Journal of KIIT. Vol. 22, No. 3, pp. 23-28, Mar. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.3.23

CNN 모델을 활용한 검색 빈도 데이터 기반 신조어 분류 알고리즘

김민정*[1], 김현수*[2], 유석종**

A Classification Algorithm for Newly-Coined Words based on Search Frequency Data using CNN Model

Min-Jung Kim*[1], Hyun-Soo Kim*[2], and Seok-Jong Yu**

요 약

SNS상에서 신조어의 사용이 일상화되고 있으며, 특히 사회집단별로 사용하는 신조어에도 차이가 있는 것으로 파악되고 있다. 본 연구에서는 급증하는 신조어의 출현 경향과 사례를 분석하여 신조어로 인해 발생하는 의사소통 문제를 개선하고자, 네이버 데이터랩에서 제공하는 검색 빈도 데이터를 활용하여 CNN 딥러닝 모델에 기반한 신조어 분류 알고리즘을 제안한다. 제안 알고리즘의 성능 분석을 위해 네이트판, DC인사이드, 네이버뉴스에서 크롤링한 데이터 셋에 적용한 결과, 약 82%의 신조어 분류 정확도를 확인할 수 있었다. 또한 오픈소스 라이브러리인 Streamlit을 사용하여 신조어의 출현 빈도 순위와 관련 정보를 시각화하는 웹서비스 시스템을 구현하였다.

Abstract

The use of newly coined words on social media has become commonplace, and we have observed that there are differences in the use of these words among different social groups. In this study, we analyze the trends and examples of the rapidly increasing appearance of newly coined words. In order to improve communication problems caused by newly coined words, we propose a newly coined word classification algorithm using a CNN deep learning model and search word frequency data provided by Naver Data Lab. When we applied this proposed algorithm to datasets crawled from Nate Pann, DC Inside, and Naver News, we were able to confirm an accuracy rate of approximately 82% for newly coined word classification. Furthermore, we have implemented a web application that visualizes the frequency of appearance rankings and related information of newly coined words using Streamlit, an open-source library.

Keywords

newly-coined word, deep learning, classification, search frequency, naver datalab

* 숙명여자대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcid.org/0009-0001-8670-2032
- ORCID[2]: https://orcid.org/0009-0006-1943-8960
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Dec. 12, 2023, Revised: Jan. 05, 2024, Accepted: Jan. 08, 2024
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Womens's University,
Cheongpa-ro 100, 47-gil, Cheongpa-ro, Yongsan-gu, Seoul, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

신조어는 새로 생긴 말, 또는 새로 귀화한 외래어로 정의되며, 새로운 사회적 이슈 등 특정 요인의 결과로 다양한 영역에서 나타나는 언어 트렌드 변화의 대표적 산물이다. 1년에 300~500개의 새로운 신조어가 생성되어 신문 또는 방송에 등장하고 있다[1][2]. 이와 같이 빠른 신조어의 생성 변화 현상은 세대별, 성별, 관심사별로 사회적 집단 간에 의사소통의 어려움을 가중시키고 있는 추세이다[3][4]. 예를 들어 젊은 층은 '중꺾마', '점메추' 등의 일상 신조어를 사용하는 반면, 중년 이상은 '김영란법', '노도강'등 정치, 경제 관련 신조어를 주로 사용하는 경향을 보인다. 이와 같은 계층별 신조어 분리 사용으로 인한 문제점을 개선하기 위해서는 다양한 커뮤니티에서 생성되는 최신 신조어의 출현 경향 및 사용 행태를 분석하여 제공하는 서비스가 요구된다. 그림 1은 젊은 층에서 사용하는 최신 신조어들의 예이다.

그림 1. 신조어 모음의 예
Fig. 1. Example of newly-coined words

네이버 데이터랩[5]은 네이버에서 검색되는 단어의 검색 빈도 데이터를 제공하는 서비스이다. 네이버 데이터랩에서 신조어(Newly-coined word)와 일반어(General word)의 검색 빈도 데이터를 조회해보면 그래프 패턴에 분명한 차이점을 발견할 수 있다. 본 연구에서는 신조어와 일반어 단어들의 검색 빈도 패턴을 딥러닝 모델로 학습하여 신조어를 분류하는 알고리즘을 개발하고자 하였다.

