깊은 언어 모델의 반성적 글쓰기에 대한 응용

# 초록

사회과학은 주로 전문가의 판단에 의존하는 자동화된 시스템보다는 전문 판단이 필요한 인지적으로 복잡하고 고도로 정교하거나 모호한 문제들을 다룹니다. 본 연구에서 다루는 사례 중 하나는 학생 교사의 글쓰기에서의 반성 분석입니다. 우리는 이러한 도전에 대한 기존의 경험을 공유하면서 기계 학습을 위한 데이터 수집에서 어떻게 성공적으로 대응할 수 있는지에 대해 논의합니다. 일반 언어 이해를 위해 사전 훈련된 혁신적인 딥 러닝 아키텍처를 기반으로하면 낮은 확신의 샘플에서 76.56-79.37%에서 높은 확신의 경우 97.56-100%의 정확도에 도달할 수 있습니다. 우리는 모든 자원과 모델을 오픈 소스로 공개하며 이 모델을 사용하여 대학생의 반성에 대한 이전에 미지의 가설을 분석합니다. 우리의 작업은 고등 교육 글쓰기에서 반성을 객관적으로 측정하는 도구 세트를 제공하며, 이는 100개 이상의 언어에서 적용 가능하며 정확도 손실은 0-4.2%로 측정됩니다. 깊은 모델의 탁월한 정확성 덕분에 제시된 도구 세트는 이전에 사용할 수 없었던 응용 프로그램을 가능하게 하며, 이는 준자동화된 학생 피드백 제공이나 교육 과정의 체계적인 변화의 학생 반응의 효과 측정과 같은 응용이 포함됩니다.

# 키워드

딥러닝 · 자연어 처리 · 반성 데이터셋 · 반성 분류 · 반성 일지 분석 · 일반화된 선형 혼합 모델

# 1 소개

최근 몇 년 동안 기계 학습 방법의 정확도가 상당히 증가하여 자동화의 가능성이 높아졌습니다. 이러한 추세는 특히 일반 언어 이해(GLU)를 위해 사전 훈련된 모델 덕분에 자연어 처리(NLP) 분야에서 현명한 주석자들을 능가하는 기계가 등장했습니다. 그러나 사회과학의 특수한 응용 분야에서는 혁신적인 NLP 기술의 적용이 특정 장벽에 직면하며 보급이 뒷전이 될 수 있다고 여겨집니다.

이 논문은 교육 분야에서 딥 러닝 방법을 사용한 특정 응용 분야에 중점을 두고 있습니다. 이는 특히 학생 교사를 위한 글에서 반성적인 범주를 식별하는 것입니다. 반성적 쓰기의 주요 도구 중 하나인 반성적 실천은 전문 개발에 적용 가능한 보편적인 도구입니다. 그러나 반성적 글쓰기 학습은 복잡한 주제이며 많은 도전에 직면합니다. 기계 학습 방법의 적용은 대학 환경에서 반성적 글쓰기의 사용에 직면하는 몇 가지 도전에 대한 해결책을 제공합니다.

특히 반성 식별의 경우 저 내적 일관성 및 교차 주석자 일관성의 형태로 문제가 발생합니다. 이러한 상황은 주석자 간의 내부 모델이 호환되지 않을 수 있거나 선택한 주석자를 위한 모델이 일관되지 않을 때 발생할 수 있습니다. 이러한 문제로 인해 통계 모델은 일관성 있는 주석자의 주관적이면서 가장 일관된 또는 과대 표현된 체계를 맞추기 위해 선택하거나 주석자의 다양성을 제대로 맞출 수 없게 됩니다.

우리는 반성 범주의 충분히 명확하지 않은 정의로 인한 장애일 수 있다는 가설을 확인하고자 합니다. 우리의 방법론에서는 이 가설에 대한 대응 방법을 학습하며, 주석자의 일관성을 극대화하려고 시도합니다. 우리는 주석자 투표를 결합하거나 높은 확신을 가진 샘플만 선택하는 고전적인 접근 방식 대신 매우 높은 불일치로 식별된 비대표적인 샘플을 피하고 주석자가 범주를 일관되게 선택하지 않은 샘플에 대해 토론하고 결국 합의하게 합니다.

데이터 수집 방법의 변경이 모든 최종 모델의 정확도에서 주요 이득을 제공하는 것으로 관찰됩니다. 예를 들어, 랜덤 포레스트의 경우 최상의 성능을 내는 비신경적 분류기에서 이득은 정확도의 26.65%를 나타냅니다. 우리는 결과 데이터 세트의 익명 버전을 자세히 설명하며 이를 자유롭게 사용할 수 있게 합니다.

우리는 반영성 추출을 문장 분류 문제로 정의하고, 얕은 분류기(Section 3.4.1에 소개)와 깊은 분류기(Section 3.4.2에 소개)라고 지칭하는 두 가지 주요 모델 그룹을 구별합니다. 깊은 분류기의 경우 교차 언어 설정에서 실험도 수행합니다. 이는 분류기가 특정 언어에서 반영성 식별을 훈련하지만 다른 언어에서 평가됩니다. 영어-체코 교차 언어 설정에서 우리의 다국어 모델이 Conneau et al. (2020)에 나열된 100개 이상의 사전 훈련된 언어 중 어떤 것이든 적용될 수 있음을 보여줍니다.

결과의 원활한 재현성에 상당한 관심을 기울이며, 이에 대한 자세한 지침을 프로젝트 리포지토리2(Section 3.5)에서 제공합니다. 우리는 훈련된 모델을 사용하여 관련 문헌에서 선택한 몇 가지 흥미로운 탐구적 가설을 조사하는 응용 프로그램을 진행합니다. 이들은 반영 일지의 반영 양과 학생 성적 간의 관계(Section 3.5.1) 및 학생 일지의 반영성 유형의 시간에 따른 발전(Section 3.5.2)을 조사합니다. 이러한 분석은 또한 우리의 모델이 추가 연구에서 어떻게 사용될 수 있는지를 보여줍니다. 모든 분석의 소스 및 결과물은 공개적으로 사용 가능하며 (Section 3.5)을 클릭하면 재현할 수 있습니다.

# 2 학생 교사의 반성 실천과 반성적 글쓰기

반성 실천은 주로 경험을 통한 학습의 과정으로 정의되며 미래 행위에 대한 자체 및 실천에 대한 새로운 통찰력을 얻기 위한 것입니다 (Boud et al., 1985). 이는 전문 개발을 위한 전략으로, 대부분의 교사 교육 프로그램에서 중요한 기반을 제공합니다 (Cochran-Smith, 2005). 반성 실천은 교육에서 중요한 개념이지만 수행하기는 어렵고 가르치기도 어렵습니다 (Finlay, 2008). 학생 교사들의 반성 실천을 위한 가장 중요한 도구 중 하나는 반성적 글쓰기입니다. 반성적 실천을 글쓰기를 통해 이루어지는 것은 여러 저자에 의해 기록되어 있습니다 (예: Loughran & Corrigan, 1995; Mena-Marcos et al., 2013; Moon, 2006; Bolton, 2010). 반성적 글쓰기의 일반적인 형태에는 반성적 또는 학습 일지의 구성이 포함됩니다 (Hatton & Smith, 1995; Bain et al., 2002; Ukrop et al., 2019), 포트폴리오 (Zeichner & Wray, 2001; Darling, 2001), 블로그 (Stiler & Philleo, 2003) 및 학습 네트워크 (Cardenas, 2014). 이러한 형태는 Moon (2006)이 세 가지 범주로 정리한 몇 가지 기본적인 특성을 가지고 있습니다: 좀 더 구조화된 저널, 개별, 대화 또는 공동 저널, 그리고 인쇄 또는 전자 저널 (또는 오디오 또는 비디오 테이프와 같은 다른 형태). 우리의 연구에서 반성적 글쓰기라는 용어를 구조화되지 않은 개별 전자 반성 일지로 정의합니다. 이는 학생들의 교육 실습 중에 자신의 생각과 반응을 기록하는 것입니다 (Lee, 2008). 이 범위에서 반성 일지는 반성 세미나에서 그룹 토론을 지원하는 데 사용됩니다. 이러한 세미나는 학생들의 실천에 중점을 둔 그룹 반성에 관련이 있습니다. 이 접근법은 Tan (2013)이 반성적 글쓰기를 반성 대화와 통합하고 대학 지도 교사가 감독하는 그룹 토론의 시작으로 반성적 글쓰기를 사용하는 것과 유사합니다.

# 2.1 반성적 글쓰기의 효율성

학생 교사의 반성적 글쓰기 실천은 여러 가지 중요한 이점을 가지고 있습니다. 예를 들어, Krol (1996)은 이것이 "반성을 촉진하며 학생과 교사 간의 효과적인 대화의 원천"이라고 강조하고 있습니다. 또는 반성적 글쓰기가 개인적인 인식론을 평가하기 위한 자체 평가로 사용될 수 있다고 (Hume, 2009; Lee, 2008) 언급합니다. 후자는 반성적 글쓰기가 또한 학생들의 동료 토론/평가의 기회를 제공할 수 있다고 말합니다 (Hedlund, 1989; LaBoskey, 1994; Colton & Sparks-Langer, 1993). 반성적 글쓰기는 또한 학생들이 반성을 통해 경험을 어떻게 구조화하는지에 대한 통찰력을 제공하는 데 도움이 됩니다 (Bean & Stevens, 2002; Maloney & Campbell-Evans, 2002; Wallin & Adawi, 2018). 전반적으로 이 접근법은 학생 교사들이 교사 정체성을 개발하는 데 도움이 되며 학생 교사들의 교수 및 학습에 대한 교훈적인 통찰력을 제공합니다 (Whipp et al., 1997).

