기계학습 기반 국내 뉴스 헤드라인의 정확성 검증 연구

백지수[1]○, 이승언[2,3], 한지영[3], 차미영[2,3]
[1]한국방송통신대학교, [2]IBS 데이터사이언스 그룹, [3]한국과학기술원
white.100.js@gmail.com, marinearchon159@gmail.com, jiyoung.han@kaist.ac.kr, mcha@ibs.re.kr

Objectivity in Korean News Reporting : Machine Learning-Based Verification of News Headline Accuracy

Jisoo Baik[1]○, Seung Eon Lee[2,3], Jiyoung Han[3], Meeyoung Cha[2,3]
[1]Korean National Open University, [2]IBS Data Science Group, [3]KAIST

요 약

뉴스 헤드라인에 제3자의 발언을 직접 인용해 전언하는 이른바 '따옴표 저널리즘'이 언론 보도의 객관주의 원칙을 해치는지는 언론학 및 뉴스 구독자에게 중요한 문제이다. 이 연구는 온라인 포털사이트를 통해 실시간 유통되는 한국어 기사의 정확성을 판별하기 위한 기계학습(Machine Learning) 모델을 제안한다. 이 연구에서 제안하는 모델은 Edit Distance와 FastText 기법을 활용해 기사 제목과 본문 내 인용구의 유사성을 측정하고, XGBoost 모델을 활용해 최종 분류한다. 아울러 이 모델을 통해 229만 건의 뉴스 헤드라인에 대해 직접 인용구가 포함된 기사가 취재원의 발언을 주관적인 윤색없이 독자들에게 전하고 있는지를 판별했다. 이뿐만 아니라 딥러닝 기반의 KoELECTRA 모델을 활용해 기사의 제목 내 인용구에 대한 감성 분석을 진행했다. 분석 결과, 윤색이 가미되지 않은 직접 인용형 기사의 비율이 지난 20년 동안 10% 이상 증가했으며, 기사 제목의 인용구에 나타나는 감정은 부정 감성이 긍정 감성의 2.8배 정도로 우세했다. 이러한 시도는 앞으로 계산사회과학 방법론과 빅데이터에 기반한 언론 보도의 평가 및 개선에 도움을 주리라 기대한다.

주제어: 한국어 뉴스 데이터, 따옴표 저널리즘, 뉴스 헤드라인, 정확도 판단, 인용구, XGBoost, KoELECTRA

1. 서론

있는 그대로의 사실(Fact)만 보도해야 한다는 객관주의(Objecitivism)는 언론에 통용되는 기본 원칙이다[1-3]. 하지만 국내 언론은 '따옴표 저널리즘'[1]을 남용하면서 객관주의 원칙을 도구화하고 있다는 비평을 받아왔다[4-6]. 여기에는 기자가 취재원으로부터 얻은 정보를 사실 여부를 검증하지 않고 단순 인용하는 보도 관행이 언론에 사실 검증 책임을 회피할 여지를 준다는 비판이 담겼다. 앞선 저널리즘 연구들은 어떤 기사가 기자의 주관이 개입되지 않고 객관성을 띠는가를 판단하는 여러 기준을 제시했지만, 대체로 특정 주제에 한정된 수백 건 안팎의 기사 표본을 수작업으로 분석하는 연구 방식을 취해 연구 결과를 일반화하기 어렵다는 한계가 있었다.

이 연구에서는 따옴표 저널리즘의 현황 및 장기간의 변화를 파악하기 위해 직접 인용구가 뉴스 헤드라인에 포함되는지 여부를 머신러닝을 통해 검증한다. 이를 위해 인용구의 형태를 인공신경망으로 학습하고 국내 뉴스 중 제목에 직접 인용구가 있는 기사 229만여 건을 추출해 연구 대상으로 삼아, 20년에 걸친 뉴스 데이터로부터 인용 표현의 현황 및 변화를 탐지한다.

이를 위해 크게 두 가지 분류 모델을 고안했다. 먼저 기사 제목 내 직접 인용구가 형식적으로나마 객관주의 원칙을 따르고 있는 기사가 얼마나 되는지 XGBoost[7] 기반 모델을 통해 검증한다. 이를 활용해 기자가 기사 본문에서 인용한 취재원의 발화를 제목에도 단순 요약하여 있는 그대로 옮긴 '직접 인용형' 기사와 본문에 없는 말을 지어내 직접 인용구로 붙인 '작문형' 기사로 구분한다. 다음으로 제목의 직접 인용구가 전달하는 뉘앙스가 중립적이지 않은 기사가 얼마나 되는지 확인하기 위해 최신 한국어 자연어 처리 모델인 KoELECTRA[8] 기반 모델을 활용한 감성 분석(Sentiment Analysis)을 통해 '부정·긍정·중립' 등 3개 클래스로 분류한다.