J. W. Kim et al., 연구[6]에서는 자연어 처리 기술을 사용하여 네이버뉴스 데이터로부터 신조어 단어들을 추출하였다. 그러나 이 연구에서는 이벤트성으로 등장하는 이슈어(Trending word)에 대해서는 명시적인 분류를 하지 않았다. 본 연구에서는 신조어와 일반어 이외에 이슈어를 분류 항목에 포함하여 구분하고자 하였다. 또한 학습 데이터셋으로 네이버뉴스 외에 네이트판과 DC인사이드를 추가하였다.

본 연구에서는 웹사이트에서 크롤링한 말뭉치 데이터로부터 신조어를 분류하기 위하여 검색 빈도 데이터와 로지스틱 회귀, CNN 알고리즘을 활용하였다. 본 연구를 통하여 새롭게 출현하는 신조어들의 의미와 용례를 쉽게 파악할 수 있으며, 언어 사용 실태 분석과 같은 다양한 연구에 활용할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 기존 신조어 관련 서비스 및 최신 연구 동향에 대해 살펴보고, 3장에서 제안하는 검색 데이터 기반 신조어 분류 알고리즘에 관해 기술하였다. 4장에서는 CNN 알고리즘을 소개하고, 실험 및 성능 평가 결과를 제시하였다. 5장 결론에서 본 연구의 한계점 및 향후 연구 방향에 대하여 제시한다.

Ⅱ. 관련 연구

2.1 신조어 관련 서비스

신조어의 의미를 제공하는 서비스에는 '네이버 오픈사전', '위키백과 대한민국 인터넷 신조어 목록' 등이 있다[7]. 네이버 오픈사전은 사용자 참여형 사전으로 약 207개의 신조어가 등록되어 있다. 위키백과 또한 사용자 참여형 사전으로 270개의 신조어가 등록되어 있으나, 단어의 의미 외에 용례 정보는 제공하지 않는다. 이 사이트들은 공통적으로 사용자가 신조어를 직접 등록하는 방식으로 정보 신뢰도가 낮고, 제공되는 신조어의 개수가 적으며 업데이트가 느리다는 문제점이 있다. 이러한 문제는 실제로 신조어가 사용되는 다양한 커뮤니티 사이트에서 데이터를 크롤링하고 머신러닝 기술로 신조어를 수집 분류하는 방법을 통해 개선할 수 있다.

2.2 신조어 분류 연구

기존의 신조어 분류 연구로, J. W. Kim et al., 연구[6]에서는 신조어 추출을 위해 soynlp를 이용한 비지도 학습 방법을 사용하였고, H. J. Kim의 연구[8]에서는 로지스틱 회귀를 활용하였다. W. Li의 연구[9]에서는 바이두 인덱스의 검색 트렌드 기능을 이용하여 중국 인터넷 신조어의 발생 시기, 연대적 검색 트렌드 등을 제시하였다. 그러나 본 연구에서 제안하는 검색 빈도 데이터 기반의 신조어 분류 알고리즘에 대한 연구는 현재까지 사례를 찾기 보기 어렵다.

2.3 CNN 분류 알고리즘

CNN 모델은 일반적인 다층 신경망 딥러닝 프레임워크로 이미지 분류에 탁월한 성능을 보인다[10]. CNN은 일반적으로 광학 문자, 붓꽃 등 실제 사물에 대한 이미지를 분류하는 데에 활용되지만, 본 연구에서는 검색 빈도 데이터 그래프 자체를 이미지로 간주하여 그 패턴을 기준으로 분류하고자 하였다. 이를 위해 네이버 데이터랩에서 제공하는 단어의 검색 빈도 데이터 그래프를 이미지 파일(PNG)로 변환하여 CNN 모델의 학습을 수행하였다.

III. CNN을 활용한 검색 빈도 데이터 기반 신조어 분류 알고리즘

3.1 신조어 분류 시스템

제안 시스템의 신조어 분류 과정은 다음과 같다.

1) 세 개의 웹사이트에서 크롤링한 말뭉치를 전처리하여 초기 데이터셋을 생성한다.

2) 데이터셋에서 전처리하여 수집한 단어에 대한 검색 빈도 데이터를 네이버 데이터랩 API를 통하여 획득한다.

3) 훈련 데이터셋에서 추출한 각 단어의 검색 빈도 그래프를 분석하여 신조어, 일반어, 이슈어로 라벨링한다.

