혐오 표현 탐지를 위한 불균형 데이터 처리 기법 비교: 오버 샘플링을 중심으로

A Study on Imbalanced Data Handling Techniques in Hate Speech Detection: Focused on Oversampling

최해민 [1]
Haemin Choi

[1]성균관대학교 데이터사이언스융합전공
E-mail: hm20201009@gmail.com

요 약

소셜 미디어를 중심으로 온라인에서 발생하는 폭력적, 차별적 언어를 식별 및 억제하는 작업은 필수적이며, 이와 같은 혐오 표현 탐지를 위한 연구가 활발히 진행되고 있다. 하지만 혐오 표현 데이터셋의 특성상 2 개 이상의 클래스를 가지는 경우, 주류 클래스인 정상 샘플에 비해 혐오표현 클래스 샘플의 수가 적은 데이터 불균형성이 나타나는 경향이 있다. 본 연구는 이를 고려해 오버 샘플링을 중심으로 혐오 표현 탐지를 위한 불균형 데이터 처리 기법을 비교한다. 다중클래스 분류를 위한 머신러닝 알고리즘 중 다항분포 나이브 베이즈, 서포트 벡터 머신이 랜덤 오버 샘플링과 SMOTE 계열 기법 적용시 큰 성능 향상을 보였으며, 딥러닝 알고리즘 중 LSTM 기반 분류기에 랜덤 오버 샘플링을 적용한 결과 주류 클래스에 대한 과적합 문제가 해결됨을 확인하였다.

키워드 : 혐오 표현 탐지, 텍스트 분류, 불균형 데이터 처리, 오버 샘플링, 자연어 처리

1. 서 론

혐오 표현 탐지는 온라인에서 발생하는 폭력적이고 차별적인 언어를 식별하고 억제하기 위한 필수적인 작업으로, 특히 소셜 미디어와 같은 플랫폼에서 개인과 집단에게 미치는 심리적, 사회적 피해를 줄이기 위해 중요하다[1]. 그러나 혐오 표현 데이터셋은 비율적으로 '혐오 표현'에 해당하는 데이터가 상대적으로 적고, 일반적 발언이 많은 형태의 클래스 불균형 문제를 가진다. 이는 일반적으로 혐오 표현의 발생 빈도가 낮아 데이터 수집 과정에서 불균형이 발생하기 때문이다. 이러한 클래스 불균형 문제는 모델이 소수 클래스(혐오 표현)를 제대로 학습하지 못하게 하여 탐지 성능을 저하시킬 수 있으며[2], 다양한 오버 샘플링 기법이 이를 개선하기 위한 방법으로 연구되고 있다.

본 연구에서는 혐오 표현 탐지를 위한 텍스트 분류에서 데이터셋의 불균형을 극복하기 위한 다양한 불균형 데이터 처리 기법을 비교한다. 그 중에서도 랜덤 오버 샘플링(ROS), 합성 소수 샘플링 기술(SMOTE)과 같은 오버 샘플링 기법을 중심으로 데이터 불균형을 처리하고 혐오 표현 탐지기의 성능에 어떤 기여를 하는지 확인하였다.

혐오 표현 탐지를 위해 다양한 머신러닝 분류알고리즘을 이용하였으며, 어떤 분류 알고리즘과 오버 샘플링 기법의 조합이 혐오 표현 탐지의 성능 개선을 돕는지 탐색한다. 본 연구의 분석 결과는 혐오 표현 데이터셋에 특성상 내재하는 클래스 불균형 문제를 극복하고, 이를 통해 실생활에서 필요성이 대두되는 혐오 표현 탐지 문제의 효율적인 개선 방안을 제안한다.

2. 제안 방법

일반적으로 샘플링 기반의 데이터 불균형 처리 기법은 오버 샘플링, 언더 샘플링으로 나뉜다. 오버 샘플링은 소수 클래스의 데이터를 늘려 다수 클래스와의 비율을 맞추며, 언더 샘플링은 다수 클래스의 데이터를 제거하여 소수 클래스와의 비율을 맞춘다. 하지만 언더 샘플링에서는 데이터의 손실이 일어날 수 있다는 점과 일반적으로 오버 샘플링이 더 좋은 성과를 보인다는 점을 고려하여, 본 연구에서는 오버 샘플링 기법을 중심으로 분석을 진행한다.

오버 샘플링 기법으로는 랜덤 오버 샘플링, SMOTE 등이 존재한다. 랜덤 오버 샘플링(ROS)은 소수 클래스의 데이터를 무작위로 선택해 복제하여 비율을 맞춘다. SMOTE 는 소수 클래스 샘플 사이에서 합성 데이터를 생성하는 방법으로, 기존 데이터를 복제하는 대신 원래 데이터 분포를 유지하면서 새로운 데이터를 만들어 낸다. 임의의 소수 클래스 샘플을 기준으로 K 개의 가장 가까운 이웃 샘플을 찾으며, 선택된 샘플과 이웃 간의 거리를 바탕으로 두 샘플 사이 임의의 위치에 새로운 샘플을 생성한다.

