JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 12: 1800~1808, Dec. 2022

# 소셜 미디어(SNS) 데이터 증강을 활용한 효과적인 여론조사 예측 모델 분석

황선익[1] · 오하영[2*]

# Analyzing Effective Poll Prediction Model Using Social Media (SNS) Data Augmentation

**Sunik Hwang[1] · Hayoung Oh[2*]**

[1]Graduate Student, Datascience, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associate Professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

## 요 약

선거기간이 되면 많은 여론조사 기관에서 후보자별 지지율을 조사하여 배포한다. 과거에는 여론조사 기관에 의존하여 지지율을 조사할 수밖에 없었지만, 현대 사회에서는 인터넷이나 모바일 SNS나 커뮤니티를 통해 국민 여론이 표출된다. 따라서 인터넷상에 표출된 국민 여론을 자연어 분석을 통해서 파악하면 여론조사 결과만큼 정확한 후보자 지지율을 파악할 수 있다. 따라서 본 논문은 인터넷 커뮤니티 게시글 데이터를 통해 유저들의 정치 관련 언급을 종합하여 선거기간 후보자의 지지율을 추론하는 방법을 제시한다. 게시글에서 지지율을 분석하기 위해 KoBert, KcBert, KoELECTRA모델을 활용하여 실제 여론조사와 가장 상관관계가 높은 모델 생성 방법을 제시하고자 한다.

## ABSTRACT

During the election period, many polling agencies survey and distribute the approval ratings for each candidate. In the past, public opinion was expressed through the Internet, mobile SNS, or community, although in the past, people had no choice but to survey the approval rating by relying on opinion polls. Therefore, if the public opinion expressed on the Internet is understood through natural language analysis, it is possible to determine the candidate's approval rate as accurately as the result of the opinion poll. Therefore, this paper proposes a method of inferring the approval rate of candidates during the election period by synthesizing the political comments of users through internet community posting data. In order to analyze the approval rate in the post, I would like to suggest a method for generating the model that has the highest correlation with the actual opinion poll by using the KoBert, KcBert, and KoELECTRA models.

**키워드** : 여론조사, SNS, 자연어처리, 선거 예측

Keywords : Opinion polls, social media, natural language processing, election predictions

**Received** 9 November 2022, **Revised** 25 November 2022, **Accepted** 1 December 2022

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com, Tel:+82-2-583-8585)
Associate Professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.12.1800 print ISSN: 2234-4772 online ISSN: 2288-4165

# Ⅰ. 서 론

## 1.1. 개요

본 연구는 국민들이 자주 이용하는 온라인 커뮤니티에서의 지난 20대 대통령 선거 관련 게시물과 같은 공개된 게시글을 기반으로 bert 등을 이용한 텍스트 분석기법을 이용한다. 텍스트를 긍정과 부정 감정으로 분류하여, 설문 조사 기반의 여론 조사를 더 정교하게 예측할 수 있는 비정형적 언어 기반의 긍정과 부정 감성 분석을 제안한다.

기존의 설문 기반 여론 분석 모형은 다량의 면접원을 통해서 시간과 비용을 들여가며 여론 조사를 했다. 하지만 본 연구는 이에 대한 대안으로 온라인 커뮤니티 텍스트를 분석하여 사용자의 긍정과 부정 감정을 통한 정치적 지지성향 분류 기법을 제안한다.

제20대 대통령 선거 기간 중 대선 후보와 관련된 게시글을 수집하여 긍정, 부정 감성을 기반으로 한 여론의 추이와 연관성에 대한 분석을 수행한다. 게시물 단위 문장의 긍정과 부정을 판별하기 위해 pretrain된 bert계열 모델을 이용해 downstream하여 네이버 영화 리뷰 감성 분석(NSMC)으로 fine-tuning하였다.

트랜스포머 양방향 인코더 표현(Bidirectional Encoder Representations from Transformers, BERT) 모델[1]은 대규모 데이터가 집약된 언어 모델로서 전체 데이터 문장의 단어 중 일부 단어에 마스킹을 하고 해당 단어를 추론하는 형태로 모델링을 한다. 학습 과정에서 문장의 의미, 분류 등 추론이 가능한 정보를 모델링한다고 알려져 있다[2]. 따라서 BERT는 자연어처리 문제를 해결하는 가장 대중적인 모델로서 사용되고 있다[3].

특히 BERT는 다양한 텍스트가 등록되는 인터넷 상 게시글이나 댓글의 감성 분석에도 용이하다. 게시글이나 댓글의 감정을 '긍정', '부정'으로 분류하거나 유사한 감정 체계로 나누는 데도 활용할 수 있다. 이를 통해서 특정한 사회적 이슈에 대한 인터넷 상의 여론을 확인하는 데도 활용을 할 수 있다. 인터넷 상의 여론과 실제 여론의 유사성을 확인한다면 향후 인터넷 게시글이나 댓글을 자연어처리한 결과로 사회적 이슈에 대한 여론 향방을 파악할 수 있다.

본 논문에서는 사회적 의사표현이 가장 선명하게 일어났다고 볼 수 있는 지난 2022년 대통령 선거에 게시된 인터넷 상의 글을 대상으로 BERT 등을 이용한 자연어처리를 수행하였다. 대선 후보가 언급된 게시글을 분류하고, 해당 게시글의 긍정적 평가율을 분석 모델을 이용해 계산함으로써 대선 후보별 호감도를 측정하였다.