4) CNN 모델로 데이터를 학습하고 신조어, 이슈어, 일반어로 다중 클래스 분류를 수행한다.

5) Streamlit으로 구현된 웹사이트에 신조어 순위, TF-IDF, 검색 빈도 그래프, 단어 뜻, 예문 정보를 출력한다.

그림 2는 본 연구에서 제안하는 시스템의 전체 구조를 나타낸 것이다.

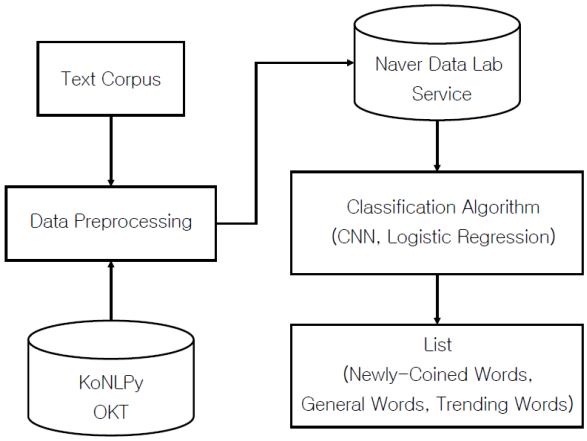


그림 2. 검색 데이터 기반 신조어 분류 시스템 구조
Fig. 2. Structure of newly-coined words classification system based on search data

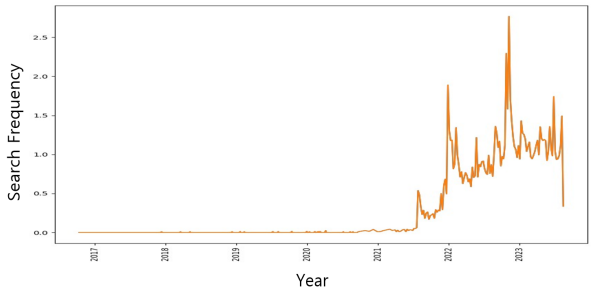
3.2 데이터 수집 및 전처리

신조어 분류를 위한 데이터 수집을 위해 네이트판, DC인사이드, 네이버뉴스를 대표 사이트로 선정하였다. Python BeautifulSoup 라이브러리를 사용한 웹 크롤링을 통해 최근 일주일간의 말뭉치를 수집하였다. 수집한 말뭉치의 범위는‘네이트판 베스트 게시물의 제목, 본문, 댓글’,‘DC인사이드의 베스트 게시물의 댓글’,‘네이버뉴스의 기사 제목과 본문’을 포함한다.

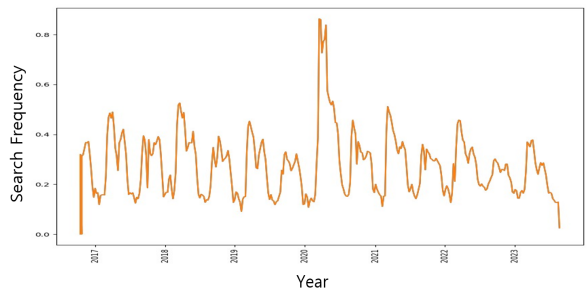
일반 단어사전을 기반으로 하는 형태소 분석기를 신조어에 적용하였을 때 온전한 단어 추출에 어려움이 발생한다. 가령‘마용성’을 기존 형태소 분석기로 토큰화하면‘마’와‘용성’의 형태로 나뉘어 추출된다. 이 문제를 해결하기 위하여 KoNLPy의 OKT를 활용하여 어절 추출을 진행한 뒤 공백이 포함된 어절과 중복된 단어를 제거하고, ranks.nl의 불용어, 한글자 단어, 외국어, 특수문자, 숫자, 자체 제작 불용어는 제거하였다. 다음으로 일반어 제거를 위하여 표준국어대사전에 등재된 단어들과 검색 빈도에서 일반어의 패턴을 보이는 단어를 제거하였다.

