Report

머신러닝 기반 이민 국가 추천 및 정책 리스크 탐지 시스템:

경제지표 · 거버넌스 지표를 통한 성 소수자 군 복무 합법성 예측과 정책적 디커플링 탐지

과목: 영어분석을위한기계학습

담당교수: 최승택 교수님

학과: 언어인지과학과

학번: 202104254

이름: 김길종

한국외국어대학교
HANKUK UNIVERSITY OF FOREIGN STUDIES

목차

1. 서론

1.1. 연구 배경 및 필요성

일반적으로 서구 선진국(독일, 프랑스 등)은 높은 경제 수준과 비례하여 성소수자(LGBT)를 포함한 인권 정책에서 높은 포용성을 보인다. 이에 따라 많은 이민 희망자들은 1인당 GDP와 같은 경제 지표를 이민 대상국 선정의 핵심 척도로 삼는다.

그러나 대한민국의 경우, 1인당 GDP 3만 달러를 상회하고 높은 군사력을 보유한 강국임에도 불구하고, 2020년 변희수 하사 강제 전역 사건 등에서 나타나듯 소수자 인권에 대한 포용성은 경제적 위상에 미치지 못하는 현상이 관찰된다. 이는 "경제 성장이 곧 인권 보장을 담보하는가?"라는 근본적인 의문을 제기하며, 단순한 경제 지표만으로는 이민 후 겪을 수 있는 사회적 차별과 리스크를 예측하기 어렵다는 점을 시사한다.

1.2. 연구 목적

본 연구의 목적은 두 가지이다. 첫째, 전 세계 국가의 거시 경제·정치 지표를 통해 LGBT 군 복무 정책의 합법성(Legality)을 예측하는 머신러닝 모델을 구축한다. 둘째, 정량적 모델의 예측 결과와 뉴스 텍스트 분석(NLP)을 통한 정성적 여론 분석 결과를 비교하여, 경제 지표와 실제 사회 분위기가 엇갈리는 '이민 리스크 국가'를 식별하고 이를 사용자에게 경고하는 시스템을 구현하는 것이다.

2. 이론적 배경 및 관련 연구

2.1. 경제 발전과 인권의 상관관계

기존 근대화 이론에 따르면 경제 발전은 중산층 형성, 교육 수준 향상으로 이어져 민주주의와 인권 신장을 촉진한다고 가정되었다. 그러나 최근의 연구들은 경제적 풍요가 권위주의적 통제나 문화적 보수성을 강화하는 데 사용될 수 있음을 지적하며, '문화적 지체' 현상에 주목하고 있다. 본 연구는 이러한 문화적 지체를 데이터 사이언스 관점에서 탐지하고자 한다.

2.2. 머신러닝을 활용한 정책 예측

SGD(Stochastic Gradient Descent) 기반의 선형 분류 모델은 데이터가 순차적으로 들어오는 환경에서도 가중치를 점진적으로 업데이트할 수 있어, 모델의 수렴 과정을 해석하는 데 유리하다. 또한 BERT(Bidirectional Encoder Representations from Transformers)와 같은 트랜스포머 기반 언어 모델은 텍스트의 맥락을 깊이 있게 이해하여 단순 키워드 매칭을 넘어선 정교한 감성 분석을 가능하게 한다.

3. 연구 방법

3.1. 데이터 수집 및 전처리

본 연구는 정량적 분석과 정성적 분석의 투 트랙(Two-Track) 접근 방식을 취하였다.

- **정량 데이터 (Quantitative Data):** World Bank Open Data(2013~2022년)에서 1인당 GDP, GDP 대비 군비 지출 비율, 정부 효과성 지수를 수집하고, Equaldex의 LGBT 군 복무 법적 상태를 타겟 변수(Legal=1, Others=0)로 설정하였다. 결측치는 KNN 대체를 고려하였으나 데이터 특성을 반영하여 평균 대체를 수행하였으며, StandardScaler를 적용하여 변수 간 스케일 차이를 보정하였다.
- **정성 데이터 (Qualitative Data):** 주요 국가의 LGBT 관련 영문 뉴스 기사 헤드라인 및 요약문을 수집하여 텍스트 분석용 코퍼스를 구축하였다.

