문서 편향성 분류 모델 구축 및 가짜 뉴스 탐지에서의 활용*

이재훈○ 김미숙**
세종대학교
jm05222@naver.com, misuk.kim@sejong.ac.kr

Classification Model of Biased Documents and Its Application in Fake News Detection*

Jaehoon Rhee ○, Misuk Kim**
Sejong University

요 약

무분별한 대량의 정보가 제공되는 현대사회에서는 객관적이고 중립적인 정보에 대한 필요성이 강조되고 있다. 그 중에서도 가짜뉴스는 최근 COVID-19에 관련한 잘못된 정보들로 인해 사회적 혼란이 가중되면서 그 심각성이 대두되었다. 따라서 본 연구에서는 정보의 객관성과 중립성을 판단하는 모델을 구축하고, 이를 가짜뉴스탐지에도 적용하여 문서의 언어적 편향성과 가짜뉴스 사이의 유의미한 상관관계를 도출하였다.

1. 서 론

무분별한 대량의 정보가 제공되는 현대사회에서 정보의 객관성 혹은 중립성을 판단하는 모델을 구축하는 것은 중요한 연구분야이다. 이러한 모델을 구축하기 위해서는 문서의 편향성 여부가 레이블링된 데이터가 필요한데, 위키피디아가 해당 정보를 일부 제공하고 있다. 위키피디아 문서 편집에 적용되는 Neutral Point of View (NPOV)[1]는 중립적 시각을 의미하는 용어로 2001년 2월 16일에 최초로 만들어져 지금까지 유지되고 있는 정책이다. NPOV는 객관적이고 중립적인 시각으로만 정보를 제공함으로써 독자들의 판단에 어떠한 개입도 하지 않고 사실전달에만 목적으로 두고 있다. NPOV 정책 다양한 기준이 있는데, (1)의견을 사실처럼 쓰지 말 것, (2)논란의 여지가 많은 주장을 사실처럼 쓰지 말 것, (3)사실을 의견처럼 쓰지 말 것, (4)단정적인 단어는 피할 것, (5)다양한 관점을 비중에 맞춰 설명할 것이 있다. 이러한 편향된 정보가 많이 반영된 문서로는 가짜뉴스(Fake News)를 고려해 볼 수 있다. 가짜뉴스는 잘못 알려진 정보, 조작된 정보들을 포함한 뉴스 등을 말한다. 오늘날에는 잘못된 정보가 포함된 수많은 소식이 뉴스나 SNS를 통해 전파되기 때문에 대부분의 사람들이 정보들을 여과없이 받아들이게 되어 가짜뉴스의 심각성이 대두되고 있다. 실제로 2020년 4월에는 5G 전자파가 인체면역을 약화시키고 코로나19를 확산시킨다는 가짜뉴스가 돌아 영국에서는 통신탑 및 방송기지국 방화사건이 일어나기도 하였다. 즉, 정보를 받아들이는 입장에서는 다방면의 검증된 배경지식을 갖고 있지 않는 이상 정확하게 가짜뉴스를 탐지하기는 쉽지 않고, 전문가의 인력을 동원하여 탐지한다 하더라도 상당한 비용과 시간이 소모된다.

따라서 본 연구는 NPOV 정책으로 구축된 데이터를 활용하여 편향된 정보가 포함되어 있는지에 대한 여부를 판단하는 모델을 구축하고, 이를 가짜뉴스를 탐지하는 모델로 응용하고자 한다. 이는 뉴스에 대해 사실 검증을 위한 배경지식 없이도 가짜뉴스를 탐지의 척도로 활용될 수 있다는 것을 정량적으로 확인하였다.

2. 관련 연구

국외에서는 위키피디아의 NPOV 정책으로 구축된 데이터를 기반하여 다양한 연구가 수행되고 있다. 대표적으로 NPOV를 위반한 유형별로 분류하는 연구[2], 분류와 더불어 NPOV를 위반을 유발하는 구절을 나타내고 수정 문장을 생성하는 연구[3]가 있다.

3. 연구 방법

3.1. 데이터 증강 (Data augmentation)

An Easier Data Augmentation(AEDA)[4]는 텍스트 데이터 증강의 방법 중 하나로, 텍스트 데이터 사이에 랜덤하게 문장부호를 삽입하는 방식이다. 이 방식은 텍스트의 의미를 훼손하지 않으면서 간단하게 데이터 증강이 가능하는 장점이 있다. 본 연구에서는 위 방식을 통해 학습에 활용될 데이터의 크기를 3배로 늘려 모델

* This research was supported and funded by Institute of Information & Communications Technology Planning & Evaluation (No. 2021000469).
** corresponding author

학습에 활용하였다.