이러한 결과에도 불구하고 일부 의견은 반성적 글쓰기의 효율성이 애매하다고 주장합니다. 예를 들어, Alger (2006)는 많은 반성적 활동들이 "반성을 촉진하는 데 잠재적이지만, 이러한 반성적 활동에 참여함으로써 교사가 반성적 기질이나 자세를 개발한다는 것을 보여주는 연구 증거는 거의 없다"고 언급합니다. Mena-Marcos et al. (2013)는 "고의적인 반성이 전문 지식 구축을 지원할 수 있지만 이는 드물게 발생한다"고 주장합니다. 많은 고등 교육의 연구들은 발달 심리학을 고려하지 않으며 학생들의 반성 판단도 나이에 영향을 받는다는 것을 간과합니다 (참조: King & Kitchener, 2004).

이 논의에서 체계적인 문헌 검토는 안내서가 될 수 있습니다. 이러한 연구에서는 반성적 글쓰기가 반성적 실천 및 교육 성과를 지원하는 효과적인 도구로 나타났다고 합니다. Lindroth (2015)는 학생 교사들의 맥락에서 반성적 글쓰기의 효과를 문서화하여 이를 지원합니다. Dyment와 O'Connell (2011)은 고등 교육 환경에서도 반성적 글쓰기가 효과적임을 발견했습니다.

몇 가지 모호한 점들이 있음에도 불구하고 이전 연구의 많은 양이 학생들의 반성적 사고나 능력 개발과 반성 실천 간의 관계를 지적했습니다. 그러나 여전히 연구의 개방적인 문제는 반성 실천과 교수의 실제 성과 간의 관계입니다. 반성과 성과 간의 측정이 이루어질 때 종종 자체 평가 설문 조사 (Fallon et al., 2003) 또는 관찰 (Cattaneo & Motta, 2020)를 통해 이루어집니다. 예를 들어, Cohen-Sayag와 Fischl (2012)는 연구를 기반으로 "결과는 학생 교육 수준과 반성적 글쓰기 수준 간의 링크가 명확하지 않다는 것을 나타냅니다"라고 주장합니다. 적극적인 반성적 글쓰기는 한 그룹에서 반성 수준을 향상시켰지만 그들의 교수 행위를 상응하는 수준으로 향상시키지 않았습니다. 그럼에도 불구하고 임계적 수준의 반성을 달성한 학생 교사들은 교수 행위를 향상시켰습니다. 이러한 이론적 발견을 기반으로 우리는 탐구적 가설 중 하나를 구성하였습니다 (Section 3.5.1), 즉 반성적 비율이 특정 임계값을 초과하는 글쓰기 그룹에서 멘토에 의한 학생들의 지각 성과가 이 임계값 이하의 그룹과 비교하여 현저히 높다는 것을 나타냅니다.

# 2.2 반성의 수준이 낮고 반성적 글쓰기의 촉진

이전 연구 결과는 많은 초보 교사들에게는 그들의 글에서 높은 수준의 반성을 나타내기가 어려울 수 있다는 의심을 제기하고 있습니다. (Ryken and Hamel, 2016, p. 31)는 "교사 반성에 대한 연구에서 일관된 결과는 교사 후보들 사이에서 높은 수준의 반성이 드물다는 것입니다 (Klein, 2008; Larrivee & Cooper, 2006; Lee, 2005; Mena-Marcos et al., 2013; Pedro, 2005; Shoffner, 2008; Ward & McCotter, 2004)"라고 지적합니다. 이 상황은 생성된 반성의 형태와는 독립적인 것으로 보입니다. Lepp et al. (2020)는 교사 후보의 서면 또는 비디오 저널이 가장 흔한 반성 수준인 서술적 수준과 동일하다고 설명합니다.

또한 학생들은 자신의 가정을 정당화하는 반성적 실천의 이해를 폄하하기 쉽다고 보고되었습니다 (Loughran, 2007). 이러한 상황은 retroreflection (Kolb, 2014)로 언급되며, 이러한 반성은 행동으로 녹아들기가 매우 드뭅니다. 학생이 반응적인 행동을 수행하더라도 그것은 기계적으로 이루어지며 반성 없이, 장기적인 영향이 없이 진행됩니다.

이러한 반성의 이해를 습득하는 것은 학생들과 초보 교사들에게 개발적인 관점에서 이러한 사고 방식을 적용하는 것이 어려울 수 있다는 사실과 관련이 있을 수 있습니다 (참조: King & Kitchener, 2004).

이전 관점에서, 멘토, 워크시트 또는 다른 지원 도구를 통해 반성적 실천을 촉진하는 것이 중요해 보입니다. Houston (2016)은 "일부 연구들은 지지 도구가 반성적 글쓰기를 돕는 효과적인 수단임을 나타내지만 이 분야에서는 적은 연구가 이루어졌다"고 밝혔습니다. Hanafi (2019); Pasternak과 Rigoni (2015)는 또한 반성적 글쓰기 교육이 후보 교사의 반성적 입장을 촉진하는 데 중요하다는 것을 보여주었습니다. 일부 저자들은 반성적 글쓰기 역량을 개발하기 위한 연습을 제공합니다. 예를 들어, 학생들은 서술적 수준부터 비평적 수준까지 서로 다른 세 가지 텍스트를 분석합니다. 작업은 반성과 관련된 것으로 보이는 문장을 밑줄 치고 공동 토론이 이어집니다 (참조: Moo, 2006; Hanafi, 2019).

멘토나 대학에서 지도하는 강사가 제공하는 지원의 일환으로는 서면 반성에 대한 직접적인 피드백을 제공하는 것이 포함됩니다. Spalding et al. (2002)은 "학생 강사에 대한 개인화된 피드백이 그들의 성장에 가장 중요하다"고 고려했습니다. 이 관점에서 학생들의 반성 글을 자동으로 평가할 수 있는 능력은 특히 대학의 많은 교사 교육 프로그램에서 완전히 개인화된 피드백이 불가능한 많은 학생들에게 필요한 개인화된 피드백을 제공하거나 편리하게 제공할 수 있는 방법으로 보입니다.

우리의 연구에서 학생들은 반성적 글에 대한 피드백을 받지 않습니다. 이 사실과 이론적 발견을 기반으로 우리는 반성 일지의 제출 간에 카테고리 수(반성의 수준을 나타내는 지표)가 변하지 않을 것으로 기대합니다 (Section 3.5.2에서 설명된 가설).

# 2.3 반성 일지 평가에서의 기계 학습 접근법

반성 일지의 평가는 주로 질적이거나 양적 내용 분석의 형태를 취해왔습니다 (Fox et al., 2019; Lepp et al., 2020; Mena-Marcos et al., 2013). 이러한 평가는 일반적으로 반성적 글쓰기의 선택된 세그먼트를 분류하는 형태로 수동으로 수행됩니다. 평가의 일반적인 목표는 반성적 글쓰기의 연구 분석이거나 학생에 대한 형성적 피드백으로 사용하는 것입니다. 언급된 대로 글쓰기의 반성 능력이나 기술을 개발하기 위해서는 학생에 대한 필수적인 개별화된 피드백이 필요합니다 (Spalding et al., 2002 참조).

반성적 글쓰기의 자동 평가를 자동화하려는 첫 번째 시도는 NLP 기술이 사회과학으로 전환되면서 나타나기 시작했습니다. 초기에는 주로 사전 기반 접근법을 기반으로 한 평가 기술을 관찰하고 있습니다 (예: Bruno et al., 2011; Chang et al., 2012; Chou and Chang, 2011; Cui et al., 2019) 또는 규칙 기반 접근법 (예: Gibson et al., 2016; Shum et al., 2017). 몇몇 열거된 연구에 대한 자세한 개요는 Ullmann (2019)에서 찾을 수 있습니다.

# 2.3.1 문헌 고찰: 방법론과 개요

다음 텍스트는 기계 학습 알고리즘을 기반으로 한 반성 일지의 자동 평가에 대한 연구를 개요합니다. 문헌 검색은 "reflective writing" 및 "machine learning" 키워드를 기반으로 하였으며, Web of Science와 Google Scholar 데이터베이스에서 검색을 수행했습니다(WOS = 18 + GS = 100 건의 연구). 연구 수집 기간은 2000년부터 2020년까지 입니다. 제목 및 초록을 검토하여 98 건의 연구를 제외하고 20 건의 연구로 결과가 나왔습니다. 이러한 논문들을 신중하게 읽은 후에는 아래에 나열된 기준을 충족하지 않는 11 건의 연구를 제거하여 관련성 있는 연구 9 건을 보유하게 되었습니다.

이러한 연구는 다음과 같은 기준을 충족합니다: (1) 논문이 반성적 글쓰기 주제에 중점을 둠; (2) 기계 학습 알고리즘, 특히 딥 러닝 알고리즘을 사용하여 반성적 범주(Classification Tasks)를 자동으로 인식하는 응용이 포함됨; (3) 주석은 인간 평가자에 의해 이루어짐. 제외 기준은 (1) 논문이 반성적 글쓰기에 중점을 두지 않고, 예를 들어 학문적 글쓰기에만 중점을 둔 경우 (García-Gorrosti- eta et al., 2018) 등; 2) 자동 인식이 비적응적인 기술, 즉 데이터 세트에 자동으로 조정할 수 없는 기술, 예를 들어 사전 기반이나 규칙 기반 접근법을 사용한 경우; (3) 데이터 세트 주석이 인간에 의해 수행되지 않고 자동으로 생성된 경우 (Beaumont & Al-Shaghdari, 2019) 등이었습니다.