이 연구의 학술적 목표는 다음과 같다.

(1) 기존 기사 분석 연구들이 정보의 진위 판단(예: 특정인의 주장에 대한 진위 여부)에 중점을 두고 있다면, 뉴스 의견성 지수(인용형 제목의 정확성 연구)는 취재원의 말이 기사에서 어떻게 활용되는가를 통해 기사의 편향성을 분석할 수 있다는 관점의 전환을 제시한다.

(2) 기존 연구는 현직 기자의 상당수가 자신의 의견을

1) 따옴표 저널리즘은 취재원의 발화를 직접 인용한 문장 위주로 제목과 본문을 구성하는 보도 관행을 비판적으로 일컫는 표현이다[6].

익명의 취재원을 활용해 기사에 인용한 경우가 있다고 증언했는데[9] 이러한 사례들은 취재원을 말을 제목에 직접 인용함으로써 기사 내용을 주관적으로 윤색하고[6, 10], 보도의 객관성과 신뢰성을 확보하기 위해 사실 보도의 형태를 차용해 정파성을 드러내는 도구적 객관주의[11] 보도 행태를 잘 보여준다. 따라서 인용형 제목의 정확성 분석은 한국 언론의 매체 편집 방향보다 통합적이고 체계적인 방법으로 언론의 정파성을 유형화하려는 학문적 노력에 기여할 수 있다.

2. 관련 선행 연구

국내 언론의 따옴표 저널리즘 관행을 비판한 저널리즘 분야 선행 연구자들은 큰따옴표(“”)를 이용한 직접 인용구 중심의 기사 작성이 기사의 객관성과 신뢰성을 해친다고 지적한다. 취재원의 실제 발언을 ‘발언했다는 사실’ 그대로 옮겨 중립성을 보장하는 것처럼 보이지만, 이는 한국에서만 발견되는 ‘형식적 객관주의’에 지나지 않는다는 것이다[4]. 특히 다수 선행 연구가 기사 제목에서의 따옴표 저널리즘을 경계하는데, 이는 취재원의 일부 발언만 과장해 전달하거나 취재원의 발언을 윤색하는 과정에 기자 개인의 주관이 개입될 여지가 있기 때문이다[5].

국내 언론계의 주장은 상반된다. 현실의 사실을 기사로 재구성하기 위해 선별된 정보원을 최대한 사실 그대로 인용할 필요가 있고, 보도가 필요한 의견을 기사에 담을 때 편향성을 회피하기 위해서도 직접 인용구를 사용할 수밖에 없는 상황이 존재한다는 것이다[12]. 이에 반해, 언론계의 주장을 언론학자들은 모순이라고 지적하는데, 직접 인용구를 제목에 사용하는 것이 언론사의 주관을 전달하기 위한 도구일 뿐이라는 점을 언론계 스스로 자인하는 것에 지나지 않는다는 것이다[12].

일부 연구는 기사 제목에 직접 인용구를 포함하는 것이 어떤 측면에서 기사의 객관성을 저해하는지를 분석했다. 그 예로, 제목 내 직접 인용구의 본문 실재 여부, 기사 본문과의 관련성, 주제(제목 내 직접 인용구가 개인의 의지나 감성을 인용하는지 여부) 등을 기사의 객관성 판단 기준으로 제시했다[6].

한국어 정보 처리 분야에서도 선행 연구와 비슷한 방법으로 국내 뉴스의 객관성을 분석한 연구들이 존재했다. 이를테면 경제 분야의 1천여 건을 대상으로 기사 제목 내 인용 형태와 의견성을 분석한다거나[13] 특정 언론사의 기사와 의도적으로 생성된 가짜뉴스로 구축한 5만 건의 기사로 제목과 본문이 다른 뉴스를 탐지하는 연구[14] 등이 수행된 바 있다.

