공학석사학위논문

약지도 방식을 활용한
한국어 텍스트 분류

Korean Text Classification via Weak Supervision

2023 년 2 월

서울대학교 대학원
산업공학과

이 수 연

약지도 방식을 활용한
한국어 텍스트 분류

Korean Text Classification via Weak Supervision

지도교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2022 년 12 월

서울대학교 대학원

산업공학과

이 수 연

이수연의 공학석사 학위논문을 인준함

2022 년 12 월

위 원 장 이 재 욱 (인)

부위원장 조 성 준 (인)

위 원 이 성 주 (인)

초록

텍스트 분류는 토픽 분류, 감정 분석 등의 다양한 과제 및 여러 영역에서 활용도가 높은 중요한 과제이다. 이를 해결하기 위한 많은 모델은 지도 학습 기반으로 대량의 레이블이 지정된 데이터를 활용해 분류기 (classifier)를 학습시키는 구조이다. 따라서 레이블이 지정된 데이터가 부족한 영역에의 적용은 제한적이다. 특히나 한국어 자연어 처리를 위한 레이블링 된 데이터는 매우 부족하므로 기존의 많은 모델을 활용하기 어렵다. 그러나 레이블이 지정되지 않은 데이터는 보다 쉽게 구축할 수 있으므로 이러한 데이터를 텍스트 분류에 효과적으로 활용하는 것은 중요한 문제이다. 본 논문에서는 이를 해결하기 위해 클래스 이름 등과 같이 매우 적은 정보만을 이용해 레이블이 지정되지 않은 데이터를 분류하고자 하는 약지도 방식 (Weakly-supervised)의 분류 모델을 선택했다. 이러한 약지도 분류 모델에 준지도 (Semi-supervised) 학습의 성능을 개선하고자 많이 활용되었던 데이터 증강 방법론 및 자체 학습 (Self-train)을 적용하여 한국어를 분류할 수 있는 구조를 제안한다. 본 논문에서는 실제 레이블 대신 pseudo label을 생성하는 기존의 모델을 선택하였다. 생성된 pseudo label을 ground truth로 가정하여 학습을 진행한 후, 업데이트된 모델을 이용해 증강된 레이블 되지 않은 데이터에 대해 자체 학습을 진행한다. 토픽 분류와 감정 분석 데이터셋을 이용해 실험을 진행한 결과 모든 데이터셋에 대해 데이터 증강 기법을 적용하지 않았을 때보다 성능이 개선됨을 확인할 수 있었다.

주요어: 텍스트 분류, 약지도, 데이터 증강, Self-train

학번: 2021-24560

목차

초록 ii

목차 iii

표 목차 v

그림 목차 vi

제 1 장 서론 1

제 2 장 선행연구 5

2.1 약지도 방식의 텍스트 분류 연구 5

2.2 데이터 증강을 활용한 준지도 텍스트 분류 연구 10

2.3 데이터 증강 관련 연구 14

제 3 장 제안 방법 16

3.1 약지도 방식의 텍스트 분류 모델 18

3.2 데이터 증강 방법론 22

제 4 장 실험 및 결과 26

4.1 실험 데이터 26

4.2 실험 설계 31
4.3 실험 결과 및 분석 34

제 5 장 결론 45

참고문헌 47

Abstract 52

표 목차

표 1.1 텍스트 분류를 위한 영어 및 한국어 데이터셋 2
표 3.1 본 논문에서 활용한 CNN 모델 구조 20
표 3.2 감정 분석 데이터셋에서 사용한 클래스 이름 22
표 4.1 KLUE-TC 데이터셋 27
표 4.2 AI-Hub 도서 요약 데이터셋 28
표 4.3 감정 분석용 데이터셋 요약 28
표 4.4 실험 데이터셋 요약 29
표 4.5 데이터셋 예시 30
표 4.6 데이터 증강 설정 31
표 4.7 모델별 변수 설정값 32
표 4.8 클래스별 키워드 33
표 4.9 EDA[26]를 활용해 증강된 데이터 예시 35
표 4.10 Pre-train Epochs에 따른 WeSTClass[14] 모델의 F1 score (macro/micro) ... 36
표 4.11 Supervision source에 따른 WeSTClass[14] 모델의 F1 score (macro/micro) . 39
표 4.12 데이터셋에 따른 모델 성능 비교 (F1-macro/micro score) 40
표 4.13 WeSTClass[14] 모델을 통해 추출된 키워드 예시 41
표 4.14 X-class[16] 모델을 통해 추출된 키워드 예시 42
표 4.15 데이터 증강 방법론을 적용한 모델의 F1 score (macro/micro) 43
표 4.16 제안 방법론을 사용하여 실제 개선된 예시 44

그림 목차

그림 2.1 Li et al.[13] 연구의 Seed-guided Topic Model 구조 6
그림 2.2 Mekala and Shang[8] 연구에서 제시하는 모델 구조 8
그림 2.3 Meng et al.[9] 연구에서 제시하는 MCP 구조 9
그림 2.4 Zhang et al.[19] 연구에서 제시하는 MATCH 방법론 10
그림 2.5 Laine and Aila[20] 연구에서 제시하는 방법론 11
그림 2.6 Xie et al.[21] 연구의 Training Objective 12
그림 2.7 Chen et al.[22] 연구의 전체 구조 13
그림 2.8 Tarvainen et al.[24] 연구의 Mean Teacher 방법론 14
그림 2.9 Wei and Zhou[26] 연구의 잠재 공간 표현 15
그림 3.1 본 논문에서 제안하는 Framework 17
그림 3.2 Meng et al.[14] 연구에서 제시하는 모델 구조 18
그림 3.3 Wang et al.[16] 연구에서 제시하는 모델 구조 21
그림 3.4 CNN Classifier 앙상블 구조도 24
그림 3.5 증강 데이터를 이용한 Classifier Self-train 과정 25
그림 4.1 Pre-train Epochs에 따른 WeSTClass[14] 모델의 F1-macro score 37
그림 4.2 데이터 증강 방법론을 적용한 모델의 F1-macro score 43

제 1 장 서론

텍스트 분류는 다양한 자연어 처리 기법을 통해 텍스트를 미리 정의된 범주로 분류하는 것으로 다음과 같은 여러 분야에 활용되고 있다. (1) 뉴스 기사 등의 텍스트의 토픽을 구분하는 토픽 분류, (2) 텍스트의 긍정/부정의 감정을 분석하는 감정 분석, (3) 사용자의 선호를 분석하여 아이템을 추천해주는 추천 시스템[1] 등이다. 또한, 딥러닝 기법을 활용하여 의학 텍스트를 분류하거나[2] 특허 문서를 분류하는[3] 등 적용되는 분야도 매우 다양하다.

텍스트 분류 기법은 Li et al. (2022)[4]에 의하면 크게 머신러닝 기반과 딥러닝 기반으로 나눠볼 수 있다. 먼저 머신러닝 기반 기법은 Naïve Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) 등의 전통적인 방법이다. 다음으로 딥러닝 기반 기법에는 Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Transformers, BERT나 GPT와 같은 Pre-trained 언어 모델을 활용하는 방법 등이 있다.

기존의 많은 방법은 지도 학습 (Supervised) 기반으로 레이블이 지정된 대량의 학습 데이터가 필요하다. 따라서 레이블이 지정된 데이터가 부족한 영역에는 활용하기가 어렵다는 문제점이 있다. 이를 해결하기 위해 준지도 방식 (Semi-supervised)이나 약지도 방식 (Weakly-supervised)의 텍스트 분류 기법들이 연구되고 있다. 준지도 텍스트 분류는 레이블이 지정된 데이터 (labeled)와 레이블이 지정되지 않은 데이터 (unlabeled)를 함께 활용하는 것이다. 약지도 텍스트 분류는 레이블링 되지 않은 데이터만을 이용하는 것이다.

이와 같이 활용도가 높은 텍스트 분류 기법을 고도화하고자 하는 연구가 계속되어 왔다. 이에 비해 한국어 텍스트 분류에 관한 연구는 부족하다. 그 이유로는 한국어의 특수성과 텍스트 분류 목적으로 레이블이 지정된 한국어

데이터셋이 부족하다는 문제점을 들 수 있다. 한국어 텍스트 분류를 위해 주로 사용되는 데이터셋을 살펴보면 뉴스/토픽 분류를 위한 KLUE Benchmark[5]와 감정 분석을 위한 NSMC (Naver Sentiment Movie Corpus)[6], 네이버 쇼핑 데이터셋[7] 등이 있다. 아래의 표 1.1에 주로 사용되는 데이터셋을 정리하였다. 이를 통해 영어권보다 데이터셋의 수도 적을 뿐 아니라 데이터셋의 크기도 별로 크지 않다는 것을 확인할 수 있다.

표 1.1: 텍스트 분류를 위한 영어 및 한국어 데이터셋

구분	영어	한국어
토픽 분류	AG NEWS (127,600) DBPedia (630,000) 20News (18,846)	KLUE-TC (63,823)
감정 분석	IMDB (50,000) Yelp (6,990,280) Amazon (4,000,000)	NSMC (200,000) 네이버 쇼핑 (200,000)

* 데이터셋 이름 (데이터셋 크기)

이처럼 한국어 텍스트 분류 기법을 효율적으로 발전시키기에는 한국어 데이터셋이 부족한 상황이다. 수동으로 레이블을 지정하여 지도 학습에 사용되는 데이터셋을 생성할 수 있지만 이는 비용과 시간이 많이 소요되는 작업이다. 레이블을 일일이 지정하기 위해서는 전문가 논의를 통해 레이블에 대해 의논하고 배정해야 하기 때문이다. 반면 레이블이 지정되지 않은 데이터셋은 텍스트 분류가 활용되는 여러 영역에서 쉽게 수집할 수 있다. 또한 의학, 특허와 같은 보다 특수한 영역에서도 수집할 수 있어[8] 텍스트 분류의 영역을 넓힐 수 있다.

따라서 레이블이 지정된 데이터셋이 부족한 상황에서 레이블이 지정되지 않은 한국어 데이터셋을 적극적으로 이용할 수 있는 텍스트 분류 기법을 발전시키는 것은 중요한 문제이다. 이를 위해 앞서 언급한 것처럼 준지도 방식의 텍스트 분류 기법이나 약지도 방식의 텍스트 분류 기법을 활용할 수 있다. 그러나 준지도 텍스트 분류는 여전히 레이블링 된 데이터가 일부 필요한 방식이다. 일부만을 레이블링하기 위해서도 레이블에 대한 명확한 정의와 전체 데이터 특성에 대한 분석이 필요하다. 그렇기 때문에 해당 방식은 전체 데이터에 대한 구체적인 파악이 어렵고 전체 레이블 클래스만을 확인할 수 있는 경우 적용이 어렵다는 한계가 있다. 또한, 전체 데이터 중 레이블링 된 데이터의 비중이 높아질수록 성능이 개선된다. 그러므로 본 논문에서는 더욱 다양한 경우에 적용할 수 있는 약지도 방식 (Weak supervision)의 텍스트 분류 기법을 활용해서 위의 문제에 접근했다. 약지도 방식은 전체 데이터를 확인하기 어려워도 레이블 클래스만 주어지면 적용할 수 있으므로 준지도 방식에 비해 범용성이 높다고 판단하였다. 이때 기존의 약지도 텍스트 분류 모델의 성능을 향상할 수 있는 Self-train과 레이블이 지정되지 않은 데이터의 증강을 결합한 구조를 제안한다. 이러한 방식을 통해 데이터의 크기를 효과적으로 확장하여 성능을 개선하고자 한다.

본 논문은 레이블이 지정된 한국어 데이터셋이 부족한 상황에서 레이블이 지정되지 않은 데이터셋을 효과적으로 활용하여 텍스트를 분류하고자 시작되었다. 이를 해결하기 위해 데이터 증강이 활용된 모듈이 추가된 약지도 방식의 한국어 텍스트 분류 구조를 제시한다. 본 연구의 공헌은 다음과 같다.

먼저 주로 준지도 방식의 학습에 사용되었던 데이터 증강을 통해서 레이블이 지정되지 않은 데이터를 효과적으로 분류하고자 했던 접근법 ([20], [21], [22])을 약지도 방식에 적용하였다. 본 연구를 통해 데이터 증강을 적용하면 기존의 약지도 방식의 텍스트 분류 모델만을 적용하는 것보다 성능을 개선할 수 있음을 보인다. 다음으로 레이블링 되지 않은 데이터 활용의 필요성이 높은 한국어

텍스트 대상으로 약지도 분류 모델을 적용한 연구로 기존 한국어 텍스트 분류 연구와 차이가 있다. 마지막으로 추후 다양한 분야의 데이터셋으로의 확장성이 높다는 점에서 의의가 있다.

본 논문은 5장으로 구성된다. 제 2장에서는 약지도 방식 및 데이터 증강과 관련된 선행 연구를 살펴본다. 제 3장에서는 본 논문에서 활용하고자 하는 연구 방법을 제안한다. 제 4장에서는 실험에 사용하고자 하는 데이터를 확인하고 실험 결과를 살펴본다. 마지막으로 제 5장에서는 결론과 한계점 및 향후 연구방향을 제시한다.

제 2 장 선행연구

2.1 약지도 방식의 텍스트 분류 연구

약지도 방식의 텍스트 분류란 레이블이 지정된 텍스트 데이터를 이용하지 않고 레이블이 지정되지 않은 데이터와 추가로 레이블 이름이나 레이블과 관련된 소수의 키워드만을 활용해서 분류하는 것이다.

