ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 50, No. 12, pp. 1063-1070, 2023. 12
https://doi.org/10.5626/JOK.2023.50.12.1063

메신저 데이터 저자 프로파일링을 위한 한국어 구어체 텍스트 기반 성별 분류 모델

(Gender Classification Model Based on Colloquial Text in Korean for Author Profiling of Messenger Data)

강 지 혜 † 김 민 호 †† 권 혁 철 †††
(Jihye Kang) (Minho Kim) (Hyuk-Chul Kwon)

요 약 소셜 네트워크 서비스(SNS)를 이용한 의사소통이 폭발적으로 증가함에 따라 메신저 기능을 통해 텍스트 데이터가 방대하게 발생하고 있다. 반면 최근 자연어 처리(Natural Language Processing) 분야의 발전으로 감성 분류, 욕설 탐지, 챗봇 등 다양한 애플리케이션이 개발되어 제공되고 있으나, 한국어 구어체 텍스트에서 발화자의 성별, 연령대와 같은 저자의 다양한 특징을 분류하려는 시도는 전무한 상황이다. 본 연구에서는 한국어 구어체를 활용하여 저자 프로파일링을 위한 성별 분류 모델을 제안한다. 발화자의 성별 분류를 위해 카카오톡 대화 데이터를 기반으로, 한국어 댓글로 학습한 KcBERT(Korean Comments BERT)에 일상대화와 유사한 '네이트판(Nate Pan)' 데이터를 추가로 학습하여 Domain Adaptation을 진행한다. 그 후 어휘 외적인 정보를 결합한 모델로 실험한 결과 약 95%의 정확도를 달성하여 성능이 향상됨을 보였다. 본 연구에서는 Domain Adaptation을 위해 자체 수집한 '네이트판(Nate Pan)' 데이터 세트와 국립국어원 제공 데이터 세트를 활용하고, 모델의 학습과 평가를 위해서 AI HUB의 '한국어 SNS' 데이터 세트를 이용한다.

키워드: 자연어처리, SNS 데이터, 디지털포렌식, 프로파일링, BERT, 도메인 적응

Abstract With explosive social network services (SNS) growth, there has been an extensive generation of text data through messenger services. In addition, various applications such as Sentiment Analysis, Abusive text Detection, and Chatbot have been developed and provided due to the recent development of Natural Language Processing. However, there has not been an attempt to classify various characteristics of authors such as the gender and age of speakers in Korean colloquial texts. In this study, I propose a gender classification model for author profiling using Korean colloquial texts. Based on Kakao Talk data for the gender classification of the speaker, the Domain Adaptation is carried out by additionally learning 'Nate Pan' data to KcBERT(Korean Comments BERT) which is learned by Korean comments. Results of experimenting with a model that combines External Lexical Information showed that the performance was improved by achieving an accuracy of approximately 95%. In this study, the self-collected 'Nate Pan' data and the 'daily conversation' data provided by the National Institute of the Korean Language were used for domain adaptation, and the 'Korean SNS' data of AI HUB was used for model learning and evaluation.

Keywords: nlp, sns data, digital forensic, profiling, bert, domain adaptation

· 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1F1A1075637126218206530010 2)
· 이 논문은 2022 한국컴퓨터종합학술대회에서 '한국어 구어체 텍스트 기반의 저자 프로파일링을 위한 성별 분류 모델'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 부산대학교 정보융합공학과 학생
str_wisdom@pusan.ac.kr
†† 비 회 원 : 부산카톨릭대학교 소프트웨어학과 교수
minho@cup.ac.kr
††† 종신회원 : 부산대학교 정보컴퓨터공학부 교수(Pusan Nat'l Univ.)
hckwon@pusan.ac.kr
(Corresponding author임)

논문접수 : 2022년 11월 15일
(Received 15 November 2022)
논문수정 : 2023년 8월 31일
(Revised 31 August 2023)
심사완료 : 2023년 9월 11일
(Accepted 11 September 2023)

정보과학회논문지 제50권 제12호(2023. 12)

1. 서 론

최근 자연어처리(Natural Language Processing, NLP) 기술이 발전하고 컴퓨터나 모바일을 매개로 하여 이루어지는 SNS 대화 메신저가 비대면 의사소통 수단으로 자리 잡으면서 다양한 소셜 네트워크 서비스에서 메신저 기능을 제공하고 있으며, 또한 콘텐츠를 제공하는 서비스들도 오픈 채팅방과 같은 서비스를 제공하면서 엄청난 규모로 데이터가 축적됨에 따라 범죄자의 특성을 추론하는 수사기법으로 사용되는 '프로파일링(Profiling)'을 텍스트 데이터에 접목하여 텍스트의 저자를 분석하려는 움직임이 나타나고 있다.