3.2. BERT

Bidirectional Encoder Representations from Transformers(BERT)[5]는 Google에서 발표한 사전학습 언어모델이다. 본 연구에서는 BERT-base 사전 학습 모델을 이용하여 3.1. 절에서 구축한 학습데이터를 다양하게 파인튜닝에 활용하여 문장의 NPOV위반 여부를 판단하는 분류모델을 구축하였다.

3.3. T-test

T-test 는 두 집단의 평균을 비교하는 모수적 통계방식이다. 본 연구에서는 두 집단이 통계적으로 유의미하게 차이가 나는지를 검사하는 방법으로 활용하였다. 어떤 집단을 서로 비교하는지에 대해서는 4.3에서 자세히 설명하였다.

4. 데이터셋

4.1. WIKIBIAS

WIKIBIAS corpus는 [3]에서 소개된 데이터셋이다. 이 데이터셋은 위키피디아의 문서 수정 목록에서 NPOV에 관련한 문장 수정기록들을 수집하여 수정 전 문장은 1, 수정 후 문장은 0으로 레이블링 한 데이터 셋으로 예시는 표 1과 같다.

표 1. WIKIBIAS 데이터 예시

sentence	label
Evolution is the source of the vast diversity of extant and extinct life on Earth.	1
Evolution may be the source of the vast diversity of extant and extinct life on Earth.	0

특히 그 중에서도 WIKIBIAS-manual은 수집된 데이터 중에서 NPOV에 관련된 데이터인지에 대한 여부를 주석자들을 통해 한 번 걸러진 데이터셋이다. NPOV와 관련 없는 문장 쌍들은 모두 0으로 레이블링하였다. 훈련셋, 검증셋, 테스트셋을 포함하여 약 8,198개의 문장이 있으며, 본 연구에서는 훈련셋과 검증셋을 합쳐 약 6,094개를 데이터 증강과 모델 훈련 및 검증에 활용하였고, 테스트셋은 모델 성능 평가의 척도로 활용하였다. 또한 훈련 데이터셋에 필터링 되지 않은 잡음 데이터 10,000개를 추가하여 학습에 사용하였다.

4.2. Politifact

Politifact[6]는 다양한 매체에서 전파된 뉴스의 진실성을 총 6개의 클래스로 레이블링한 데이터셋이다. 뉴스의 출처, 발언자 등과 같은 meta-data도 포함되어 있으며, 클래스는 ‘pants-fire’, ‘false’, ‘barely-true’, ‘half-true’, ‘mostly-true’, ‘true’ 로 6개의 class로 구성되어 있다. 표 2는 Politifact 데이터 예시를 나타낸다.

표 2. Politifact 데이터 예시

document	class
Black Lives Matter of Atlanta Charged with Wire Fraud, Money Laundering and Allegedly Using almost 500k in Donations For Personal Use.	false
If you look at the average teacher pay compared to the average pay of your citizens, Virginia ranks last.	true

이 중에서 news의 출처가 영상이나 이미지인 경우에는 내용을 묘사하여 만들어진 문장들이기 때문에 제외하였다. 본 연구에서는 Politifact 데이터를 NPOV 정책 위반 여부와 가짜뉴스 사이의 상관관계를 분석하는데 활용하였다.

5. 실험 및 실험결과

5.1. 실험 설정

본 실험에서는 NPOV 정책 위반 문장 분류를 위해 BERT-base모델을 사용하였으며, 배치사이즈 32, 입력 문장 최대길이 128, 최적화 함수는 BertAdam, 학습률 1e-5로 설정하여 모델을 학습하였다.

5.2. 분류모델 성능

본 실험에서는 진행한 실험에서는 BERT에 원본 데이터만 활용한 모델(BERT), 원본데이터에 증강데이터를 추가한 모델(BERT_aug), 원본데이터와 증강데이터에 잡음데이터까지 추가한 모델(BERT_aug_noise)의 성능을 실험하였으며, 결과는 표 3과 같다.

표 3. Bias 여부 분류 모델 성능

model	label	precision	recall	f1	support	acc.
Base model[3]	-	0.70	0.39	0.52	-	0.68
BERT	0	0.70	**0.87**	**0.78**	1252	0.70
	1	0.71	0.46	0.55	852	
BERT_aug	0	**0.76**	0.69	0.72	1252	0.69
	1	0.60	**0.67**	**0.63**	852	
BERT_aug_noise	0	0.72	0.84	0.77	1252	**0.71**
	1	0.69	0.52	0.59	852	

실험결과 분류 정확도는 BERT_aug_noise가 71%로 가장 높았다. [3]에서 기존 실험의 정확도가 68%로 나온 것에 비해 더 나은 성능을 보였다. 또한 [3]의 실험에서 약 42만개의 noise데이터를 추가하여 약 71%의 정확도를 기록한 반면, 본 실험에서는 증강데이터 약 12,000개, noise데이터 10,000개 만을 추가하여 71%의 정확도를 달성하였다.

