

KcBERT와 TabNet 스택킹 앙상블을 활용한 하이브리드 야구
승패 예측: KBO 리그 사례를 중심으로최홍규*¹, 노민우*², 이승우*³, 이한준**Hybrid Baseball Game Result Prediction using KcBERT and
TabNet Stacking Ensemble: Evidence from the KBO LeagueHongkyu Choe*¹, Minwoo Noh*², Seungwoo Lee*³, and Hanjun Lee**

요 약

스포츠 승패 예측 분야에서 기존 모델은 주로 정형화된 기록에 의존하여 경기 외적인 여론 흐름과 같은 비정형적 요소를 반영하지 못한다는 한계가 있다. 본 연구는 2020~2024년 KBO 리그 기록과 MLB PARK 커뮤니티 게시글 약 100만 건을 통합하여 새로운 승패 예측 프레임워크를 제안한다. KcBERT를 사용하여 팬들의 감성 지수를 정량화하고, 정형 데이터 학습에 특화된 TabNet 기반의 스택킹 앙상블 모델을 구축하였다. 특히 1, 3, 5일간의 여론 시차 변수를 생성하고 시드 초기화 실험을 통해 딥러닝 모델의 초기 가중치 민감도 문제를 완화한 뒤 검증 데이터 기준 최적 모델을 선정하여 안정성을 확보하였다. 실험 결과, 제안 모델은 최고 70.2%의 정확도로 감성 지표가 제외된 모델(58.60%)보다 11.6% 향상되었다. 분석 결과 경기 직전의 여론이 통계적으로 의미 있는 영향력을 보였으며, 이는 커뮤니티 여론이 승부 예측의 설명 변수로 활용될 수 있음을 시사한다.

Abstract

Traditional sports prediction models often fail to capture unstructured factors, such as external public opinion trends, due to reliance on structured records. This study proposes a novel win-loss prediction framework integrating 2020 - 2024 KBO League records with approximately one million MLB PARK posts. We quantified fan sentiment using KcBERT and constructed a TabNet-based Stacking Ensemble specialized for tabular data. Notably, we generated 1, 3, and 5-day time-lag features for public opinion. To address deep learning initial weight sensitivity, we conducted multiple random seed initialization experiments and secured stability by selecting the optimal model based on validation performance. Experimental results show that the proposed model achieved up to 70.2% accuracy, outperforming the model without sentiment indicators (58.60%) by 11.6%. Analysis revealed that sentiment immediately preceding the match had a significant influence, confirming community opinion as a meaningful explanatory variable.

Keywords

machine learning, KcBERT, TabNet, stacking ensemble, sentiment analysis, sports analytics

* 명지대학교 경영정보학과 학사과정
- ORCID¹: <https://orcid.org/0009-0000-8442-5214>
- ORCID²: <https://orcid.org/0009-0001-0224-6253>
- ORCID³: <https://orcid.org/0009-0004-9963-2688>
** 명지대학교 경영정보학과 교수(교신저자)
- ORCID: <https://orcid.org/0000-0002-9005-3661>

• Received: Feb. 23, 2026, Revised: Mar. 15, 2026, Accepted: Mar. 18, 2026
• Corresponding Author: Hanjun Lee
Dept. of Management Information Systems, Myongji University, 34,
Geobukgol-ro, Seodaemun-gu, Seoul, 03674, Korea
Tel.: +82-2-300-0772, Email: hjlee1609@gmail.com

I. 서 론

최근 빅데이터 분석 기술이 발전함에 따라, 스포츠 산업에서도 데이터를 기반으로 전략적 의사결정을 내리는 접근이 필수적인 요소로 자리 잡고 있다. 특히 야구는 '기록의 스포츠'라 불릴 만큼 방대한 경기 데이터가 축적되어 있어 객관적 전력 분석이 가장 활발하게 이루어지는 종목이다. 초기에는 선수 평가에 머물렀던 데이터 활용은 최근 머신러닝 및 딥러닝 기술과 결합하여 경기 결과를 사전에 예측하는 고도화된 모델링 연구로 진화하고 있다[1].

그러나 기존 연구들은 평균자책점, 타율, 출루율 등 경기 내적인 정량적 데이터에만 의존하여 모델을 구축한다는 한계를 지닌다. 스포츠 경기 결과는 선수들의 신체적 기량뿐만 아니라 팀의 분위기, 선수의 심리 상태, 팬들의 응원 열기 등 경기 외적인 무형의 요소들이 복합적으로 상호작용한 결과물이다. 특히 지속적인 경기가 이루어지는 시즌 야구에서는 정형화된 기록 데이터만으로는 당일 경기의 불확실성과 변동성을 온전히 설명하는 데 한계가 존재한다. 이에 따라 학계에서는 정량적 지표가 포착하지 못하는 잠재적 승패 요인을 발굴하기 위해 텍스트 마이닝을 통해 대중의 반응을 분석하려는 시도가 이어지고 있다[2].

본 연구는 이러한 맥락에서 국내 프로야구 팬덤 커뮤니티에 주목하였다. 스포츠 분야에서 온라인 커뮤니티는 경기 전후 즉각적이고 활발한 소통이 이루어지는 곳으로 이곳에서 형성된 여론은 의견 공유를 넘어 경기에 대한 외적인 사회적 분위기를 대변하는 성격을 띤다. 팬들의 게시글에 담긴 기대, 실망, 비판, 응원 등의 감정은 경기의 현재 흐름을 반영하는 선행 지표가 될 수 있으며 이를 정량화하여 모델에 반영한다면 기존 통계 기반 모델의 예측 성능을 개선할 수 있을 것으로 기대된다.