반면 SNS를 통한 대화 데이터나 메신저 데이터는 사용자가 일상에서의 간편한 의사소통을 위해 한글 자모의 초성만을 이용한 표현(이하 "초성체"라 한다.)이나 줄임말, 신조어 및 이모티콘과 같은 다양한 비언어적 표현을 빈번하게 사용하는 구어(口語)체적 특성을 강하게 띤다. 따라서 대화 관련 텍스트 데이터를 활용한 기존 연구의 경우 문어 중심으로 개발되어 SNS 대화 데이터에 적용하고 연구하는 데 한계가 있다.

따라서 본 논문에서는 국내 메신저 애플리케이션 중 점유율이 가장 높은 '카카오톡'의 데이터를 분석하고, 메신저 언어의 다양한 특성 중 '한국어 구어체 텍스트 기반 성별 분류'에 유용한 정보를 분석함으로써 저자 프로파일링 시스템을 설계 및 구축하고자 한다.

'저자 프로파일링(Author profiling)'이란 문체 및 내용을 기반으로 저자의 다양한 특성을 파악하거나, 작성자를 식별하기 위해 주어진 텍스트 집합을 분석하는 것을 의미한다.

SNS 메신저 데이터는 구어체 데이터의 특성상 가변적이고, 비정형화되어 있어 기존의 연구에 따라 저자 프로파일링을 하는 데는 어려움이 있다. 하지만 현시대를 대표하는 의사소통 수단으로 자리 잡은 메신저 데이터는 다양한 의사소통 양상을 관찰할 수 있으므로 이에 맞는 분석 기법이 필요하다.

본 논문에서 제안하는 저자 프로파일링을 위한 성별 분류는 2인의 대화에서 저자의 특성을 파악하기 때문에 상대방 간 관계, 친밀도, 영향력 분석 등 다양한 과제에 적용할 수 있다.

또한 포렌식(forensic) 관점에서 볼 때, 개인정보를 도용한 후 메신저를 사용하여 범죄를 유도하는 경우에 저자 프로파일링을 적용하여 일차적인 필터링(filtering)을 통해 대화를 제한하는 방식으로 활용할 수 있다. 이러한 방식을 통해 범죄 예방에 이바지할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 국내외 텍스트 데이터를 통한 저자 프로파일링 연구에 대해 살펴본 후, 3장에서는 한국어 구어체 데이터를 기반으로 한 성별 분류 모델 실험환경을 설명한다. 다음으로 4장에서는 제안 모델과 데이터 분석을 통한 Domain Adaptation을 적용한 방법을 설명하고, 5장에서 실제 데이터에 적용하여 비교 실험하고, 마지막으로 6장에서는 결론과 향후 연구를 논한다.

2. 관련 연구

대표적인 저자 프로파일링을 위한 텍스트 기반 성별 분류 연구로 'Bots and Gender Profiling'[1]이 있다. 이는 PAN[1]의 공유 Task인 'Authorship Analysis(저자 분석)' 분야의 연구로, 영어와 스페인어로 작성된 트위터 메시지(Twitter Message)를 기반의 텍스트 저자의 챗봇 여부와 성별을 분류하는 과제이다.

영어로 작성된 트위터 메시지에서 작성자가 챗봇이 아닌 사람일 경우 성별 분류 결과 Logistic Regression과 N-gram을 사용하여 84.3%의 정확도를 달성하였다.

Janneke van de Loo 외 연구진은 텍스트 기반 연령 및 성별을 예측하는 저자 프로파일링[2]을 연구하였다. 소셜 네트워크 서비스에서 발생한 유해한 콘텐츠와 행위를 탐지하기 위해 네덜란드 채팅 게시물 약 380,000개를 대상으로 저자의 연령대를 상대적으로 분리하도록 이진 연령 분류기(binary age classifier)로 훈련했다.

이들은 연령 경계가 높아질 때 macro-averaged(매크로 평균) F-scores가 증가한다는 것을 발견했고, 저자 프로파일링이 소셜 미디어의 유해 콘텐츠 탐지에 어떻게 적용될 수 있는지 보여주었다.

국내에서는 저자 프로파일링과 유사한 연구로, 텍스트에서 문체적 특징을 추출하여 저자를 밝혀내는 저자 판별(authorship attribution)[3] 연구가 있다.

최지명[4]은 블로그의 영화 리뷰를 대상으로 텍스트에 나타나는 언어적 자질들의 빈도 분포와 기계학습(machine learning) 알고리즘을 이용하여 한국어 텍스트 저자의 자동 판별 방법론을 연구하였다.

박찬엽[5]은 페이스북 내에 성별과 연령을 프로필에 공개한 사용자의 포스트를 수집하여 6층의 컨볼루션 층과 3층의 Fully-Connected Layer으로 구성된 CNN(Convolutional Neural Networks)으로 저자 판별을 시도하였다. 50,000회를 기점으로 테스트 세트의 에러율은 약 8%를 보였다.