다음으로 실제 여론조사 결과와 같은 기간 대선 후보의 긍정적 평가율을 비교하였다. 여론조사 결과와 인터넷 상 후보별 호감도의 상관관계를 분석하여 실제 여론조사를 대체할 수 있는 분석 모델을 연구하였다.

그리고 실제 여론조사 결과와 같은 기간 온라인 커뮤니티 상의 후보자에 대한 긍정과 부정 게시글의 상관관계를 더 강화할 수 있는 데이터 증강 기법을 사용하였다. 상호참조해결 모델을 이용하여 분석 모델에 따라 상관관계를 향상시킬 수 있는 모델을 확인하였다. 이를 바탕으로 대선 후보별 여론조사 결과를 상대적으로 더 효과적으로 파악할 수 있는 분석 모형을 제시한다.

# Ⅱ. 관련 기법

분석은 BERT의 KOBERT, KCBERT와 ELECTRA의 koELECTRA를 이용하였다. 현존하는 자연어처리 모델 가운데 BERT기반 모델과 BERT보다 마스킹 활용에서 자유로운 ELECTRA모델은 전이학습을 통해 모델 학습 성능이 뛰어나다고 알려져있다. 그리고 위의 3가지 모델 모두 인터넷 게시글 말뭉치를 바탕으로 모델을 생성하였다. 본 연구는 분석 대상 자료가 인터넷 게시글이므로 인터넷 게시글을 바탕으로 생성된 말뭉치가 감정 분류에 효과적일 것으로 예상하였다. 마지막으로 본 모델에서 사용할 말뭉치인 NSMC(네이버 영화평 분류 코퍼스)에서 해당 모델들은 타모델 대비 높은 분류 정확도(accuracy)를 보이므로 분석 대상 모델로 적합하다고 판단되었다.

## 2.1. BERT

KOBERT : SKTBrain에서 공개한 한국어 BERT모델로서, 한국어 위키 5백만 문장과 한국어 뉴스 2천만 문장을 학습한 모델이다. "한국어"에 대해 많은 사전 학습이 이루어져 있고, 감정을 분석할 때, 긍정과 부정만으로 분류하는 것이 아닌 다중 분류가 가능한 것이 강점이다. 인터넷 문장과 정형화된 기사 문장이 혼용돼 있어서 다양한 형태의 한국어 문장에 대응할 수 있다[4].

KCBERT : 네이버 뉴스의 댓글과 대댓글 데이터로 학습한 BERT모델이다. 댓글 데이터로 학습한 모델이어서 온라인에서의 자연어처리에 강점을 보인다. 총합

약 1억 5천만개, 메타데이터를 제외한 순수 텍스트 기준 약 15GB의 댓글 데이터셋이다[5].

토크나이저는 HuggingFace의 Tokenizers 라이브러리를 통해 BERT WordPiece 토크나이저를 학습되었으며 토크나이저의 Vocab 수는 총 3만개로 진행하였으며, 한국어의 개별 문자를 가능한 많이 포함해 Out of Vocab이 최대한 발생하지 않도록 학습된 모델이다.

### 2.2. ELECTRA

koELECTRA : BERT 이후에 등장한 언어모델로서, BERT가 가진 학습 데이터 사용의 비효율성을 극복하기 위해 탄생한 모델이다. BERT에서는 학습 과정에서 전체 입력 토큰 중 [MASK]로 가려진 15%의 토큰들만 학습에 사용하였고 이 때문에 데이터 효율성이 떨어지는 현상이 발생했다. 하지만 ELECTRA는 이를 극복하기 위해 [MASK]로 가려지지 않은 나머지 85% 토큰에 대해서도 학습을 진행하게 되는데 이 경우 BERT 대비 초기 학습 속도와 성능 면에서 우수하다고 증명되었다. 이러한 ELECTRA 모델의 특성을 대규모 한국어 말뭉치로 학습한 모델이 KoELECTRA이다[6].

### 2.3. 데이터 증강 기법

데이터 증강 기법은 적은 데이터로 예측 모델을 생성할 때 데이터를 증강시켜서 기존 모델에 비하여 더 효과적으로 결과를 예측할 수 있는 모델을 만들 수 있는 방법이다. 의미역 결정 학습 문장과 단어 치환을 수행할 어절을 마스킹한 문장을 하나로 연결하여 사전학습된 BERT 모델의 입력으로 사용한다. 이때 마스킹 할 어절의 비율(k, a)은 사용자가 선택할 수 있으며 중심어를 제외한 어절이 임의로 선택된다. 단어 치환의 결과는 마스킹 되었던 어절을 제외하고 마스크 언어 모델의 확률이 가장 높은 어절을 선택하여 출력하게 된다[7].

## Ⅲ. 대상 데이터 설명

### 3.1. 데이터 설명

데이터셋은 국내 커뮤니티 가운데 가장 이용자가 많은 디시인사이드의 '국내야구' 갤러리의 게시글을 대선 기간인 2021.11.~2022.3.까지 전체 크롤링 한 것이며 전체 데이터 수는 2,987,541개이다.

