JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 5, pp. 1041-1047, May. 2023

Check for updates

경기 기록 통계와 뉴스 기사에 대한 자연어처리를 결합한 야구 승부 예측 시스템

김 민 종[1] · 이 현 아[2*]
[1]금오공과대학교 컴퓨터소프트웨어공학과 학사과정
[2*]금오공과대학교 컴퓨터소프트웨어공학과 교수

Baseball Match Prediction Combining Game Record Statistics with Natural Language Processing for News Articles

MinJong Kim[1] · Hyunah Lee[2*]

[1]Undergraduate Course, Department of Computer Software Engineering, Kumoh National Institute of Technology, Gumi, Gyeongbuk, 39177, Korea
[2*]Professor, Department of Computer Software Engineering, Kumoh National Institute of Technology, Gumi, Gyeongbuk, 39177, Korea

[요 약]

대량의 데이터가 잘 정의된 형태로 축적된 스포츠 분야는 기계학습을 적용하기에 매우 적합한 응용 분야로 이전 경기에 대한 통계에 기반하는 다양한 승부 예측이 시도되어왔다. 하지만 이 방식은 시시때때로 변화하는 선수들의 상태와 팀 분위기 등 경기 외적 요인을 반영할 수 없다. 본 연구에서는 기록의 스포츠라고 불리는 야구의 승부 예측에 이전 경기 통계와 함께 경기 전 뉴스 기사를 활용한 방법을 제안한다. 제안 시스템에서는 자연어처리 사전학습모델인 KoBERT를 활용하여 뉴스 기사의 긍부정 이진 분류와 뉴스 기사 임베딩 벡터 각각을 얻어 통계 정보에 추가로 적용하였다. 결과에서는 통계기반 시스템은 0.6508의 정확도를 보인 것에 비해, 뉴스 기사의 긍부정 이진 분류를 추가하여 0.7222, 기사 임베딩 벡터를 추가하여 0.7430의 정확도를 얻어, 자연어처리 도입으로 인한 성능 향상을 확인할 수 있었다.

[Abstract]

The field of sports, which provides a large amount of accumulated data in a well-defined form, is suited to machine learning applications, so match predictions based on those statistics have been attempted, but they cannot reflect external factors, such as the changing conditions of players and team. In this paper, we propose the use of pre-game news articles along with game statistics for baseball match prediction. In the proposed system, positive-negative binary classification of news articles and news article embedding vectors is achieved by the natural language processing pre-trained model KoBERT and added to statistical information for prediction. While the statistics-based system showed an accuracy of 0.6508, the accuracy rate obtained by adding binary classification of news articles was 0.7222 and that by adding the article embedding vector was 0.7430.

색인어 : 야구 승부 예측, 클래스 가중치, 자연언어처리, 뉴스 기사, 윈도우 크기

Keyword : Baseball Match Prediction, Class Weight, Natural Language Processing, News Articles, Window Size

http://dx.doi.org/10.9728/dcs.2023.24.5.1041

Received 20 March 2023; **Revised** 18 April 2023
Accepted 24 April 2023

***Corresponding Author; Hyunah Lee**

Tel: +82-54-478-7546
E-mail: halee@kumoh.ac.kr

Ⅰ. 서 론

최근 딥러닝의 기술 발전은 자연어 처리와 그 응용 분야의 비약적인 성능 향상에 기여하고 있다. 특히 대량의 데이터로부터 학습한 BERT[1]나 KoBERT[2]와 같은 사전학습 언어모델은 비정형 데이터인 텍스트로부터 정형 데이터를 효과적으로 추출하여 자연어가 가지고 있는 여러 중의성을 해결하는 장점을 가지며, 다양한 응용문제들에 적용되어 우수한 결과를 보여준다.

예측 시스템은 기계학습으로 실용적인 결과를 얻기에 적합한 분야이다. 풍부한 데이터가 축적된 스포츠 분야에서의 경기 통계에 기반한 승부 예측은 인공지능의 다양한 기법을 적용하여 그 유효성을 판단하기에 매우 좋다[3],[4]. 특히 야구는 기록의 스포츠라고 불릴 정도로 수집된 기록의 종류가 다양하고 그 양이 풍부하여 활용도가 높다. 이런 특징으로 야구 경기의 기록에 기반하여 승부를 예측하는 다양한 연구들이 진행되어, 통계 데이터 제공을 위한 연구[5], 경기 기록 데이터에 딥러닝 기술을 통한 승부 예측이나 선수가 보여줄 미래의 실적을 예측하는 연구[6]-[8] 등이 시도되어왔다.