약지도 방식은 Meng et al. (2020)[9]에 의하면 크게 세 가지로 구분할 수 있다. 1) Dataless 분류 접근법, 2) 토픽 모델링 기반 접근법, 3) 신경망 기반 접근법이다. 첫 번째로 Chang et al. (2008)[10]에 의해 제안된 Dataless classification은 Wikipedia를 World knowledge source로 활용한다. Explicit Semantic Analysis (ESA)를 이용하여 의미론적 관점에서 레이블 이름과 주어진 문서를 분석해 문서와 관련된 컨셉을 생성한다. 이렇게 얻어진 의미론적 해석을 활용해 Classifier를 학습시킨다. Song and Roth (2014)[11]는 Dataless hierarchical classification을 제시했다. 이는 주어진 데이터셋과 레이블을 semantic space에 임베딩하여 문서와 잠재 레이블 간의 의미론적 유사도를 계산하는 방식으로 이루어진다.

두 번째로 토픽 모델링 기반 접근법은 주어진 키워드, 시드 단어 (seed words)를 활용하여 분류하는 연구이다. Chen et al. (2015)[12]는 카테고리 설명 단어 (Category description words)만을 활용해 텍스트를 분류하는 설명적 LDA (Descriptive LDA; DescLDA)를 제안했다. 먼저 주어진 카테고리 설명 단어로 생성된 소수의 문서에 대한 Descriptive Dirichlet priors를 추론한다. 그다음 Category-aware topics를 학습하여 문서에 카테고리 레이블을 배정하게 된다. Li et al. (2016)[13]의 연구에서는 그림 2.1에서 상세하게 확인할 수 있는 Seed-guided

Topic Model (STM)이 제안되었다. 이는 앞의 연구([12])와 유사하게 레이블이 지정되지 않은 문서 집합과 카테고리의 의미와 관련된 시드 단어를 활용한다. Category word probability와 Initial document category distribution을 추정하고 Topic Influence를 통해 문서의 레이블을 예측하는 모델이다.

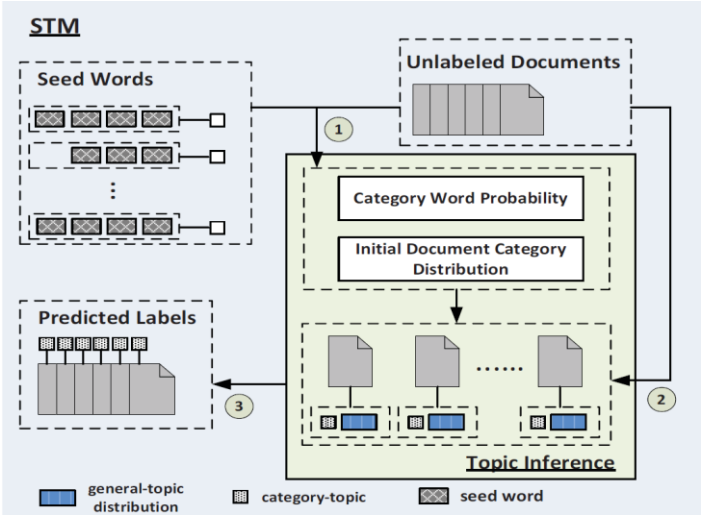

그림 2.1: Li et al.[13] 연구의 Seed-guided Topic Model 구조

마지막으로 신경망 기반 접근법은 Neural model을 이용해서 더욱 고도의 분석을 하고자 하는 연구 흐름이다. 주어진 문서에 대해 pseudo label을 생성해 지도 학습처럼 (문서, 레이블) 기반으로 Neural classifier를 학습시키는 방식이다. 이는 소량의 시드 정보를 함께 활용하는 연구 ([8], [14], [15])와 레이블 이름만을 활용하는 연구 ([9], [16])로 나누어 볼 수 있다. Meng et al. (2018)[14]은 Weakly-Supervised Text Classification (WeSTClass) 모델을 제시했다. 해당 모델은 클래스의 semantics를 모델링하여 생성된 pseudo documents를 통해 Self-training 방식으로 Neural classifier를 학습시킨다. 소량의 시드 정보는 구축할 수 있다는

가정에서 나온 방법론으로, 다음과 같은 세 종류의 시드 정보를 활용한다. (1) 각 클래스에 대한 표현적 키워드, (2) 12개 이하의 레이블링 된 문서 데이터, (3) 클래스 이름이다.

약지도 방식으로도 앞서 서술된 Song and Roth (2014)[11]의 연구처럼 계층적 관계를 반영하여 텍스트 분류를 하고자 한 연구가 있었다. [14]의 연구를 확장하여 Meng et al. (2019)[15]은 클래스 간의 관계를 고려하여 계층적으로 분류하기 위한 WeSHClass 모델을 제시했다. 해당 모델도 [14]와 유사하게 시드 정보를 활용해 각 클래스의 분포 기반으로 pseudo document를 생성하여 Self-train 하는 구조이다. 차이점은 클래스 계층 구조의 각 노드에서 Local classifier를 훈련시키고, 통합하여 구축된 Global classifier가 재귀적인 방식으로 최종 예측에 사용된다는 점이다.

Mekala and Shang (2020)[8]은 기존 연구처럼 시드 단어를 활용하지만, 문맥(context)를 반영하고자 했고 사전 학습된 언어 모델을 사용했다는 점에서 기존 연구와 차이점을 보인다. User-provided seed words를 이용하여 Text classifier를 훈련시키는 Contextualized Weak supervision (ConWea)을 제시했다. 먼저 BERT와 같은 사전 학습된 언어 모델을 Contextualized learning 기법으로 활용한다. 이러한 기법과 시드 정보를 결합하여 Contextualized corpus를 구축한다. 또한 시드 단어를 계속해서 확장하여 Contextualized seed words를 생성한다. 이와 같은 정보를 이용해 레이블링 되지 않은 contextualized 문서에 대해 pseudo label을 생성하여 Text classifier를 훈련시키게 된다. 이때 Classifier로는 문서 간의 계층적 구조를 고려하는 Hierarchical Attention Networks (HAN)을 이용한다.

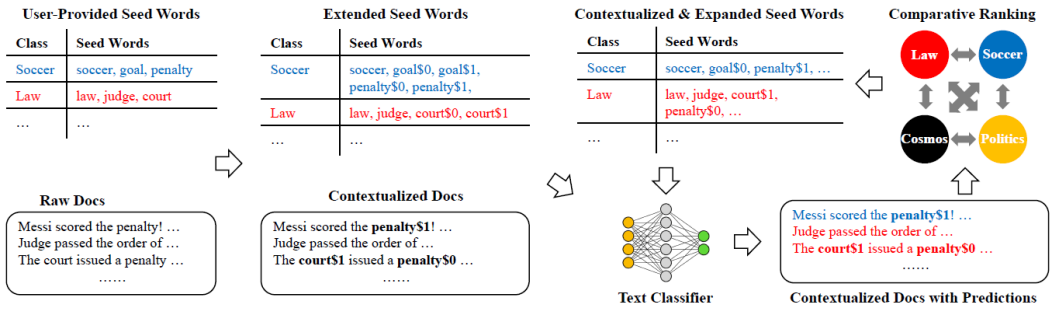

그림 2.2: Mekala and Shang[8] 연구에서 제시하는 모델 구조

여기서 더 나아가 시드 정보도 활용하지 않는 약지도 방식의 연구가 제시되었다. Meng et al. (2020)[9]의 연구에서는 오직 레이블 이름만을 활용하여 레이블이 지정되지 않은 데이터셋을 분류하는 LoTClass (Label-Name-Only Text Classification)이 제시되었다. BERT와 같은 사전 학습된 언어 모델을 Category understanding과 분류를 위한 Feature representation learning 모델로 사용한다. 먼저 사전 학습된 Masked Language Model (MLM)을 레이블이 지정되지 않은 데이터에서 주어진 레이블 이름을 대체할 수 있는 단어를 예측하도록 훈련시켜서 각 클래스의 Category vocabulary를 구축한다. 다음으로 Masked Category Prediction (MCP) task로 Contextualized word-level category supervision을 생성한다. 이를 통해 MCP head가 마스킹 된 단어의 카테고리를 예측하도록 훈련시킨다. 마지막으로 이렇게 훈련된 모델을 전체 데이터셋에 대해 Self-train 시킴으로써 모델을 일반화한다.

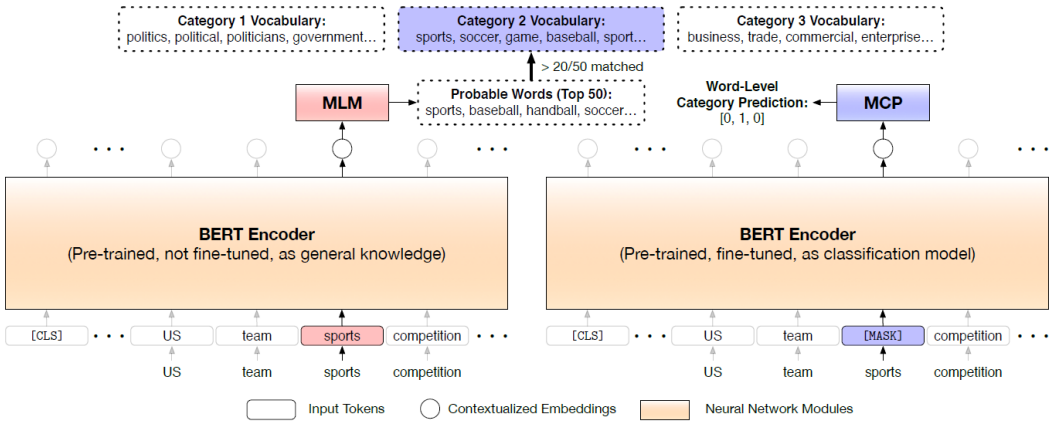

그림 2.3: Meng et al.[9] 연구에서 제시하는 MCP 구조

Wang et al. (2021)[16] 역시 오직 레이블의 이름만을 활용하는 Extremely weak supervision 모델인 X-class를 제시했다. 앞의 연구[9]와 다른 점은 클래스와 문서의 관계를 고려하여 문서를 표현한다는 점이다. 먼저 BERT와 같은 사전 학습된 언어 모델을 이용해서 주어진 레이블과 유사한 단어들을 추가하며 Class representation을 학습한다. 또한, 이를 활용하여 Document representation을 표현하고, 클러스터링 기법을 적용해 pseudo training set을 구축한다. 이러한 pseudo set으로 Classifier를 학습시키는데, 이 모델에서는 BERT classifier를 사용했다.

이 외에도 단순 텍스트, 문서 데이터 외에도 작성자나 태그 등과 같은 메타데이터를 함께 활용해 약지도 텍스트 분류를 하고자 하는 연구도 존재한다. Zhang et al. (2018)[17]의 연구는 먼저 레이블 및 여러 메타데이터와 단어, 문서가 같은 잠재 공간에 있도록 임베딩한다. 이를 기반으로 단어와 메타데이터를 모두 포함하는 학습용 문서를 합성하여 생성한 후 Classifier를 학습시키는 구조이다. Mekala et al. (2020)[18]에서 제안하는 META (Metadata-empowered weakly-supervised text classification) framework는 기존의 약지도 분류 방법론과 유사한

구조를 가진다. 주어진 시드 단어에 기반하여 생성된 pseudo label을 활용해 Classifier를 학습시키는 것이다. 이때 메타데이터를 추가 소스로 활용하고 네트워크 모티프라는 개념을 도입했다는 점에서 차이가 있다. 적절한 메타데이터의 조합을 포착하여 시드 모티프를 추출하는 것이다.

앞서 서술한 계층 구조와 메타데이터를 결합하여 약지도 방식의 텍스트 분류 모델을 향상하고자 하는 방법론도 있다. Zhang et el. (2021)[19]의 연구에서는 메타데이터와 레이블의 계층 구조를 모두 활용하는 MATCH 솔루션을 제안한다. 상세 정보는 그림 2.4를 통해 확인할 수 있다. 먼저 동일한 임베딩 공간에서 주어진 문서와 메타데이터의 임베딩을 사전 학습시킨다. 다음으로 완전 연결된 어텐션 (Fully-Connected Attention)을 이용해 문서와 메타데이터 간의 관계를 포착한다. 이후 해당 논문에서 제안하는 여러 정규화 기법을 이용해 주어진 레이블 간의 계층을 반영하여 최종 레이블 예측을 도출하게 된다.

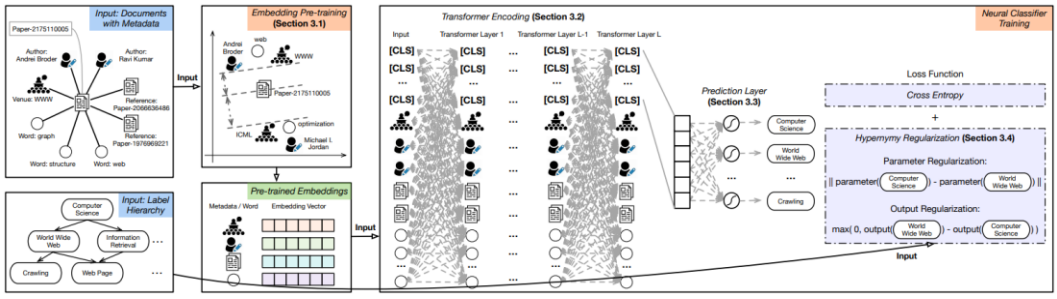

그림 2.4: Zhang et al.[19] 연구에서 제시하는 MATCH 방법론

2.2 데이터 증강을 활용한 준지도 텍스트 분류 연구

본 절에서는 지속해서 수행되고 있는 데이터 증강을 활용하여 준지도 방식 (Semi-supervised)의 텍스트 분류 성능을 개선하고자 하는 연구에 대해 살펴보고자 한다.

