JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 8: 999~1005, Aug. 2020

순환 신경망과 합성곱 신경망을 이용한 뉴스 기사 편향도 분석

오승빈[1*] · 김현민[1] · 김승재[1]

Analyzing Media Bias in News Articles Using RNN and CNN

Seungbin Oh[1*] · Hyunmin Kim[1] · Seungjae Kim[1]

[1*]Student, Incheon Academy of Science and Arts, Incheon, 22009 Korea

요 약

오늘날의 검색 포털은 뉴스의 창구로서는 가장 큰 비율을 차지하지만, 중립성에 대해서는 의문이 제기되고 있다. 이는 포털 뉴스가 편향된 정보의 소비를 유도할 수 있기 때문이다. 본 논문은 뉴스 기사의 정치적 편향도를 딥러닝을 이용하여 측정하는 방법에 대하여 소개한다. 이는 기사를 비판적으로 바라보는 시각을 뉴스 독자에게 제공할 것이다. 구체적으로, 국회 회의록에서 추출한 키워드에 편향도를 부여하고, 이를 기반으로 기사의 편향도를 분석하여 머신러닝용 데이터를 구축하였다. 최종적으로 순환 신경망과 합성곱 신경망을 융합한 딥러닝을 통해 기사의 편향도를 계산하는 것을 목표로 하였다. 학습한 모델의 정확도를 분석한 결과 문장별 편향의 좌/우편향 판정은 95.6%의 정확도를 보였으나, 신문기사 전체에서는 46.0%의 정확도를 보였다. 이는 기존의 여러 편향성 연구와 다르게 특정 주제에 한정되지 않고 기사의 보수-진보 편향성을 분석할 수 있도록 한다.

ABSTRACT

While search portals` 'Portal News' account for the largest portion of aggregated news outlet, its neutrality as an outlet is questionable. This is because news aggregation may lead to prejudiced information consumption by recommending biased news articles. In this paper we introduce a new method of measuring political bias of news articles by using deep learning. It can provide its readers with insights on critical thinking. For this method, we build the dataset for deep learning by analyzing articles' bias from keywords, sourced from the National Assembly proceedings, and assigning bias to said keywords. Based on these data, news article bias is calculated by applying deep learning with a combination of Convolution Neural Network and Recurrent Neural Network. Using this method, 95.6% of sentences are correctly distinguished as either conservative or progressive-biased; on the entire article, the accuracy is 46.0%. This enables analyzing any articles' bias between conservative and progressive unlike previous methods that were limited on article subjects.

키워드 : 국회 회의록, 뉴스 기사, 딥러닝, 키워드 추출, 편향성

Keywords : National Assembly proceedings, News article, Deep learning, Keyword extraction, Bias

Received 13 May 2020, **Revised** 20 May 2020, **Accepted** 3 June 2020

*** Corresponding Author** Seungbin Oh(E-mail:sboh1214@gmail.com, Tel:+82-32-890-6700)
Student, Incheon Academy of Science and Arts, Incheon, 22009 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.8.999 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

뉴스는 가장 대중적이고 파급력이 강력한 매스미디어이며, 여론 형성에 큰 영향을 주기 때문에 기사는 중립적인 입장에서 작성하여야 한다. 하지만 최근 몇 년 동안 뉴스 기사의 조회 수와 수익을 늘리고자 왜곡된 사실을 전하거나 자극적인 내용으로 구성된 경우가 많아지고 있다. 수많은 뉴스 중 제목만 보고 클릭하게 되는 포털의 특성상, 이런 현상은 편향된 정보만 소비하게 되는 '필터 버블'을 일으킨다. 포털 이용자의 필터 버블은 한국 언론이 여론을 형성하는 역할로서 권력자와 맺어진 유대관계에 매우 종속적인 성격을 보인다는[1] 성질과 맞물려 더욱 심각해진다. 과거에도 '어용언론'과 같이 언론 프레임을 이용한 선동 의혹이나 정치적 중립성의 위배와 같은 수많은 논란의 불씨가 되었다. 포털 뉴스 공급자들도 이러한 정치적 논란에서 벗어날 수 없었다. 뉴스 배열에서 헤드라인에 배치된 뉴스를 조작하거나, 추천 알고리즘을 통한 정보 편식 유도로 언론 조작이 가능하기 때문이다. 하지만 이러한 문제는 비판적 읽기 과정을 통해 방지할 수 있으며, 기사의 편향성 인지는 기사를 비판적으로 바라볼 수 있게 한다.