3. 실험환경

3.1 실험 데이터

본 논문에서 실험에 사용한 '한국어 SNS' 데이터는

1) 디지털 텍스트 포렌식 및 문체 스타일 탐지 등에 대한 행사를 진행하며 task를 공유한다.

AI HUB[6]에서 공개한 발화와 발화자 정보로 이루어진 약 200만 건의 데이터로 생산과 서비스 과정에 소비자를 참여하게 해 아이디어를 얻고 반영하는 크라우드 소싱 방식을 사용하여 수집한 대화 원문에서 개인정보 비식별화, 대화 주제 분류 및 검수 과정을 거쳐 구축되었다. 학습데이터는 데이터 누수(Data Leakage)를 방지하기 위해 학습, 검증 및 평가 데이터 세트로 분리하여 사용한다.

또한, 기존의 한국어 저자 분석을 위한 데이터는 주로 문어 중심으로 개발되었고, 개인정보 보호에 따른 수집의 어려움으로 구축에 한계가 있으므로, 학습 및 평가에 사용하는 데이터와 별개로 Domain Adaptation을 위해 추가적인 도메인 Corpus(일상대화 데이터)를 수집·구축하여 실험을 진행한다.

3.2 어휘 외적 정보 분석

성별 분류를 위해 대화 참여자 간의 관계성은 어휘 사용, 문체의 변화뿐 아니라 발화의 길이, 이모티콘 등과 같이 어휘 외적 자질의 사용 양상에 따라 차이가 발생한다고 가정한다. 가설 검증을 위해 학습에 사용하지 않는 국립국어원에서 제공하는 '모두의 말뭉치'[7]를 사용하여 검증한다.

해당 데이터는 '성별', '직업', '거주지', '발화내용' 등을 포함한 데이터로 본 연구에서 사용하는 데이터와 유사한 점이 있다.

3.2.1 비언어적 표현

본 논문에서 주목한 SNS 대화 데이터의 특징은 한 문장에서도 개행(Enter Key) 횟수가 문어체에 비해 많다는 것과, 'ㅋ', 'ㅠ'와 같은 초성체나 의성어를 빈번하게 사용한다는 것이다.

따라서 비언어적 표현인 발화당 개행 횟수 및 초성체 등을 발화자별로 추출하여 웃음 (ㅎ, ㅋ), 슬픔 (ㅜ, ㅠ), 당황 (. , ;)을 기준으로 구분하였다.

모두의 말뭉치 데이터로 실험해 본 결과, 표 1과 같이 남녀별 구어체를 사용하는 데 있어서 차이가 있음을 확인할 수 있었다. 전체적으로 남성과 비교하면 여성이 구어체를 더 많이 사용하는 경향이 있으며 문장의 길이가 더 긴 반면 남성은 여성과 비교하여 더 많은 개행을 사용함을 알 수 있다.

표 1 어휘 외적 표현 통계 정보 (모두의 말뭉치)
Table 1 Statistical information on extrapolation of vocabulary (everyone's corpus)

	Male	Female
ㅋ	0.0213	0.0374
ㅎ	0.0128	0.0128
ㅠ	0.0012	0.0025
ㅜ	0.0029	0.0058
.	0.0154	0.0175
,	0.0006	0.0008
enter	42.1425	35.6847
length	465.5234	478.5315

표 2 어휘 외적 표현 통계 정보 (학습데이터)
Table 2 Statistical information on nonvocabulary expressions (learning data)

Speaker Type		Enter	Len	Laugh	Cry	Comma
Male-Male	A	0.0916	96.2256	4.4701	1.2311	1.9625
	B	0.0919	95.9597	4.4824	1.2457	2.0089
Male-Female	A	0.0985	86.2459	5.2412	1.5319	2.5104
	B	0.0934	94.2912	6.3319	1.6898	2.7599
Female-Male	A	0.0934	94.1849	6.3386	1.6937	2.7612
	B	0.0987	86.1176	5.2335	1.5385	2.5002
Female-Female	A	0.0899	93.2990	7.6343	1.8285	3.0739
	B	0.0899	93.1968	7.6240	1.8235	3.0767

위 실험 결과를 따라서 학습에 사용하는 데이터에 적용했을 경우, 어휘 외적 표현의 통계정보는 표 2와 같으며, '여성'에 비해 '남성'의 전체 발화 길이별 개행 횟수가 많고 그 외 오타나 어휘 수준은 유의미한 차이가 나타나지 않았다.

반면 비언어적 표현 통계 정보 중 'ㅎㅎ', 'ㅋㅋ'와 같은 웃음을 뜻하는 초성체의 경우 남성에 비해 여성이 더 많이 사용함을 확인했으며, 특히 '여성-여성'의 대화에서 크게 발생했다.