하지만 이러한 연구들은 이전 경기들에서의 통계 정보인 스탯(stats)에 기반하는 방식으로, 시시때때로 변화하는 선수들의 상태와 팀 분위기 등의 경기 외적 요인을 반영할 수 없다. 본 연구에서는 비정형 데이터인 뉴스 기사 텍스트에 자연어 처리를 적용하여 정보를 추출하고 이를 기존 경기 기록 통계에 결합하는 방식으로 경기 외적인 요인을 야구 승부 예측에 반영하고자 한다. 자연어 처리를 통한 승부 예측에서는 긍부정 이진 분류와 문서 임베딩의 두 가지 방식을 적용하여 학습을 위한 입력 벡터를 추출한다.

경기 직전에 긍정적인 기사들이 많이 발행되는 팀은 다음 경기에서 좋은 경기 결과를 얻을 것으로 기대된다. 온라인 스포츠 뉴스는 기사에 대한 독자의 감성을 선택할 수 있는 '좋아요'나 '화나요' 등의 감성 태그를 제공하며 이 태그들의 통계는 해당 기사의 긍부정 여부를 판단하는데 활용할 수 있다. 제안하는 방식의 첫 번째 접근에서는 뉴스 기사에 부착된 감성 태그를 활용하여 뉴스 기사 텍스트에 대한 긍부정 이진 분류를 학습하고, 이로부터 획득한 팀별 긍정 점수를 통계 정보에 추가하는 방식으로 경기 외적인 요인을 반영한 승부 예측을 시도한다.

최근의 자연언어처리의 사전학습모델을 활용하면 문서의 텍스트 정보를 임베딩 벡터로 쉽게 변환할 수 있다[1]-[2]. 팀별 뉴스에 대한 임베딩 벡터는 팀 분위기를 내포하고 있을 것으로 기대되므로 이 또한 승부 예측에 좋은 정보가 될 수 있다. 제안하는 방식의 두 번째 접근에서는 팀별 뉴스 기사의 임베딩 벡터를 통계 정보에 추가하여 승부 예측에 경기 외적인 정보를 반영한다.

승부 예측에서 뉴스를 사용할 때 경기 당일에서부터 며칠 전까지 발행된 기사를 경기 외적인 요인으로 활용할지도 중요 고려사항이 될 수 있다. 제안 시스템에서는 다양한 윈도우 사이즈의 기사를 적용하여 경기 외적 요인이 승부 예측에 영향을 주는 지속 시간을 분석한다.

논문의 구성은 다음과 같다. 2장에서는 한국프로야구(KBO)의 승부 예측을 위한 통계 정보와 뉴스 기사 수집에 관하여 설명한다. 3장에서는 기존 경기 통계에 기반한 승부 예측 방식과 성능을 보이고, 4장에서는 제안하는 자연어 처리를 결합한 승부 예측 방식과 성능에 대해서 보인다. 5장에서는 결론과 향후 연구에 관해 설명한다.

Ⅱ. 승부 예측을 위한 데이터 수집

제안하는 시스템은 한국프로야구의 경기 통계 정보(stats)과 함께 야구 관련 뉴스 기사를 사용하여 승부를 예측한다.

	date	home	home score	away	away score	result	stadium	home pitcher	away pitcher
0	20200505	KIA	2	KW	11	LOSE	Gwangju	Yang Hyeon-Jong	Jakob Daniel Brigham
1	20200505	LG	8	DS	2	WIN	Jam-sil	Cha Woo-Chan	Alcántara
2	20200505	SS	0	NC	4	LOSE	Daegu	Baek Jung-Hyun	Drew James Rucinski
3	20200505	KT	2	LT	7	LOSE	Suwon	Odrisamer Despaigne Orue	Straily
4	20200505	SK	0	HH	3	LOSE	Incheon	Nicholas Gordon Nick Kingham	Warwick Anthony Saupold
...	...	...	...	...	...	...	...	...	...
1097	20210710	SS	3	LT	2	WIN	Daegu	Won Tae-In	Park Se-woong
1098	20210710	SK	1	HH	3	LOSE	Incheon	Oh Wonseok	Kim Minwoo
1099	20210711	SS	11	LT	0	WIN	Daegu	Baek Jung-Hyun	Straily
1100	20210711	KIA	2	KT	0	WIN	Gwangju	Lee Euilee	Odrisamer Despaigne Orue
1101	20210711	SK	8	HH	2	WIN	Incheon	Wilmer Font Gómez	Yun Daekyung

그림 1. 수집한 KBO 경기 통계의 일부
Fig. 1. Part of collected game stat(statistics) of KBO

표 1. 투수 스탯의 각 자질명과 자질에 대한 설명
Table 1. Name and its description of pitching stat

Name	Description
win	Number of wins
loss	Number of loss
save	Number of save as a relief pitcher
hld	Number of hold as a relief pitcher
blown	Number of tie or loss with save condition
game	Count of games
gs	Number of running as a game starting pitcher
ip	Number of inning pitched
k	Number of strikeout

