AI 한국어 패치하기

KUBIG
DATA SCIENCE & AI

Team | forAIgner
21기 강서연, 20기 윤시호, 20기 이우진

KUBIG
DATA SCIENCE & AI

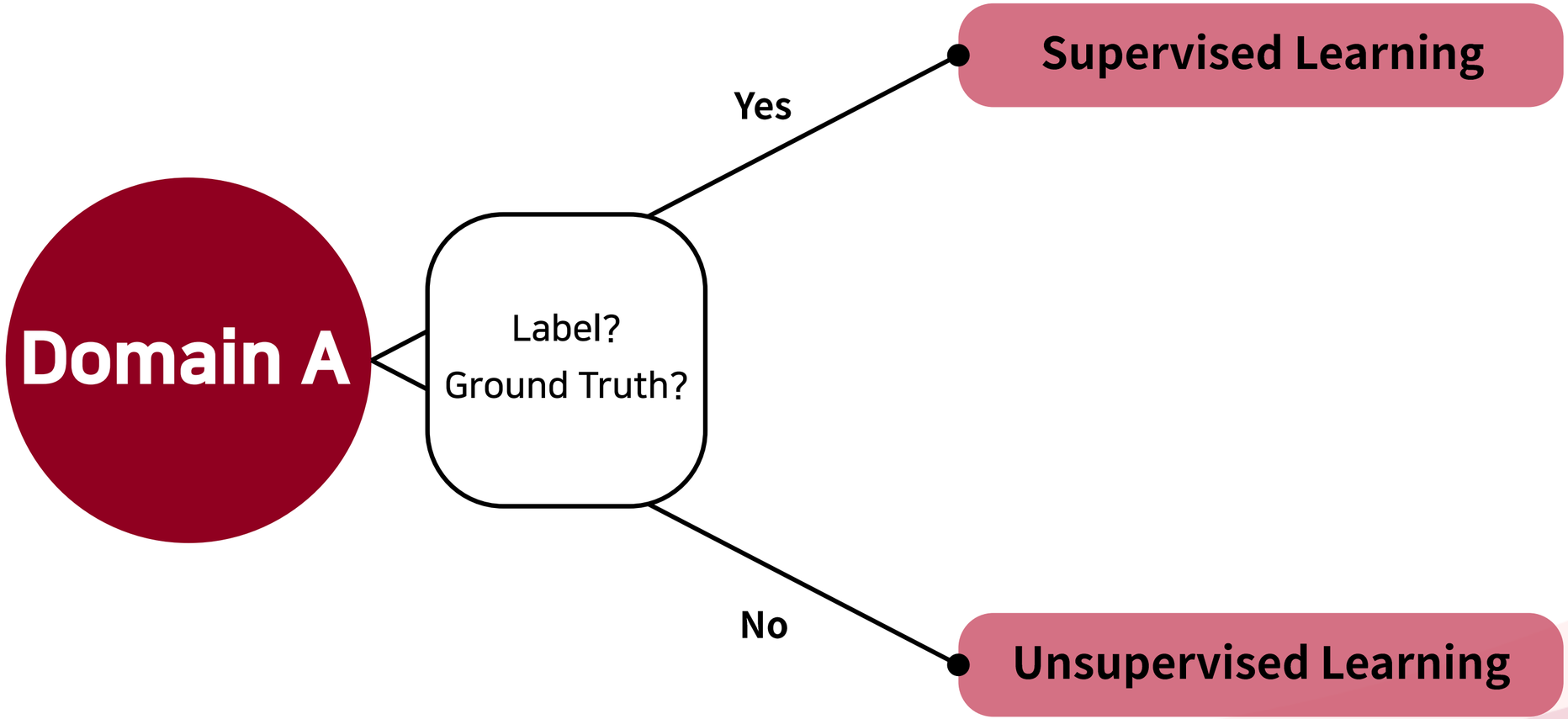
CONTENTS

KUBIG
DATA SCIENCE & AI

01. 문제 정의 및 해결 방안

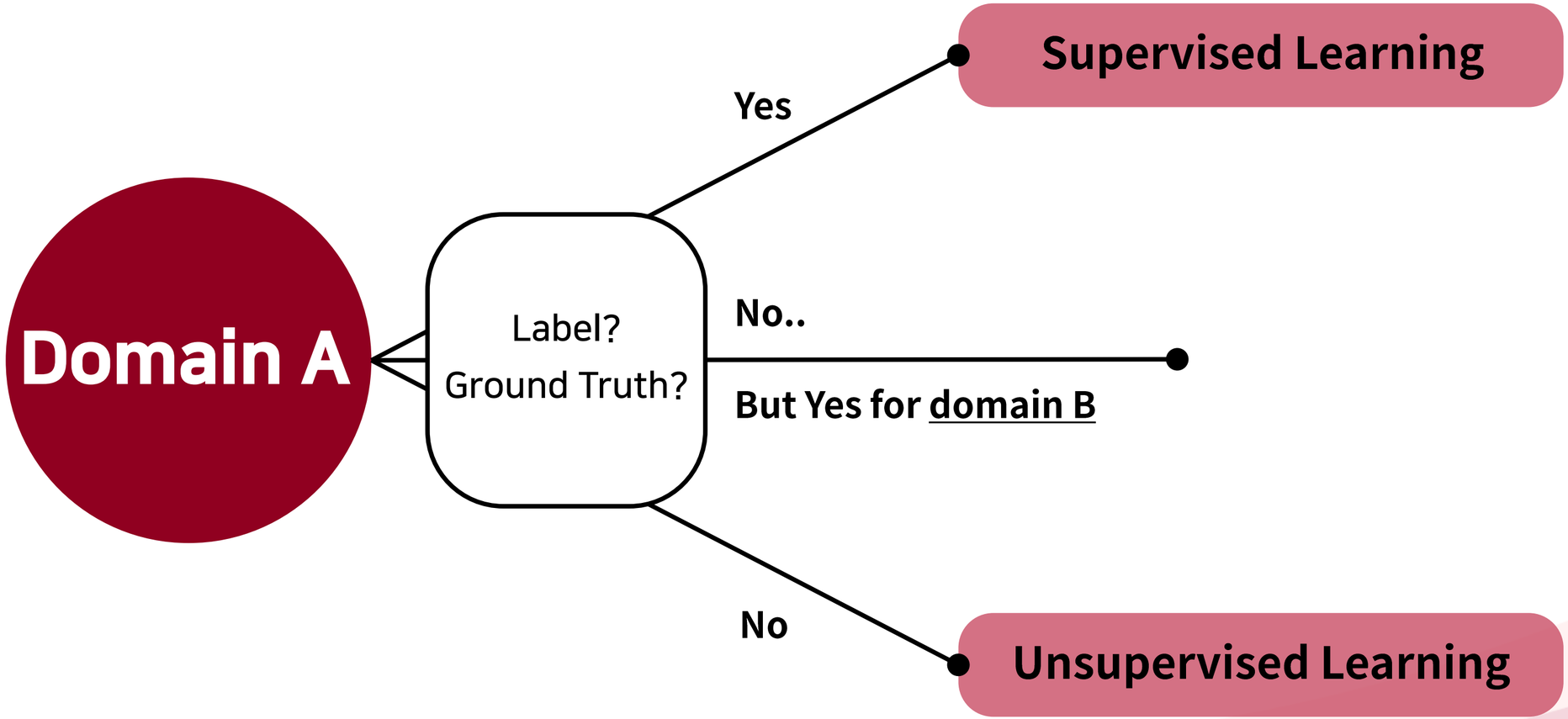
01. 문제 정의 및 해결 방안

Problem



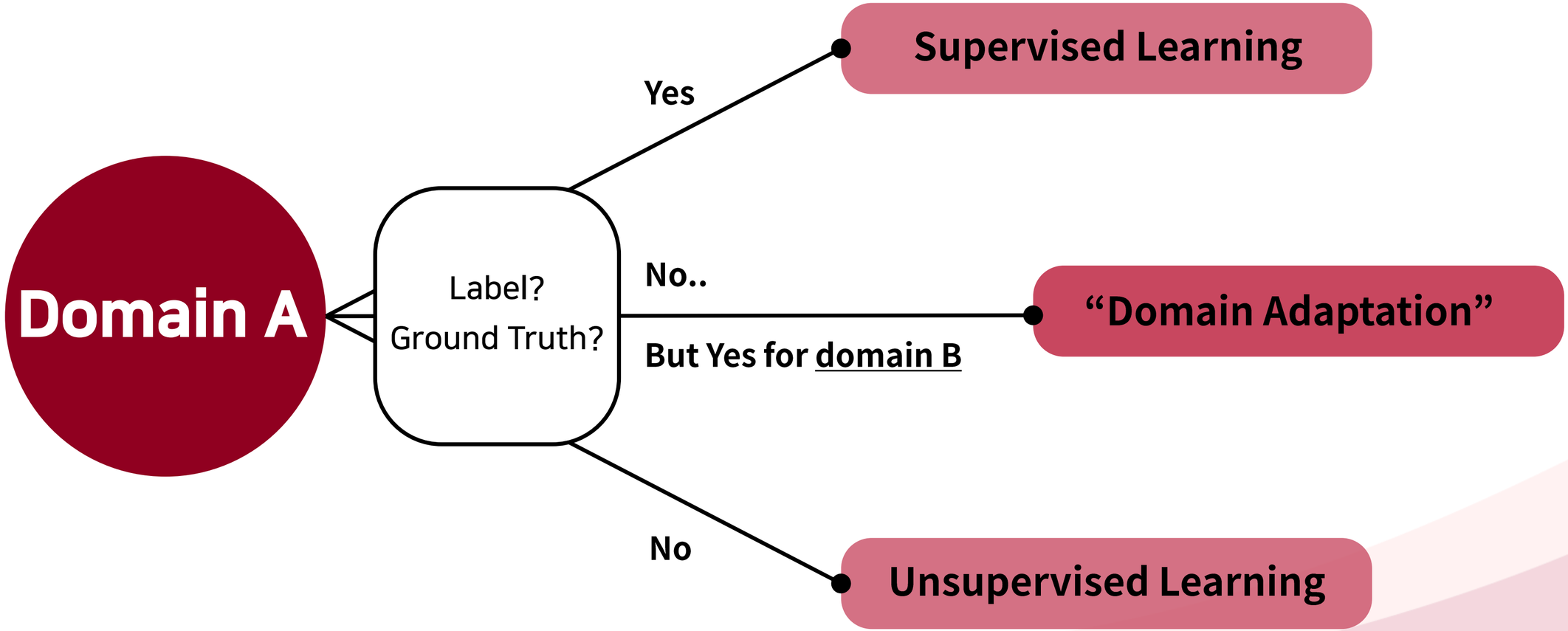
01. 문제 정의 및 해결 방안