3.2.2 오타 개수

Python용 띄어쓰기 및 맞춤법 교정 라이브러리 'Py-Hanspell'[8]을 활용하여 데이터 세트의 발화자별로 검사를 진행하였다. 기존 발화와 맞춤법 검사 후 발화의 차이를 이용해서 오타 개수를 단순 카운트로 도출한다. 오타 데이터 추출 예시는 다음 표 3과 같다.

대화 참여자의 성별에 따른 오타의 개수는 다음 표 4와 같다. 성별에 따른 오타 개수는 이성(남성-여성, 여성-남성)일 경우 오타의 횟수가 빈번하게 등장함을 알

표 3 오타 개수 추출 예시
Table 3 Examples of typo count extraction

ChatID	Speaker ID	Spell Check		# of Typos
		Before	After	
1	A	고마워 공주 사랑행	고마워 공주 사랑해	1
	B	오빠옆엔 내가있자나 프사 맘에 들었닝 털복숭이	오빠 옆엔 내가 있잖아 프사 맘에 들었니 털북숭이	5

표 4 오타 분석
Table 4 Typing analysis

Speaker Type		Average Number	Number by Speech Length
Male-Male	A	6.3013	0.0655
	B	6.2915	0.0656
Male-Female	A	6.3832	0.0740
	B	7.0567	0.0748
Female-Male	A	7.0455	0.0748
	B	6.3606	0.0739
Female-Female	A	6.5220	0.0699
	B	6.5147	0.0699

수 있다. 이는 이성 관계일 경우 연인 관계가 포함되어 있어 ㅇ붙힘(밥 먹었엉?), 애교(언눙) 등이 동성에 비해 빈번하게 사용하기 때문으로 추측된다.

본 연구에서 사용한 'Py-Hanspell'은 특정 도메인에 특화된 라이브러리가 아니기 때문에 교정 내용에 약간의 오류가 포함되어 있다. 이러한 오류에도 발화마다 같은 방식으로 교정했으므로, 성별 발화의 차이를 확인하기 위해 단순 카운트를 적용하였다.

3.2.3 어휘 수준

어휘 수준은 국립국어원[9] '한국어 기초사전'에서 어휘 등급(초급, 중급, 고급)을 활용하였다. 각 발화별 형태소 분석은 Bert Tokenizer의 경우 Word Piece 모델을 사용하므로 어휘 등급 사전과 비교를 할 수 없어 형태소 분석 패키지 KoNLPy(Korean NLP in Python)의 Okt 라이브러리를 사용했다. 한국어 기초사전의 어휘 등급 중 '없음'을 제외한 총 19,662개의 품사별 어휘와 각 발화의 형태소 분석 결괏값을 비교하여 도출하였다.

어휘 수준의 수치화를 위해서 단순 빈도수와 기초사전 어휘 등급에 따른 점수(1~3)를 주었다. 그 결과 남성은 이성과의 대화에서 더 높은 수준의 어휘를 사용하고, 여성은 동성과의 대화에서 더 높은 수준의 어휘를 구사함을 알 수 있었다. 또한 동성과의 대화일 경우 비슷한 수준의 어휘를 구사하는 것으로 나타났다.

4. 한국어 구어체 텍스트 기반 저자 프로파일링을 위한 성별 분류 모델

4.1 제안 모델

본 논문에서 제안하는 '한국어 구어체 텍스트 기반의 저자 프로파일링을 위한 성별 분류 모델'의 구조도는 그림 1과 같으며 데이터 정제, 예측 모델, 분류 모델로 구성되어 있다.

먼저 데이터 정제 단계에서는 초기 데이터를 대화방 및 발화자별로 결합하여 모델이 학습할 수 있도록 하며, 데이터 정제 시 '띄어쓰기 없는 표현 형태'와 같은 메신저 대화의 고유한 특성을 파괴하지 않고 분석하기 위해 최소한의 작업을 수행한다. 또한, 어휘 외적 정보를 추출하여 분석한 후 분류 모델 학습에 사용할 수 있도록 한다.

표 5 한국어 기초사전 데이터의 예
Table 5 Examples of basic Korean dictionary data

Entry	Part of Speech	Grade	Meaning
가게	Noun	Beginner	작은 규모로 물건을 펼쳐 놓고 파는 집
가공	Noun	Advanced	원료나 재료를 새로운 제품으로 만듦.
가끔	Adverb	Beginner	어쩌다가 한 번씩

표 6 어휘 수준 추출 예시
Table 6 Example of lexical level extraction

Speech	Morphological		Grade	Grade Score
	morpheme	Part of Speech		
오늘 밥 먹어?	오늘	Noun	Beginner	1
	밥	Noun	Beginner	1
	먹다	Verb	Beginner	1
	?	Punctuation	None	0

예측 모델에서는 2인 발화자의 발화 정보를 대화 쌍으로 사용하여 학습하고 각 발화자의 성별('남성-남성', '남성-여성', '여성-남성', '여성-여성')에 대해 예측한다. 그 후 Domain Adaptation을 위해 추가적으로 메신저 데이터와 유사한 도메인 Corpus를 구축하고, BERT 계열 모델 중 초기 베이스라인 성능이 가장 높은 모델을 선정하여 진행한다.