하지만 커뮤니티 데이터는 은어, 비속어, 신조어, 불완전한 문장 구조 등이 빈번하게 사용되어 일반적인 자연어 처리 모델로는 정확한 감성 분석을 수행하는 것에 어려움이 있다. 이를 해결하기 위해 한국어 댓글 데이터로 사전 학습된 KcBERT(Korean Comments BERT)를 사용하여 야구 커뮤니티의 구어체 텍스트를 효과적으로 분석하고자 하였다.

따라서 본 연구의 목적은 2020년부터 2024년까지의 KBO 리그 데이터와 국내 최대 야구 커뮤니티(MLBPARK)의 텍스트 데이터를 결합한 승부 예측 모델을 제안하는 것이다. 구체적으로는 KcBERT를 통해 산출된 팬들의 감성 지수를 변수화하고, 이를 정형 데이터 처리에 특화된 딥러닝 모델인 TabNet 기반의 스타킹 앙상블 기법에 적용하여 승패 예측의 정확도를 최적화하고자 한다. 나아가 경기 직전 1일, 3일, 5일간 누적된 여론이 실제 경기 결과에 미치는 시차 효과를 확인함으로써 팬 커뮤니티의 텍스트 데이터가 스포츠 승부 예측의 새로운 설명 변수로서 지니는 가치를 입증하고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 검토하고, 3장에서는 데이터 수집과 모델링을 설명한다. 4장에서는 실험 결과 및 성능 비교 결과를 제시하고 5장에서는 시사점과 향후 연구 방향을 논의한다.

II. 선행연구

초기 스포츠 분야의 데이터 분석은 주로 경기 승리 요인을 규명하는 데 집중되었다. 선수나 팀의 기록이 승리와 결과에 미치는 영향을 분석하여 승패의 결정적 요인을 파악하는 것이 주된 목표였으며, 다양한 스포츠 중에서도 야구 경기는 통계와 밀접한 관련이 있는 스포츠로 이러한 통계를 활용한 연구가 지속적으로 이어져왔다[3][4].

현재는 단순한 승리 요인 분석에서 나아가 이전의 성적을 바탕으로 머신러닝 및 딥러닝을 활용한 승패 예측 분야의 연구가 지속적으로 이루어지고 있다. 주로 과거의 경기 기록과 선수들의 누적 성적을 바탕으로 로지스틱 회귀분석과 같은 통계적 방법에서 최근에는 인공지능망(ANN, Artificial Neural Network)을 활용한 연구가 진행되고 있다. 그중에서 한국프로야구 리그인 KBO 승부 예측 연구에 따르면 약 61%~64%의 정확도를 기록한 바 있다[5]. 한 가지의 예측 모델만 설계하는 것이 아니라 선수 기록에 관련된 데이터를 시즌별로 나누어 시즌별로 특화된 여러 개의 예측 모델을 설계하는 연구도 진행된 바 있다[6]. 이러한 흐름은 국내뿐만 아니라 국제 연구에서도 활발히 나타나고 있다. 해외 야구

리그인 메이저리그(MLB)를 대상으로 한 해외 연구들은 주로 세이버메트릭스(Sabermetrics) 기반의 방대한 야구 통계 지표에 서포트 벡터 머신(SVM, Support Vector Machine), 랜덤 포레스트(Random Forest), 인공신경망(ANN) 등 다양한 기법을 접목하여 경기 결과를 예측해 왔다[7][8]. 그러나 스포츠 경기가 가진 데이터의 특성으로 인하여 수십 개의 정교한 정량적 변수와 딥러닝 알고리즘을 사용하더라도 모델의 예측 정확도는 대체로 55%에서 62% 수준에 머무르고 있다. 이러한 연구들은 특정 구단이나 시즌에 한정된 데이터를 활용하여 제한된 환경에 특화된 모델 설계에 집중하였다. 이는 정량적 데이터만을 활용한 접근이 예측 성능 개선에 제약을 가질 수 있음을 시사한다.

데이터 분석 기술의 발전으로 여러 형태의 데이터 처리가 가능해짐에 따라 스포츠 승패 예측 분야에서도 분석의 범위가 확장되고 있다. 대표적으로 텍스트 데이터를 결합한 연구가 시도되고 있다. 그 중에서도 소셜 네트워크 서비스에 작성되는 경기 반응에 대한 비정형 텍스트를 텍스트 마이닝 기법으로 분석하여 팀의 분위기와 내재적 흐름을 파악하려는 시도는 기존 전력 분석의 한계를 보완하는 새로운 연구 방향으로 주목받고 있다[9]. 2020년부터 2021년의 경기 기록에 대한 통계와 뉴스 기사에 대한 자연어 처리를 결합한 야구 승부 예측 시스템을 제안한 선행연구에서는 비정형 텍스트 데이터를 활용함으로써 예측 성능이 향상됨을 확인하였으며, 이를 통해 이러한 접근의 유효성을 제시하였다[10].

본 연구는 스포츠 분야에서 기존 선행 연구들이 주로 정량적인 경기기록에 의존하여 경기 승패를 예측해 온 한계를 극복하고자 팬 커뮤니티 데이터를 활용하여 국내 프로야구 경기인 KBO의 승부 예측 모델을 구축하고자 한다. 뉴스나 경기 기록 데이터와 같이 객관적 사실 위주의 정보는 팀을 둘러싼 경기 외적인 분위기나 팬들의 반응과 같은 여론 및 사회적 반응의 흐름을 충분히 반영하지 못한다는 한계가 있다. 따라서 이러한 기존 데이터의 한계와 다른 방향성을 제시하고자 뉴스보다 반응 속도가 빠르고 감정 표현이 드러나는 팬 커뮤니티 데이터를 핵심 분석 자료로 활용하였다. 특히 KcBERT를 도입하여 커뮤니티 특유의 은어를 효과적으로 분석한 뒤 이를 예측 모델에 통합하고자 하였다. 최종적으로 기존의 야구 통계 지표와 결합하여 다중 앙상블 예측 모델로 통합함으로써 기존 연구들과 차별화된 예측 성능론을 제시하고자 한다.