Problem



01. 문제 정의 및 해결 방안

Problem



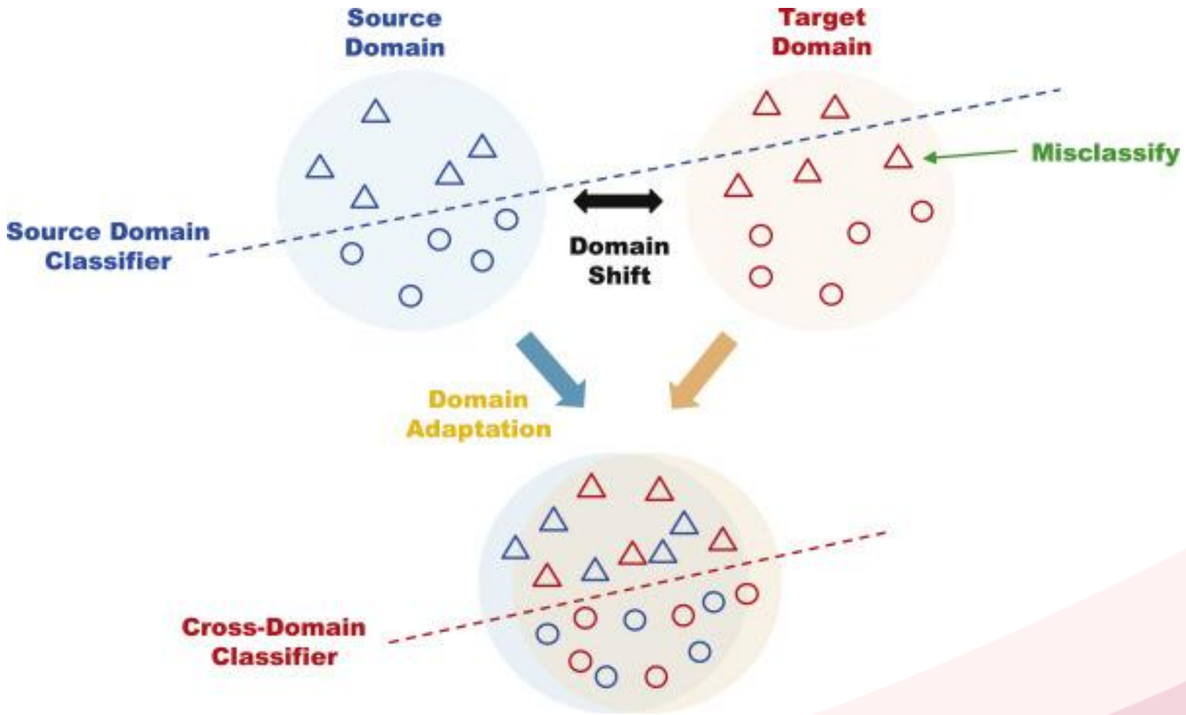
01. 문제 정의 및 해결 방안

Domain Adaptation

- **정형 데이터**
  - (Source) A 생산 라인의 품질 예측 모델을
  - (Target) B 생산 라인에 Domain Adapatation

- **이미지 데이터**
  - (Source) 실제 사진에 대한 segmentation 모델을
  - (Target) 그림 이미지에 Domain Adaptation

- **자연어 데이터**