Name	Description
bb	Number of walk (Based on Ball)
hr	Number of home run
babip	Batting Average on Balls In Play
lob	Left On Bases
era	Earned Run Average
ra	Run Average
fip	Fielding Independent Pitching
kfil	KBReport[10]'s Fielding Independent Pitching(fip)
war	Wins Above Replacement

	name	team	w	l	save	hld	blown	game	gs	ip	k	bb	hr	babip	lob	era	ra	fip	kfip	war
0	Park Se-woong	LT	0.125000	0.000000	0.000000	0.0	0.000000	0.500000	0.500000	2.650000	1.090000	0.415000	0.260000	0.037750	8.337500	0.778750	0.018750	0.723750	0.710000	0.020000
1	Lee In-Bok	LT	0.125000	0.000000	0.000000	0.0	0.000000	0.250000	0.125000	0.762500	0.888750	0.355000	0.000000	0.051125	9.087500	0.532500	0.015000	0.338750	0.320000	0.025000
2	Kim Won-Jung	LT	0.111111	0.222222	0.444444	0.0	0.111111	0.888889	0.000000	1.000000	1.111111	0.666667	0.333333	0.022222	7.577778	0.777778	0.011111	0.827778	0.818889	-0.001111
3	Enderson Daniel Franco	LT	0.000000	0.111111	0.000000	0.0	0.000000	0.444444	0.333333	1.666667	0.866667	0.400000	0.266667	0.036222	9.055556	0.600000	0.016667	0.697778	0.690000	-0.006667
4	Straily	LT	0.111111	0.111111	0.000000	0.0	0.000000	0.444444	0.444444	2.466667	1.235556	0.264444	0.176667	0.030111	10.222222	0.308889	0.107778	0.440000	0.385556	0.082222
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
303	Jakob Daniel Brigham	KW	0.400000	0.000000	0.000000	0.0	0.000000	0.400000	0.400000	2.600000	1.524000	0.276000	0.276000	0.042400	15.680000	0.692000	0.080000	0.914000	0.890000	0.046000
304	Joshua Allen Smith	KW	0.500000	0.000000	0.000000	0.0	0.000000	0.500000	0.500000	3.500000	1.285000	1.285000	0.645000	0.050000	41.650000	1.285000	0.150000	2.745000	2.895000	-0.005000
305	Park Gwanjin	KW	0.000000	0.000000	0.000000	0.0	0.000000	0.500000	0.000000	0.050000	0.000000	13.500000	0.000000	0.333500	16.650000	27.000000	-0.065000	6.170000	7.230000	-0.010000
306	Moon Sunghyun	KW	0.000000	0.000000	0.000000	0.0	0.000000	0.250000	0.000000	0.250000	0.000000	4.500000	0.000000	0.000000	25.000000	0.000000	0.007500	2.335000	2.715000	-0.007500
307	Cho Younggun	KW	0.000000	0.000000	0.000000	0.0	0.000000	1.000000	0.000000	1.000000	27.000000	18.000000	0.000000	1.000000	100.000000	0.000000	0.030000	3.340000	2.350000	0.010000

그림 2. 수집한 투수 스탯 데이터의 일부
Fig. 2. Part of collected stat(statistics) of pitchers

2-1에서는 경기 통계 데이터의 수집에 대해서, 2-2에서는 뉴스 기사 수집에 관하여 설명한다.

2-1 야구 경기 통계 데이터 수집

본 연구에서는 2020년 5월에서 2021년 7월까지의 KBO의 경기 통계를 사용한다. 그림 1에서는 추출된 데이터의 일부를 보인다. 데이터는 [경기 일자(date), 홈 팀(home), 홈 팀 점수(home score), 원정팀(away), 원정팀 점수(away score), 홈팀 기준 경기 결과(result), 구장(stadium), 홈 선발 투수(home pitcher), 원정팀 선발 투수(away pitcher)]의 아홉 개의 메타 정보로 구성된다. 경기 정보는 네이버 스포츠의 야구 카테고리의 일정 탭[9]에서 수집하였으며, 대상 기간의 총 1,102회의 경기에서 선발 투수 정보가 없거나 결측치가 존재하는 경기를 제외하고 765회의 경기를 얻었다. 야구 경기의 특성상 무승부는 매우 드물게 발생하며 승부 예측에서는 무승부보다는 승패에 관한 판단이 중요하므로, 수집된 경기 중 무승부는 제외하고 승패 결과가 얻어진 757회의 경기에서 9개의 메타 정보를 수집하여 저장한다. 수집된 정보에 기반하여 각 팀의 승률과 상대 전적을 계산하여 별도로 저장한다.

승부에 가장 중요한 정보로 여겨지는 선발 투수에 대한 스탯도 수집하였다. KBReport.com(케이비리포트)[10]에서는 각 투수가 뛴 경기별로 표 1과 같은 18개의 정보를 제공한다. 각 투수가 출전한 모든 경기에 대한 스탯의 평균으로 해당 투수에 대한 스탯 데이터를 생성하였으며, 그림 2와 같이 해당 기간에 KBO에서 활동한 307명의 투수에 대한 스탯을 수집하였다.

