

다중 형태소 분석기 불일치 기반 온라인 커뮤니티 신조어 탐지: 프로야구 도메인 사례 연구

이규진*¹, 김군도*², 오효정**

A Multi-Morpheme Analyzer Disagreement Approach for Detecting Newly-Coined Words in Online Community Texts: a Case Study of Korean Professional Baseball

Gyu-Jin Lee*¹, Gundo Kim*², and Hyo-Jung Oh**

이 논문은 2026년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음

요 약

프로야구와 같이 온라인 팬덤 커뮤니티를 중심으로 통용되는 신조어는 빠르게 생성되고 확산되는 반면, 생명 주기가 짧고 변형이 잦아 기존의 사전 기반 방식으로는 탐지가 어렵다. 본 연구는 이러한 한계를 보완하기 위하여 형태소 분석기 간 불일치를 활용한 신조어 탐지 모델을 제안한다. 프로야구 도메인에서 수집한 게시물 제목의 명사 추출 결과를 바탕으로 일치도가 낮은 소수 의견에 주목하여 신조어 후보를 선정한 뒤 필터링 과정을 거쳐 최종 신조어를 선정하였다. 수작업으로 구축한 200개 문장 평가셋에 기반해 제안 모델의 유효성을 검증한 결과, 정확도는 88.89%, 재현율은 78.05%로 나타났다. 또한 최종 선정된 신조어를 형태적·의미론적 두 가지 유형으로 분류하였으며, 모델을 9개 구단 커뮤니티에 확장 적용하여 총 1,791개의 신조어를 선정함으로써, 형태소 분석기 간 불일치 정보에 기반한 제안 모델의 유용성 및 범용성을 확인하였다.

Abstract

This study proposes a newly-coined words detection model based on disagreement among morphological analyzers, focusing on professional baseball fandom communities. Because online newly-coined words are rapidly generated, short-lived, and frequently modified, they are difficult to detect using dictionary-based methods. Post titles were analyzed with five morphological analyzers, and noun candidates extracted by only a small number of analyzers were selected as newly-coined word candidates. After normalization and filtering, evaluation on manually labeled candidates from 200 sentences showed a precision of 88.89% and a recall of 78.05%. Applying the model to nine additional team communities identified 1,791 newly-coined words, suggesting that analyzer disagreement is an effective cue for domain-specific newly-coined word detection.

Keywords

newly-coined words detection, morpheme analyzer, minority, text mining, online community

* 전북대학교 기록관리학과 석사과정

- ORCID¹: <https://orcid.org/0009-0008-2775-5178>

- ORCID²: <https://orcid.org/0009-0001-9326-6988>

** 전북대학교 문헌정보학과 교수,

문화융복합아카이빙연구소 공동연구원 (교신저자)

- ORCID: <https://orcid.org/0000-0001-8067-2832>

· Received: Jun. 02, 2026, Revised: Jun. 27, 2026, Accepted: Jun. 30, 2026

· Corresponding Author: Hyo-Jung Oh

Dept. of Library & Information Science, Jeonbuk National University,

567 Baekje-daero, Deokjin-gu, Jeonju, Jeollabuk-do, (54896) Korea

Tel.: +82-63-270-3208, Email: ohj@jbnu.ac.kr

I. 서 론

현대의 신조어는 표현이 간결하고 전달 내용이 명확하여 직관적이며, 독특한 결합 양상과 사용 맥락을 반영하는 효율적인 언어 형태로, 기존 단어의 축약·변형·결합을 포함하는 확장된 개념으로 이해할 수 있다[1]. 반면 생성 속도가 빠르고 사용 기간이 짧으며 변형이 빈번하게 발생하기 때문에 어휘 사전에 의미가 충분히 정리되어 있지 않은 경우가 많다. 이로 인해 문맥의 내용 추론이 어렵고 집단 간의 의사소통의 어려움을 심화시키는 요인으로 작용하기도 한다[2].

특히 야구나 게임과 같이 ‘팬덤’ 커뮤니티를 중심으로 소비되는 신조어는 생성 주기가 짧고, 축약이나 변형의 형태로 나타나는 경우가 많다. 또한 맞춤법과 문법이 엄격하게 지켜지지 않으며, 대부분 구어체 표현으로 사용되어 어절의 경계가 불분명하다는 특성을 지닌다. 따라서 표준어 기반의 사전 등 제어에 의존하는 형태소 분석만으로는 빠르게 조형되는 신조어를 탐지하는 데 한계가 있으며, 특히 ‘온라인 커뮤니티’라는 텍스트 유형의 언어적 특성을 반영한 신조어 탐지 방법이 필요하다.

본 연구에서는 다수의 형태소 분석기에서 도출된 결과 간 불일치 정보를 활용하여 신조어를 추출하는 방법을 제안한다. 형태소 분석기 간 높은 일치도를 보이는 다수 의견(Majority)이 아닌, 오히려 분석 결과가 상이하게 나타나는 소수 의견(Minority)에 주목하여 해당 명사를 신조어 후보로 포착함으로써, 형태소 분석기 간 불일치 정보를 신조어 탐지의 주요 단서로 활용하는 역발상적 시도이다. 특히 본 연구에서는 뉴스에도 등장했던 김태형 감독을 친근하게 부르는 별명인 ‘튼둥’과 같이[3], 이미 일상적인 언어로 자리 잡은 야구 신조어가 출현하는 프로야구 도메인을 분석 대상으로 선정하여 제안 모델의 유효성을 검증하고자 하였다.

구체적인 연구 수행 과정은 다음과 같다. 프로야구 분야에서 대표적인 온라인 팬덤 커뮤니티를 선정하여 게시글을 수집하고 게시글 제목을 중심으로 다수의 형태소 분석기를 적용한 뒤, 분석기 간 일치도가 낮은 소수 의견을 중심으로 신조어 후보를 추

출하였다. 이후 단어 경계 정제와 필터링 과정을 거쳐 최종 신조어를 선정하고 접사와 의미 유형에 따라 분류하였다. 나아가 특정 구단 커뮤니티를 기반으로 설계한 신조어 탐지 방법을 다른 구단 커뮤니티에 확장 적용함으로써 제안한 방법의 범용성을 검증하였다.