분류 모델은 예측 모델의 결과와 데이터 정제 과정에서 추출한 어휘 외적 정보 중에서 유용한 자질을 결합하여 학습에 사용하며, 이는 예측 모델의 결과를 보정하는 역할을 한다.

4.2 예측 모델 선정

본 연구의 성별 분류 모델은 구글에서 2018년에 공개한 BERT[10] 모델을 이용한다. 공개된 한국어 사전 학습 BERT 모델 중에서 구어체에 적합한 모델을 선정하기 위해 SKTBrain에서 공개한 KoBERT[11]와 KoELECTRA[12], KcBERT[13], KcELECTRA[14]에서 성별 분류 성능을 비교했으며, 이 중에서 네이버 뉴스 댓글과 대댓글로 사전 학습된 KcBERT(Korean comments BERT)가 가장 높은 성능을 보였으며 메신저 데이터와 같이 오탈자, 신조어 등의 표현이 빈번하게

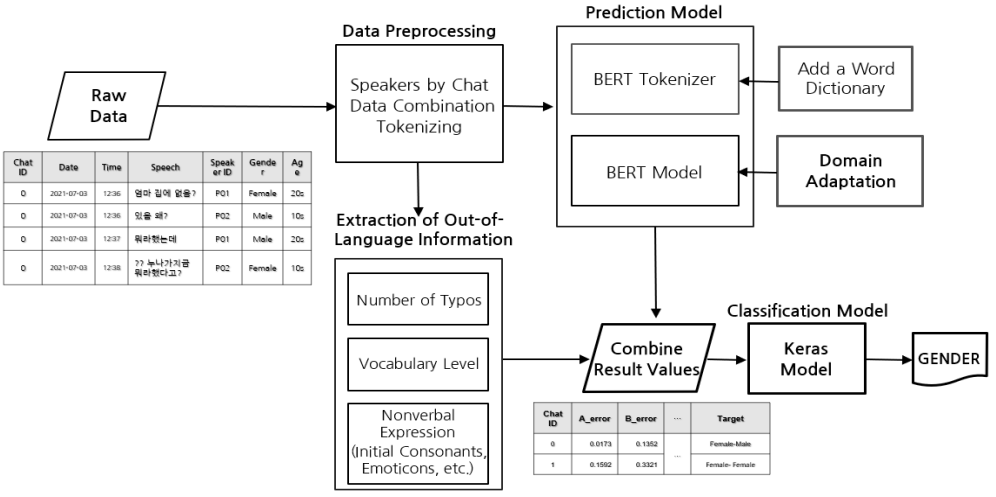

그림 1 제안 모델 구조도

Fig. 1 Schematic diagram of the proposed model

등장하는 구어체 텍스트에 대해서 가장 좋은 사전지식을 함유하고 있을 것으로 판단하여 학습 모델로 선정했다. 각 모델의 실험 결과는 다음과 같다.

반면, KcBERT 모델의 토픽이나 장르를 특정하는 언어 분포는 학습에 사용된 뉴스 관련 단어들이 주를 이루고 있으므로, 본 논문의 Task의 Domain에 맞는 언어 분포로 옮겨오는 Domain Adaptation을 위해 단어사전 구축 및 MLM(Masked Language Modeling) 학습을 추가로 진행했다.

4.3 Domain Adaptation 적용

Domain Adaptation은 실험에 사용된 사전 학습 모델의 토픽이나 장르를 특정하는 언어 분포는 학습에 사용된 뉴스 관련 단어들이 주를 이루고 있으므로, 본 논문의 Task의 Domain에 맞는 언어 분포로 옮겨오는 것으로 Domain Adaptation에 사용하기 위해 국내 포털사이트인 '네이트(Nate)'에서 운영하는 일상생활에 관한 이야기나 고민을 나누는 개방형 게시판인 '네이트 판'에서 카테고리 채널 19개를 선정하여 웹 크롤링을 통해 말뭉치 데이터 495,668개를 수집하였다. 추가로 학습 및 검증에 사용하지 않는 메신저 데이터와 비슷한 양상을 보이는 '국립국어원 일상 대화 말뭉치 2020'을 활용한다. 해당 말뭉치는 약 500시간 동안의 음성 대화를 텍스트로 전사하여 구성되었다. 데이터는 총 2,739명의 발화자의 대화 ID, 발화내용, 발화자의 성별, 직업과 같은 발화자 정보를 포함하고 있다. 또한, 국립국어원에서 최근에 공개한 '국립국어원 온라인 대화 말뭉치'와 '국립국어원 메신저 말뭉치'도 추가하여 사용한다. '국립국어원 온라인 대화 말뭉치'는 총 3,836건으로 구성되어 있으며, '국립국어원 메신저 말뭉치'는 총 74,665건의 분량으로 구성되어 있다.