뉴스 기사의 편향성 분석은 크게 코드북과 같이 미리 정해져 있는 규칙에 따라 분석하는 '내용 분석' 기법과 기사에서 무슨 정보가 어떻게 제공되는지를 분석하는 '프레임 분석' 기법으로 나뉜다. 해외의 경우, 함보르그가 제시한 자동화 방법[2] 중 가장 유력한 방법중 하나인 취재원 선택의 분석을 주로 이용하나 이 방법은 대부분의 취재원이 익명으로 처리되는 우리나라 언론의 특성상 제대로 된 효과를 발휘하기 어렵다고 판단되었다. 이외에도 감정사전을 이용하거나[3] 단어 선택 횟수를 분석하여 제작한 방법도 존재하지만, 이런 방법은 기술적 한계로 다양한 기사의 편향성을 자동으로 분석하는 데 어려움이 존재한다. 이런 이유로 대한민국의 언론 상황에 맞는 언론의 편향도 자동 측정 기법은 현재 존재하지 않는다는 것을 알 수 있다. '인터넷 뉴스 소비 습관의 시각적 확인을 통한 편향 완화 연구'(황구현, 2017)를 비롯한 기존의 연구에서 편향을 측정하기 위해서는 언론사나 뉴스의 주제를 기반으로 한 기초적 편향만 분석할 수 밖에 없었다.

하지만 본 논문에서 제안하는 기법은 여러 정치적 주제에 대한 편향성 측정에 대한 자동화를 가능하게 한다. 본 기법을 활용하기 위하여 이 연구에서는 국회 회의록에서 키워드를 추출하고, 이념지도에 기반하여 국회의원의 발언에 편향도를 부여하여 키워드별 편향도를 산출하였다. 이후 기사에서 키워드를 추출하여 산출한 편향도를 기반으로 키워드 분석을 시행하였다. 본 기법은 이렇게 얻은 Labeled Data를 통하여 인공신경망을 훈련시켜 대부분의 기사에서 정치적 보수-진보 편향성을 산출한다.

Ⅱ. 정치적 편향성의 분석

2.1. 데이터 수집

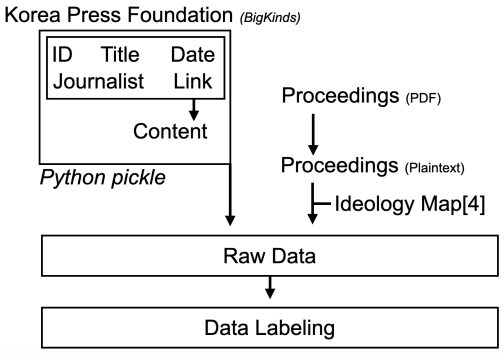

Fig. 1 Creation of Labeled Data

그림 1은 뉴스 기사와 국회회의록의 수집 및 가공 과정을 순서도로 나타낸 것이다.

키워드 추출을 위하여 이 연구에서는 국회 누리집에서 이용 가능한 20대 국회 회의록을 다운로드 받은 뒤, 이 중 상임위원회, 특별위원회 및 본회의록을 Apache Tika를 이용해 텍스트 형태로 추출하였다. 이후 정규표현식을 이용하여 회의 개의 이후부터 산회 선포까지의 회의 기록만을 분리하였다. 한 위원회 회의가 여러 개의 회의로 나누어진 경우에는 회의를 중지하는 부분까지의 기록만을 분리하였다.

국회의원의 편향도는 조선일보가 서울대 한규섭 교수(폴랩)에 의뢰하여 발표한 '20대 국회의원 이념지도'[4]를 기반으로 계산하였다. 본 편향도는 W-NOMINATE 통계기법을 이용하여 17대 국회부터 20대 국회까지 총 9478건에 대한 의원들의 표결 행태를 분석하여 제작되었다.

분석을 위해 한국신문협회 중앙지 회원사 중 한국언론진흥재단(BigKinds)에서 제공하는 기사 URL 정보가 정확하지 않아 수집할 수 없었던 매일경제, 머니투데이, 한국일보를 제외하고 국제신문과 매일신문을 추가한 24개 신문사를 표본 프레임으로 사용하였다. 표본추출 기간은 최근 정치 활동을 분석할 수 있는 기간인 2018년 1월 1일~2019년 11월 1일까지 679일로 하여 기사를 수집하였다. 분석 자료의 수집은 다음과 같은 3단계로 나누어진다.

