통합적 감정·편향·진위 분석을 통한 중립적 재작성

양혜원, 최원교, 김태완
동덕여자대학교 데이터사이언스전공
e-mail : {20221631, 20221645, kimtwan21}@dongduk.ac.kr

Neutral Rewriting of News Headlines via Integrated Emotion, Bias, and Veracity Analysis

Hye-Won Yang, Won-Kyo Choi, Tae-Wan Kim
Data Science Major, Dongduk Women's University

요 약

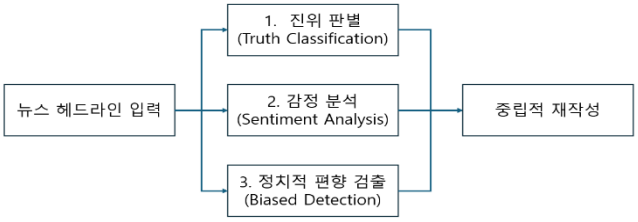
본 논문은 영어 뉴스 헤드라인을 대상으로 진위 판별, 감정 분석, 정치적 편향 검출, 중립적 재작성을 통합한 자연어처리 파이프라인을 제안한다. 각 단계별로 DistilBERT, DeBERTa-v3, T5-base 모델을 비교·선택하였으며, 진위·감정·편향 분석 결과를 확률 값 현태로 T5 재작성 모델에 입력하여 의미적 중립화를 수행한다. 최종 출력은 감정적 표현과 이념적 왜곡이 완화된 'Neutralized Sentence' 형태로 제공된다. 또한, 분석 결과를 레이더 차트로 시각화하여 뉴스별 신뢰도, 감정 강도, 편향 정도를 직관적으로 확인할 수 있으며, 재작성 과정의 해석 가능성을 높였다. 본 연구는 뉴스 콘텐츠의 객관적 평가와 중립적 표현 자동화를 위한 기반을 제공하며, 자국어 뉴스나 실시간 뉴스 스트림에도 확장 가능함을 보여준다.

1. 서론

최근 온라인 뉴스 및 SNS 를 통합 정보 확산이 가속화되면서 가짜 뉴스(Fake News), 편향적 보도(Biased Reporting)로 인한 사회적 문제가 심화되고 있다. 특히 짧은 뉴스 헤드라인은 정보의 출처나 진위를 충분히 파악하기 어렵고, 감정적 또는 정치적 언어 사용 시 사실 왜곡이 발생하기 쉽다.

이에 본 연구는 영어 뉴스 헤드라인을 대상으로 진위 판별(Fake/Real), 감정 분석(Positive/Neutral/Negative), 정치적 편향 검출(Left/Center/Right)을 수행한 뒤 중립적 재작성(Neutral Rewriting)을 진행하는 통합 자연어처리 파이프라인을 제안한다. 본 연구의 목표는 다단계 분석을 통해 뉴스 헤드라인의 신뢰성과 객관성을 향상시키는 것이다. 이를 위해 각 단계별로 두 가지 이상의 언어모델을 비교 실험하고, 최적의 성능을 보인 모델을 파이프라인에 통합하였다.

기존 연구는 뉴스의 진위 판별, 감정 분석, 정치적 편향 검출, 중립적 재작성 등을 각각 독립적으로 다루어 왔다. 그러나 이러한 단일 과제 접근은 언어적 편향, 감정, 사실성 간의 상호 연관성을 포괄적으로 이해하는 데 한계가 있다. 본 연구는 이를 보완하기 위해 진위·감정·편향 분석을 단계적으로 수행한 후, 그 결과를 기반으로 중립적 재작성을 수행하는 통합 파이프라인을 구축하였다.



2. 연구 데이터 및 학습 설정

2.1 진위 판별 (Fake News Detection)

진위 판별 단계에서는 FakeNewsNet(Shu et al., 2020) 데이터셋을 활용하였다. 이 데이터셋은 PolitFact 와 GossipCop 두 플랫폼에서 수집된 뉴스 기사와 관련 트위터 데이터를 포함하며, 각 기사는 진짜 혹은 가짜로 구분되어 있다. 전체 데이터는 학습 18,556 개, 테스트 4,640 개로 구성하였으며, 진위 여부는 fake_label(0:real, 1:fake)로 병합하였다.

모델로는 DistilBERT 와 BERT-base 를 모두 적용하여 성능을 비교하였다.

모델	Epoth	배치	학습 손실(loss)	테스트 정확도
DistilBERT	2	16	0.252 → 0.205	0.8498
DistilBERT	3	16	0.135 → 0.193	0.8593
DeBERTa-v3-base	3	16	0.279 → 0.318	0.8580

비교 결과, 두 모델 모두 85~86% 수준의 테스트 정

확도를 보여 유사한 성능을 나타냈다. 그러나 DistilBERT 는 학습 속도와 자원 효율 측면에서 우수하였고, 이를 기반으로 최종 파이프라인에 적용하였다. 이러한 비교 실험을 통해 모델 선택의 근거를 명확히 하고, 향후 연구에서 재현 가능한 모델 검증 절차(Valid evaluation process)를 확보하였다.

