Journal of the Korean Institute of Industrial Engineers
Vol. 50, No. 1, pp. 36-46, February 2024.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2024.50.1.036

비지도 학습 기반 가짜뉴스 생성 프레임워크 개발 및 데이터셋 품질 평가 지표의 효과성 연구

김중훈 · 박새란 · 이지윤 · 김재희 · 강필성†

고려대학교 산업경영공학부

Development of an Unsupervised Learning-Based Fake News Generation Framework and Study the Effectiveness of Dataset Quality Evaluation Metrics

Joonghoon Kim · Saeran Park · Jiyoon Lee · Jaehee Kim · Pilsung Kang

School of Industrial & Management Engineering, Korea University

With the advent of the internet and social media, fake news has become easily generated and disseminated, causing severe societal issues. The development of robust fake news detection models to mitigate the damages incurred from fake news necessitates a substantial volume of high-quality fake news data. However, there exists a significant shortage of publicly available datasets that embody such quality. In this study, we propose an unsupervised learning-based framework for generating fake news, solely leveraging real news data. This proposed framework is structured into four key phases: similar news exploration, fake news title generation, labeling error filtering, and data quality verification. Moreover, we introduce a novel set of evaluation metrics to assess the quality of the generated fake news data in terms of reliability, diversity, and complexity. Through extensive experiments, we have validated the effectiveness of the proposed framework and evaluation metrics, paving the way for more nuanced approaches to the creation and analysis of fake news datasets.

Keywords: Natural Language Processing, Fake News Generation, Evaluation Metrics

1. 서 론

가짜뉴스의 확산은 현대 사회에서 심각한 문제로 대두되고 있다. 인터넷 및 소셜 미디어의 발전으로 인해 정보의 양과 전파 속도가 기하급수적으로 증가하였고, 이러한 환경은 가짜뉴스가 더욱 쉽게 생성 및 전파될 수 있는 토양을 제공하고 있다. 가짜뉴스는 대중의 인식과 의사 결정 과정에 중요한 영향을 미치며, 사회적 분열과 혼란을 유발할 수 있는 위험성을 내포하고 있다. 특히, 2016년 미국 대선에서 가짜뉴스가 유권자들의 행태에 영향을 미쳤다는 주장이 제기되어, 그 심각성과 문제점이 세계적으로 주목받게 되었다. 또한, 최근 COVID-19와 관련된 가짜뉴스로 인한 거짓 정보의 확산이 문제시되고 있어, 가짜뉴스 탐지의 중요성이 더욱 강조되고 있다(Jang *et al.*, 2021; Oh *et al.*, 2022).

가짜뉴스는 연구의 목적과 관점에 따라 다양한 방식으로 정의되고 있다. 일반적으로, 가짜뉴스는 '사실과 다른 거짓 정보를 담고 있는 뉴스' 또는 '뉴스 제목과 본문의 연관성이 없는 뉴스'라는 개념을 공통적으로 포함하고 있다(Hwang *et al.*,

이 논문은 2023년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2022R1A2C2005455)의 성과물임. 또한, 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00471, 모델링 & 최적화 기반 오류-free 정보인프라 자율제어 기술 개발).
† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3383, Fax: 02-929-5888,
E-mail: pilsung_kang@korea.ac.kr
2023년 10월 17일 접수; 2023년 11월 28일 게재 확정.

2017; Oh *et al.*, 2022; Jwa *et al.*, 2019; Yoon *et al.*, 2019). 본 연구에서는 '뉴스 본문과 제목 사이의 연관성이 없거나 뉴스 제목에 본문에서 확인할 수 없는 정보가 포함된 뉴스'를 가짜뉴스로 정의하였다.

최근, 머신러닝 및 딥러닝 기법을 활용한 가짜뉴스 탐지 모델에 관한 연구가 활발히 이루어지고 있지만 (Shim *et al.*, 2019; Yoon *et al.*, 2019; Jwa *et al.*, 2019), 탐지 모델 학습을 위한 가짜뉴스 데이터셋 구축에 관한 연구는 상대적으로 부족한 실정이다. 관련 연구에서는 인간이 작성한 가짜뉴스를 활용하는 방법(Huang *et al.*, 2022) 또는 수집된 뉴스를 직접 분류하여(Shim *et al.*, 2019) 가짜뉴스로 활용하는 경우가 많지만, 이러한 가짜뉴스 데이터는 양이 제한적이며 수집 비용이 상당하다는 문제점이 있다. 더욱이, 현재 공개된 한국어 가짜뉴스 데이터셋은 양적으로 매우 부족하며, 대부분 인간이 직접 생성하였다는 단점이 있다. 이러한 한계점을 극복하기 위해, 메타데이터를 활용하여 유사한 뉴스의 제목을 원래 뉴스의 제목으로 교체하는 방식으로 가짜뉴스를 생성하는 방법이 제안되었다(Jang *et al.*, 2021). 그러나, 이러한 방식은 생성된 가짜뉴스가 유사뉴스에 의존적이라는 문제와 모델이 쉽게 구별할 수 있는 형태의 가짜뉴스가 생성될 가능성이 있다는 문제가 있다. 또한, 가짜뉴스 데이터셋의 품질을 평가하는 지표에 관한 연구도 부족하여, 생성된 가짜뉴스 데이터의 효과성이나 탐지 모델 학습에 미치는 효과를 객관적으로 판단하기 어려운 상황이다.

본 연구에서는 이러한 문제점을 해결하기 위해 실제 뉴스 데이터를 활용하여 가짜뉴스를 자동으로 생성하는 프레임워크와 생성된 가짜뉴스의 품질을 다양한 관점에서 평가할 수 있는 지표를 제안한다. 해당 프레임워크는 유사뉴스 탐색, 가짜뉴스 제목 생성, 라벨링 오류 필터링, 데이터 품질 검증의 4단계로 구성된다. 유사뉴스 탐색 단계는 수집한 데이터셋 내에서 진짜뉴스와 가장 유사한 뉴스 데이터를 탐색하는 단계이다. 가짜뉴스 제목 생성 단계는 가짜뉴스 제목 생성을 위한 모델을 훈련하고, 원본뉴스와 유사뉴스 탐색 단계에서 찾은 유사뉴스의 정보를 혼합하여 가짜뉴스 제목을 생성하는 단계이다. 여기서, 원본뉴스는 가짜뉴스 생성 대상이 되는 진짜뉴스를 의미한다. 라벨링 오류 필터링 단계는 생성된 가짜뉴스 중에서 내용상 가짜뉴스가 아닌 데이터를 필터링하는 단계이다. 이 단계에서는 잘못 라벨링 된 데이터를 식별하여 제외한다. 데이터 품질 검증 단계는 생성된 가짜뉴스 데이터의 품질을 신뢰성, 다양성, 그리고 난이도의 측면에서 평가하는 단계이다. 데이터의 품질을 검증하기 위해 OLER, Coverage, Difficulty라는 세 가지 새로운 평가 지표를 제안하였으며, 이를 통해 생성된 가짜뉴스 데이터의 품질을 다양한 측면에서 객관적인 지표로 평가할 수 있다.

본 연구의 주요 기여점은 다음과 같다. (1) 진짜뉴스 데이터만을 활용하여 고품질의 가짜뉴스 데이터를 생성할 수 있는 프레임워크를 제안한다. 제안된 프레임워크는 언어와 데이터 종류와 관계없이 일반화 가능한 방법론이다. (2) 생성된 데이터셋의 객관적인 품질 평가 지표를 제안한다. 데이터셋의 신뢰성, 다양성, 난이도 등을 객관적인 지표로 평가하고, 가짜뉴스 탐지 모델 학습에 적합한 데이터인지를 판단할 수 있다. (3) 다양한 실험을 통해 제안된 프레임워크의 타당성과 효과성을 입증하였으며, 프레임워크 내 단계별 최적의 방법을 탐색하였다.

