음주운전 사건 형종 분류 모델 성능 개선: 가중치 랜덤 샘플링을 적용한 라벨 불균형 해소 및 요약 모델을 사용한 데이터 증강 기법 적용

김형, 김학성
KAIST 소프트웨어대학원프로그램

probro@kaist.ac.kr, bani0913@kaist.ac.kr

Performance improvement of drunk driving penalty classification model: Using weighted random sampling to resolve label imbalance and using summarization model for data augmentation

Kim Hyung, Kim Hak Sung
KAIST Software Graduate Program

요 약

최근 법률 서비스는 기술의 발전을 통해 '리걸 테크(Legal Tech)'라는 산업으로 발전하였다. 이는 사용자들에게 인공지능 법률 질의응답 상담 등 다양한 서비스를 제공하고 있다. 하지만 이런 관심과 수요에도 리걸 테크에서 사용하는 법률 데이터는 양이 부족하고 정형화되어있지 않아 서비스의 성능 개선이 녹록지 않다. 본 연구에서는 인공지능 법률 질의응답 서비스 내 자연어 처리 모델의 성능 개선에 대해 다룬다. 실험을 위한 모델로 음주운전 사건 피의자에 합당한 형종을 예측하는 분류기를 선정했다. 분류기의 학습 데이터는 lbox-open[1] 내 데이터를 사용했다. BERT 모델로 형종 분류기를 구성하였고, 아무 기법을 적용하지 않았을 때 형종 예측 정확도는 71~73%로 나타났다. 그리고 라벨 불균형을 가중치 랜덤 샘플러로 해소하고, 요약 모델을 사용해 데이터를 증강시킨 결과 각각 75~85%, 80~92%의 정확도 상승을 확인했다. 위 연구를 통해, 판결문과 같이 정형화된 텍스트를 모델 학습에 사용할 경우 데이터 정제를 통해 성능 향상을 달성할 수 있음을 제시하고자 한다.

I. 서 론

고전적으로 법률 서비스는 다른 서비스보다 대면 영업의 비중이 크고 업무의 복잡도 또한 높은 편이다. 하지만 인공지능을 비롯한 기술의 발전을 통해 위 경향들이 어느정도 해결되었고, 그 결과 기술과 법률 서비스가 통합된 리걸 테크 산업이 대두되었다. 이런 리걸 테크는 법률 정보를 사용하기 쉬운 형태로 일반 사용자들에게 제시해 사용자로 하여금 법률 서비스를 이용하는 결정을 내리는 데 도움을 주기도 한다. 과거에는 '타이니 로(Tiny Law, 개인 간 분쟁, 소규모 법적 분쟁 등)'[2]에 대해 전문적인 법률 서비스를 필요로 하는 경향이 적었지만, 최근에는 일반인들의 법률에 대한 이해도와 관심도가 증가함에 따라 위와 같은 타이니 로에 대해서도 변호사를 선임하려는 분위기가 조성되어 있다.

하지만 이런 수요에도 불구하고, 인공지능 법률 서비스는 데이터 양이 적고 라벨 편향성이 큰 법률 도메인의 특성으로 인해 그 성능을 향상시키는 데 어려움이 있다. 따라서 본 연구는 리걸 테크 내 인공지능 법률 서비스의 성능 개선을 목표로 한다. 여러 종류의 서비스 중 음주운전 사건의 범죄 사실을 통해 형종(징역/집행유예/벌금) 이진 분류를 수행하는 분류기를 테스트 모델로 설정하였는데, 이 경우 다른 형사 사건에 비해 혈중 알코올 농도에 따른 형종 판정이 뚜렷하게 나타나는 사건이므로 성능 개선 방법론 적용 시 차이점이 도드라질 것으로 판단했기 때문이다.

따라서 위 분류기의 성능을 개선시키기 위해 두 가지 방법론을 적용하고 각각의 결과를 분석하려 한다.

II. 본론

이 단락에서는 데이터 정제 및 분류 성능을 개선하기 위한 방법론 적용 내용과 실험 설정에 대해 설명한다.