본 연구의 차별성은 다음과 같다. 첫째, 사전에 등재되지 않은 표현을 직접 탐색하는 기존 방식과 달리, 형태소 분석기 간 불일치 양상을 신조어 탐지의 단서로 활용하였다. 둘째, 한국어 온라인 커뮤니티 텍스트에서 나타나는 조사 결합과 어절 변형을 고려하여 단어 경계 정제와 도메인 기반 필터링을 결합하였다. 셋째, 수작업 평가셋 기반으로 제안 모델의 유효성을 검증함과 더불어 타 구단 커뮤니티에 확장 적용함으로써 제안 방법의 범용성을 확인하였다. 궁극적으로 본 연구는 온라인 커뮤니티로부터 빠르게 생성·확산되는 신조어의 형태적·의미론적 특성을 규명하기 위한 기반 마련에 목적을 둔다.

II. 이론적 배경 및 선행 연구

2.1 신조어

언어학적으로 새로운 단어는 신조어(Neologisms)와 임시어(Nonce words)로 구분된다[4]. 신조어는 언어 공동체에서 일정 수준 수용되어 어휘 체계에 편입될 가능성이 있거나 지속성을 지닌 표현을 의미하는 반면, 임시어는 특정 상황에서 일시적으로 생성되어 사용되는 표현을 가리킨다[4]. 그러나 본 연구와 같이 새롭게 조형된 표현을 추출하는 데 초점을 두는 연구에서는 해당 어휘가 공동체에 정착하였는지보다 새로운 표현의 ‘생성’ 자체에 초점을 두는 경향이 있어, ‘Neologisms’ 대신 ‘Newly-Coined Words’라는 용어가 사용되기도 한다[5]. 실제 온라인 커뮤니티에서는 두 범주가 명확히 구분되지 않는 경우가 많고 새롭게 등장한 표현이 지속성을 지닐지를 즉시 판단하기도 어렵다. 따라서 본 연구에서 지칭하는 ‘신조어’는 온라인 텍스트에서 새롭게 생성된 표현의 탐지에 초점을 두고, ‘Newly-Coined Words’라는 용어의 의미를 사용한다.

2.2 텍스트 마이닝 기반 어휘 분석

텍스트 마이닝은 대규모 비정형 텍스트 데이터를 대상으로 형태소 분석, 단어 빈도 분석, 공기 관계 분석 등을 수행함으로써 텍스트 내부에 잠재된 유의미한 정보와 지식을 추출하고 기존에 드러나지 않은 패턴과 관계를 발견하는 데 활용되는 기법이다[6][7]. 특히 SNS를 비롯한 온라인 커뮤니티와 같이 다수의 사용자로부터 대량의 텍스트가 지속적으로 생성·축적되는 환경에서 유용하게 활용될 수 있다. 이러한 특성을 고려할 때, 텍스트 마이닝은 빠르게 생성되고 변형되는 온라인 신조어를 탐지하고 그 사용 양상과 특성을 분석하는 데 적절한 방법론적 기반을 제공한다.

기존의 텍스트 마이닝을 활용한 신조어 추출과 관련된 선행연구들을 살펴보면 특정 형태소 분석기에 기반해 후보를 추출한 뒤, 추출된 어휘의 동시 출현 단어를 참조하거나 감성적 특성을 파악하는 등의 방법이 주를 이루고 있다[2][8]. 또한, 한국어는 하나의 어절 안에서 어휘 형태소와 조사, 접미사, 어미 등이 결합하는 교착어적 특성을 지니므로 기존 연구에서는 명사 뒤에 등장하는 조사, 접미사, 문서 내 출현 빈도 등을 결합한 확률 모델을 통해 미등록어 후보를 판정하였다[9][10]. 국외 논문으로는 트위터 말뭉치를 통해 신조어 후보를 추출하고, 출현 빈도와 고유 사용자 수 등을 기준으로 필터링을 진행한 후 최종 후보를 선정한 연구가 있다[11]. 다만, 분절 경계가 뚜렷한 영어와 달리, 한국어는 조사 결합, 어절 단위의 변형, 형태소 분석기별 추

출 결과 차이가 크게 나타난다. 이에 본 연구는 단일 형태소 분석기의 추출 결과와 미등록어 판정에 의존하지 않고, 다수의 형태소 분석 결과 간 불일치 정보를 역으로 활용하여 신조어를 추출하고, 이를 접사와 의미 유형의 측면에서 분류한다는 점에서 차별성을 갖는다.

III. 불일치 기반 신조어 탐지

본 장에서는 형태소 분석기 간 불일치 정보를 활용하여 신조어를 탐지하는 과정을 단계적으로 설명한다. 전체 과정은 그림 1과 같으며, 데이터 수집 및 전처리(3.1절), 신조어 후보 탐지(3.2절), 필터링(3.3절), 유형 분류(4.1절)의 네 단계로 구성된다.

3.1 데이터 수집 및 전처리

신조어 탐지를 위한 데이터는 프로야구 관련 게시글이 활발하게 생성되는 온라인 커뮤니티인 ‘디시인사이드’와 ‘Mlbpark’에서 수집하였다. 디시인사이드의 경우 ‘야구 갤러리’가 야구뿐만 아니라 정치·사회·문화 등 다양한 주제를 포괄하는 자유 게시판의 성격을 지니고 있으므로, 야구 관련 게시글 비중이 높은 ‘구단 갤러리’로 수집 범위를 한정하였다. Mlbpark는 한국과 미국 프로야구에서 활약하는 한국 선수 관련 정보를 공유하는 대표적 야구 전문 플랫폼으로, 폭넓은 이용자층과 활발한 게시 활동을 고려하여 수집 대상에 포함하였다.