한국언론진흥재단에서 키워드 기반 검색자료에서 원문기사의 링크, 언론사 및 게재 날짜를 확보한다. 이후 Requests Library를 바탕으로 데이터 수집기(Data scraper)를 직접 제작하여 데이터를 수집하였다. 이를 통하여 HTML 형식의 온라인 개재물을 수집하였다. 마지막으로 Modest 기반의 Selectolax라는 CSS selector를 이용하여 HTML 파일에서 기사를 원문에 가깝게 추출하여 결과를 Python pickle 형태로 저장하였다. 수집된 데이터 중 구분자(ID), 언론사명(Press), 기사제목(Title), 본문기사내용(Content), 게재일시(Date) 및 기자명(Journalist)만 Python의 Pickle 형식으로 저장하였다. 본문기사내용 이외의 정보는 한국언론진흥재단에서 제공된 정보를 사용하였다.

2.2. 키워드 편향도 측정

뉴스에서 편향도는 어떤 사실을 바라보는 관점의 차이이며 이는 논조의 차이를 통해 드러난다[5]. 그리하여 빈도수의 관점에서 정파적으로 차이가 나는 표현을 식별하고 통계적인 방법을 통해 정치성향을 할당한다. 이에 따라 각 표현이 정치 성향 점수를 갖게 되며 이를 뉴스가 얼마나 사용하느냐에 따라 정치성향을 추정할 수 있게 된다[6]. 이러한 연구에 기반하여 본 연구에서는 국회회의록에서 추출한 키워드에 발언자(국회의원)의 편향성에 기반하여 편향도를 부여한 후, 키워드를 뉴스에서 찾아 기사 및 문장별로 편향도를 계산하여 이를 머신러닝을 위한 Labeled Data로 사용하고자 하였다. 본 연구에서의 정치적 성향은 '20대 국회의원 이념지도'[4]에 기반한다. 이념지도는 의원들의 편향도를 정치 성향을 언급할 때 주로 사용되는 진보, 보수의 정도에 따라 -50~50 사이의 숫자로 나타내고 있으며, 본 논문에서는 진보를 좌편향, 보수를 우편향으로 표현한다.

$$\begin{cases} \tilde{f}_{pc} = \alpha_p + \beta_p y_c + \epsilon_{pc} \\ \widetilde{f_{pc}} = \dfrac{f_{pc}}{\sum_p f_{pc}} \end{cases} \quad (1)$$

수식 (1)의 p는 키워드, c는 국회의원을 가리키는 첨자이다. f_{pc}는 국회의원 c가 키워드 p를 사용한 횟수이며, 이에 따라 $\tilde{f}_{pc}$는 국회의원 c가 사용한 키워드 전체 사용 횟수 중에서 키워드 p가 차지하는 비율이다. α_p와 β_p는 선형회귀를 통해 키워드별로 추정되는 파라미터이고 의원의 소속과 키워드 사용비율의 관계를 나타낸다. 키워드 편향성의 측정방식을 참고한 논문 [6]에서는 y_c를 국회의원의 여야에 따른 더미변수(여당=1, 야당=0)를 사용하였지만 본 연구에서 국회의원의 이념지도[4]를 통해 국회의원의 편향성 데이터를 갖췄으므로 더미변수 대신 의원들의 편향도를 0~1사이의 실수로 매핑한 값을 사용한다. 가장 좌편향적인 의원이 0의 값을, 가장 우편향적인 의원이 1의 값을 가진다. 키워드의 편향도를 나타내는 값은 β_p로, 이 값이 양수이면 키워드 p가 우편향, 0이면 중도, 음수이면 좌편향이며 그 절대값이 클수록 더욱 편향적임을 나타낸다.

그리하여 2018년 1월부터 2019년 10월까지의 500개의 국회본회의, 상임위원회, 특별위원회의 회의록으로부터 형태소 분석기 Khaiii를 이용해 키워드를 추출한 결과 총 712개가 추출되었다. 키워드 추출 기준은 논문[6]의 부록을 참조하였다. 이 중 선형회귀의 정확도를 위해 5회 이상 언급된 144개의 키워드만 분석에 사용하였다. 그리하여 아래의 표1에 144개의 키워드 중 10회 이상 언급된 키워드 일부의 언급된 횟수와 편향도를 나열하였다.