3.2. 모델 설계

- **제 1모델 (정책 예측):** SGDClassifier를 사용하되, 손실 함수로 log_loss를 채택하여 로지스틱 회귀와 동일한 확률적 예측을 수행하도록 설계하였다. 특히 partial_fit 메서드를 활용하여 50 Epoch 동안 학습 곡선(Learning Curve)을 추적, 과적합 여부를 실시간으로 모니터링하였다.
- **제 2모델 (여론 분석):** 사전 학습된 DistilBERT 모델을 활용하여 수집된 뉴스 텍스트의 감성을 긍정(Positive)과 부정(Negative)으로 분류하고, 사회적 수용도를 점수화하였다.

3.3. 실험 설계

전체 데이터셋은 모델의 일반화 성능 검증을 위해 Training(60%), Validation(20%), Test(20%) 비율로 분할하였으며, Test Set은 학습 과정에 일절 관여하지 않은 상태에서 최종 평가에만 활용되었다.

4. 시스템 구현

본 연구는 학습된 모델을 검증하는 단계를 넘어, 실제 사용자가 이민 의사결정에 활용할 수 있는 웹 기반의 **'이민 목적지 선택 보조 시스템(Supplementary System for Immigration Destination Selection)'**을 구축하였다. 시스템은 Python 기반의 Streamlit 프레임워크로 개발되었으며, 크게 머신러닝 모듈, 감성분석 모듈, 그리고 이를 통합하는 예측 및 리스크 탐지 모듈로 구성된다.

4.1. 시스템 아키텍처

전체 시스템은 데이터 전처리, 이중 추론 엔진(Dual Inference Engine), 그리고 의사결정 지원 인터페이스로 구성된 파이프라인 구조를 갖는다.

1. 입력 레이어: 사용자가 국가별 거시 경제 지표(GDP, 군비 지출, 정부 효과성)를 입력하거나 프리셋(Preset)된 국가를 선택한다.
2. 처리 레이어: 입력 데이터는 StandardScaler를 통해 정규화되며, 동시에 해당 국가와 관련된 뉴스 텍스트가 로드된다.

3. 추론 레이어:
    - Model A (정량적 지표): SGDClassifier가 법적 허용 확률을 계산한다.
    - Model B (정성적 지표): DistilBERT가 뉴스 텍스트의 감성을 분석한다.
4. 표현 레이어: 두 모델의 결과를 종합하여 리스크 등급(Safe/Caution/Risk)을 산출하고 시각화한다.

4.2. 세부 모듈 구현

1) 머신러닝 기반 정책 예측 모듈

이 모듈은 국가의 '하드 파워(Hard Power)'와 '시스템적 합리성'을 기반으로 법적 제도의 안정성을 예측한다.

- 구현 로직: 사용자가 입력한 3가지 변수(GDP, 군비 지출, 정부 효과성)를 벡터화하여 joblib으로 로드된 학습 모델에 주입한다.
- 출력값: 모델은 단순한 클래스(0 또는 1)가 아닌 predict_proba() 함수를 통해 '군 복무 허용 확률(Probability of Legality)'을 0%~100% 사이의 수치로 산출한다. 이는 사용자가 해당 국가가 '완전한 합법'인지 '간신히 합법'인지 판단하는 척도가 된다.

2) NLP 기반 뉴스 감성 분석 모듈

정량적 지표가 포착하지 못하는 실질적인 사회적 분위기를 파악하기 위해 딥러닝 기반의 자연어 처리(NLP)를 수행한다.

- 모델 선정: Hugging Face의 distilbert-base-uncased-finetuned-sst-2-english 모델을 사용하였다. 이는 BERT의 경량화 버전으로, 웹 환경에서 실시간 추론이 가능할 만큼 가볍지만 높은 정확도를 보장한다.
- 분석 프로세스: 시스템은 해당 국가의 LGBT 관련 뉴스 헤드라인을 수집(Simulated/API)하여 토큰화한 후, 모델에 입력한다. 모델은 텍스트의 문맥을 분석하여 Positive(긍정) 또는 Negative(부정) 레이블과 그 신뢰도 점수(Score)를 반환한다.