먼저 Laine and Aila (2016)[20]는 여러 네트워크 앙상블의 성능 개선 효과를 단일 네트워크에도 적용해보고자 한 연구이다. 이를 위해 단일 네트워크 내의 개별 하위 네트워크의 앙상블을 이용했다. 서로 다른 epochs, 정규화, 입력 데이터 증강 조건에서의 network-in-training의 outputs를 이용하여 레이블링 되지 않은 데이터에 대한 레이블 예측값의 consensus를 생성하는 Self-Ensembling 방법론을 제시했다.

두 가지 모델을 제시하는데 먼저 Π-model은 Training input에 대해 두 번 Stochastic augmentation과 Dropout을 적용하여 두 개의 예측값을 생성한다. Supervised cross-entropy Loss는 Labeled inputs만을 사용해서 계산된다. Unsupervised Loss로는 두 예측값 사이의 Mean squared difference를 계산하고, 이러한 두 Loss의 가중 합을 계산한다. 두 번째 모델인 Temporal Ensembling은 Π-model과 달리 여러 네트워크의 평가 outputs를 앙상블 outputs로 통합한다. 따라서 epoch 당 입력에 대해 학습 시 평가를 한 번만 진행한다. Unsupervised Loss는 이전의 네트워크 앙상블 평갓값에 의존한다. 아래 그림 2.5에서 두 모델의 자세한 구조를 확인할 수 있다.

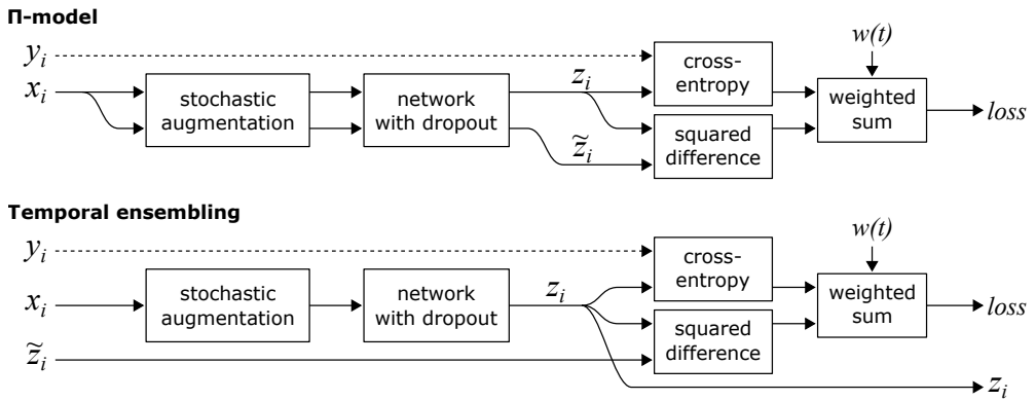

그림 2.5: Laine and Aila[20] 연구에서 제시하는 방법론

다음으로 Xie et al. (2020)[21]의 연구에서 제안된 Unsupervised Data Augmentation (UDA)는 Consistency training framework를 활용하여 주로 지도 학습에서 사용되던 데이터 증강 방법론을 준지도 학습으로 확장한 것이다. Consistency learning은 데이터에 noise를 추가하여 작은 변화에 강건하도록 정규화하는 것이다. 해당 논문에서는 noise 추가 대신 고품질의 데이터 증강으로 대체해도 가능하다고 서술한다. 상세 구조는 아래의 그림 2.6을 통해 확인할 수 있다. 레이블링 된 데이터와 레이블링 되지 않은 데이터를 함께 활용하여 학습을 진행한다. 먼저 레이블링 된 데이터 x_1 을 통해서는 예측값과 실제 값 사이의 Supervised cross-entropy Loss $(p_\theta(y|x))$ 를 구성한다. Unsupervised consistency Loss는 원래 데이터 x_2 와 증강된 데이터 $\hat{x}$ 의 확률 분포 간의 차이인 KL Divergence로 구성된다. Final Loss는 식 (2.1)처럼 두 Loss의 합으로 정의된다.

$$\min_\theta \mathcal{J}(\theta) = \mathbb{E}_{x_1 \sim P_L(x)}\left[-log_{p_\theta}(f^*(x_1)|x_1)\right] + \lambda \mathbb{E}_{x_2 \sim P_U(x)} \mathbb{E}_{\hat{x} \sim q(\hat{x}|x_2)}\left[CE(p_{\tilde{\theta}}(y|x_2)) \,\|\, p_\theta(y|\hat{x})\right] \quad (2.1)$$

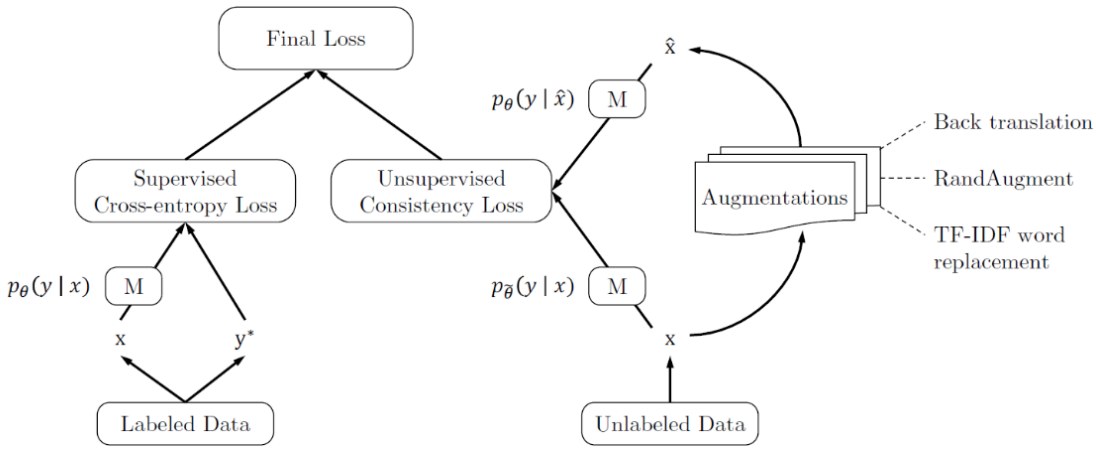

그림 2.6: Xie et al.[21] 연구의 Training Objective

마지막으로 Chen et al. (2020)[22]은 레이블링 된 데이터, 레이블링 되지 않은 기존 데이터, 증강 데이터를 혼합하는 MixText 모델을 제안했다. 이때 데이터 증강을 위해 새롭게 제시된 TMix 방법론을 통해 레이블이 없는 데이터를 증강한다. 이러한 TMix 방법론은 Zhang et al. (2017)[23]의 연구에서 제시된 MixUp 방법론을 텍스트 영역으로 확장한 것이다. Textual hidden space에서 각 텍스트의 Hidden representation에 Interpolation을 적용하여 증강한다. MixText는 먼저 세 가지 종류의 데이터를 활용해 (2.2)와 같이 Supervised Loss와 Consistency Loss를 계산한다. 다음으로 레이블이 없는 데이터에 대해 Confident labels를 생성하게 하도록 (2.3)에 표현된 것처럼 Entropy minimization을 도입한다. 마지막으로 이를 통합한 전체 Loss function은 (2.4)로 정의된다.

$$L_{TMix} = \mathbb{E}_{x,x' \in X} KL(min(y, y') \| p(TMix(x, x'))) \tag{2. 2}$$

$$L_{margin} = \mathbb{E}_{x \in X_u} max\,(0, \gamma - \|y^u\|_2^2) \tag{2. 3}$$

$$L_{MixText} = L_{TMix} + \gamma_m L_{margin} \tag{2. 4}$$

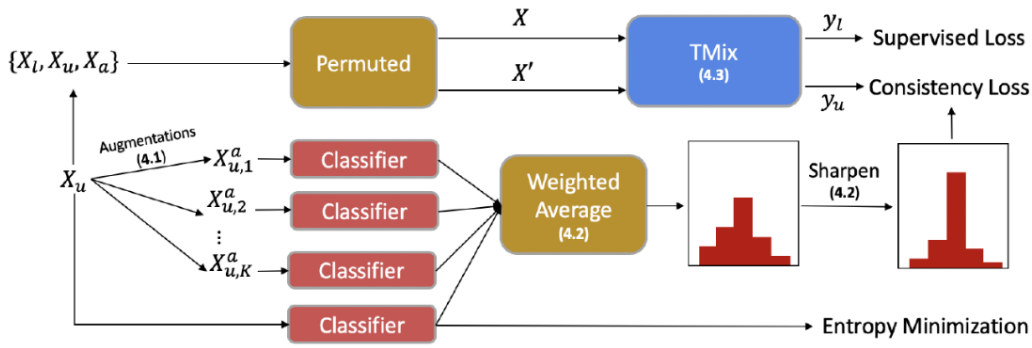

그림 2.7: Chen et al.[22] 연구의 전체 구조

Tarvainen et al. (2017)[24]의 연구에서는 Mean Teacher 모델을 제안한다. 이는 Laine and Aila (2016)[20]의 Temporal Ensembling 방법을 개선한 것으로, Teacher 모델과 Student 모델을 도입하여 차별점을 둔다. Student 모델 weights의 Exponential moving average를 이용해 Teacher 모델의 weights를 업데이트하여 학습시키는 구조이다. 이를 통해 모델 예측의 정확도를 향상할 수 있다.

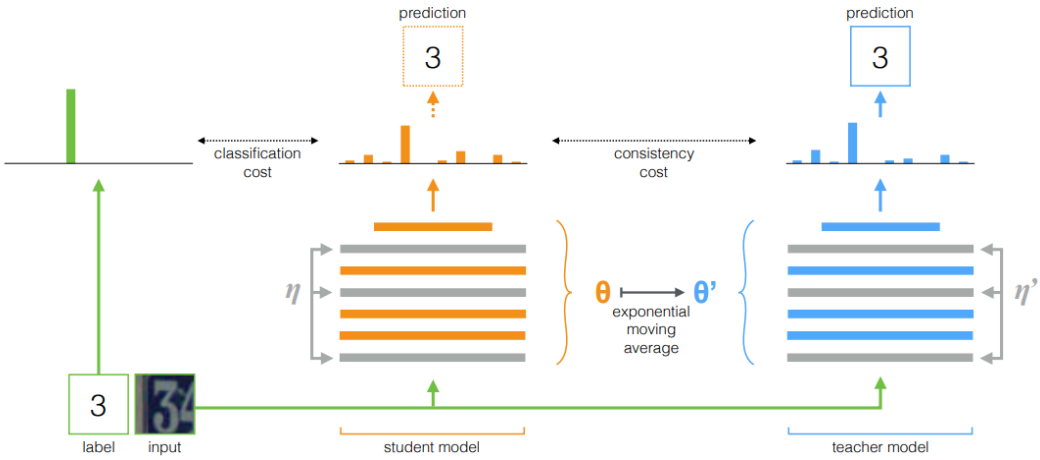

그림 2.8: Tarvainen et al.[24] 연구의 Mean Teacher 방법론

이 외에도 Berthelot et al. (2019)[25]의 MixMatch 방법론 등 다양한 방식의 데이터 증강 기법을 통해 준지도 학습을 향상하려는 연구가 지속적으로 이루어지고 있음을 알 수 있다.

2.3 데이터 증강 관련 연구

본 절에서는 데이터 증강과 관련된 연구를 살펴보고자 한다. Wei and Zhou (2019)[26]에서 제안된 Easy Document Augmentation (EDA)는 텍스트 분류 성능

개선을 위한 4개의 operations로 구성된다. 첫 번째, 유의어 교체(Synonym Replacement; SR)는 임의로 불용어가 아닌 n개의 단어를 선택해 유의어로 교체하는 것이다. 두 번째, 임의 삽입(Random Insertion; RI)은 임의 단어의 유의어를 문장 내 임의의 위치에 삽입하는 것이다. 세 번째, 임의 교체(Random Swap; RS)는 임의의 단어 2개를 선택해 위치를 교체하는 것이다. 네 번째, 임의 삭제(Random Deletion; RD)는 확률 p로 임의 단어를 삭제하는 것이다.

이러한 증강 방법론을 적용했을 때 문장의 의미가 크게 변화하면 원래의 클래스 레이블이 유지되지 않을 수도 있다. 이를 확인하기 위해 해당 논문에서는 t-SNE를 이용해 시각화를 해보았는데 그 결과는 아래의 그림 2.9와 같다. 해당 그림을 통해 EDA로 증강된 문장들이 원 문장들의 레이블에 따른 잠재 공간 표현(latent space representation)과 매우 유사하게 표현되고 있음을 확인할 수 있었다. 따라서 본 기법을 적용해도 클래스 레이블이 보존된다는 결과를 얻을 수 있다. 이와 더불어 실제로 이 방법을 이용한 텍스트 분류 모델의 성능이 개선되었다고 서술하고 있다.

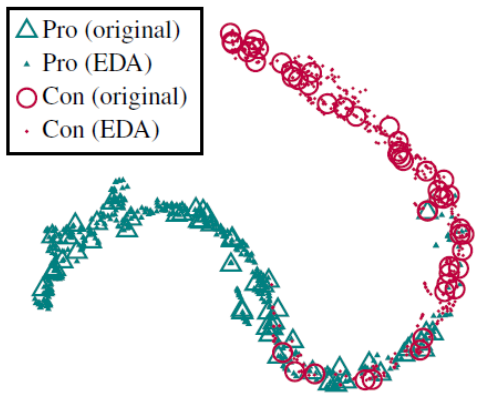

그림 2.9: Wei and Zhou[26] 연구의 잠재 공간 표현

제 3 장 제안 방법

2장에서 살펴본 선행 연구를 통해 다양한 방식의 데이터 증강을 이용한 방법론이 준지도 방식의 텍스트 분류에는 적용되어 왔지만 약지도 방식의 텍스트 분류에는 적용되지 않고 있음을 확인할 수 있었다. 이에 본 연구에서는 기존의 약지도 텍스트 분류 모델을 선정하여 해당 모델의 성능을 개선할 수 있는 데이터 증강을 활용한 Self-train 과정을 결합한 분류 모델 구조를 제안한다.

