스페인어-영어 코드스위칭 감성분석을 위한 재라벨링 접근법: 데이터 품질 개선의 영향 분석

신바다[O1], 김선오[2]
[1]단국대학교 모바일시스템공학과
[2]단국대학교 컴퓨터공학과
1127bada@gmail.com, suno8386@dankook.ac.kr

Re-labeling Approach for Spanish-English Code-switching Sentiment Analysis: Impact of Data Quality Improvement

Bada Shin[O1], Sunoh Kim[2]
[1]Department of Mobile Systems Engineering, Dankook University
[2]Department of Computer Engineering, Dankook University

요 약

스페인어-영어 코드스위칭 감성분석은 SNS 콘텐츠 분석 등 실용적 응용에서 중요하나, 기존 연구는 주로 모델 개선에 집중하며 데이터 품질 문제를 간과해왔다. 본 연구는 데이터 중심 접근법을 통해 트위터 기반 LINCE SA 데이터셋의 라벨링 품질을 분석하고 재라벨링 효과를 검증하였다. 원본 6,319개 샘플 중 100개 검증 결과 17% 라벨링 오류를 발견하였으며, 주요 패턴은 Positive와 Neutral 간 혼동이 전체 오류의 65%를 차지하는 것이었다. Hispanic-American 문화적 맥락을 고려한 재라벨링을 통해 5,567개 정제 데이터셋을 구축하였다. mBERT 기반 통제 실험에서 재라벨링 데이터셋은 기존 56.6%에서 60.6%로 4.0%p 성능 향상을 달성하였다. 데이터 품질 개선의 효과를 검증한 후, 정제된 고품질 데이터의 잠재력을 최대한 활용하기 위해 mBERT-XLM-R Late Fusion 앙상블을 적용하여 67.15%의 최종 성능을 달성하였다. 본 연구는 데이터 품질 개선이 모델 성능 향상의 기반이 됨을 입증하였으며, 구축된 고품질 데이터셋을 커뮤니티에 공개하여 관련 연구 발전에 기여하고자 한다.

1. 서 론

스페인어-영어 코드스위칭은 미국 Hispanic-American 커뮤니티의 일상적 의사소통 패턴으로, 특히 SNS와 같은 비공식적 디지털 공간에서 빈번하게 나타난다 (Das & Gambäck, 2013). 미국 내 Hispanic 인구는 2023년 기준 6,520만 명으로 전체 인구의 19% 이상을 차지하는 핵심 소비자 집단으로 (Pew Research Center, 2024), 이들의 SNS 감성 정보는 브랜드 관리와 맞춤형 마케팅에 중요한 자산이 되고 있다.

* 본 연구는 2024 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업 지원을 받아 수행되었음(2024-0-00035)

그러나 코드스위칭 텍스트의 감성분석은 동일 문장 내 두 언어 혼재로 인한 어휘적 모호성과 통사적 복잡성으로 인해 기존 단일 언어 모델들에게 상당한 도전을 제기한다 (Sitaram et al., 2020). 최근 연구들은 주로 mBERT를 Spanglish 데이터로 사전훈련하거나 멀티태스크 학습 등 모델 중심 접근에 집중해왔으나, LinCE 벤치마크에서 감성분석이 가장 어려운 태스크로 나타나는 등 (Xie et al., 2025) 여전히 성능 한계가 존재한다.

본 연구의 주요 기여는 다음과 같다. 첫째, LINCE SA 데이터셋의 체계적 품질 분석을 통해 17% 라벨링 오류를 발견하였다. 둘째, Hispanic-American 문화적 맥락을 반영한 5,567 개 고품질 데이터셋을 구축하였다. 셋째, 데이터 품질 개선만으로 4.0%p 성능 향상을 달성하였으며, 앙상블 기법과

결합하여 최종 67.15% 성능을 달성함으로써 기존 베이스라인 56.6% 대비 10.55%p 의 상당한 향상을 이루었다.

2. 데이터셋 분석 및 재라벨링 과정

2.1 원본 데이터셋 품질 문제

본 연구는 LINCE (Linguistic Code-switching Evaluation) 벤치마크의 감성분석 데이터셋을 기반으로 하였다 (Aguilar et al., 2020). 원본 18,789 개 샘플 중 고품질 코드스위칭 기준을 적용하여 6,319 개를 추출하였다. 추출 기준은 스페인어와 영어의 동시 포함, 각 언어당 최소 2 개 토큰, 순수 이중언어 구조로 설정하였다.

데이터 품질 평가를 위해 100 개 샘플을 무작위 추출하여 수동 검증을 실시한 결과, 17%의 체계적 라벨링 오류율이 확인되었다. 주요 오류 패턴은 **Positive 와 Neutral 간 혼동이 전체의 65%**로 나타났으며, 세부 원인으로는 자조적 표현이나 유머의 과도한 긍정 분류, Hispanic-American 소셜미디어 문화적 맥락 누락, 표면적으로 중립적이지만 내재된 감정의 미탐지, 부정적 감정 강도의 과소평가 등이 확인되었다.