3.3 신조어의 유형 분류

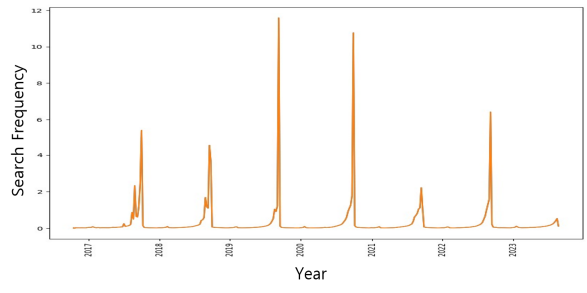
본 연구에서는 단어 유형을 네이버 데이터랩의 검색 빈도 패턴에 따라 신조어, 일반어, 이슈어로 구분하였다. 데이터가 제공되는 총 7년간의 기간 중 최근에 급격하게 검색 빈도가 높아진 후 일정 기간 유지되는 단어를 신조어로 분류하고(그림3-a), 전체 데이터 기간 동안 검색 빈도가 고르게 나타나는 단어는 일반어(그림3-b)로 분류한다. 그리고 검색량이 간헐적으로 급격히 높아지는 단어를 이슈어(그림3-c)로 판단한다.



(a) 신조어 ('인스스')
(a) Newly-coined word ('Instagram story')



(b) 일반어 ('과제')
(b) General word ('Homework')



(c) 이슈어 ('추석')
(c) Trending word ('Thanksgiving day')

그림 3. 단어 유형별 검색 빈도 그래프
Fig. 3. Frequency graph of word types

IV. 실험 및 성능 평가

4.1 실험 및 분류 결과

수집된 단어들을 일반어, 이슈어, 신조어 범주로 라벨링하여 로지스틱 회귀와 CNN 모델로 지도 학습을 진행하였다. 데이터셋은 8:2로 분할하여 훈련 세트과 검증 세트로 사용하였다. 로지스틱 회귀 모델은 사이킷런의 선형모델 라이브러리를 사용하여 구현하였다. CNN 모델은 은닉층에 3개의 합성곱층과 최대 풀링층 쌍, 2개의 밀집층을 두었고, 출력층에는 소프트맥스 활성화 함수를 사용하였으며 옵티마이저로는 Adam을 사용하여 구현하였다. 표 1은 네이트판 데이터셋에 대해서 CNN과 로지스틱 회귀 모델의 신조어 분류 성능 지표를 비교한 것이다.

표 1. 분류성능평가지표
Table 1. Classification performance metrics

Model	CNN		Logistic regression	
Dataset	Train	Test	Train	Test
Accuracy	0.88	0.82	0.60	0.51
Precision	0.89	0.86	0.55	0.46
Recall	0.83	0.83	0.49	0.47
F1 Score	0.85	0.81	0.49	0.46

그림 4와 같이, CNN의 분류 정확도가 모든 지표에 대하여 로지스틱 회귀보다 유의미하게 우수함을 보였으며, 이 결과에 따라 CNN 모델을 제안 시스템의 기본 신조어 분류 알고리즘으로 채택하였다.

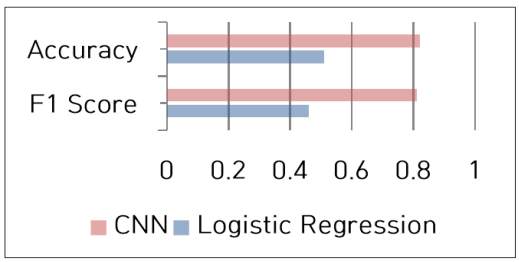


그림 4. 두 모델의 정확도와 F1 스코어 비교
Fig. 4. Comparison of accuracy and F1 score between two models

4.2 신조어 분류 결과

표 2는 CNN으로 다중 분류한 신조어, 일반어, 이슈어의 예시이다. 각 사이트별로 추출한 신조어의 특징에 차이가 있다. 네이트판(NP)과 DC인사이드(DC)에서는 전형적인 신조어가 주로 등장하는 반면, 네이버뉴스(NN)에서는'생성형'등 최근 이슈어와 신조어가 혼동되어 분류되고 있다.