기존의 모델은 간단한 시드 정보와 레이블이 지정되지 않은 데이터가 주어지면 약지도 방법론을 활용해 pseudo label을 생성하여 이를 이용해 Classifier를 훈련하는 구조이다. 여기에 Classifier의 앙상블 구조 및 Self-train 과정과 레이블링 되지 않은 데이터의 증강을 적용해 효과적으로 모델의 성능을 향상할 수 있다. 전체 framework를 요약하면 아래의 그림 3.1과 같다.

다음 절에서는 변형하여 활용하고자 하는 약지도 분류 모델과 데이터 증강 모듈을 추가하여 변형한 모델로 나누어서 서술한다.

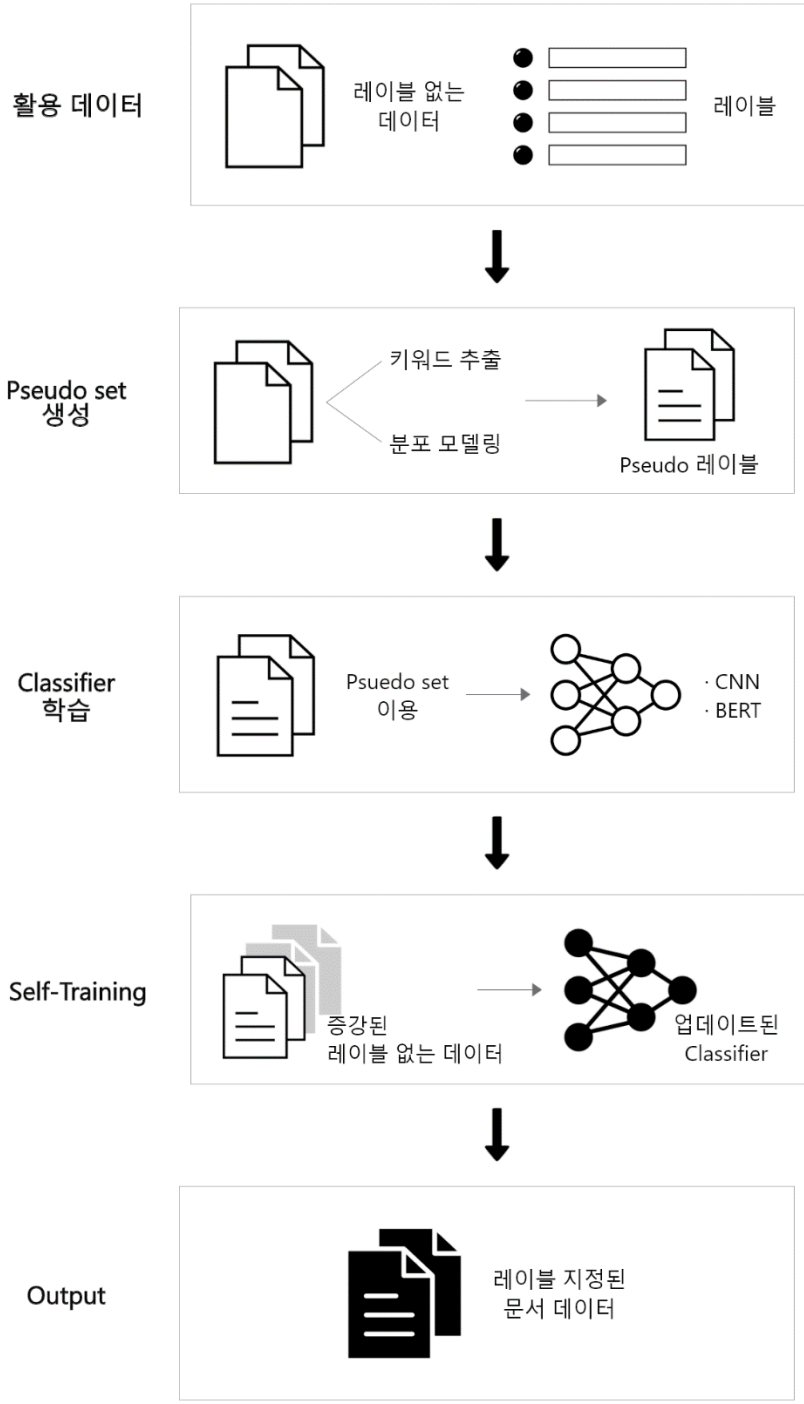

그림 3.1: 본 논문에서 제안하는 Framework

3.1 약지도 방식의 텍스트 분류 모델

2.1에서 서술한 여러 종류의 약지도 텍스트 분류 모델 중 다양한 정보를 파악할 수 있는 신경망 (Neural network) 기반의 모델에 관해 연구하고자 한다. 시드 정보를 활용하는 모델과 레이블 이름만을 활용하는 모델을 하나씩 선택하여 실험을 진행했다. 각 모델의 구조에 대해 간략하게 서술한다.

첫 번째, 시드 정보를 이용하는 모델로는 다양한 Supervision source를 적용한 Meng et al.[14]의 모델을 활용하였다. 레이블 이름, 각 클래스 (레이블)에 대한 표현적 키워드, 12개 이하의 레이블링 된 문서 데이터 중 하나의 시드 정보와 레이블이 지정되지 않은 데이터가 주어진다. 해당 연구에서 제안하는 WeSTClass (Weakly-Supervised Text Classification) 모듈은 Pseudo-document generation 모듈과 Self-training 모듈로 구성된다. 전체 모델 구조도는 아래 그림 3.2와 같다.

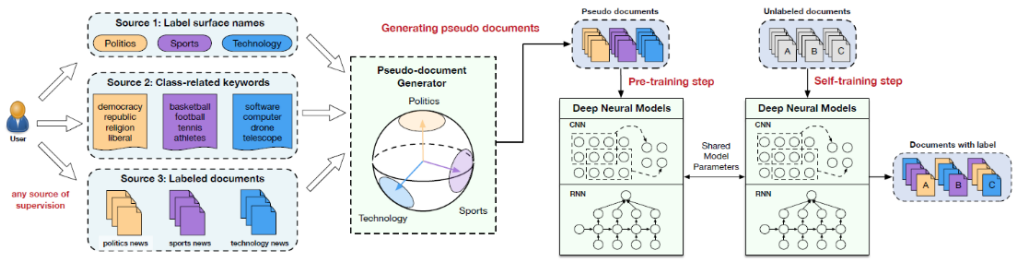

그림 3.2: Meng et al.[14] 연구에서 제시하는 모델 구조

먼저 pseudo-document를 생성하기 위해서 Skip-Gram 모델을 활용해 corpus의 단어들을 임베딩한다. 이를 통해 주어진 문서와 단어가 결합된 semantic space에 있도록 한다. 이러한 semantic space에서 제공된 시드 정보를 이용해 각 클래스와 연관된 키워드들을 추출하고, 각 클래스를 고차원의 구형 분포로 모델링한다. 이후 각 클래스의 분포에 기반한 Generative mixture model을 제시하여 같은 구형

분포에 존재하게끔 pseudo-document를 생성한다. 이에 대한 pseudo label로는 모델이 과적합 하지 않게 하도록 one-hot vector가 아닌 확률 분포를 부여한다.

다음은 Self-training 모듈로, 이전 단계에서 생성된 pseudo-document를 이용해 신경망 모델을 훈련시킨다. Class j로부터 생성된 pseudo-document D_i^* 에 대한 pseudo label을 l_{ij} 라 하고 Classifier의 output을 Y 라 한다. 다음 (3.1)의 식을 사용하여 모델의 예측값과 pseudo label 사이의 KL divergence Loss를 계산하여 학습시킨다.

$$loss = KL(L||Y) = \sum_i \sum_j l_{ij}\, log \frac{l_{ij}}{y_{ij}} \quad (3.\ 1)$$

이후 레이블이 지정되지 않은 실제 문서 데이터를 반복적으로 Bootstrapping 하는 방식으로 Self-training을 진행하여 모델을 개선한다. 이때 [14]의 연구에서는 신경망 모델로 CNN과 RNN을 사용하였다. 그러나 논문에 서술된 대부분의 실험 결과에서 CNN이 우수한 성능을 보여 본 연구에서도 CNN을 기본 모델로 이용하였다. 사용한 CNN의 구조는 표 3.1과 같다.

표 3.1: 본 논문에서 활용한 CNN 모델 구조

Layer (Type)
input (InputLayer)
embedding (Embedding)
conv1d (Conv1D)
conv1d_1 (Conv1D)
conv1d_2 (Conv1D)
conv1d_3 (Conv1D)
global_max_pooling1d (GlobalMaxPooling1D)
global_max_pooling1d_1 (GlobalMaxPooling1D)
global_max_pooling1d_2 (GlobalMaxPooling1D)
global_max_pooling1d_3 (GlobalMaxPooling1D)
concatenate (Concatenate)
dense (Dense)
dense_1 (Dense)

두 번째, 레이블 이름만을 이용하여 분류하는 모델로는 Wang et al.[16]의 연구를 활용하였다. 해당 연구는 Class-Oriented Representation, Document-Class Alignment, Text Classifier Training 3단계로 구성된다.

먼저 corpus 내의 각 단어에 대해 BERT와 같은 사전 학습된 언어 모델과 주어진 클래스 정보를 활용하여 Contextualized wore representation을 생성한다. 이러한 Contextualized representation을 평균하여 단어들의 Static representation을 계산한다. 이를 기반으로 주어진 클래스와 연관된 키워드를 반복적으로 추가하여 Class representation을 구축한다. 마지막으로 Attention 기법을 적용해 Class-oriented Document representation을 구성한다.

다음으로 Gaussian Mixture Model (GMM)을 이용해 문서들을 클래스 개수 k개로 클러스터링한다. 앞서 구한 representations의 noise를 고려하여 PCA (Principal Component Analysis)를 적용해 최종 클러스터와 클래스를 align 시킨다. 문서와 클래스 간의 alignment를 통해 문서에 대한 pseudo labels를 생성할 수 있다. 이때 각 문서가 속한 클러스터의 확률을 이용해 가장 confident 한 예제를 선택해 pseudo training set을 구축한다. 이러한 pseudo labels를 ground truth로 놓고 Text classifier를 학습시키게 된다. 해당 논문에서는 BERT를 이용하였다.

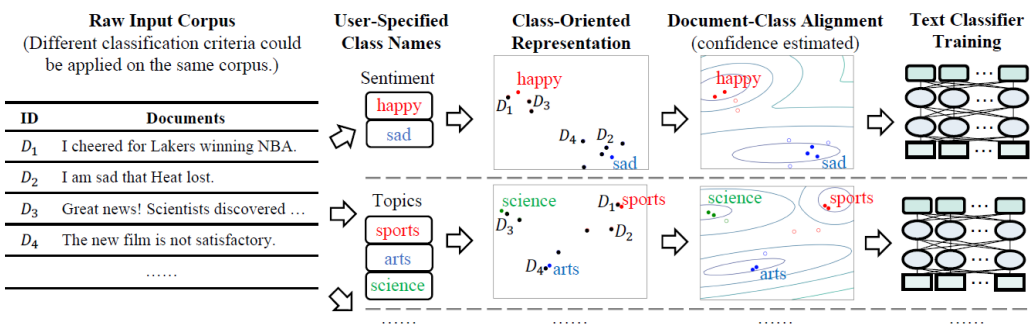

그림 3.3: Wang et al.[16] 연구에서 제시하는 모델 구조

두 모델 모두 약지도 방식이므로 레이블 없이 문서를 이해하기 위해 각 문서에 클래스 이름 자체가 등장할 때 클래스와 관련된 키워드를 추출하는 방식으로 학습이 진행된다. 따라서 클래스 이름이 문서 전체에서 한 번도

등장하지 않으면 모델이 키워드를 이해할 수 없게 된다. 감정 분석 데이터셋은 긍정/부정의 클래스로 구성되어 있지만, 실제 리뷰에는 그러한 단어가 사용되지 않는다. 이를 해결하기 위해 본 연구에서는 표 3.2에 서술된 것처럼 각 감정 분석 데이터셋의 도메인과 연관된 단어를 클래스 이름으로 치환하여 사용하였다.

표 3.2: 감정 분석 데이터셋에서 사용한 클래스 이름

데이터셋	클래스
NSMC (영화 리뷰)	재미있다 / 재미없다
Shopping (구매 리뷰)	좋아요 / 별로예요

3.2 데이터 증강 방법론

3.1절에서 살펴본 모델들은 서로 다른 방식으로 시드 정보를 이용하여 주어진 문서에 대한 pseudo training set을 생성하여 Classifier를 훈련시키는 구조이다. 본 연구에서는 이러한 기존 구조에 증강된 데이터를 효과적으로 활용할 수 있는 구조를 추가하여 성능을 개선하고자 한다.

