# 인공 신경망 기반 법익 침해 뉴스 여부 예측 모델

**최성환**
**서울대학교 컴퓨터공학부**

# News Article's Violation of Law Interest Predict Model Based on Artificial Neural Network

Seonghwan Choi
Dept. of Computer Science and Engineering, Seoul National University

**요 약**

미디어 경쟁 시대 황색언론의 전례 없는 난립에 의해, 언론에 의한 개인 및 사회의 법익이 침해되는 사례가 늘어나고 있다. 하지만 언론중재위원회를 통한 공식적 제재 수단은 위원회의 정기적 합의를 통해서만 이뤄지며 시의성이 떨어진다는 비판이 있다. 또한 기하급수적으로 늘어나고 있는 법익 침해 기사에 대해 적절히 대응할 수 있을지 의문이다. 이에 인공 신경망에 기반한 뉴스의 법익 침해 예측 모델을 구현하고자 한다. 언론중재위원회의 시정위원회 시정권고 기사 리스트를 기반으로, 기사 제목을 입력으로 하여 각 기사가 개인의 법익을 침해하는지, 사회의 법익을 침해하는지, 또는 법익을 침해하지 않는지 감독학습으로 학습한다. 학습 모델은 FNN, CNN, GRU, GRU+CNN 의 네 가지 경우로 구현하며, Word2vec 라이브러리를 통해 임베딩한다. 이를 통해 인공 신경망의 언론 생태계에의 적용 가능성을 확인하고자 한다.

## 1. 서 론

최근 1 인 미디어와 매체의 개인화가 일어남에 따라 미디어 생태계에 유입되는 정보의 총량이 늘어났다. 이는 국민의 알 권리와 대국민 집단으로서의 책임감 간에서 무거운 책임감을 느껴야 하는 언론으로 하여금 치열한 경쟁시장으로 이끌어 자극적, 선정적 기사를 쓰게 한다. 분명한 사회적 문제로, 다양한 형태로 개인과 사회적 법익을 침해한다. 본 프로젝트에서는 언론중재위원회의 언론 시정권고자료를 기반으로 기사제목을 통해 기사의 법익 침해 여부를 예측하는 인공 신경망 모델을 구현하고자 한다.

유사한 문제의식으로부터 시작된 기존 연구들은 대부분 뉴스의 진위 여부 판별에 집중한다. 뉴스의 다양한 법익 침해 양상 중에서 가짜뉴스 문제를 기계학습으로 해결하려는 시도로, 문서의 형태, 문서의 공유 네트워크 형태 및 공유 양상, 언어적 특성 기반 가짜뉴스 판별 등으로 나뉜다. 윤태웅과 안현철(2018)은 서울대학교 팩트체크센터의 가짜뉴스 데이터를 기반으로, 뉴스의 토픽과 관련인 등을 모델링해 SVM 으로 진위 여부 예측을 시도했다. 하지만 문장의 수치화에 있어, 단어의 빈도수에 기반한 TF-IDF 을 사용해 단어와 문장의 맥락을 반영하지 못한 한계가 있다. 심재승 외 2 명(2019)은 이 문제를 해결하기 위해 Doc2vec 라이브러리를 사용해 문장을 정량화했다. 하지만 서울대학교 팩트체크센터의 200 개 데이터를 사용해 학습하는 등, 아직까지 가짜뉴스에 대한 명확한 기준이 없고, 따라서 데이터가 부족해 인공 신경망 시스템이 갖는 의의가 높지 않다는 한계가 있다.

본 프로젝트에서 한정하는 문제는 다음과 같다. 대한민국 언론중재위원회에서 정기적으로 발표하는 시정권고소위원회 시정권고 심의안건 회의결과를 근거 데이터로 하며, 데이터는 기사제목, 카테고리, 연도로 한정한다. 법익 침해 유형을 [개인적 법익 침해, 사회적 법익 침해, 침해 없음]으로 구분해 예측한다. 이는 언론중재위원회 자료의 침해 여부 분류에 따른 것이다. '침해 없음'에 해당하는 데이터는 네이버 뉴스 사이트로부터 기간별로 법익침해가 있는 기사의 비율과 동일하게 추출해 사용한다.

자연어 데이터를 임베딩하기 위해 gensim 라이브러리의 Word2vec 을 사용한다. FNN, CNN, GRU, GRU+CNN 의 네 모델을 기반으로 감독학습으로 뉴스의 법익침해 여부를 학습한다. 인공 신경망 기반 예측 모델의 언론 생태계에 대한 적용의 방법론적 접근인 본 프로젝트를 통해 법익 침해 및 황색언론적 뉴스의 양상 및 특징을 파악하고 인공 신경망의 적용 가능성을 확인해보고자 하는 데 목표가 있다.