Table. 1 Part of extracted keywords which were mentioned more than 10 times in National Assembly proceedings

Keyword	Count	Bias ($\beta_p \times 10^4$)
Honourable Citizens	321	-7
Government Inquiry	82	-77
Minju Pyeonghwa Party	28	225
Proper Country	31	37
National Assembly Vice Chairman	46	-18
Quality Jobs	20	-238
Fast Track	30	-67

Keyword	Count	Bias ($\beta_p \times 10^4$)
Working National Assembly	14	31
Minister of SMEs and Startups	10	488
ASF	16	153
Blacklist	14	-102
Ballistic Missile	15	-261
Internet Primary Bank	12	-346
Ratification	12	-10
Control Tower	14	-212
Minister of Labor	11	-26
Bareun Mirae Party	17	43
Guideline	13	62

2.3. 뉴스 편향도 측정

수식 (1)을 통해 구한 키워드 별 파라미터를 사용하여 뉴스 및 문장의 편향도를 추정한다. 즉, $\tilde{y}_n$이 뉴스 혹은 문장 n의 편향도라고 할 때, 다음의 식을 만족하는 $\tilde{y}_n$을 찾고자 한다.

$$\tilde{y}_n = \arg\min_{y_n} \sum_p (\widetilde{f_{pn}} - \widehat{\alpha_p} - \widehat{\beta_p} y_n)^2 \quad (2)$$

$$\tilde{y}_n = \frac{\sum_p \widehat{\beta_p}(\tilde{f}_{pn} - \widehat{\alpha_p})}{\sum_p \widehat{\beta_p^2}} \quad (3)$$

$\tilde{y}_n$은 뉴스 혹은 문장 n의 최종적인 편향도로, 수식 (2)와 같이 뉴스 혹은 문장에 포함된 모든 키워드 p에 대해 식(1)의 편향도 측정의 선형회귀 식과 같은 꼴을 가지는 $(\widetilde{f_{pn}} - \widehat{\alpha_p} - \widehat{\beta_p} y_n)^2$의 합이 최소가 되도록 하는 y_n의 값이다. 이러한 $\tilde{y}_n$은 수식(3)에서 문장 혹은 뉴스 내의 모든 키워드 p의 사용 비율 $\widetilde{f_{pn}}$과 수식 (1)을 통해 구한 α_p, β_p 를 통해 도출된다. 수식(1)의 β_p와 같이 $\tilde{y}_n$의 값이 양수이면 뉴스 혹은 문장 n이 우편향, 0이면 중도, 음수이면 좌편향이며 그 절대값이 클수록 더욱 편향적임을 나타낸다. 위의 수식들을 이용하여 한국언론진흥재단에서 추출한 뉴스와 뉴스에 포함된 각각의 문장의 편향도를 계산하여 Labeled Data를 제작하여 딥러닝 모델 학습에 사용한다.

Ⅲ. 딥러닝 모델의 구성

3.1. 머신러닝

3.1.1. 합성곱 신경망

완전연결 구조의 신경망은 가중치가 너무 많아 복잡도가 높다. 따라서 학습이 매우 더디고 과잉적합에 빠질 가능성도 크다. 합성곱 신경망(CNN, convolution neural network)은 부분연결 구조를 가지고 있기 때문에 모델의 복잡도를 획기적으로 낮춘다. 이미지 인식, 신호처리, 영상처리에서 주로 사용된다. 합성곱 신경망은 주로 입력층(input layer), 합성곱층(convolutional layer), 풀링층(pooling layer), 전결합층(fully connected layer), 출력층(output layer)으로 이루어져 있다. 이 구조는 인간의 시각 뉴런과도 비슷한 구조를 갖추고 있다. 이는 영상과 같은 행렬 구조 또는 3차원 이상의 텐서 구조까지 처리할 수 있다.