2.2절에서 살펴본 연구들은 소량의 레이블링 된 데이터에 대한 Supervised Loss를 기준으로 레이블링 되지 않은 데이터에 대한 학습을 진행한다. 해당 학습 과정에서 구축된 Unsupervised Loss를 Supervised Loss에 추가하여 함께 학습하는 것이다. 이때 Supervised Loss를 구하는 데에 사용된 모델을 공유하여 레이블이 지정되지 않은 데이터에 대한 학습을 진행하는 ([24]) 등의 방식이 이용되었다. 본 연구에서는 이를 약지도 방식의 텍스트 분류 모델에도 적용하고자 한다. 레이블이 지정된 데이터에 대한 학습을 기준으로 이용하는 대신 pseudo label이 지정된

데이터에 대한 학습을 기준으로 삼게 된다.

먼저 CNN이나 BERT와 같은 Classifier를 이용해 pseudo label이 지정된 데이터에 대한 학습을 진행한다. 이를 통해서 Classifier의 weights를 업데이트한다.

다음으로 원 데이터에 대해 데이터 증강을 적용한다. 본 연구에서는 실제 레이블링 되지 않은 데이터를 증강한다면 class-oriented 한 증강을 할 수 없을 것이라고 가정하였다. 이러한 가정에 따라 각 데이터에 대해 임의로 증강이 진행되는 EDA[26] 방법론 (유의어 교체, 임의 삽입, 임의 교체, 임의 삭제)을 이용했다. 이렇게 레이블이 지정되지 않은 원 데이터셋과 증강된 데이터셋을 통합한 데이터셋을 $\mathcal{D} = \{\mathcal{D}^{unlabeled}, \mathcal{D}^{augmented}\}$로 표현한다.

업데이트된 Classifier를 해당 데이터셋 $\mathcal{D}$ 에 대해 Self-train 하여 변화에 강건해질 수 있게 하고, 학습 데이터의 크기가 증가함에 따라 성능이 향상된다. 이를 통해 기존의 약지도 분류 모델만을 사용했을 때보다 성능이 개선된 것을 확인할 수 있었다.

모델별로 살펴보면 WeSTClass[14] 모델은 3.1절의 식 (3.1)과 같이 pseudo-document와 모델의 예측값 간의 KL divergence Loss를 계산하여 Classifier를 사전 학습시킨다. 이렇게 생성된 pseudo-document 및 pseudo label은 실제 레이블값이 아니기 때문에 noisy 한 데이터라고 볼 수 있다. 따라서 본 논문에서는 이러한 영향을 완화하기 위해 식 (3.2)처럼 3개의 CNN Classifier를 앙상블 하여 각 Classifier 모델의 예측값을 평균하는 방식으로 학습을 진행한다. 이에 대한 구조도는 아래의 그림 3.4에 표현하였다.

$$Ensemble(\hat{y}) = average(y_i, i \in (0,3)) \quad (3.2)$$

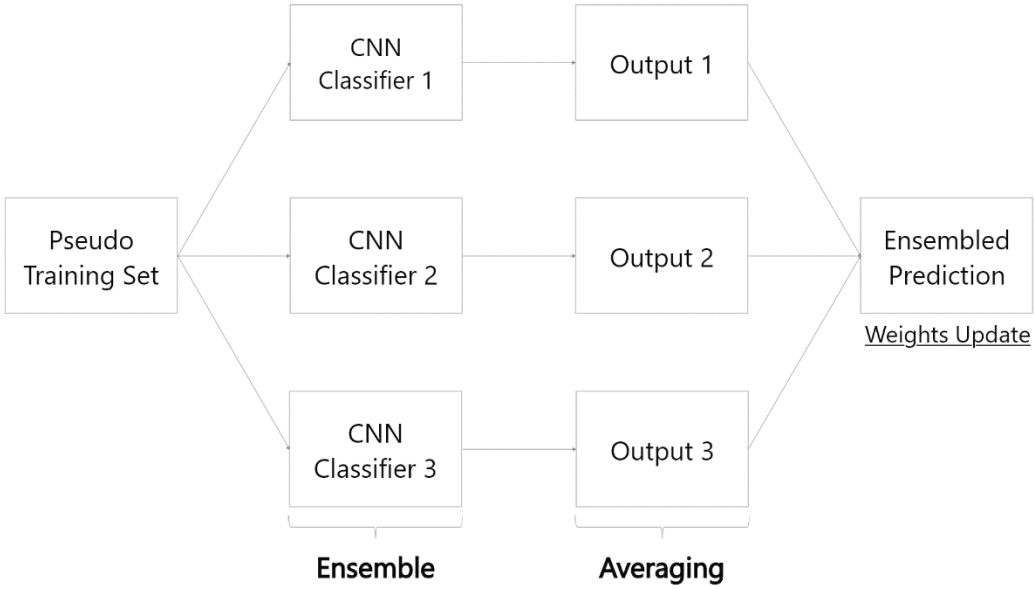

그림 3.4: CNN Classifier 앙상블 구조도

위와 같은 과정을 거쳐 업데이트된 Classifier의 weights를 이용하여 $\mathcal{D}_i$에 대해 output Y_i를 예측하고, 학습된 Classifier로 pseudo label y_i^*을 생성한다. 다음으로 사전 학습 시처럼 식 (3.3)과 같이 예측값 Y_i와 pseudo label y_i^* 간의 KL divergence Loss를 계산하여 Self-train을 진행한다.

$$loss = KL(y^* \| Y) = \sum_i y_i^* \, log \frac{y_i^*}{Y_i}, \; i \in \mathcal{D} \tag{3.3}$$

X-class[16] 모델 변형 역시 생성된 pseudo label을 ground-truth로 취급한다. Pseudo label과 BERT Classifier 예측값 사이의 Loss를 계산하는 방식으로 학습이 이루어지는 것이다. 대신 KL divergence Loss가 아닌 Cross Entropy Loss를

계산한다.

업데이트된 BERT Classifier를 이용해 전체 데이터셋 $\mathcal{D}$에 대해 output Y_i를 예측하고, X-class 모델을 활용해 pseudo label y_i^*를 도출한다. 식 (3.4)와 같이 예측값 Y_i와 pseudo label y_i^* 사이의 Cross Entropy Loss를 계산하여 학습을 진행하게 된다.

$$CE\ Loss = -\sum_{i=1}^{k} y_i\, log(P_i)\,, k\ classes \tag{3.4}$$

지금까지 서술한 과정을 통해 모델을 개선함으로써 성능이 향상된다. 아래의 그림 3.5에 본 과정을 표현하였다.

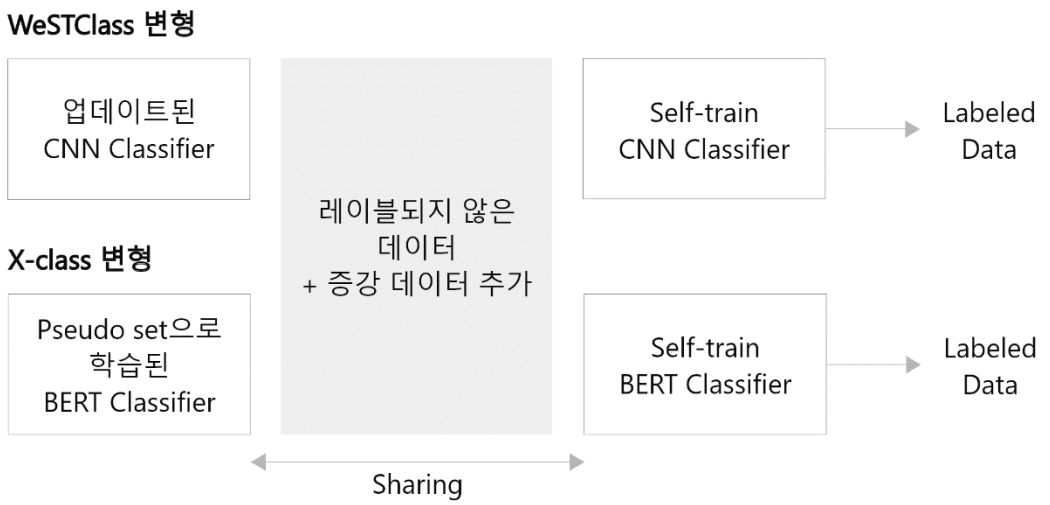

그림 3.5: 증강 데이터를 이용한 Classifier Self-train 과정

제 4 장 실험 및 결과

4.1 실험 데이터

본 연구에서는 다양한 분야에 사용되는 텍스트 분류 중 토픽 분류와 감정 분석 두 가지 분류 과제를 선택하였다. 실험의 다양성을 위하여 동일 과제 내의 각 데이터셋이 다른 도메인에서 구성될 수 있도록 선정하였다. 토픽 분류를 위해서는 KLUE Benchmark[5] 중 Text classification 목적으로 구축된 KLUE-TC 데이터셋과 AI-Hub에서 제공하는 도서자료 요약 데이터셋[27]을 가공한 데이터셋을 사용했다. 감정 분석을 위해서는 Naver Sentiment Movie Corpus (NSMC)[6] 데이터셋과 Naver Shopping Review[7] 데이터셋을 활용했다.

먼저 토픽 분류 데이터셋에 대해 서술하고자 한다. KLUE (Korean Language Understanding Evaluation)[5]는 한국어 언어 모델의 NLU 능력을 평가하기 위해 구축된 벤치마크이다. 문장 유사도 (Sentence Textual Inference; STS), 개체명 인식 (Named Entity Recognition; NER) 등 다양한 목적의 데이터셋을 제공한다. 그중 KLUE-TC는 2016.1~2020.12 기간 동안의 연합뉴스 온라인 기사의 헤드라인을 수집하여 7개의 클래스로 분류한 데이터셋이다. 현재 Training과 Test 데이터셋을 제공하고 있어 이를 합하여 이용했다. 또한 KLUE-TC는 전체 7개 클래스 (세계, 스포츠, 정치, IT/과학, 경제, 생활문화, 사회)로 구성되어 있다. 그러나 '생활문화' 클래스와 같이 한 클래스 내에 여러 주제의 데이터들이 포함된 경우도 존재한다. 따라서 본 논문에서는 실험의 단순성을 위해 데이터 규모와 내용이 비슷한 세계, 스포츠, 정치, IT/과학 4개의 클래스 (31,365개)를 사용했다.

표 4.1: KLUE-TC 데이터셋

토픽	개수
세계	9,155
스포츠	8,320
정치	8,101
IT/과학	5,789
경제	7,466
생활문화	17,100
사회	8,834
전체 합계	54,785
4개 클래스 합	31,365

AI-Hub 도서자료 요약 데이터[27]는 다양한 주제의 한국어 도서 원문 데이터로부터 추출한 20만 개의 문단과 각 문단에 대한 요약문 20만 건으로 구성된 데이터셋이다. 생성 요약 모델을 위해 구축된 데이터셋이지만 요약문과 주제를 추출한다면 토픽 분류 데이터셋으로 활용할 수 있을 것으로 고려되어 선택하였다. 실험을 위해 데이터 중 주제 분류 (해당 원문의 KDC 분류명)와 요약문 (원문 문단에 대한 생성 요약)을 추출해 분류 목적의 데이터셋을 구축했다. 사회과학, 기술, 철학, 법학 등 여러 주제로 이루어져 있으나 이 중에서 유사한 주제들을 묶어 법률, 교육, 예술, 과학 4개의 클래스의 데이터셋으로 구성했다.

표 4.2: AI-Hub 도서 요약 데이터셋

토픽	개수	내용
법률	12,851	법무 및 검찰, 법학
교육	13,445	교육일반, 유아 및 초·중등교육
예술	8,735	공연예술 및 매체예술, 예술, 문화예술
과학	7,921	과학기술진흥, 과학기술연구
합계	42,952	

다음으로 감정 분석 데이터셋이다. Naver Sentiment Movie Corpus (NSMC)[6] 데이터셋은 '네이버 영화'에서 스크래핑한 한국어 영화 리뷰 데이터로 140자 이하의 짧은 리뷰들로 구성되었다. 전체 10점 평가 기준에서 5~8점의 중립 리뷰는 제외되었고, 9~10점의 긍정 리뷰와 1~4점의 부정 리뷰만 포함되어 있다. 긍정, 부정 각각 약 10만 개 정도로 총합 20만 개 리뷰로 구성된 데이터셋이다.

Naver Shopping Review[7] 데이터셋은 네이버 쇼핑에서 2020.6~2020.7 기간의 제품별 리뷰와 별점을 수집한 데이터셋이다. 별점 3점의 중립 리뷰는 제외하고 긍정 (별점 4, 5점)과 부정 (별점 1, 2점) 리뷰만을 레이블링하였다. 데이터 크기는 약 10만 개 정도로 총합 20만 개의 리뷰로 이루어졌다. 아래의 표 4.3을 통해 전처리 적용 후 각 감정 분석용 데이터셋의 크기를 요약하였다.

표 4.3: 감정 분석용 데이터셋 요약

구분	NSMC	Naver Shopping
긍정	97,446	99,953
부정	97,678	99,955
합계	195,124	199,908

* 전처리 적용 후 기준

토픽 분류 데이터셋과 비교했을 때 감정 분석 데이터셋은 상대적으로 크기가 크다. 따라서 유사한 기준에서 비교하기 위해 감정 분석 데이터셋의 일부를 샘플링하여 초기 실험을 진행했다. 표 4.4를 통해 실험에 사용한 전체 데이터셋의 개요 및 전체적인 요약을 확인할 수 있다.

표 4.4: 실험 데이터셋 요약

구분	KLUE-TC	AI-Hub	NSMC	Shopping
분류 기준	토픽 (뉴스 제목)	토픽 (도서 요약)	감정 (영화 리뷰)	감정 (쇼핑 리뷰)
클래스 개수	4	4	2	2
클래스	세계, 스포츠, 정치, 과학	법률, 교육, 예술, 과학	재미있었다, 재미없었다	좋아요, 별로예요
개수	31,365	42,952	49,088	50,000
평균 길이	7.36	27.84	7.73	9.98

각 데이터셋은 다른 분야의 텍스트로 구성되어 있으므로 데이터셋마다 다른 특성을 보인다. 표 4.5를 통해 각 데이터셋 별로 클래스 2개씩을 선택하여 예시를 정리하였다.

