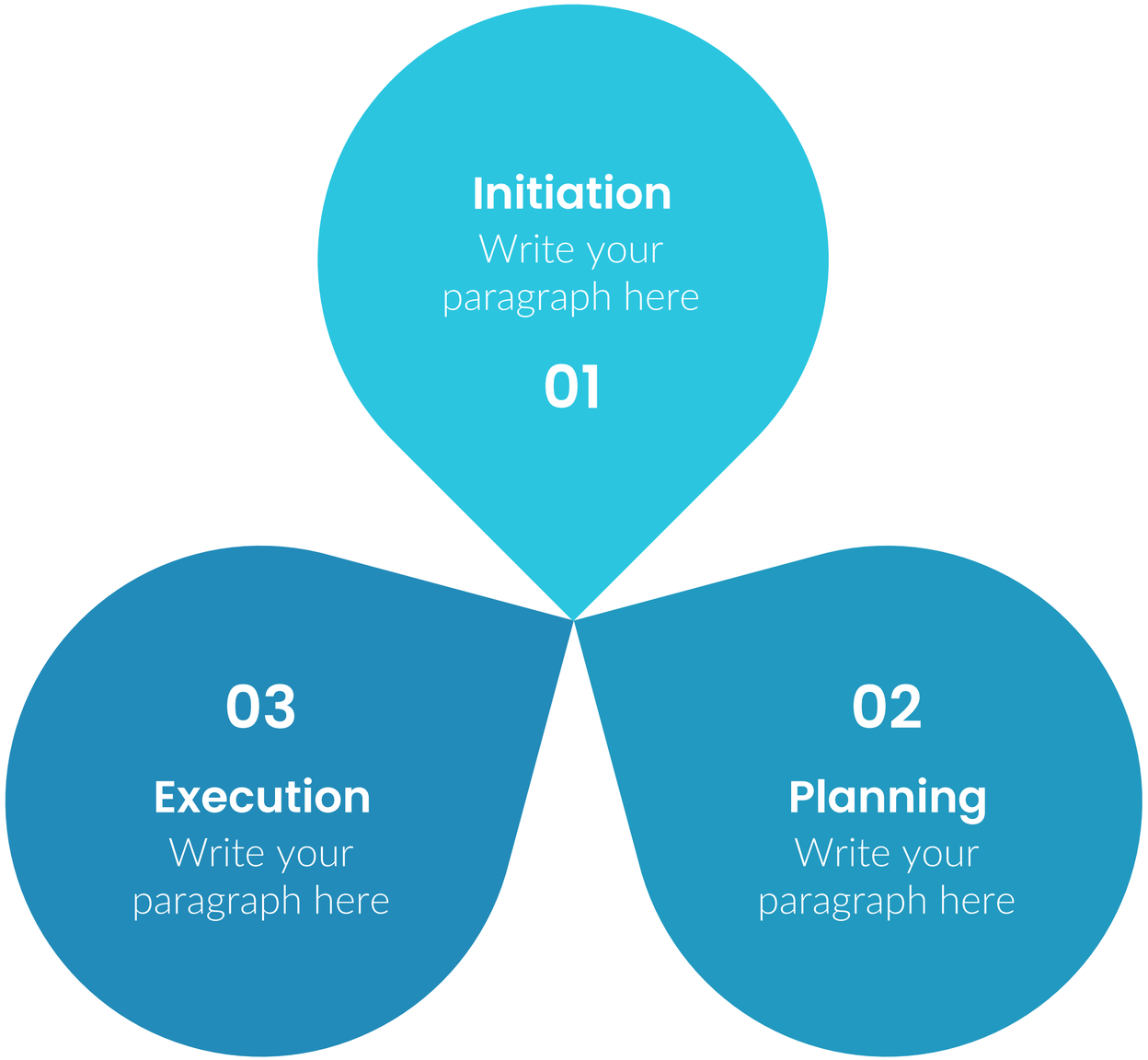
Planning
Write your paragraph here

Execution
Write your paragraph here

Control
Write your paragraph here

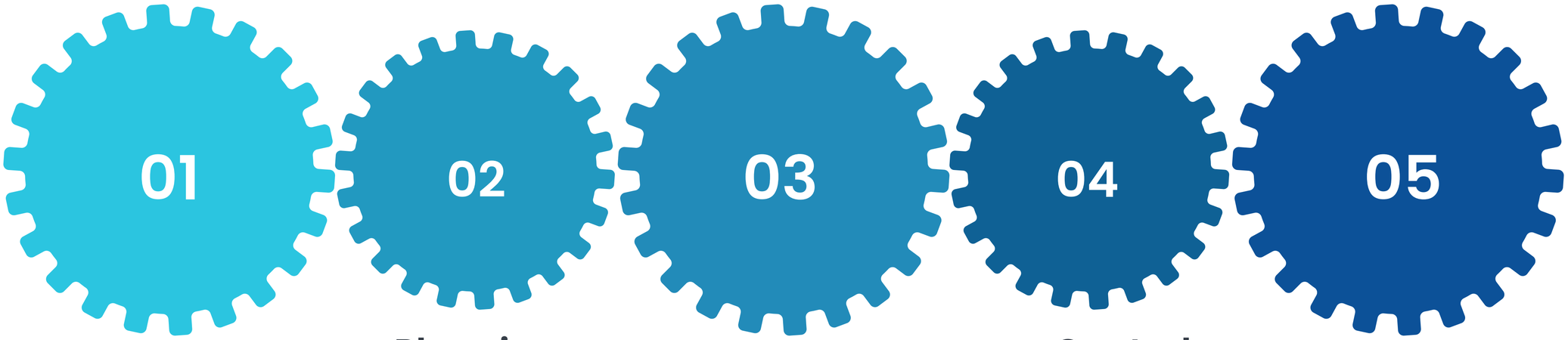
Process Infographics

Marketing is the study and management of exchange relationships.



Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.