## 2. 모델 설계

### 2.1 데이터 전처리

법익침해 사례에 해당하는 데이터셋은 언론중재위원회 시정권고소위원회에서 정기적으로 발표하는 시정권고 심의안건 회의결과 보고서에서 뉴스 제목이 직접 제시된 2012 년 9 월 이후부터 2019 년 9 월까지의 자료로 확보했으며, 법익침해 해당사항이 없는 데이터셋은 동일 기간 네이버 뉴스 데이터 중 매월 초의 기사들을 크롤링했다. 데이터셋 크기는 법익침해 데이터셋 5,465 개, 일반 데이터셋 605,882 개이다.

전처리 과정에서 추후 카테고리에 의한 학습 데이터 균등 샘플링을 위해 두 데이터셋의 카테고리를 [정치, 경제, 사회, 생활문화, 세계, IT 과학, 그 외]의 일곱 개로 재정리했다. 법익침해 사례 데이터셋은 세분화돼 있는 법익침해 사항을 언론중재위원회의 분류에 기반해 [개인적 법익 침해, 사회적 법익 침해]의 두 가지로 통합했다. 법익침해 비해당 데이터셋은 법익침해 사항이 '없음'으로 분류하여 결과적으로 모든 데이터셋을 '개인적', '사회적', '없음'의 세 가지 분류로 라벨링했다. 이후 Mecab 형태소 분석 엔진으로 제목 데이터에서 형태소를 추출, 불용어를 제거했다. 모든 처리가 끝난 두 데이터셋으로부터 각각 20%의 데이터에 평가 데이터로 사용할 것임을 태깅했다.

데이터셋의 임베딩은 gensim 라이브러리의 Word2Vec 엔진을 사용했다. 한 데이터셋의 토큰 갯수를 10 개로 제한하고, 10 개보다 긴 데이터는 절삭, 작은 데이터는 앞 토큰의 반복을 통해 길이를 맞췄다. Word2Vec 을 통해 하나의 단어를 32 원소를 갖는 벡터로 대응했다.

### 2.2 학습 데이터 선택

현재 데이터셋은 법익침해 여부에 따라 크기의 편차가 매우 커서 그대로 학습을 진행하면 모델이 법익침해 없음에 해당하는 분류로 편중될 가능성이 있다. 이에 학습 데이터로 뉴스기사의 시의성과 카테고리 특이성을 고려해 법익침해 비해당 데이터셋에서 법익침해 해당 데이터셋의 년도와 카테고리를 축으로 하여 같은 비율만큼 랜덤 샘플링했다. 랜덤 샘플링은 2.1 에서 평가 데이터셋으로 태깅되지 않은 대상에 한정했다.

또한 이렇게 학습 데이터셋을 선택함으로 인해 생기는 법익침해 비해당 데이터셋의 낭비를 줄이고 오버피팅을 방지하고자 학습 과정에서 주기적으로 데이터셋의 샘플링을 다시 했다. 즉, 매 학습마다 법익침해 해당 데이터셋은 동일하지만 법익침해 비해당 데이터셋은 주기적으로 균등비율 샘플링되어 달라지도록 했다. 샘플링 비율은 Figure 1 의 표를 따른다.

| Year/ Category | 정치 | 경제 | 사회 | 생활문화 | 세계 | IT 과학 | Others |
|---|---|---|---|---|---|---|---|
| 2012 | 0 | 0 | 60 | 0 | 0 | 0 | 0 |
| 2013 | 0 | 0 | 223 | 25 | 0 | 0 | 0 |
| 2014 | 3 | 0 | 261 | 15 | 2 | 0 | 0 |
| 2015 | 18 | 29 | 213 | 61 | 12 | 14 | 3 |
| 2016 | 60 | 16 | 363 | 237 | 17 | 11 | 12 |
| 2017 | 54 | 51 | 471 | 186 | 20 | 20 | 28 |
| 2018 | 88 | 23 | 375 | 492 | 32 | 8 | 7 |
| 2019 | 104 | 28 | 415 | 262 | 42 | 6 | 5 |

**Figure 1) 법익침해 해당 데이터셋의 카테고리, 연도별 데이터 수**

### 2.3 모델 및 실험 설계

본 연구에서는 FNN, CNN, GRU, GRU+CNN 혼합 모델을 연구 대상으로 한정했다. CNN 모델의 경우 Convolution Kernel 사이에 1*1 Convolution Kernel 을 두어 채널 수를 줄임으로써 연산양을 줄였다. 모든 모델에 Adam Optimizer 를 사용했으며 Xavier 가중치 초기화를 적용해 적절한 초기조건 조성을 꾀했다. 또한 모든 모델의 FNN 층에 p=0.5 의 드랍아웃을 적용했다.