우리의 개요는 두 개의 딥 러닝 기술을 사용한 Carpenter et al. (2020) 및 Ullmann (2019)을 제외하고 대부분의 연구가 자동 인식 및 예측을 위해 전통적인 기계 학습 알고리즘을 사용했음을 보여줍니다. 대부분의 연구에서 가장 성능이 우수한 알고리즘은 Random Forests(4 건의 연구)와 Naïve Bayes(3 건의 연구)입니다. 예측 성능은 일반적으로 정확도, F1 점수, 정밀도 및 리콜의 측정으로 보고됩니다. 알고리즘의 성공은 정확도 = .68–.96, 정밀도 = .52–.96, F1 점수 = .41–87의 범위를 가지고 있습니다.

반성 분류에 대한 주석 체계는 각 연구마다 고유하며 현재로서는 상호 비교를 위한 노력을 찾을 수 없습니다. 몇몇 경우에는 특정 맥락에 대한 질적 연구를 기반으로 범주가 생성되었습니다 (예: Carpenter et al., 2020); 다른 경우에는 범주가 이론적으로 유도되었습니다 (예: Jung & Wise, 2020). 이 메타 스키마 내에서 우리는 반성의 깊이에 중점을 둔 하위 범주와 반성의 범위에 중점을 둔 다른 하위 범주를 나타냅니다(Ullmann, 2019 참조). 반성의 범위 하위 범주 내에서 가장 일반적인 범주는 설명(직접적인 출현 4 건), 감정(직접적인 출현 3 건) 및 분석(직접적인 출현 3 건)입니다.

평가자 간 신뢰성은 Cohen의 𝜅(4 건의 연구), Krippendorff의 가중치 없는 𝛼(2 건의 연구) 및 군간 상관 관계(2 건의 연구)와 같은 표준 변수로 표현됩니다. 반성 데이터 세트는 대부분 대학에서 비롯되었지만 교사 또는 중등학교 학생이 반성의 저자로 참여하는 경우도 있습니다. 주석된 데이터 세트의 크기는 301에서 10,000 문장까지 다양하며(M = 2,954.56, Mdn = 1,966, SD = 2,943.61) 제안된 분류 방법을 벤치마킹하는 테스트 데이터 세트를 선택하는 데 가장 일반적으로 사용되는 기술은 표준 교차 검증 또는 전통적인 80% 대 20% 또는 다른 비율의 분할입니다.

주석 스키마가 다양하다는 사실에도 불구하고, 우리의 작업에서는 Ullmann (2019)의 주석 메타 스키마를 따르기로 결정했습니다. 이 메타 스키마는 24개의 반성적 글쓰기 모델의 분석을 기반으로 하며, 반성과 관련된 8개 범주를 포함하고 있습니다: Reflection, Description of an Experience, Feelings, Personal belief, Awareness of difficulties, Perspective, Outcome – Lessons learned, Outcome – Future intention. 이 스키마의 이점은 체계적인 분석을 기반으로 다양한 범주를 정당화하여 주석자가 이해하기 쉽게 제공하는 것입니다. 또 다른 중요한 요소는 Ullmann (2019)의 최고 성능 모델과 비교할 수 있는 기술 사양을 제공한다는 점입니다.

# 3. 방법론

이 섹션에서는 우리의 연구 문제와 가설에 대한 배경을 설명합니다 (3.1절), 그 후에는 이에 대한 대응 조치의 단계별 개요를 제시합니다 (3.2절). 우리는 데이터 수집 프로세스에 대한 철저한 설명을 진행합니다 (3.3절). 마지막으로, 우리는 NLP 맥락에서 반성 식별의 기술적 측면과 가설을 평가하는 데 사용하는 통계적 방법에 대한 기술적 측면을 설명합니다 (3.4절 및 3.5절).

3.1 연구 문제와 가설

이 섹션은 주 연구 문제와 가설을 주 연구 문제와 가설의 형태로 정리합니다.

기계 학습 알고리즘을 통한 반성 일지의 평가에 대한 문헌 검토를 기반으로하면 동료 연구의 몇 가지 공통적인 결점을 확인할 수 있습니다. 문제의 복잡성에 관계없이 연구들은 주로 기본적이거나 직접적으로 관련이 없는 텍스트 표현으로 계속 진행됩니다. 이후에는 대부분 선형 분류기 또는 문제와 구조적 관련이 없는 신경 아키텍처를 사용하여 예측이 수행됩니다.

우리는 연구들 사이에서 점진적인 질적 발전을 확인하지 못합니다. 아마도 이는 한정된 재현성 옵션 때문일 것입니다. Ullmann (2019)을 제외하고는 관련 연구의 데이터 소스가 공개적으로 사용 가능하지 않다고 명시합니다. 게다가 어떤 연구도 동일한 반성 범주 메타 스키마를 공유하지 않으며 이는 각 연구의 질적 비교를 허용하는 것입니다. 몇몇 연구만이 실제 응용 프로그램을 통해 진행되었으며, 여기에는 Knight et al. (2020)와 Ullmann (2019)의 ReflectR 작업만 확인할 수 있습니다. 비록 관련 문헌에서 데이터에 대한 지원이 필요한 다양한 이론들이 있지만요.

우리는 문제의 의미적 복잡성을 인정하며 주로 Transformer의 딥 랭귀지 모델의 응용에 중점을 둡니다 (Vaswani et al., 2017). 그러나 이러한 모델의 품질을 문헌에서 최고로 작동하는 깊이 조절된 전통적인 접근 방식의 성능과 비교적 비평적으로 비교합니다. 우리의 중심적인 연구 문제(RQ)는 결과의 질적 측면을 다룹니다:

RQ: 자동 분류로 학생 교사의 일지에서의 반성을 어느 수준까지 달성할 수 있을까요?

**(Section 2.3.1)에서 언급된 이유로 Ullmann (2019)의 주석 체계를 선택합니다. 편리성을 위해이 체계의 원래 이름을 줄여 다음과 같은 주석이 달린 범주를 얻습니다: Reflection, Experience, Feelings, Beliefs, Difficulties, Perspective, Learning, Intention. Ullmann (2019)이 제안한 체계의 유일한 사용자 정의는 Reflection 범주와 일치하지 않는 문장을 나타내는 Others 범주의 추가입니다. 이 범주 덕분에 모델은 비반성적인 문장도 구별할 수 있어 기계 학습 모델을 사용하여 텍스트에서 반성의 상대적인 비율을 평가할 수 있습니다. 우리는 이 자동 분류를 선택된 반성 이론적 연구에 기반한 일련의 탐색적 분석에서 사용합니다. 첫 번째 분석은 학생 교사의 교수행위와 관련하여 비판적인 반성 수준에 도달하는 것만이 더 나은 결과와 관련이 있다는 주장을 기반으로 합니다 (Cohen-Sayag & Fischl, 2012), 대부분의 학생 교사가 설명 수준의 반성에만 도달한다는 주장 (Lepp et al., 2020) (2.1절 참조). 이를 확인하기 위해 가설 1 (H1)을 제시합니다.**

H1: “비판적 반성” 임계값이 존재하여 학생들의 성과에 통계적으로 유의미한 차이를 볼 수 있습니다

**교수 행위의 품질을 추정하기 위해 학생 교사의 평가를 사용하며, 이는 현장에서 학생 멘토들이 작성한 설문 조사를 통해 이루어집니다. 만족스러운 품질로 반성을 식별할 수 있는 능력 덕분에 우리는 학생들의 성과를 해당 시간에 대한 멘토의 보고서에서 볼 수 있는 저널의 통계와 관련시킬 수 있습니다.**

**두 번째 가설은 반성 작성에서만 개인 및 개인 맞춤형 지원이 더 나은 결과로 이어진다는 주장과 관련이 있습니다 (Spalding et al., 2002, p. 1393). 우리는 가설 2 (H2)를 설정하여 예상되는 상태를 반영합니다. 즉, 작성자가 지원 없이 시간이 지남에 따라 반성을 개선한다는 것입니다.**

H2: 반성 일지의 범주 수는 반성 일지 제출의 순서에 따라 달라집니다

**우리의 설정에서 어떤 학생 교사들은 반성 일지의 일련의 일지를 수집했는데, 그 동안 학생들은 자신의 글에 대한 개인적인 피드백을 받지 못했습니다. 따라서 학생들이 더 많은 반영을 위해 피드백을 필요로 한다면, 시간이 지남에 따라 범주 수에 중요한 변화가 예상되지 않을 것이며 순서에 따라 중요한 경향이 나타나지 않을 것으로 기대됩니다.**

# 3.2 방법론 개요

이 섹션은 중심적인 연구 질문과 탐험적 가설에 대한 답변을 목표로 한 연구에서 수행된 단계들에 대한 개요를 제공합니다. 이러한 단계는 데이터 수집에서 기계 학습 평가 및 분류된 반성을 기반으로 한 통계 테스트에 이르기까지 다양합니다. 각 단계는 이후 섹션에서 더 자세히 설명될 것입니다.

**반성 일지 수집**: 우리는 체코의 학생 교사들로부터 1,070개의 반성 일지를 수집했습니다. 이 중 300개의 반성 일지를 무작위로 선택했으며 동일한 저자의 일지를 우선했습니다.

**데이터 주석 - 첫 번째 라운드**: 선택된 일지는 Doccano 주석 소프트웨어(Nakayama et al., 2018)에 로드되었습니다. 6명의 주석가들은 일지에서 반성을 식별하기 위해 훈련을 받았습니다. 반성 범주의 체계는 Ullmann (2019)의 체계를 따랐습니다. 이 체계는 주석가들을 위한 자세한 매뉴얼로 운용되었습니다.