SMOTE 는 소수 클래스 예측 성능을 개선하는데 유용한 도구로 여겨지며, 변형에 따라 K-means SMOTE, ADASYN, SVMSMOTE 와 같은 기법들이 존재한다. K-means SMOTE 는 소수 클래스 데이터를 여러 클러스터로 나눈 후, 각 클러스터 내에서 SMOTE 를 적용해 클러스터별 특성을 반영한 합성 데이터를 생성한다. ADASYN(Adaptive Synthetic Sampling)은 소수 클래스 중에서 학습이 어려운 샘플에 더 많은 합성 데이터를 추가하여 모델이 어려운 샘플에 집중하도록 한다. SVM SMOTE 는 서포트 벡터 머신과 SMOTE 를 결합한 기법으로, 소수 클래스의 데이터 중 SVM 의 결정 경계에 가까운 샘플에 초점을 맞춰 합성 데이터를 생성한다.

본 연구에서는 머신러닝 알고리즘으로 다항분포 나이브 베이즈, 서포트 벡터 머신, 랜덤 포레스트 분류기,

XGBoost 분류기를 선택하며, 오버 샘플링 기법 중 ROS, SMOTE, K-means SMOTE, ADASYN 을 적용하여 혐오 표현 텍스트 분류 성능 개선 정도를 살펴본다. 또한 LSTM 과 같은 딥러닝 알고리즘을 위한 불균형 텍스트 데이터 처리 기법 적용 사례가 충분치 않은 점에 집중하여, ROS 를 통해 분류 성능이 어느 정도로 개선되는지 탐색한다.

3. 실험 및 결과

3.1 실험 데이터

혐오 표현 분류기 구현을 위한 데이터셋으로는 HateXplain[3]을 사용하였다. HateXplain 은 2020 년에 공개되었으며, 소셜미디어 Twitter 와 Gab 으로부터 다양한 포스트를 수집해 구축되었다. 혐오 표현(hate speech) 감지와 같은 혐오성 콘텐츠 분석을 위한 벤치마크 데이터셋으로 여겨진다. 각 텍스트 데이터 샘플은 어노테이터들에 의해 'normal', 'offensive', 'hatespeech' 세 가지 클래스로 라벨링 되어 있으며, 원본 데이터셋은 19201 개의 샘플로 구성된다. 본 연구에서는 'offensive', 'hatesppech' 두 클래스가 전체에서 59%를 차지한다는 점을 고려하여, 실험 가설과 real-life situation 을 반영해 불균형성을 더욱 부여하기 위해 두 클래스가 전체에서 각 21%, 총 42%를 차지하도록 다운 샘플링 하였다. 학습 데이터를 오버 샘플링 할 때, 머신러닝 분류기 사용 시에는 클래스 비율이 1:0.6:0.6 이 되도록, LSTM 분류기 사용 시에는 1:1:1 이 되도록 오버 샘플링 하였다.

3.2 실험 결과

첫 번째 실험은 각 머신러닝 분류기를 대상으로 오버 샘플링 기법을 미적용한 경우와 적용한 경우 소수 클래스에 대한 F1-score 를 비교하는 방식으로 이뤄졌으며, Text Vectorization 을 위해 TF-IDF 가 사용되었다. 표 1 은 각 머신러닝 분류기의 오버 샘플링 기법 적용에 따른 'offensive' 클래스에 대한 F1-score 를 나타낸 것이다. 다항분포 나이브베이즈의 경우 오버샘플링 미적용시 0.02 였던 점수가 ROS 및 SMOTE 계열의 기법 적용시 0.21 까지 향상됨을 확인할 수 있다. 다른 분류기의 경우 괄목할 만한 성능 향상은 없었으나, 약간의 개선이 조금씩 있음을 확인할 수 있다. 표 2 는 'hate' 클래스에 대한 F1-score 를 나타낸 것으로, 나이브베이즈 분류기의 성능이 전반적으로 큰 향상을 보이며 특히 ADASYN 적용의 경우 0.56 까지 개선됨을 알 수 있다. 또한 SVM 분류기가 0.34 로부터 0.60 을 웃도는 성능 향상을 보였으며, 특히 KMeans SMOTE 적용시 0.63 의 최고 성능을 보였다. 두 소수 클래스에 대해 랜덤포레스트와 XGBoost 의 경우 전반적으로 성능 개선은 있으나 미미한 양상을 보였다.