각 모델은 특성에 따라 다른 모수들을 사용하되, 10 번의 학습 데이터 샘플링에 대해 10 번의 에폭으로 학습을 진행했다. 설계한 모델을 바탕으로 전처리 과정에서 지정해둔 전체 데이터셋 중 따로 태깅한 20% 데이터셋을 평가 데이터셋으로 하여 정확도를 평가했다(In-Vocab 데이터). 또한 새로운 단어가 포함된 데이터셋에 대한 성능을 평가하기 위해 2019 년 10, 11 월 언론중재위원회 시정권고 심의안건 회의결과 보고서 데이터셋으로 구성된 평가 데이터셋을 추가 구성해 평가했다(Out-of-Vocab 데이터).

## 3. 실험 결과

### 3.1 모델별 학습 추이

모델별 학습 데이터셋 손실, 평가 데이터셋 손실의 변화 추이는 Figure 2 와 같다. 사용한 평가 데이터셋은 In-Vocab 데이터셋이다. FNN 의 경우 가장 안정되고 편차가

적은 하강 양상을 보였으며, GRU 가 적용된 두 모델은 비슷한 하강 양상을 보였다. 주목할 만한 부분은 CNN 모델의 하강 양상인데, 다른 모델에 비해 높은 값에서 손실값에서 수렴했으나 학습 데이터셋보다 평가 데이터셋에 더 잘 적합해 뛰어난 일반화 성능을 보였다. 또한 FNN 모델에서는 데이터셋이 새로 샘플링될 때마다 학습 데이터셋 손실은 소폭 상승, 평가 데이터셋 손실은 소폭 하락하는 양상을 보였는데, 이는 타 모델에서는 보이지 않는 양상으로 FNN 모델이 현재 학습하고 있는 데이터셋에 대해 과적합하고 있음을 보여준다.

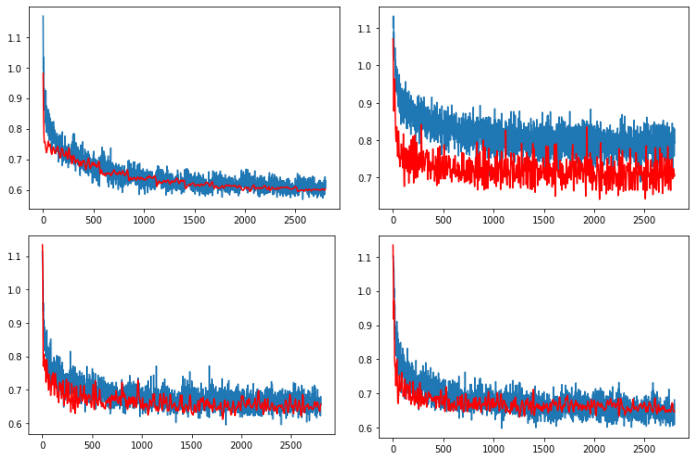

**Figure 2) Train Loss(Blue), Test Loss(Red) – 왼쪽 위부터 시계 방향으로 FNN, CNN, GRU+CNN, GRU**

### 3.2 In-Vocab 데이터 성능 평가

모델별 In-Vocab 데이터셋에 대한 예측성능 평가 결과는 Figure 3 과 같다.

| Model | Params | Accuracy | |
|---|---|---|---|
| | | Train | Test |
| FNN | lr=0.002, batch=240 | **96.46%** | **94.80%** |
| CNN | lr=0.002, batch=240 | 80.62% | 83.82% |
| GRU | lr=0.005, batch=240 | 90.61% | 89.35% |
| GRU+CNN | lr=0.005, batch=240 | 92.00% | 90.41% |

**Figure 3) 모델별 In-Vocab 데이터셋 예측 정확도**

FNN 모델의 경우 학습 및 평가 데이터셋 모두에서 가장 좋은 예측 성능을 보였다. GRU 가 적용된 두 모델의 경우 유사한 성능을 보였다. CNN 모델은 타 모델과는 달리 학습 데이터셋보다 평가 데이터셋에서 더 나은 정확도를 보였다. 3.1 의 결과와 같은 맥락에서 CNN 모델의 뛰어난 일반화 성능을 확인할 수 있다.