본 논문의 구성은 다음과 같다. 먼저, 제2장에서는 가짜뉴스에 대한 정의, 생성, 평가, 탐지에 관련된 선행 연구를 소개하며, 본 연구와의 차이점을 비교한다. 제3장에서는 본 연구에서 제안하는 방법론의 전체적인 프레임워크를 단계별로 자세하게 살펴본다. 제4장에서는 실험에 사용된 데이터 및 실험 환경을 설명하고, 정량적 및 정성적인 실험 결과를 제시하고 분석한다. 마지막으로 제5장에서는 본 연구의 결론을 도출하고 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 가짜뉴스 정의

가짜뉴스는 관점, 분야 또는 연구 범위에 따라 다양하게 정의되고 있다. Hwang *et al.*(2017)은 가짜뉴스를 '의도적으로 다른 사람을 속이기 위한 거짓 정보가 포함된 뉴스'로 정의하고 있다. Oh *et al.*(2022)은 '뉴스 제목에 거짓 정보가 포함된 뉴스'를 가짜뉴스로 정의하고 있으며, Jwa *et al.*(2019), Yoon *et al.*(2019)은 '뉴스 제목과 본문의 내용이 관련이 없는 뉴스'를 가짜뉴스로 정의하고 있다. 본 연구의 목적은 외부 정보없이 주어진 뉴스 제목과 본문만으로 가짜뉴스를 생성하는 것이기 때문에 가짜뉴스를 '뉴스 본문과 제목 사이의 연관성이 없거나 뉴스 제목에 본문에서 확인할 수 없는 정보가 포함된 뉴스'로 정의하였다.

2.2 가짜뉴스 생성

Huang *et al.*(2022), Shu *et al.*(2021), Nagoudi *et al.*(2020)은 가짜뉴스 본문을 생성하기 위해 원본뉴스에는 존재하지 않는 문장을 삽입하거나, 원본뉴스와 유사한 내용의 외부 지식을 활용하거나, 본문의 특정 단어들을 교체하는 방식을 사용하고 있다. Jang *et al.*(2021)은 메타 데이터를 활용하여 원본뉴스의 제목을 유사한 내용을 가진 다른 뉴스의 제목과 교체하는 방식으로 가짜뉴스 제목을 생성하고 있다. 이러한 연구들은 외부 지식의 활용을 위해 복잡한 모델링 과정이 요구되며, 본문 내용을 변형하기 위해 사람이 직접 주석(annotation) 작업이 필요하다는 한계점을 가지고 있다. 이러한 한계점을 극복하기 위해 외부 지식 활용 및 가짜뉴스 생성을 위한 주석 작업 수행 없이 본 연구는 주어진 유사한 뉴스 정보만을 활용하여 본문과 내용이 일치하지 않는 가짜뉴스 제목을 직접 생성하는 방법론을 제안하였다.

2.3 가짜뉴스 평가

Pillutla *et al.*(2021)은 인간에 의해 생성된 가짜뉴스와 모델에 의해 생성된 가짜뉴스 간의 유사도를 측정하는 지표를 제안하였다. Huang *et al.*(2022)은 이 방법론으로 생성된 데이터셋과 기존 데이터셋을 활용하여 각각의 탐지 모델을 훈련시키고, 탐지 성능을 비교함으로써 가짜뉴스 데이터셋의 효과성을 평가하였다. 그러나 이러한 연구들은 가짜뉴스를 평가하기 위해 인간이 직접 작성한 가짜뉴스가 필요하며, 유사도 평가를 통한 가짜뉴스의 난이도 측정 과정에서, 유사도가 높은 경우 해당 뉴스가 실제로는 진짜뉴스일 가능성을 고려하지 않는다는 한계점이 있다. 본 연구는 이러한 한계점을 극복하기 위해 가짜뉴스의 품질 및 난이도를 평가하는 새로운 지표를 개발하였으며, 다양한 관점에서 가짜뉴스를 평가하는 방안을 제안하였다.

2.4 가짜뉴스 탐지

Jang *et al.*(2021), Yoon *et al.*(2019), Jwa *et al.*(2019)은 뉴스 제목과 본문을 입력 데이터로 사용하여 가짜뉴스 여부를 판별하는 모델을 제안하였으며, GRU(Chung *et al.*, 2014), BERT (Devlin *et al.*, 2018)와 같은 딥러닝 기반의 방법론들을 활용하여 탐지 모델에 관한 연구를 수행하였다. 본 연구는 가짜뉴스 생성 방법론을 제안하므로 탐지 모델을 활용한 평가가 필수적이다. 따라서, 본 연구는 가짜뉴스 데이터셋의 효과성을 평가하기 위해 BERT를 탐지 모델로 활용하여 실험을 수행하였다.

3. 제안 방법론

본 연구에서는 <Figure 1>과같이 유사뉴스 탐색, 가짜뉴스 제목 생성, 라벨링 오류 필터링 그리고 데이터 품질 검증의 네 단계로 구성된 프레임워크를 제안한다.

3.1 유사뉴스 탐색

가짜뉴스 제목 생성 모델의 입력을 구성하는 단계에서 원본뉴스와 가장 유사한 뉴스를 탐색한다. 이 과정은 원본뉴스에 존재하지 않는 유사뉴스의 정보를 모델에 제공함으로써, 자연스러우면서도 쉽게 구분하기 힘든 가짜뉴스 제목을 생성하기 위함이다. 이를 위해, 원본뉴스와 같은 카테고리에 속하는 다른 모든 뉴스와의 유사도를 산출하였으며, 이 중에서 가장 높은 유사도를 보이는 뉴스를 선택하여 원본뉴스 - 유사뉴스 쌍을 형성하였다.

유사도 계산에는 TF-IDF와 같은 통계 기반 방법론과 ColBERT(Khattab and Zaharia *et al.*, 2022)와 같은 딥러닝 기반 방법론을 활용하였다. 일반적으로, 통계 기반 방법론은 단어 출현 빈도와 문서 빈도를 기준으로 문서의 중요도를 측정하지만, 단어의 순서나 문맥을 충분히 반영하지 못한다는 한계가 있다. 반면, 딥러닝 기반 방법론은 사전 학습된 언어 모델의 풍부한 지식을 활용하여 단어의 순서와 문맥을 고려한 중요도 계산이 가능하지만, 모델 훈련이 필요하며 연산 시간이 상대적으로 더 길다는 단점이 존재한다.

TF-IDF는 텍스트 검색이나 텍스트 마이닝 분야에서 광범위하게 활용되는 방법론이다. 이 방법론은 다수의 문서로 구성된 문서 집합에서 각 문서 내의 단어의 상대적 중요도를 측정하는 통계적 지표로 활용된다. 식 (1)에서 $TF(t,d)$는 문서 d에서 단어 t의 출현 빈도를 의미하며, $IDF(t,D)$는 전체 문서 집합 D에서 단어 t가 등장하는 문서 수의 역수를 의미한다. TF-IDF 값은 문서 내에서 특정 단어의 빈도와 역문서 빈도의 곱으로 계산된다.

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

ColBERT(Khattab and Zaharia *et al.*, 2022)는 텍스트 검색 분야에서 제안된 모델이며, BERT(Devlin *et al.*, 2018)를 기반으로 구성되어 있다. 이 방법론은 검색 쿼리와 검색 대상 문서를 각각 인코딩한 후, 토큰 단위의 임베딩 유사도를 계산하여 문서 간의 의미적 유사성을 측정한다. 이러한 접근 방식은 사전 학습된 BERT의 풍부한 언어 지식을 활용하여 단어의 복잡한 의미를 충분히 반영하여 문서 간의 세밀한 의미적 유사성을 측정할 수 있다는 장점이 있다.