표 7 어휘 수준 분석

Table 7 Vocabulary level analysis

Speaker Type		Average Vocabulary Level	Level by Speech Length
Male-Male	A	1.4634	0.0152
	B	1.4603	0.0152
Male-Female	A	1.4098	0.0163
	B	1.3925	0.0148
Female-Male	A	1.3964	0.0148
	B	1.4125	0.0164
Female-Female	A	1.4335	0.0154
	B	1.4341	0.0154

표 8 한국어 PLM 모델 성별 분류 성능

Table 8 Gender classification performance of the Korean PLM model

Model	Accuracy	Precision	Recall	F1-Score
KoBERT	80.17	74.25	86.25	79.80
KoELECTRA-Base-v3	82.89	87.47	85.54	86.49
KcBERT-base	89.20	87.90	88.43	88.16
KcELECTRA-base	87.73	81.14	92.66	86.52

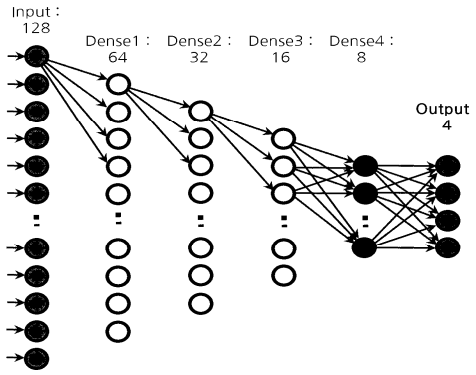

그림 2 분류 모델 구조도
Fig. 2 Classification model structure diagram

4.4 분류 모델 설계

분류 모델은 예측 모델에서 반환된 확률값에 어휘 외적 정보를 추가하여 예측모델 결과의 정확도를 보정하는 역할을 한다. 분류 모델은 Keras 패키지의 Sequential 모델이며 그림 2와 같다. 해당 모델은 방법론을 검증하기 위한 것으로 단순한 구조로 설계하여 검증하였다.

5. 실험 평가 및 결과

5.1 실험 평가 방법

5.1.1 성능평가지표

분류 모델의 성능평가는 표 9의 혼동 행렬(Confusion matrix)의 값을 각각 구하여 수식 1의 정밀도(Precision)와 수식 2의 재현율(Recall)로 계산하여 측정한다. 혼동행렬은 유형별로 정확한 예측과 잘못된 예측의 수를 한 번에 보여주는 표다. 또한, 혼동행렬에 나타나는 예측의 수들(TP, FN, FP, TN)을 조합하여 각종 지표들을 계산한다.

정밀도란 모델이 정답이라고 분류한 것 중에서 실제 정답인 것의 비율을 말하며, 재현율은 실제 정답인 것 중에서 모델이 정답이라고 예측한 것의 비율이다. F1 Score는 수식 3과 같이 정밀도와 재현율의 조화 평균이다.

$$Precision(P) = \frac{TP}{TP+FP} \quad \text{(Formula 1)}$$

$$Recall(R) = \frac{TP}{TP+FN} \quad \text{(Formula 2)}$$

$$F1Score = 2 \times \frac{P \times R}{P+R} \quad \text{(Formula 3)}$$

- TP(True Positive) : 실제 값이 Positive인 경우를 올바르게 예측
- TN(True Negative) : 실제 값이 Negative인 경우를 올바르게 예측

표 9 혼동행렬 분류 정리표
Table 9 Confusion matrix classification summary table

		Real Answer	
		True	False
Classification Results	True	True Positive	False Positive
	False	False Negative	True Negative

표 10 Domain Adaption 성능평가
Table 10 Domain adaptation performance evaluation

Model	Acc	Precision	Recall	F1
KcBERT-base	89.20	87.90	88.43	88.16
KcBERT-base + Domain Adaptation	93.71	90.36	91.18	90.93

- FP(False Positive) : 실제 값이 Negative인 경우를 Positive로 잘못 예측
- FN(False Negative) : 실제 값이 Positive인 경우를 Negative로 잘못 예측

5.2 예측 모델 실험 결과

예측 모델은 MLM(Masked Language Modeling) 추가학습을 통한 Domain Adaption을 적용한 모델로 93.71%의 정확도로 기존 KcBERT의 분류 정확도 89.2%에서 약 4%의 성능 향상을 보인다.