2.2 감정 분석 (Sentiment Analysis)

감정 분석 단계에서는 TweetEval Sentiment 서브셋을 사용하였다. 이 데이터셋은 트윗 문장을 대상으로 긍정(positive), 중립(neutral), 부정(negative)의 세 가지 감정 클래스로 분류된다. 공식 데이터 구성에 따라 학습 45,615 개, 검증 2,000 개, 테스트 12,284 개 샘플을 사용하였다.

모델 성능 비교를 위해 DistilBERT 및 DeBERTa-v3-base 두 가지 모델을 실험하였으며, 학습 Epoch 수에 따른 성능 변화를 관찰하였다.

모델	Epoth	배치	학습 손실(loss)	테스트 정확도
DistilBERT	3	16	0.636 → 0.401 → 0.426	0.664
DistilBERT	4	16	0.489 → 0.0632	0.6619
DeBERTa-v3-base	3	16	0.576 → 0.283 → 0.525	0.72

비교 결과, DeBERTa-v3-base 모델이 전체적으로 가장 높은 정확도(72%)와 균형 잡힌 정밀도・재현율・F1-score(모두 0.72)를 보였다.

DistilBERT 는 상대적으로 가벼운 구조로 학습 효율성이 높았으나, 최종 성능은 DeBERTa 에 비해 다소 낮았다. 따라서 최종 파이프라인에는 DeBERTa-v3-base 모델을 감정 분석 모듈로 채택하였다.

2.3 정치적 편향 검출 (Political Bias Detection)

정치적 편향 검출 단계에서는 입력 데이터 구성에 따라 모델 성능 차이를 비교하였다. 타이틀과 헤딩만 병합한 DistilBERT 모델은 학습 손실이 Epoch 진행에 따라 다소 증가하며 학습이 불안정한 모습을 보였고, 테스트 정확도는 0.443 으로 낮게 나타났다. 반면 타이틀, 헤딩, 본문을 모두 포함한 DistilBERT 모델은 학습 손실이 안정적으로 수렴하며 테스트 정확도가 0.600 으로 약 15%p 향상되었다. ALBERT-base-v2 모델 또한 동일한 데이터 구성에서 학습 손실이 소폭 증가하였고, 테스트 정확도는 0.535 로 나타났다.

모델	데이터 구성	Epoth	배치	학습 손실(loss)	테스트 정확도
DistilBERT	타이틀 + 헤딩	3	16	0.894 → 0.983	0.443
DistilBERT	타이틀 + 헤딩 + 본문	3	16	1.03 → 1.00	0.600
ALBERT-base-v2	타이틀 + 헤딩 + 본문	3	16	0.978 → 1.06	0.535

이를 바탕으로 본 연구에서는 성능과 안정성을 고려하여 DistilBERT 모델(타이틀+헤딩+본문)을 최종 편향 검출 모델로 선택하였다.

2.4 중립적 재작성(Neutral Rewriting)

중립적 재작성 단계에서는 T5-base 모델을 이용한 단계적 파인튜닝(stepwise fine-tuning)을 수행하였다. Step 1 에서는 1,000 개씩 누적 학습(Batch 1~4)으로 학습을 진행하였고, 학습 손실과 검증 손실 모두 점진적으로 감소하며 안정적인 수렴을 보였다. Step 2 에서는 데이터 수를 5,000 개로 늘리고 시드를 고정하여 재학습을 수행하였으며, 학습 손실은 감소하고 검증 손실은 안정적으로 유지되었다. Step 3 에서는 데이터 수를 10,000 개로 확장하고 시드 고정을 해제하여 재학습한 결과, 학습 및 검증 손실 모두 안정적인 범위에서 유지되었으며, ROUGE-1 과 ROUGE-L 점수 또한 의미적 정보 손실 없이 중립적 재작성 품질이 일정 수준으로 유지됨을 확인하였다.

단계	데이터 구성	Epoth	학습 손실(loss)	검증 손실(loss)	ROUGE-1	ROUGE-L
Step 1	1,000개씩 누적 학습 (Batch 1~4)	1~4	0.4565 → 0.3842	0.3656 → 0.3012	1.46 → 1.30	1.43 → 1.31
Step 2	5,000개 (train 4,750 / test 250, 시드 고정)	3	0.3543 → 0.2835	0.3166 → 0.3251	1.72 → 1.59	1.72 → 1.60
Step 3	10,000개 (시드 고정 X)	2	0.2734 → 0.3081	0.3066 → 0.3081	1.58 → 1.63	1.59 → 1.63

3. 통합 파이프라인 설계

본 연구의 통합 파이프라인은 네 단계로 구성된다. 먼저 뉴스의 진위를 판별하는 진위 분류(Truth Classification), 이어 뉴스의 감정을 분석하는 감정 분석(Sentiment Analysis), 그리고 뉴스의 정치적 편향을 평가하는 편향 검출(Bias Detection)이 수행된다. 마지막으로 이전 단계의 결과를 종합하여 뉴스 문장을 중립적이고 사실적인 형태로 변환하는 중립적 재작성이 이루어진다.