III. 연구 방법

본 연구는 2020년부터 2024년까지의 KBO 정규 리그 데이터를 바탕으로, 정형 데이터(경기 및 선수 기록)와 비정형 데이터(커뮤니티 여론 텍스트)를 융합한 승패 예측 모델을 제안한다. 제안하는 방법론의 전체적인 파이프라인은 그림 1과 같으며 다음과 같이 구성된다. 먼저 경기 기록과 커뮤니티 게시글 데이터를 수집한 뒤, 정형 데이터의 전처리 및 파생 변수 생성을 수행한다.

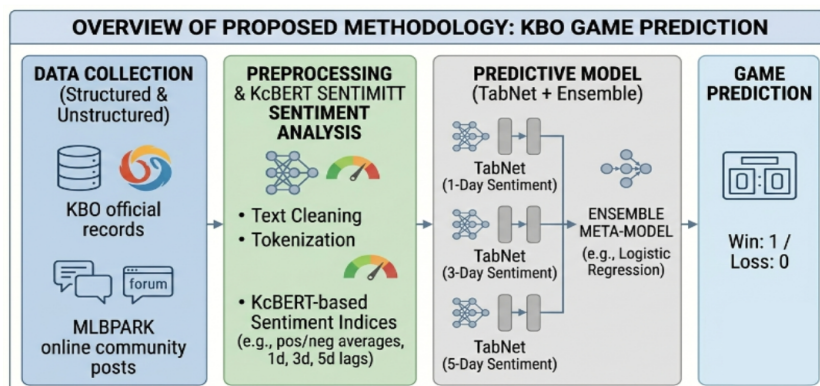


그림 1. 연구 흐름도

Fig. 1. Overall research process

동시에 KcBERT를 활용하여 비정형 텍스트로부터 관측 기간별(1일, 3일, 5일) 감성 지수를 산출함으로써 최종 학습 데이터셋을 구축한다. 이후 구축된 융합 데이터를 TabNet 기반의 스테킹 앙상블 모델에 학습시켜 최종 승패를 예측하고 모델의 성능을 평가하는 단계로 이어진다.

3.1 데이터 수집 및 전처리

연구의 정형 데이터는 KBO 공식 기록실의 기록을 사용하였다. 경기 일자, 대진 팀(홈/원정), 구장 정보 등의 경기 정보와 선수 기록(ERA, Earned Run Average, WHIP, Walks and Hits Per Inning Pitched 등)을 포함한다. 투수 지표는 선발 투수의 기록만 사용하였으며, 전 시즌의 누적 평균 데이터를 활용하였다[10]. 전 시즌 기록이 존재하지 않는 선수의 경기는 분석 대상에서 제외하였으며, 이에 따라 최종적으로 사용된 경기의 수는 2,653경기가 반영되었다. 수집된 데이터는 다음과 같은 절차에 따라 모델 학습에 적합한 형태로 전처리하였다. 먼저 승리 여부는 홈팀 승리를 기준으로 1, 패배를 0으로 정의하여 이진 분류 문제로 설정하였으며, 무승부 경기는 데이터셋에서 제외하였다. 야구와 같은 스포츠 경기는 두 팀 선수 간의 상대적인 전력 차이가 승패를 결정짓는 핵심 요인이다. 따라서 본 연구에서는 홈팀과 원정팀의 선발 투수의 스탯 절대값을 그대로 사용하지 않고, 두 팀 간의 차이를 나타내는 파생 변수를 생성하여 활용하였다. 구체적으로 홈팀과 원정팀 선발 투수의 평균자책점 차이(diff_era), 시즌 누적 홈런 수의 차이(diff_hr), 팀 타율의 차이(diff_avg) 등 전력의 우위를 직관적으로 비교할 수 있는 총 49개의 차분 변수를 모델의 입력값으로 구성하였다. 이를 통해 모델이 단순한 기록 입력이 아니라 상대 팀 대비 전력의 우세 정도를 학습하도록 설계하였다.

비정형 데이터는 국내 최대 야구 커뮤니티인 MLBPARK(<http://mlbpark.donga.com/>)의 ‘한국야구타운’ 게시판의 데이터를 웹 크롤링으로 수집하였다. 해당 커뮤니티는 동아일보가 운영하는 동아닷컴 산하의 스포츠 커뮤니티로 2001년부터 현재까지 20년

넘게 운영된 커뮤니티이다. 특히 경기 전후로 팬들의 즉각적인 반응이 게시글 형태로 활발히 생성되는 공간이다. 따라서 데이터 편중을 방지하기 위해 경기 일자별 최대 4,000건으로 제한된 수의 게시글을 수집하였으며, 광고성 글이나 시스템 자동 생성 게시물과 같이 중복되는 게시글은 제외하였다. 이후 수집한 데이터를 구단별로 나눈 뒤 제목의 텍스트 데이터만 사용하여 최종적으로 2020년부터 2024년의 정규리그 시즌 전체 기간 동안 게시된 총 1,048,576건의 커뮤니티 게시물 텍스트 데이터를 확보하였다.