3. 데이터 수집

이 연구는 국내 뉴스 데이터셋으로 수행한 일부 한국어 정보처리 분야의 선행 연구들보다도 광범위하고 큰 규모의 실제 국내 뉴스를 데이터셋으로 사용해 분류 결과의 일반화를 시도했다. 국내 온라인 플랫폼을 통해 축적된 한국어 뉴스 데이터 중 주제와 관계없이 수집된 1800만여 건의 기사 중 제목에 직접 인용구가 존재하는 약 230만 건의 기사를 연구 대상 데이터셋으로 삼았다.

구체적으로 2000년대 이후 4개 정권이 3년 차에 접어든 시점부터 1년 동안 국내 언론이 보도한 1802만6897건의 기사를 연구 표본 데이터로 수집했다. 이중 제목에 직접 인용구가 포함된 기사 229만8694건(12.75%)을 추출해 데이터셋을 구축했다. 국내 온라인 포털사이트 네이버뉴스 플랫폼에서 유통되는 온라인 기사를 웹크롤링 방식으로 모았다.

네이버에서는 언론사의 분류에 따라 여섯 종류의 섹션(정치, 경제, 사회, 국제, 생활·문화, IT·과학)으로 기사를 노출하는데, 두 개 이상의 섹션에서 같은 기사가 중복으로 수집될 경우 각각 한 건의 데이터로 간주했다. 일례로 동일 기사가 사회 섹션과 정치 섹션에 중복 노출될 경우 같은 기사이더라도 데이터가 총 두 건 존재하는 것으로 처리했다. 수집한 텍스트에서 제목에 직접 인용구를 포함한 기사를 추출할 때에는 파이썬(Python)의 정규표현식(Regular Expression) 모듈을 이용했다. 실험에서는 모델마다 전체 데이터의 80%를 모델 훈련(Train) 데이터로, 10%를 검증(Validation) 데이터로, 나머지 10%를 시험(Test) 데이터로 사용했다.

데이터셋 내 기사의 구체적인 보도 시점은 표 1과 같다. 조기 대선으로 당선된 문재인 대통령을 제외한 나머지 각 대통령의 재임 기간 기사는 각 대통령의 취임일(대통령 당선 이듬해 2월25일)로부터 만 2년째 되는 날부터 1년 동안 보도된 기사를 모았다. 문 대통령의 취임만 2주년은 2019년 5월이지만 연구의 편의를 위해 앞선 세 대통령 재임 기간의 데이터와 수집 기준 시점을 당해 2월25일부터 1년간으로 일치시켰다.

정권별 전체 수집 기사 수 대비 직접 인용구 포함 제목 기사의 비중을 살펴본 결과, 그 비중은 정권별로 유의미한 차이가 나타나지 않는다고 판단했다. (4개 정권별 비율의 분포에 대한 카이제곱검정 결과 χ^2 = 0.13969, p-value = 0.99, DF = 3를 나타냄.) 제목에 직접 인용구를 포함한 기사의 각 시기별 비율은 이명박 정부 3년차에 11.81%로 가장 낮았고, 노무현 정부 3년차에 13.59%로 가장 높았다.

표 1 : 연구 대상 기사 보도 시점과 각 정권 시점별 직접 인용구 포함 제목 기사 비율

대통령	취임일	연구 대상 기사 보도 시점	직접 인용구 포함 제목 비율(%)
노무현	2003. 2. 25.	2005. 2. 25. ~ 2006. 2. 24.	13.59
이명박	2008. 2. 25.	2010. 2. 25. ~ 2011. 2. 24.	11.81
박근혜	2013. 2. 25.	2015. 2. 25. ~ 2016. 2. 24.	13.2
문재인	2017. 5. 10.	2019. 2. 25. ~ 2020. 2. 24.	12.64