이전 세 단계의 결과값 – 즉 진위(truth), 감정(sentiment), 편향(bias) – 은 각각 확률 값(confidence) 형태로 산출되어 T5 재작성 모델의 프롬프트 입력으로 통합된다. 모델은 다음과 같은 지시형 프롬프트

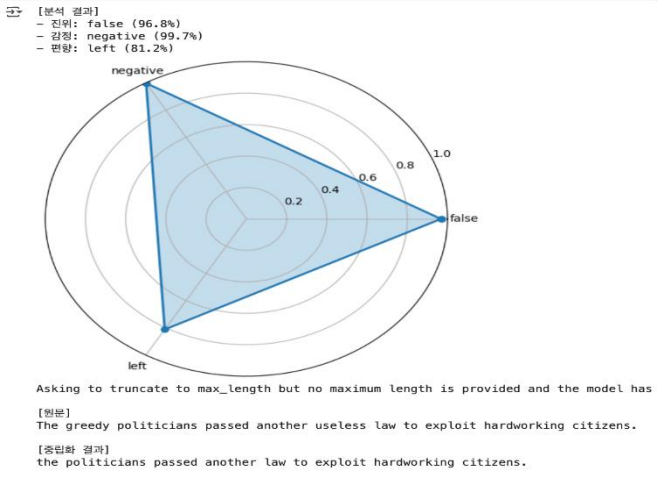
> “Rewrite the following news text in a neutral and factual tone.
> Analysis summary: Truthfulness={truth_label}, Sentiment={sentiment_label}, Bias={bias_label}.
> Original: {text}”

(prompt)를 입력으로 받아 실행된다.

이를 통해 단순한 문체 변환이 아니라, 뉴스의 진위・감정・편향 정보를 고려한 의미적 중립화(Semantic Neutralization)가 이루어지며, 최종 출력은 감정적 표현이나 이념적 왜곡이 완화된 형태인 ‘Neutralized Sentence’로 반환된다.

4. 결과 시각화 및 해석

각 분석 단계의 예측 결과는 레이더 차트 형태로 시각화하여 뉴스별 신뢰도(Truthfulness Confidence), 감정 강도(Sentiment Polarity), 이념적 편향(Bias Orientation)을 직관적으로 표현하였다. 각 분석 결과에는 예측 클래스와 함께 softmax 확률 기반의 confidence score 가 제공되어, 모델이 특정 클래스(예: fake, negative, left)에 얼마나 강하게 쏠려 있는지를 정량적으로 보여준다.



이러한 시각화 결과는 단순한 모델 출력에 그치지 않고, 중립화 모델의 프롬프트 구성에도 활용되어, 결과적으로 해석 가능한(Explainable) 재작성 과정을 구현하였다. 이를 통해 사용자는 모델의 판단 근거를 시각적으로 확인하명서, 뉴스 문장의 중립화 과정을 투명하게 이해할 수 있다.

5. 결론

본 연구는 진위 판별, 감정 분석, 편향 검출, 중립적 재작성을 결합한 통합 자연어처리 파이프라인을 제안하였다. 각 단계별 실험 결과, DistilBERT 와 DeBERTa-v3 모델이 비교적 안정적이고 균형 잡힌 성능을 보였으며, 이를 기반으로 T5 재작성 모델이 신뢰도와 감정, 편향 정보를 반영한 문장 중립화를 성공적으로 수행하였다.

다만, 본 연구의 모델 성능은 일부 감정 중립 문장이나 미묘한 편향 표현에서 제한적 정확도를 보였으며, 재작성 품질 역시 데이터셋 규모와 프롬프트 설계에 영향을 받았다. 이에 향후 연구에서는 프롬프트 최적화, 데이터셋 확장 및 다국어 뉴스 적용 등을 통해 모델 성능과 중립화 품질을 개선할 필요가 있다.

본 연구는 뉴스 콘텐츠의 객관적 평가 및 중립적 표현 자동화를 위한 기반을 제공하며, 향후 실시간 뉴스 스트림이나 다국어 환경에도 적용 가능성을 제시한다.

참고문헌

[1] Jae-Hoon Rhee, Mi-Sook Kim, "Fake News Detection Using Document Bias and Sentiment Analysis," Proceedings of the Korean HCI Conference, Gangwon, 2023.02.01.

[2] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H., "FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media," arXiv preprint arXiv:1809.01286, 2020.

[3] Wang, W. Y., "“Liar, liar pants on fire”: A new benchmark dataset for fake news detection," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 422–426, 2017.

[4] Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L., "TweeTEval: Evaluating Tweet Representations on Sentiment, Emotion, and Emoji Prediction," Proceedings of the 28th International Conference on Computational Linguistics (COLING), 2020.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL-HLT, pp. 4171–4186, 2019.

[6] Pryzant, R., Bhagavatula, C., Hovy, E., & Roth, D., "Automatically neutralizing biased sentences in text," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 7784–7798, 2020.