**데이터 주석 - 일관성 평가**: 각 주석가 쌍마다 상호 간 신뢰성을 평가하고 다른 질적 분석을 수행했습니다. 우리는 수집된 데이터의 품질이 추가적인 사용에 부적절하다고 양적으로 결론 내렸습니다.

**데이터 주석 - 두 번째 라운드**: 첫 번째 라운드의 주석된 집합에서 일관적인 반성 레이블에 속할 가능성이 높은 7,128개의 문장을 선택했습니다. 방법론은 전체 문장에 대한 단일 레이블의 주석으로 단순화되었습니다. 주석가들이 문장의 범주에 동의하지 않은 경우, 주석가들은 토론하고 단일 레이블에 동의하기로 요청되었습니다.

**자동 반성 식별 - 평가**: 우리는 선택된 기계 학습 분류기를 이러한 주석이 달린 문장의 90%에서 훈련시키고 독립적인 5% 보류 집합에서 평가했습니다. 나머지 5%는 하이퍼파라미터 튜닝에 사용되었습니다. 우리는 문헌을 기반으로 언어 처리를 위한 여러 고전적인 분류기 알고리즘을 선택했습니다(섹션 2.3 참조): Random Forest, Naïve Bayes, Support Vector Machine, Logistic Regression, 그리고 몇 가지 잘 수행되는 딥 모델, 다국어 분류를 지원하는 BERT (Devlin et al., 2019) 및 XLNet (Conneau et al., 2020)이 포함됩니다.

**연구 가설 - 통계 평가**: 가장 성능이 우수한 분류기를 사용하여 원래의 1,070개의 반성 일지 집합을 기반으로 한 반성에 대한 탐험적 가설을 평가했습니다.

# 3.3 데이터셋 수집 및 주석

데이터 수집 프로세스는 그림 1에 나열된 단계를 따릅니다. 체코 학생 교사들로부터 1,070개의 반성 일지(33,859문장, 반성 일지의 길이: 평균 2,817 문자, 표준 편차 1,287)를 얻었습니다(5년제 석사 과정의 4학년, n = 220). 학생들은 교육 실습 중에 1년 동안 다섯 번의 반성 일지를 작성했습니다. 이들 일지에서 학생들은 교육 실습 경험에 관한 자유로운 글을 제공하도록 요청받았습니다. 학생 교사들이 제공한 주제별 반성 일지의 분포는 다음과 같습니다: Czech (n = 112), English (122), Geography education (88), History (54), Biology education (67), Civics education (147), Physics and chemistry education (66), Music education (39), Art education (115), Russian and French (31), Special education (56), German (74), Vocational subjects (44), Mathematics education (55). 이 데이터 세트인 CEReD - Czech-English Reflective Dataset Authors(Štefánik & Nehyba, 2021)는 Github에서 사용 가능하며, 섹션 3.2의 각주 참조.

다음 섹션에서는 이러한 저널 세트에 대한 반성 주석을 수집하는 과정을 묘사하며, 데이터 수집의 이전 결과와 주석 수집 방법론과 관련된 문제에 대응한 두 번째 반복에 대한 범위를 제공합니다.

첫 번째 단계에서 주석가들에게 임의의 텍스트 단위를 주석 달도록 요청했습니다. 이 텍스트 단위의 의미가 선택된 반성 범주 중 하나에 속하는지를 결정합니다. 주석가가 해당하는 경우 여러 범주에 대해 하나의 일지 내용에 대해 할당할 수 있었으며 주석이 달린 텍스트 조각은 임의의 상호 오버레이를 가질 수 있었습니다. 그런 다음 주석이 달린 각 일지는 익명화되어 문장으로 분할되었습니다. 이는 다음 단어의 구두점과 대문자 사용을 결합한 휴리스틱스를 기반으로합니다. 그런 다음 주어진 문장의 단어의 1/2 이상을 동일한 범주에 할당한 주석가의 대다수인 경우 해당 문장을 9가지 사전 정의된 범주 중 하나에 할당했습니다. 동일한 일지에 정확히 세 가지 다른 주석가가 주석을 달도록 보장하여 매치가 없도록했습니다. 이러한 주석이 달린 문장의 자동 분류의 초기 실험을 수행한 후 보류 중인 문장의 분류기 정확도가 매우 낮다는 것을 관찰했습니다. 최고의 성능을 내는 Random Forest 분류기도 정확도의 48.85%에 불과했습니다. 이 상황의 측정된 상호 주석자 불일치가 우리의 기대를 초과하는 주석자의 혼란을 나타낸다는 것을 평가하기 위해 우리는 5명의 주석가에 대한 쌍별로 Cohen의 𝜅를 측정했습니다. 이 𝜅는 10.45%에서 50.76%까지 범위를 보여주었습니다. 앞에서 언급한 투표 전략이 주석자의 발산을 완화하는 데 미치는 효과를 평가하기 위해 우리는이 전략을 사용하여 할당된 범주를 사용하여 100개의 문장을 무작위로 선택하고보다 경험이 풍부한 전문 주석가5를 사용하여 할당된 레이블을 다시 평가했습니다. 우리는 36.6%의 할당된 주석이 전문 주석가에 의해 합의되지 않았음을 발견했습니다. 측정된 상호 주석자 불일치가 주석자 간의 상호 불일치보다는 주석자의 내부 불일치로 인한 것이라는 가설을 더하여 이 상황을 조사했습니다. 이 상황의 측정된 상호 주석자 불일치가 주석자 간의 상호 불일치보다는 주석자의 내부 불일치로 인한 것이라는 가설을 더하여 이 상황을 조사했습니다. 이러한 상황의 측정된 상호 주석자 불일치가 주석자 간의 상호 불일치보다는 주석자의 내부 불일치로 인한 것이라는 가설을 더하여 이 상황을 조사했습니다. 상황의 측정된 상호 주석자 불일치가 주석자 간의 상호 불일치보다는 주석자의 내부 불일치로 인한 것이라는 가설을 더하여 이 상황을 조사했습니다. 이러한 상황에서 얻은 측정치를 사용하여 우리는 모든 주석을 다시 수집하기로 결정했습니다. 우리는 방법론을 다음과 같이 조정했습니다. 주석자의 대다수가 참 범주에 동의한 문장 만 선택하여 가능한 한 분리되고 결정적 인 데이터 세트를 수집하도록 노력했습니다. 이러한 문장에 대해 두 가지 성과가 가장 뛰어난 주석가 두 명으로부터 중복하여 주석을 수집하도록 요청했습니다. 이 주석자가 개인 주석을 기반으로한 무작위 숲 모델의 분류 성능. 또한 주석가들에게 주어진 범주 내에서 주어진 샘플의 전형성을 7에서 1 사이의 비십진 값으로평가하도록 요청했습니다. 여기서 7은 주어진 범주의 가장 전형적인 샘플을 나타내며, 전형성의 샘플, 즉 확신 1은 주어진 범주의 가장 전형적이지 않은 샘플을 나타냅니다. 이 값을 확신이라고 지칭하며 이 확신은 평가에서 샘플을 분할하는 데 사용됩니다. 완료되면 주석자가 할당한 범주와 일치하지 않은 경우를 측정한 결과가 나왔습니다. 7,128개의 문장 중 418개의 경우에 해당하여 문장의 5.86%에 해당합니다. 주석가들에게 할당된 범주에 대한 토론을 진행하고 일치하지 않는 문장 각각에 대해 단일하고 가장 관련 있는 범주에 동의하도록 요청했습니다. 이 과정의 결과로 우리는 커뮤니티를 위해 수집하고 공개한 데이터 세트에 가장 내부 일관성이 높은 두 주석가에 의한 "전형성"의 추정값도 포함되었습니다. 데이터베이스의 일부 문장은 반복적이거나 상호 수정이 적은 것으로 나타났습니다. 이러한 이론적 가능성이 없도록 표준 보류 중 평가 세트를 제공하기 위해 문장을 알파벳순으로 정렬하고 마지막 5%만 사용합니다. 그림 3, 4 및 6에보고 된 결과는이 보류 중 문장의 정확도 또는 최소 지정 확신의 하위 집합에 대한 것입니다. 재현성을 위해 테스트 분할은 프로젝트 저장소에서 재사용 및 비교 가능하도록 공개적으로 제공됩니다(섹션 3.5 참조).

3.4 반영성 분류

이 섹션에서는 NLP 분야에서 반성 식별 문제를 찾아보고 이에 대한 해결책으로 실험하는 분류 알고리즘 및 해당 텍스트 표현에 대한 설명을 시작합니다.

우리는 반성 식별을 텍스트 분류 또는 Named Entity Recognition (NER) 문제로 다가갈 수 있다고 인식합니다. 분류 접근 방식에서는 시스템이 우리의 경우 하나의 문장으로 표현된 토큰 시퀀스를 단일 범주로 할당합니다. 반성의 표현이 반드시 단일 문장과 일치하지 않을 수 있기 때문에 전통적인 분류 접근 방식은 주어진 범주에 덜 관련된 부분을 포함할 수 있습니다. 나중에 언급된 NER 접근 방식은 일반적으로 단어와 같은 텍스트의 임의의 단위를 분류하는 것을 허용하므로 동일한 범주가 단어 시퀀스를 걸쳐 확장 될 수 있습니다. 이 작업에서는 결과의 직관성, 이전 연구와의 비교 용이성, 주석된 데이터의 양에 대한 더 낮은 요구 사항 및 주석자를 더 일관되게 주석으로 제한할 수 있는 능력 때문에 문장 분류로만 문제를 다가가기로 합니다(섹션 3.3 참조). 그러나 NER 접근 방식이 가져올 수있는 언급 된 혜택을 인정합니다. 분류된 문장 자체의 표현과 함께 시스템은 문장 자체의 반성 범주의 중요한 결정자로 판단되는 문맥을 고려할 수 있습니다. Fig. 2에서 특정 교육 샘플을 통해 설명한 것처럼 우리는 분류된 문장의 두 선행 및 하행 문장의 문맥을 포함하여 실험을 진행합니다. 이 추가 텍스트 세그먼트는 분류된 문장에서 방법 특정 방식으로 분리되며 각 분류기 유형에 대해 별도로 설명합니다. 텍스트 분류에 일반적으로 사용되는 비신경 기계 학습 알고리즘을 기반으로하는 "얕은"이라는 두 가지 분류 알고리즘 범주와 선형 분류기 스택을 사용하여 최종 예측을 구성하는 "깊은"이라는 두 가지 분류 알고리즘 범주로 실험합니다.