### 3.3 Out-of-Vocab 성능 평가

임베딩 과정에서 OOV 문제 발생 시 해당 단어를 하나의 더미 단어에 대응해 변환했다. 모델별 Out-of-Vocab 데이터셋에 대한 예측성능 평가 결과는 Figure 4 와 같다. 모델 비교를 위해 모든 데이터에 대해 3 개 라벨 중 하나로 동일하게 예측하는 모델을 Baseline 으로 둔다.

| Model | Params | Accuracy |
|---|---|---|
| FNN | lr=0.002, batch=240 | 32.32% |
| CNN | lr=0.002, batch=240 | **57.93%** |
| GRU | lr=0.005, batch=240 | 47.56% |
| GRU+CNN | lr=0.005, batch=240 | 52.44% |
| Baseline | - | (추정)33.33% |

**Figure 4) 모델별 Out-of-Vocab 데이터셋 예측 정확도**

Out-of-Vocab 데이터셋에서 FNN 모델의 성능이 가장 낮게 측정됐다. FNN 모델은 Baseline 보다 낮은 정확도를 보이며 데이터의 중요 단어 단서에 대해 민감하게 반응했다. CNN 모델은 In-Vocab 데이터셋에 대한 예측 정확도가 가장 낮았던 것과는 반대로 가장 높은 정확도를 보였다. CNN 모델의 가중치 공유에 기반한 필터링 특성에 의해 모델이 데이터의 세부 사항보다 맥락의 파악에 중점을 두게 된 것으로 보인다.

GRU+CNN 모델의 경우도 GRU 단독 모델에 비해 높은 정확도를 보였다. 하지만 모든 모델에서 In-Vocab 평가 데이터셋에 비해 큰 폭으로 성능이 하락했다.

## 4. 결론 및 논의

본 연구에서는 인공신경망 기반 예측 모델 중 FNN, CNN, GRU 를 빌딩블록으로 사용해 모델 특성을 살펴보았다. 본 연구에서의 모델 성능은 Well-formed 된 환경(In-Vocab 환경, 손실 하강에서의 낮은 분산으로 대표되는 안정성 등)에서의 예측 성능, 과적합 방지의 두 가지로 나눌 수 있다. 또한 과적합은 두 가지 항목으로 나누어 볼 수 있는데, 첫째는 In-Vocab 데이터 내 학습/평가 데이터셋 간의 과적합 문제이며 둘째는 Out-of-Vocab 단어가 발생했을 때 강건하게 반응하는지 여부이다.

FNN 모델은 Well-formed 환경에서의 예측 성능은 가장 좋았으나 과적합 문제도 가장 크게 발생했으며 특히 OOV 문제 발생 시 성능이 크게 하락했다. FNN 의 특성상 입력 데이터의 전체 맥락보다는 특정 단어 등의 세부사항에 집중하게 되어 환경 변화에 대한 민감도가 높아진 것으로

보인다. CNN 모델은 과적합 문제에 잘 대응했으나 예측 성능은 다른 모델에 비해 떨어졌다. FNN 과 CNN 모델을 비교함으로써 과적합과 예측 성능 간의 Trade-off 가 있음을 확인했다.

GRU 모델은 이 Trade-off 를 적절히 해결했다. 과적합 문제에서 FNN 에 비해 월등히 나으면서도 예측 성능은 크게 떨어지지 않았다. 또한 3.2 에서 GRU+CNN 모델의 예측 성능이 GRU 단독 모델에 비해 나은 나은 모습을 보임을 통해 GRU+CNN 의 조합이 과적합 방지 및 모델 성능 모두의 향상을 꾀할 수 있다는 결론이 도출 가능했다.

본 연구는 NLP 문제에 있어 여러 인공신경망 구조들의 장단점을 파악했다는 점에 의의가 있다. 또한 OOV 문제에 대해 CNN 이 대응 방법이 될 수 있음을 보였으며, NLP 문제의 정확도-과적합 Trade-off 문제를 사례를 통해 다루었다는 점에 의의를 둔다.

NLP 에 있어 OOV 문제는 필연적이다. 특히 본 연구에서 다룬 뉴스 데이터셋의 경우 시의성 있는 단어가 주로 사용되어 신조어나 단어의 의미 변화에 민감하다. 뉴스 데이터에서의 OOV 해당 단어는 특히 인명, 사건명 등 해당 데이터의 핵심 정보를 포함하는 경우가 많아 기존의 더미벡터나 영벡터, 주변 단어로부터의 추정벡터 사용 등의 방법으로 뉴스 데이터가 좋은 성능을 내지 못할 것임을 생각할 수 있다. 실제로 본 연구에서도 더미벡터를 사용한 OOV 해당 단어의 임베딩으로 좋은 예측결과를 얻지 못했다. 결국 시의성이 작용하는 데이터의 경우 임베딩 모델의 지속적인 Fine-tuning 을 통한 업데이트가 필수적이며, 이는 주기적인 예측 모델의 재학습을 야기한다는 한계가 있다.

**참고문헌**

[1] 윤태욱, 안현철 (2018). 텍스트 마이닝과 기계 학습을 이용한 국내 가짜뉴스 예측. Journal of Information Technology Applications & Management, 25(1), 19-32.

[2] 심재승, 원하람, 안현철 (2019). 국내 뉴스 기사를 활용한 Doc2vec 기반 지능형 가짜뉴스 분류모델. 한국지능정보시스템학회 학술 대회논문집, 1-3.