4. 세부 연구 주제별 사용 모델

4.1. 제목 내 직접 인용구의 본문 실재 여부 판별

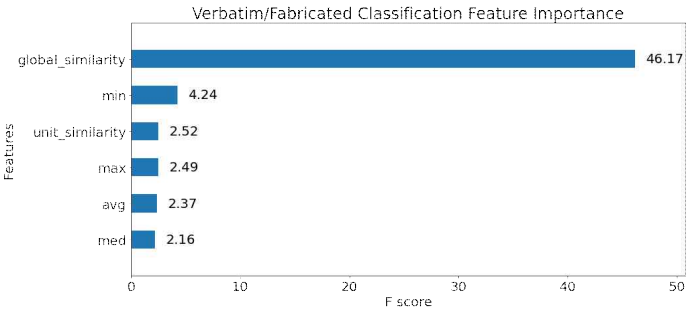
직접 인용구 제목에 포함된 기사가 객관적으로 사실을 전달하려면 제목의 직접 인용구가 기사 본문에도 동일하게 존재한다는 선행 연구[6]에 주목했다. 제목 내 직접 인용구가 해당 기사의 본문에 등장한 직접 인용구와 형태나 의미적 차이가 미미한 기사를 '직접 인용형(Verbatim)', 그 외 기사를 '작문형(Fabricated)'으로 정의하고 데이터셋을 이진 분류하는 트리(Tree) 구조의 XGBoost 모델을 구성했다.[7]

이 연구의 모델은 이를테면 <IMF, "한국경제 앞날 밝다">라는 제목의 기사의 본문에도 "한국경제 앞날이 밝다"와 동일한 구절을 포함한 직접 인용구가 존재한다면 직접 인용형 기사로 분류하도록 고안됐다.

모델의 입력 데이터로는 제목과 본문에서 큰따옴표("")로 묶인 직접 인용구만 정규 표현식으로 추출한 뒤, 추출한 문장을 워드 임베딩(Word Embedding)하여 만들어진 벡터 간 유사성 거리를 이용했다. 구체적으로는 기사마다 제목 인용구 벡터와 2개 이상의 본문 인용구 벡터 간 유사성 거리를 측정해 모델에 입력했다. 제목에는 직접 인용구가 있으나 본문에는 직접 인용구가 없는 단신 기사는 분석 과정에서 데이터 전처리를 통해 제외했다. 그 결과 실제 분류 분석에 이용된 데이터는 172만2176건으로 집계됐다.

제목 인용구와 본문 인용구 간의 유사성 거리 측정은 두 가지 방법을 동시에 사용했다. 먼저, 파이썬 Edit Distance 모듈로 레벤슈타인 거리(Levenshtein Distance)[15]의 최소·최대·평균·중앙값을 분류 예측에 활용했다. 레벤슈타인 거리는 '편집 거리'라고도 불리는 지표로, 이를 활용하면 문자열 매칭을 정량화할 수 있다. 예를 들어 5개 글자 차이로 서로 다른 의미를 가지는 'Microsoft'라는 단어와 'Soft'의 레벤슈타인 거리를 5로 표현하는 식이다. 연구팀은 이러한 원리를 이용하여 레벤슈타인 거리로 제목 인용구와 본문 인용구가 비슷한지를 판단할 수 있을 것이라고 판단했다.

하지만 단순 문자열 매칭만으로 유사도를 판단하기 어려운 사례들이 연구 중 발견되었다. 이를테면 '집값 폭등'과 '부동산 가격 상승'처럼 같은 의미이지만 형태의 차이가 큰 사례가 적지 않았다. 이 때문에 레벤슈타인 거리만으로는 이같은 사례까지 정확히 분류하기 어렵다고 판단했다. 따라서 단어의 통사와 형태까지 학습할 수 있는 패스트텍스트(FastText) 모델[16]을 이용해 구한 단어 단위 벡터 간 코사인 거리도 레벤슈타인 거리와 동시에 고려했다. 이때 제목과 본문 내 인용구 각각에 포함된 단어들 간의 코사인 거리로 평균을 구한 것이 Unit Similarity, 본문 내 인용구를 종합한 내용에 포함된 단어의 코사인 거리와 제목 내 인용구에 포함된 단어의 코사인 거리를 구해 비교한 것이 Global Similarity이다. 그림 1은 연구에서 사용한 XGBoost 모델의 요인 중요도를 나타내며, 사용한 모든 유사도 지표 중 패스트텍스트 기반 Global Similarity가 가장 중요한 요인으로 작용한 것을 볼 수 있다.



[그림 1] 직접 인용형·작문형 기사 분류에 사용한 XGBoost 모델의 Feature Importance 그래프

이 연구에 사용한 XGBoost 모델은 사전에 사람이 수작업으로 라벨을 붙인 2026건의 훈련용 데이터(train dataset)를 학습했다. 훈련용 데이터로 이용한 기사는 연구 초반 구축한 데이터셋에서 무작위 추출했다. 라벨링 작업을 위해 언론학 전공 코더(coder) (2명)을 고용했다.