2-2 구단별 뉴스 기사 수집

네이버 스포츠 야구 카테고리의 최신 뉴스 탭[11]에서 수집된 경기 통계와 같은 기간의 뉴스 기사를 수집하였다. 한 기사 안에 여러 구단에 대한 정보가 혼재된 뉴스 기사에서는 승부 예측의 대상이 되는 팀에 대한 정보를 획득하기 어렵지만, 네이버 야구 뉴스에서는 그림 3과 같이 구단별 검색 옵션을 제공하여 구단별 기사를 용이하게 수집할 수 있다. 해당 옵션을 통해 화보 기사를 제외하고 10개 구단에 관하여 기사 본문 텍스트가 제공되는 총 40,674개의 뉴스 기사를 수집하였다. 그림 4는 구단별 뉴스 개수를 보인다. 구단별 평균 뉴스 기사의 개수는 4067.4개로 나타났으며, 해당 기간에서는 키움(KW)의 기사가 가장 많이 발행되어 5,313개가 수집되었으며, NC의 기사가 가장 적게 발행되어 2,207개가 수집되었다.

기사 본문에 존재하는 특수 문자나 기사 마지막 부분의 기자 이름과 이메일 주소와 같이 경기 결과에 영향을 미치지 않는 문자열은 정규표현식을 이용하여 전처리하여 제거하였으

며, 수집된 기사의 평균 길이는 243.23어절로 나타났다.

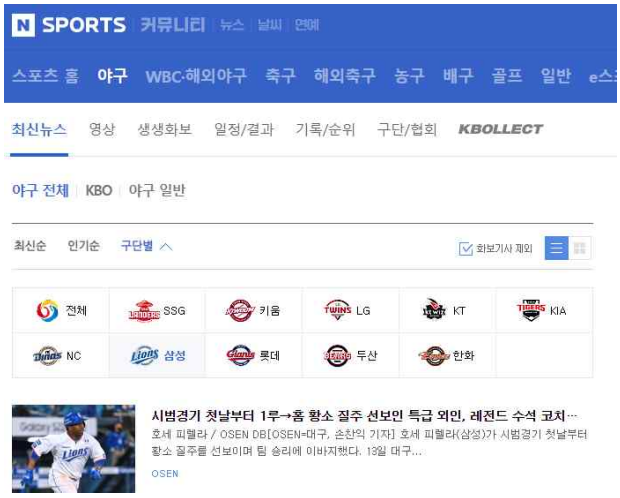

그림 3. 네이버 스포츠 야구 최신 뉴스의 구단별 뉴스 화면
Fig. 3. Naver Sports baseball news screen for each team

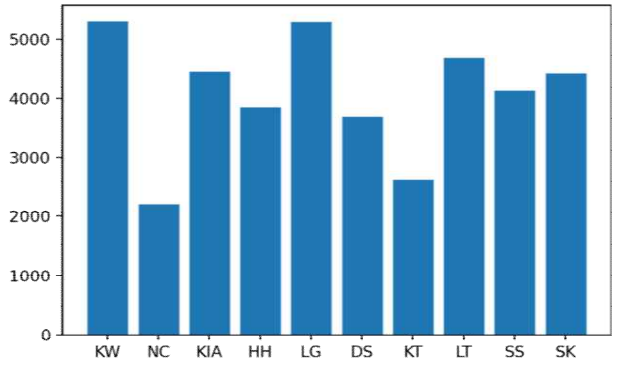

그림 4. 구단별 뉴스 기사 개수
Fig. 4. Number of articles by club

Ⅲ. 통계 데이터에 기반한 승부 예측 시스템

본 장에서는 KBO 경기 통계만을 활용한 승부 예측 시스템의 구축 방식을 소개한다. 특히, 클래스 가중치를 적용하여 승부 예측을 진행하는 방법을 소개한다.

3-1 입력 벡터의 구성

기존의 승부 예측 연구에서는 특정 팀의 경기 전적을 사용하여 승부를 예측하는 방식을 대부분 사용하고 있으나, KBO에서 각 팀은 특정 팀과 한 시즌에 16번의 경기를 진행하여 팀별 전적을 사용하는 방식은 학습에 필요한 만큼의 충분한 경기를 시즌별로 얻을 수 없다. 만일 정보 부족 문제를 해결하기 위해 여러 시즌의 통계를 연결하여 사용한다면 시즌마다 달라지는 감독과 선수에 의해 정보의 정확성이 부정확해질 수 있다.