TF-IDF로는 원본뉴스의 제목, 본문, 제목 - 본문과 원본뉴스의 유사도 비교 대상이 되는 비교뉴스의 제목, 본문, 제목 - 본문 간의 유사도를 각각 계산하였다. ColBERT로는 원본뉴스의 제목과 비교뉴스의 제목 - 본문 간의 유사도를 계산하였다. 여기서 제목 - 본문은 제목과 본문을 결합하여 유사도를 계산한 경우를 의미한다.

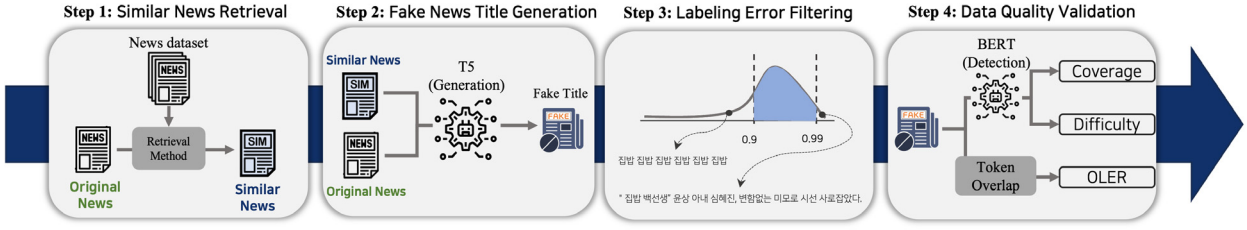

Figure 1. Unsupervised Learning-based Fake News Generation Framework

Table 1. Similar News Search Performance of TF-IDF and ColBERT

Method	Source News	Target News	Top1-Accuracy
TF-IDF	Title	Title	0.9928
	Title	Content	0.4746
	Title	Title-Content	0.9668
	Content	Title-Content	0.6943
	Content	Title-Content	0.9961
	Title-Content	Title-Content	**0.9953**
ColBERT	Title	Content	0.9250

Top1-Accuracy는 텍스트 검색 분야에서 일반적으로 사용되는 평가 지표로, 주어진 쿼리에 대해 가장 관련성이 높은 문서를 정확하게 식별하는 능력을 측정한다. <Table 1>에서 Top1-Accuracy는 전체 뉴스 데이터셋 내에서 원본 뉴스와 가장 유사도가 높은 뉴스가 실제로 원본 뉴스인지를 판단하는데 사용된다. 높은 Top1-Accuracy 값은 텍스트 유사도 계산을 통해 전체 뉴스 중에서 원본뉴스를 효과적으로 탐색하고 식별할 수 있음을 나타낸다.

<Table 1>의 실험 결과를 확인해 보면, TF-IDF를 사용하여 측정한 원본뉴스의 제목 - 본문과 비교뉴스의 제목 - 본문 간의 유사도 측정 성능이 가장 뛰어나다는 것을 확인할 수 있다. 또한, TF-IDF가 ColBERT보다 더 우수한 성능을 보여주었다. 이러한 결과는 ColBERT가 모델 입력 길이의 제한으로 인해 뉴스 정보의 일부만을 사용해야 했다는 점과 GPU 리소스 제한으로 인해 모델 학습 시 배치 크기가 제한적이었던 점 때문으로 추정된다. 이러한 이유로, 원본뉴스와 비교뉴스에서 제목과 본문 모두를 활용하는 TF-IDF 방법을 최종적인 유사뉴스 탐색 방법론으로 선정하였다.

3.2 가짜뉴스 제목 생성

본 연구에서는 진짜뉴스 데이터만을 활용하였다. 이를 위해 모델을 <Figure 2>에 제시된 것과 같이 진짜뉴스 본문을 입력으로 받아, 해당 본문에 적합한 진짜뉴스 제목을 생성할 수 있도록 미세조정하였다. 이러한 방식으로 학습된 모델에 원본뉴스-유사뉴스 쌍을 입력으로 제공하면, 두 본문의 정보를 적절하게 혼합하여 새로운 뉴스 제목을 생성한다. 이렇게 생성된 제목은 <Table 2>에서 보여지는 바와 같이 가짜뉴스 제목으로 분류할 수 있다.

Figure 2. Training Process of the News Title Generation Model and Fake News Generation Process

Table 2. An Example of a Fake News Title Generated by Framework

	Original News	Similar News
Title	앞머리 내린 이제훈, 풋풋한 매력 선보여	'시그널' 김혜수, "이제훈, 질리지 않는 얼굴… 목소리도 좋아"
Content	이제훈이 새로운 드라마를 앞두고 앞머리를 내린 헤어스타일을 선보여 풋풋한 매력을 뽐내고 있다. 이제훈이 내년 초 방송 예정인tvN 드라마 '내일 그대와' 촬영에 들어가면서 헤어스타일에 변화를 준 것으로 알려졌다. …	김혜수는 '시그널' 스페셜 토크에서 상대배우 이제훈과의 연기 호흡과 드라마 출연 이유를 밝혔다. 상대 배우 이제훈에 대해서는 "질리지 않는 얼굴이다"며 "이제훈과 미팅을 하는데 목소리가 좋더라."라고 감탄했다. …
Fake News Title	앞머리 내린 이제훈, 김혜수와 연기 호흡 맞춰	

(1) 뉴스 제목 생성 모델 학습

본 연구에서는 뉴스 제목 생성 모델의 학습을 위해 T5(Raffel *et al.*, 2020)를 사전 학습 모델로 활용하였다. T5 모델은 Transformer구조(Vaswani *et al.*, 2017)를 기반으로 하며, 입력과 출력을 모두 자연어 형식으로 일관되게 처리할 수 있는 특징이 있다. 이 모델은 주로 자연어 생성 작업에 특화되어 있으며, 질의응답, 텍스트 분류, 요약 등의 다양한 과업에서 뛰어난 성능을 보인다(Raffel *et al.*, 2020). 본 연구에서는 T5 모델 학습 과정에서 진짜뉴스의 본문을 입력 데이터로 사용하였으며, 이를 통해 모델이 본문의 내용을 파악하고 다양한 정보를 고려하여 적절한 뉴스 제목을 생성할 수 있도록 훈련되었다.

(2) 모델 입력 방식

제목 생성 모델은 진짜뉴스의 본문을 입력 데이터로 받아 진짜뉴스 제목을 생성하도록 학습되었다. 이에 따라, 가짜 뉴스 제목을 구성하는 데에 있어 원본뉴스와 유사뉴스의 정보를 효율적으로 혼합하여 모델 입력으로 적용하는 단계가 중요한 핵심 요소로 간주된다. 원본뉴스와 유사뉴스의 내용이 적절히 조합되어야만 원본뉴스에 존재하지 않는 정보를 포함한 가짜 뉴스 제목을 만들어낼 수 있기 때문이다. 본 연구에서는 이러한 목적을 달성하기 위해 두 가지의 입력 순서와 세 가지의 입력 형태를 고려한 실험을 <Figure 3>과 같이 진행하였다.

뉴스 기사는 대체로 두괄식 구조를 보이며, 따라서 뉴스 본문의 초반 부분과 제목 사이에 높은 상관관계가 존재한다. 이러한 특성으로 인해 가짜뉴스 제목 생성 모델은 입력 데이터의 초반 부분에 더욱 중점을 두어 제목을 생성하는 경향이 있다. 이를 고려하여, 문장 순서 및 뉴스 내용의 순서가 가짜뉴스 생성에 미치는 영향을 확인하기 위해 모델에 입력되는 원본뉴스와 유사뉴스의 순서를 조정하는 실험을 수행하였다. 이 실험은 'Forward'와 'Backward' 두 가지 방식에 대해 진행되었다.