3.1.2. 순환 신경망

시간성(temporal property)을 가지는 데이터를 순차 데이터(sequential data)라고 한다. 순차 데이터는 시간에 따라 내용이 변하므로 동적이며 길이가 가변적이라는 특징이 있다. 순환 신경망(RNN, recurrent neural network)은 이러한 시간성 데이터를 효과적으로 처리하는 학습 모델이다. 순환 신경망은 시간성, 가변 길이, 문맥 의존성을 가진다. 순환 신경망은 입력층, 은닉층, 출력층이 있는데, MLP와 달리 은닉 노드 사이의 '순환 에지'가 존재한다는 것이 특징이다. 이 순환 에지를 가짐으로서 시간성, 가변 길이 및 문맥 의존성이라는 세 가지 기능을 모두 갖춰 t-1 순간에 발생한 정보를 t 순간으로 전달하게 된다.

3.1.3. 케라스

케라스(Keras)는 Python으로 작성된 고수준 신경망 프레임워크로 최근 Tensorflow의 공식 고수준 API로 지정되었다. 빠른 실험에 특히 중점을 두고 있어 아이디어를 결과물로 최대한 빠르게 구현할 수 있는 프레임워크이다. 컨볼루션 신경망, 순환 신경망, 그리고 둘의 조합까지 모두 지원된다. Keras는 단순하고 자명한 API를 제공하여 사용자 친화적인 이용 경험을 제공하며 모듈화되어 있어 새로운 머신 러닝 모델 제작에 용이하다.

3.2. 머신러닝 모델

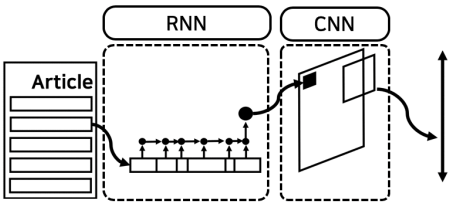

Fig. 2 Model for learning article bias

위에서 기사별, 문장별로 계산한 Labeled Data를 활용하여 뉴스 기사의 편향성을 객관적으로 측정할 수 있는 머신러닝 모델을 제작했다. 기사는 제목과 문장으로 이루어진 여러 문단으로 구성되어 있다. 따라서 본 연구에서는 그림 2와 같이 순환 신경망과 합성곱 신경망을 융합한 모델을 사용하였다. 순환 신경망은 각각의 유닛이 순환적으로 연결된 특징을 가지고 있어 시변적 동적 특징을 모델링할 수 있다. 따라서 케라스의 단어 수준 원-핫 인코딩을 이용하여 기사의 문장을 전처리하고, 순환 신경망을 이용하여 문장별 편향도 값을 학습한다. 이후 시각적 이미지를 분석하는 데에 사용되는 합성곱 신경망을 이용하여 신문의 제목과 각 문장의 편향도를 시각적으로 배치한 후 뉴스 기사의 편향도를 학습한다.

문장의 편향도를 분석하는 모델에는 Bidirectional LSTM Cell을 100개의 Cell로 1층 사용한 뒤 Fully Connected Layer를 통해 1개의 값을 도출하였다. 기사의 편향도를 분석하는 모델에는 Filter Size가 2, Kernel Size가 2×2인 2차원 합성곱층과 Pooling Size가 2×2인 2차원 풀링층을 2번 적용한 뒤 전결합층을 통해 1개의 편향도 전체 값을 도출하였다. 과적합을 방지하기 위해 모든 노드에 20%의 Dropout을 적용하였다. 손실 함수는 제곱평균 값을 사용하였으며, 최적화 알고리즘은 learning rate를 0.0001로 조정한 Adam Optimizer를 사용하였다. 또한 전체 학습 데이터의 20%를 검증 데이터로 분리하였다.

Ⅳ. 결과 및 분석

4.1. Labeled Data에서의 편향도 분포

뉴스 기사에서 키워드를 찾아 뉴스 기사와 문장의 편향도를 분석한 후 편향도들의 분포에 관련된 통계 자료를 제작하였다.