  - (Source) 뉴스 기사 요약 모델을
  - (Target) 법률 문서에 Domain Adaptation



01. 문제 정의 및 해결 방안

Project

- **목표**

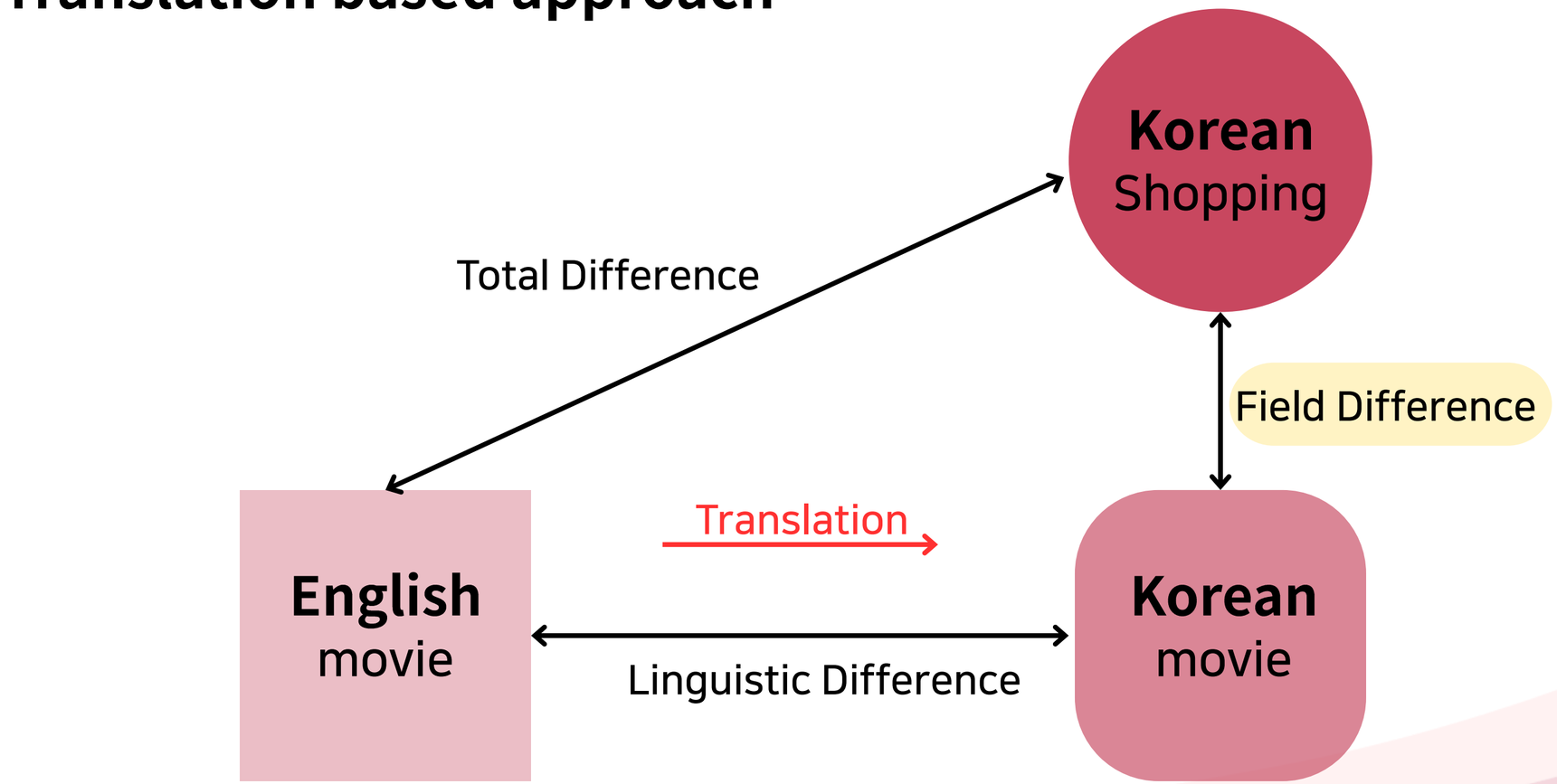
  네이버 쇼핑 리뷰 텍스트 긍정 / 부정 감성 이진 분류

- **데이터**
  - 영어 텍스트 데이터(LabeId): IMDb, SST, Dynasent
  - 한국어 텍스트 데이터(Unlabeld): 네이버 쇼핑 리뷰 텍스트 데이터
- **제약 사항**
  - 0.5B 규모 이상의 고성능 사전 학습 모델 사용 금지