3.2 자연어 처리를 통한 감성 분석

연구에서 수집한 야구 커뮤니티 데이터는 자음과 모음 기반의 이모티콘, 은어와 같이 문법에 맞지 않거나 불완전한 문장 구조가 빈번하게 나타나는 구어체 특성을 가진다. 기존의 한국어 BERT 모델인 KoBERT(Korea BERT)는 주로 위키피디아나 뉴스 기사와 같은 정제된 문어체를 기반으로 학습되었기 때문에 커뮤니티 특유의 비정형 텍스트를 분석하는데 있어 문맥적 의미를 충분히 포착하지 못하는 한계가 있다. 이에 본 연구는 네이버 뉴스 댓글과 대규모 커뮤니티 텍스트를 기반으로 사전 학습된 KcBERT 모델[11]을 사용하였다. KcBERT는 1억 1천만 개 이상의 뉴스 댓글 데이터를 바탕으로 학습되어 신조어와 오타자, 그리고 구어체 표현이 포함된 문맥을 이해하는 데 특화되어 있다. 특히 야구 커뮤니티 게시글은 뉴스 댓글과 유사하게 짧은 호흡의 문장과 감정적인 표현이 주를 이루므로 문어체 기반 모델보다 KcBERT의 토큰라이저가 해당 도메인의 언어적 특성을 더 효과적으로 해석할 수 있다고 판단하였다.

추가적으로 KcBERT 모델을 기반으로 야구 커뮤니티 데이터의 도메인에 맞는 긍·부정 감성 분석을 수행하기 위해 추가 학습을 적용하였다. 학습에는 대규모 커뮤니티 데이터의 수동 라벨링에 따르는 시간적·비용적 제약을 극복하고 도메인 적응성을 높이기 위해 소량의 정답 데이터를 바탕으로 미학습 데이터에 대한 예측치를 부여하는 의사 라벨링

기법[12]을 적용하였다. 이는 모델이 높은 신뢰도로 예측한 샘플을 다시 학습 데이터에 포함시켜 감성 분류 경계를 정교화하는 방식이다. 본 연구에서는 이를 비정형 데이터 구축 과정에 적용하였다. 가장 먼저 수집된 기본 데이터에 대해 기본 KcBERT 모델로 추론을 수행하여 긍정 및 부정 확률을 산출하였다. 그 이후 라벨링 과정에서 발생할 수 있는 노이즈를 최소화하기 위해, 모델의 예측 확률이 0.85 이상인 데이터만을 신뢰 구간으로 선별하여 분류의 명확성을 높이고자 하였다. 이 과정에서 특정 시기에 편향되지 않도록 전체 수집 기간인 1,102일 내에서 일 수별로 최대 30개의 데이터를 무작위로 추출하였으며, 긍정 및 부정 클래스의 개수를 동일하게 맞추는 과정을 수행하였다. 그 결과로 부정과 긍정 레이블이 각각 11,645건씩 포함된 총 23,290건의 데이터셋을 구축하였다.

구축된 총 23,290건의 데이터셋은 모델의 객관적인 성능 평가를 위해 공부정 클래스의 비율을 동일하게 유지하면서 학습용 80%와 검증용 20%로 분할하였다. 학습 단계에서 KcBERT 모델의 사전 학습된 가중치를 초기값으로 설정하고, 파라미터를 미세 조정하였다. 이때 최적화 도구로는 AdamW 알고리즘을 사용하였다[13]. AdamW는 기존 Adam 옵티마이저가 가중치 감소를 적용하는 과정에서 발생하는 일반화 성능 저하 문제를 개선한 알고리즘으로 BERT와 같은 대규모 트랜스포머 모델의 파인 튜닝 시 과적합을 억제하고 수렴 안정성을 높이는 데 효과적이다. 학습 효율을 위해 입력 시퀀스의 길이는 최대 128 토큰으로 제한하였으며, 배치 크기는 32로 설정하였다. 학습은 총 3회에 나누어 진행되었으며, 학습 종료 시마다 검증 데이터셋을 통해 성능을 모니터링하였다. 실험 결과, 2회차 학습 시점에서 검증 손실이 최소화되며 가장 안정적인 일반화 성능을 보였다. 따라서 최종적으로 선택된 모델은 검증 데이터셋에 대하여 0.8483의 정확도를 기록하였으며, 이는 모델이 야구 커뮤니티의 비정형 텍스트 속 긍정과 부정의 뉘앙스를 의미 있게 판별할 수 있음을 시사한다. 최종 파라미터 조합은 표 1에서 보는 바와 같다.

표 1. KcBERT의 최종 파라미터 조합

Table 1. Final Parameter Settings for KcBERT Fine-tuning

Parameter	Learning rate	Epochs	Optimizer
	Batch size	Sequence length	Number of labels
Value	2e-5	3	AdamW
	32	128	2

학습이 완료된 모델을 활용하여 전체 데이터에 대한 감성 추론을 수행하였으며, 산출된 개별 게시글의 긍·부정 확률값은 경기 데이터와 결합하기 위해 일자별·구단별로 집계하였다. 구체적으로, 경기 당일을 기준으로 홈팀과 원정팀을 구분하고 해당 일자에 속하는 게시글들의 긍정 및 부정 지표를 각각 평균하여 홈팀 긍정 평균(avg_pos_home), 원정팀 부정 평균(avg_neg_away)과 같은 변수를 구성하였다. 이후 팬들의 여론이 경기력에 미치는 시차 효과를 검증하기 위해, 생성된 기초 변수를 기준으로 직전 1일, 3일, 5일 동안의 평균값을 각각 산출하여 최종 파생 변수로 활용하였다. 아래 표 2는 파생 변수의 기초 통계량을 보여준다. 분석 결과, 경기 1일 전 홈팀과 원정팀의 긍정 및 부정 감성 점수 평균은 0.50-0.57 사이에 분포하고 있으며, 관측 기간이 길어질수록 표준편차가 다소 감소하여 누적된 감성 흐름이 안정화되는 경향을 보였다.