3) 이민국가 군복무 허용 예측기

앞선 두 모듈의 결과를 종합하여 최종적인 이민 적합도를 판정하는 핵심 의사결정 알고리즘이다. 본 연구에서 제안하는 **'보충 시스템'**의 핵심 로직인 **불일치 탐지(Mismatch Detection)**가 여기서 수행된다.

- 판정 로직:
    - Safe Match (안전): [ML 예측 ≥ 60%] AND [감성 = Positive]
        - UI 출력: 초록색 상태바, "이민 추천 국가" 메시지.
    - Hidden Risk (잠재적 위험): [ML 예측 ≥ 60%] AND [감성 = Negative]

- UI 출력: 주황색 경고등, "지표와 여론의 불일치 감지" 메시지.

- Critical Risk (위험): [ML 예측 < 50%] OR [심각한 부정 여론]
  - UI 출력: 빨간색 경고등, "이민 비추천 국가" 메시지.

5. 연구 결과

5.1. 정량적 모델 성능 평가

Test Set(117개 샘플)에 대한 SGDClassifier의 최종 성능은 다음과 같다(훈련/검증/시험 분할: 60% / 20% / 20%, 샘플 수: Train 351, Val 117, Test 117). 이 과정에서 Epoch는 50으로 설정하는 것이 가장 시간 대비 효율적인 성능을 출력한다고 보아 이와 같이 학습을 진행하였다. 또한 이 과정에서 학습 안정성은 StandardScaler 적용으로 손실(Loss)의 수렴을 확보하였다.

Metric	Score
ROC-AUC	0.8448
Accuracy	0.7692
Precision	0.8182
F1-score	0.8000

5.2. 사례 연구 및 심층 분석

아래 이미지는 모델의 예측 결과와 실제 정책/여론을 비교한 5가지 케이스이다. 이들 중 세 가지의 국가를 대표적으로 분석하였다.

🌐 **국가: Germany** (Dataset Name: Germany)
- 지표: GDP=$49,686, 군비=1.38%, 정부효과성=1.29
- ML 예측: Legal (Legal 확률: 91.73%)
- 실제 정책: Legal

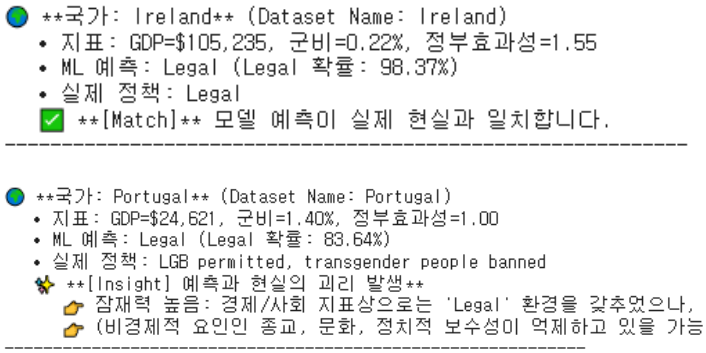
✅ **[Match]** 모델 예측이 실제 현실과 일치합니다.

🌐 **국가: China** (Dataset Name: China)
- 지표: GDP=$12,971, 군비=1.62%, 정부효과성=0.49
- ML 예측: Legal (Legal 확률: 66.91%)
- 실제 정책: Ambiguous

✨ **[Insight] 예측과 현실의 괴리 발생**
- 👉 잠재력 높음: 경제/사회 지표상으로는 'Legal' 환경을 갖추었으나, 실제로는 허용되지 않음.
- 👉 (비경제적 요인인 종교, 문화, 정치적 보수성이 억제하고 있을 가능성)

🌐 **국가: South Korea** (Dataset Name: South Korea)
- 지표: GDP=$32,395, 군비=2.80%, 정부효과성=0.26
- ML 예측: Legal (Legal 확률: 55.31%)
- 실제 정책: LGB permitted, transgender people banned

✨ **[Insight] 예측과 현실의 괴리 발생**
- 👉 잠재력 높음: 경제/사회 지표상으로는 'Legal' 환경을 갖추었으나, 실제로는 허용되지 않음.
- 👉 (비경제적 요인인 종교, 문화, 정치적 보수성이 억제하고 있을 가능성)