KUBIG
DATA SCIENCE & AI

02. Translation-based Approach

02. Translation-based Approach



02. Translation-based Approach

Learning with MMD loss

- **MMD(Maximum Mean Discrepancy)**
  - 두 확률분포의 거리를 나타내는 지표
  - Loss로 사용 시, 두 데이터의 latent space를 정렬시킬 수 있음

$$\mathrm{MMD}^2(P,Q) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i,x_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j),$$

$$\text{where } x_i \sim P,\ y_j \sim Q$$

  - 여기서 커널은 Gaussian 커널을 사용

02. Translation-based Approach

Learning with MMD loss

- **Objective**
  - Source(한국어 번역, IMDb)와 Target(한국어, 네이버 쇼핑)의 embedding representation을 정렬
- **Embedding** Model
  - klue/bert-base
- **Data**
  - 한국어 번역 데이터(Labeled): Translated IMDb(20K)
  - 한국어 데이터(Unlabeled): 네이버 쇼핑 리뷰(20K)
- **Training**
  - Embedding + MLP Classifier
  - $Total\ Loss = \mathcal{L}_{BCE} + \beta\ \mathcal{L}_{MMD}$

02. Translation-based Approach

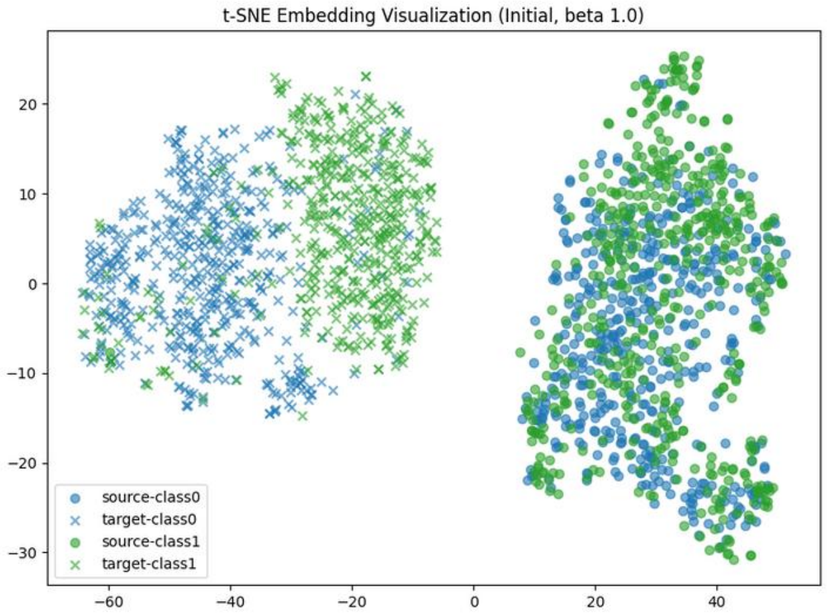
Result

- **Multi lingual direct (linguistic + field diff.)**
  - Multi lingual embedding(Fix) + MLP Classifier
  - 영어 데이터로 학습
- **Translation direct (field diff.)**
  - Korean embedding + MLP Classifier
  - 번역 데이터로 학습
- **Rich-resource case (Ideal)**
  - Korean embedding + MLP Classifier
  - 네이버 쇼핑 데이터(Labeled, 20K)로 학습
- **Low-resource case**
  - Korean Embedding + MLP Classifier
  - 네이버 쇼핑 데이터(Labeled, 2K)로 학습

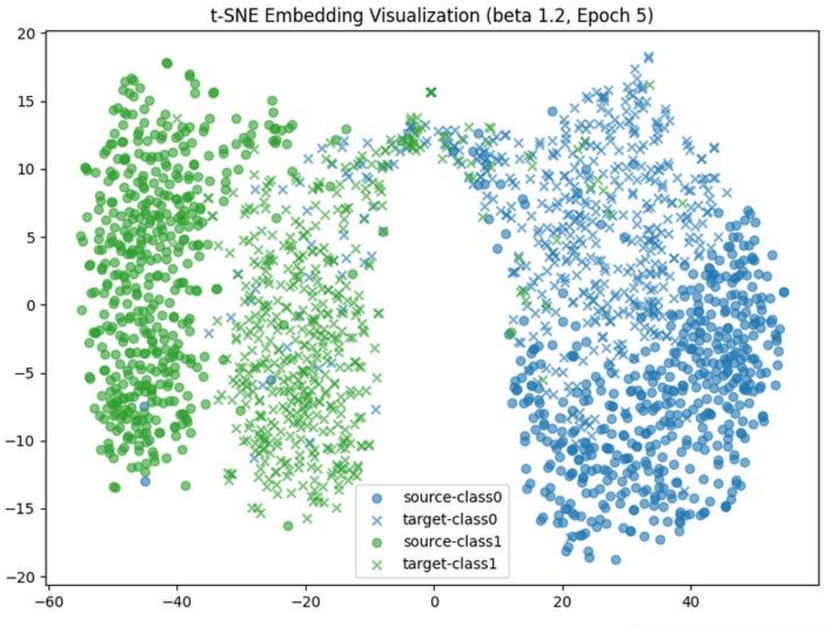
	Accuracy	F1-Score	AUC-ROC
Multi lingual direct	0.7948	0.8147	0.8949
Translation direct	0.8409	0.8393	0.9368
Rich-resource case	**0.9316**	**0.9327**	**0.9771**
Low-resource case	0.9240	0.9239	0.9684
With MMD	0.9078	0.9078	0.9613

02. Translation-based Approach

Result



학습 전



학습 후

KUBIG
DATA SCIENCE & AI

03. Translation-free Approach

03. Translation-free Approach

2-stage pipeline

1) DANN
(Domain-Adversarial
Neural Network)

- Domain invariant feature 학습
- 소스 도메인과 타겟 도메인의 분포를 특징 공간 상에서 정렬
- 도메인 간의 격차를 1차적으로 해소