표 2. 감성 점수 기초 통계량

Table 2. Descriptive statistics of sentiment scores

Variable	Mean	Std	Min	Max
Pos_1_Day_Home	0.540	0.134	0.142	0.860
Neg_1Day_Home	0.569	0.117	0.200	0.880
Pos_3_Day_Home	0.505	0.098	0.200	0.850
Neg_3_Day_Home	0.543	0.088	0.230	0.824
Pos_5_Day_Home	0.493	0.087	0.241	0.850
Neg_5_Day_Home	0.526	0.078	0.271	0.803
Pos_1_Day_away	0.536	0.129	0.166	0.861
Neg_1Day_away	0.573	0.114	0.200	0.900
Pos_3_Day_away	0.504	0.098	0.198	0.850
Neg_3_Day_away	0.544	0.087	0.230	0.820
Pos_5_Day_away	0.493	0.087	0.240	0.850
Neg_5_Day_away	0.526	0.078	0.280	0.800

추가적으로, 양 팀에 대한 팬덤의 관심도 격차를 반영하고자 양팀의 게시글 차이 변수

(post_count_d_day_diff)를 일자 수에 맞게 추가하였다. 이는 경기 n 일간 홈팀과 원정팀의 게시글 수 차이를 산출한 것으로 화제성의 상대적 우위를 나타낸다. 따라서 본 연구의 예측 모델에 투입된 최종 독립변수와 종속변수의 세부적인 범주와 명세는 표 3과 같다. 결과적으로 특정 시점의 단발적인 반응과 누적된 감성 흐름뿐만 아니라, 각 구단을 향한 대중의 관심도 집중 현상이 통합적으로 고려하여 팬들의 반응이 실제 경기 승패와 어떠한 관계를 맺는지 입체적으로 분석하고자 한다.

3.3 모델링 및 학습

본 연구는 데이터셋의 특성을 고려하여 정형 데이터 학습에 특화된 딥러닝 모델인 TabNet을 기반 모델로 사용하였다. 또한, 감성 정보가 경기력에 미치는 시차 효과를 다각도로 포착하고 예측의 안정성을 최적화하기 위해 스택킹 앙상블(Stacking ensemble) 기반의 최적 모델 탐색 시스템을 구축하였다.

TabNet은 의사결정나무의 장점인 해석 가능성과 심층 신경망의 장점인 표현 학습 능력을 결합한 아키텍처이다[14]. 본 연구에서는 TabNet의 순차적 어텐션 메커니즘을 활용하여 모델이 매 단계마다 승패 예측에 중요한 변수를 스스로 선택하도록 유도하였다. 이는 수십 개의 경기 지표와 감성 변수 중에서 핵심 인자에 가중치를 집중시켜 일반화 성능을 높이는 데 기여한다.

단일 시점의 데이터만으로는 여론의 형성 과정과 경기력 간의 복합적인 관계를 설명하기 어렵다는 점을 보완하기 위해 기간별로 특화된 모델을 설계

하고 이후에 결합하는 2단계 계층적 스택킹 앙상블 구조를 설계하였다. 먼저 1일(1-day), 3일(3-day), 5일(5-day) 간격의 시계열 윈도우로 집계된 데이터셋을 각각 학습한 3개의 독립적인 TabNet 모델을 구축하였다. 각 모델은 해당 기간의 데이터 특성에 최적화된 승리 확률을 산출한다. 이렇게 각 TabNet 모델이 예측한 확률값을 새로운 입력 변수로 하여 최종적으로 로지스틱 회귀모델이 승패를 판별한다. 최종 예측 식은 (1)과 같이 정의된다. 이때 Y 는 모델이 최종적으로 판별한 홈팀의 승리 확률 결과값을 의미하며, o 는 입력값을 0과 1 사이의 확률값으로 변환하는 시그모이드 함수를 나타낸다. P 는 각각 1일, 3일, 5일의 관측 기간을 시차 변수로 학습한 독립된 단일 TabNet 모델들이 출력한 홈팀 승리 확률 예측값이다. ω 는 각 시점별 모델의 신뢰도를 반영하는 가중치이며, 이 가중치를 학습하여 단일 모델의 편향을 보정한다. 이 가중치 값의 크기는 승패 예측에 있어 단기(1일), 중기(3일), 장기(5일) 데이터 중 어느 시점의 정보가 상대적으로 높은 예측 기여도를 나타내는지에 대한 지표로 해석할 수 있다.

$$y = \sigma(w_0 + w_1 P_{1Day} + w_2 P_{3Day} + w_3 P_{5Day}) \quad (1)$$

TabNet과 같은 심층 신경망은 초기 가중치에 따라 학습 수렴도와 성능의 편차가 발생하는 특성이 있다. 따라서 이러한 초기화 불안정성을 해결하고자 서로 다른 난수 시드를 적용한 다중 실험을 수행하였다. 학습 과정에서 검증 데이터셋의 손실이 최소화되는 파라미터를 채택함으로써 테스트 데이터에 대한 과적합 가능성을 완화하고 모델의 재현성과 일반화 성능을 최적화하였다.

표 3. 변수 설명

Table 3. Description of variables

Category	Variable
Player statistics	Differences in statistics between home and away starting pitchers: Earned Run Average (era_diff), Walks and Hits Per Inning Pitched (whip_diff), Base on Balls / Walks (bb_diff), Strikeouts (so_diff), total 49 variables
Community sentiment	Post count per team(Post count), Positive and negative ratios extracted from text data (e.g., pos_1_day, neg_1_day), calculated over 1-day, 3-day, and 5-day observation periods prior to the match, total 18 variables
Target variable	Match result (label: Home team win=1, loss=0), total 1 variable

IV. 연구 결과

4.1 모델 성능 평가

본 연구에서 제안한 TabNet 스택킹 앙상블 모델의 성능을 검증하기 위해 최종적으로 완성된 모델과 기존 머신러닝 모델(Random Forest, XGBoost) 및 단일 TabNet 모델의 성능을 비교하였다. 데이터셋은 7:3 비율로 분할하였으며, random_state를 고정하여 공정한 비교를 수행하였다.