스포츠 경기는 각 팀의 소속 지역에서 경기를 치러 홈팀과 원정팀을 구분할 수 있다. 특정 두 팀 간(예를 들어 키움 대 SSG) 승부를 예측하기 위해서는 KBO 10개 구단의 45개 쌍에서 각각 15개 이하의 경기 정보만 사용할 수 있지만, 승부 예측을 홈팀과 원정팀 간 승패를 판단하는 문제로 보고 각 팀과 선발 투수의 스탯을 사용하면 더욱 풍부한 데이터를 승부 예측에 사용할 수 있다. 이에 착안하여 제안하는 방식에서는 특정 팀 간의 승부를 판정하는 방식이 아닌, 홈팀을 기준으로 승패를 판단하여 데이터 부족 문제를 완화한다.

야구에서는 투수의 능력이 경기 승패에 가장 큰 영향을 미치는 요소로 꼽힌다. 이를 반영하여 승부 예측을 위한 입력 벡터는 홈팀의 승률과 원정팀의 승률, 홈팀의 원정팀 상대 전적을 백분율로 계산한 3개의 값에, 홈팀의 선발 투수와 원정팀의 선발 투수의 18개의 스탯을 이어 붙인 전체 길이 39의 벡터를 사용한다.

3-2 다중 분류 모델과 클래스 불균형 가중치 적용

학습에는 Multi-Classification을 이용하고 카테고리는 홈 기준 [승리, 패배]로 구성한다. epoch는 100, batch_size는 16, learning_rate는 0.001, 손실함수는 cross entropy loss를 사용하였고 옵티마이저는 adamW를 사용한다. 입력층의 차원이 (39, 1024)인 선형 신경망에 sigmoid 활성화 함수를 적용하고 과적합 방지를 위한 0.5 계수를 가진 dropout 레이어와 출력층에는 차원이 (1024, 2)인 선형 신경망에 softmax 활성화 함수를 적용한 신경망을 구성하였다. 학습 데이터는 568경기, 검증 데이터는 189경기로 구성하였다.

승부 예측의 검증과 평가에서는 대상 경기에서 홈팀의 승리나 패배를 정확하게 예측한 비율을 지표로 평가하며 본 논문에서는 이를 정확도로 표현한다. 학습 결과 검증 단계에서 최고성능으로 0.5714의 정확도를 얻었으며 그 이상의 값은 나타나지 않았다. 검증 데이터에 대해 가중치를 적용하지 않은 결과의 혼동 행렬은 표 2와 같았다.

표 2. 클래스 가중치 미적용 시 검증 결과 혼동 행렬
Table 2. Confusion matrix of result without class weight

Real / Prediction	Win	Lose
Win	54	38
Lose	43	54
Total	97	92

표 3. 클래스 가중치 적용 시 검증 결과 혼동 행렬
Table 3. Confusion matrix of result with class weight

Real / Prediction	Win	Lose
Win	58	32
Lose	34	65
Total	92	97

야구 경기는 홈팀이 홈구장에서 경기할 때 승률이 더 높다. 수집된 전체 데이터에서 총 홈팀의 승리 데이터는 401개, 패배 데이터는 356개로 승리 클래스가 패배 클래스보다 많으므로 클래스 불균형이 발생한다. 수집한 데이터에서 승리 클래스의 개수가 패배 클래스보다 더 많으므로 클래스별 가중치 적용 여부에 대한 성능을 비교한다.

클래스 가중치 적용에서는 전체 데이터 개수를 각 클래스의 개수로 나누어 클래스 가중치를 생성하였다. 승리, 패배의 가중치는 학습 과정에서 승리에 1.8381, 패배에 2.1930을 부여했다. 학습 결과 검증 단계에서 최고성능은 0.6508의 정확도를 보여주었으며 그 이상의 값은 나타나지 않았다. 표 3은 검증 데이터에 가중치를 적용한 결과의 혼동 행렬을 보인다. 혼동 행렬에서는 클래스를 균등하게 구성하기 위해 가중치를 적용하여 재구성된 데이터에 의한 결과를 보인다.

표 4는 가중치 적용 유무에 따른 정확도를 비교한다. 결과에서 클래스 가중치를 적용하는 것이 적용하지 않았을 때보다 성능이 높게 나타나, 제안하는 시스템에서는 클래스 가중치를 적용한 결과를 기준으로 승부 예측을 진행한다.

표 4. 클래스 가중치 적용에 따른 통계 기반 승부 예측 정확도
Table 4. Accuracy of stat-based match predictions with or without class weight

without class weight	0.5714
with class weight	0.6508

Ⅳ. 자연어 처리를 적용한 승부 예측 시스템

통계 정보를 기반으로 승부를 예측하는 경우 팀별 부상선수나 팀 분위기와 같은 경기 외적인 요인을 반영할 수 없다. 뉴스 기사는 이러한 외적 요인을 제공하는 좋은 정보원이 되지만 비정형 데이터의 한계로 승부 예측에 직접 사용하기 쉽지 않다. 이번 장에서는 뉴스 기사에 대한 자연어 처리를 통해 경기 외적 요인을 추가 반영한 승부 예측 시스템의 구축 방식을 설명하고 통계 정보만을 사용한 결과와 비교한다.