### 3.2. 데이터 전처리

전체 크롤링 데이터를 단어별로 구분하여 카운팅을 하였다. 특정 명사의 언급을 확인하기 위해 토크나이징을 하지 않고 띄어쓰기 여부로 단어를 판별하였다. 100회 이상 등장한 단어의 빈도수와 비율을 분석한 결과, 대선후보자에 대한 실명 언급 비율이 높음을 확인할 수 있었다. 전체 단어 중 '이재명'에 대한 언급 비율은 0.5%로 7번째로 많았고, '윤석열'에 대한 언급 비율은 0.49%로 9번째로 많았다.

인터넷 상의 게시글이라 비속어나 유행어가 자주 등장하였지만 특정 후보자 실명에 대한 언급빈도가 높음을 확인할 수 있었다. 후보자 별명이나 비속어가 섞인 단어보다 실명 언급 빈도가 상당히 높음을 확인할 수 있었다. 다만 전체 글에서 해당 단어가 차지하는 비율이 낮으므로 데이터 증강을 통한 데이터 확대를 고려해볼 필요가 있다.

### 3.3. 데이터 분석

여론조사는 공직선거법 제96조와 제108조로 관리되고 있다. 위의 법에 따라 선거에 관한 여론조사 결과를 공표 또는 보도하는 때에는 선거여론조사기준으로 정한 사항을 함께 공표 또는 보도하여야 한다. 이에 따라서 선거에 관한 여론조사를 실시한 기관은 조사설계서나 결과분석 자료 등을 선거 이후 6개월 동안 보관하여야 한다. 그리고 해당 자료를 중앙선거여론조사심의위원회 홈페이지에 등록해야 한다.

본 연구는 중앙선거여론조사심의위원회에 등록된 21대 대통령선거 여론조사 결과를 바탕으로 분석을 진행하였다. 다만 등록된 여론조사 기관만 90여개에 달하기 때문에 신뢰성 있는 조사기관을 대상으로 선정하였다. 매출액 기준으로 규모가 큰 한국리서치, 엠브레인퍼블릭, 리얼미터, 한국갤럽 등 4개사의 조사를 대상으로 선정하였다. 그리고 정기적으로 유사한 문항으로 진행된 조사를 선정하였으며 그 대상은 아래와 같다[8].

**Table. 1** Institutions under investigation

| polling agency | Requester | number of surveys |
|---|---|---|
| real meter | ohmynews | 19 times |
| gallup korea | self-investigation | 11 times |
| 4 companies | | 17 times |

각 여론조사 기관은 대체로 정기적으로 대통령 후보자에 대한 지지도 또는 선호도 조사를 진행하였으며 성별, 세대별 세부조사표까지 결과로 제공하고 있다. 이 같은 조사는 보통 매주 시행되며 특별한 경우 격주로 진행되기도 한다. 대통령후보자 윤곽이 드러난 2021년 11월부터 대통령선거 일주일 전 여론조사공표금지기간인 2022년 3월2일까지의 조사를 대상으로 하였다. 4사 공동의 경우 ㈜엠브레인퍼블릭, 케이스탯리서치, ㈜코리아리서치인터내셔널, ㈜한국리서치사의 공동조사이다.

본 연구에서는 인터넷 상 게시글을 통한 대통령 후보자에 대한 긍정적 게시글 정도와 실제 여론조사 결과를 비교하여 인터넷 상 게시글을 통해서 사전에 여론조사 결과를 예측할 수 있는 모델을 고안하고자 한다.

실제 여론조사 결과는 대통령 선거기간인 2021.11.~2022.2.까지 발표된 여론조사 결과를 대상으로 했으며 매출액, 인지도 등을 고려하여 정기적인 여론조사를 시행하는 한국갤럽, 리얼미터, 4사공동조사(한국리서치, 엠브레인퍼블릭, 케이스탯리서치, 코리아리서치인터내셔널 등)의 여론조사 결과를 대상으로 하였다.

## Ⅳ. 모델링

본 연구에서는 BERT를 이용한 인터넷 게시글 감정분석을 수행하였다. 3개 모델을 사용하였다. 인터넷 게시글을 바탕으로 모델링한 KcBERT, SKT에서 뉴스기사와 인터넷 게시글을 바탕으로 모델링한 KoBERT, 그리고 BERT보다 좋은 성능을 보인다고 알려진 ELECTRA계열의 KoELECTRA모델을 이용해서 분석을 진행한다.

레이블 설정을 위한 파인튜닝은 대표적인 한국어 감정분석 코퍼스인 네이버 영화평 분류 코퍼스(NSMC)를 이용하여 진행하였다.

- NSMC : 네이버 영화평에 올라온 리뷰에 대해 긍정과 부정 감정을 분류해놓은 말뭉치이다. Train데이터 15만개, Test데이터 5만개로 구성돼 있으며 본 연구에서는 Train데이터의 20%를 Validation데이터로 사용하였다. Loss를 기준으로 적합한 Epoch를 설정하였다.

### 4.1. KcBERT 모델을 이용한 긍정 감성 분석

KcBERT모델과 NSMC말뭉치를 이용하여 대통령후보자가 언급된 게시글의 감정 분석 결과는 아래와 같다. 감정은 1에 가까울수록 긍정적이며, 0에 가까울수록 부정적이다.

매주 이뤄지는 여론조사와 비교분석하기 위해 주단위로 감정분석 결과를 정리하였다.