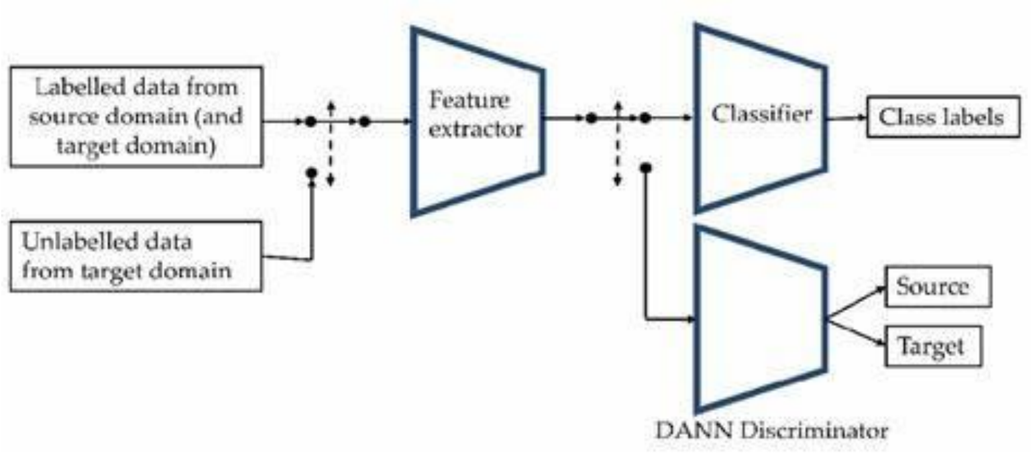
2) Free-Match

- 타겟 도메인 특화 성능 최적화
- 사전에 정렬된 모델을 기반으로 레이블 없는 타겟 데이터를 활용 (Semi-Supervised Learning; SSL)

03. Translation-free Approach

DANN

Adversarial Learning을 통해 Domain invariant feature를 학습



- **Feature Extractor (*xlm-roberta-base*):** 입력을 저차원 feature 벡터로 매핑
- **Sentiment Classifier:** feature 벡터를 기반으로 감성 분석 수행
- **Domain Discriminator:** feature 벡터가 소스 도메인에서 왔는지, 타겟 도메인에서 왔는지 구별

03. Translation-free Approach

DANN

Sentiment Classifier

Source 데이터의 레이블을
정확히 예측하도록 학습

VS

Domain Discriminator

Source와 Target을
최대한 잘 구별하도록 학습

$$Total\ Loss = \mathcal{L}_{sentiment} + \lambda\, \mathcal{L}_{domain}$$

03. Translation-free Approach

DANN

감정 분류기 + 도메인 판별기 혼란이라는 두 가지 목표를 동시에 학습

(gradient reversal layer 사용)

- 도메인 구분이 불가능한 feature 생성
- 도메인 정보가 제거된, 순수하게 분류 작업에만 유용한 feature를 생성

$$Total\ Loss = \mathcal{L}_{sentiment} + \lambda\, \mathcal{L}_{domain}$$

03. Translation-free Approach

Free - Match

DANN은 소스 도메인에서 학습된 지식을 정보가 없는 타겟 도메인으로
성공적으로 adaptation할 수 있으나,
도메인 정렬이 반드시 타겟 도메인에서의 최고 성능으로 직결되지는 않음

타겟 도메인 데이터 자체의 구조를 활용해
성능을 한 단계 더 끌어올리기 위한 **Free-Match** 도입!

03. Translation-free Approach

Free - Match

Key Principle

Semi-Supervised Learning(SSL)에서 자기 적응 임계값 (SAT) 활용

- **Semi-Supervised Learning(SSL)**

  소수의 레이블 데이터와 다수의 레이블 없는 데이터를 함께 사용해 모델을 학습하는 방법

- **Pseudo-Labeling**

  레이블 없는 데이터에 대해서, 모델 스스로 생성한 예측값을 임시 정답(Pseudo-Label)으로 사용

03. Translation-free Approach

Free - Match

Fix-Match

Fixed Threshold

모델의 예측 신뢰도가 사전에 정의된 값(예: 0.95)을
넘을 때만 유사 레이블로 인정하고 학습에 사용

Free-Match

Self Adaptation Threshold

threshold값을 고정하지 않고,
모델의 학습 상태에 따라 매 순간 동적으로 조절

03. Translation-free Approach

Free - Match

- Global Confidence

  모델의 전반적인 예측 confidence.

  학습이 진행될수록 threshold의 기본값을 상향 조정

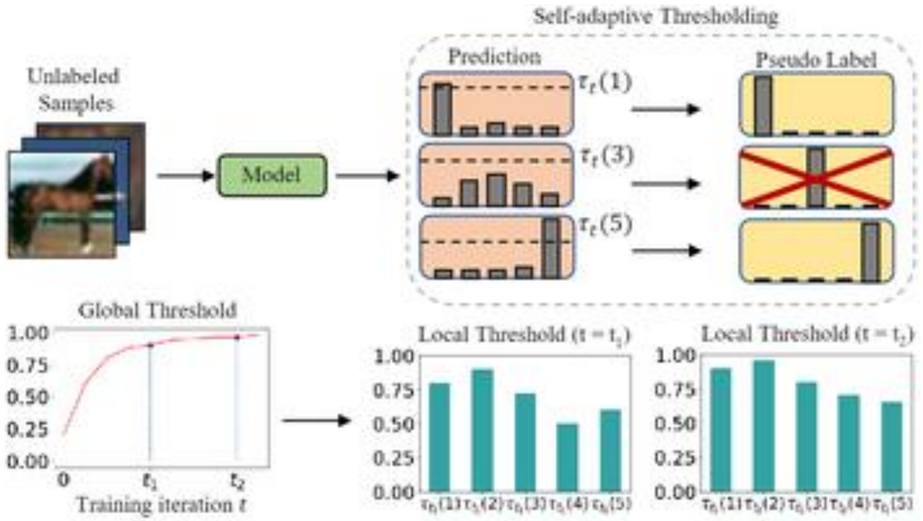
- Local Confidence

  클래스별 예측 분포.