당초 수작업 라벨링 과정에서는 제목의 직접 인용구의 각색 정도를 판단해 5개 등급으로 분류했으며, 코더간 신뢰도(Krippendorff's alpha)는 .93으로 매우 높았다(5등급 분류 기준은 표 2를 참조). 그러나 제목에 직접 인용구를 포함한 기사가 형식적으로나마 객관주의를 따르려면 제목의 인용구에도 취재원의 발언에 사용된 모든 형태소를 그대로 옮겨야 한다고 판단했다.

표 2 : 제목 직접 인용구의 각색 정도에 따른 5등급 분류

분류		설명
Verbatim	완전 직접인용형	본문 인용구 일부 또는 전체를 그대로 옮긴 제목.
	단순변형형	본문 인용구에서의 의미 변화 없이 동의어로 어휘를 교체한 제목. 본문 직접 인용문에 직접 언급되지 않은 주어, 목적어 등을 채워넣은 제목.
Edited	윤색형	본문 인용구와 인용구에 의미(뉘앙스) 변화가 있는 제목.
Fabricated	요약형	본문 내용 중 직접 인용되지 않은 내용을 직접 인용 형식으로 제목에 포함한 경우.
	작문형	본문에 제목 인용구의 내용이 등장하지 않고 개략적인 내용도 없는 경우.

따라서 XGBoost 모델을 설계하고 훈련시킬 때에는 제목과 본문의 인용 형태 간 일치도가 가장 높은 '완전 직접 인용형' 기사인지, 제목 인용구에 내용 왜곡이 발생한 기사인지('요약형' 또는 '작문형'인지) 여부만 고려했다. 표 2와 같이 수작업 라벨 중 모델 학습에 실질적인 영향력이 적다고 판단한 '단순변형형'과 '윤색형' 544건은 제외하고 Verbatim과 Fabricated 두 가지로 훈련용 라벨을 재분류하여 모델을 학습시켰다. 훈련용 데이터 2026건에 대한 전처리 과정에서 인용구 추출이 이뤄지지 않은 9건의 기사도 제외하여 실제 훈련에는 Verbatim 라벨 데이터 273건, Fabricated 라벨 데이터 851건 등 총 1668건이 이용됐다.

4.2. 제목 내 직접 인용구 감정 분석

직접 인용구가 포함된 제목으로 기사를 보도할 때 제목의 직접 인용구가 전달하는 내용의 뉘앙스가 부정적일 경우 특정 대상에 대한 부정적 인식을 전파할 수 있어 기사의 객관성을 해칠 수 있다.[5] 이같은 지적에 초점을 맞춰 감성 분석 모델을 설계했다. 언론이 취재원이 발화한 비판·비방·우려·지적 등의 부정적 발언을 직접 인용해 제목에 내세울 경우, 해당 보도는 의도적으로 제3의 대상에 대해 부정적인 인식을 퍼뜨리는 도구로 전락할 수 있다는 점 때문이다.

감성 분석에는 최신 한국어 사전학습(Pre-training) 모델인 KoELECTRA 모델[8]을 사용했다. 원본 모델인 ELECTRA는 딥러닝 기반 자연어 처리 분야에서 'State-of-the-art'로 꼽힌다. 생성자(Generator)가 만들어낸 가짜 토큰(Replaced Token)을 학습할 문장에 포함한 뒤 각 토큰이 원본(Original)인지 아닌지를 탐지하는 Replaced Token Detection(RTD)을 통해 기존의 사전학습 모델 BERT에 비해 연산 효율과 예측 성능이 높은 것으로 알려져 있다.[17]

이 연구는 파이썬 Transformers 라이브러리를 이용해 한국어 ELECTRA 모델인 KoELECTRA를 사용했다. 모델의 입력 데이터로는 각 기사 제목에서 직접 인용구만 정규 표현식으로 추출한 뒤 따옴표를 뗀 문장들을 사용했다. 4.1.에서 예시로 든 기사 제목을 다시 예로 들면 '한국경제 앞날 밝다'라는 문장을 모델에 입력해 분류 분석을 수행했다. 각 인용구를 분류하는 클래스는 부정(Negative)과 긍정(Positive), 중립(Neutral) 등 세 가지로 설정했다. Label은 부정은 0, 긍정은 1, 중립은 2로 정의했다. KoELECTRA 모델 훈련을 위한 샘플 데이터는 총 1만585건으로 역시 코더의 수작업 라벨링하여 사용했다. 각각 부정 3536건, 긍정 1286건, 중립 5763건으로 라벨링하여 이중 8580 건을 실제 훈련에 이용했다.