**Table. 2** NSMC Sentiment Analysis Weekly Time Series

| Base date | Yoon | Lee |
|---|---|---|
| 2021-11-01 | 0.469605 | 0.354361 |
| 2021-11-08 | 0.404812 | 0.400125 |
| 2021-11-15 | 0.420063 | 0.405626 |
| 2021-11-22 | 0.407976 | 0.422688 |
| 2021-11-29 | 0.417716 | 0.410493 |
| 2021-12-06 | 0.404533 | 0.409725 |
| 2021-12-13 | 0.425614 | 0.411854 |
| 2021-12-20 | 0.378487 | 0.400797 |
| 2021-12-27 | 0.40776 | 0.417468 |
| 2022-01-03 | 0.420539 | 0.410544 |
| 2022-01-10 | 0.405351 | 0.40371 |
| 2022-01-17 | 0.396041 | 0.40991 |
| 2022-01-24 | 0.42698 | 0.415606 |
| 2022-01-31 | 0.426991 | 0.421421 |
| 2022-02-07 | 0.438658 | 0.395515 |
| 2022-02-14 | 0.445017 | 0.424023 |
| 2022-02-21 | 0.426025 | 0.424166 |
| 2022-02-28 | 0.410787 | 0.418204 |

이를 시각화한 그래프는 아래와 같다.

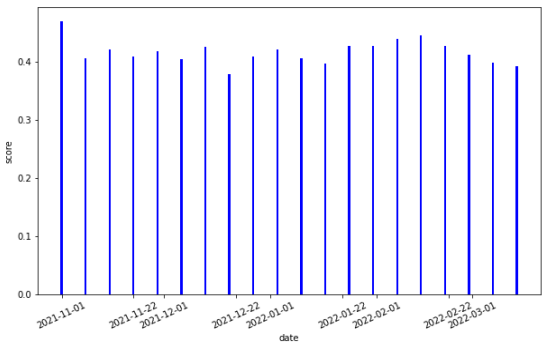


**Fig. 1** Time series of Yoon Seok-yeol's sentiment analysis result

### 4.2. KoBERT 모델을 이용한 긍정 감성 분석

KoBERT모델과 NSMC말뭉치를 이용하여 대통령후보자가 언급된 게시글의 감정 분석 결과는 아래와 같다.

감정은 1에 가까울수록 긍정적이며, 0에 가까울수록 부정적이다.

매주 이뤄지는 여론조사와 비교분석하기 위해 주단위로 감정분석 결과를 정리하였다.

**Table. 3** NSMC Sentiment Analysis Weekly Time Series

| Base date | Yoon | Lee |
|---|---|---|
| 2021-11-01 | 0.242167 | 0.341968 |
| 2021-11-08 | 0.306925 | 0.397518 |
| 2021-11-15 | 0.282097 | 0.378159 |
| 2021-11-22 | 0.242412 | 0.292748 |
| 2021-11-29 | 0.227298 | 0.28812 |
| 2021-12-06 | 0.280569 | 0.336046 |
| 2021-12-13 | 0.239668 | 0.319533 |
| 2021-12-20 | 0.252604 | 0.3163 |
| 2021-12-27 | 0.275474 | 0.330588 |
| 2022-01-03 | 0.260724 | 0.353883 |
| 2022-01-10 | 0.267228 | 0.354722 |
| 2022-01-17 | 0.305878 | 0.345505 |
| 2022-01-24 | 0.274268 | 0.308038 |
| 2022-01-31 | 0.268503 | 0.305375 |
| 2022-02-07 | 0.252508 | 0.314482 |
| 2022-02-14 | 0.270021 | 0.30809 |
| 2022-02-21 | 0.251757 | 0.29245 |
| 2022-02-28 | 0.279868 | 0.312206 |

이를 시각화한 그래프는 아래와 같다.

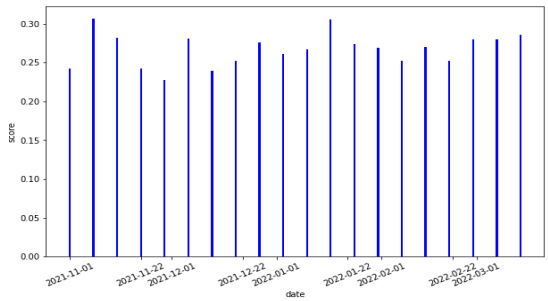


**Fig. 2** Time series of Yoon Seok-yeol's sentiment analysis result

### 4.3. KoELECTRA 모델을 이용한 긍정 감성 분류

KoELECTRA모델과 NSMC말뭉치를 이용하여 대통령후보자가 언급된 게시글의 감정 분석 결과는 아래와 같다. 감정은 1에 가까울수록 긍정적이며, 0에 가까울수록 부정적이다.

매주 이뤄지는 여론조사와 비교분석하기 위해 주단위로 감정분석 결과를 정리하였다.