이러한 초기 품질 검증 결과를 바탕으로 전체 데이터셋에 대한 체계적 재라벨링을 수행한 결과, 실제로 Positive 와 Neutral 간의 상호 변경이 전체 재라벨링의 45.8%를 차지하여 초기 분석의 정확성이 검증되었다. 추가적으로 데이터 구조 분석에서 훈련셋과 검증셋 간 동일한 sample_id 를 가진 376 개의 중복 쌍이 발견되어 총 752 개 샘플에서 정보 누출 위험이 확인되었으며, 이는 전체 데이터의 11.9%에 해당하는 수치이다.

2.2 재라벨링 과정

데이터 무결성 확보. 원본 데이터셋에서 동일한 sample_id를 가지지만 서로 다른 텍스트 내용을 포함하는 샘플들이 발견되었다. 이는 원본 데이터 구축 과정에서 train/validation/test 분할 후 독립적인 ID 할당으로 인해 발생한 것으로 추정된다. 이러한 ID 충돌은 교차 검증 시 데이터 누출 위험을 야기할 수 있어, 보수적 접근을 통해 해당 샘플들을 제거하였다. 최종적으로 5,567개의 구조적으로 무결한 데이터셋을 확보하여 원본 대비 88.1%를 보존하였다.

전처리 정책. 데이터 정제 과정에서 URL 과 멘션 정보는 의미적 노이즈로 간주하여 제거하였으나, 해시태그는 #happy, #stressed 와 같이 감성 분석에 유용한 단서를 제공하므로 보존하였다. 이러한 선택적 전처리는 소셜미디어 텍스트의 감성적 표현력을 최대한 유지하면서도 불필요한 메타데이터를 제거하기 위함이다.

재라벨링 원칙. Hispanic-American 소셜미디어 문화적 맥락을 우선 고려하여 스페인어 능력을 갖춘 연구자가 각 샘플의 감정 표현과 문화적 뉘앙스를 종합적으로 검토하였다. 주요 재라벨링 기준으로는 텍스트 내용 우선 원칙, JAJAJAJA와 같은 웃음 표현의 긍정 지시를 포함한 문화적 표현 해석, 코드스위칭 지점의 감정 강조 패턴 인식, 모호한 사례의 보수적 라벨링을 통한 허위 긍정 방지 등을 적용하였다.

재라벨링 과정에서 763 개 샘플(13.7%)이 수정되었다. 주요 변경 패턴은 Positive 에서 Neutral 로 194 개(25.4%), Neutral 에서 Positive 로 156 개(20.4%), Neutral 에서 Negative 로 133 개(17.4%), Positive 에서 Negative 로 114 개(14.9%)로 나타났다. 특히 Positive 와 Neutral 간 상호 변경이 전체의 45.8%를 차지하여, 초기 품질 검증에서 발견된 주요 오류 패턴인 65%가 실제 데이터 전반에서도 확인되었다. 이는 각각 과도한 Positive 분류, 숨겨진 Positive 맥락 발견, Negative 감정 미탐지, 심각한 오분류를 나타낸다. 최종 **Refined Dataset** 은 Positive 54.3%에 해당하는 3,021 개, Neutral 27.3%에 해당하는 1,519 개, Negative 18.4%에 해당하는 1,027 개의 분포를 달성하여 원본 대비 더 균형잡힌 구조를 확보하였다.

대표적인 재라벨링 사례는 다음과 같다. 첫째, 정보 전달 목적의 "I've never heard of that : No tener conocimiento sobre algo P1 : I think I'll buy me a loco burrito . P2 : Loco burrito ? I've never heard of that"에서 'loco burrito' 키워드로 인한 시스템의 Positive 오판을 Neutral 로 수정하였다. 둘째, "#selfie #goinghome #happy A descansar un poco! #weekend"에서 해시태그 나열로 인해 Neutral 로 오판된 것을 명확한 Positive 표현으로 수정하였다. 셋째, "Me acostumbre ya a tenerte aqui (sad emoticons) Im depressed y estas pr"에서 핵심 감정이 영어 코드스위칭 'I'm depressed'으로 표현되었으나, 자동화 시스템이 스페인어 부분의 중립적 표현에만 집중하여 전체 맥락의 부정적 감성을 놓치는 오류를 Negative 로 수정하였다.

3. 실험 설계 및 결과

3.1 실험 설계

본 연구는 데이터 품질과 모델 복잡도의 개별 효과를 측정하기 위한 통제된 실험을 설계하였다. 변수의 독립성을 보장하기 위해 단계적 비교 분석을 수행하였다.

베이스라인 모델: mBERT(bert-base-multilingual-cased)를 기본 모델로 선택하였다 (Devlin et al., 2019). 이는 코드스위칭 연구의 표준 실험 도구로, 재현성 확보와 데이터 품질 효과의 통제된 측정을 위한 적절한 선택이다.