'Forward'는 원본뉴스 본문을 유사뉴스 본문 앞에 배치하는 방식이며, 'Backward'는 반대로 원본뉴스 본문을 유사뉴스 본문보다 뒤에 배치하는 방식이다.

또한 본 연구에서는 원본뉴스와 유사뉴스의 내용을 효과적으로 혼합하는 방법을 탐색하기 위해 'Chunking', 'Rotation', 'Summarization' 세 가지 입력 형태에 대한 실험을 수행하였다. 생성 모델이 입력 길이에 제한을 가지고 있으므로, 원본뉴스와 유사뉴스는 각각 모델의 최대 입력 길이의 절반까지만 활용되었다.

'Chunking'은 원본뉴스와 유사뉴스의 본문을 각각 처음부터 모델의 최대 입력 길이의 절반에 해당하는 부분까지 잘라낸 후, 이를 조합하여 모델의 입력으로 사용하는 방법이다. 다시 말해, 각 뉴스 본문은 앞부분에서부터 일정 길이만큼만 선택되며, 모델의 최대 입력 길이를 초과하는 부분은 배제된다. 'Rotation'은 원본뉴스와 유사뉴스의 본문에서 한 문장씩 교대로 선택하여 모델의 입력으로 활용하는 방법이다. 'Summarization'은 원본뉴스와 유사뉴스 본문에 대한 요약문을 각각 생성하여 이를 모델의 입력으로 활용하는 방법이다. 이 과정에서 한국어 요약 과업에 특화되어 미세조정된 모델을 활용하였다. 또한, 'Chunking' 및 'Rotation'과 마찬가지로 생성 모델의 입력 길이 제한으로 인해, 요약문의 최대 길이는 뉴스 제목 생성 모델의 최대 입력 길이의 절반 이하로 설정하여 생성하였다.

3.3 라벨링 오류 필터링

본 연구에서는 원본뉴스 - 유사뉴스 쌍을 기반으로 생성 모델을 통해 생성된 뉴스 제목은 가짜뉴스 제목으로 분류한다. 그러나, 생성된 뉴스 제목이 진짜뉴스 제목으로 취급될 수 있는 경우는 라벨링 오류가 발생한 것으로 인식될 수 있다. 이러한 오류는 유사뉴스와 원본뉴스 본문의 내용이 같거나 상당히 유사한 경우에 발생할 수 있다. 라벨링 오류는 데이터셋의 품질을 저하시키며, 가짜뉴스 탐지 모델 학습에 부정적인 영향을 미칠 수 있으므로 제거되어야 한다. 이러한 라벨링 오류를 감소시키기 위해 BERTScore(Zhang *et al.*, 2019)를 활용하여 생성된 뉴스 제목의 라벨링 오류를 식별하고 평가하였다.

BERTScore(Zhang *et al.*, 2019)는 텍스트 생성 품질을 평가하는 평가 지표로서, 후보 문장(candidate sentence)과 기준 문장

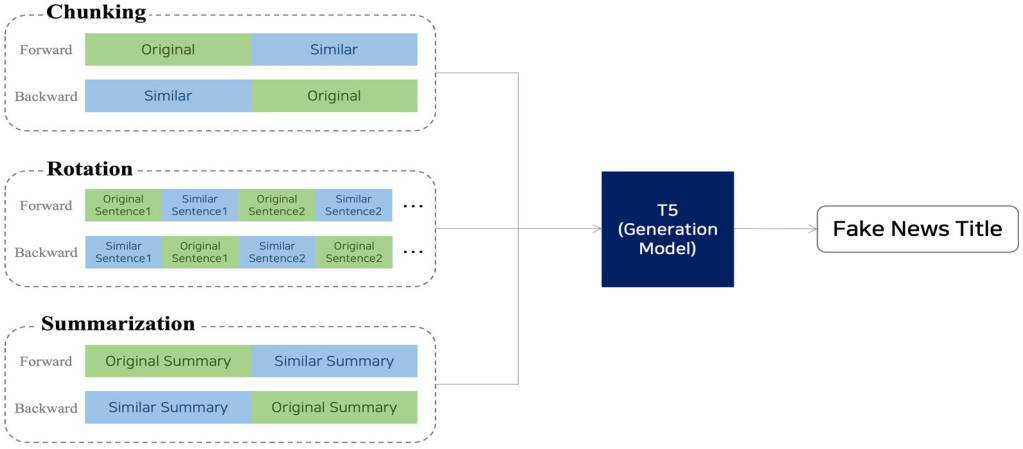

Figure 3. Input Construction Method for the Fake News Title Generation Step

(reference sentence)간의 유사도를 정량적으로 측정한다. BERTScore는 기존의 BLEU(Papineni *et al.*, 2002) 및 ROUGE(Lin *et al.*, 2004) 지표와는 달리, 사전학습 모델인 BERT를 사용하여 문장 간의 유사도를 산출하고, TF-IDF를 통해 가중치를 조정한다. BERTScore는 단순히 문장 내 단어들의 표면적인 비교에 그치지 않고, 단어의 의미와 문장 구조의 다양성을 반영하여 유사도를 측정할 수 있는 특징이 있다. BERTScore는 0에서 1 사이의 값으로 계산되며, 후보 문장과 기준 문장 사이의 유사도가 높을수록 1에 가까운 값을 산출하고, 유사도가 낮을수록 0에 가까운 값을 산출한다.

BERTScore가 1에 근접할수록 가짜뉴스 제목과 진짜뉴스 제목 간의 유사도가 높다는 것을 나타내며, BERTScore가 정확히 1인 경우에는 가짜뉴스 제목과 진짜뉴스 제목이 동일함을 의미한다. 반대로, BERTScore가 0에 가까울수록 가짜뉴스 제목과 진짜뉴스 제목 간의 유사도가 낮으며, 이는 두 제목 사이에 명백한 차이가 있다는 것을 의미한다.

라벨링 오류 필터링을 위해, 원본뉴스 제목과 생성된 가짜뉴스 제목 사이의 BERTScore를 계산하였다. 원본뉴스 제목과 과도하게 유사한 가짜뉴스 제목을 제거하기 위해 BERTScore가 0.99 이상인 데이터를 제외하였다. 또한 원본뉴스 제목과 유사도가 매우 낮거나 문법적으로 올바르지 않은 형태로 생성된 가짜뉴스 제목을 필터링하기 위해, BERTScore가 0.9 이하인 데이터를 제외하였다. 예를 들어, BERTScore가 0.9보다 작은 경우에는 <Table 2>의 예시에서 원본뉴스 제목이 '시그널 김혜수, "이제훈, 질리지 않는 얼굴… 목소리도 좋아"'인 경우에 가짜뉴스 제목이 '앞머리 내린 '이제훈', 풋풋한 매력 선보여'와 같이 유사뉴스 제목 그대로 생성되는 경우 혹은 '이제훈이제훈이제훈'과 같이 비문이 생성되는 경우를 확인하였다. 이러한 방법을 통해 라벨링 오류를 큰 폭으로 줄일 수 있을 뿐만 아니라, 과도한 단어의 반복이나 문법적 오류로 인해 자연스럽지 않은 가짜뉴스 데이터를 걸러내어 데이터의 품질을 향상할 수 있었다.

3.4 데이터 품질 검증

제안된 프레임워크의 목적은 고품질의 가짜뉴스 데이터셋을 생성하는 것이므로, 생성된 데이터의 품질 검증은 필수적인 절차로 간주한다. 가짜뉴스 데이터셋의 신뢰성, 다양성 및 난이도를 측정하기 위해, OLER, Coverage, Difficulty라는 세 가지 평가 지표를 도입하였다. OLER은 데이터셋의 신뢰성을, Coverage는 데이터셋의 다양성을, 그리고 Difficulty는 데이터셋의 난이도를 평가할 수 있는 지표로써 활용된다. 이러한 평가 지표들을 통해 생성한 가짜뉴스 데이터셋의 품질을 다각도로 평가하고 비교할 수 있으며, 이를 통해 최적의 고품질 가짜뉴스 데이터셋 구성이 가능하다.