**Table. 4** NSMC Sentiment Analysis Weekly Time Series

| Base date | Yoon | Lee |
|---|---|---|
| 2021-11-01 | 0.336646 | 0.265443 |
| 2021-11-08 | 0.280282 | 0.294786 |
| 2021-11-15 | 0.293777 | 0.305255 |
| 2021-11-22 | 0.302306 | 0.324918 |
| 2021-11-29 | 0.289702 | 0.324359 |
| 2021-12-06 | 0.278772 | 0.306583 |
| 2021-12-13 | 0.305698 | 0.311828 |
| 2021-12-20 | 0.277386 | 0.300496 |
| 2021-12-27 | 0.276842 | 0.316854 |
| 2022-01-03 | 0.290398 | 0.320004 |
| 2022-01-10 | 0.269805 | 0.298524 |
| 2022-01-17 | 0.264864 | 0.296518 |
| 2022-01-24 | 0.292363 | 0.316927 |
| 2022-01-31 | 0.26506 | 0.312308 |
| 2022-02-07 | 0.313289 | 0.292191 |
| 2022-02-14 | 0.294263 | 0.317963 |
| 2022-02-21 | 0.286225 | 0.322255 |
| 2022-02-28 | 0.286798 | 0.308699 |

이를 시각화한 그래프는 아래와 같다.

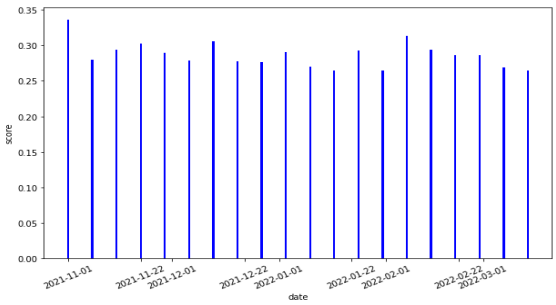


**Fig. 3** Time series of Yoon Seok-yeol's sentiment analysis result

### 4.4. 실제 여론조사와의 상관관계 평가

KcBERT모델과 NSMC말뭉치를 이용하여 대통령후보자가 언급된 게시글의 감정 분석 결과를 베이스로 하고, 데이터증강 기법을 이용한 분석결과를 비교하였다. 데이터 증강은 상호참조해결 방법을 사용하였다.

상호참조해결은 기존 의미역 결정 학습 문장과 단어 치환을 수행할 어절을 마스킹한 문장을 하나로 연결하여 사전학습된 BERT 모델의 입력으로 사용한다. 이 때 마스킹 할 어절의 비율(k, a)은 사용자가 선택할 수 있으며 중심어를 제외한 어절이 임의로 선택된다. 단어 치환의 결과는 마스킹 되었던 어절을 제외하고 마스크 언어 모델의 확률이 가장 높은 어절을 선택하여 출력하게 된다[7].

상호참조해결에서는 증강 모델을 생성할 때와 증강 문장을 생성할 때의 마스킹 비율을 각기 다르게 설정할 수 있다. 본 연구에서는 마스킹 비율을 각각 0.1과 0.15를 적용하여 비교해보았다. 온라인 커뮤니티 게시글의 긍정·부정 감정 분석 모델은 kcbert, kobert, koelectra를 사용하였다. 여론조사는 리얼미터, 갤럽, 4사공동조사를 이용하였다.

### 4.5. 데이터 증강을 통한 결과 비교

위의 실제 여론조사 결과와 KcBert, KoBert, Koelectra를 이용한 인터넷 게시글 분석 간 상관관계를 분석하였다. Baseline과 상호참조해결을 통한 데이터 증강 결과를 비교하였다.

데이터 증강은 NCMC 말뭉치를 train데이터셋만 증강시키고 test데이터셋은 증강시키지 않았다. 그 이유는 증강 데이터의 명확한 검증을 위해서다. test데이터셋까지 증강시킬 경우, 증강 데이터로 인해서 모델의 정확한 loss를 구하기 어려워질 수도 있다. 그리고 상호참조모형에서 상호참조말뭉치를 적용시키는 대신 NSMC train데이터셋 자체 텍스트 데이터로 증강을 실행하였다. 한국어로 된 상호참조해결 분야 양질의 말뭉치는 구하기 어려울뿐더러, 인터넷 게시글의 경우 비속어가 많아서 실질적 데이터 증강 효과를 기대하기 어려웠기 때문이다.

증강에 사용할 MLM모델도 크게 3개로 나눠서 적용을 하였다. KcBert, KoBert, Koelectra모델을 이용하여 각각 데이터 증강을 하였다. 데이터 증강은 우선 NCMC 말뭉치를 train데이터셋만 증강시키고 test데이터셋은 증강시키지 않았다. 그 이유는 증강 데이터의 명확한 검증을 위해서다. test데이터셋까지 증강시킬 경우, 증강 데이터로 인해서 모델의 정확한 loss를 구하기 어려워질 수도 있다. 그리고 상호참조모형에서 상호참조말뭉치를 적용시키는 대신 NSMC train데이터셋 자체 텍스트 데이터로 증강을 실행하였다. 한국어로 된 상호참조해결 분야 양질의 말뭉치는 구하기 어려울뿐더러, 인터넷 게시글의 경우 비속어가 많아서 실질적 데이터 증강 효과를 기대하기 어려웠기 때문이다.

증강에 사용할 MLM모델도 크게 3개로 나눠서 적용을 하였다. KcBert, KoBert, Koelectra모델을 이용하여 각각 데이터 증강을 하였다. 마스킹 비율은 크게 두 프로세스에서 사용된다. 첫번째가 상호참조해결 모델 생성할 때의 마스킹 비율(k)이고, 두 번째가 실제 증강 데이터 생성할 때이다. 각각 마스킹 비율(a)은 0.1과 0.15를 각각 적용하였다. 분석에 투입될 증강 데이터의 양의 비율(r, 원데이터:증강데이터 비율)은 1:1, 2:1, 3:1로 설정하였다.