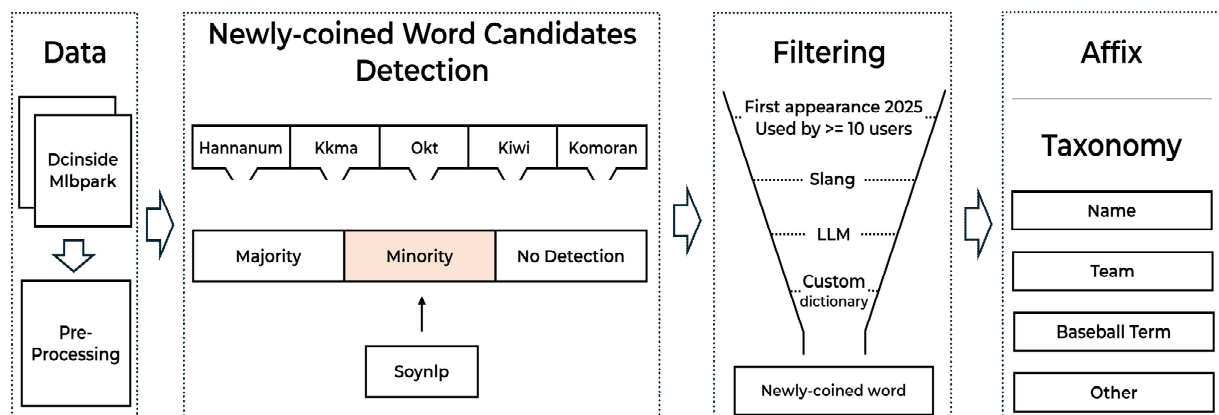


그림 1. 불일치 기반 신조어 탐지 워크플로우

Fig. 1. Workflow of newly-coined words detection model

관중 규모가 클수록 팬덤 활동과 온라인 커뮤니티 내 게시글 생산이 상대적으로 활발하여 신조어 생성 양상을 관찰하기에 유리할 것이다. 본 연구에서는 이 같은 점을 고려하여, 분석 대상 구단으로 2023년부터 2025년까지 3개 시즌 연속으로 관중 규모가 가장 컸던 LG 구단을 선정하였다[12]. 다만 단일 구단 중심의 분석이 지닐 수 있는 편향 가능성을 보완하기 위해, 제안한 탐지 방법을 나머지 9개 구단 커뮤니티에 확장 적용하여 일반화 가능성을 함께 검토하였다.

수집 기간은 2024년 1월부터 2025년 12월까지로 설정하였으며, 분석 대상은 게시글 제목으로 한정하였다. 수집 대상 게시판 특성상, 게시글 제목은 작성자의 의도와 핵심 내용을 직접적으로 반영하는 반면, 본문은 제목과 동일한 내용을 반복하거나 짧은 단어로만 구성되는 경우가 많아 분석 효과가 낮기 때문이다.

수집 결과, 총 4,655,275건의 게시글 제목 데이터를 획득하였으며, 수집된 데이터는 형태소 분석의 정확도를 높이기 위해 다음 두 가지 전처리를 수행하였다. 먼저 정규표현식을 활용하여 한글·영문·숫자 및 공백을 제외한 URL, 이메일, 특수문자 등을 제거하였으며, 다음으로 연속된 공백은 단일 공백으로 정규화하였다.

3.2 신조어 후보 탐지

3.2.1 토큰화 과정

본 절은 동일한 문장에 대해 서로 다른 형태소 분석기를 적용할 경우, 분석기별 사전 구성과 알고리즘 차이에 의해 결과가 상이하게 나타난다는 점에 주목한다. 특히 특정 어절이 소수의 분석기에서만 명사로 인식되는 경우, 해당 어절은 사전에 등재되지 않은 미등록어, 즉 신조어(Newly-coined word) 후보일 가능성이 높다는 가설을 세워 탐지 모델을 고안하였다. 활용한 형태소 분석기는 Kiwi(0.22.2)[13]와 KoNLPy[14]의 Hannanum, Kkma, Komoran, Okt 총 5개이다. 각 분석기의 명사 추출 결과를 비교하여, 동일 어절에 대한 분석 결과의 일치 수준을 다음과 같이 두 가지 범주로 구분하였다: 1) 다수

의견(Majority)은 4~5개 분석기에서 공통적으로 추출된 경우이며, 2) 소수 의견(Minority)은 일부 1~3개 분석기에서만 추출된 경우이다. 다수 의견(Majority)은 대체로 사전에 포함된 어휘가 중심이 되므로 신조어 후보가 포함될 가능성이 상대적으로 낮다. 반면 소수 의견(Minority)은 분석기 간 불일치가 크게 나타나는 집단으로, 신조어 후보가 포함될 가능성이 상대적으로 높은 범주에 해당한다.

가설을 검증하기 위해 200개 문장을 대상으로 수작업 평가셋을 구축하였다. 3인의 신조어 분류 결과를 교차 검증하여 총 218개의 신조어 후보를 추출하였으며, 욕설, 고유명사 등을 제외하고 2025년에 최초로 등장한 41개의 표현을 최종 신조어로 선별하였다. 이를 형태소 분석기 간 일치도 수준과 비교해 본 결과, 소수 의견으로부터 176(35+87+54)개 신조어 후보를 선정, 전체 후보 중 80.74%를 확보할 수 있었으며, 이 중 최종 신조어 34(10+15+9)개, 82.93%가 포함되어 형태소 분석기 간 불일치가 신조어 탐지에 유의미한 단서로 활용될 수 있음을 확인하였다(표 1 참조).