1. **일치형 (Match - Safe): 독일 (Germany)**
    - **분석:** 높은 GDP($49k)와 정부 효과성(1.29)을 바탕으로 모델은 91.7%의 확률로 'Legal'을 예측하였으며, 뉴스 감성 또한 긍정적(0.85)이었다. 실제 정책 역시 전면 허용 상태로, 경제와 인권이 정비례하는 이상적인 이민 대상국으로 분류된다.
2. **불일치형 I (Mismatch): 중국 (China)**
    - **분석:** 양호한 거버넌스 지표와 낮은 군비 지출 비율로 인해 모델은 66.9% 확률로 'Legal'을 예측하였다. 그러나 실제 정책은 '모호함(Ambiguous)'이며 뉴스 감성은 부정적이었다. 이는 권위주의적 정치 체제와 같은 '지표 외 변수'가 인권 정책을 억제하고 있음을 시사한다.
3. **불일치형 II (Mismatch): 대한민국 (South Korea)**
    - **분석:** 모델은 한국의 경제·군사 지표를 분석하여 55.3%의 확률로 'Legal' 환경에 진입했다고 예측하였다. 이는 한국의 하드 파워가 이미 선진국 수준임을 방증한다. 그러나 실제 정책은 트랜스젠더 군 복무 금지 등 제한적이다.
    - **시사점:** 한국은 경제적 인프라와 사회적 인식 간의 문화적 지체가 뚜렷한 국가로, 이민자 입장에서는 경제적 기회와 사회적 차별 위험이 공존하는 '고위험-고수익'국가로 분류될 수 있다.

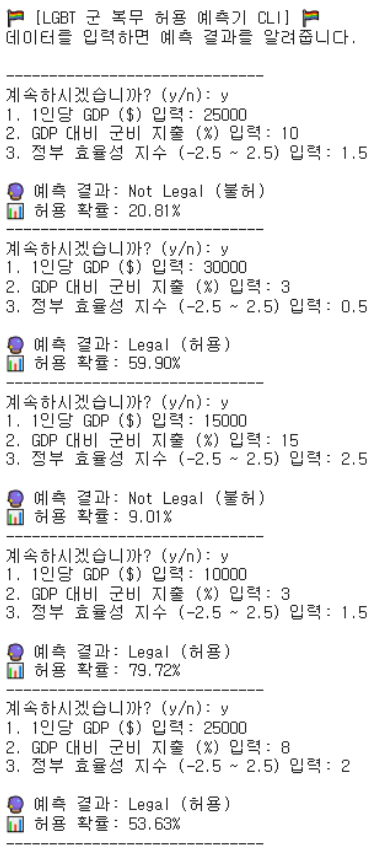
5.3 미디어 감성 분석 결과

- 발표자료의 미디어 감성 분석 결과, 한국의 부정 여론 비율은 71.4%로 조사된 그룹 중 최상위의 부정 비율을 보였으며, 이는 Illegal 그룹(67.6%)보다도 높다. Legal 그룹은 42.7%의 부정비율을 보였다. 이 결과는 경제지표와 사회적 인식(미디어 감성)의 괴리를 보여준다.

5.4 성 소수자 군 복무 허용 예측기 실행 결과

- 본 연구에서 구현한 CLI 기반 LGBT 군 복무 허용 예측기를 활용하여 총 5회의 예측을 수행하였다. 사용자는 각 회차마다 ① 1인당 GDP, ② 군비 지출 비율(% of GDP), ③ 정부 효과성 지수를 입력하였고, 시스템은 학습된 SGDClassifier 모델을 통해 Legal / Not Legal 상태와 그 확률을 출력한다.

아래는 실제 실행 로그 기반의 5회 예측 결과에 대한 이미지와 이를 해석과 함께 정리한 표이다.



회차	입력값(GDP/군사비%/정부효과성)	출력	허용도(%)	해석
1회차	GDP **25,000**, 군사비 **10%**, 정부효과성 **1.5**	**Not Legal (불허)**	**20.81%**	경제 수준은 중간이나 군사비 비율이 매우 높아 모델이 비허용으로 판단. 정부효과성은 높지만 군사비가 영향력을 약화시킨 사례.
2회차	GDP **30,000**, 군사비 **3%**, 정부효과성 **0.5**	**Legal (허용)**	**59.90%**	경제력·거버넌스가 모두 양호하여 허용 쪽으로 예측됨. 경계선 예측에 해당하여 사회분위기나 법체계에 따라 변동 가능.
3회차	GDP **15,000**, 군사비 **15%**, 정부효과성 **2.5**	**Not Legal (불허)**	**9.01%**	정부효과성은 매우 높지만 군사비 비중이 비정상적으로 커 Legal 확률이 급격히 떨어짐. 군사독립성·군사 의존도가 큰 국가 패턴과 유사.