표 4.5: 데이터셋 예시

데이터셋	예시
KLUE-TC	Class: 스포츠 Text: FA 최대어 정지석·문성민·양효진 원소속팀 잔류로 가닥
	Class: 세계 Text: 사우디 이라크 스포츠도시 건설에 1조원 지원
AI-Hub	Class: 법률 Text: 특허와 무관한 물건이나 포장 또는 광고에 허위로 특허표시를 하거나, 이를 양도 또는 대여하는 행위는 모두 처벌 대상이다. 허위표시죄는 공익을 해치는 행위로 비친고죄이다.
	Class: 교육 Text: 학교급과 무관히 학생들의 진로정보 수요가 다양한 영역에 분포돼있고 대학 및 직업 경로까지 요구되며 그 필요도가 70%이상이다. 이는 학생 미래 설계에 진로정보의 통합적인 제공이 필요함을 시사한다.
NSMC	Class: 긍정 (재미있다) Text: 지금까지 본 영화 중 마음이 가장 따뜻해지는 영화
	Class: 부정 (재미없다) Text: 목소리 보이지는 않지만 이건 아니다 싶다 너무 어색하고 모든 면에서 독창성이 없다
Shopping	Class: 긍정 (좋아요) Text: 싱싱해서 좋네요. 맛도 좋고 아주 최상품입니다. 다음에 또 구매 예정입니다
	Class: 부정 (별로예요) Text: 외장하드가 작동 안 하네요 환불 교환처리 귀찮아서 못했어요 지인걸로 해보면 작동되는데 케이블 불량이에요

4.2 실험 설계

WeSTClass[14], X-class[16] 두 모델에 대해서 실험은 크게 데이터 증강을 적용하지 않은 원 모델과 데이터 증강 방법론을 적용한 모델에 대한 실험으로 나누어 진행했다.

먼저 데이터 증강은 EDA[26] 방법론 (유의어 교체, 임의 삽입, 임의 교체, 임의 삭제)을 사용하여 진행했다. 원 데이터셋과 증강된 데이터셋의 비율이 1:8이 되도록 구성하였다. 즉, 원 데이터셋과 증강된 데이터셋을 합한 전체 데이터셋은 원 데이터셋 크기 대비 9배가 되는 것이다. 증강 시에는 표 4.6과 같이 2단계로 나누어서 (1) 원 데이터셋에서 증강 데이터셋을 구성하고, (2) 해당 데이터셋에서 다시 증강하여 최종 데이터셋을 구성했다.

표 4.6: 데이터 증강 설정

구분	상세 내용
1단계	원 데이터 문장에 대해 유의어 교체, 임의 삽입, 임의 교체, 임의 삭제 (4배)
2단계	원 데이터셋과 증강된 데이터셋을 합한 전체 데이터셋 (5배) 중 랜덤으로 샘플링하여 선택된 문장에 대해 유의어 교체, 임의 삽입, 임의 교체, 임의 삭제 (4배)

다음으로 한국어에 맞게 모델을 변환하였다. WeSTClass[14] 모델 실험에서는 전체 framework 중 단어 임베딩을 계산하는 부분에서 형태소 분석 단계가 필요하다. 한국어 형태소 분석을 위해 Okt, 꼬꼬마 (Kkma), 코모란 (Komoran) 등 여러 Tokenizer를 시도해보았는데, Tokenizer에 따른 유의미한 결과 차이는 발견하지 못했다. 실험에는 그중 가장 성능이 좋았던 KoNLPy 패키지의 Okt Tokenizer를 사용했다. 문장 단위 분리를 위해서는 KSS 패키지의 sentence split을 이용했다. 불용어 (Stopword)로는 Ranks NL 사이트의 'Korean Stopwords' 목록[28]에 데이터셋 전처리 과정에서 파악된 별도 불용어 단어를 추가하여

사용했다.

X-class[16] 모델 실험에서는 사전 학습된 언어 모델로 'bert-base-multilingual-cased' 모델을 활용했다. 이 외에 모델 학습을 위해 적용한 변수들은 아래와 같다.

표 4.7: 모델별 변수 설정값

모델	변수	설정값
WeSTClass	Pretrain epochs	20, 30
	Learning rate	1e-3
	Optimizer	SGD
X-class	Train epochs	3
	Learning rate	1e-5
	Optimizer	AdamW

다음으로 WeSTClass[14] 모델은 레이블 이름 외에도 레이블과 관련된 시드 단어 (키워드)와 레이블링 된 문서 약 10개에 대한 정보가 필요하다. 감정 분석 데이터셋의 레이블은 3.1절에서 서술한 것처럼 리뷰 데이터의 속성에 따라 영화 리뷰는 '재미있다/재미없다'로, 쇼핑 리뷰는 '좋아요/별로예요'로 변경하였다.

키워드를 선택할 때는 최대한 간편하게 모델을 활용할 수 있게 하도록 각 클래스에 대해 단순 빈도수 기반으로 키워드를 선택하였다. 키워드 개수에 따른 성능을 분석하기 위하여 1, 3, 5개를 선택해 실험을 진행했다. 또한, 키워드 선택에 따른 모델 성능을 파악하기 위해 키워드 3개 설정 기준으로 서로 다른 두 개의 세트를 선정했다. 각 클래스별 키워드와 관련된 상세한 내용은 표 4.8에서 확인할 수 있다. 이때 맨 앞에 있는 단어가 키워드 1개 선택 시 모델에 주어진 키워드이고, 키워드 5개 선택 시에는 괄호 속의 단어까지 포함하면 된다.

마지막으로 레이블링 된 문서는 10, 20개를 랜덤 샘플링해 각 경우에 대한 결과를 확인해보았다.

표 4.8: 클래스별 키워드

데이터셋	클래스	키워드
KLUE-TC	과학	(1) 개발, AI, 기술 (5G, 인공지능) (2) 개발, 5G, 인공지능
	스포츠	(1) 감독, 월드컵, 선수 (아시안게임, 시즌) (2) 감독, 아시안게임, 월드컵
	세계	(1) 이란, 트럼프, 시리아 (총리, 홍콩) (2) 이란, 트럼프, 총리
	정치	(1) 대통령, 정부, 국회 (북한, 김정은) (2) 대통령, 북한, 정부
AI-Hub 도서 요약	법률	(1) 범죄, 법적, 피해자 (대법원, 사건) (2) 범죄, 피해자, 대법원
	교육	(1) 학교, 학생, 학습 (교사, 교육과정) (2) 학교, 학생, 교사
	예술	(1) 문화, 영화, 미디어 (애니메이션, 콘텐츠) (2) 문화, 애니메이션, 영화
	과학	(1) 연구, 기술, 바이오 (데이터, 특허) (2) 연구개발, 과학기술, 데이터
NSMC	재미있다	(1) 좋은, 아름다운, 감동 (명작, 최고의) (2) 최고의, 좋은, 아름다운
	재미없다	(1) 아깝다, 지루하다, 뻔한 (어설픈, 최악) (2) 최악의, 아깝다, 지루하다
Shopping	좋아요	(1) 재구매, 빠르고, 만족합니다 (가성비, 저렴하게) (2) 재구매, 가성비, 만족합니다
	별로예요	(1) 반품, 아쉽네요, 사지 마세요 (불편, 비추) (2) 반품, 불편, 비추

마지막으로 실험 결과를 평가하기 위해서 분류 실험에서 주로 활용되는 F1-macro score, F1-micro score 두 개의 지표를 사용했다. 토픽 분류 데이터셋의 경우 데이터셋이 불균형한 경우가 있기 때문에 정확도 (Accuracy)가 아닌 F1 score를 활용했다.

F1 score는 (4.1)과 같이 Precision과 Recall의 조화평균이다. F1-macro score는 식 (4.2)에 서술된 것처럼 k개의 각 클래스에 대한 F1 score를 계산하고 이를 산술평균한 것이다. F1-micro score는 클래스별로 F1 score를 구하는 것이 아니라 전체 클래스에 대한 F1 score를 계산한 것이다.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4.1)$$

$$F1 - macro\ score = \frac{1}{k} * \sum_{i=1}^{k} F1\ score_i \quad (4.2)$$

4.3 실험 결과 및 분석

본 절은 증강된 데이터 확인 및 데이터 증강을 적용하지 않은 원 모델에 대한 결과와 데이터 증강을 적용한 결과에 대한 분석으로 구성된다. 원 모델에 대한 실험으로는 적합한 사전 학습 에포크 (pre-train epochs) 탐색 실험과 레이블, 키워드, 일부 레이블링 된 문서 중 가장 성능이 좋은 Supervision source 탐색, 데이터셋에 따른 성능 분석을 위한 실험으로 나뉜다. 마지막으로 증강된 데이터를 활용해 본 논문에서 제안하는 방법론의 성능을 확인한다. 추가로 제안된 방법론을 적용하여 달라진 예측 결과 일부의 예시를 확인하여 분석하고자 한다.

본 논문에서는 EDA[26]를 활용하여 데이터를 증강했는데 증강된 데이터 예시를 표 4.9에 제시하였다. 예시를 보면 임의로 단어를 삭제하거나 순서를 교체하는 과정에서 어색한 문장이 발생했다. 그러나 의미적으로는 크게 다르지 않고 레이블이 보존된 데이터가 생성된 것을 확인할 수 있다.

표 4.9: EDA[26]를 활용해 증강된 데이터 예시

데이터셋	증강된 데이터 예시
KLUE-TC	[과학] AI 기술기업이 목표…엔씨소프트 자체 브랜드 자 만든다 [스포츠]월드컵 잉글랜드·독일 프로팀 스카우트 손흥민·황희찬에 관심 [정치] 靑 특사단 만찬 뒤 밤늦게 귀환 예정속보 [세계] 시리아 정부군 휴전 선언 후 반군 지역 공습 중단
AI-Hub	[예술] 문화재로 미 지정된 나머지『구급간이방』의 가치는 높으나 연구자 간의 견해 불일치 등의 문제로 인해 국어학적 서지학적 연구의 진전 이후 것이 바람직하다. [과학] 3,4차 산업혁명으로 사람들의 일자리를 로봇이 대체할 것이라는 우려가 미래에 연구 보도되었다. 그러나 그것은 주관적인 의견이므로 신뢰하기 어렵다.
NSMC	[긍정 (재미있다)] 1956년 작품이라는게 또한 않을정도로 디테일하다! 안소니 퀸의 명연기 믿기지가 물론 [부정 (재미없다)] 발연기 진짜 못보겠다 도저히 이렇게 연기를 못할거라곤 상상도 못했네
Shopping	[긍정 (좋아요)] 사진상보기에는 일단은 깔끔하고 사이즈도 원하던사이즈라 딱좋네요 색상도 티비다이랑 맞췄는데 좋아요 이제집가서 잔스크래치만 확인해보면 될듯해요! 예쁜상품 저렴하게 만들어주셔서 [부정 (별로예요)] 선생님이 친절은 하신데 수업 장소가 너무 지저분해서 놀랐습니다. 깜짝 청소... 필요합니다. 쓸까 말까 하다가 써요...

먼저 제안 방법론 적용 전의 실험 결과이다. 첫 번째로 WeSTClass[14] 모델의 사전 학습 에포크에 변화를 주어 실험을 진행해보았다. WeSTClass 논문[14]에서는 토픽 분류 데이터셋 대상으로는 20 epochs, 감정 분류 데이터셋 대상으로는 30 epochs로 설정하여 CNN Classifier를 학습시켰다. 본 논문에서는 Supervision source는 변경하지 않고, 모든 데이터셋에 대해 20, 30, 50, 100 epochs로 나누어 실험을 진행했다. 실험 결과는 아래의 표 4.10과 그림 4.1에서 확인할 수 있다. 표에서 볼드는 가장 좋은 결과, 밑줄은 해당 분류에서 가장 좋은 결과를 의미한다. 그래프는 각 데이터셋과 Source에 대해 epoch별로 F1-macro score를 표현했다.