뉴스 기사의 활용에서는 경기 직전 며칠까지의 뉴스를 사용할지에 대해 다양한 window size를 적용하고, 기사 텍스트 정보에서 얻은 홈팀과 원정팀의 긍정 점수를 통계 스탯에 결합하는 방식과 기사 텍스트의 문서 임베딩을 직접 통계 스탯에 결합하는 두 가지 방식을 적용하고 비교한다. window size는 경기 이전 3일과 5일, 7일을 적용하여 실험한다. 아래에서는 기사 텍스트에서 긍정, 부정 이진 분류를 활용한 예측과 문서 임베딩을 활용한 예측 각각을 설명한다.

4-1 감성 태그 기반 뉴스 기사 긍부정 감성 분류 적용 모델

네이버 스포츠 뉴스 야구 카테고리의 최신 뉴스 탭[11]은 구단별 뉴스 기사에 대해서 그림 5와 같이 독자들이 기사를 읽고 의견을 표현한 감성 태그 통계가 제공된다. 긍정 태그가 많이 부착된 구단일수록 다음 경기에서 좋은 경기 결과를 기대할 수 있다. 하지만 이러한 태그는 모든 기사에 대해 충분한 숫자로 부착되지는 않는다.

본 연구에서는 네이버 스포츠 기사의 감성 태그 중 [좋아요(like), 팬이에요(I'm a fan)]를 긍정으로, [슬퍼요(sad), 화나요(angry)]를 부정으로 간주하고, 긍정 태그들의 수가 부정 태그들의 수보다 크면 해당 기사를 대상 구단에 대한 긍정 성향으로 판단하여 각 기사의 긍부정을 태그한다. 태그된 기사들이 수집되면 학습을 통해 감성 태그가 부착되지 않은 기사에 대해서도 긍부정 분류 점수를 얻을 수 있다. 제안 시스템에서는 각 팀에 관한 기사로부터 팀별 긍정 점수를 얻고 이를 기존 경기 통계에 결합하여 승부 예측을 시도한다.

제안 시스템에서는 KoBERT 언어모델에 뉴스 기사의 전문을 입력으로 하여 긍부정으로 분류하는 이진 분류를 fine-tuning하여 사용한다. fine-tuning에 필요한 데이터로는 2022년 4월부터 6월까지 네이버 스포츠 뉴스의 야구 카테고리에서 수집된 23,193개 기사 중 긍정 데이터 1만 8,020개, 부정 데이터 5,173개를 사용한다. 수집한 데이터의 17,394개는 학습 데이터로 사용하고 나머지 5,799개는 검증 데이터로 나누어 사용한다. 긍정 데이터가 부정 데이터보다 더 많으므로 분류기 학습에도 클래스 가중치를 적용한다. 학습에 적용한 가중치는 기존 클래스 가중치 적용 방법과 같은 방법을 이용하였다. 학습 데이터에 긍정 데이터는 13,472개, 부정 데이터는 3,922개로 가중치는 긍정에 1.2911, 부정에

News Text	Reader Sentiment Tag Statistics
… KT is suffering from a pitching gap in the first week of opening. The losing streak at the beginning of the season and the sense of deja-vu during past 3 years are great… (… KT가 개막 일주일 만에 투타 엇박자에 시달리고 있다. 지난 3년 동안 발목을 잡은 시즌 초반의 연패와 기시감이 크다 …)	
…SSG Landers rookie left-hander Oh Won-seok won the match against KT Wiz native ace Koh Young-pyo. After the opening, SSG ran a 4-game winning streak… (…SSG랜더스 신예 좌완 오원석이 KT위즈 토종에이스 고영표와 맞대결에서 완승했다. SSG는 개막 후 파죽의 4연승을 질주했다.…)	

*To convey the original tone of the original newspaper article, Korean texts and its translations are paralleled in the left column.
그림 5. 네이버 스포츠 뉴스 기사의 독자 감성 태그 통계
Fig. 5. Reader sentiment tag statistics of Naver Sports News articles

4.4349를 적용하였다.

학습에서 optimizer는 adamW를 사용하고 손실함수는 Binary Cross Entropy를 이용하였다. 학습률은 1e-5로 설정하였고 batch size는 16, epoch는 10으로 설정하고 binary-classification task로 학습을 진행하였다. 학습 결과 이진 분류에 대해 0.9285의 정확도를 얻었다.

학습한 KoBERT 긍부정 분류기를 이용하여 해당 경기 일자 이전의 홈 팀, 원정팀 각각의 뉴스 기사를 긍정, 부정으로 분류하여 긍정 데이터 비율을 계산한다. 얻어진 홈 팀의 긍정 비율과 원정팀의 긍정 비율 두 개의 값을 통계 기반 승부 예측 시스템의 입력 벡터에 추가한다. 긍부정 분류기를 적용한 텍스트 기반 승부 예측 시스템은 입력층을 (41, 1024)로 변경한 것 이외에는 모두 통계 기반 시스템과 같다.