표 1. 일치도 수준별 신조어 후보 및 최종 선정 신조어 분포

Table 1. Distribution of labeled and selected newly-coined words by agreement level

Agreement		Candidates (%)	Corrected (%)
Majority	5	9 (4.13%)	0 (0.00%)
	4	20 (9.17%)	5 (12.20%)
Minority	3	35 (16.06%)	10 (24.39%)
	2	87 (39.91%)	15 (36.59%)
	1	54 (24.77%)	9 (21.95%)
No detection	0	13 (5.96%)	2 (4.87%)
Total		218 (100%)	41 (100%)

표 2는 실제 형태소 분석 과정에서 소수 의견이 어떤 양상으로 나타나는지 구체적으로 살펴보기 위한 사례로, 신조어 후보가 포함된 게시글 제목 2건을 대상으로 각 형태소 분석기가 신조어 후보를 원형에 가깝게 추출하는지, 조사와 결합된 어절로 추출하는지, 또는 음절 단위로 분절하는지 등을 비교하였다. 표 내의 굵은 글씨(Example 1의 ‘위뇨’, ‘튀튼환’, ‘썩참’, Example 2의 ‘댕이어빠’, ‘위뇨’ 등)는

프로야구 커뮤니티에서 사용되는 신조어 후보를 의미하며, ‘-’는 소수 의견을 나타낸 결과이다.

표 2의 Example 1을 살펴보면 ‘오늘’과 같은 일반 명사는 대부분 형태소 분석기에서 안정적으로 추출(다수 의견)되는 반면, 신조어 후보는 분석기별로 다르게 처리되는 양상을 보였다. 예를 들어 ‘위튼환’과 ‘씩참’은 일부 분석기에서 완전한 형태로 추출하지 못하고 ‘위’, ‘환’, ‘참’과 같이 음절 단위로 분절되었다. Example 1과 2에서 공통으로 등장하는 ‘위노’는 Kiwi, Kkma에서는 동일한 형태로 추출되었으나, Hannanum에서는 ‘위노만’, ‘위노보단’과 같이 조사가 결합된 형태로 추출되었다. Okt의 경우 Example 1에서는 ‘위노’로 Example 2에서는 ‘위’, ‘노보’, ‘단’으로 분절하여 결합된 조사에 따라 추출 결과가 달라지는 모습을 보였다. 이러한 결과는 신조어가 기존 형태소 분석기의 사전과 분석 규칙에 안정적으로 포착되지 않으며, 분석기별로 상이한 결과를 산출한다는 점을 보여준다.

표 2. 온라인 커뮤니티 문장 분석 사례

Table 2. Examples of morpheme analysis in online community sentences

Morpheme analyzer	Nouns
Example 1	오늘 위노만 위튼환이랑씩참보다 나으면 개추
Hannanum	오늘, -위노만, -위튼환이랑, -씩참보다, -나, -개추
Kiwi	오늘, -위노, -위튼, -환, -참, -개추
Kkma	오늘, -위노, -위, -환, -환이랑, -이랑, -개
Komorán	오늘, -위, -환, -이랑, -참, -개, -추
Okt	오늘, -위노, -위튼환, -씩, -개추
Example 2	댕이어빠가 지금 대주자 뛰어도 위노보단 잘할듯
Hannanum	-댕이어빠, -위노보단, 듯
Kiwi	-댕이어빠가, 대주자, -위노, 듯
Kkma	-댕이어빠, -지금, 대주자, -위노, 듯
Komorán	-빠, 대주자, 듯
Okt	-댕, -어빠, -지금, 대주자, -위, -노보, -단

Newly-coined word candidates: **Bold**, Minority: -

표 2를 통해 형태소 분석기별 특성도 파악할 수 있었는데, Hannanum은 ‘위노만’, ‘위튼환이랑’, ‘씩

참보다’, ‘위노보단’과 같이 신조어 후보에 조사가 결합된 형태를 하나의 명사 단위로 유지하는 경향을 보였다. 이는 후처리가 필요하다는 한계를 갖지만, 신조어 후보의 형태를 비교적 온전하게 보존한다는 점에서 장점으로 해석할 수 있다. Kkma는 ‘위노’, ‘댕이어빠’ 등 일부 신조어 후보를 비교적 정확하게 추출하였으나, 복합명사를 원형과 분절 형태로 함께 산출하는 경향이 나타났다. Kiwi는 ‘위노’를 안정적으로 추출하였으나, ‘위튼환’, ‘씩참’과 같은 표현은 완전한 형태로 탐지하지 못하였다. Komoran과 Okt는 신조어 후보를 작은 단위로 분절하거나 일부만 명사로 인식하는 경향을 보였다.

결과적으로 표 2는 소수 의견이 단순히 형태소 분석 오류만을 의미하는 것이 아니라, 신조어 후보가 포함될 가능성이 높은 단서임을 보여준다. 즉, 온라인 커뮤니티 신조어는 분석기 간 불일치를 유발하며, 이러한 불일치 양상은 신조어 후보 탐지의 효과적인 단서로 활용할 수 있음을 함의한다. 다만, 소수 의견은 조사가 결합된 형태를 함께 추출하거나 단어를 분절하는 한계를 보이며 ‘이랑’, ‘지금’과 같이 신조어가 아님에도 소수 의견에 포함되는 경우가 있으므로, 단어 경계를 정제하여 신조어의 원형을 복원하는 후처리 과정과 신조어가 아닌 단어를 제거하기 위한 필터링 과정이 필요하다.

3.2.2 단어 경계 정제를 통한 신조어 원형 복원

상기한 바와 같이 소수 의견은 분석기가 안정적으로 분석하지 못하는 신조어 후보를 탐지하는 데 유용한 단서가 된다. 다만 조사가 결합된 어절을 신조어 후보로 선정할 경우, 동일한 신조어가 조사 결합 형태에 따라 서로 다른 명사로 처리될 수 있으므로 신조어 후보의 원형을 확보하기 위해 본 연구에서는 Soynlp의 Word Extraction[15]을 활용하여 단어 경계 정제 후처리를 수행하였다.