분석 대상은 여론조사기관별로 대선 후보의 전체 지지율과 남성, 여성 지지율이다.

(예시: score_kcbert_증강_a01_r11_lee란, kcbert를 이용한 데이터 증강과 긍·부정 감정 분석을 진행하였으며 마스킹 비율(a)은 0.1, 증강 데이터 비율은 1:1)

KcBERT모델을 이용하여 모델 생성 시 마스킹 비율(k) 0.15, 증강 데이터 생성 시 마스킹 비율(a) 0.1, 0.15에 r은 1:1, 2:1, 3:1을 적용한 분석 결과는 다음과 같다(대표 그래프 1개 표시).

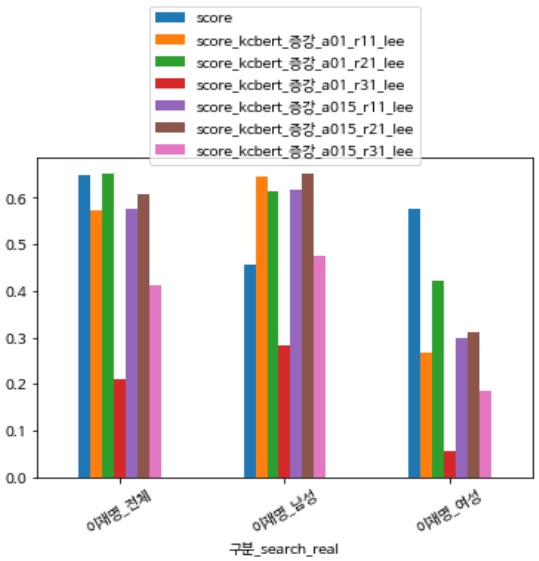


**Fig. 4** Realmeter Correlation Comparison

상호참조해결을 통한 데이터증강 시 대부분 Baseline 결과값보다 더 나빠진 상관관계를 보였으므로 실질적

인 데이터 증강효과는 없는 것으로 보인다.

KoBERT모델을 이용하여 모델 생성 시 마스킹(k) 비율 0.15, 증강 데이터 생성 시 마스킹(a) 비율 0.1, 0.15에 r은 1:1, 2:1, 3:1을 적용한 분석 결과는 다음과 같다(대표 그래프 1개 표시).

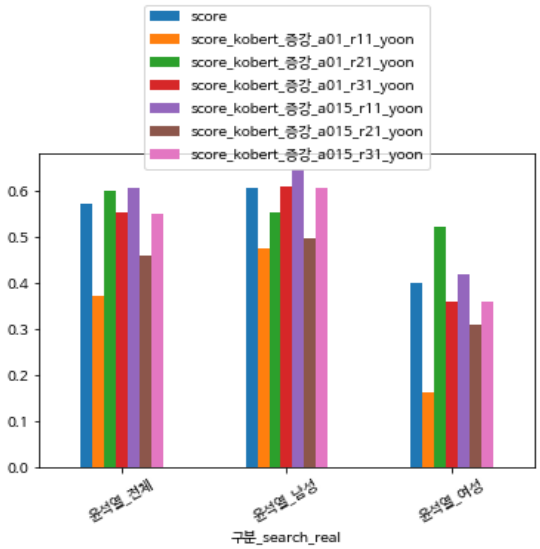


**Fig. 5** Realmeter Correlation Comparison

KoBERT모델을 이용하여 인터넷 게시글을 분석한 결과, Baseline 결과에 비하여 데이터 증강 시 실질적인 상관관계 향상 효과가 있는 것으로 나타난다. 특히 증강 데이터 생성 시 마스킹(a) 비율 0.1, 증강 데이터 비율 2:1일 때, 4사 여론조사에서 윤석열 후보 관련 조사를 제외하고 모든 조사에서 데이터 증강 시 실제 여론조사와의 상관관계가 증가하는 것으로 나타났다.

Koelectra모델을 이용하여 모델 생성 시 마스킹 비율(k) 0.15, 증강 데이터 생성 시 마스킹 비율(a) 0.1, 0.15에 r은 1:1, 2:1, 3:1을 적용한 분석 결과는 다음과 같다(대표 그래프 1개 표시).

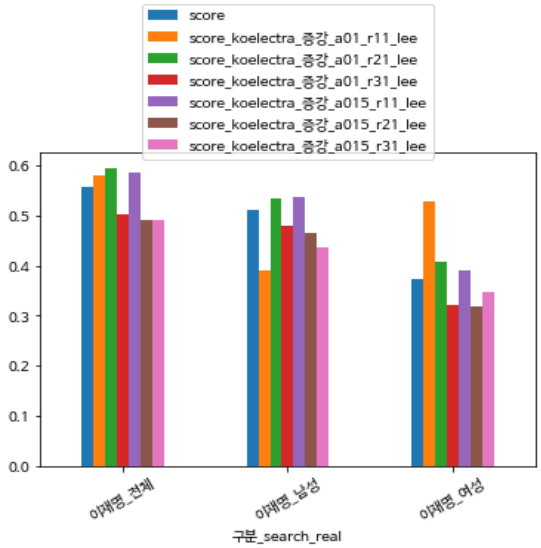


**Fig. 6** Realmeter Correlation Comparison

Koelectra모델을 이용하여 인터넷 게시글을 분석한 결과, Baseline 결과에 비하여 데이터 증강 시 실질적인 상관관계 향상 효과가 있는 것으로 나타난다. 특히 증강 데이터 생성 시 마스킹(a) 비율 0.1, 증강 데이터 비율 2:1일 때, 리얼미터 여론조사에서 윤석열 후보 관련 조사를 제외하고 모든 조사에서 데이터 증강 시 실제 여론조사와의 상관관계가 증가하는 것으로 나타났다.