표 4.10: Pre-train Epochs에 따른 WeSTClass[14] 모델의 F1 score (macro/micro)

Epochs / Supervision source		Topic		Sentiment	
		KLUE-TC	AI-Hub	NSMC	Shopping
20	Labels	0.441/0.495	0.496/0.530	**0.692/0.693**	**0.856/0.856**
	Keywords	**0.790/0.788**	0.620/**0.640**	0.634/0.638	0.838/0.838
	Docs	0.771/0.771	0.624/0.640	0.687/0.688	0.616/0.616
30	Labels	0.429/0.479	0.472/0.497	0.683/0.684	0.852/0.852
	Keywords	0.782/0.778	0.615/0.635	0.665/0.665	0.839/0.840
	Docs	0.782/0.779	0.608/0.619	0.687/0.688	0.615/0.615
50	Labels	0.429/0.473	0.487/0.522	0.674/0.675	0.848/0.848
	Keywords	0.790/0.787	**0.624**/0.633	0.663/0.663	0.834/0.834
	Docs	0.777/0.774	0.615/0.634	0.673/0.686	0.604/0.604
100	Labels	0.414/0.459	0.477/0.505	0.665/0.667	0.843/0.845
	Keywords	0.779/0.778	0.614/0.623	0.634/0.638	0.837/0.837
	Docs	0.774/0.771	0.608/0.619	0.687/0.688	0.594/0.594

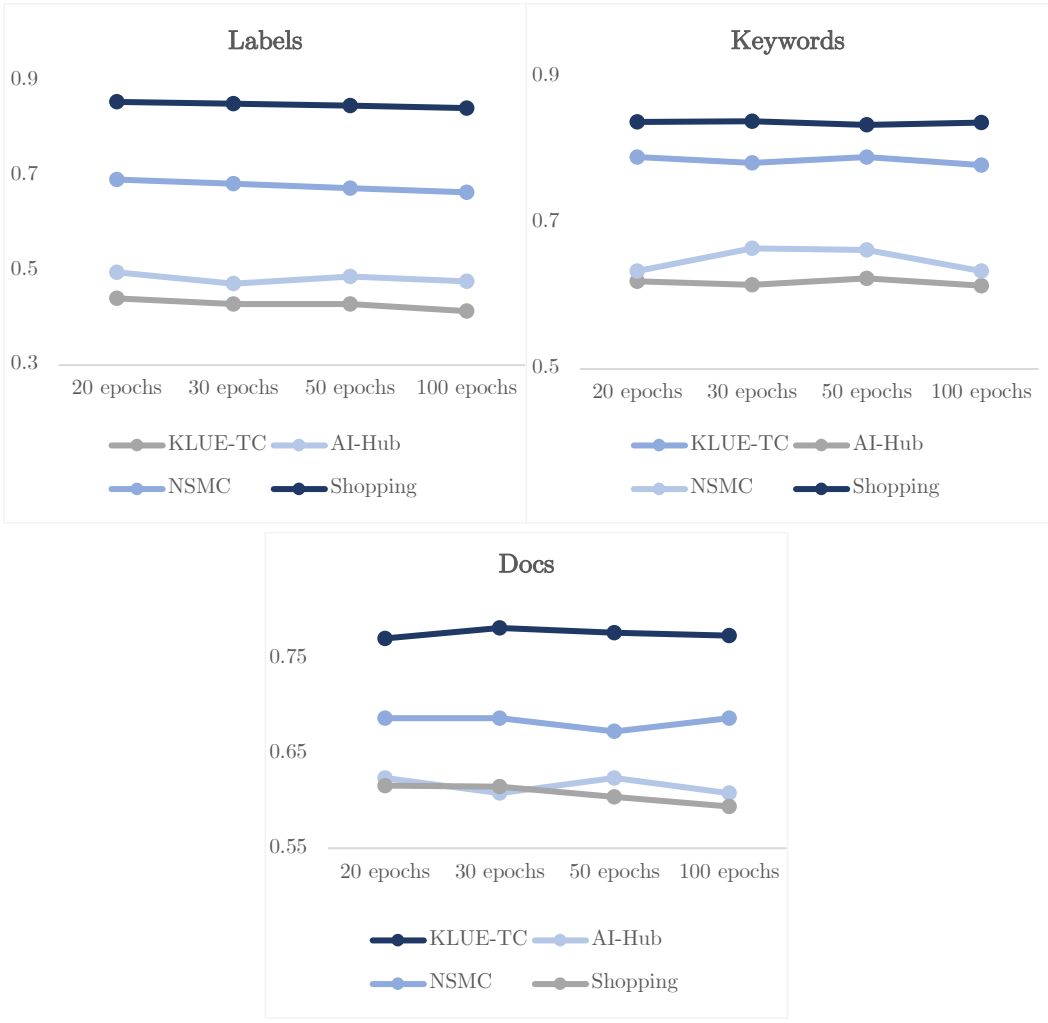

그림 4.1: Pre-train Epochs에 따른 WeSTClass[14] 모델의 F1-macro score
(위: (좌) Labels, epoch별 F1-macro, (우) Keywords, epoch별 F1-macro,
아래: Docs, epoch별 F1-macro)

결과를 보면 epoch가 늘어남에 따른 성능 향상을 크게 확인할 수 없었고, 대부분은 epoch가 20이거나 30일 때 가장 좋은 성능을 보였다. 다만 도서 요약 데이터셋의 경우에만 keyword를 추가 supervision source로 활용 시 예외적으로 50 epochs일 때 가장 좋은 성능을 보였다. 따라서 epoch를 늘려도 성능이 크게

향상되지 않는다고 판단할 수 있다. 이에 따라 다음의 실험들은 epoch 20과 30으로 진행한 후 좋은 성능을 선택하여 비교했다.

두 번째로 WeSTClass[14] 모델의 Supervision source에 따른 실험 결과를 분석했다. 이 중 제일 좋은 성능을 보이는 source를 선택하여 추후 실험도 이 기준에 따라 진행하였다. 4.1절에서 서술한 각 데이터셋에 대해 F1-macro, F1-micro score를 계산했고, 아래의 표 4.11에서 상세 결과를 확인할 수 있다. 볼드는 전체 source 중에서 가장 좋은 결과, 밑줄은 해당 분류에서 가장 좋은 결과를 의미한다.

토픽 분류 데이터셋의 경우 레이블 이름만을 활용하는 것보다 키워드를 함께 활용하는 것이 월등히 좋은 성능을 보였다. 특히 KLUE-TC의 경우 매우 큰 차이로 개선되었음을 확인할 수 있다 (F1-macro score 0.441 → 0.793). 그러나 감정 분석 데이터셋은 레이블과 키워드를 사용하는 것의 차이가 크게 나타나지 않았다. 그 이유로는 먼저 리뷰 데이터셋은 레이블의 수가 두 개이기 때문에 모델이 예측하기가 쉽다는 이유가 있다. 또한 구어체 데이터이므로 자주 사용하는 단어의 수가 보다 제한적일 것이라고 추론해볼 수 있다.

다음으로 키워드 개수를 1, 3, 5개로 변화시켜서 실험을 해보았다. 차이가 거의 나지 않는 데이터셋도 있지만 모든 데이터셋에서 3개일 때 가장 우수한 성능을 나타냈다. 이러한 결과에는 본 연구에서 단순 빈도수 기반으로 단어를 선정한 것이 영향을 미쳤을 수 있다. 빈도수로 단어를 선택하다 보니 클래스의 핵심을 담은 단어가 포함되지 않았을 가능성도 있기 때문이다. 키워드 3개 기준으로 키워드로 제공하는 단어를 변경했을 경우 단어의 선택에 따라 성능이 달라지는 것 역시 확인할 수 있었다.

레이블링 된 문서를 사용하는 경우 토픽 분류 데이터셋에서는 레이블만 사용하는 것보다 성능이 개선되었다. 그러나 키워드를 이용하는 것보다 성능 개선 정도가 작았다. 또한, 참조하는 레이블링 된 문서의 개수가 클수록 우수한 성능을 보였지만 이는 준지도 방식과 유사하므로 다음 실험부터는 적용하지 않았다.

표 4.11: Supervision source에 따른 WeSTClass[14] 모델의 F1 score (macro/micro)

		Topic		Sentiment	
		KLUE-TC	AI-Hub	NSMC	Shopping
Labels		0.441/0.495	0.496/0.530	**0.692/0.693**	**0.856/0.856**
Keywords	n=1	0.783/0.782	0.571/0.586	0.606/0.608	0.809/0.810
	n=3(1)	0.790/0.788	0.620/0.640	0.665/0.665	0.838/0.838
	n=3(2)	**0.793/0.791**	**0.632/0.650**	0.673/0.673	0.855/0.855
	n=5	0.783/0.781	0.607/0.624	0.660/0.665	0.853/0.853
Docs	n=10	0.771/0.771	0.624/0.640	0.687/0.688	0.616/0.616
	n=20	0.786/0.783	0.632/0.644	0.688/0.688	0.702/0.703

세 번째로 데이터셋의 차이 관점에서 모델의 성능 차이를 살펴보았다. 표 4.12를 보면 AI-Hub의 도서 데이터가 다른 데이터에 비해 상대적으로 평균 텍스트 길이가 긴 것을 확인할 수 있다. 해당 데이터에 대해서는 Contextualized representation이 보다 수월한 BERT를 이용하는 X-class[16] 모델의 성능이 가장 우수했다. 반면 X-class[16] 모델은 텍스트 길이가 10을 넘지 않는 다른 데이터셋에 대해서는 성능이 높지 않았다. 이를 통해 텍스트가 길수록 시드 정보 없이도 BERT를 이용해 class-oriented 한 특성을 잘 추출할 수 있다고 추론할 수 있다.

또한 추가 정보에 따른 결과도 다르게 나타났다. 먼저 레이블 이름만을 이용하는 X-class[16]와 레이블 이름만을 추가 정보로 이용하는 WeSTClass[14] + Labels를 비교해보면 토픽 분류 데이터셋의 경우에는 X-class[16]가 WeSTClass[14]보다 우수한 성능을 보였다. 이에 비해 감정 분석 데이터셋에 대해서는 WeSTClass[14]가 더 좋은 결과를 나타냈다. 다음으로 텍스트 길이가 긴 AI-Hub 도서 자료 데이터셋을 제외하고는 키워드를 추가 정보로 받은

WeSTClass[14]가 가장 좋은 결과를 보였다. 이를 통해 데이터셋의 특성에 따라 적합한 모델과 그에 따른 성능이 다르게 나타난다는 것을 확인할 수 있다.

표 4.12: 데이터셋에 따른 모델 성능 비교 (F1-macro/micro score)

	Topic		Sentiment	
	KLUE-TC	AI-Hub	NSMC	Shopping
Avg. length	7.36	27.84	7.73	9.98
WeSTClass + Lables	0.441/0.495	0.496/0.530	**0.692/0.693**	**0.856/0.856**
WeSTClass + Keywords	**0.793/0.791**	0.632/0.650	0.673/0.673	0.855/0.855
X-class	0.576/0.579	**0.744/0.735**	0.576/0.636	0.680/0.681

WeSTClass[14]와 X-class[16] 두 모델 모두 클래스를 효과적으로 분류하기 위해 주어진 레이블 이름에 대해 연관 키워드를 추출하는 방식으로 학습을 진행한다. 표 4.13을 통해 WeSTClass[14] 모델이 추출한 키워드 예시를 작성하였고, 표 4.14에는 X-class[16] 모델에서 추출한 키워드 일부를 서술하였다.

각 모델의 예시를 확인해보면 해당 클래스와 연관된 키워드를 모델이 잘 추출하고 있음을 확인할 수 있다. 다만 키워드를 추출하는 방식이 다르기 때문에 다음과 같은 차이점을 보인다. 먼저 표 4.13을 통해 확인할 수 있는 WeSTClass[14] 모델은 해당 클래스와 유사한 의미의 단어들도 함께 추출한다. 반면 표 4.14에 나타난 X-class[16] 모델은 클래스 명이 포함된 키워드 위주로 추출한다.

표 4.13: WeSTClass[14] 모델을 통해 추출된 키워드 예시

모델	클래스	키워드
KLUE-TC	과학	'네이버', '인공지능', 'AI', '개발', '차세대', '서비스', 'KIST', '스마트', '카카오', '솔루션', '구글', '블록체인', '네트워크', ...
	스포츠	'대표팀', '아시안게임', '여자배구', '월드컵', '손흥민', '배구', '출전', '승리', '합류', '선수', '남자배구', '복귀', '패배', '김연경'...
	세계	'미군', '억류', '혁명수비대', '리비아', '반정부', '시위대', '시위', '테러조직', '예멘', '석방', '이스라엘', '이라크', '난민', '내전', ...
	정치	'김정은', '여야', '국회', '문대통령', '정부', '원내대표', '민주', '북한', '총선', '당', '회동', '장관', '총리', '조국', '개헌', ...
AI-Hub 도서	법률	'피해자', '범죄', '기본권', '성폭력', '사생활', '프라이버시', '가해자', '법적', '방지', '발생의', '위험성', '성매매', ...
	교육	'학생', '학교', '학습', '수업', '진로', '교수', '교사', '학력', '교실', '수준별', '성취', '교과별', '흥미', '학습방법', '교과', '성취수준',...
	예술	'영화', '애니메이션', '배급', '제작', '캐릭터', '상영', '방송', '창작', '영화제작', '만화', '공동제작', '예술', '3D', '작품', ...
	과학	'혁신', '응용', '제품', '신기술', '바이오헬스', '기술개발', '첨단', '과학기술', '나노', '인지컴퓨팅', '신약', '헬스케어', ...
NSMC	재미있다	'아름다운', '여운', '감동적인', '멋진', '동화 같은', '사랑스러운', '예쁜', '마음', '대작', '웰메이드', '환상적인','훈훈해지는', ...
	재미없다	'유치하고', '산만하고', '진부하다', '어이없는', '삼류영화', '뻔하고', '쓰레기다', '낭비', '드럽게', '아깝고','최악의',...
Shopping	좋아요	'빠른 배송에', '만족스러운', '꼼꼼해요', '착한가격', '좋아하십니다', '아주좋아요', '저렴해요', '추천합니다',...
	별로예요	'개판이네요', '버릴려구요', '그냥 씀', '버립니다', '짜증남', '엉성합니다', '귀찮네요', '불량입니다', '불쾌하네요', ...

표 4.14: X-class[16] 모델을 통해 추출된 키워드 예시

모델	클래스	키워드
KLUE-TC	과학	'과학', '과학기술', '과학계', '과학자', '과학자들', '혁신기술', '과기부', '기술개발', '과제', '기초연구', '기술', '신기술', ...
	세계	세계', '세계적', '세계의', '세계최초', '국내서', '전세계', '국내', '전국', '국제적', '국제', '유럽서', '전국서', '해외서', '미국서', ...
AI-Hub 도서	예술	'예술', '예술과', '예술의', '예술에', '예술을', '예술이', '예술가와', '예술적', '예술가', '예술인', '예술계의', '문화예술의'...
	과학	'과학', '과학적', '과학의', '과학을', '과학에', '기초과학', '과학으로', '과학기술의', '과학기술이', '혁신성장', '연구기관'...
NSMC	재미없다	재미없음', '재미없어', '재미없고', '재미없는데', '재미없었다', '재미없네', '재미없다고', '재미없는', '재미없어서', ...
Shopping	별로예요	'재미있다', '재미있다고', '재미있고', '재미있었다', '재미있어', '재미있게', '재미있네', '재미있음', '재미있긴', '재미있어서',...