Word Extraction은 별도의 사전 없이 학습 데이터에 나타난 문자열의 빈도와 결합 패턴을 분석하여 단어 경계를 추정하는 방법으로, Branching Entropy를 통해 특정 문자열 뒤에 이어질 수 있는 문자의 엔트로피를 계산한다. 다만 신조어 후보의 경계를 적절히 추정하기 위해서는 해당 표현이 실제 문맥

에서 어떻게 사용되는지를 반영한 충분한 학습 데이터가 필요하다. 따라서 3.1에서 수집한 게시글 제목 데이터를 학습 데이터로 활용하여 단어의 경계를 학습하고, 조사가 결합된 후보를 정제하였다.

표 3은 Soynlp의 Branching Entropy를 적용한 결과의 사례를 나타낸 것이다. Entropy 점수의 변화가 크게 나타나는 지점을 경계로 포착하여 ‘댕이어’, ‘댕이어빠’는 ‘댕이어빠’로, ‘워노보단’은 ‘워노’로, ‘뤼튼환이’는 ‘뤼튼환’으로 정제되었다. ‘댕이어’ 뒤에는 이어질 수 있는 문자의 종류가 제한적이기 때문에 상대적으로 낮은 엔트로피를 보이는 반면, ‘댕이어빠’ 뒤에는 ‘가’, ‘는’, ‘도’ 등 다양한 조사가 결합할 수 있어 엔트로피가 증가한다. 반대로 ‘댕이어빠’처럼 조사가 결합된 형태에서는 뒤따를 수 있는 문자가 제한되므로 엔트로피가 다시 낮아지게 된다. 이는 Soynlp가 단어 경계를 학습하여 신조어 후보와 조사·어미를 분리하는 데 활용될 수 있음을 보여준다. 그러나 선정된 후보셋에는 일반어, 고유명사, 욕설, 기존 야구 용어, 선수명 등이 포함될 수 있으므로, 최종 신조어 판정을 위해 필터링 과정이 요구된다.

표 3. 조사 제거를 통한 원형 복원 사례

Table 3. Examples of restoring base forms by removing post-positions

Original		Apply	
Word	Entropy	Word	Entropy
댕이어	0.7667	댕이어빠	3.1762
댕이어빠가	1.6094		
워노보단	0.6730	워노	3.8730
뤼튼환이	2.2183	뤼튼환	3.9845

3.3 필터링 및 선정 결과

그림 2는 신조어 후보인 ‘댕이어빠’, ‘유메키’ 등 5개의 단어를 예시로 초기 후보군에서 조건에 부합하지 않는 표현을 순차적으로 제거하고, 최종적으로 신조어만을 선별하는 과정을 나타낸다. 먼저 2025년 시즌에 새롭게 등장한 표현을 선별하기 위해 최초 등장 시점을 기준으로 정제하였다. 또한 특정 사용자의 일시적 표현을 제외하기 위해 해당 단어를 사

용한 고유 사용자가 10명 이상인 경우만 최종 후보에 포함하였다(①)[11]. 필터링에 적절한 고유 사용자 수를 설정하기 위해 고유 사용자를 5명, 10명, 15명, 20명으로 두고 필터링을 진행하였다. 5명과 10명의 후보 수는 70% 이상 차이가 있었지만, 최종 신조어는 동일하였고, 15명, 20명으로 설정한 경우에는 포착되지 않은 신조어가 있어 10명 이상으로 사용자를 설정하였다.

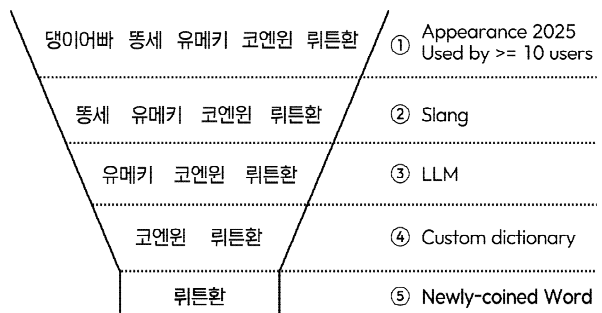


그림 2. 신조어 후보의 단계별 필터링 및 선정 과정
Fig. 2. Multi-stage filtering and selection of newly-coined word candidates

이후 한국어 욕설 필터 정규식[16]을 적용하여 욕설과 비방 표현을 제외하였다(②). 다음으로 생성형 AI 기반 맥락 분석을 통해 신조어가 아닌 고유명사를 제거하였다(③). 예를 들어 시구자 이름과 같이 일시적으로 언급된 고유명사는 형태소 분석기에서 미등록어로 처리될 수 있으나, 본 연구에서 정의하는 신조어가 아니므로 제외된다. 마지막으로 KBO 홈페이지에서 제공하는 선수 등록 현황[17]과 「공식 야구 규칙 2025」[18]의 야구 용어를 기반으로 용어 사전을 구축하여 기존 야구 관련 고유명사 및 선수명을 제외하였다(④).

그림 2의 예시 단어를 통해 필터링 과정을 살펴보면, 최초 등장 시점이 2024년인 ‘댕이어빠’(①), 선수 비방 표현인 ‘똥세’(②), 시구자 ‘유메키’(③), 선수명 ‘코엔원’(④)과 같은 표현이 순차적으로 제거되고 최종적으로 ‘뽀튼환’과 같은 신조어가 남게 된다(⑤). 제안한 불일치 기반 신조어 탐지 모델의 유효성을 검증하기 위해 3.2절에서 일치도 수준별 분포를 확인한 200개의 문장을 대상으로 소수 의견 추출, Soynlp 정제, 필터링 과정을 수행한 결과 최종 신조어는 36개가 추출되었다.

표 4는 200개 문장에 나타난 218개 신조어 후보를 대상으로, 수작업으로 선정한 41개의 최종 신조어와 신조어 탐지 모델을 통해 선정된 36개의 최종 신조어를 정확도와 재현율로 정리한 결과이다. 정확도(88.89%)와 재현율(78.05%) 모두 높게 나타나 제안한 모델이 신조어 후보를 탐지하는 데 효과적임을 검증하였다.