표 4. 모델별 성능 비교

Table 4. Performance comparison by models

Model	Accuracy	F1-score	Precision
RF	0.612	0.608	0.609
XGB	0.635	0.629	0.631
TabN	0.702	0.698	0.700

실험 결과는 표 4에서 보는 바와 같이, 기존 트리 기반 머신러닝 모델(Random Forest, XGBoost)은 60% 초반대의 정확도를 보인 반면, 본 연구에서 제안한 모델은 약 70.2%의 정확도를 달성하였다. 이는 정형 데이터뿐만 아니라 커뮤니티 감성 데이터의 시차적 정보를 효과적으로 결합한 결과로 해석할 수 있다.

표 5. 변수별 성능 비교

Table 5. Performance comparison by feature

Setup	Included features	Accuracy
Base	Player stats only	0.586
Addition	Base + 1_Day	0.674
	Base + 3_Day	0.595
	Base + 5_Day	0.600
Ablation	Ensemble - 1_Day	0.590
	Ensemble - 3_Day	0.685
	Ensemble - 5_Day	0.667
Ensemble	Base + All Days	0.702

나아가 본 연구에서 도입한 커뮤니티 감성 데이터가 모델의 예측력에 미치는 기여도를 평가하기 위하여 입력 변수 조절을 수행하였다. 표 5는 정형 변수와 각 기간별 비정형 감성 변수의 조합에 따

른 모델의 성능 변화 보여준다. 실험 결과, 커뮤니티 관련 데이터를 완전히 제외하고 선발 선수의 정량적 경기 통계만을 학습한 베이스라인 모델의 예측 정확도는 58.60%에 그쳤다. 반면, 모든 관측 기간의 감성 지표를 결합한 최종 앙상블 모델은 70.20%의 정확도를 기록하며 확인한 성능 향상을 보였다. 특히 각 기간별 변수의 영향력을 살펴본 결과, 경기 직전 1일차 감성 지표가 단일 모델 추가 시 67.40% 가장 높은 성능을 보여주었으며, 반대로 최종 앙상블에서 1일차 지표를 제외했을 때 정확도가 59.00%로 크게 하락하는 현상이 관찰되었다. 이는 비정형 텍스트 데이터, 그중에서도 단기적인 여론의 흐름이 기존의 정량적 경기 지표로는 설명할 수 없는 예측 변수로 활용 가능성을 시사한다. 결과적으로 본 연구는 스포츠 승패 예측 분야에 비정형 커뮤니티 및 여론 지표를 결합하는 방법론이 예측 모델 성능 고도화 작업에서 역할 수행 가능성이 있음을 확인하였다.

4.2 변수 중요도 및 감성 지표의 영향력 분석

본 연구는 스택킹 앙상블 방식을 통해 TabNet 모델이 단기(1일), 중기(3일), 장기(5일) 데이터 중 어느 시점의 정보가 가장 결정적인 역할을 하는지 규명하였다. 따라서 1일, 3일, 5일 데이터를 학습한 개별 모델들의 예측값은 최종 결합 모델로 입력되어 최종 승패를 산출한다. 이때 결합 모델 내부의 가중치인 회귀계수를 확인하면 모델이 1일, 3일, 5일 중 어떤 시점의 정보를 가장 신뢰하고 핵심적인 근거로 삼았는지 파악할 수 있다. 분석 결과, 1일차 모델에 부여된 가중치가 2.154로 가장 높게 나타났다. 이어서 5일차 모델이 1.023, 3일차 모델이 0.842이다. 이는 커뮤니티의 즉각적인 반응과 전날의 경기력이 다음 날 승패에 가장 높은 선행 지표로 작용함을 시사한다. 반면, 3일 및 5일 데이터는 보조적인 지표로서 기능하며, 1일차 모델이 놓칠 수 있는 중장기적인 흐름을 보완하는 역할을 수행하는 것으로 해석된다. 이러한 앙상블 모델의 사용으로 단순히 산술 평균을 취하는 것이 아니라, 예측력이 높은 특정 시점에 가중치를 집중시키는 적응형 판단을

내렸다. 따라서 팬들의 여론과 팀 분위기가 승패에 반영되는 최적의 시차를 통계적으로 확인한 뒤 적용하였으며 이를 통해 본 연구의 모델이 이러한 시차적 특성을 효과적으로 학습했음을 보여준다.

TabNet의 Mask 기능을 활용하여 개별 모델 내부에서 어떠한 변수가 실질적인 예측 근거로 작용했는지 파악하기 위해 그림 2, 그림 3 및 그림 4와 같이 모델별 변수 중요도를 분석하였다. 특히, 중장기적인 여론의 흐름과 경기 데이터를 포괄하고 있는 5일 모델의 변수 중요도 분석 결과에서는 그림 4의 붉은 색 막대에서 보여지는 바와 같이, 커뮤니티 감성 점수($avg_neg_score_5_days_away$ 등)가 상위권에 위치하고 있음을 확인할 수 있다. 이는 정량적인 경기 기록뿐만 아니라 팬들의 여론이 승패 예측에 유의미한 설명 변수로 작용함을 뒷받침한다.