(1) Overall Labeling Error Rate(OLER)

Overall Labeling Error Rate(OLER)는 데이터셋 내 라벨링 오류의 정도를 측정하기 위한 평가 지표이다. 본 연구에서, 가짜뉴스 제목은 원본뉴스 본문에 포함되지 않거나 본문과 일치하지 않는 정보를 포함하는 뉴스 제목으로 정의되었다. 따라서, 생성된 뉴스 제목의 모든 단어가 원본뉴스 본문에서 발견될 경우, 이를 가짜뉴스로 분류하는 것은 어려운 경우가 많다고 판단하였다.

이를 측정하기 위해 원본뉴스 본문과 가짜뉴스 제목 사이의 형태소 단위 단어 중복 여부를 계산하였다. OLER은 원본뉴스 본문의 단어와 가짜뉴스 제목 내 단어 사이의 중복 여부를 통해 계산된다. 단어 중복 여부는 $word\ overlap$을 통해 측정되며, 가짜뉴스 제목의 단어들이 원본뉴스 본문의 단어들에 모두 포함되어 있으면 1로, 그렇지 않으면 0으로 계산된다(식 (4)). 이를 기반으로 OLER은 전체 데이터셋의 수 N과 $word\ overlap$의 발생 비율로 계산된다(식 (5)).

$$F_i = \{word | word \in fake\ news\ title_i\} \tag{2}$$

$$R_i = \{word | word \in real\ news\ content_i\} \tag{3}$$

$$word\ overap_i = \begin{cases} 1 & \text{if } F_i \subset R_i, \\ 0 & else \end{cases} \tag{4}$$

$$OLER = \frac{\sum_i word\ overlap_i}{N} \tag{5}$$

그러나, OLER을 라벨링 오류 필터링에 활용하지 않았다. 이는 가짜뉴스 제목의 단어들이 모두 원본뉴스 본문에 포함되어 있음에도 불구하고, 내용상의 불일치가 발생할 때 정확한 평가가 어렵기 때문이다. 예시로, 원본뉴스 제목이 '내일 비가 온다'일 때, 가짜뉴스 제목이 '내일 비가 안 온다'인 경우를 들 수 있다. 본문에 '안'이라는 단어가 등장했다면 가짜뉴스 제목의 단어들이 모두 원본뉴스 본문에 존재하지만, 내용 간의 불일치로 인해 가짜뉴스로 분류될 수 있다. 이러한 이유로, OLER은 라벨링 오류의 전반적인 정도를 파악하는 평가 지표로 활용될 수 있다.

(2) Coverage

Coverage는 모델이 학습 데이터에 대해서만이 아니라 학습되지 않은 데이터에 대해서도 효과적으로 가짜뉴스를 식별할 수 있는 능력이 있는지를 평가하는 지표로 사용된다. 본 연구에서는 이 지표를 통해 생성 방법론이 교체 방법론보다 더 우수한 데이터셋을 생성하는지를 평가하고자 하였다. 여기서 '교체 방법론'은 유사한 뉴스 제목을 원본뉴스 제목과 교체하여 가짜뉴스 제목을 생성하는 방법을 지칭하며, '생성 방법론'은 제안된 프레임워크를 통해 가짜뉴스 제목을 직접 생성하는 방법을 지칭한다. 교체 방법론은 원본뉴스 제목과 유사한 뉴스 제목을 단순히 교체하는 방법이므로, 이렇게 구성된 가짜뉴스 제목의 품질이 생성된 제목에 비해 낮다고 판단될 수 있다. 따라서, Coverage를 통해 생성 방법론을 사용하여 생성된 가짜뉴스 데이터로 학습된 모델이 다양한 유형의 가짜뉴스를 정확하게 식별할 수 있는지를 평가한다. 이는 결국 생성된 가짜뉴스 데이터가 모델에 더 높은 강건성을 부여할 수 있는지를 평가하는 것과 같다.

Coverage는 교체 방법론의 데이터로 학습된 모델과 생성 방법론의 데이터로 학습된 모델 간의 성능을 비교하여 산출된다(식 (6)).

$$Coverage = \frac{ACC(D_G, D_S)}{ACC(D_S, D_G)} \quad (6)$$

여기서 D_G는 생성 방법론으로 만들어진 가짜뉴스 데이터셋을, D_S는 교체 방법론으로 만들어진 가짜뉴스 데이터셋을 나타낸다. $ACC(D_i, D_j)$는 D_i 데이터셋으로 학습된 모델의 D_i 데이터셋에 대한 성능을 나타낸다. 만약 Coverage가 1을 초과하면, 이는 생성 방법론이 교체 방법론보다 다양한 유형의 가짜뉴스를 더 효과적으로 탐지하는 데 도움이 되는 데이터셋을 만들 수 있다는 것을 의미한다.

(3) Difficulty

Difficulty는 생성된 가짜뉴스의 난이도를 측정하기 위한 핵심적인 지표이다. 이 지표는 가짜뉴스 데이터셋의 품질 평가에서 중추적인 역할을 수행하며, 기존의 탐지 모델은 난이도가 높은 가짜뉴스를 정확하게 식별하기 어렵다는 전제하에, 생성된 가짜뉴스 데이터셋에 대한 탐지 모델의 성능을 기반으로 산정된다(식 (7)).

$$Difficulty = ACC(D_G, D_G) \quad (7)$$

탐지 모델로는 BERT(Devlin *et al.*, 2018)를 사용하였다. Difficulty 값이 낮다는 것은 해당 가짜뉴스 데이터셋의 난이도가 더 높다는 것을 의미한다.

4. 실험 및 결과

4.1 실험 설계

(1) 데이터셋

본 연구의 실험에는 AIHub에서 제공하는 '낚시성 기사 탐지 데이터셋'을 활용하였다. 이 데이터셋은 총 364,333개의 데이터를 포함하고 있으며, 본 연구에서는 이 중 원천 데이터인 314,202건의 진짜뉴스 데이터만을 선별하여 사용하였다. 뉴스 제목 생성 모델의 학습, 검증 및 평가를 위해 전체 데이터셋을 291,466개의 학습 데이터, 36,434개의 검증 데이터 및 36,433개의 평가 데이터로 분할하였다. 또한, 전체 데이터셋 중 132,000개를 카테고리 분포에 따라 층화추출하여 가짜뉴스 데이터 생성에 활용하였다. 가짜뉴스 데이터셋 생성 후, 각각의 방법론에 대한 라벨링 오류 필터링을 수행하였으며, 그 결과는 <Table 3>에서 확인할 수 있다. 필터링 과정을 거친 가짜뉴스 데이터 중 66,000개의 데이터를 학습, 검증 및 평가 단계에 사용하기 위해 각각 40,000개, 13,000개, 13,000개로 층화추출하였다. 마찬가지로, 진짜뉴스 데이터도 학습, 검증 및 평가 단계에서 각각 40,000개, 13,000개, 13,000개의 데이터를 사용하였다.