표 4. 라벨링 문장 기반 신조어 탐지 성능

Table 4. Performance of newly-coined words detection on labeled sentences

		Actual	
		Positive	Negative
Predicted	Positive	32(TP)	4(FP)
	Negative	9(FN)	173(TN)

Precision	$32(TP) / [32(TP) + 4(FP)] = 88.89\%$
Recall	$32(TP) / [32(TP) + 9(FN)] = 78.05\%$

특히 가장 성능이 좋았던 단일 형태소 분석기(Kiwi)의 미등록어 탐지 기능을 사용하여 신조어 후보를 포착하는 방식과 본 연구에서 제안하는 다중 형태소 분석기 불일치 기반 방식의 차이를 비교한 결과, Kiwi는 총 157개의 단어를 미등록어 후보로 추출하였고, 제안한 필터링을 통해 29개의 최종 신조어가 선정되었다. 결과적으로 Kiwi의 미등록어 탐지 기능은 96.55%의 높은 정확도를 보였으나, 재현율(68.29%)이 낮게 측정된(F-score: 80.00) 반면, 제안 모델은 단일 분석기에서 미등록어로 판정되지 않더라도 상이하게 추출되는 소수 의견을 활용함으로써 더 많은 신조어를 포착하였음(F-score: 83.12)을 알 수 있다.

제안한 모델을 LG 구단 커뮤니티에서 수집한 제목 데이터의 전체 코퍼스에 적용하여 최종적으로 199개의 신조어를 탐지하였다. 표 5는 필터링을 거쳐 최종 선별된 LG 구단의 신조어 199개를 대상으로 일치도별 정확도를 분석한 결과이다. 전체 199개 신조어 가운데 단일 분석기에서만 추출된 경우가 169개로 전체의 약 85%를 차지, 2개 분석기에서 추출된 경우는 27개, 3개 분석기에서 일치한 경우는 단 3개에 불과하였다. 이는 신조어의 대부분이 분석기 간 일치도가 낮은 영역에 분포하고 있음을 보여

주는 것으로, 본 연구에서 주목한 형태소 분석기 간 불일치 정보가 온라인 커뮤니티 신조어 탐지에서 유효한 단서로 활용될 수 있음을 입증한다.

표 5. 최종 선정 신조어의 정확도

Table 5. Precision of the selected newly-coined word

Agreement		Precision (%)
Minority	3	3/199 (1.51%)
	2	27/199 (13.57%)
	1	169/199 (84.92%)

IV. 신조어 유형 분류 및 확장 적용

본 장에서는 제안한 모델을 기반으로 탐지된 신조어의 언어적 특성을 분석하고, 타 구단 커뮤니티에 대한 적용 결과를 검토함으로써 모델의 유용성과 범용성을 논의한다.

4.1 신조어의 형태적·의미론적 유형 분류

최종 선정된 야구 신조어는 단순히 새로운 문자열이 아니라, 프로야구 팬덤 내부의 인물 지칭, 구단 정체성, 평가 표현 등을 반영하는 언어 단위로 기능한다. 따라서 신조어 형성 방식과 의미 기능을 파악하기 위해 형태적 측면에서는 접사 사용 양상을, 의미론적 측면에서는 신조어가 지시하는 대상을 기준으로 유형 분류를 수행하였다.

표 6은 LG 구단 커뮤니티에서 탐지된 199개의 신조어를 형태적·의미론적 특성에 따라 분류한 결과이다. 먼저 형태적 특성을 살펴본 결과, 프로야구 신조어에는 특정 구단이나 팬덤을 지시하거나 평가·감정의 의미를 부여하는 접사가 함께 나타나는 것을 확인할 수 있었다. 예를 들어 ‘취-’는 LG 구단과 관련된 접두사이며, ‘대-’는 선수나 경기 상황에 대한 긍정적 평가를 나타내는 접두사에 해당한다. ‘-어게인’과 같은 접미사는 기존 표현에 추가적인 의미를 덧붙이는 방식으로 사용되는 양상을 보였다. 이는 프로야구 신조어가 단순히 축약이나 변형에 그치지 않고, 팬덤 내부의 정체성, 평가, 감정을 반영하며 형성된다는 점을 보여준다.

표 6. 신조어 접사 및 의미 유형 분류

Table 6. Classification of newly-coined words by affix and semantic type

Category	Count	Example	Prefix	Suffix	Co-occurrence
Nickname	147	쥐현민	쥐		오다(5), 가다(5)
		대용준	대		보다(2), 가다(2)
		뤼튼환			치다(32), 보다(17)
Team	40	엘어게인		어게인	부르짖다(2), 웃기다(2)
		오문박			치다(16), 타선(15)
Baseball term	9	뤼튼찬스			쓰다(6), 볼넷(2)
Other	3	엘지빵		빵	까다(6), 먹다(4)
Total	199				

의미 유형 분류에서는 신조어의 표현 형태만이 아니라 실제 사용 문장, 공기어, 격틀 구조를 함께 검토하였다. 동일한 표현이라도 문맥에 따라 인물, 팀 지칭어 등으로 다르게 해석될 수 있기 때문에 신조어가 문장 내에서 주어, 목적어, 보어, 수식 대상 등 어떤 위치에 나타나는지, 어떤 조사 및 서술어와 결합하는지를 분석하였다. 예를 들어 격조사(‘-이/가’, ‘-은/는’)와 결합하여 경기 결과나 수행 능력을 나타내는 서술어와 함께 쓰이는 경우에는 선수·감독 등 개별 인물을 지칭하는 별명 유형으로 판단할 수 있으며, ‘타선’ 등 집단 관련 공기어와 함께 사용되는 경우에는 팀 유형으로, ‘쓰다’, ‘나오다’ 등 장비 사용, 경기 행위와 관련된 서술어와 결합하는 경우에는 야구 용어 유형으로 분류할 수 있다.