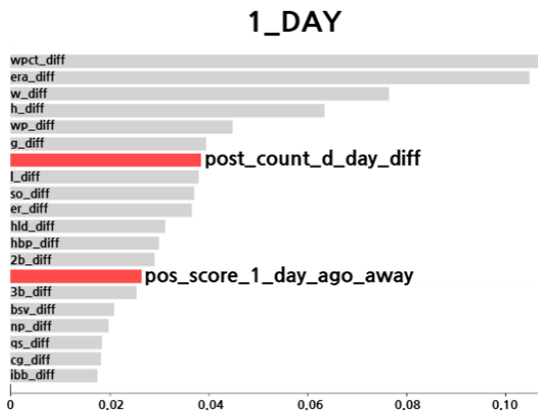


그림 2. 1일 관측 모델의 변수 중요도

Fig. 2. Feature importance of the 1-day model

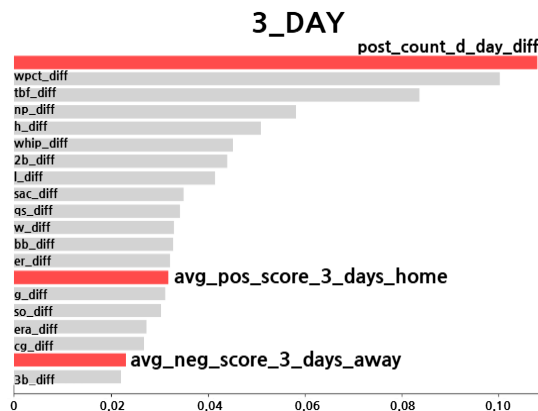


그림 3. 3일 관측 모델의 변수 중요도

Fig. 3. Feature importance of the 3-day model

5_DAY

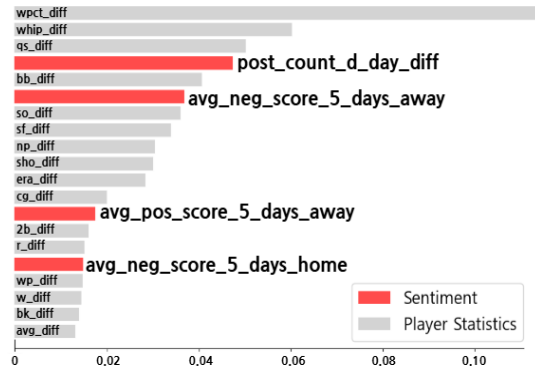


그림 4. 5일 관측 모델의 변수 중요도

Fig. 4. Feature importance of the 5-day model

더불어 게시글 수 차이($post_count_d_day_diff$)가 세 모델 그래프에서 상위 중요도를 기록한 점은 주목할 만하다. 이는 커뮤니티 내의 정보량이 많고 반응이 활발할수록 해당 여론 데이터가 승패 예측의 신뢰도를 높이는 근거로 작용했음을 시사한다. 이를 통해 커뮤니티 여론 분석이 단순한 요소를 넘어 스포츠 승패 예측 모델의 성능을 향상시키는 실질적인 도구로서 잠재력을 가지고 있음을 확인할 수 있다.

V. 결론 및 시사점

본 연구는 KBO 리그 경기 결과 예측에서 기존의 정형 데이터 중심 접근을 넘어, 팬 커뮤니티의 텍스트 데이터를 활용한 비정형 정보까지 분석 범위를 확장하였다. 구체적으로는 한국어 특화 언어모델인 KcBERT를 이용해 팬 게시글의 긍·부정 감성을 정량화하고, 이를 TabNet 기반의 스테킹 앙상블 모델에 결합함으로써 경기 결과 예측 성능을 향상시키고자 하였다.

이러한 접근은 기존 스포츠 승패 예측 연구가 주로 정량적 경기 기록과 통계 지표에 의존해 왔다는 점에서 차별성을 지닌다. 본 연구는 팬 커뮤니티에서 형성되는 비정형 텍스트 정보를 감성 지표로 변환하고 이를 기존 정형 데이터와 통합함으로써, 스포츠 예측 연구에서 데이터 융합 기반의 새로운 분석 프레임워크를 제시하였다는 의의가 있다. 특히 정형 스렛과 비정형 감성 지표를 함께 학습할 수 있도록 TabNet 기반의 앙상블 구조를 설계함으로써,

서로 다른 유형의 데이터를 효과적으로 반영할 수 있는 예측 모델을 구축하였다. 실험 결과, 팬 커뮤니티 감성 지표를 결합한 모델은 기존 정형 데이터 기반 모델보다 우수한 예측 성능을 보였다. 이는 팬 커뮤니티 여론이 경기 결과 예측에 유의미한 정보로 활용될 수 있음을 보여준다.

본 연구의 결과는 스포츠 경기 결과가 단순한 전력 지표의 합만으로 설명되기 어렵고, 경기 외적 분위기와 팬 여론 또한 복합적으로 작용할 수 있음을 보여준다. 이는 그동안 응원이나 잡담 수준으로 여겨지던 커뮤니티 데이터가 실제로는 경기력 예측에 기여할 수 있는 정보 자원임을 확인한 결과이며, 데이터 과학과 스포츠 산업의 융합 측면에서도 의미 있는 시사점을 제공한다. 따라서 향후 스포츠 구단의 전력 분석에서는 선수 및 팀의 객관적 기록뿐 아니라 팬 여론과 같은 외부 반응에 대한 모니터링 역시 보조적 지표로 고려할 필요가 있다.