두 번째 실험으로는 문장의 장기 의존성을 학습하는 딥러닝 알고리즘인 LSTM 을 이용하여 랜덤 오버 샘플링 기법을 적용한 후 분류 성능을 확인하였으며, 워드 임베딩 기법으로는 Word2Vec 이 사용되었다. 먼저 오버 샘플링을 하지 않은 데이터와 첫 실험과 동일하게 1:0.6:0.6 로 랜덤 오버 샘플링된 데이터로 모델 학습 및 검증을 진행하였으나, 데이터 불균형성으로 인해 모델이 주류 클래스에 과적합 되어 검증 정확도가 전 학습 과정에서 일정하게 나타남을 확인하였다. 이는 모델이 주류 클래스만을 분류하여 나타난 결과로, 과적합 방지를 위해 클래스 비율이 1:1:1 이 되도록 학습 데이터에 대해 랜덤 오버 샘플링을 수행한 후 모델 학습 및 검증을 진행하였다. 그림 1 은 학습 반복에 따른 LSTM 분류기의 훈련 및 검증 정확도를 나타내고 있다. 훈련 정확도에 비해 검증 정확도는 변동성이 있기는 하나, 과적합 문제가 해결되고 최대 0.60 까지 검증 정확도가 향상됨을 확인하였다. 학습 후 각 클래스의 F1-Score 는 0.62, 0.64, 0.23 으로, 소수 클래스 중 하나인 'offensive' 클래스에 대하여 특히 높은 점수를 보였다.

4. 결론

본 연구에서는 소셜 미디어 기반의 자연어에서 혐오 표현 탐지를 위한 다중클래스 텍스트 분류를 진행하는 데에 있어, 데이터셋의 불균형성을 극복하기 위해 소수 클래스에 대한 오버샘플링 기법을 적용한 결과를 비교하였다. 그 결과 머신러닝 알고리즘 중 다항분포 나이브 베이즈, 서포트 벡터 머신 분류기가 ROS, SMOTE 계열의 기법 적용시 성능이 향상됨을 확인하였다. 또한 딥러닝 알고리즘인 LSTM 분류기의 경우 ROS 를 적용한 결과 주류 클래스에 대한 과적합이 해결됨을 확인하였다. 본 연구에서는 LSTM 활용시 ROS 만을 적용하였으나, 향후 연구에서는 Synonym Replacement, Random Insertion, Random Swap 등 데이터 증강 기법을 적용해 불균형 텍스트 데이터의 분류 성능을 높여볼 수 있을 것으로 기대된다.

	No OS	ROS	SMOTE	KMeans	ADASYN
NB	0.02	**0.19**	**0.21**	**0.19**	**0.21**
SVM	0.34	0.37	0.38	0.35	0.37
RF	0.31	0.34	0.33	0.28	0.33
XGB	0.35	0.40	0.39	0.37	0.39

표 1. 머신러닝 분류기 'offensive' 클래스 F1-score
Table 1. F1-score of 'offensive' class

	No OS	ROS	SMOTE	KMeans	ADASYN
NB	0.21	**0.51**	**0.51**	**0.47**	**0.56**
SVM	0.34	**0.60**	**0.61**	**0.63**	**0.61**
RF	0.66	0.67	0.67	0.66	0.68
XGB	0.67	0.67	0.69	0.67	0.69

표 2. 머신러닝 분류기 'hate' 클래스 F1-score
Table 2. F1-score of 'hate' class

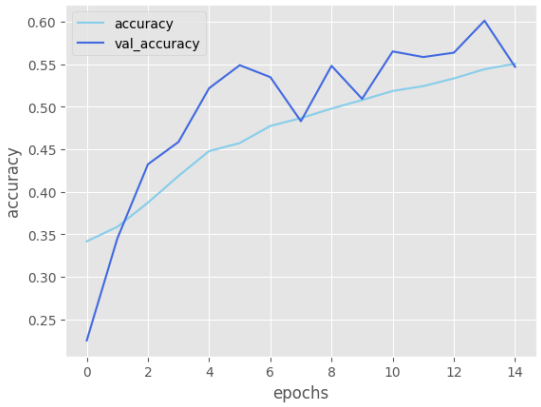

그림 1. LSTM 분류기의 훈련 및 검증 정확도
Fig 1. Training and Validation Accuracy of LSTM

참 고 문 헌

[1] S. Anna and W. Michael, "A Survey on Hate Speech Detection using Natural Language Processing," *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp.1–10, 2017.

[2] P. Cristian, and B. Mihaela, "Dealing with Data Imbalance in Text Classification," Procedia Computer Science, vol.159, pp.736-745, 2019.

[3] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P. and Mukherjee, A. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection", *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), pp. 14867-14875, 2020.