MLM(Masked Language Modeling) 추가학습을 통한 KcBERT 기반의 성별 예측 모델은 성별 분류의 정확도 93.71, F1 Score 90.93으로, 기존의 KcBERT-base

표 11 대화방 오분류 예시
Table 11 Example of misclassification of chat rooms

Chat ID	Speaker	Speech Content	Pred	actual
1	A	곡물과자 드실라우?ㅎ 찬장맨 위칸에 있어서~ 항상 까치발 들고 꺼내고 있엉ㅋㅋㅋㅋㅋ 웅? 진짜?!?! 웅우~~ 밀가루도 안들어가고~~	F	M
	B	오옷 네! 감사합니다 ㅋㅋ이거있는줄 몰랐웅 은근 고소하고 맛난 과자~ 오옷좋네	F	F
2	A	진짜 무슨 문제잇냐 너? 생각없이말하네 뭔 내노력이없엇다는말 언제햇냐고 하고자빠졋네 왜저래진짜 니가 말햇잖아	M	F
	B	뭐 어쩌라고 억울해서 미칠거같은데 나도 니 대답안한거 기분나빴는데 참았는데 지는 뭐	F	M

모델보다 모든 지표에서 향상되는 결과를 도출하였다. 하지만 표 11과 같이 대화방 1의 발화자 B의 성별이 '여성'임에도 '남성'으로 예측하여, 어휘 외적 정보를 결합하여 추가로 학습한 결과와 비교해보고자 한다. KcBERT의 예측 결과에 비언어적 표현과 같은 어휘 외적 정보를 결합하여 모델을 학습하고 분류하여 미세한 예측결과를 보정 해줄 수 있을 것으로 가정하고, 데이터 추출 및 실험을 설계하였다. 실제 예측 모델의 결과는 각 발화자에 대한 점수 쌍이며 이어지는 분류 모델의 입력값으로 활용된다.

5.3 분류 모델 실험 결과

본 논문에서 제안하는 성별 분류 모델은 표 12에서 KcBERT + Domain Adaptation + 분류 모델이다. MLM 추가학습을 통한 Domain Adaption 모델은 93.71%의 정확도로 기존 KcBERT의 분류 정확도 89.2%에서 약 4%의 성능 향상을 보인다.

분류 모델은 기존 대화 내용을 기반으로 출력된 예측 모델의 점수 쌍과 어휘 외적인 정보를 추출하여 결합한 데이터를 입력으로 사용하는 모델로 95.07%의 정확도로, 기존 KcBERT-base 모델보다 약 6%의 정확도 향상을 보였다.

표 13은 표 12의 대화방별 class g 확률 값으로, 대화방 1의 실제 정답인 '여성-여성'의 확률은 0.488554로 가장 높은 확률값으로 분류되었고, 대화방 2의 실제 정답인 '여성-남성'의 경우 0.762770으로 제대로 분류됨을 알 수 있다.

이로써 어휘 외적 요소를 추출 및 결합한 분류 모델 Output의 각 클래스의 확률값 분석을 통해 어휘 외적 정보를 통한 보정 효과를 확인할 수 있다.

5.4 예측 모델과 분류 모델 결합 실험 결과

'한국어 구어체 텍스트 기반 저자 프로파일링을 위한 성별 분류 모델'은 예측 모델과 분류 모델을 결합함으로써 각 발화자의 성별('남성-남성', '남성-여성', '여성-남성', '여성-여성')에 대한 예측 성능을 향상한 모델로 예측 모델의 각 성별 예측 점수를 분류 모델에 적용한다.

예측 모델은 MLM(Masked Language Modeling) 기반 Domain Adaption을 적용하여 성능을 보정했으며 예측 모델의 결과값을 분류 모델의 입력으로 활용하고 초성체와 같은 비언어적 표현에 대한 정보를 추가하여 구어체에 적합하도록 구현하였다.

기존 KcBERT-base 모델은 정확도 89.20%인데 비해 제안하는 결합모델은 95.07%로 성능이 높을수록 추가적인 성능의 향상이 어려움에도 5% 이상의 정확도 증가를 함으로 제안하는 기법의 유효성을 검증했다.

6. 결론 및 향후 과제

본 논문에서는 국내 메신저 애플리케이션 중 사용량이 가장 많은 '카카오톡' 데이터를 분석하여 한국어 구어체 텍스트의 저자 프로파일링을 위한 성별 분류 연구에 적용하였다.

이를 위해 남녀 텍스트에서의 어휘 외적인 차이가 있을 것이라고 가정하고, 외적인 정보를 분석 및 추출하고 구축하여 구어체 텍스트 기반의 저자 프로파일링을 위한 성별 분류 시스템을 설계하고자 하였다.

본 논문은 사전 학습 모델인 KcBERT에 Domain Adaptation을 적용하였고, 발화 자체에서 외적 자질 분석을 통해 파생변수를 도출하고 결합을 시도했다. 결과적으로 Domain Adaptation 및 예측 모델과 분류 모델을 결합하여 기존 대비 시스템의 성능이 향상했고, 초성체 등 남녀 간 비교적 차이를 보이는 어휘 외적 정보와 결합하여 사용함으로써 성능 향상을 보였다.