상기 기준에 따라 야구 분야의 신조어는 표 6에 나타난 바와 같이 별명(Nickname), 팀(Team), 야구 용어(Baseball term), 기타(Other)의 네 가지 의미 유형으로 분류되며, 이들 유형별 조형 패턴을 학습해 특성을 이해한다면, 새로 탐지된 신조어의 의미 분류 역시 자동화 가능함을 알 수 있다.

4.2 신조어 탐지 방법의 확장 적용

제안한 탐지 모델의 범용성, 즉 특정 구단 커뮤니티에만 국한된 것이 아님을 확인하기 위해 동일한 절차를 나머지 9개 프로야구 구단 커뮤니티에

확장 적용하였다. 그 결과 LG 구단을 포함한 10개 구단 커뮤니티에서 총 1,791개의 신조어가 추출되었다. 표 7에 따르면 HANWHA가 402개로 가장 많았으며, LOTTE 271개, LG 199개, SAMSUNG 194개 순으로 나타났다. NC는 82개로 가장 적은 수를 보였다. 이러한 차이는 구단별 게시글 생산량, 팬덤 규모와 참여 연령대, 커뮤니티 활성화 등의 요인과 관련될 가능성이 있을 것으로 해석할 수 있다.

표 7의 구단별 대표 예시를 살펴보면, 프로야구 신조어 형성에는 지역적·문화적 요소가 함께 반영되는 것을 확인할 수 있다. LOTTE의 ‘양쓰카이’는 양석환 선수를 부산 방언으로 표현한 사례이며, KIA의 ‘기영핑’은 ‘캐치! 티니핑’이라는 대중문화 콘텐츠와 선수명이 결합된 사례이다. 또한 ‘번저-’, ‘-어게인’과 같이 동일한 접사가 여러 구단에서 반복적으로 활용되는 양상은 신조어 형성 방식이 팬덤 내부에서 하나의 언어적 관습으로 공유되고 있음을 시사한다.

이러한 신조어 중 상당수는 선수 정보, 경기 기록, 대중문화 등 외부 지식을 통해 의미 유추가 가능한 구조를 지닌다. 따라서 외부 데이터를 연동한 지식그래프를 구축할 경우, 야구 팬덤 신조어의 형성 맥락과 의미 관계를 체계적으로 표현할 수 있을 것이며, 나아가 이를 기반으로 야구 입문자도 쉽게 활용할 수 있는 팬덤 신조어 사전으로의 확장이 가능할 것으로 기대된다.

표 7. 구단별 신조어 예시

Table 7. Examples of newly-coined words by team

Team	Words	Examples
DOOSAN	117	두춧가루, 맹어게인
HANWHA	402	번저손, 폰와류문
KIA	132	기영핑, 똑딱덤
KIWOOM	125	김기특, 뭉뭉
KT	139	대손잔, 케릴라
LG	199	뤼튼환, 엘어게인
LOTTE	271	양쓰카이, 번저탕
NC	82	뽕송섭, 타마카세
SAMSUNG	194	삼파루파, 차외감
SSG	130	쓱정빈, 류고전회
Total	1,791	

V. 결론 및 향후 과제

본 연구는 프로야구 온라인 커뮤니티의 신조어를 탐지하기 위해 형태소 분석 불일치 기반 탐지 방법을 제안하였다. 디시인사이드 구단 갤러리와 Mlbpark에서 수집한 총 4,655,275개의 게시글 제목을 대상으로, 다섯 가지 형태소 분석기의 명사 분석 결과 간 불일치를 신조어 탐지의 핵심 단서로 활용하였다. 라벨링을 통해 제안 방법을 평가한 결과 정확도 88.89%, 재현율 78.05%를 보여, 형태소 분석기 간 불일치 정보가 프로야구 팬덤 신조어 탐지에 유효한 단서로 기능할 수 있음을 확인하였다. 이는 형태소 분석 성능이 좋은 분석기라 하더라도 신조어 탐지에서는 분석기 간 불일치 정보 자체가 유의미한 탐지 단서가 될 수 있음을 보여준다. 또한 제안한 방법을 9개 구단 커뮤니티에 확장 적용한 결과 총 1,791개의 신조어가 추출되었으며, 구단을 지시하는 접두사와 의미를 부가하는 접미사 등 야구 팬덤 내 신조어의 특징적인 형성 양상을 확인하였다. 이러한 결과는 본 방법이 프로야구 팬덤 커뮤니티 뿐 아니라, 유사한 언어적 특성을 지닌 다른 스포츠 및 다양한 온라인 커뮤니티로 확장 적용이 가능할 것으로 기대된다.

다만 본 연구는 몇 가지 한계를 지닌다. 첫째, 특정 온라인 커뮤니티만을 데이터 수집 대상으로 삼았기 때문에 특정 팬덤의 언어 사용 양상에 편향될 가능성이 있다. 둘째, 분석 대상을 명사로 한정함에 따라 다른 품사에 해당하는 신조어나 명사구 유형의 신조어는 충분히 반영하지 못하였다. 셋째, 2024년 이전에 생성되었더라도 해당 연도에 커뮤니티에 등장하지 않은 표현은 2025년 신조어로 분류될 수 있어 최초 등장 시점 판별에 한계가 있다. 따라서 향후에는 수집 기간을 확대하여 연도별 신조어의 생성 및 변화를 보다 정밀하게 추적할 필요가 있다. 마지막으로, 신조어 유형 분류의 세분화와 개별 신조어에 대한 의미 정보 보장 역시 중요한 과제로 남아 있다. 후속 연구에서는 접사 및 유형 분류 결과를 바탕으로 신조어 간의 관계를 구조화하고, 외부 데이터를 연동한 지식 그래프로 시각화함으로써 신조어 연구를 위한 분석 도구로 확장하고자 한다.