Table 3. The Number of Data and Ratio after Labeling Error Filtering

Method		Size of filtered dataset	Filtering Rate
Forward	Chunking	95,630	27.6%
	Rotation	101,370	23.2%
	Summarization	130,804	0.8%
Backward	Chunking	130,471	1.2%
	Rotation	116,149	12.0%
	Summarization	120,238	8.9%
TF-IDF		127,490	3.4%

(2) 실험 환경

본 연구에서 가짜뉴스 생성 모델로 활용된 모델은 huggingface 플랫폼에 공개된 'KETI-AIR/ke-t5-base-newslike'로 훈련 과정에서는 배치 크기를 4, epoch를 7로 설정하였으며, 최대 입력 길이와 최대 생성 길이는 각각 512와 32로 지정하였다. 이 과정에는 NVIDIA RTX 2080Ti 그래픽 카드 2장이 사용되었다. 또한, 요약문 생성 모델로는 'lcw99/t5-base-korean-text-summary' 모델이 사용되었으며, 이 모델은 AIHUB의 "요약문 및 레포트 생성 데이터:를 바탕으로 미세조정되었다. 이 모델은 'Summarization' 입력 구성 방법론에서 활용되었고, 별도의 추가 학습 없이 활용되었다. 본 연구에서는 원본뉴스와 유사뉴스의 제목 및 본문을 입력 데이터로 활용하여 가짜뉴스 제목을 생성하였다. 원본뉴스와 유사뉴스의 본문만을 입력 데이터로 사용하여 가짜뉴스 제목을 생성한 경우의 성능을 <Table 7>에서 분석하였다. 라벨링 오류 필터링 단계에서는 BERTScore 계산을 위해 'klue/roberta-large' 모델을 사용하였고, 마지막 레이어의 출력값을 Representation으로 활용하였다. 데이터 품질 검증 지표인 Coverage와 Difficulty를 측정하기 위한 탐지 모델로 'klue/bert-base'를 채택하였다. 이 모델의 학습 시에는 배치 크기를 8, 학습 단계를 5,000으로 설정하였으며, Optimizer로는 Adam이 사용되었다. 이 과정에서는 NVIDIA RTX 2080Ti 그래픽 카드 1장이 활용되었다.

4.2 실험 결과

(1) 주요 실험 결과

본 연구에서 제안한 프레임워크의 방법론별 성능 결과는 <Table 4>에 상세히 제시되어 있다.

<Table 4>에서의 OLER 지표의 결과를 통해, 입력 형태와 관계없이 'Forward' 방법론은 Backward '방법론'에 비해 상대적으로 높은 라벨링 오류율을 보이는 것을 확인할 수 있다. 이러한 결과는 'Forward' 방법론이 원본뉴스의 실제 제목을 재현하는 경우가 많음을 시사한다. 추가로, 'Forward' 방법론의 OLER값은 뉴스 제목 생성 모델이 뉴스 본문의 초반 부분에 집중하여 제목

을 생성하는 경향이 있다는 것을 나타낸다. 이는 일반적으로 뉴스 본문이 두괄식으로 작성되며, 제목이 본문의 첫 문단과 강한 연관성을 지니기 때문이다. 그러므로, 'Forward' 방법론의 원본뉴스의 본문이 유사뉴스보다 먼저 입력되는 특성으로 인해, 유사뉴스의 정보 반영이 상대적으로 적으며 원본뉴스의 정보가 과도하게 반영되는 경향이 있다는 것으로 해석할 수 있다. 또한, 원본뉴스와 유사뉴스를 교대로 삽입하는 'Rotation' 방법론의 OLER지표를 분석해보면, 'Forward' 방법론 및 'Backward' 방법론과의 값 차이가 뚜렷한 것을 알 수 있다. 이는 생성 모델이 본문의 초반 부분, 특히 첫 문장을 중점으로 제목을 생성하는 경향성이 있다는 것을 의미한다. 그리고, 'Backward' 방법론 중에서는 'Chunking', 'Rotation', 'Summarization' 순으로 OLER이 값이 낮아 라벨링 오류율이 상대적으로 낮다는 것을 확인할 수 있다.

Table 4. Main Results

Method		OLER	Coverage	Difficulty
Forward	Chunking	28.1%	0.99	0.6305
	Rotation	22.8%	1.01	0.6268
	Summarization	38.2%	0.87	0.8193
Backward	Chunking	5.7%	1.03	0.8519
	Rotation	8.1%	1.08	0.8148
	Summarization	15.6%	0.93	0.9058
TF-IDF		4.0%	-	0.8710

<Table 4>의Coverage 지표를 살펴보면, 'Backward' 방법론이 'Forward' 방법론에 비해 상대적으로 높은 값을 보이는 것을 확인할 수 있다. 이러한 결과는 'Forward' 방법론으로 생성된 가짜뉴스 데이터에 라벨링 오류가 상대적으로 많이 포함되어 있으며, 데이터 다양성이 부족함을 나타낸다. 이러한 다양성 부족은 모델이 교체 방법론으로 생성된 가짜뉴스 데이터를 정확하게 탐지하는 데 어려움을 초래한다. 즉, 'Forward' 방법론을 통해 생성된 데이터셋은 원본뉴스의 정보를 과도하게 반영하여 라벨링 오류가 빈번하게 발생하고, 이에 따라 유사뉴스의 정보가 높은 비중으로 반영된 교체 방법론의 가짜뉴스 데이터에 대한 탐지 능력을 학습시킬 수 없다는 것을 의미한다. 또한, 'Backward' 방법론 중, 'Chunking'과 'Rotation' 방법론에서는 Coverage 값이 1을 초과함을 알 수 있다. 이 결과는 'Chunking'과 'Rotation' 방법론이 교체 및 생성 방법론을 통해 생성된 가짜뉴스 데이터를 정확하게 탐지할 수 있는 능력을 탐지 모델에 성공적으로 학습시킬 수 있다는 것을 의미한다. 더욱이, 'Rotation' 방법론이 'Chunking' 방법론에 비해 가짜뉴스 데이터의 다양성이 약간 더 높게 나타나는 것을 확인할 수 있는데, 이는 'Rotation' 방법론이 가짜뉴스 데이터의 다양성을 더욱 향상할 수 있는 효과적인 방법론임을 시사한다.

'Summarization' 방법론은 뉴스 본문을 요약한 후 해당 요약문을 기반으로 뉴스 제목을 생성하는 방법론으로, <Table 4>에서 확인할 수 있듯이 교체 방법론으로 학습된 탐지 모델과의 Difficulty 차이가 미미하거나, 'Backward'의 경우에는 더 높은 값을 보이는 것을 알 수 있다. 이는 'Summarization' 방법론이 TF-IDF보다 모델이 탐지하기 쉬운 가짜뉴스 데이터셋을 생성하는 경향이 있음을 의미한다. 요약문은 뉴스 본문의 전반적인 내용을 종합하는 특성이 있으며, 뉴스 본문의 초반부와 연관성이 높은 제목의 특성을 고려하면, 'Summarization' 방법론들의 Difficulty가 높게 나타나는 원인을 이해할 수 있다.

Difficulty 지표만을 단독으로 고려할 경우, 'Forward' 방법론들은 'Backward' 방법론들에 비해 낮은 값을 보이므로 더 복잡하고 어려운 가짜뉴스 제목을 생성하는 것으로 결론지을 수 있다. 그러나 OLER과 Coverage 지표를 함께 고려하면, 이러한 결과는 'Forward' 방법론들에 포함된 라벨링 오류의 빈도가 높기 때문이라는 결론을 도출할 수 있다. 구체적으로는, 라벨링 오류는 탐지 모델이 진짜뉴스와 가짜뉴스를 구분하는 기준을 혼란스럽게 만들며, 가짜뉴스 탐지 모델 학습에 부정적인 영향을 주게 된다. 이러한 결과는 라벨링 오류를 정확하게 필터링하고 평가하는 작업의 중요성을 강조하며, 이는 정성평가 결과에서도 확인할 수 있다.