[전처리] 데이터셋 설정

본 연구에서는 lbox-open Github[1]에서 가져온 150k 개의 precedent_corpus 를 사용한다. 이 데이터 중 학습에 사용하기 위한 음주운전 사건을 정제하기 위해 두 가지 조건을 적용했다. 첫째로, 음주운전 사건을 다룰 예정이므로 관련 법률 제 44 조 제 1 항을 기준으로 데이터를 정제했다. 둘째로, 음주운전 법률 개정을 통해 혈중 알코올 농도에 따른 형종 판정 기준이 변하였으므로 최근 개정일(2019.6.25) 이후의 데이터를 정제했다. 그 결과, 전체 데이터 중 2838 개의 데이터로 줄어들었다. 해당 데이터 내 라벨별 분포는 아래와 같다.[표 1]

라벨	데이터 갯수
징역	333
집행유예	1967
벌금	538
계	2838

표 1. 음주운전 사건 판결문 데이터셋 내 라벨별 분포

[실험 1] Base 모델 선정 및 학습 데이터 분류

실험에서 사용할 분류 모델의 Base 모델은 BERT 모델 중 bert-base-multilingual-cased[3]을 사용했다. bert-large 나 다른 Base 모델을 사용하지 않은 이유는, 규모가 상대적으로 작을 경우 아래 기술될 개선 방법론의 영향을 크게 받기 때문이다. 위 모델을 기반으로 징역/집행유예/벌금 3 개 형종에 대한 각각의 이진 분류 모델을 제작하였고, 앞서 정제한 3384 개의 데이터를 분류하는 실험을 진행했다.

3 개의 이진분류기는 pytorch lightning[4]의 모듈을 이용했다. 분류기의 구조를 작성하기 위한 LightningModule 클래스와 데이터를 train/valid/test 셋으로 구분하기 위한 LightningDataModule 클래스를 적용했다.

[실험 2] 가중치 랜덤 샘플링으로 라벨 불균형 해소 후 분류

3 개의 이진 분류 모델의 입력은 [표 1]에서 언급한 것과 같이 어떤 값을 예측하느냐에 따라 라벨별 데이터가 균일하지 않다. 이 경우 3 개의 모델이 상대적으로 과적합이 발생하거나 local-minima 에 빠지는 오류를 범할 수 있다.

따라서 데이터셋 생성 클래스 내에 라벨 빈도의 역순에 따라 가중치를 부여하는 가중치 랜덤 샘플러를 추가하였다. 이를 통해 각각의 이진 분류기는 학습 시 참/거짓에 대한 데이터를 대등한 비율로 학습하여 편향적인 데이터 학습에서 발생하는 문제를 해결할 수 있다.

[실험 3] 요약 모델을 활용한 데이터 증강 후 분류

라벨 불균형 해소를 위해 가중치 기반 샘플링을 적용했으나, 이 경우 높은 빈도의 라벨 데이터 중 일부를 모델이 학습하지 못하는 문제가 발생한다.

이를 해결하고자 빈도가 낮은 라벨의 판결문 데이터를 증강하는 방법을 제시했다. 다양한 방법의 자연어 데이터 증강 방법이 있지만, 본 연구에서는 KoBART[5]를 적용한 텍스트 요약 모델을 이용해 증강하는 방법을 선택했다.

약 2000 자에 가까운 학습 데이터를 200/400/800 자 내외로 요약한 후 원본에 추가하여 사용하는 방법을 선택하였으며, 이 경우 빈도가 낮은 데이터는 최대 3 배 이상의 추가 데이터를 사용해 학습할 수 있다. 또한 이 경우 라벨별 출현 빈도 역전을 방지하기 위해 실험 2 의 가중치 랜덤 샘플링 기법도 적용하여 학습을 진행했다.

앞서 소개한 3 종류의 실험에 대해 가장 기본적인 성능 지표인 정확도(Accuracy)는 아래와 표와 같이 나타났다.

분류 모델	실험 1	실험 2	실험 3
징역형 이진분류	73.24	85.56	92.24
집행유예형 이진분류	72.01	75	91.67
벌금형 이진분류	71.81	83.27	80.11

표 2. 실험 3 종에 대한 테스트셋 정확도(Accuracy)

실험 3 의 경우 가장 높은 정확도를 달성하였고, 해당 모델의 정확도와 재현율(Recall), 정밀도(Precision), 그리고 재현율과 정밀도의 조화평균인 F1 score 는 아래와 같다.