개선 모델: 정제된 데이터의 잠재력을 최대한 활용하기 위해 mBERT 와 XLM-R (Conneau et al., 2020)의 Late Fusion 앙상블을 구현하였다. 이는 Sharma et al. (2022)의 방법론을 참고하되, 기존 연구와 달리 각 모델에 2 계층 분류 헤드(768→256→3)를 도입하여 개별 모델의 표현력을 추가적으로 개선하였다. 개별 모델의 로짓 출력을 MLPClassifier 메타 모델로 결합하였다 (그림 1).

훈련 설정: mBERT 는 학습률 1e-5, 에포크 5, 배치 크기 16 으로 설정하였다. XLM-R 은 학습률 3e-5, 에포크 6, 배치 크기 16 으로 최적화하였다. 모든 실험에서 AdamW 옵티마이저와 손실함수 기준 조기 종료를 적용하였으며, Weighted F1 Score 를 평가 지표로 사용하였다.

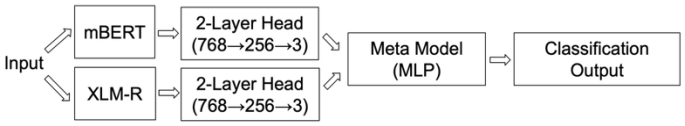

그림 1 . Late Fusion Architecture

3.2 결과 및 분석

실험 결과는 데이터 품질 개선과 모델 개선의 독립적 기여도를 명확히 보여준다.

표 1. 데이터셋 성능 비교

데이터셋	F1 Score
Original	56.6%
Refined	60.6%(+4.0%p)

표 2. 모델 성능 비교

모델	F1 Score
mBERT	60.6%
Fusion	67.15%(+6.55%p)

데이터셋 효과 분석. 단독 mBERT 모델에서 데이터 품질 개선만으로 4.0%p 성능 향상을 달성하였다(표 1). 이는 체계적 재라벨링을 통한 데이터 품질 개선이 모델 성능 향상의 핵심 요인임을 실증적으로 증명한다.

방법론 효과 분석. Late Fusion 앙상블에서는 메타모델을 통한 학습된 결합으로 개별 모델인 mBERT 57.36%와 XLM-R 61.04%를 상회하여 67.15%를 달성하였다(표 2). 단계적 분석을 통해 데이터 품질 개선의 독립적 효과 4.0%p 와 앙상블 기법의 추가 효과 6.55%p 를 확인하였으며, 기존 Late Fusion 연구의 1,707 개 대비 3.3 배 확장된 5,567 개 데이터셋에서 결과의 일반화 가능성을 검증하였다.

오류 분석 결과, 코드스위칭 지점에서의 복합적 감정 표현을 완전히 해석하지 못하는 경우가 관찰되었다. 특히 언어별로 다른 감정을 표현하는 패턴이나, "cabrón", "loco" 등 맥락에 따라 의미가 변화하는 문화적 표현의 화용론적 의미 파악에서 한계를 보였다. 이는 코드스위칭 특화 모델 연구의 필요성을 시사한다.

4. 결론

본 연구는 스페인어-영어 코드스위칭 감성분석에서 데이터 중심 접근법의 효과성을 실증하였다. LINCE SA 데이터셋의 체계적 분석을 통해 17% 라벨링 오류를 발견하고 Hispanic-American 문화적 맥락을 반영한 재라벨링으로 5,567 개 고품질 데이터셋을 구축하였다. 단계적 실험을 통해 데이터 품질 개선의 독립적 효과 4.0%p 와 모델 개선의 추가적 효과 6.55%p 를 분리 측정하여 총 10.55%p 성능 향상을 달성하였다. 이는 모델 아키텍처 중심의 기존 연구 동향에서 고품질 데이터 확보의 우선순위를 재조명하는 결과이다. 구축된 고품질 데이터셋을 연구 커뮤니티에 공개하여 코드스위칭 감성분석 연구 발전에 기여하고자 한다.

참고문헌

[1] Das, A. and Gambäck, B., "Code-Mixing in Social Media Text: The Last Language Identification Frontier?", TAL, 2013.

[2] Pew Research Center, "Who is Hispanic?", Pew Research Center Short Reads, September 12, 2024.

[3] Sitaram, S., "A Survey of Code-switched Speech and Language Processing", arXiv:1904.00784, 2020.

[4] Xie, K., "Enhancing Multilingual Language Models for Code-Switched Input Data", arXiv:2503.07990, 2025.

[5] Aguilar, G., "LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation", arXiv:2005.04322, 2020.

[6] Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[7] Conneau, A., et al., "Unsupervised Cross-lingual Representation Learning at Scale", arXiv:1911.02116, 2020.

[8] Sharma, G., et al., "Late Fusion of Transformers for Sentiment Analysis of Code-Switched Data", Findings of EMNLP, 2023.