5. 활용 모델의 성능

5.1. XGBoost 모델의 직접 인용형·작문형 분류 성능

직접 인용구를 지닌 제목 기사의 라벨을 학습한 XGBoost 모델은 시험 데이터셋에 대해 정확도 0.90, AUROC 스코어 0.85를 나타냈다. 구체적인 분류 성능은 Macro 평균 기준 표 3과 같다. 연구의 중점 관심사인 완전 직접 인용형 기사 분류에 대해서도 F1-score 0.78을 보장했다.

표 3 : 시험 데이터를 기준으로 측정한 직접 인용형 기사 분류 결과

	Precision	Recall	F1	AUROC
Levenstein	0.74	0.74	0.74	0.74
Levenstein +FastText	0.88	0.85	0.86	0.85

5.2. KoELECTRA 모델의 부정·긍정·중립 분류 성능

KoELECTRA 모델로 기사 제목 인용구의 어조를 예측한 결과 Test Macro-F1 Score가 0.78임을 확인했다.

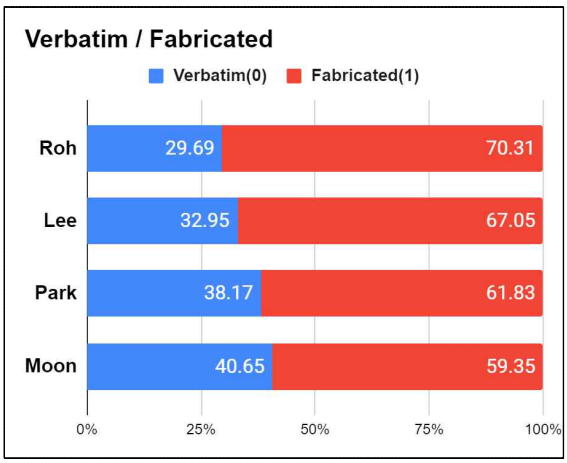
6. 분석 결과

6.1. 직접 인용형·작문형 기사 분류 결과

XGBoost 모델을 직접 인용구를 가진 제목 기사 데이

터셋에 적용한 결과, 본문에 직접 인용구가 없는 단신 기사를 제외한 172만2176건 중 64만5643건(37.5%)이 본문의 직접 인용구가 제목에 동일한 형태로 옮겨진 직접 인용형 기사로 분류되었다. 또 데이터셋 중 절반 이상에 해당하는 107만6533건(62.5%)의 기사가 본문에 없는 내용이 제목에 직접 인용구로 붙여진 작문형 기사로 분류되었다.

데이터셋을 구축할 때 기사의 보도 시점을 고려한 만큼 각 정권별 직접 인용형·작문형 기사 비중의 추이도 살폈다. 그 결과 그림 2와 같이 노무현 정부 이래 국내 언론 보도에서 직접 인용형 기사의 비율이 꾸준히 높아지고 있음을 확인했다. 노무현 정부 당시 30%가 채 되지 않았던 직접 인용형 기사 비율은 이명박 정부에서 30%대 초반, 박근혜 정부에서 30%대 후반까지 늘어난 뒤 문재인 정부에 들어서는 40%를 넘겼다.

