통합적 감정 · 편향 · 진위 분석을 통한 중립적 재작성

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

Hyewon Yang*, Wonkyo Choi*, and Taewan Kim*

***Data Science Major, Dongduk Women's University**

{20221631, 20221645, kimtwan21}@dongduk.ac.kr

ABSTRACT

본 연구는 영어 뉴스 헤드라인을 대상으로 진위 판별, 감정 분석, 편향 검출을 통합한 자연어처리 파이프라인을 제안한다. 각 단계별 분석 결과를 T5 재작성 모델에 입력하여 의미적 중립화를 수행하며, 최종 출력은 감정과 이념적 왜곡이 완화된 'Neutralized Sentence'로 제공된다. 또한, 분석 결과를 레이더 차트로 시각화하여 뉴스별 신뢰도와 감정·편향 강도를 직관적으로 확인할 수 있다.

INTRODUCION

최근 온라인 뉴스와 SNS에서는 헤드라인 중심으로 정보가 소비되며, 클릭 유도나 정치적 성향에 따른 과장·감정적 표현이 독자의 인지 편향과 정보 왜곡을 유발한다. 본 연구는 기존의 단일 분석 중심 연구들과 달리, 진위 · 감정 · 편향을 통합 수행하고, 이를 기반으로 중립적 재작성을 수행하는 자연어처리 파이프라인을 제안하여 뉴스 신뢰성과 객관성 향상을 목표로 한다.

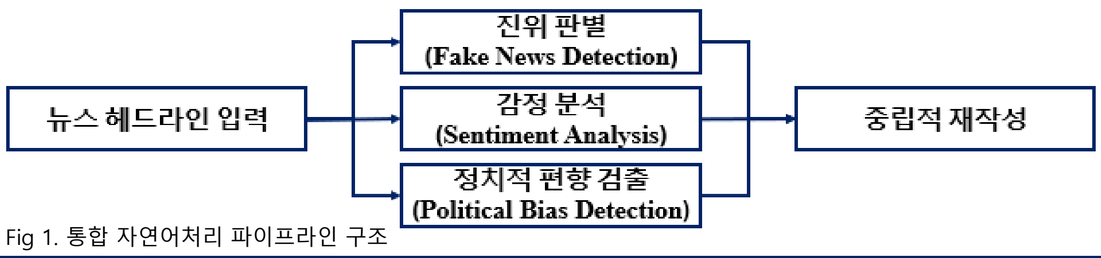

Fig 1. 통합 자연어처리 파이프라인 구조

METHOD

▶ 진위 판별 (Fake News Detection)

진위 판별 단계에서는 FakeNewsNet 데이터셋(Fake/Real, 학습 18,556개·테스트 4,640개)을 사용하였으며, DistilBERT와 BERT-base를 비교한 결과 두 모델 모두 **약 85~86% 정확도**를 보였으나, 학습 속도와 효율이 우수한 **DistilBERT**를 최종 모델로 선택하였다.

▶ 감정 분석 (Sentiment Analysis)

감정 분석 단계에서는 TweetEval Sentiment 데이터셋(학습 45,615·검증 2,000·테스트 12,284)을 사용하였으며, DistilBERT와 DeBERTa-v3-base를 비교한 결과, **DeBERTa-v3-base**가 **정확도 72%**로 균형 잡힌 성능을 보여 최종 모델로 선택되었다.

▶ 정치적 편향 검출 (Political Bias Detection)

정치적 편향 검출에서는 타이틀·헤딩·본문을 모두 활용한 **DistilBERT** 모델이 테스트 정확도 **0.63**으로 가장 안정적이어서 최종 모델로 선택되었다. 편향과 같이 고난이도 주제의 경우 정확도가 높지 않더라도, 재작성 모델은 편향 분석 결과의 절대값보다는 경향(label + confidence)만을 참고하므로, 불확실성이 있는 경우에도 출력이 자연스럽게 중립적 방향으로 수렴하는 특징을 보인다.

▶ 중립적 재작성 (Neutral Rewriting)

중립적 재작성 단계에서는 T5-base 모델을 활용하여 단계적 파인튜닝(stepwise fine-tuning)을 수행하였다. 데이터 규모를 점진적으로 1,000 → 5,000 → 10,000개로 확대하며 학습 및 검증 손실을 안정적으로 유지하였고, ROUGE-1과 ROUGE-L 점수 역시 의미적 정보 손실 없이 일정 수준을 유지하여 중립적 재작성 품질이 안정적임을 확인하였다.