반면 실험에 사용된 데이터는 SNS 메신저 데이터로 구어체적 특성이 있으나 실제 구어체와 다르며 음성을 통한 구어체 분석에는 음성 인식 문제가 발생할 수 있다. 또한, 구어체적 특성 중 크게 영향을 미치는 사투리와 SNS에서 활발히 사용하는 이모티콘 처리가 미흡하다는 한계점이 있다.

향후 연구로는 사투리 및 이모티콘 처리를 추가하여 해당 연구방법을 연령대 분류에 확장하고, 대화 참여자 간의 친밀도 및 영향력 분석 등을 통해 메신저 텍스트 데이터 기반 저자 프로파일링 연구를 진행하고자 한다.

표 12 성별 분류 모델의 성능 비교
Table 12 Performance comparison of gender classification models

Model	Acc	Precision	Recall	F1
KcBERT-base	89.20	87.90	88.43	88.16
KcBERT-base + Domain Adaptation	93.71	90.36	91.18	90.93
Proposed Model (KcBERT-base + Domain Adaptation + Classification Model)	95.07	90.75	91.50	91.12

표 13 '표 12'대화방별 class의 확률값
Table 13 Probability values of classes by chat room in Table 12

Chat ID	Male-Male	Male-Female	Female-Male	Female-Female
1	0.0011	0.1643	0.3460	0.4886
2	0.0003	0.2099	0.7628	0.0270

References

[1] F. Rangel and P. Rosso, “Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter,” *CEUR Workshop Proceedings (CEUR-WS.org)*, Vol. 2380, pp. 1-36, 2019.

[2] J. V. D. Loo, G. D. Pauw and W. Daelemans, “Text-Based Age and Gender Prediction for Online Safety Monitoring,” *International Journal of Cyber-Security and Digital Forensics*, vol. 5, pp. 46-60, 2015.

[3] Han Na-rae, "Korean author discrimination using frequency information," *Journal of the Cognitive Science Society*, Vol. 20, No. 2, pp. 225-241, 2009.

[4] Choi Myung-ji, "Determination of Korean text authors using machine learning algorithms," a domestic master's degree thesis, Yonsei University graduate school, 2015.

[5] Park Chan-yeop, Jang In-ho, and Lee Joon-ki, "Determination of gender and age of web text authors using deep learning: Focusing on SNS users," *Journal of the Korea IT Service Association*, Vol. 15(3), pp. 147-155, 2016.

[6] AI HUB, URL:https://aihub.or.kr/aidata/30718

[7] National Institute of Korean Language (2021), National Institute of Korean Language Daily Conversation Corpus 2020 (version 1.2), URL: https://corpus.korean.go.kr

[8] Py-Hanspell, Github repository, URL:https://github.com/ssut/py-hanspell

[9] National Institute of Korean Language, URL: https://krdict.korean.go.kr/

[10] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In NAACL-HLT, 2019.

[11] Lee Joon-beom, "KcBERT: BERT Learned from Korean Comments," a collection of papers at the 32nd Korean and Korean Information Processing Conference, pp. 437-440, 2020.

[12] SKTBrain, “Korean bert pre-trained cased (kobert),” URL:https://github.com/SKTBrain/KoBERT, 2019.

[13] J. Park, “KoELECTRA: Pretrained ELECTRA Model for Korean,” Github repository, URL: https://github.com/monologg/KoELECTRA, 2020.

[14] Lee Joon-beom, “Kcelectra: Korean comments electra”, https://github.com/Beomi/KcELECTRA, 2021.

강 지 혜

2014 전남대학교 행정학과 학사 졸업
2022 부산대학교 정보융합공학과 석사 졸업. 관심분야: 자연어처리, 머신러닝, 인공지능

김 민 호

2007 부산대학교 전자전기정보컴퓨터공학부 학사 졸업. 2009 부산대학교 컴퓨터공학과 석사 졸업. 2022 부산대학교 정보융합공학과 박사 졸업. 2020~현재 부산가톨릭대학교 컴퓨터정보공학과 조교수. 관심분야: 한국어정보처리, 자연어처리, 머신러닝, 인공지능

권 혁 철

1982년 서울대학교 컴퓨터공학과 학사 졸업. 1984년 서울대학교 컴퓨터공학과 석사 졸업. 1987년 서울대학교 컴퓨터공학과 박사 졸업. 1992년~1993년 (미) Stanford 대학 CSLI 방문 교수. 1987년~현재 부산대학교 정보컴퓨터공학부, 인지과학 협동과정 교수. 2020년~현재 부산대학교 인공지능대학원 교수. 관심분야: 자연언어처리, 정보검색, 인공지능