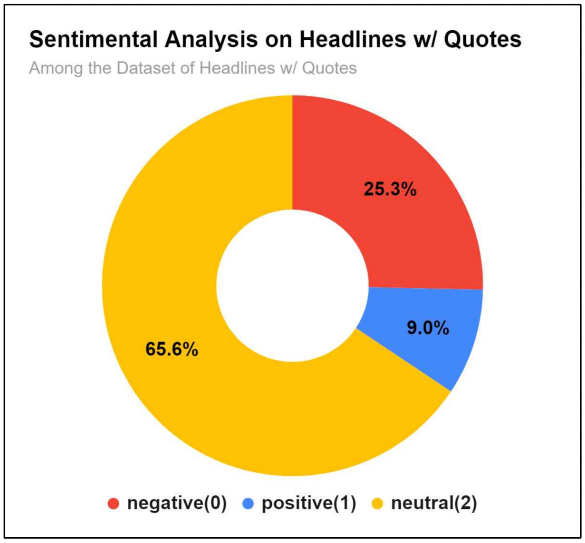


[그림 2] 직접 인용형(Verbatim), 작문형(Fabricated) 분류 결과의 정권별 추이

6.2. 제목 내 직접 인용구의 감성 분석 분류 결과

KoELECTRA 모델을 활용한 감성 분석 결과, 그림 3처럼 전체 데이터셋의 25.3%(58만2221건)가 부정적인 직접 인용구를 제목에 사용했다고 분류됐다. 또 전체의 65.6%(150만8680건)의 기사는 중립적인 직접 인용구를 포함했다고 분류됐다. 긍정적인 내용의 직접 인용구를 포함한다고 분류된 기사는 9.04%(20만7793건)에 그쳤다.

연구에서 사용한 KoELECTRA 모델은 모든 정권에서 약 25%씩은 제목의 직접 인용구가 부정적인 내용을 담고 있다는 분류 결과를 내놓았다. 아울러 각 정권마다 부정 : 긍정 : 중립의 비중을 25 : 9 : 65 비율로 일정하게 나타냈다.



[그림 3] 직접 인용형 제목의 인용구에 대한 감성 분석 분류 결과

7. 결론

한국 언론의 정파성이 1997년 김대중 정부 이후 두드러졌다는 연구 결과[18, 19]에 따라 이후 언론 정파성을 규명하기 위한 저널리즘 연구는 대부분 사드 논쟁[20], 세월호 참사[21] 등 특정 사회 현안에 대한 언론사 간 논조 비교에 초점이 맞춰졌다. 한국 언론이 취하는 도구적 객관주의 보도 행태에 대한 연구도 분석 대상이 1면 기사[5] 또는 특정 주제로 한정돼 있고 적은 양의 데이터를 수작업으로 분석하는 방법에 의존했다. 이같은 선행 연구는 한국 언론 보도의 특이점을 밝혔다는 의의가 있지만 언론 보도의 특징을 사례 연구로 파편화하는 한계를 낳았다.

이 연구는 앞선 연구에서 나타난 한정된 연구 표본의 한계를 국내 뉴스 빅데이터 처리 관점에서 접근해 극복하려 했다. 연구팀은 국내 언론 보도 관행에서 직접 인용구를 포함하는 제목 사용 현황을 파악하고, 직접 인용형태를 데이터를 기반으로 평가하는 모델을 제시했다.

연구 결과 국내 언론에서 객관주의 원칙을 지키려는 노력이 증가하고 있다고 판단했다. 각 정권에서의 전체 보도량 대비 제목에 직접 인용구를 포함한 기사의 비율이 큰 차이가 없던 반면, 제목 내 직접 인용구와 본문 내 직접 인용구가 일치하는 직접 인용형 기사의 비율이 20년 사이 10% 이상 증가했음이 이를 방증한다. 감성 분석 결과로는 부정 감성의 인용 비율이 긍정 감성보다 큰 경향을 확인했다.

추후 연구에서는 이와 같은 변화가 형식적 객관주의뿐만 아니라 뉴스의 질적 향상을 가져오는 건강한 변화인지는 언론학 이론에 기반한 세부 분석을 통해 검토할 필

요가 있다. KoELECTRA를 이용한 분류 결과 제목에 직접 인용구가 있는 기사의 35% 가량에서 편향성이 있다고 나타난 만큼 실제로 언론사나 기자의 주관이 개입된 표현이 증가했는지에 대한 추가 분석이 필요하다. 특히 개별 기사에 대한 세부 연구를 통해 언론이 자극적이거나 부정적 내용을 형식적 객관주의로 포장해 특정 사안에 대한 비판 여론을 의도적으로 퍼트리는 현상이 늘어난 것은 아닌지 검토해야 한다.

마지막으로 데이터 기간에 포함된 세월호 사건과 조기대선 등 역사적 사건을 거치면서 나타난 한국 사회의 양극화가 직접 인용형 기사 비율의 증가에 영향을 미쳤을 가능성도 고려해야 한다. 언론사가 첨예한 가치를 중립적으로 보도하기 위한 방법으로 직접 인용의 형태로 형식적 객관주의를 취해왔을 가능성이 있다. 실제로 사회 섹션의 직접 인용형 기사 비율은 노무현 정부 때에는 정치 섹션을 제외한 다른 섹션에서의 비율과 비슷한 26% 수준이었지만, 문재인 정부 시기에는 38% 수준까지 늘어났다. 이는 따옴표 저널리즘이 두드러지게 나타나는 정치 분야 기사에 비해, 시민사회와 SNS(소셜네트워크서비스)의 성장으로 각계의 의견 분출이 활발해진 것과도 무관하지 않다. 따라서 추후 연구에서는 제목에 직접 인용구가 포함된 기사의 경우 발화자가 존재하는지, 언론사가 대표성 있는 취재원을 인용해 기사의 객관성을 검토해야 한다.