# 3.4.1 얕은 분류기

Ullmann (2019)의 예에 따라, 먼저 간단한 얕은 분류기로 실험을 진행합니다. 관련 연구 개요를 기반으로 다음 범주의 분류기를 선택합니다: (i) 선형 판별자를 기반으로 한 것: 서포트 벡터 머신, 로지스틱 회귀 및 로지스틱 회귀, (ii) 조건부 확률 분류기인 나이브 베이즈, (iii) Tree-based: Random Forest 분류기. Ullmann (2019)의 예에 따라 이러한 분류기들은 하나의 핫-워드 표현, 즉 문장 크기에 맞는 최적의 성능을 위해 하이퍼파라미터가 조정된 100~2000 예측기 단어 또는 토큰 범위의 단일 단어 표현을 사용합니다. 눈에 띄게 성능이 우수한 Random Forest 분류기는 문맥과 문장의 가장 빈번한 상위 800개 토큰을 포함하는 단어 가방 표현을 사용합니다. 문맥을 표현에 포함할지 여부를 실험적으로 조사합니다. 문맥이 포함된 경우 해당 문맥은 분류된 문장과 동일한 방식으로 표현되며 두 표현이 연결됩니다. 최상의 성능을 보이는 Random Forest 분류기의 경우, 문맥의 동일한 단어 표현을 포함하면 정확도가 70.8%에서 73.6%로 향상됩니다. 각 분류기의 특정 매개변수에 대한 하이퍼파라미터 튜닝을 수행하고 최상의 성능을 보이는 시스템의 정확도를 Fig. 3에 보고합니다. 모든 평가된 얕은 분류기의 정확도 개요는 부록 1에서 확인할 수 있습니다.

양식의 맨 위

양식의 맨 아래

# 3.4.2 심층 분류기

두 번째로, 우리는 Transformer 패밀리의 선택된 최첨단 신경 시퀀스 분류 모델로 실험을 진행합니다 (Vaswani et al., 2017). 이러한 분류기는 일반적인 언어 이해와 관련된 목적으로 대량의 텍스트 말뭉치에서 사전 훈련되었습니다. 예를 들어, 마스크된 언어 모델링 (MLM) (Devlin et al., 2019)이나 소음 제거(Lewis et al., 2020)와 같은 목적으로 대부분의 언어 이해에 관련된 큰 말뭉치에서 사전 훈련되었습니다. 문장 분류 문제에 대해 심층 신경 분류기를 적합시킬 때 일반적으로 사용되는 분류 문제에 대한 엔트로피 손실 목적을 사용합니다 (Cox, 1958). Transformer 모델의 잘 알려진 표현 관례를 따릅니다. 입력 문장을 SentencePiece 하위 단어의 단어 가방으로 나타냅니다 (Kudo & Richardson, 2018). SentencePiece 서브워드는 이전의 단어 가방 표현과 다르며 일부 단어는 내부적으로 서브워드로 분할됩니다. 서브워드는 단어의 굴절에서처럼 자주 발생하는 접두사 및 접미사입니다. 사용하는 서브워드 어휘가 각 모델의 사전 훈련에 사용된 어휘를 포함하도록 합니다. 얕은 분류기의 표현 외에도 Transformer는 입력 단어의 상호 순서 인코딩을 사용합니다. 신경 분류 모델에게 분류된 문장의 문맥을 제공하는 것이 중요하다고 생각합니다. 문장을 분류된 문장에서 문맥으로 이동하는 위치에 특수 토큰 "<s>"과 함께 컨텍스트 세그먼트를 추가합니다. 이는 Devlin et al. (2019)와 유사하게 수행됩니다. 문맥 창이 일관되게 분류된 문장과 함께 제공되지 않으면 신경 분류기가 수렴하기 어렵고 얕은 분류기와 비교했을 때 크게 성능이 떨어집니다. UDA (Xie et al., 2020) 데이터 증가 기술을 사용하여 문맥에 대한 필요성을 일부 제거했지만 이러한 분류기는 여전히 문맥을 사용하는 동일한 아키텍처보다 5~10%의 정확도 점수가 낮습니다. 각 모델에 대해 훈련 배치 크기, 학습률 및 에포크 수를 체계적으로 조정하지만 모델은 대부분의 시도에서 거의 일관된 성능을 보입니다. 영어 분류에 대한 성능은 Fig. 4에서 최적의 매개변수를 기반으로 검증 세트의 성능을 보고하고, 체코 분류는 Fig. 3에서 확인할 수 있습니다. 기술적인 설명의 확장 버전은 부록 1에서 참조하십시오.

양식의 맨 위

양식의 맨 아래

# 3.5 탐색적 가설의 방법론

훈련된 분류기를 사용하여 주석이 없는 새로운 반영 일지에서 반영성을 식별합니다. 이러한 식별된 반영의 특성의 분포 또는 시간적 역학을 기반으로 분석을 수행합니다. 분석은 학습 신뢰도가 4 이상인 XLM-RoBERTa 분류기를 사용하여 수행되었습니다.

각 탐색적 분석의 전체 절차가 완전히 재현 가능하도록 합니다. 익명화된 데이터 소스, 신경망 분류 모델 및 텍스트에서 반영 세그먼트를 한 줄로 추출할 수 있는 파이썬 라이브러리와 함께 각 분석은 해당 노트북에서 단일 클릭으로 실행할 수 있습니다. 각 가설에 해당하는 노트북은 Github 리포지토리에서 찾을 수 있습니다.

# 3.5.1 H1: 학생들의 성과에 통계적으로 유의한 차이를 볼 수 있는 “비판적 반성” 임계치가 존재한다

우선, 우리의 데이터셋에서 반영 일지의 세트를 해당 평가와 연결합니다. 여기서 평가는 지도 교사인 멘토가 동시에 제출한 것입니다. 학생의 성과는 멘토로부터 수집한 설문지를 통해 보고됩니다. 이 설문지는 7개의 질문으로 구성되어 있으며 6점 Likert 척도의 답변이 있습니다. 여기서 6은 최고의 성과를 나타내며 1은 최악의 성과를 나타냅니다. 설문 조사에 참여한 참가자 수는 n = 223입니다. 전체 데이터셋 및 설문지 항목은 프로젝트 리포지토리에서도 사용할 수 있습니다.

수집된 결과물에 대한 질적 평가를 수행했습니다. 이에 대한 병렬 분석은 설문지의 한 요인과 구성 요소를 시사하며, 크론바흐 알파는 0.91이고 항목 총 상관 관계는 0.66에서 0.77 사이입니다. 크론바흐 알파의 권장 크기는 0.7 이상이고 항목 총 상관 관계는 0.30 이상이어야 합니다(Nunnally & Bernstein, 1994; Cristobal et al., 2007). 따라서 우리는 설문지의 전반적인 점수를 학생 교사의 실습에서의 품질 성과의 지표로 사용할 수 있습니다. 멘토가 학생 성과를 객관적으로 평가하기 위해 최선의 노력을 기울였지만 멘토의 평가는 일부 혼동 요인에 영향을 받을 수 있습니다(멘토의 인턴십 기간, 교육 수준, 학생과의 관계 등). 우리는 각 평가의 순위를 합산하고 값을 ⟨0, 1⟩로 정규화합니다. 이 값을 학생의 성과, 즉 해당 일지를 작성할 때의 학생의 성과로 참조합니다.

일지와 해당 성과를 모두 보유한 상태에서 우리는 다음과 같이 "비판적 반성" 수준을 윤곽 지을 수 있는 상대적 반영의 값을 체계적으로 식별하기 위한 문제를 제시합니다. 우리는 학생의 성과를 가장 잘 예측하기 위해 단일 예측 변수로 반영의 상대적 비율(반영 문장 수를 전체 문장 수로 나눈 값)을 고려하고 주어진 일지와 관련된 보고된 성과를 대상 변수로 합니다. 이 프레임워크에서 우리는 회귀 의사 결정 트리를 단일 예측 변수에 맞추어 목표 또는 종속 변수를 가장 잘 예측하기 위해 사용합니다. 학생의 성과가 높은 그룹이 반영의 양에 의해 구별될 수 있다고 가정하므로 이 방식으로 "비판적 반성" 임계값을 학생 성과의 최적 분리 값으로 체계적으로 탐색합니다. 설명을 위해 최적의 상대적 반영 분리 값(X[0])을 갖는 의사 결정 트리도 Fig. 5에 시각화되어 있습니다.

의사 결정 트리를 맞춘 후에는 찾은 반영 임계값을 사용하여 해당하는 상대적 반영 수준에 따라 분리된 두 그룹의 성과가 통계적으로 다른지 여부를 검정합니다. 구체적으로는 두 그룹의 평균 성과 사이에 통계적으로 유의한 차이가 있는지 검정하는 T-테스트를 수행합니다.

또한 각 범주에 대해 별도로 프로세스를 반복합니다. 각 범주의 상대 비율을 기반으로 한 별도의 분석 결과는 프로젝트 리포지토리에서 이 분석의 노트북에서 찾을 수 있습니다.

# 3.5.2 H2: 반영 일지의 범주 수는 반영 일지 제출의 순서에 달라진다

우리는 각 범주에 대한 인과 모델 세트를 구성합니다. 이 모델 세트는 예측 변수로 시간 순서 및 저널의 길이를 사용하며 각 예측된 범주의 발생 횟수를 결과 변수로 사용합니다. 저널의 길이는 정규화에 사용되며 각 저널마다 길이 및 총 분류된 항목(문장) 수가 다르기 때문입니다.