### 4.6. NSMC 데이터 증강 결과

데이터 증강 시 koelectra와 kobert모델에서 전반적인 상관관계 향상이 이뤄지고 kcbert에서는 데이터 증강에 따른 상관관계 향상 효과가 거의 없는 것으로 나타났다. 아래는 데이터 증강 모델 생성 시 마스킹(k)비율을 0.15로 하였을 때 각 모델별로 데이터 생성 시 마스킹 비율(a)과 증강 데이터 비율(r)에 따른 상관관계를 정리한 표이다.

Kcbert로 증강 데이터 생성 후 온라인 커뮤니티 게시글과 여론조사 결과 비교 시, 대부분의 모델에서 증강에 따른 상관관계 향상 효과가 없었다.

**Table. 5** Comparison of Lee Jae-myung Candidate Augmentation Model

| polling agency | kcbert | | | | | | |
|---|---|---|---|---|---|---|---|
| | base line | a=0.1 r=1:1 | a=0.1 r=2:1 | a=0.1 r=3:1 | a=0.15 r=1:1 | a=0.15 r=2:1 | a=0.15 r=3:1 |
| 4 companies | 0.472 | 0.574 | 0.653 | 0.21 | 0.578 | 0.607 | 0.413 |
| real meter | 0.649 | 0.534 | 0.548 | 0.27 | 0.519 | 0.476 | 0.441 |
| Gallup Korea | 0.794 | 0.162 | 0.718 | -0.112 | 0.471 | 0.241 | 0.039 |

Kobert로 증강 데이터 생성 후 온라인 커뮤니티 게시글과 여론조사 결과 비교 시, 마스킹 비율(a) 0.1, 증강 데이터 비율(2:1)일 때 두 후보의 온라인 커뮤니티 게시글과 여론조사 결과에서의 상관관계가 6개 비교 항목(여론조사기관별, 후보별 비교) 중 5개 항목에서 증가하는 것으로 나타났다. 해당 모델로 데이터 증강을 통해 상관관계를 도출할 경우 평균적으로 9.3%정도 상관관계가 증가하는 것으로 나타난다.

**Table. 6** Comparison of Lee Jae-myung Candidate Augmentation Model

| polling agency | kobert | | | | | | |
|---|---|---|---|---|---|---|---|
| | base line | a=0.1 r=1:1 | a=0.1 r=2:1 | a=0.1 r=3:1 | a=0.15 r=1:1 | a=0.15 r=2:1 | a=0.15 r=3:1 |
| 4 companies | 0.614 | 0.665 | 0.662 | 0.629 | 0.461 | 0.769 | 0.642 |
| real meter | 0.622 | 0.371 | 0.741 | 0.55 | 0.367 | 0.6 | 0.61 |
| Gallup Korea | 0.63 | 0.375 | 0.762 | 0.809 | 0.529 | 0.94 | 0.8 |

Koelectra로 증강 데이터 생성 후 온라인 커뮤니티 게시글과 여론조사 결과 비교 시, 마스킹 비율(a) 0.1, 증강 데이터 비율(2:1)일 때 두 후보의 온라인 커뮤니티 게시글과 여론조사 결과에서의 상관관계가 6개 비교 항목(여론조사기관별, 후보별 비교) 중 5개 항목에서 증가하는 것으로 나타났다. 해당 모델로 데이터 증강을 통해 상관관계를 도출할 경우 평균적으로 12%정도 상관관계가 증가하는 것으로 나타난다. 특히 koelectra모델을 이용하여 증강을 시도할 경우, 증강을 통한 상관관계 향상이 이뤄지지 않은 1개 항목(4사 조사, 윤석열 항목)도 baseline과 비교하여 거의 동일한 수준의 상관관계를 유지하는 점을 봤을 때, 실질적으로 koelectra모델에서 마스킹 비율(a) 0.1, 증강 데이터 비율 2:1로 증강을 시도했을 경우 적어도 baseline 이상의 상관관계 향상 효과를 항상 기대할 수 있다.

**Table. 7** Comparison of Yoon Seok-Yeol Candidate Augmentation Model

| polling agency | koelectra | | | | | | |
|---|---|---|---|---|---|---|---|
| | base line | a=0.1 r=1:1 | a=0.1 r=2:1 | a=0.1 r=3:1 | a=0.15 r=1:1 | a=0.15 r=2:1 | a=0.15 r=3:1 |
| 4 companies | 0.447 | 0.371 | 0.44 | 0.469 | 0.367 | 0.471 | 0.461 |
| real meter | 0.361 | 0.488 | 0.483 | 0.449 | 0.513 | 0.384 | 0.482 |
| Gallup Korea | 0.326 | 0.433 | 0.344 | 0.33 | 0.418 | 0.227 | 0.471 |