  모델이 어려워하는 클래스의 threshold는 낮추고,

  자신있는 클래스의 threshold는 높여

  학습 기회를 공평하게 부여



03. Translation-free Approach

Free - Match

Exponential Moving Avaerage (EMA)

- SAT를 안정적으로 계산하기 위해 과거 추세 반영
- 개별 EMA 값 계산 방식

$$S_t = \alpha \cdot S_{t-1} + (1-\alpha) \cdot V_t$$

→ α를 통해 과거 정보를 얼마나 반영할 것인지 조절

- global confidence와 local confidence 각각에 대해 EMA 적용

→ Threshold가 개별 batch에 의존하는 대신, 장기적인 추세에 따라 안정적으로 변화

03. Translation-free Approach

Free - Match

Algorithm

1. Pseudo Labeling

   레이블 없는 타겟 데이터에 대해 prediction → 가장 높은 확률의 클래스를 pseudo label로 선정

2. SAT Calculating

   global confidence와 local confidence를 업데이트하여, 클래스별 동적 threshold 계산

3. Masking

   각 데이터의 예측 신뢰도가 동적 threshold를 넘을 경우에만 학습에 사용

4. Backpropagation

   $$Total\ Loss = \mathcal{L}_{SL} + \lambda\, \mathcal{L}_{SSL}$$

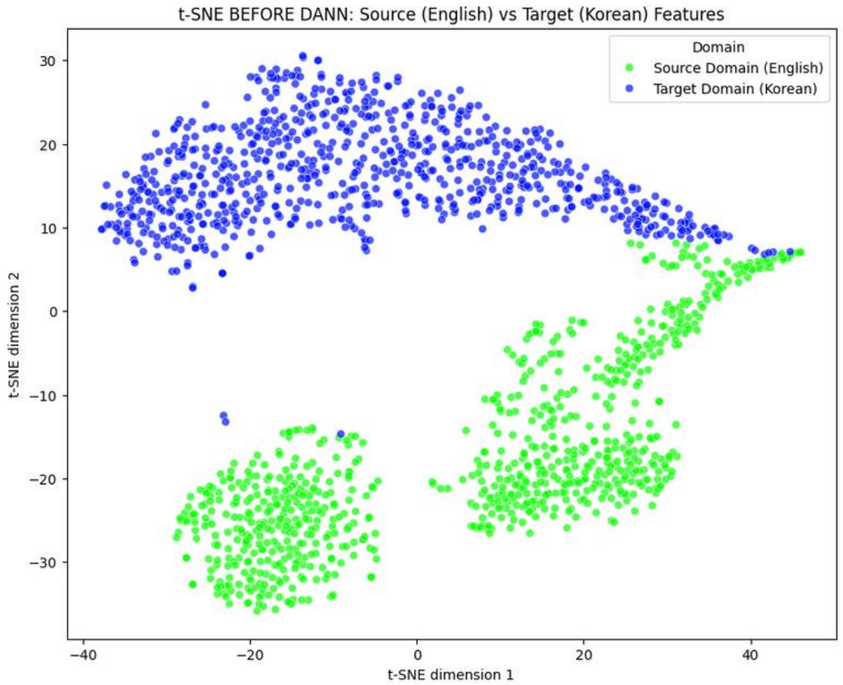
03. Translation-free Approach

Results

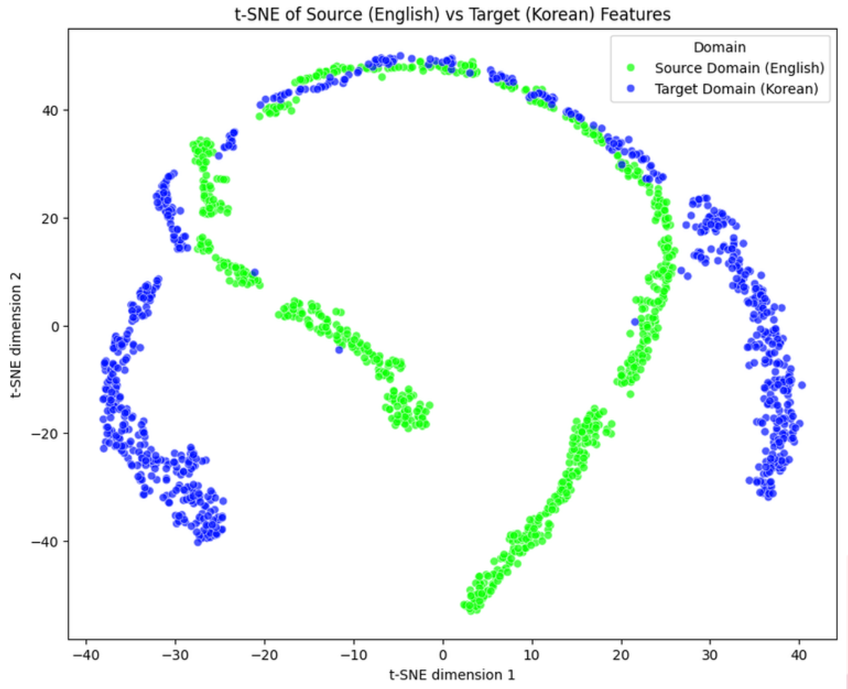
	Accuracy	F1-score	AUC-ROC
Multi lingual direct	0.7948	0.8147	0.8949
Rich-resource case (한국어 모델)	0.9316	0.9327	0.9771
DANN	0.8648	0.8644	0.9354
DANN + Free-Match	0.9064	0.9064	0.9451
With MMD	0.9078	0.9078	0.9613