또한 실험 성능을 살펴보면, 'Rotation'이 'Chunking'과 'Summarization' 방법론에 비해 더 복잡한 가짜뉴스 제목을 효과적으로 생성한다는 것을 확인할 수 있다. 이는 'Forward'와 'Backward' 방법론 모두 'Rotation'의 Difficulty 값이 가장 낮게 나타나며, 교체 방법론인 TF-IDF와 비교할 때 최소 약 5%p에서 최대 약 23%p까지의 탐지 성능 차이를 보인다는 점에서도 확인할 수 있다. 이와 더불어, 'Chunking'은 'Rotation'에 비해 원본뉴스 또는 유사뉴스의 정보가 초반부에 집중되는 경향이 있어, 'Forward' 방법론에서는 라벨링 오류가 가장 빈번하게 발생하며, 'Backward' 방법론에서는 TF-IDF의 성능과 큰 차이를 보이지 않는다는 것을 확인할 수 있다.

정량평가 결과를 종합적으로 분석하면, 'Summarization' 방법론이 뉴스 본문의 내용을 요약하여 제목을 생성하기 때문에 어려운 가짜뉴스 제목 생성에는 적합하지 않다는 것을 확인할 수 있다. 반면, 단순 제목 교체 방법론인 TF-IDF보다는 'Chunking'과 'Rotation'과 같은 생성 방법론이 높은 난이도의 가짜뉴스 제목 생성에 더 효과적이라는 것을 알 수 있다. 또한, 원본뉴스의 정보가 초반부에 주로 반영되는 'Forward' 방법론들이 'Backward' 방법론들에 비해 더 많은 라벨링 오류를 유발하며, 'Rotation' 방법론이 'Chunking'에 비해 어려운 가짜뉴스 제목 생성에 더 효과적이다는 것을 확인할 수 있었다. 따라서, 제안된 프레임워크의 입력 구성 방법론들 중에서 'Backward' 방식으로 입력 순서를 정하고, 'Rotation' 방식으로 입력 형태를 구성하는 것이 가장 효과적이며 고난이도의 가짜뉴스 제목을 생성할 수 있는 방법이라는 결론을 도출하였다.

(2) 정성 평가 결과

라벨링 오류와 각 방법론 간의 난이도 차이를 실질적으로 평가할 수 있는 지표로서, 인간 평가자에 의한 정성 평가를 시

행하였다. 본 연구의 정성 평가는 라벨링 오류 필터링 후의 데이터셋에서 뉴스 카테고리별로 무작위로 15개의 항목을 추출하여, 총 105개의 데이터에 대해 평가가 이루어졌다. 평가 과정에는 3명의 평가자가 참여하였으며, 각 평가자는 TF-IDF를 포함한 7가지 방법론으로 생성된 가짜뉴스 제목 35개에 대해 평가를 수행하였다. 라벨링 오류 비율은 가짜뉴스 제목과 실제 뉴스 본문을 비교하여, 실제 뉴스에 없거나 일치하지 않는 정보가 포함되었는지를 평가함으로써 측정하였다. 또한, 평균 순위는 라벨링 오류가 발견되지 않은 경우에 한 해 각 방법론 간의 상대적인 난이도를 고려하여 순위를 부여함으로써 산출하였다. 즉, 평균 순위는 라벨링 오류 제외 후, 가짜뉴스 제목 간의 상대적인 난이도 순위를 의미한다.

<Table 5>의 라벨링 오류 비율을 살펴보면, 'Forward' 방법론들이 'Backward' 방법론들에 비해 눈에 띄게 더 많은 라벨링 오류를 보이는 것을 확인할 수 있다. 이와 대조적으로, 교체 방법론인 TF-IDF는 상대적으로 적은 라벨링 오류를 보인다. 라벨링 오류가 비교적 적은 'Backward' 방법론 간의 난이도를 비교할 때, 'Rotation' 방법론이 가장 높은 평균 순위를 차지함을 알 수 있으며, 이는 'Rotation' 방법론이 어려운 가짜뉴스를 가장 효과적으로 생성하는 방법론이라는 것을 시사한다. 그리고 'Summarization' 방법론은 TF-IDF에 비해 난이도가 낮은 가짜뉴스 제목을 생성한다는 것을 알 수 있다. 또한, 'Forward'와 'Backward' 방법론들의 평균 순위를 비교하면, 'Forward' 방법론들의 평균 순위가 전반적으로 더 높다는 것을 알 수 있다. 이는 라벨링 오류 비율이 높아, 평가 과정에서 제외된 데이터가 많았기 때문으로 추정된다. 정량 평가 결과를 종합적으로 고려해보면, 정량 평가와 정성 평가 결과 간의 연관성이 높다고 결론지을 수 있다.

Table 5. Human Evaluation Results

Method		Labeling Error Rate	Average Rank
Forward	Chunking	77%	1.7
	Rotation	62%	1.6
	Summarization	51%	3.0
Backward	Chunking	19%	2.1
	Rotation	24%	1.7
	Summarization	21%	3.3
TF-IDF		14%	2.3

정량 평가와 정성 평가의 연관성을 분석하고 제안한 평가 지표들의 타당성을 보이기 위해, OLER, Difficulty, 라벨링 오류 비율, 평균 순위 간의 상관관계를 분석하였다. <Figure 4>의 결과에서 알 수 있듯이, 라벨링 오류와 관련된 지표인 OLER과 라벨링 오류 비율 간의 높은 상관성(0.75)이 존재한다는 것을 알 수 있다. 이는 OLER이 라벨링 오류를 정확하게 반영하고 있음을 의미한다. 또한, 난이도를 나타내는 지표인 Difficulty와 평균 순위 간에도 높은 상관성(0.72)이 존재한다는 것을 알 수 있다. 마찬가지로, 이는 Difficulty가 실제로 가짜뉴스의 난이도를 정확하게 반영할 수 있음을 의미한다. 이러한 결과는 본 연구에서 도입한 평가 방법이 가짜뉴스 데이터에 대한 인간의 판단과 높은 수준으로 일치하며, 제안된 평가 지표들이 실질적인 유효성을 가지고 있다는 것을 입증한다.

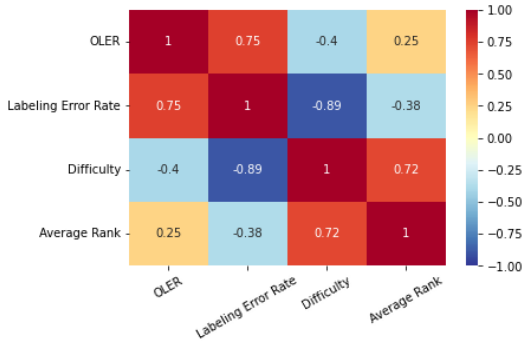

Figure 4. Correlation between Main Results and Human Evaluation Results

(3) 유사뉴스 탐색 방식에 따른 효과성 비교

본 연구에서는 유사뉴스 탐색 단계에서 TF-IDF를 사용하여 원본뉴스와 가장 유사한 뉴스를 선택하여 가짜뉴스를 생성하는 방식(Top1)과 세 번째로 유사한 뉴스를 선택하여 가짜뉴스 제목을 생성하는 방식(Top3)의 효과를 비교하기 위한 실험을 수행하였다. 이러한 실험은 원본뉴스와 구별하기 어려운 유사뉴스 정보를 가짜뉴스에 주입하는 것이 무작위 정보 주입보다 가짜뉴스 생성에 더 효과적이라는 가정을 검증하기 위한 것이다.