Table. 2 Statistics on news article bias

Article count	482763
Percentage of articles with zero bias	60.100%
Percentage of articles with bias larger than 300	1.300%
Average bias	-7.216
Standard deviation of bias	81.071

Table. 3 Statistics on sentence bias

Sentence count	11702126
Percentage of sentences with zero bias	96.039%
Percentage of sentences with bias larger than 300	0.180%
Average bias	-4.720
Standard deviation of bias	31.607

표2는 뉴스 기사 편향도 분포에 관한 표, 표3은 문장 편향도 분포에 관한 표이다. 편향도가 0인 문장이 기사에서는 60.1%, 문장에서는 96.039%로 큰 비율을 차지했다. 키워드가 144개 정도밖에 되지 않아 키워드를 포함하지 않는 기사와 문장이 많아 중도인 기사와 문장이 많은 것으로 추정된다. 편향도 평균값의 경우 문장과 기사에서 각각 -7.216, -4.720으로 모두 약간의 좌편향적인 수치가 도출되었다. 표준편차는 각각 81.071, 31.603로 크지 않다. 대부분 0과 가까이 분포해있기 때문에 크지 않은 것으로 추정된다. 또한 Matplotlib을 통해 기사와 문장의 편향도의 분포를 시각화하여 아래와 같은 그래프를 얻었다.

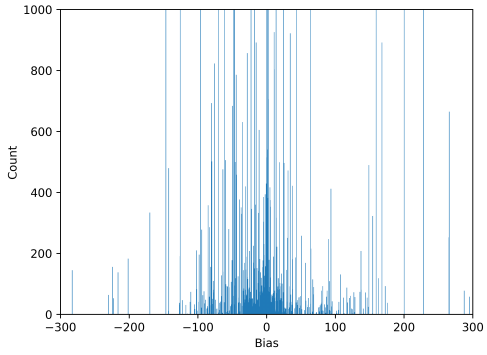

Fig. 3 Article Bias Distribution of Labeled Data

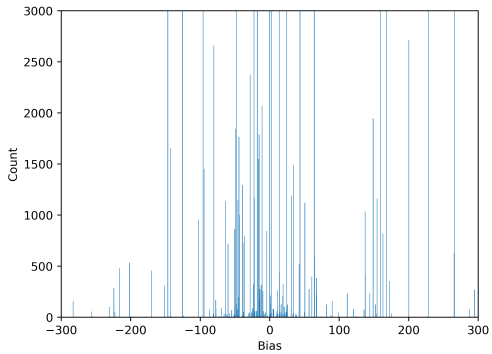

Fig. 4 Sentence Bias Distribution of Labeled Data

그림3은 뉴스 기사 편향도 분포에 관한 그림, 그림4는 문장 편향도 분포에 관한 그림이다. 뉴스 기사와 문장에서 편향도가 수천의 값을 가지는 경우가 있었으나 이는 기사 혹은 문장 내부에 키워드가 너무 적게 포함되어 있어 발생하는 것으로 추정된다. 표2, 표3에서 볼 수 있듯이 기사에서 1.3%, 문장에서 0.18%을 제외하면 모두 -300에서 300 사이의 편향도를 가진다. 그리하여 -300~300을 벗어나는 값은 각각 -300과 300으로 처리하였다. 마지막으로 머신러닝에 데이터를 전달하기 전에 편향도 값을 -300~300에서 -2~2로 매핑하였다.

4.2. 머신러닝 학습 결과

최종적으로 위에서 제작된 Labeled Data를 통해 머신러닝을 수행한 결과 손실 함수에 따라 매우 다른 결과를 얻었다. 문장 단위 좌/우편향 여부 정확도 시험(CNN Binary Accuracy Test)은 95.6%, 기사 단위 좌편향/우편향 여부 정확도 시험(RNN Binary Accuracy Test)는 46.0%를 기록하였다. 편향도 오차의 RMS값은 각각 0.054, 0.4651이다. Labeled Data가 처음부터 0에 집중되어있어서 좌/우 정확도는 낮았다. 그러나 RMS값 자체는 크지 않은 것으로 보아 상당히 정확한 예측을 했음을 알 수 있다.

Ⅴ. 결론 및 제언

본 연구는 머신러닝을 통해 객관적으로 뉴스의 편향도를 분석하는 것을 목표로 삼았다. 이를 위하여 웹에서 뉴스와 국회 회의록을 수집하여 기본 데이터를 구성하였다. 이후 국회 회의록에서 추출한 키워드에 편향도를 부여해 이를 기반으로 각 뉴스와 문장마다 편향도를 측정하였고, 이를 머신러닝을 위한 Labeled Data로 사용하였다. 최종적으로 위의 데이터를 활용하여 RNN와 CNN을 융합한 머신러닝 모델을 통해 데이터를 학습하였다.