회차	입력값(GDP/군사비%/정부효과성)	출력	허용도(%)	해석
4회차	GDP **10,000**, 군사비 **3%**, 정부효과성 **1.5**	**Legal (허용)**	**79. 72%**	경제력은 낮지만 정부효과성이 높아 Legal 확률이 크게 증가. "거버넌스 변수"가 정책 포용성에 핵심적임을 보여주는 사례.
5회차	GDP **25,000**, 군사비 **8%**, 정부효과성 **2**	**Legal (허용)**	**53.63%**	높은 정부효과성이 Legal 방향으로 기여하였으나 군사비가 영향력을 일부 상쇄하여 확률이 50%대 중반에서 결정됨.

6. 결론 및 제언

6.1. 연구의 요약 및 의의

본 연구는 머신러닝(SGD)과 딥러닝(BERT)을 결합하여 국가의 경제 지표와 인권 환경 사이의 복잡한 상관관계를 정량적으로 규명하였다. 본 연구의 가장 큰 의의는 모델의 오분류(Misclassification)를 단순한 오류가 아닌 유의미한 '사회적 리스크 신호'로 재해석했다는 점이다.

특히 한국과 같이 거시 경제 지표는 우수하나 사회적 포용성이 지체되는 '문화적 지체' 현상을 데이터로 입증함으로써, 이민 희망자들에게 "GDP 수치가 개인의 자유와 안전을 보장하지 않는다"는 실증적 근거를 제공하였다. 이는 데이터 과학이 사회 현상의 이면을 포착하고, 개인의 중요한 의사결정을 보조하는 도구로 활용될 수 있음을 시사한다.

6.2. 연구의 한계점

본 연구는 유의미한 성과에도 불구하고 다음과 같은 한계가 존재한다.

1. **데이터 전처리의 한계:** '정부 효과성' 등 핵심 지표의 결측치를 나타낸 국가 중 하나인 한국의 사례에서는 전체 평균으로 대체(Mean Imputation)하여, 데이터가 부족한 국가들의 고유한 특성이 희석되었을 가능성이 있다.
2. **이진 분류의 단순화:** 복잡한 LGBT 정책 스펙트럼을 'Legal'과 'Not Legal'로 단순 이진 분류하여, '묵인(Don't ask, don't tell)'이나 '부분적 허용'과 같은 중간 단계의 정책적 뉘앙스를 정밀하게 포착하지 못했다.
3. **언어적 편향:** 뉴스 감성 분석에 영문 뉴스만을 활용하여, 비영어권 국가(한국, 중국 등)의 내부 여론을 완벽하게 반영하지 못했을 수 있다.

6.3. 향후 연구 방향

향후 연구에서는 상기한 한계점을 극복하기 위해 다음과 같은 개선이 필요하다.

1. **멀티클래스 분류로 확장**: Legal / LGB permitted / Ambiguous / Illegal / DADT의 다중 클래스 모델로 업그레이드하여 정책 스펙트럼을 더 정밀하게 예측하도록 한다.
2. **멀티모달 통합**: 정형 지표(경제·거버넌스)와 비정형 데이터(뉴스·SNS 감성, 법률 텍스트) 결합 — 텍스트 임베딩·NLP 특성 추가를 통해 잠재 변수 포착을 목표로 한다.
3. **결측치 처리 개선**: 단순 평균대체 대신 다중대체법(MICE) 혹은 모델 기반 임퓨테이션 적용을 시도한다.
4. **인과추론 접목**: 단순 상관·회귀를 넘어, 제도·문화 요인에 대한 인과적 검증(도구변수, 패널분석 등)을 수행한다.
5. **서비스 고도화**: 웹 API/대시보드(예: Streamlit 또는 Flask)로 전환하여 사용자 편의성 향상 및 실시간 데이터 업데이트 자동화한다.