표 5는 3-4절의 클래스 가중치를 적용한 통계 기반 승패 이진 분류 모델에 기사 텍스트의 긍부정 이진 분류를 추가한 모델의 window_size에 따른 정확도를 보인다. 통계 기반 모델의 최고 정확도 0.6508에 비해 큰 성능 향상을 얻을 수 있었으며, window_size가 5일 때 가장 높은 정확도를 보였다.

표 5. 기사 텍스트 긍부정 이진 분류를 결합한 승부 예측의 WINDOW_SIZE별 정확도

Table 5. Accuracy by WINDOW_SIZE of match prediction combining positive-negative binary classification of news text

WINDOW_SIZE	3 days	5 days	7 days
Accuracy	0.7190	0.7222	0.6815

4-2 뉴스 기사 문서 임베딩 적용 모델

기사 텍스트에서 얻은 홈팀과 원정팀의 긍정 기사 비율을 적용하는 방식은 기사에 내포된 다양한 정보를 반영하는데 한계가 있다. 이에 비해 뉴스 기사의 전체 텍스트 정보를 승부 예측에 사용하면 더 정확한 승부 예측을 기대할 수 있다.

자연어 처리에서 단어 임베딩은 단어의 표현을 실숫값 벡터로 나타내는 것을 의미한다. 생성된 단어 벡터를 이용하여 딥러닝의 신경망 가중치를 학습하거나 문장 벡터, 문서 벡터를 만들 수 있다. 이 방식으로 얻은 기사의 문서 벡터를 직접 승부 예측에 적용하면 기사 전체의 풍부한 정보를 승부 예측에 활용할 수 있다.

임베딩 방법에는 Word2Vec, FastText 등 다양한 방법론이 존재하는데 제안하는 시스템에서는 BERT[1]의 한국어 언어모델 KoBERT[2]를 사용한다. KoBERT는 단일 언어모델이 다국어 언어모델보다 성능이 더 향상되는 실험 결과[12]를 토대로 다국어 Bert 모델의 한국어 성능 한계를 극복하기 위해서 개발되었다. Word2Vec나 FastText는 임베딩 레이어를 랜덤하게 초기화한 뒤 대량의 데이터에 임베딩 알고리즘을 이용하여 사전 학습된 단어 임베딩을 생성한다. 이 방법은 태스크에 사용하기 위한 데이터가 적다면 성능 향상을 기대할 수 있지만, 단어가 하나의 벡터에 대응되어 문맥이나 다의어, 동음이의어를 고려하지 못한다는 단점이 존재한다. 하지만 BERT는 셀프 어텐션을 이용하여 입력을 구성하는 모든 단어를 참고하므로 문맥을 반영한 임베딩을 생성하는 장점으로 기존 임베딩 방식에 비해 큰 성능 향상을 보인다.

KoBERT base-v1 기준 문서 임베딩 벡터의 크기는 (1, 512, 768)이 생성된다. 생성된 벡터의 크기가 너무 크기 때문에 학습 시간이나 과적합 문제가 생길 수 있어 이를 해결하기 위해 제안 시스템에서는 문서 벡터의 평균에 Pooling 연산을 이용한다. 2차원 Pooling은 2차원 필터를 이동시키며 필터가 적용되는 영역 내의 특징들을 요약하는 작업을 의미한다. 주로 너무 많은 특징을 출력하는 CNN(convolutional neural networks)에서 학습할 매개변수의 수와 네트워크에서 수행되는 계산을 줄이는 데 사용되며 특정 영역에 있는 특징을 요약하는 데도 사용된다[13]. Pooling을 사용하면 연산의 횟수는 줄어들지만, 성능 감소는 거의 발생하지 않는다.

제안 시스템에서는 2차원 Average Pooling을 사용하고 커널 사이즈는 (469,768) stride = 1로 설정하여 Pooling 연산의 최종 결과 문서 임베딩 벡터의 크기를 (1, 44)로 만든다. Pooling을 이용한 텍스트 기반 승부 예측 시스템은 홈팀, 원정팀 각각 뉴스 기사의 임베딩 벡터에 Pooling을 적용한다. 입력 차원은 경기 스탯의 39개 값에 홈팀과 원정팀 각각의 임베딩 벡터 44개 값을 연결한 (127, 1024)로 변경하고 이외 모델 구조 및 학습 파라미터는 통계 기반 시스템과 같다.

표 6은 window_size에 따른 Pooling 연산을 이용한 텍스트 기반 승부 예측 시스템의 정확도를 보인다. 통계 기반 모델의 최고 정확도 0.6508에 비해 전체적으로 향상된 성능을 나타냈으며, window_size=5일 때는 긍부정 이진 분류를 사용한 결과의 0.7222보다 높은 0.7430의 정확도를 보여 임베딩 모델을 적용하여 더 정확한 승부 예측 결과를 얻을 수 있었다.