기존에는 기사의 편향성을 분석하기 위해서 코더를 이용하여 한 주제에 대해 분석한 뒤 교차검증을 진행하여야 하였다. 그러나 본 방식은 국회회의록에서 키워드를 추출하고, 뉴스와 문장을 분석하여 Labeled Data를 지속적으로 제작할 수 있다. 또한 본 측정 방식을 참고한 연구와의 차별점으로, 키워딩 과정에서 이념지도를 추가적으로 사용하여 측정의 정확성을 높였다. 이렇게 추가된 데이터를 학습하여 계속 변화하는 사회적 이슈에 대응할 수 있으므로 본래 의도인 비판적 기사 소비 유도에 적합할 것으로 보인다. 이와 더불어 발언자 편향도의 기준 자료와 토론, 연설 등의 기초 자료를 구성할 수 있다면 정치적 편향성 외에도 탈원전-친원전과 같은 다른 분야에서도 적용할 수 있을 것으로 기대한다.

본 연구의 한계점은 다음과 같다. 첫째, Labeled Data 제작 과정에서의 편향도 분석 방식을 보완할 필요가 있다. Data Labeling에서 키워드를 추출할 때 복합명사와 감성어구 키워드만 추출하므로 편향성이 담겨있을 만한 서술어나 표현은 놓칠 수 있다. 또한, 발언자의 편향성에만 기반한 키워드 편향도 분석법도 더욱 개선되어야 할 필요가 있다. 예를 들어, ‘패스트트랙’의 경우 좌편향 키워드로 나타났지만 실제로는 우편향 뉴스에서도 공격의 목적으로도 사용될 수 있는 단어이다. 둘째, 기사 원문을 최근의 알고리즘으로 임베딩할 필요가 있다. 머신러닝 모델 부분에서는 Keras의 기본 단어 임베딩 모델을 이용하여 전처리를 시행하였다. 이는 단어 수준의 원-핫 인코딩이므로 단어 사이의 유사도를 표현하지 못하는 한계가 있다.[7] 따라서 문장 기반의 BERT나 ELMo 모델을 사용한다면 더욱 성능을 개선할 수 있을 것이다. [8]

ACKNOWLEDGEMENT

This research was results of a study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

We would like to thank our advisor Prof. Jong-Seok Lee for helping our research by guiding us, a first-timer in research, to the right direction.

REFERENCES

[1] S. J. Kim, and Y. G. Jeong, "A Study on the Characteristics of Political Bias of Korean Press : Focused on the Analysis of 19th Presidential Election Coverage," *Korean Journal of Communication & Information*, vol. 88, pp.110-145, 2018.

[2] T. Hamborg, K. Donnay, and B. Gipp, Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, vol. 20, no. 4, 391-415. 2019.

[3] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, Sentiment analysis in the news. arXiv Prepr. arXiv1309.6202 (2013).

[4] Chosun News. [20th National Assembly Ideology Map] Comparing the 17~20th National Assembly Ideology [Internet]. Available: https://news.chosun.com/site/data/html_dir/2018/01/08/20180108010 43.html.

[5] M. Gentzkow, and J. M. Shapiro, "What drives media slant? Evidence from U.S. daily newspapers," *Econometrica*, vol. 78, no.1, pp. 35-71, 2010.

[6] D. W. Choi, "Internet Portal Competition and Economic Incentives to Tailor News Slant," *The Korean Journal of Industrial Organization*, vol. 25, no. 2, pp. 1-40, Jun. 2017.

[7] Tensorflow. tf.keras.preprocessing.text.Tokenizer[Internet]. Available:https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer.

[8] S. Y. Hong, S. H. Na, J. H. Shin, and Y. K. Kim, "BERT and ELMo for contextualized word embeddings in Korean Dependency Parsing," *The Korean Institute of Information Scientists and Engineers*, 2019.6, 491-493(3 pages).

오승빈(Seungbin Oh)

2018년 ~ 현재 : 인천과학예술영재학교 재학중
※관심분야 : PC 및 모바일 어플리케이션 개발, 머신러닝(본문과 같이)

김현민(Hyunmin Kim)

2018년 ~ 현재 : 인천과학예술영재학교 재학중
※관심분야 : 컴퓨터 하드웨어 및 반도체, 머신러닝(본문과 같이)

김승재(Seungjae Kim)

2018년 ~ 현재 : 인천과학예술영재학교 재학중
※관심분야 : 정보보안, 컴퓨터 그래픽스, 머신러닝(본문과 같이)