Initiation
Write your paragraph here

Planning
Write your paragraph here

Execution
Write your paragraph here

Control
Write your paragraph here

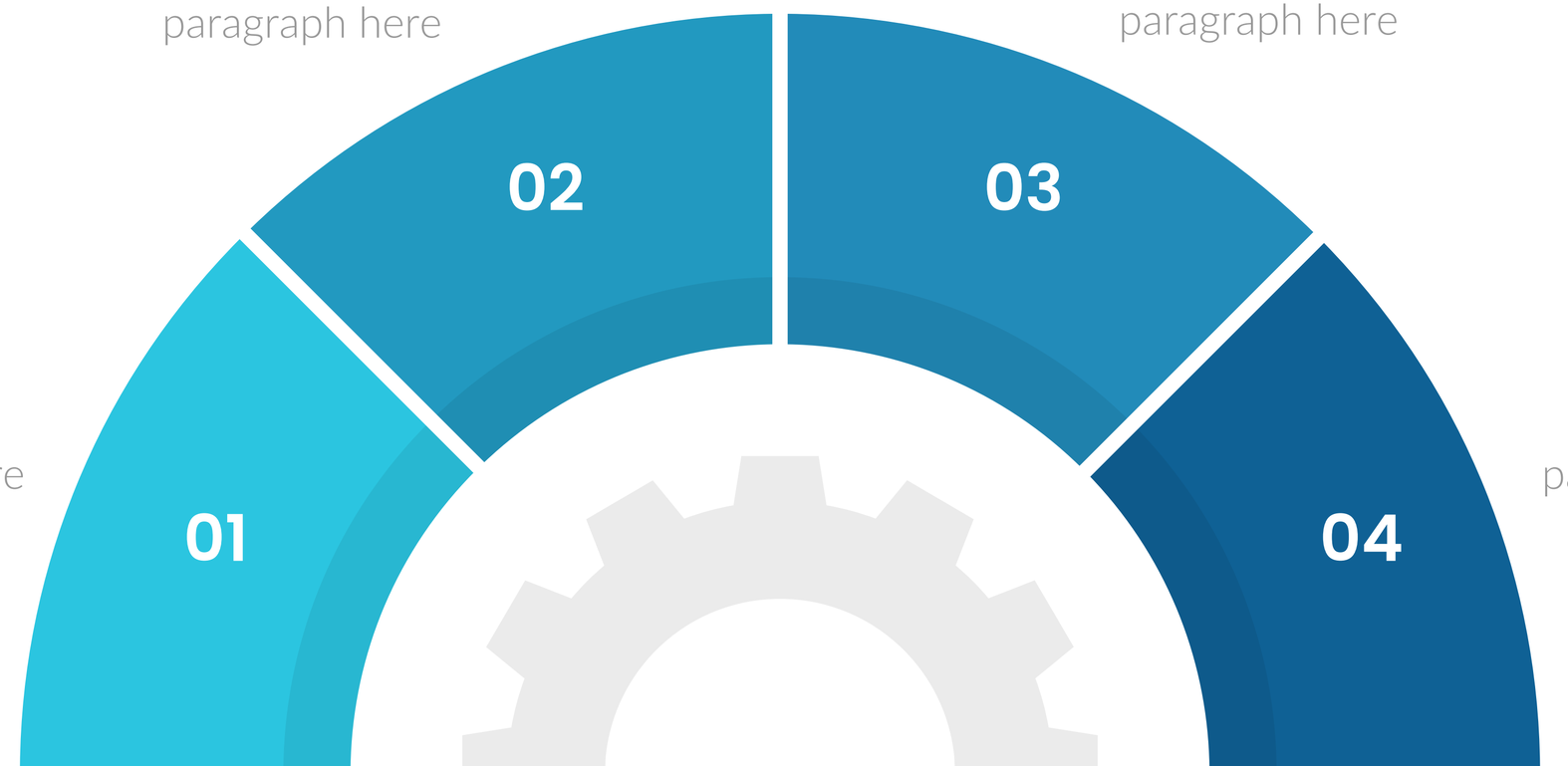
Closure
Write your paragraph here

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.

Planning
Write your paragraph here

Execution
Write your paragraph here

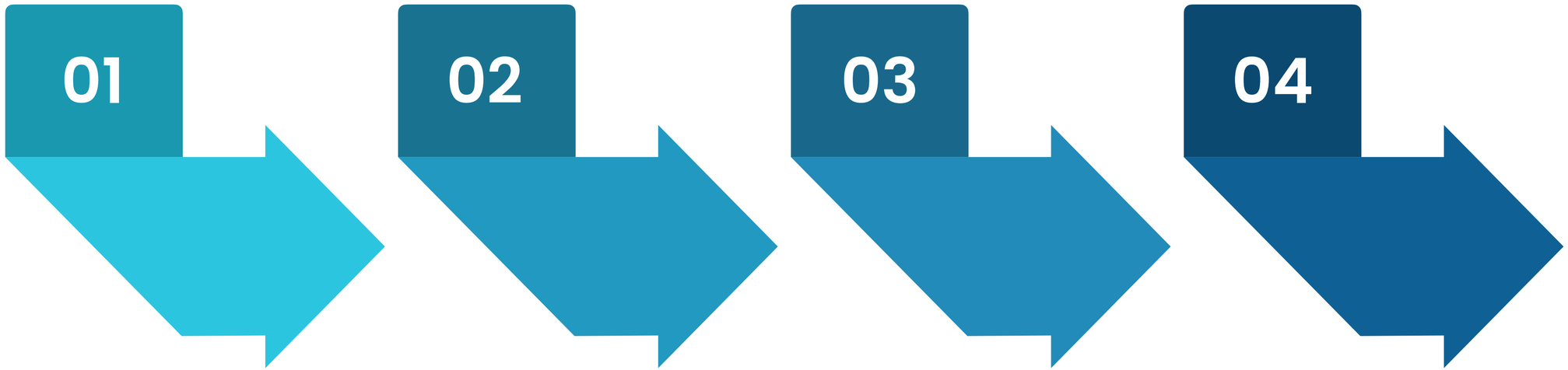


Initiation
Write your paragraph here

Control
Write your paragraph here

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.



Initiation
Write your paragraph here

Planning
Write your paragraph here

Execution
Write your paragraph here

Control
Write your paragraph here

Process Infographics

Marketing is the study and management of exchange relationships. Marketing is the business process of creating relationships with and satisfying customers.

Initiation
Write your paragraph here

Planning
Write your paragraph here

Execution
Write your paragraph here

Control
Write your paragraph here