Table 6. Comparison results from Top1 and Top3

Method		Top1		Top3	
		OLER	Difficulty	OLER	Difficulty
Forward	Chunking	28.1%	0.6305	27.6%	0.5575
	Rotation	22.8%	0.6268	21.5%	0.6787
Backward	Chunking	5.7%	0.8519	2.9%	0.8986
	Rotation	8.1%	0.8148	4.8%	0.8675
TF-IDF		4.0%	0.8710	1.2%	0.9254

<Table 6>의 결과를 살펴보면, OLER값은 Top1 방식과 Top3 방식 사이에 큰 차이가 없음을 확인할 수 있다. 그러나, 난이도 지표인 Difficulty를 기준으로 볼 때, 유의미한 차이를 보이며, 이는 Top1 방식의 가짜뉴스의 난이도 측면에서 더 효과적이라는 것을 시사한다. 이 결과는 원본뉴스와 높은 연관성을 가진 정보를 활용하여 가짜뉴스를 생성하는 접근 방법이 난이도가 높은 가짜뉴스 생성 과정에서 중요한 요소라는 사실을 나타낸다. 추가로, 교체 방법론인 TF-IDF 간의 결과를 비교를 통해, 이러한 차이가 더욱 명확하게 드러나게 된다.

(4) 가짜뉴스 제목 생성 전략에 따른 효과성 비교

본 연구에서는 본문만을 활용하여 가짜뉴스 제목을 생성하는 전략과 제목과 본문을 동시에 활용하여 가짜뉴스 제목을 생성하는 전략 간의 효과성을 비교하기 위한 실험을 수행하였다.

<Table 7>의 결과에 따르면, 제목과 본문을 동시에 활용하여 가짜뉴스 제목을 생성하는 전략이 본문만을 활용하는 방법에 비해 OLER 및 Difficulty 지표에서 상당히 뛰어난 수치를 보이고 있다. 이러한 결과는 뉴스 제목을 가짜뉴스 제목 생성에 직접적으로 활용함으로써 실제 뉴스 제목과 유사한 스타일의 문장을 효과적으로 생성할 수 있게 되어, 생성된 가짜뉴스 제목의 난이도가 상승한다는 것을 시사한다. 더불어, OLER 지표에서의 큰 차이는 제목과 본문을 함께 사용함으로써 가짜뉴스 제목 생성 모델에 더욱 명시적인 정보를 제공하게 되어, 가짜뉴스 제목 생성에 긍정적인 영향을 미친다는 것을 암시한다.

Table 7. Comparison Results from Title-Content and Content

Method		Title-Content		Content	
		OLER	Difficulty	OLER	Difficulty
Forward	Chunking	28.1%	0.6305	50.7%	0.7604
	Rotation	22.8%	0.6268	39.8%	0.6648
Backward	Chunking	5.7%	0.8519	12.2%	0.8512
	Rotation	8.1%	0.8148	16.9%	0.8203

5. 결론

본 연구에서는 비지도 학습 기반의 전략을 통해 진짜뉴스 데이터를 활용하여 고품질의 가짜뉴스 데이터를 생성하는 프레임워크를 제안하였다. 제안된 프레임워크는 가짜뉴스 탐지 모델의 연구와 개발에 효과적으로 활용될 수 있으며, 언어와 데이터 종류와 관계없이 일반화 가능한 방법론으로서 특히 가짜뉴스 탐지 분야에서 고품질 가짜뉴스 데이터 확보에 크게 기여할 것으로 예상된다. 제안된 프레임워크는 난이도가 높은 가짜뉴스를 생성하고, 라벨링 오류를 필터링함으로써 생성된 가짜뉴스 데이터의 품질을 향상시킨다. 더불어, 본 연구는 데이터 품질을 폭넓게 평가할 수 있는 새로운 평가 지표와 방법론을 제시한다. 다양한 실험을 통해 제안한 프레임워크의 유효성과 효과성을 검증하였고, 가짜뉴스 생성에 관한 근본적인 이해를 제공하였다.

본 논문의 프레임워크는 가짜뉴스 탐지 모델의 더욱 효과적인 개발을 촉진하고, 가짜뉴스로 인한 피해를 감소시킬 수 있는 가능성을 제시한다. 또한, 본 연구의 결과를 바탕으로, 생성 모델이 직접 가짜뉴스 제목을 생성하도록 훈련시키는 새로운 연구 방향을 제시할 수 있으며, 이를 통해 더욱 발전된 가짜뉴스 탐지 방법론의 개발이 기대된다.

참고문헌

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014), Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv preprint arXiv:1412.3555.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018), Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805.

Huang, K., McKeown, K., Nakov, P., Choi, Y., and Ji, H. (2022), Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation, arXiv preprint arXiv:2203.05386.

Hwang, Y., and Kwon, O. (2017), Conceptualizing and Regulating Fake News Means: Self-regulation of Internet Service Providers, *With a Focus on Korea. Media and Law*, **16**(1), 53-101.

Jang, J., Cho, H., Lee, J., and Kim, M. (2021), Development of a Hierarchical Deep Learning Model for Fake News Detection with Different Statements and Building a Fake News Dataset. Development and Fake News Dataset Construction, *Korea Intelligence Conference Proceedings*, 1939-1941.

Jwa, H., Oh, D., Park, K., Kang, J. M., and Lim, H. (2019), Exbake: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (bert), *Applied Sciences*, **9**(19), 4062.

Khattab, O. and Zaharia, M. (2020), Colbert: Efficient and Effective Passage Search via Contextualized Late Interaction over Bert, In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39-48.

Lin, C. Y. (2004), ROUGE: A Package for Automatic Evaluation of Summaries, *In Text Summarization Branches Out*, pages 74-81, Barcelona, Spain. Association for Computational Linguistics.

Nagoudi, E. M. B., Elmadany, A., Abdul-Mageed, M., Alhindi, T., and Cavusoglu, H. (2020), Machine Generation and Detection of Arabic Manipulated and Fake News, arXiv preprint arXiv:2011.03092.

Oh, M., Lee, J., Choi, H., Jin, J., and Chun, K. (2022), A Study on Machine Learning-based Fake News Detection Method in Health and Welfare, *Korea Institute For Health And Social Affairs*, 2022-48.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002), Bleu: A Method for Automatic Evaluation of Machine Translation, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics, 311-318.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021), Mauve: Measuring the Gap between Neural Text and Human Text Using Divergence Frontiers, *Advances in Neural Information Processing Systems*, **34**, 4816-4828.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020), Exploring the Limits of Transfer Learning with a Unified Text-to-text Trans-former, *The Journal of Machine Learning Research*, **21**(1), 5485-5551.

Shim, J. S., Won, H. R., and Ahn, H. (2019), A Study on the Effect of the Document Summarization Technique on the Fake News Detection Model, *Intelligence and Information Research*, **25**(3), 201-220.

Shu, K., Li, Y., Ding, K., and Liu, H. (2021), Fact-enhanced Synthetic News Generation, In *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 13825-13833.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez,

A. N., Kaiser, Ł., and Polosukhin, I., (2017), Attention is All You Need, *Advances in Neural Information Processing Systems*, 30.
Yoon, S., Park, K., Shin, J., Lim, H., Won, S., Cha, M., and Jung, K. (2019), Detecting Incongruity between News Headline and Body Text Via a Deep Hierarchical Encoder, In *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 791-800.

저자소개

김중훈 : 중앙대학교 응용통계학과에서 2022년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석사과정에 재학 중이다. 연구 분야는 자연어 처리 및 추천시스템이다.

박새란 : 세종대학교 데이터사이언스학과에서 2023년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석사과정에 재학 중이다. 연구분야는 자연어처리 및 대화형 추천시스템이다.

이지윤 : 광운대학교 정보융합학부에서 2023년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석사과정에 재학 중이다. 연구분야는 자연어처리 및 대화형 추천시스템이다.

김재희 : 성균관대학교 소비자학과에서 2022년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석박사통합과정에 재학 중이다. 연구분야는 Information Retrieval 및 Large Language Model이다.

강필성 : 서울대학교 산업공학과에서 2003년 학사, 2010년 박사학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 정교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.