분류 모델	징역	집행유예	벌금
Accuracy	92.24	91.67	80.11
Precision	81.98	98.79	78.95
Recall	98.72	88.5	75.72
F1 score	89.58	93.36	77.3

표 3. 실험 3 모델의 테스트셋 상세 성능지표

III. 결론

본 연구에서는 인공지능 모델을 활용한 음주운전 사건 형종 분류모델의 성능 개선을 라벨 불균형 해소 및 데이터 증강을 통해 달성하였다.

실험 1 에서는 초기 모델인 bert-base-multilingual-cased 를 기반으로 한 이진 분류기를 통해 71~73%의 정확도를 보였다. 실험 1 에서 고려되지 않은 라벨 불균형 문제를 라벨 출현 빈도별 역순으로 가중치를 부여해 실제 학습 시 참/거짓 데이터를 균등하게 학습할 수 있도록 실험 2 에 적용하였으며, 그 결과 기존 모델 대비 4~11% 개선된 성능을 나타냈다.

하지만 오버 샘플링을 하지 않는 이상, 가중치 랜덤 샘플링의 경우 상대적으로 빈도가 높은 라벨의 데이터는 학습이 제대로 일어나지 않을 수 있으므로, 빈도가 낮은 라벨의 데이터를 KoBART 를 기반으로 한 텍스트 요약 모델에 입력해 200/400/800 자로 데이터를 증강했다. 이 경우 낮은 빈도의 데이터의 경우 최대 3 배 이상의 데이터를 학습에 사용할 수 있게 되었고, 실험 결과 80~92%의 정확도를 나타내며 성능이 개선되었다. 또한 가장 높은 정확도를 달성한 실험 3 의 징역형 이진분류 모델의 경우, F1 score 도 89.58%를 달성하며 본 실험의 유의미함을 증명하였다.

다만 벌금형 이진분류의 경우, 실험 2 의 정확도가 실험 3 의 정확도보다 높게 나타난 것을 표 2 를 통해 확인할 수 있다. 이 경우, 파인튜닝하지 않은 KoBART 모델을 요약에 사용하는 과정에서 벌금형 데이터 중 주요 키워드가 소실되었을 가능성이 예상된다. 추후 특정 도메인에서 사용하는 요약 모델의 경우도 일정 수준의 파인튜닝 과정이 필요하다는 것을 나타냈다고 판단한다.

결론적으로 위 연구 내용은 라벨이 편향되어있으며, 전문적인 어체와 단어가 등장하며, 많은 양의 데이터가 없는 법률 도메인 연구에서 학습 데이터를 어떻게 준비할수 있는지에 대한 방향을 제시한다. 나아가 법률 데이터가 아니더라도 위 세 특징을 포함하는 다른 텍스트 데이터에 위 기법을 사용하는 것도 의미가 있을 것으로 보인다. 향후 위 방법론들은 음주운전이 아닌 다른 형사사건을 다룰 때 사용되거나, 나아가 형사사건보다 더 다양한 범주의 범죄사실이 발생하는 민사사건의 텍스트를 학습 데이터로 변환하는 경우에도 유용하게 쓰이기를 기대할 수 있다.

ACKNOWLEDGMENT

해당 연구는 KAIST 소프트웨어대학원프로그램의 교과 석사 졸업 프로젝트이며, 담당 지도교수인 KAIST 전산학부 신인식 교수와 KAIST 김재철 AI 대학원 서민준 교수의 자문을 바탕으로 진행되었다.

참 고 문 헌

[1] LBox Co.Ltd., lbox-open, GitHub repository, https://github.com/lbox-kr/lbox-open, 2022.

[2] Dominic Woolrych, “Why ‘tiny law’ excites this legal start-up leader”, https://www.thelawyermag.com/au/news/general/why-tiny-law-excites-this-legal-start-up-leader/208287, 2019.

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018

[4] Coleman, A., Parno, B., Howell, J., Grier, C., & Perl, H. PyTorch Lightning: Lightweight PyTorch wrapper for high-performance AI research. Journal of Open Source Software, 4(38), 2019.

[5] SK telecom, KoBART, GitHub repository, https://github.com/lbox-kr/lbox-open, 2020.