표 2. CNN 모델의 다중 분류 결과
Table 2. Classification results of CNN model

Types	Examples of classified words
Newly -coined words	갓생, 업보빔, 웅니, 지팔지꼰 (NP)
	마라탕후루, 버튜버, 순살아파트 (DC)
	블루푸드, 노랜딩, 생성형 (NN)
General words	시댁집, 시엄니, 위생상, 추천함 (NP)
	연구조사, 주변국, 혐오자, 고영양 (DC)
	소속팀, 여행시간, 재확립, 휴직제도 (NN)
Trending words	강력범죄자, 성형설, 연말무대 (NP)
	검찰관계자, 반일불매, 전쟁위기, 정식연재 (DC)
	개막공연, 양식어류, 환영오찬, 추석대목 (NN)

4.3 신조어 정보 사이트 구현

본 연구 결과를 바탕으로 분류된 신조어의 실시간 정보 웹사이트를 구현하였다. 사이트별 신조어의 등장 빈도 순위, TF-IDF, 최초 출현일, 검색 빈도수 그래프, 단어 뜻, 출현 문장 정보를 제공한다. 신조어의 의미는 생성형 AI인 구글 바드의 API를 통해 획득하였다.

그림 5. 확장 랭킹 (네이트판)
Fig. 5. Extended ranking (Nate pann)

신조어의 활용 사례는 해당 신조어가 추출된 실제 출처 문장을 통해 제공한다. 또한 사용자가 신조어를 입력하면 실시간으로 신조어 여부를 분류 알고리즘으로 판별 후 등록하는 기능을 구현하였다. 그림 5는 추출된 신조어에 대한 확장 정보 화면이다.

V. 결론 및 향후 과제

본 연구에서는 기존 사례가 없는 검색 빈도 데이터 기반 신조어 분류 알고리즘을 제안하였다. 신조어와 일반어의 검색 빈도 패턴의 차이에 착안하여 CNN 모델로 지도 학습 분류 알고리즘을 구현하였다. CNN의 분류 정확도가 약 82%로 측정되어 효과적으로 신조어 탐색이 가능함을 확인하였다.

본 논문의 기대효과는 다음과 같다. 첫째, 최신 신조어의 출현 경향을 쉽게 파악할 수 있다. 둘째, 신조어 정보 제공을 통해, 신조의 의미 파악이 용이하고 의사소통의 어려움을 개선할 수 있다. 셋째, 검색 빈도 데이터를 활용한 후속 연구에 도움을 줄 수 있다. 본 연구의 한계점으로 네이버뉴스 데이터의 경우 신조어와 비슷한 그래프 양상을 보이는 이슈어가 혼동되어 함께 추출된다는 점을 들 수 있다. 또한 제안 시스템에서는 영어와 초성어로 이루어진 신조어는 배제하고 있다.

References

[1] O. C. Raquel, "English Neologisms in Modern Times", Department of English, German and Translation and Interpretation Studies, Mar. 2022.

[2] K. I. Nam, "New Words of 2019 investigation", National Institute of Korean Language, Dec. 2019.

[3] Incruit, https://people.incruit.com/news/popupnewsprint.asp?newsno=3414267 [accessed: Nov. 22, 2023]

[4] L. Šomanová, "Words Recently Coined And Blended: Analysis Of New English Lexical Items", Masaryk University, Jun. 2017.

[5] Naver Datalab, https://datalab.naver.com/keyword/trendSearch.naver [accessed: Nov. 22, 2023]

[6] J. W. Kim, J. W. Jeong, and M. Y. Cha, "Automatic New Korean Words Extraction Using Portal News Headlines", Proc. of HCI Korea 2020, pp. 163-166, Feb. 2020.

[7] H. L. Park, S. J. Jo, H. B. Han, and S. J. Yu, "Text Mining-based Sentiment Analysis of Newly-Coined Words and Implementation of SMU Sentimental Dictionary", Journal of KIIT, Vol. 20, No. 2, pp. 21-28, Feb. 2022. https://doi.org/10.14801/jkiit.2022.20.2.21.

[8] H. J. Kim, "Extraction Method of New Word from Online Community: The Application of New Method to Morphological Analysis", The Graduate School of Information, Yonsei University, Feb. 2019.

[9] W. Li, "The Integration of Chinese Internet Neologisms into Everyday Language", LUND University, Jun. 2023.

[10] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network", EURASIP Journal on Image and Video Processing, Vol. 2019, No. 40, Feb. 2019. https://doi.org/10.1186/s13640-019-0417-8.

저자소개

김 민 정 (Min-Jung Kim)

2019년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심분야 : 머신러닝, 생성형 AI, 자연어처리

김 현 수 (Hyun-Soo Kim)

2019년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심분야 : 머신러닝, 데이터분석

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 전산과학과(이학사)

1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)

2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)

2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수

관심분야 : 데이터마이닝, 추천시스템, 정보시각화