사사 문구

이 논문은 과학기술정보통신부의 재원으로 기초과학연구원의 지원(IBS-R029-C2)과 한국연구재단의 지원을 받아 수행된 기초연구사업임.
(No. NRF-2017R1E1A1A01076400,
NRF-2019S1A5B5A01040041).

참고문헌

[1] J. Boyer, How Editors View Objectivity, Journalism Quarterly, 58(1), pp. 24-28, 1981.
[2] W. Donsbach and B. Klett, Subjective objectivity. How Journalists in Four Countries Define a Key Term of Their Profession, International Communication Gazette, 51(1), pp. 53-83, 1993.
[3] 남재일, "시민인륜의 저널리즘을 위한 이론적 탐색: 객관주의 저널리즘에 대한 성찰과 모색", 민주주의와 인권, 제17권, 제4호, pp. 233-272, 2017.
[4] 이나연, "과학적 객관주의, 형식적 객관주의, 한국형 형식적 객관주의", 한국언론학보, 제62권, 제2호, pp. 112-142, 2018.
[5] 박재영, 이완수, "인용(quotation)과 취재원 적시(attribution)에 대한 한미(韓美) 신문비교", 한국언론학보, 제51권, 제6호, pp. 439-468, 2007.
[6] 이준웅, 양승목, 김규찬, 송현주, "기사 제목에 포함된 직접인용부호 사용의 문제점과 원인", 한국언론학보, 제51권, 제3호, pp. 64-90, 2007.
[7] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
[8] W. Kim and J. Kim, O. Jeong, Dialog-KoELECTRA: Korean conversational language model based on ELECTRA model, https://github.com/skplanet/Dialog-KoELECTRA , 2021.
[9] 남재일, "한국 언론윤리 현황과 과제", 한국언론재단, 2006.
[10] J. Han and G. Lee, A comparative study of the accuracy of quotation-embedded headlines in chosun ilbo and the new york times from 1989 to 2009, Korea Journal, 53(1), pp. 65-90, 2013.
[11] 최영재, "대통령 지지도와 언론보도의 상관관계", 한국방송학회 학술대회 논문집, pp. 288-293, 2007.
[12] 박대민, "사실기사의 직접인용에 대한 이중의 타당성 문제의 검토", 한국언론학보, 제59권, 제5호, pp. 121-151, 2015.
[13] 이영인, 김정욱, 한지영, 김태균, 하유이, 차미영, "딥러닝 기반 모델을 활용한 국내 경제뉴스 제목 내 인용 형태 분류 및 뉴스의 의견성 분석", 한국정보과학회 학술발표논문집, pp. 1794-1796, 2019.
[14] 장준원, 조하현, 이재영, 김미숙, "제목과 본문이 다른 가짜뉴스 탐지를 위한 계층적 딥러닝모델 개발 및 가짜 뉴스 데이터셋 구축", 한국정보과학회 학술발표논문집, pp. 1939-1941, 2021.
[15] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet physics doklady, 10(8), 1966.
[16] P. Bojanowski and E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, 5, pp. 135-146, 2017.
[17] K. Clark, M. Luong, Q. Le, C. Manning, ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, ICLR 2020, 2020.
[18] 송용회, "언론의 현실해석과 객관화 담론전략". 한국언론학보, 제51권, 제1호, pp. 229-251, 2007.
[19] 이준웅, 최영재, "한국 신문위기의 원인", 한국언론학보, 제49권, 제5호, pp. 5-35, 2005.
[20] 홍주현&손영준, "사드 루머(THAAD rumor) 보도에 나타난 한국 언론의 정파성", 한국언론정보학보, 제84권, pp. 152-188, 2017.
[21] 김영욱, 함승경&김영지, "세월호 침몰 사건의 미디어 담론 분석", 한국언론정보학보, 제83권, pp. 7-38, 2017.