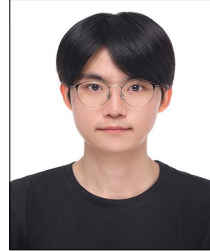
References

- [1] C. W. Kang and I. G. Sohn, "A comparative study of the morphological and phonological simplification of neologisms in English and Korean in digital environments", *The Journal of Linguistics Science*, Vol. 113, pp. 199-221, Jun. 2025. <https://doi.org/10.21296/jls.2025.06.113.199>.
- [2] 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", *The Journal of Korean Institute of Information Technology*, Vol. 20, No. 2, pp. 21-28, Feb. 2022. <http://dx.doi.org/10.14801/jkiit.2022.20.2.21>.
- [3] Teun Dong, https://imnews.imbc.com/original/mbig/6712479_29041.html. [accessed: Nov. 02, 2025]
- [4] E. Mattiello, "Analogy in word-formation: A study of English neologisms and occasionalisms", *De Gruyter Mouton*, pp. 25-35, 2017. <https://doi.org/10.1515/9783110551419>.
- [5] G. M. Nam and N. G. Kim, "Deep Learning-based Target Masking Scheme for Understanding Meaning of Newly Coined Words", *Journal of the Korea Society of Computer and Information*, Vol. 26, No. 10, pp. 157-165, Oct. 2021. <https://doi.org/10.9708/jksci.2021.26.10.157>.
- [6] H. Hassani, C. Benek, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text Mining in Big Data Analytics", *Big Data and Cognitive Computing*, Vol. 4, No. 1, pp. 1-34, Jan. 2020. <https://doi.org/10.3390/bdcc4010001>.
- [7] Y. Matsumoto, "Tourism Information Analysis of SNS Data Using Text Mining", *J. Adv. Comput. Intell. Intell. Inform.*, Vol. 29, No. 4, pp. 829-837, Jul. 2025. <https://doi.org/10.20965/jaciii.2025.p0829>.
- [8] S. J. Kim, A. Y. Lee, and D. Y. Park, "A Study on Neologism Research Trends and Perception of Social Media Users On Neologism Through Big data Analysis", *Hanminjok Emunhak*, Vol. 92, pp. 9-47, Jun. 2021. <https://doi.org/10.31821/HEM.92.1>.

- [9] K. Han, H. Hong, and H. Ji, "Probabilistic Model for Unknown Word Identification Based on Suffix Morpheme Sequences", *Journal of Digital Contents Society*, Vol. 20, No. 4, pp. 843-850, Apr. 2019. <https://doi.org/10.9728/dcs.2019.20.4.843>.
- [10] K. Park and Y. Seo, "Korean Unknown-noun Recognition using Strings Following Nouns in Words", *Journal of the Korea Contents Association*, Vol. 17, No. 4, pp. 576-584, Apr. 2017. <https://doi.org/10.5392/JKCA.2017.17.04.576>.
- [11] P. Brasolin, G. H. Franzini, and S. Spina, "'Ti blocco perché sei un trollazzo". Lexical Innovation in Contemporary Italian in a Large Twitter Corpus", *Proc. 9th Italian Conf. on CLiC-it*. Venice, Italy, pp. 100-112, Nov. 2023. <https://aclanthology.org/2023.clicit-1.14/>.
- [12] D. Y. Kim, K. H. Jang, H. S. Lee, Y. K. Chung, and H. M. Kang, "Trends in the Numbers of Attendance for Korean Professional Baseball Teams Using Dynamic Time Warping(DTW) and Time Series Clustering Analysis", *Journal of Intelligence and Information Systems*, Vol. 31, No. 4, pp. 217-234, Dec. 2025. <http://dx.doi.org/10.13088/jiis.2025.31.4.217>.
- [13] Kiwi, <https://github.com/bab2min/kiwi>. [accessed: Nov. 22, 2025]
- [14] Konlpy, <https://github.com/konlpy/konlpy>. [accessed: Nov. 22, 2025]
- [15] Soynlp Word Extractor, https://github.com/lovit/soynlp/blob/master/tutorials/wordextractor_lecture.ipynb. [accessed: Nov. 20, 2025]
- [16] Korean Curseword Regex, <https://github.com/curiousstovald/KoreanCursewordRegex>. [accessed: Nov. 24, 2025]
- [17] Korea Baseball Player, <https://www.koreabaseball.com/Futures/Player/Register.aspx>. [accessed: Nov. 24, 2025]
- [18] Korea Baseball Rule, https://6ptotvmi5753.edge.naverncp.com/KBO_FILE/ebook/pdf/2025_%EC%95%BC%EA%B5%AC%EA%B7%9C%EC%B9%99.pdf. [accessed: Nov. 24, 2025]

저자소개

이 규 진 (Gyu-Jin Lee)



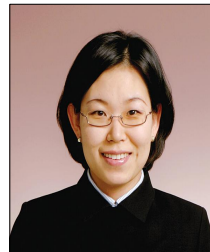
2026년 2월 : 전북대학교
 문헌정보학과(문헌정보학사)
 2026년 3월 ~ 현재 : 전북대학교
 기록관리학과 석사과정
 관심분야 : 전자기록,
 빅데이터정보처리, 텍스트마이닝

김 군 도 (Gundo Kim)



2025년 2월 : 전북대학교
 문헌정보학과(문헌정보학사)
 2025년 3월 ~ 현재 : 전북대학교
 기록관리학과 석사과정
 관심분야 : 전자기록,
 빅데이터정보처리, 텍스트마이닝

오 효 정 (Hyo-Jung Oh)



2008년 2월 : 한국과학기술원
 컴퓨터공학과(공학박사)
 2000년 2월 ~ 2015년 5월 :
 한국전자통신연구원
 인공지능연구센터 책임연구원
 2015년 5월 ~ 현재 : 전북대학교
 문헌정보학과/기록관리학과 교수
 관심분야 : 정보검색, 텍스트마이닝, 빅데이터정보처리