이유를 추론해보자면, koelectra모델은 대체토큰탐지 모델을 바탕으로 한국어 뉴스, 위키 등 데이터를 기반으로 모델링이 이뤄졌고, kobert모델은 bert모델을 바탕으로 한국어 뉴스, 인터넷 댓글 등 데이터를 수집하여 모델링이 이뤄졌다. 다만 kcbert모델은 위의 두 모델과 유사하게 bert기반으로 한국어 데이터를 수집하여 모델링이 이뤄졌으나 10글자 이하인 경우 모델링에서 데이터를 제외하였다. 데이터 증강 모델에서 전체 분석 대상 데이터인 89,391개 가운데 인터넷 게시글의 특성 상 10글자 이하인 데이터가 많다. 본 연구에서 분석 대상인 인터넷 댓글의 경우 10글자 이하인 경우가 존재하며, 그 비율은 전체 게시글의 약 10% 정도인 8,739개이다. 이 같이 10글자 이하 데이터를 모델링에서 제외함으로써 데이터 증강 시 정확도가 더욱 떨어지는 현상을 보였을 것으로 추정된다.

또한 kcbert는 한국어 위키나 네이버 댓글 등에 대한 전처리를 최소화 하였는데 이 같은 부분도 영향을 미쳤을 것이라 추정된다. kcbert와는 달리 뉴스 기사 등과 같이 정형화된 인터넷 말뭉치를 일부 활용한 kobert나 koelectra는 데이터 증강 시에 모델 예측력이 더욱 향상되는 것으로 보인다.

## Ⅴ. 결 론

분석 결과 koelectra와 kobert모델에서 데이터를 증강시킬 경우 거의 대부분 조사에서 실제 여론조사와의 상관관계가 증가하였다. 특히 전체 여론조사 결과뿐만 아니라 성별 여론조사 결과와의 상관관계도 증가하였는데 전체적 데이터 증강을 이용한 예측 정확도 향상에 기여할 수 있을 것으로 판단된다.

데이터 증강 모델 생성 시 마스킹 비율(k)을 0.15로 설정하여 모델을 생성하고, 증강 데이터 생성 시에는 input문장에서 마스킹 비율(a)을 0.1로 설정할 때 전반적인 상관관계 증가 비율이 높은 것으로 나타났다. 원데이터와 증강 데이터 비율은 2:1일 때 전반적으로 상관관계 증가 효과가 가장 큰 것으로 나타났다. 다만 데이터 증강 모델 생성 시 마스킹 비율(k)이 0.1일 경우 실질적인 상관관계 증가 효과는 미미한 것으로 나타났다.

실제 여론조사를 온라인 커뮤니티 게시글을 통해서 예측할 때 데이터 증강을 이용한다면, 여론조사기관과 관계없이 상관관계 향상 효과를 기대할 수 있는 것으로 나타났다. 특히 기존의 온라인 게시글의 긍·부정 감정의 수준과 관계없이 실제 여론조사 추세를 더 정확하게 예측할 수 있는 모델이다.

이 같은 연구 결과를 바탕으로 추후 온라인 게시글과 실제 여론조사의 상관관계뿐만 아니라 실제 여론조사 수치를 예측할 수 있는 연구로 확장시켜 나간다면 한국어 자연어처리 분석의 효용성을 더욱 향상시킬 수 있는 연구가 될 것이다.


### ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT).

(No. NRF-2022R1F1A1074696)


## Reference


[ 1 ] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional tranformers for language understanding,” in *Proceedings of NAACL-HLT 2019*, Minneapolis: MN, pp. 4171-4186, 2018.

[ 2 ] H. Shin, M. Kim, Y. M. Jo, H. Jang, and A. Cattle, “Annotation Scheme for Constructing Sentiment Corpus in Korean,” in *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, Bali, Indosesia, pp. 181-190, 2012.

[ 3 ] S. A. Lee and H. P. Shin, “A Method of Infusing Additional Features into Pre-Trained BERT Models for Sentiment Analysis,” in *Proceedings of the 2020 Korea Software Symposium*, Online, pp. 275-277, 2020.

[ 4 ] SKTBrain, developed Kobert source code Guide [Internet]. Available: https://github.com/SKTBrain/KoBERT.

[ 5 ] J. B. Lee, “KcBERT: Korean comments BERT,” in *Proceedings of the 32nd Korean and Korean Information Processing Conference*, Online, pp. 437-440, 2020.

[ 6 ] J. W. Park, developed KoELECTRA source code Guide [Internet]. Available: https://github.com/monologg/KoELECTRA.

[ 7 ] J. S. Bae, C. G. Lee, J. H. Lim, and H. K. Kim. “BERT-based Data Augmentation Techniques for Korean Semantic Role Labeling,” in *Proceedings of the Korean Computer Science and Technology Conference*, Online, pp. 335-337, 2020.

[ 8 ] NamuWiki, Korean Opinion poll overview [Internet]. Available: https://namu.wiki/w/%EC%97%AC%EB%A1%A0%EC%A1%B0%EC%82%AC.



**황선익(Sunik Hwang)**

성균관대 데이터사이언스학과 수료
한국산업은행 재직
※관심분야 : 소셜미디어 자연어처리 및 데이터 분석

**오하영(Hayoung Oh)**

서울대학교 컴퓨터공학과 박사
숭실대학교 조교수
아주대학교 부교수
U.C.Berkeley 방문연구원
성균관대학교 소프트웨어융합대학 조교수
※관심분야 : 소셜정보망 분석, 추천시스템, 데이터분석 및 인공지능