표 6. 문서 임베딩 벡터에 Avg Pooling을 적용한 텍스트 기반 승부 예측의 WINDOW_SIZE별 정확도

Table 6. Accuracy by WINDOW_SIZE of text based prediction using document embedding vector with average pooling

WINDOW_SIZE	3 days	5 days	7 days
Accuracy	0.6942	0.7430	0.7111

긍부정 이진 분류와 문서 임베딩의 모든 결과에서 window size에 따른 성능은 5일 때 가장 높았다. window size가 3일 때는 충분한 경기 외적 요인을 반영하지 못하고, window size가 7인 경우에는 이미 지난 경기 외적 요인까지 반영하여 정확도가 더 낮아진 것으로 보인다.

V. 결론

본 시스템은 경기 외적인 요인을 고려하여 야구 경기 승부 예측의 정확도를 높이는 방향으로 경기 이전의 뉴스 기사에 자연어 처리를 적용할 것을 제안하였다. 결과에서는 경기 스탯과 같은 정형 데이터와 함께 뉴스 텍스트의 비정형 데이터를 사용하여 성능 향상을 확인할 수 있었다.

본 논문에서는 딥러닝의 신경망을 선형으로만 구성하였다. 향후 연구로는 다양한 신경망을 적용하는 동시에 언어모델을 야구 도메인에 맞게 사전 학습하는 방식을 진행할 예정이다. 또한 야구뿐만 아니라 축구, 배구 혹은 스포츠 이외의 외적 요인을 고려할 수 있는 분야로 확장하는 방향으로 진행할 예정이다.

감사의 글

이 연구는 금오공과대학교 학술연구비로 지원되었음(2021년).

참고문헌

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv Preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

[2] SKT KoBERT [Internet]. Available: https://github.com/SKTBrain/KoBERT

[3] B. Min, J. Kim, C. Choe, H. Eom, and R. I. McKay, “A Compound Framework for Sports Results Prediction: A Football Case Study,” *Knowledge-Based Systems*, Vol. 21, No. 7, pp. 551-562, October 2008. https://doi.org/10.1016/j.knosys.2008.03.016

[4] A. McCabe and J. Trevathan, “Artificial Intelligence in Sports Prediction,” in *Proceedings of Fifth International Conference on Information Technology: New Generations*, Las Vegas, NV, USA, pp. 1194-1197, April 2008. https://doi.org/10.1109/ITNG.2008.203

[5] S. R. Bailey, J. Loeppky, and T. B. Swartz, “The Prediction of Batting Averages in Major League Baseball,” *Stats*, Vol. 3, No. 2, pp. 84-93, April 2020. https://doi.org/10.3390/stats3020008

[6] K. Jeong, J. Kim, and Y. Han. “A Prediction of Baseball Game Results Using Recurrent Neural Networks,” in *Proceedings of the Korea Information Processing Society Conference*, Vol. 24, pp. 873-876, 2017. https://doi.org/10.3745/PKIPS.y2017m11a.873

[7] A. S. Yaseen, A. F. Marhoon, and S. A. Saleem, “Multimodal Machine Learning for Major League Baseball Playoff Prediction,” *Informatica*, Vol. 46, No. 6, 2022. https://doi.org/10.31449/inf.v46i6.3864

[8] H. C. Sun, T. Y. Lin, and Y. L. Tsai, “Performance Prediction in Major League Baseball by Long Short-term Memory Networks,” *International Journal of Data Science and Analytics*, Vol. 1, No. 12, 2022. https://doi.org/10.48550/arXiv.2206.09654

[9] Naver Sports KBO Schedule [Internet]. Available: https://m.sports.naver.com/kbaseball/schedule/index?category=kbo

[10] Baseball Record Room KB Report [Internet]. Available: http://www.kbreport.com/

[11] Naver Sports KBO News [Internet]. Available: https://sports.news.naver.com/kbaseball/news/index?isphoto=N&type=team&team=kbo

[12] Google-research’s Fine-tuning Result for BERT [Internet]. Available: https://github.com/google-research/bert/blob/master/multilingual.md

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communication of the ACM*, Vol. 60, No. 6, pp. 84-90, May 2017. https://doi.org/10.1145/3065386

김민종(MinJong Kim)

2018년～현 재 : 금오공과대학교 컴퓨터소프트웨어공학과 학사과정

※관심분야 : 인공지능(Artificial Intelligence), 자연어처리(Natural Language Processing) 등

이현아(Hyunah Lee)

1996년 : 연세대학교 컴퓨터과학과 (학사)
1998년 : KAIST 전산학과 (석사)
2004년 : KAIST 전산학과 (박사)

2000년～2004년: ㈜다음소프트 언어처리연구소
2004년～현 재: 금오공과대학교 컴퓨터소프트웨어공학과 교수

※관심분야 : 자연언어처리, 텍스트데이터마이닝, 정보검색 등