인과 추론 모델(참조: Gelman & Hill, 2006)로써 우리는 일반화 선형 혼합 모델(GLMMs) 세트를 구축합니다(Jiang, 2017; Stroup, 2012; Faraway, 2016). GLMMs는 단순한 모델인 ANCOVA와 비교하여 몇 가지 이점을 제공합니다. 예를 들어, 결측값 처리를 위해 리스트 삭제가 필요하지 않으며 다변량 모델은 고정 효과에 대한 테스트에서 더 높은 표현력을 제공합니다(참조: Hoffman & Rovine, 2007).

우리의 접근법을 표준 표기법에 매핑하기 쉽게 하기 위해 각 i번째 범주에 대한 j개의 샘플이 있다고 할 때, 다음과 같은 GLMMs가 사용되었습니다:

category�∈{Experience, Reflection, Feeling, Difficulty, Perspective, Belief, Learning, Intention, Other}category*i*​∈{Experience, Reflection, Feeling, Difficulty, Perspective, Belief, Learning, Intention, Other} category�count=�00+�10ordering��+�20journal length+�0�+���category*i*count​=*γ*00​+*γ*10​ordering*ij*​+*γ*20​journal length+*u*0*j*​+*eij*​

우리는 각 범주의 분포를 테스트하고 우리의 관찰과 가장 일치하는 가정을 수행합니다. Other 및 Experience 범주에 대한 모델은 Conway–Maxwell–Poisson 분포를 사용하며 Reflection에서 Intention까지의 모델은 음이항 분포로 가장 잘 맞습니다. Conway–Maxwell–Poisson 분포 및 음이항 분포는 GLMMs에서 비정상적인 분포에 사용되는 특수한 분포입니다. Conway–Maxwell–Poisson 및 음이항 분포는 조건부 분산이 조건부 평균보다 낮거나 높을 때 사용되는 과소 또는 과대 분산 카운트 데이터에 대해 흔히 사용됩니다.

이 분석은 Python (Van Rossum & Drake, 2009) 및 R 스크립트 (R Core Team, 2020)를 사용하여 수행되었으며 특히 lme4 패키지 (Bates et al., 2012), glmmTMB 패키지 (Magnusson et al., 2017), 잔차 진단을 위한 DHARMa 패키지(Hartig, 2019)를 사용했습니다.

분석 프로시저의 모든 세부 사항(분석을 위한 행렬 작성, 분포 선택, 가설 검정 및 잔차 진단)은 Section 3.5의 참조에서 재현할 수 있도록 사용 가능합니다.

# 4 결과

# 4.1 반영성 분류

이 섹션에서는 섹션 3.1에서 소개된 주요 연구 질문에 대한 정확도를 측정하는 데에 자동 시스템을 사용하여 얻을 수 있는 분류 정확도에 대해 답합니다.

Figure 3은 조사된 여러 전통적인 분류기 중에서 가장 성능이 우수한 두 "shallow" 분류기와 가장 성능이 우수한 "deep" 분류기의 정확도를 비교합니다. 이는 체코어 문장 세트의 분류에 대한 것으로, 왼쪽에는 Random Forest 분류기가, 오른쪽에는 XLM-RoBERTa 트랜스포머가 있습니다. 두 분류 모델은 레이블된 신뢰 임계값을 초과하는 문장 하위 집합에서 반복적으로 훈련되고 평가되었습니다.

유사하게, Fig. 4는 이러한 모델들이 영어 반영 문장에 대해 훈련되고 평가될 때의 분류 정확도를 보여줍니다 (왼쪽), 그리고 영어로 훈련된 모델이 체코어 반영 문장에 대해 평가될 때의 모습입니다 (오른쪽).

양식의 맨 위

양식의 맨 아래

# 4.1.1 단일 언어 환경

Fig. 3에서 얕은 분류기가 특히 낮은 신뢰 설정에 대한 강력한 기준으로 나타나며, 신뢰 점수가 3 이상인 테스트 샘플에서 55.31%의 정확도에 도달합니다. 낮은 신뢰 샘플은 또한 같은 신뢰 임계값에서 XLM-RoBERTa와 같은 심층 신경 모델의 가장 큰 이점을 볼 수 있는 부분이며, 정확도는 63.44%에서 76.56%로 범위가 있습니다. 높은 신뢰도 샘플에서 심층 모델의 정확도는 92.68%에서 97.56%로 범위가 있으며, 최고의 비-신경 분류기(Random Forrest)의 정확도는 80.49%에서 82.93% 사이입니다. Fig. 4에서 보고된 영어 설정에서의 약간의 과성능은 XLM-Roberta가 더 많은 영어 텍스트로 사전 훈련되어 있어 언어에 대한 더 나은 기본 이해를 가지고 있기 때문임을 나타냅니다. 따라서 영어 테스트 세트에서 신경 분류기의 정확도는 93.18%에서 100%로 범위가 있습니다. 영어와 체코어 간의 정확도 차이는 다국어 기본 모델을 특히 언어별 데이터로 더 많은 사전 훈련하거나 저자원 언어의 응용 프로그램에 대한 경우 정확도를 더 향상시킬 수 있음을 시사합니다. 이 차이는 주석이 달린 반영 문장의 양을 더 늘림으로써 더욱 완화될 수 있습니다. 참고로 우리 실험에서의 훈련 반영 문장의 양은 두 언어 모두 동일합니다. Table 1은 특정 범주에 대한 샘플 크기 (N)와 선택된 성능 지표 (정밀도, 재현율, F1 점수) 간의 강력한 상관 관계를 드러내며, 이는 표현되지 않은 범주에 대한 교육 데이터 확장이 성능을 크게 향상시킬 수 있음을 시사합니다. 상위 네 가장 일반적이지 않은 범주에 대한 테스트 샘플 수가 주어진 임계값을 초과하는 것보다 낮다는 사실에도 불구하고 결과는 심층 분류기가 저자원 환경에서 훨씬 더 우수한 성능을 발휘한다는 점을 보여줍니다. 또한 Experience 범주가 단순한 의사 결정 트리의 규칙 기반 시스템으로는 특히 식별하기 어렵지만 더 풍부한 신경 모델로 상황이 크게 개선됨을 관찰합니다.

# 4.1.2 다국어 설정

XLM 신경 모델은 다국어 텍스트에 대한 사전 훈련이 되어 있기 때문에 우리는 교차언어 분류에 대한 실험도 진행합니다. 즉, 영어로 훈련된 분류기가 다른 언어에서 얼마나 잘 수행되는지를 확인합니다. Fig. 4 (오른쪽)에서 볼 수 있듯이 이러한 설정에서 영어 모델은 해당 언어에 독점적으로 미세 조정된 모델과 비교할 때 정확하게 수행됩니다. 체코어 문장에서 평가된 영어 모델을 체코어 문장에서 평가된 체코어 모델과 비교하면 정확도의 감소가 0에서 4.2% 사이로 나타납니다. 이는 XLM이 사전 훈련된 언어 중 데이터를 쉽게 얻을 수 있는 언어에서 훈련된 모델이 다른 언어에도 성공적으로 사용될 수 있음을 시사합니다 (Conneau & Lample, 2019). 또한 여러 언어의 리소스가 사용 가능한 경우 다국어로 다양한 언어에서 동시에 세밀한 조정을 통해 분류 품질을 더 향상시킬 수 있다고 생각합니다.

중요한 것은 우리의 데이터 세트에서 훈련한 다국어 모델이 RoBERTa Conneau 및 Lample (2019)의 100개 사전 훈련된 언어 중 어느 것이든 소량의 정확도 손실만으로 적용될 수 있다는 결과를 보여줍니다. 우리는 이 작업의 일부로 평가된 영어 모델을 자유롭게 사용할 수 있도록 제공하여 다른 언어에서 리플렉티비티에 대한 연구를 촉진하고 대규모 주석이 달린 데이터 수집이 불가능한 경우를 고려합니다.

# 4.2 탐색적 가설

이 섹션에서는 섹션 3.5에서 소개된 탐색적 분석의 결과를 요약합니다

# 4.2.1 H1: “중요한 반성” 임계값이 존재하여 학생들의 성과에 통계적으로 유의한 차이가 있다

섹션 3.5.1에서 설명한 대로, MSE 목적 함수를 사용하여 깊이가 1인 결정 트리를 적합시켜보았습니다(그림 5 상단에 표시됨). 이를 통해 상대 반영 비율이 .57인 후보 중요 반영 임계값을 얻었습니다.

그림 5에 나와 있는 대로 이 임계값을 기준으로 나눈 두 그룹의 평균 성과는 통계적으로 유의미하게 다릅니다. 구체적으로, 이 임계값 아래에 있는 학생들과 이 임계값 위에 있는 학생들의 평균 성과 간의 상대 성능 점수 차이는 .031입니다.

이 접근을 각각의 반영 범주에 대해 별도로 반복하여 각 범주에 대한 임계값이 존재하는지 여부를 확인합니다. 결과적으로 각 범주에 대한 그룹 평균은 .95의 유의수준에서 통계적으로 다릅니다. 그러나 Experience 및 Belief 범주의 경우 Experience 및 Beliefs 발생 비율이 더 높은 그룹이 다른 그룹보다 성과가 통계적으로 유의미하게 낮다는 결과를 얻었습니다.