03. Translation-free Approach

Results



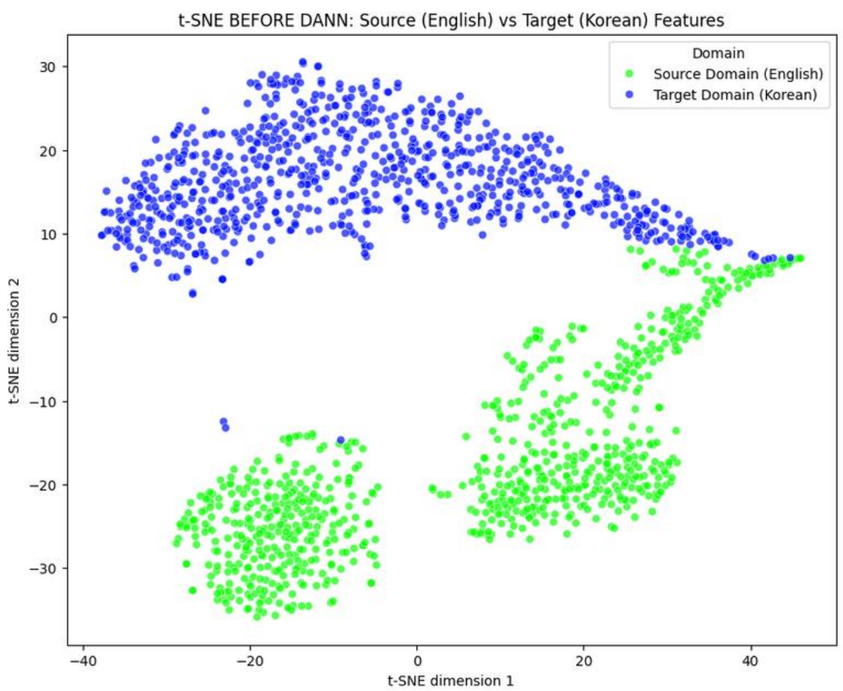
학습 전



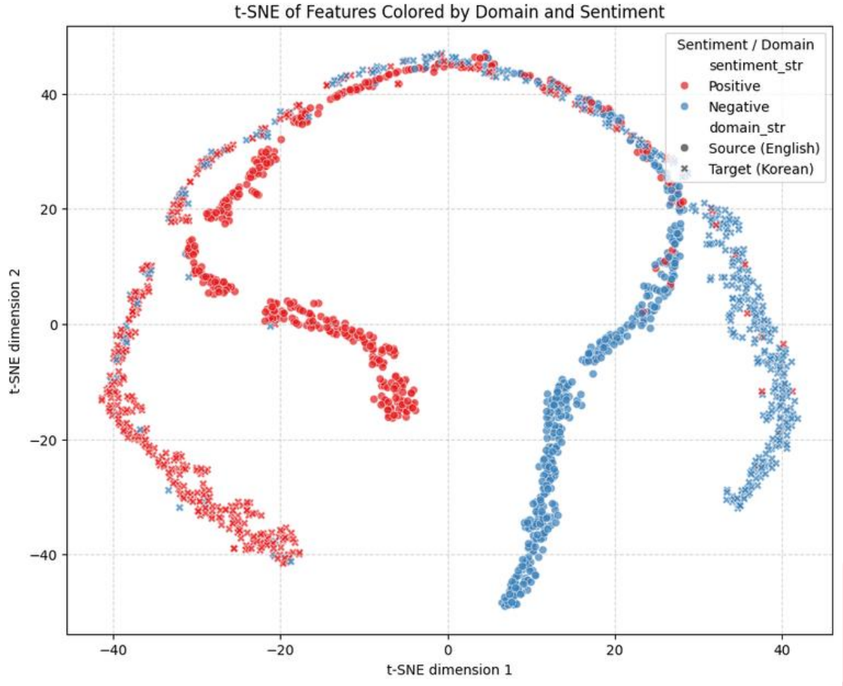
DANN 학습 후

03. Translation-free Approach

Results



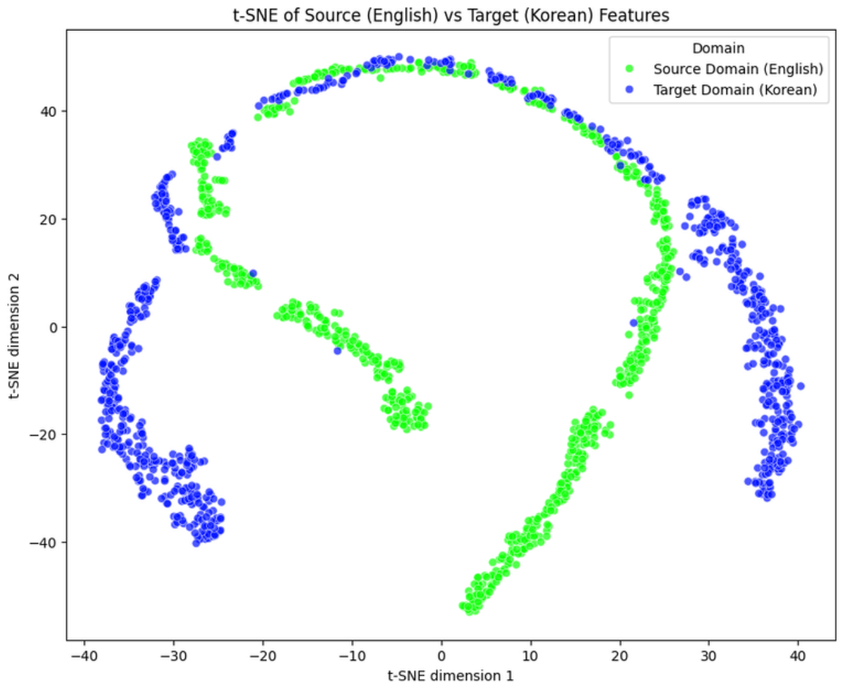
학습 전



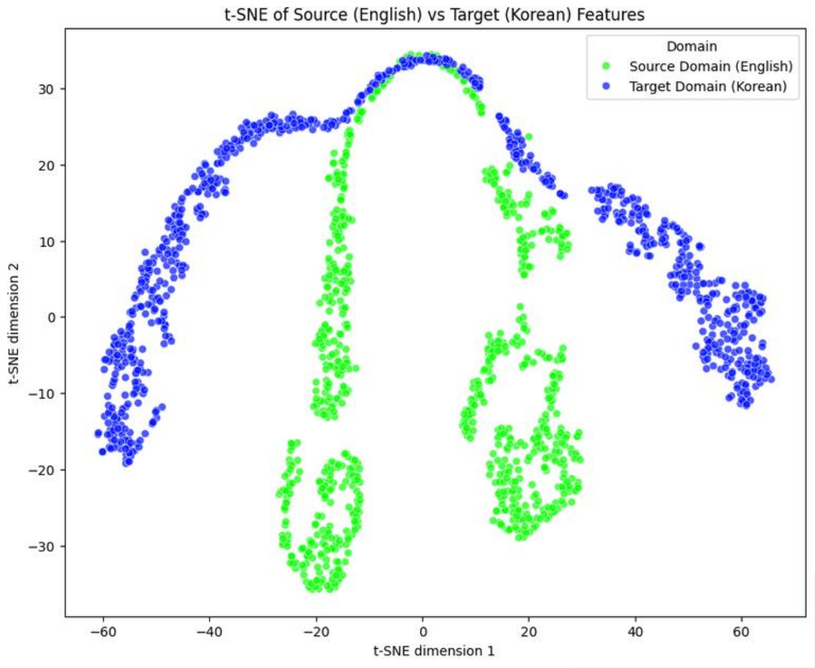
DANN 학습 후

03. Translation-free Approach

Results



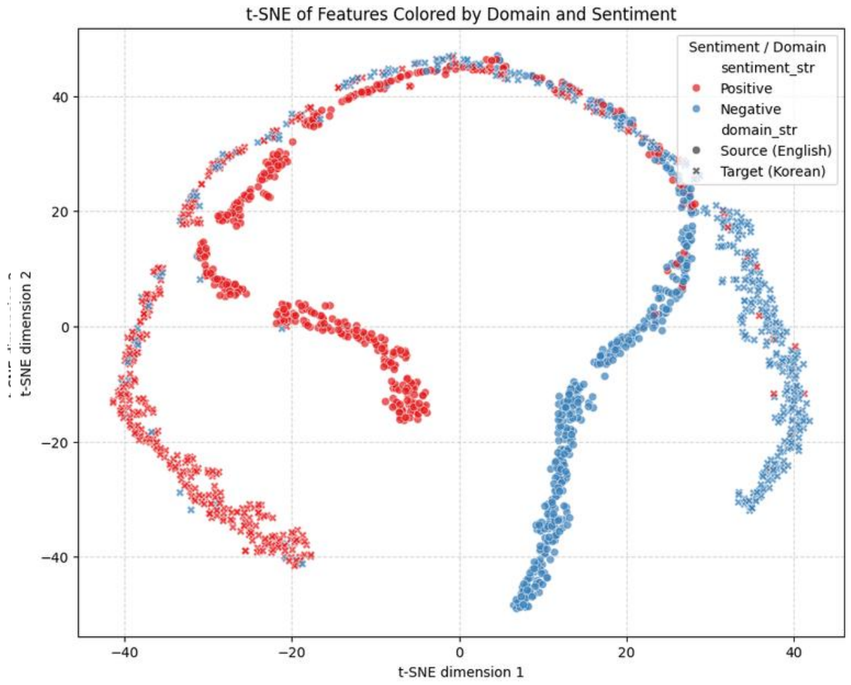
DANN 학습 후



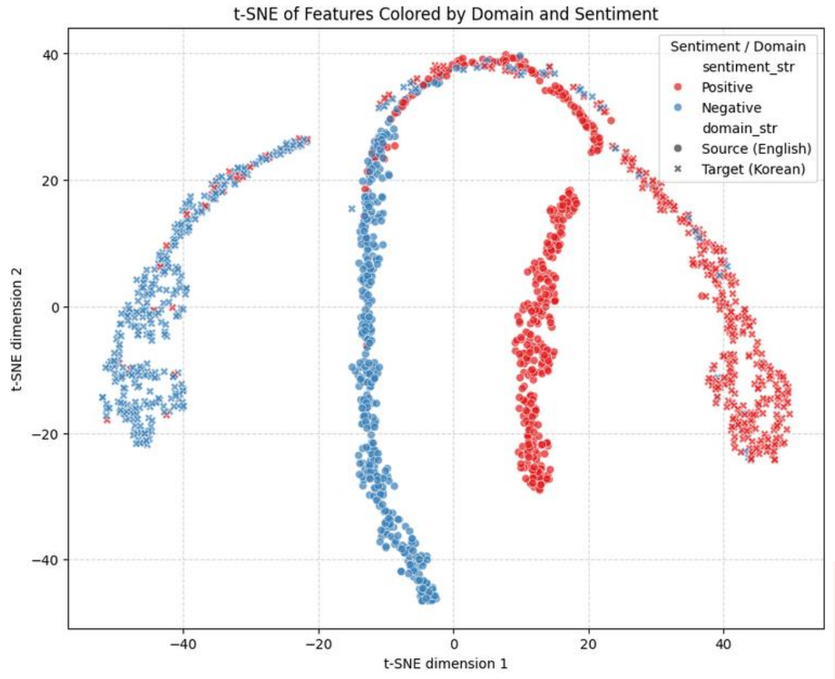
Free-Match 학습 후

03. Translation-free Approach

Results



DANN 학습 후



Free-Match 학습 후

KUBIG
DATA SCIENCE & AI

04. 결과 분석 및 향후 계획

04. 결과 분석 및 향후 계획

결과 분석

- **DANN + FreeMatch 파이프라인**
  - DANN을 통해 도메인 간 차이를 해소하여 안정적인 학습 기반 마련
  - FreeMatch가 타겟 도메인 데이터의 특성을 학습하여 성능 최적화
- **Low resource 상황에서 레이블링 비용없이 타겟 도메인에 대한 고성능 모델 구현 가능**
- **다양한 Domain Adaptation (DA) 문제에 적용 가능한 실용적인 DA 프레임워크**

04. 결과 분석 및 향후 계획

향후 계획

- **테스크 확장**
  - 이진 분류 → 다중 클래스 감성 분류 (1~5점 별점 기반)
  - 감정 범주 다양화: 기쁨 / 놀람 / 분노 / 슬픔 등으로 확장 가능

- **방법론 개선**
  - 적대 학습의 불안정성 개선: Spectral Normalization 등의 정규화 추가

KUBIG
DATA SCIENCE & AI

Thank You