다음으로 Classifier Self-train 및 증강 데이터를 적용했을 때의 결과를 살펴본다. 표 4.15에 전체 결과를 작성하였고, 볼드는 해당 데이터셋에서 가장 좋은 결과를 뜻한다. 또한, 보다 확인하기 쉽게 그림 4.2를 통해 F1-macro score를 그래프로 표현했다.

Self-train과 데이터 증강을 활용한 모델이 모든 데이터셋에 대해 적용하지 않은 모델보다 성능이 개선되었다. 그러나 X-class[16]를 변형한 모델은 NSMC 데이터셋 (F1-macro score 0.576 → 0.668)을 제외하고는 개선 폭이 크지 않았다. NSMC 데이터셋은 WeSTClass[14]를 변형한 모델을 적용했을 때도 성능이 크게 개선되어서, NSMC 데이터셋에 대해 모델들이 제대로 학습하지 못했던 것으로 보인다.

반면 WeSTClass[14]를 변형한 모델은 AI-Hub 데이터셋을 제외한 모든 데이터셋에서 가장 좋은 성능을 보였다. 특히 NSMC 데이터셋에 대해 큰 개선

폭을 보였다 (F1-macro score 0.673 → 0.738). AI-Hub 데이터셋에 대해서는 X-class[16]를 변형한 모델의 성능보다는 좋지 않았지만 기존 WeSTClass[14] 모델보다는 우수한 성능을 나타냈다 (F1-macro score 0.632 → 0.741). 이러한 결과를 통해 본 방법론의 효과성을 확인할 수 있다.

표 4.15: 데이터 증강 방법론을 적용한 모델의 F1 score (macro/micro)

	Topic		Sentiment	
	KLUE-TC	AI-Hub	NSMC	Shopping
WeSTClass + Keywords	0.793/0.791	0.632/0.650	0.673/0.673	0.855/0.855
X-class	0.576/0.579	0.744/0.735	0.576/0.636	0.680/0.681
WeSTClass + Ens. + Aug.	**0.845/0.844**	0.741/0.750	**0.738/0.739**	**0.880/0.880**
X-class + Aug.	0.592/0.590	**0.778/0.770**	0.668/0.703	0.773/0.775

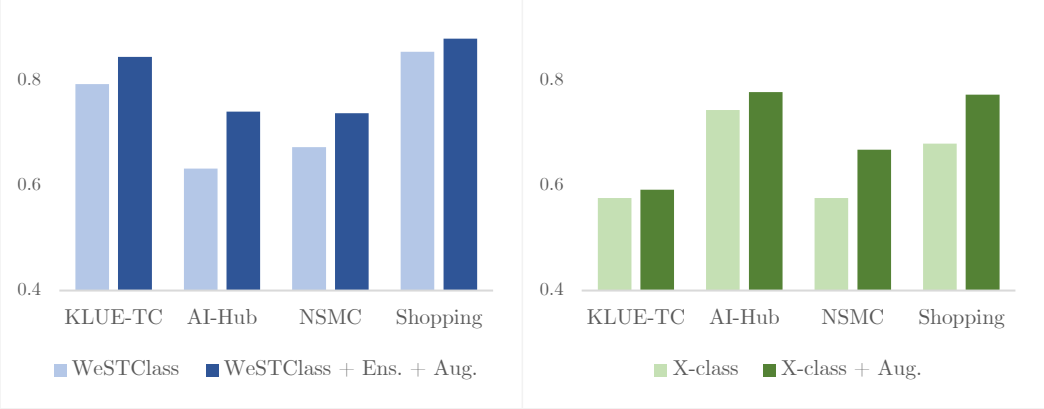

그림 4.2: 데이터 증강 방법론을 적용한 모델의 F1-Macro score
(좌: WeSTClass[14] 모델의 증강 적용 전/후, 우: X-class[16] 모델의 증강 적용 전/후)

마지막으로 실제 개선 예시를 간단하게 확인하고자 한다. WeSTClass[14] 모델로 좋은 성능을 보여주었던 KLUE-TC 데이터셋을 대상으로 했다. 해당 데이터셋의 예측 결과 일부를 본 논문에서 제시하는 방법론 적용 전에 예측된 결과와 적용 후 예측된 결과로 나누어 표 4.16에 서술하였다.

증강 전의 KLUE-TC 데이터셋 (31,365개) 중 약 12% (3,800개)의 예측이 변경되었다. 실제 예시를 보면 여러 클래스에 속할 수 있는 단어가 등장할 때 이러한 단어를 구분하는 경우가 증가하였음을 확인할 수 있다.

표 4.16: 제안 방법론을 사용하여 실제 개선된 예시

텍스트	레이블	
	기존	변경
베트남 경제 고성장 지속…2분기 GDP 6.71% 성장	스포츠	세계
美 베네수엘라 구호품 반입 촉구 안보리 결의 추진	정치	세계
트럼프 의회서 장벽예산 합의할 가능성 50% 이하	정치	세계
인니 항공당국 추락 보잉기 조종사 대화 보도 사실과 달라	정치	세계
中환구시보 美 신미사일방어전략 북미 核담판에 도움 안돼	정치	세계
2022 AG 개최지 중국 항저우로의 초대	정치	스포츠
커지는 중국 기업 채무불이행 리스크…올해 또 최고치	과학	세계
北 최부일 직책 인민보안부장→인민보안상 변경	스포츠	정치
샌더스 힐러리와 조만간 만날 것...협력 모색2보	정치	세계
KT 기가지니로 아이스크림 주문...SPC 해피오더와 연동	세계	과학
다음에 다시 만나요...아리스포츠컵 북한 선수단 출국	정치	스포츠
대통령 퇴진 4개월째 시위 수단서 저항의 상징 여성 눈길	세계	정치
블랙홀 생성비밀 풀어줄 중력파 3개 검출기서 첫 동시관측	스포츠	과학
美연준 美경제 완만한 성장 지속...단기 낙관론 유지	정치	세계

제 5 장 결론

본 논문에서는 레이블이 지정된 데이터를 활용할 수 없을 때 레이블이 지정되지 않은 데이터와 레이블 이름, 소수의 키워드만을 이용해 약지도 방식으로도 효과적으로 텍스트를 분류하는 모델을 제안하였다. 주어진 레이블에서 키워드를 추출하여 레이블이 지정되지 않은 데이터셋 대상으로 pseudo training set을 생성하는 기존의 약지도 텍스트 분류 모델 ([14], [16])을 대상으로 한다. 이러한 모델을 변형하여 먼저 Classifier를 앙상블 하는 등의 방법을 통해 pseudo training set을 학습한다. 다음으로 기존의 레이블이 지정되지 않은 데이터에 증강된 데이터를 추가하여 통합된 데이터셋 대상으로 Classifier를 Self-train 하여 Classifier의 성능을 개선한다.

이러한 방법론을 한국어로 구성된 토픽 분류와 감정 분석 각 데이터셋에 적용해보았다. 그 결과 데이터셋의 특성에 따라 키워드와 같은 추가적인 시드 정보를 활용하는 모델과 BERT 등의 사전 학습된 언어 모델을 이용하여 해당 텍스트 내의 Contextualized representation을 학습하는 모델의 성능이 다르게 나타났다. 또한, 데이터 증강 및 Self-train을 결합한 모델이 모든 데이터셋에 대해 가장 좋은 성능을 보였다. 이를 통해 본 논문에서 제시하는 방법론은 효과적으로 약지도 텍스트 분류 성능을 개선할 수 있는 방법의 하나라는 것을 확인할 수 있었다. 특히 레이블이 지정된 데이터가 부족한 한국어 연구에 크게 활용할 수 있을 것으로 기대된다.

본 연구의 한계는 다음과 같다. 첫 번째, 데이터 증강 및 Self-train에 제한된 기법을 활용하여 다른 연구와의 심도 있는 비교 분석에 한계가 있다. 증강 기법과 Classifier의 예측을 강건하게 하기 위한 기법으로 각각 EDA와 Average 앙상블을

이용했다. 두 번째, pseudo training set을 생성하는 기존의 약지도 분류 모델에 대해서만 데이터 증강을 이용하여 성능을 개선하고자 함으로써 아직 본 연구의 확장성에는 한계가 존재한다. 세 번째, 각각 토픽 분류와 감정 분석에 대해 2개의 제한된 도메인의 데이터셋만을 활용했다. 따라서 서로 다른 데이터셋에 대해 모델이 다른 성능을 나타내는 결과에 대한 원인을 파악하기 어려웠다.

향후 연구는 먼저 다양한 기법을 적용해 봄으로써 가장 적합한 조합을 탐색하여 성능을 개선할 수 있을 것이다. 예시로는 Back translation이나 Contextualized 데이터 증강 기법 등의 다양한 증강 기법과 Voting 기법 등 여러 앙상블 방법론이 있다. 다음으로 더욱 범용적인 성능 개선 방법론을 연구하여 확장성을 향상할 수 있다. 본 논문에서 활용한 약지도 분류 모델 외의 여러 방식의 약지도 모델에도 활용할 수 있는 방법론에 대한 논의가 필요할 것이다. 마지막으로 여러 데이터셋에도 공통으로 작용할 수 있는 통합적인 모델을 제안하고자 한다. 더욱 다양한 도메인과 특성이 있는 데이터셋에 적용하여 모델의 성능을 개선함으로써 데이터셋 간의 모델 성능 차이를 축소할 수 있을 것으로 기대한다.

참고 문헌

[1] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150.

[2] Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health* (pp. 246-250). IOS Press.

[3] Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics, 117*(2), 721-744.

[4] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 13*(2), 1-41.

[5] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

[6] 박은정. (2016). Naver sentiment movie corpus v1.0.

https://github.com/e9t/nsmc

[7] 이민철. (2020). 감정 분석용 말뭉치. https://github.com/bab2min/corpus/tree/master/sentiment

[8] Mekala, D., & Shang, J. (2020, July). Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 323-333).

[9] Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. (2020). Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.

[10] Chang, M. W., Ratinov, L. A., Roth, D., & Srikumar, V. (2008, July). Importance of Semantic Representation: Dataless Classification. In *AAAI* (Vol. 2, pp. 830-835).

[11] Song, Y., & Roth, D. (2014, June). On dataless hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 28, No. 1).

[12] Chen, X., Xia, Y., Jin, P., & Carroll, J. (2015, February). Dataless text classification with descriptive LDA. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).

[13] Li, C., Xing, J., Sun, A., & Ma, Z. (2016, October). Effective document

labeling with very few seed words: A topic model approach. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 85-94).

[14] Meng, Y., Shen, J., Zhang, C., & Han, J. (2018, October). Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on information and knowledge management* (pp. 983-992).

[15] Meng, Y., Shen, J., Zhang, C., & Han, J. (2019, July). Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 6826-6833).

[16] Wang, Z., Mekala, D., & Shang, J. (2021). X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp.3043-3053).

[17] Zhang, Y., Meng, Y., Huang, J., Xu, F. F., Wang, X., & Han, J. (2020, July). Minimally supervised categorization of text with metadata. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1231-1240).

[18] Mekala, D., Zhang, X., & Shang, J. (2020, January). Meta: Metadata-empowered weak supervision for text classification. In Proceedings of the 2020

Conference on Empirical Methods in Natural Language Processing (EMNLP).

[19] Zhang, Y., Shen, Z., Dong, Y., Wang, K., & Han, J. (2021, April). MATCH: Metadata-aware text classification in a large hierarchy. In Proceedings of the Web Conference 2021 (pp. 3246-3257).

[20] Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

[21] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems, 33*, 6256-6268.

[22] Chen, J., Yang, Z., & Yang, D. (2020). Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*.

[23] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

[24] Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems, 30*.

[25] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning.

Advances in neural information processing systems, 32.

[26] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv prep

[27] AI 허브 (2022). 도서자료 요약. AI 허브. https://www.aihub.or.kr/

[28] Korean Stopwords. Ranks NL. https://www.ranks.nl/stopwords/korean

Abstract

Korean Text Classification via Weak Supervision

Suyeon Lee
Department of Industrial Engineering
The Graduate School
Seoul National University

Text classification is a critical task with high utilization in various areas and tasks such as topic classification and sentiment analysis. Many models to solve this task are structured to learn classifiers using a large amount of labeled data based on supervised learning. Therefore, its application to areas where labeled data is scarce is limited. In particular, labeled data for Korean natural language processing is insufficient, making it challenging to utilize many existing models. However, since unlabeled data is easier to construct, using such data effectively for text classification is a significant problem. To tackle this problem, in this thesis, we take a weakly-supervised classification approach that classifies unlabeled data using very little information such as class names. We propose a structure that can classify Korean text by applying the widely used data augmentation methodology and self-train to improve the performance of semi-supervised learning to these weakly supervised classification models. In this thesis, we select existing models that generate pseudo labels instead of actual labels. After learning by assuming the generated pseudo labels as ground truth, self-learn is

performed on the augmented unlabeled data using the updated model. As a result of conducting experiments using the topic classification and sentiment analysis datasets, we confirm that data augmentation and self-train methodology improved performance in all datasets compared to when they were not applied.

Keywords: Text Classification, Weak Supervision, Data Augmentation, Self-train

Student Number: 2021-24560