5.3. 가짜뉴스 편향성 여부

Politifact 내의 표본들은 하나 이상의 문장으로 구성되어 있다. 따라서 본 실험에서는 각각의 문장들에 대해 분류하고 같은 문서내 문장들끼리의 평균을 문서에 대한 예측값으로 설정하였다 (그림 1).

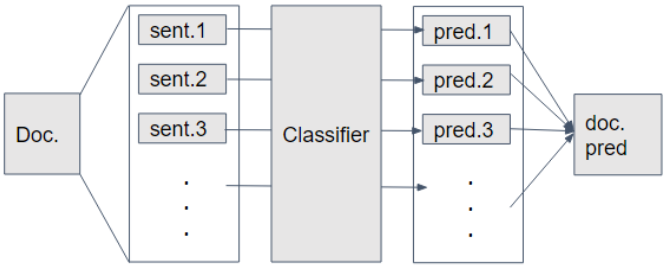

그림 1. 문서의 편향성 계산 프레임워크

Class에 따라 예측값들의 평균과 표준편차는 표 4와 같다.

표 4. class 별 편향성 값

model	statistics	pants fire	false	barely true	half true	mostly true	true
BERT	mean	0.054	0.046	0.045	0.042	0.047	0.047
	std	0.215	0.199	0.198	0.192	0.203	0.204
BERT_aug	mean	0.376	0.360	0.376	0.370	0.356	0.346
	std	0.468	0.462	0.468	0.465	0.462	0.46
BERT_aug_noise	mean	**0.164**	**0.159**	**0.138**	**0.134**	**0.135**	**0.127**
	std	0.354	0.349	0.330	0.326	0.328	0.322

세가지 모델에 대해서 모두 예측값들의 평균이 거짓에 가까울수록 높은 값을 갖는 경향을 보인다. 또한 분류성능이 가장 좋았던 BERT_aud_noise가 가장 뚜렷한 결과를 나타내었다. 여기서 표준편차가 높게 측정된 것은 문서의 86%가 한 개의 문장으로 구성되어 있어 예측값들과 평균의 오차가 크기 때문이다. 그럼에도 class별 경향성이 유의미한 차이가 있는지 확인하기 위해 각 class 별 예측값의 평균의 차이를 T-test를 통해 검증하였고 그 결과는 표 5와 같다.

표 5. 각 class pair별 T-test 결과

p-value (< 0.05)	pants fire	false	barely true	half true	mostly true	true
pants fire	1					
false	0.5476	1				
barely true	**0.0062**	**0.0111**	1			
half true	**0.0014**	**0.0020**	0.6549	1		
mostly true	**0.0024**	**0.0038**	0.7490	0.9031	1	
true	**0.0002**	**0.0003**	0.2167	0.3985	0.3482	1

실험 결과는 [‘Pants-fire’, ‘false’]와 [‘barely-true’, ‘half-true’, ‘mostly-true’, ‘true’] 사이에 유의미한 차이가 있음을 보인다. 이는 가짜뉴스가 정상적인 뉴스에 비해 NPOV 정책을 위반한 문장이 비교적 많다는 것을 의미한다.

6. 결론 및 향후 연구

본 연구에서는 WIKIBIAS corpus 데이터를 활용하여 데이터를 증강하는 방식으로 기존 연구에 비해 더 좋은 분류성능을 보이는 모델을 구축하였다. 또한 이를 가짜뉴스 탐지에 활용하여 class 별 편향성 값을 계산하였고, 각 class 별 편향성 정도가 유의미하게 차이나는 것을 T-test를 통해 확인하였다.

향후에는 여러 텍스트 데이터 증강 기법을 활용하여 학습 데이터를 구축하거나 다양한 언어모델을 활용하여 편향성을 판단하는 분류기의 성능을 향상시키는 연구로 발전할 수 있을 것으로 기대한다.

7. 참고문헌

[1] Wikipedia: Neutral point of view. (16 April 2022). Retrieved April 28, 2022, From Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.

[2] Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013, August). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1650-1659).

[3] Zhong, Y., Yang, J., Xu, W., & Yang, D. (2021, November). WIKIBIAS: Detecting Multi-Span Subjective Biases in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1799-1814).

[4] Karimi, A., Rossi, L., & Prati, A. (2021, November). AEDA: An Easier Data Augmentation Technique for Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2748-2754).

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

[6] Vo, N., & Lee, K. (2020, November). Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7717-7731).