한편, 본 연구는 정형 데이터 구성 측면에서 기존 승부 예측 연구의 일반적 방법론을 따랐으며, 전년도 시즌 누적 데이터를 활용하여 데이터의 일관성과 모델 학습의 안정성을 확보하였다. 그러나 데이터 수집의 현실적 제약으로 인해 선발 투수 중심의 제한적인 지표만 사용하였고, 불펜 투수 지표, 구장별 세부 스플릿, 무승부와 같은 야구 고유의 세부 정보를 충분히 반영하지 못했다는 한계가 있다. 향후 연구에서는 경기일자별로 갱신되는 선수 기록과 실시간 커뮤니티 반응을 함께 반영하여 승리 확률을 동적으로 제시하는 실시간 예측 모델로 확장할 필요가 있다. 또한 텍스트 데이터에 더해 경기 영상, 인터뷰 음성, 미디어 콘텐츠 등 다양한 비정형 데이터를 결합한다면 보다 정교한 다차원적 승부 예측 모델을 구축할 수 있을 것이다. 나아가 이러한 접근은 다양한 스포츠 종목으로도 확장 가능하며, 데이터 기반 의사결정 체계를 고도화하는 데 기여할 수 있을 것으로 기대된다.

References

- [1] M. Haghighat, H. Rastegari, and N. Nourafza, "A Review of Data Mining Techniques for Result Prediction in Sports", *Advances in Computer Science: an International Journal*, Vol. 2, No. 5, 7-12, Nov. 2013.
- [2] F. Godin, J. Zuallaert, B. Vandersmissen, W. D. Neve, and R. V. de Wall, "Beating the Bookmakers: Leveraging Statistics and Twitter Data for Predicting Soccer Results", *KDD Workshop on Large-Scale Sports Analytics*, pp. 1-4, Aug. 2014. <https://doi.org/10.13140/2.1.2168.0000>.
- [3] C. Kim and S. Jin, "Predicting Wins through Professional Baseball Statistics", *Journal of The Korean Data Analysis Society*, Vol. 16, No. 1, pp. 211-220, Feb. 2014.
- [4] S. Shin, K. Park, Y. Cho, and S. Coi, "A Study on Analyzing Factors Affecting the Outcome of Korean Professional Baseball Games: A Case of Samsung Lions", *Journal of The Korean Data Analysis Society*, Vol. 9, No. 4, pp. 2071-2083, Aug. 2007.
- [5] P. Kim, S. Lee, and S. Jeon, "Predicting the Outcomes of K-League Matches using Machine Learning Algorithms", *Journal of Digital Contents Society*, Vol. 25, No. 4, pp. 1027-1037, Apr. 2024. <https://doi.org/10.9728/dcs.2024.25.4.1027>.
- [6] E. Roh and J. Choi, "A Study of KBO Professional Baseball Game Prediction using Artificial Neural Networks", *Journal of The Korea Society of Information Technology Policy & Management*, Vol. 9, No. 1, pp. 335-338, Dec. 2017. <https://doi.org/10.36498/kbigdt.2020.5.2.77>.
- [7] C. S. Valero, "Predicting Win-Loss outcomes in MLB Regular Season Games - A Comparative Study Using Data Mining Methods", *International Journal of Computer Science in Sport*, Vol. 15, No. 2, pp. 91-112, Dec. 2016. <https://doi.org/10.1515/ijcss-2016-0007>.
- [8] C. C. Tung and M. C. Lee, "Exploring and Selecting Features to Predict the Next Outcomes of MLB Games", *Applied Sciences*, Vol. 12, No. 4, Art no. 2125, Feb. 2022. <https://doi.org/10.3390/app12042125>.
- [9] B. Kang, M. Huh, and S. Choi, "Performance

Analysis of Volleyball Games using the Social Network and Text Mining Techniques", Journal of the Korean Data And Information Science Society, Vol. 26, No. 3, pp. 619-630, May 2015. <http://dx.doi.org/10.746/jkdi.2015.26.3.619>.

- [10] M. Kim and H. Lee, "Baseball Match Prediction Combining Game Record Statistics with Natural Language Processing for News Articles", Journal of Digital Contents Society, Vol. 24, No. 5, pp. 1041-1047, May 2023. <https://doi.org/10.9728/dcs.2023.24.5.1041>.
- [11] J. Lee, "KcBERT: Korean Comments BERT", GitHub repository, Aug. 2020. <https://github.com/Beomi/KcBERT>.
- [12] J. Park, M. Lim, and P. Kim, "Fine-tuning Method to Improve Sentiment Classification Performance of Review Data", Smart Media Journal, Vol. 13, No. 6, pp. 44-53, Jun. 2024. <https://doi.org/10.30693/SMJ.2024.13.6.44>.
- [13] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", Proc. of the International Conference on Learning Representations, New Orleans, USA, pp. 1-18, May 2019. <https://doi.org/10.48550/arXiv.1711.05101>.
- [14] S. Ö. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning", Proc. of the AAAI Conference on Artificial Intelligence, California, USA, Vol. 35, No. 8, pp. 6679-6687, May 2021. <https://doi.org/10.1609/aaai.v35i8.16826>

노 민 우 (Minwoo Noh)



2021년 3월 ~ 현재 : 명지대학교
경영정보학과 학부과정
관심분야 : Agentic AI, 머신러닝,
자연어 처리, 클라우드 컴퓨팅

이 승 우 (Seungwoo Lee)



2021년 3월 ~ 현재 : 명지대학교
경영정보학과 학부과정
관심분야 : 스포츠 산업, 데이터
분석, AI

이 한 준 (Hanjun Lee)



2001년 2월 : 서울대학교
컴퓨터공학과(공학사)
2004년 2월 : 서울대학교
컴퓨터공학과(공학석사)
2016년 8월 : 고려대학교 경영학과
MIS 전공(경영학박사)
2020년 3월 ~ 현재 : 명지대학교
경영정보학과 부교수
관심분야 : 머신러닝, 자연어 처리, 정보시스템, 정보화
정책

저자소개

최 홍 규 (Hongkyu Choe)



2021년 3월 ~ 현재 : 명지대학교
경영정보학과 학부과정
관심분야 : 데이터 분석, 자연어
처리, Agentic AI, 데이터
사이언스