Type	Model & Accuracy
진위 판별 (Fake News Detection)	DistilBERT 0.8593
감정 분석 (Sentiment Analysis)	DeBERTa-v3-base 0.72
정치적 편향 검출 (Political Bias Detection)	DistilBERT(Title+Headline+Text) 0.63
중립적 재작성 (Neutral Rewriting)	T5-base ROUGE-L 약 4% 상승

Table 1. 단계별 모델 구성 및 정확도 비교

▶ 통합 파이프라인 (Integrated Pipeline)

본 연구의 통합 파이프라인은 네 단계로 구성된다. 뉴스의 진위, 감정, 정치적 편향을 순차 분석한 뒤, 각 단계의 확률(confidence)을 T5 재작성 모델에 입력하여 중립적 재작성을 수행한다. 이를 통해 단순 문체 변환이 아닌 의미적 중립화가 이루어지며, 최종 출력은 감정과 편향이 완화된 'Neutralized Sentence'로 제공된다.

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

EXPERIMENTAL RESULTS

본 연구에서는 제안한 통합 자연어처리 파이프라인의 중립 재작성 효과를 간접적으로 검증하였다. 예를 들어, 원문 헤드라인

> The greedy politicians passed another useless law to affect hardworking citizens.

의 분석 결과, 진위는 97.1% 확률로 거짓(false), 감정은 99.7% 확률로 부정(negative), 편향은 87.4% 확률로 좌(left)로 치우친 것으로 나타났다. 이를 중립화 모델에 입력하여 재작성한 결과,

> The politicians passed another law to affect hardworking citizens.

와 같이 중립적인 문장으로 변환되었다.
재작성 후 분석 결과, 진위는 68.0% 확률로 사실(truthful)에 가까워졌고, 감정은 55.3% 확률로 중립(neutral), 편향은 53.6% 확률로 좌측 경향이 완화되어 중립에 보다 근접하였다.

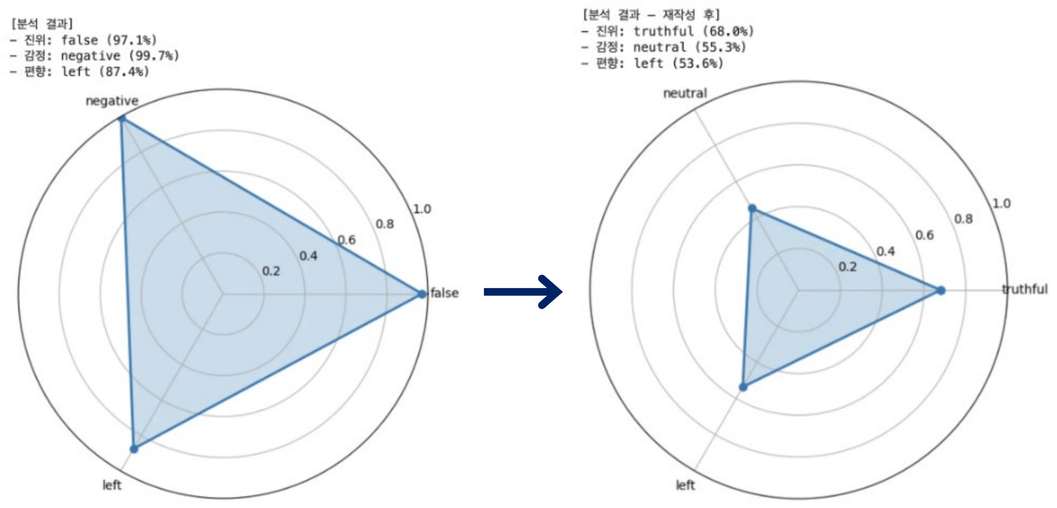

Fig 2. 중립화 전후 결과 비교 (레이더 차트)

CONCLUSION

본 연구는 뉴스 헤드라인의 **진위, 감정, 편향을 통합적으로 분석**하고 이를 기반으로 중립적 재작성을 수행함으로써, 감정적·이념적 왜곡을 완화하면서도 의미적 정보는 유지할 수 있음을 확인하였다. 레이더 차트와 확률값 시각화를 통해 중립화 전후의 변화가 직관적으로 검증되었으며, 이는 실제 뉴스 표현의 **신뢰성과 객관성 향상 가능성**을 보여준다. 제안한 접근은 세 가지 영역을 동시에 판단·조정함으로써 기존 연구와 차별화되며, 향후 **다국어 뉴스나 실시간 스트림으로의 확장 가능성**이 있다.

REFERENCES

[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