이 결과는 Belief 범주의 높은 비율이 실제로 더 나은 실무 성과와 관련이 없을 수 있다는 가정을 제시합니다. 이러한 해석은 믿음과 행동 간의 관계에 대한 현대 지식론의 관점과 일치합니다(Schwitzgebel, 2010 참조). 마찬가지로, 우리의 데이터는 단순한 경험의 기술만으로는 멘토들이 인식하는 학생들의 실무 성과와 높은 연관성이 없다는 것을 나타냅니다. 반대로, 우리의 데이터는 실무에서 높은 인지된 학생 성과와 관련이 있는 범주 비율이 Feeling, Reflection, Difficulty, Perspective, Learning, Intention과 같은 다른 범주들에 대한 반성 및 분석을 수행하는 것으로 나타납니다.

이 결과는 교육 응용 프로그램에서 반성 글쓰기를 하는 데 적용될 수 있습니다. 이를 통해 경험과 신념의 깨끗한 설명만 추구하는 것이 아니라 경험을 분석하고 학생들에게 더 높은 성과와 관련된 다른 범주에 대해 반성할 것을 촉진하는 것이 권장된다는 가정이 제기됩니다. 이 가정은 이전에 반성이 실무 성과에 영향을 미치려면 임계 값을 초과해야 한다는 주장과 일치할 수 있습니다(Cohen-Sayag & Fischl, 2012). 따라서 중요한 반성은 Reflection 범주의 깊이 및 Feeling, Reflection, Difficulty, Perspective, Learning, Intention과 같은 범주의 비율과 관련이 있을 것으로 가정할 수 있습니다.

# 4.2.2 H2: 반성 일지의 범주 수는 반성 일지 제출의 순서에 따라 달라진다

두 번째 연구 탐구 가설에 대응하여, 우리는 반성 일지의 전달 시간에 따라 각 반성 범주의 발생 확률을 계산합니다. 그런 다음 범주의 수가 전혀 변하지 않는다는 가설, 즉 첫 번째 전달 시점과 동일한 상태로 유지된다는 것을 평가합니다. 각 범주에 대해 반성 저널 제출의 순서에 따라 적절한 모델에서 베타(𝛽)에서 exp (𝛽)를 계산하여 오즈 비를 계산했습니다(표 2 참조, 테이블 5는 별지 참조).

표 2는 첫 번째 제출 이후 네 번의 제출에서만 Experience 범주가 유의미하게 영향을 받는다는 것을 보여줍니다. 이는 저널의 두 번째, 세 번째, 네 번째 및 다섯 번째 제출에서 Experience 범주의 빈도가 첫 번째 제출과 비교하여 평균적으로 약 1회 증가한다는 것을 의미합니다. 다음으로 중요한 증가는 Difficulty 범주에서 나타나며, 두 번째, 세 번째 및 다섯 번째 제출에서 평균적으로 1.5회 증가합니다. Perspective 범주에서는 첫 번째 제출과 비교하여 두 번째와 세 번째 제출에서 1회 증가합니다. Learning 범주는 저널의 세 번째 및 다섯 번째 제출에서 유의미한 증가를 경험합니다.

반면에 비-반영 범주 인 Other 범주에 대해서는 저널의 세 번째 및 다섯 번째 제출에서 약 1회 미만으로 유의미한 감소가 있습니다. Feeling 범주에서도 첫 번째 제출과 비교하여 두 번째 및 다섯 번째 제출에서 1회 미만으로 감소하며, 이는 유의미합니다. 믿음, 의도 및 반영 범주의 경우 개별 제출의 일환으로 범주 수가 증가하거나 감소합니다. 그러나 Reflection 범주에서만 첫 번째 제출과 비교하여 저널의 세 번째 및 네 번째 제출에서 유의미한 감소가 있습니다.

각 범주에 대한 예측된 데이터 세트에서의 발생 횟수에 대한 참조는 1,070개의 반성 일지에 대한 표 3을 참조하십시오. 명백히 Other 범주가 가장 많으며(저널당 9.8회), Experience, Feeling, Reflection 및 Difficulty와 같은 범주가 이어집니다. Perspective, Belief, Learning 및 Intention 범주는 평균적인 저널에서 단일 발생을 초과하지 않습니다. 저널의 평균 길이는 약 29문장입니다(M = 28.65, SD = 14.45). Experience 범주는 경험 상황의 기술적인 설명을 나타냅니다. 이는 반영의 깊이에서 가장 낮은 수준으로, 주로 반영의 선결 조건입니다. 이 관점에서 반영되지 않거나 학생 경험을 설명하는 문장의 절반 이상(16.5문장)은 반영이 아니거나 학생 경험을 단순히 설명합니다.

참조 횟수를 고려할 때 각 범주의 횟수의 증가 또는 감소는 크지 않습니다. 이러한 결과는 Spalding et al. (2002), Hanafi (2019), Pasternak & Rigoni (2015)와 일치합니다. 이러한 연구는 학생 교사의 반영 쓰기의 질이 개인화되고 구체적인 피드백 없이는 일관되게 향상되지 않는다고 주장합니다. 학생 교사의 반영 쓰기의 발전에서는 반영 범주의 감소를 기대하지 않으며, 더 발전된 학생의 경우 반영 범주의 더 큰 감소를 예상할 것으로 기대됩니다. 그러나 Other의 감소는 상대적인 반영(텍스트의 "바인더"로서의 역할) 측면에서 어느 정도 제한이 있음을 주목합니다.

# 5. 토론 및 결론

우리의 연구는 선정된 기계 학습 접근법을 반성 일지의 자동 식별에 적용하고 이를 사용하여 문헌에서 식별된 몇 가지 관련 가설에 대한 답을 찾았습니다. 이전 연구와 대조적으로, 우리는 Vaswani et al. (2017)의 transformers 아키텍처를 기반으로 한 최신의 딥 랭귀지 모델의 품질을 조사하고 기존에 최고의 성능을 발휘한 전통적인 앙상블, 통계 또는 확률 모델을 구현한 결과와 비교했습니다. Ullmann (2019)이 제안한 범주의 분류를 따라, 우리는 Random Forrest 앙상블 모델에 의해 주도된 간단한 모델의 순위를 재현했지만, 새로운 모델이 이전 연구의 최고 모델보다 중요한 질적 향상을 찾았습니다 (4.1.1 절 참조). 더불어, 우리는 새로운 신경 모델이 학습된 다국어 모델이 적용 가능한 다른 언어에 대해 '콕'으로 작동한다는 사실을 확인했습니다 (4.1.2 절 참조). 더불어, 우리는 연구 결과 중 몇 가지를 강조합니다.

* 데이터 수집 과정은 후속 기계 학습 응용프로그램의 품질에 가장 중요한 결정요인입니다. 범주의 정의와 차이가 모호할 수 있는 응용프로그램에서 데이터 일관성을 극대화하는 데이터 수집 프레임워크를 작업하는 것이 중요합니다. 구별적 특징이 연속적인 경우 신뢰도를 기반으로 샘플을 다양화하면 '명확한' 샘플과 아마도 정확한 것들을 구별할 수 있습니다.
* 4.1절에서 보여진 바와 같이 보다 복잡한 신경 언어 모델은 전통적인 기계 학습 방법과 비교하여 상당한 질적 이점을 제공합니다. 또한 신경 모델은 정제된 샘플에서 더욱 이점을 얻을 수 있습니다.
* 다국어 언어 모델은 사회과학의 기계 학습 응용 연구를 민주화할 수 있으며, 영어 이외의 언어로 작성된 연구는 우리와 같은 작업을 기반으로 한 결과를 더해 전세계적으로 기여할 수 있습니다.
* 우리의 반성 식별을 기반으로 한 분석에서는 반성의 양이 모든 관찰된 차원, 즉 실제 수행 중인 다양한 범주에서 규칙적으로 증가하지 않음을 발견했습니다. 그러나 학생이 보고한 성적과 글쓰기에서의 반성 양 간에 관계가 존재한다는 것을 보여줌으로써, 학생들에게 더 많은 반성을 도입하도록 미래의 노력을 동기부여할 필요가 있습니다.

이러한 결과뿐만 아니라 우리의 연구는 반성 일지에 투명한 주석 지침이 포함된 익명화된 주석이 달린 말뭉치, 100개 이상의 사전 훈련된 언어로 적용 가능한 신경 모델 및 반성 식별을 위한 라이브러리를 커뮤니티에 제공합니다. 이는 우리가 제안한 것보다 더 많은 교육 수준에서의 교육을 위한 유용한 정보의 사용 사례를 포함하여 향후 연구가 우리의 결과를 기반으로 쉽게 발전할 수 있다고 믿습니다.

# 5.1 영향

우리의 연구의 결과에는 향후 연구 및 교육 실무에서 식별할 수 있는 몇 가지 영향을 간략하게 요약합니다.

1. **자동 기계 학습 도구 생성을 위한 데이터 수집 방법론:** 문제가 명확하지 않은 많은 교육 및 사회 응용프로그램의 맥락에서 자동 기계 학습 도구를 만들기 위한 데이터 수집 방법론을 제시했습니다. 다단계 데이터 수집 및 샘플 신뢰도 및 주석자 합의를 기반으로 한 정제를 통해 다른 사회 응용프로그램에서도 자동 도구의 주요 장애물 중 하나를 우회할 수 있을 것으로 기대합니다.
2. **CEReD 데이터 세트의 공개:** CEReD 데이터 세트의 공개는 반영 식별 도구의 품질을 더욱 향상시키는 데 상당한 도움이 될 것입니다. CEReD는 미래 방법들 간의 상호 비교를 표준화함으로써 향후 방법들의 상호 비교를 용이하게 할 것입니다.
3. **자동화 된 피드백 제공:** 우리의 반영 식별 도구는 학생들에게 개인 맞춤형, 자동화 된 또는 반자동화된 피드백을 쉽고 정확하게 제공할 수 있게 해줍니다. 이는 용량 또는 개인적인 이유로 다른 질적 평가를 전달하기 어려운 상황에서 특히 유용합니다.
4. **교육 과정의 방법적 조정 측정:** 무료로 제공되는 다국어 모델의 높은 정확도는 교육 과정에서의 방법적 조정을 양적으로 측정하고 그 영향을 평가할 수 있게 합니다. 조정 측정은 학생 작성물의 반영 비율이나 특정 범주의 개수 변화와 같은 형태를 취할 수 있습니다. 이러한 양적 평가는 교육의 잠재적 속성에도 영향을 미칠 수 있으며 예를 들어 학생들의 학습 프로세스 참여와 관련이 있을 수 있습니다.

또한, 자동 또는 반자동화된 피드백의 효율성에 대한 관련성은 해당 피드백의 형태에 중요한 영향을 미치는 중요한 고려 사항입니다. 향후 연구에서는 해당 피드백의 형태에 대한 가능한 형태를 자세히 살펴보겠으며, 자동 점수 부여, 학생 개인적 경향을 기반으로 한 자동으로 선택된 텍스트 피드백 또는 학생의 일지에 대한 시각적 분석과 같은 접근 방법을 평가하고 이러한 방법들이 학생들의 지각된 성능 및 시간에 따른 발전에 미치는 영향을 비교할 것입니다.

# 5.2 한계

우리의 결과를 향후 연구에서 사용할 때 고려해야 할 몇 가지 한계를 인정합니다. 3.4절에서 언급한 바와 같이 반영 식별의 복잡성은 문장 분류와 일치하지 않습니다. 반영 결정 요인은 여러 문장에 걸쳐 있을 수 있으며 또는 복잡한 단일 문장에는 여러 유형의 반영이 포함될 수 있습니다. 또한, 우리의 반영 범주는 운용화를 극대화하기 위해 문헌 검토를 기반으로 선택되었지만 더 발전된 반영 범주 체계가 존재할 수 있다는 점을 인정합니다. 우리의 모델이 한 언어에서 훈련된 모델이 추가 오류를 포함한 채로 다른 언어에서 사용될 수 있다는 것을 보여줬지만, 크로스링구얼 실험은 서로 다른 교육 체계에서의 방법론적 및 문화적 차이를 반영하지 않으며, 우리의 모델을 다른 언어에 적용하기 전에 신중하게 고려되어야 합니다. 또한, 크로스링구얼 응용프로그램의 정확도는 언어에 따라 다양하며 RoBERTa 모델이 사전 훈련된 언어 특정 말뭉치의 크기에 의해 조건이 걸립니다 - 해당 논문을 참조하십시오 (Liu et al., 2019). 우리는 이 측면을 고려하고 RoBERTa 모델의 사전 훈련 언어 (체코어)를 사용하여 나열된 크로스링구얼 오류 한계를 측정했으며, 이로 인해 크로스링구얼 오류의 추가 측정값을 제공하고 있습니다. 그러나 다른 언어에서의 응용 품질은 아마도 다를 것입니다. 이러한 공변량에 의해 유발된 질적인 차이는 자체 도메인에서 우리의 방법론에 따라 몇 가지 주석을 수집하고 이러한 새로운 주석의 예측 품질을 평가함으로써 양적으로 측정될 수 있습니다. 향후 연구에서 우리의 도구 세트를 사용하는 분석에서는 분류기가 각 범주 사이에서 다양한 오류를 드러내는 점을 고려해야 합니다. 특히 덜 표현되는 범주는 더 흔한 범주보다 더 높은 오류 수준을 드러냅니다. 이러한 특성은 특히 실험이 경계에 가까운 경우 오류를 고려해야 함을 시사합니다. 우리는 소개된 데이터 세트가 문장의 단순 분할, 자동 번역 및 기술적인 기술성을 기준으로 한 몇 가지 오류를 노출한다는 점을 인정합니다. 교육적인 관점에서 우리는 우리의 작업을 반영 작성의 정의에서 한 가지 특정 개념을 기반으로 하고, 여러 다른 개념을 갖는 반영에 관한 문헌이 존재함을 알고 있습니다. 예를 들어, Bass et al. (2020) 및 Kinsella (2007)가 강조하는 것처럼 반영을 전체적이고 신체화된 과정으로 강조하는 다른 개념이 있습니다. 우리가 첫 번째 분석을 기반으로 하는 멘토에 의한 학생의 지각된 성과는 멘토와 학생 간의 관계에 의해 편향될 수 있습니다. 우리는 독립적인 관찰자에 의한 성과 평가가 학생의 자체 평가와 결합되어 편향 도입의 위험을 최소화할 수 있다고 언급합니다.

# A: 분류기: 기술적 세부정보

이 섹션에서는 선택한 얕은(shallow) 및 심층(deep) 분류기의 훈련에 대한 기술적 구체적인 사항을 설명합니다.

# A.1: 데이터 분할

데이터셋에 문맥과 관계없이 중복 문장이 없도록 하기 위해 우리는 먼저 중복 문장을 제거합니다. 중복되는 문장이 분할 간에 거의 동일한 경우를 최소화하기 위해 문장을 알파벳 순으로 정렬합니다. 문장과 해당 문맥을 포함한 정렬된 목록을 90:5:5 비율로 훈련, 검증 및 테스트 분할로 나누어 총 6,097개의 훈련 문장, 340개의 검증 문장 및 340개의 테스트 문장이 생성됩니다. 우리는 최소 신뢰도가 4인 문장으로 훈련된 분류기를 최적화하는 데 사용되는 검증 세트를 선택하는 데에, 훈련 신뢰도 임계값과 일치하는 신뢰도 임계값을 기반으로 합니다. 모든 분할에서의 신뢰도 필터링은 해당 분할에만 적용되므로 재현 가능하며 서로 다른 신뢰도 임계값에서 테스트 시에 샘플이 교환되지 않습니다.

# A.2: 얕은 분류기

얕은 분류기의 최적의 하이퍼파라미터와 그들의 정확도에 대한 비교적 결과를 보고합니다.

# A.2.1: 전처리

단어 가방(bag-of-words) 표현의 차원을 최소화하기 위해 문장과 문맥의 모든 단어를 소문자로 변환하고 어간(stemming)을 제거합니다. Bag-of-words의 어휘를 가장 많이 나오는 상위 n개의 단어로 제한하되, 50% 이상의 문장에서 나타나는 불용어는 제외합니다. 여기서 n은 각 얕은 분류기의 하이퍼파라미터로 간주됩니다.

모든 얕은 분류기에 사용되는 최종 bag-of-words 표현은 문장과 그 문맥의 bag-of-words 벡터를 연결한 것입니다.

**하이퍼파라미터**

우리는 학습된 신뢰도의 유효성 검사 데이터 세트에서 다음과 같은 얕은 분류기의 매개변수에 대한 철저한 하이퍼파라미터 검색을 수행했습니다:

* (공유) 문맥 사용 ∈ {True, False}
* (공유) 앞뒤 문맥 창 크기 ∈ {1, 2, ... , 5}
* (공유) Bag-of-Words 크기: 상위 n 토큰으로 제한된 토큰 수 ∈ {100, 150, ... , 1000}
* (공유) 특정 언어에 대한 토큰화 ∈ {True, False}
* Random Forrest: 트리 깊이 ∈ {2, 3, 4, 5}
* Random Forrest: 추정자 수 ∈ {100, 200, ... , 500}
* SVM: 커널 함수 ∈ {linear, polynomial, radial}

**결과**

도표 6은 섹션 1에 나열된 하이퍼파라미터를 튜닝한 모든 평가된 분류기의 결과를 보여줍니다.

B: 깊은 분류기

우리는 섹션 3.4.2에서 설명한 학습 프로세스에 대한 기술적 특징을 보충하여 복제하거나 더 쉽게 사용자 정의할 수 있도록 합니다.

우리는 두 가지 다국어 Transformer 패밀리 대표 모델로 실험했습니다: Multi-lingual BERT-base-cased (Devlin et al., 2019) 및 XLM-RoBERTa-large (Conneau et al., 2020). 우리는 1에서 설명한 것과 동일한 방법으로 샘플을 분할합니다.

우리는 각각 해당 모델의 지원 언어에 대한 Wordpiece (Turc et al., 2019) 또는 Senten- cepiece (Kudo & Richardson, 2018) 모델을 기반으로 텍스트 단위를 세분화합니다. 우리는 다른 시퀀스 페어 문제와 마찬가지로, 예를 들어 답 추출 또는 함의 분류와 같이 특별한 심볼 "＜s＞"로 구분된 문장의 연결을 사용합니다.

기본적으로 우리는 유효한 배치 크기가 32인 경우 신경망 분류기를 훈련합니다. 즉, 모델의 가중치는 32개의 훈련 샘플에서 집계된 그래디언트를 사용하여 조정됩니다. 우리는 웜업을 전체 훈련 단계의 10%로 설정하고 기본적으로 훈련을 20번 에포크로 스케줄링합니다. 우리는 500개의 훈련 단계 동안 인내심을 가지고 평가 정확도에 대한 조기 중지를 사용합니다. 단일 분류기의 훈련은 Nvidia Tesla T4의 두 GPU에서 약 8시간이 소요됩니다.

최종 평가를 위해 우리는 검증 정확도에서 측정된 각 50개의 훈련 단계마다 가장 높은 정확도를 가진 모델을 선택합니다. 도표 7은 최소 신뢰 임계값이 4인 샘플로 훈련된 Multilingual-BERT-base-cased와 XLM-RoBERTa-large의 검증 손실 및 정확도를 비교합니다.

계산 복잡성 및 우리의 결과가 상당한 매개변수 조정에 대해 일관되게 나타나기 때문에 우리는 딥 분류기의 매개변수에 대한 체계적인 하이퍼파라미터 검색을 수행하지 않고 최상의 지식과 직관으로 이러한 값을 설정합니다.