딥러닝과 머신러닝을 활용한 독자 반응 기반 웹툰 데뷔작 성공 예측 모델

허은영*, 김승화*, 김현희*
*동덕여자대학교 정보통계학과
e-mail: sillver.411@gmail.com,
tmdghk312@gmail.com,
heekim@dongduk.ac.kr

A Success Prediction Model for Debut Webtoon Based on Reader reaction Using Deep Learning and Machine Learning

Eun Yeong Heo*, Seung Hwa Kim*, Hyon Hee Kim*
*Dept. of Statistics and Information Science, Dongduk Women's University

요 약

본 논문에서는 매년 성장하는 웹툰 시장 속에서 신인 작가들이 성공할 수 있는 성공 요인을 밝히고자 하였다. 국내 1위 웹툰 플랫폼인 네이버 웹툰 중 데뷔작을 기준으로 완결 웹툰 212개, 연재 중인 웹툰 112개, 총 324개의 웹툰을 수집하여 연구를 진행하였다. 기존 선행연구와의 차별화를 두기 위해 독자의 직접적인 반응 중 하나인 댓글을 성공 요인에 포함하였다. 댓글에 담긴 긍정, 부정을 나타내는 주관을 탐지하기 위해 딥러닝을 이용하여 감성 분석을 실시하였다. 각 웹툰에 대한 댓글 반응을 포함하여 평균, '좋아요' 수, 장르 그리고 첫 화 댓글 수와 5화까지 평균 댓글 수를 흥행에 영향을 미치는 독립변수로 사용했다. 댓글 반응이 중요 요인인지를 확인하기 위해 각 모델 생성 시 댓글 반응을 포함한 모델과 포함하지 않은 모델을 생성하여 성능 평가를 실시하였다. 로지스틱 회귀분석, 아다 부스트, 그리고 서포트 벡터 머신 모델을 정확도와 ROC 그래프를 이용해 효율성을 비교하고, 이를 통해 댓글 반응을 활용한 로지스틱 회귀 모델이 가장 적합하다고 판단하였다. 모델 생성 결과 '좋아요' 수, 1화 댓글 수, 댓글 반응 순으로 성공 요인에 많은 영향을 미치는 것을 알 수 있었다.

1. 서론

네이버 완결 웹툰을 조회 순으로 정렬했을 때, 상위 100개의 웹툰 중 신인 작가의 데뷔작은 31개로 기존 작가들에 비해 2.2배 정도 적은 걸 확인 할 수 있었다. 선행연구에 따르면 웹툰 흥행에 미치는 영향력은 One Source Multi-Use(OSMU), 독자 평점, 작가파워 순으로 나타났다. 작가파워란 작품의 수가 많은 작가일수록 인지도가 높고 그에 따른 영향력이 클 거라는 전제를 바탕으로 전작의 작품 수를 나타내는 독립변수이다. 이러한 변수가 높은 영향력을 미친다는 것은 신인 작가들이 성공하기 어렵다는 것을 말한다. 따라서 본 논문은 작가파워를 가지지 않는 신인 작가들의 웹툰 데뷔작 성공 예측 모델을 만들고, 이를 통해 흥행에 미치는 요인의 순서를 재정립하였다.

웹툰 댓글에서 많이 사용되는 단어들 중 다수는 일반적인 감성 분석 사전으로는 판별하기 어려운 단어가 많다. 웹툰을 결제하는 수단인 '쿠키'를 사용할 때 '쿠키를 굽는다'라고 표현한다. 이 때 '굽는다'는 일반적인 감성 사전에서 중립을 나타내는 단어이다. 또한, 웹툰의 업데이트가 다른 웹툰보다 월등히 빠를 때 사용되는 '베였다'는 일반적으로 부정 단어라고 생각한다. 하지만 이와 같은 단어들은 웹툰 댓글에서 긍정을 의미한다. 뿐만 아니라 감성분석 사전에 존재하지 않는 '베도(베스트도전)'나 '최강자전'과 같은 단어들도 웹툰 댓글에서는 긍정의 의미로 자주 사용된다. 이러한 문제를 해결하기 위해 웹툰 댓글을 직접 수집하여 새로운 감성 분석 사전을 생성하였다. 이를 기반으로 딥러닝을 활용해 독자들의 반응을 수치화하였다.

독자들의 반응을 직접적으로 알 수 있는 댓글 점수를 이용하여 웹툰 성공 예측 모델을 생성하였다. 이 모델은 독자의 반응을 고려하지 않은 모델에 비해 더 높은 정확도를 가진다. 이를 통해 독자의 직접적인 반응은 웹툰의 흥행에 미치는 중요한 요인이라는 것을 알 수 있다.

2장에서는 데이터 수집 방법에 대해 알아보고, 3장에서는 딥러닝을 활용한 댓글 내용 분석의 자세한 내용을 서술했다. 4장 로지스틱 회귀분석, 아다 부스트, 그리고 서포트 벡터 머신 모델에 대한 설명과 정확도를 서술했으며, 마지막으로 5장에서는 ROC 커브 그래프와 AUC를 이용하여 가장 적합한 모델 선정 및 평가, 그리고 결론에는 웹툰 데뷔작 흥행에 가장 많은 영향을 미치는 요인을 기술하였다.

2. 데이터 수집

국내 최대 웹툰 플랫폼인 '네이버 웹툰'에서 324개의 웹툰에 대한 데이터를 수집했다. 다른 플랫폼에서 정식으로 연재한 적이 없는 그림 작가들의 데뷔작을 기준으로 웹툰을 선정했다. 네이버 웹툰은 조회 수를 공개하지 않기 때문에 7월 30일 17:00을 기준으로 그에 해당하는 순위와 작품 '좋아요' 수, 작품의 전체 평점, 장르, 첫 화 댓글 수, 5화까지의 댓글 수 평균을 수집했다.

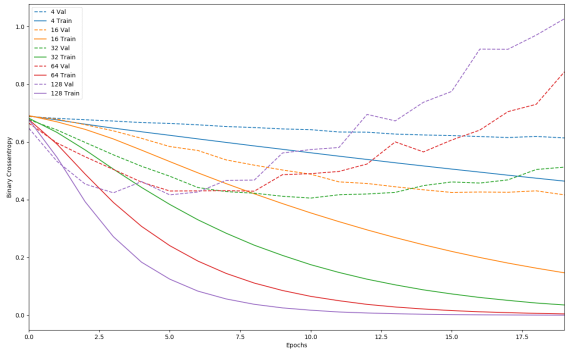
또한, 본 논문은 일반적으로 활용하는 수치 데이터가 아닌, 독자의 직접적인 반응이라고 할 수 있는 댓글 데이터를 수집하였다. 댓글은 각 편의 댓글이 1,000개 미만일 경우 모든 댓글을, 1,000개 이상일 경우 약 1,000개의 댓글을 시간 순으로 수집하였다. 8월 7일 13:00을 기준으로 하여 웹툰 당 5편씩, 즉 작품별로 최대 5,000개의 댓글을 수집하였다. 크롤링에는 파이썬의 BeautifulSoup 라이브러리를 사용하였다.

댓글 반응을 분석하기 위한 훈련 데이터와 테스트 데이터는 완결 웹툰의 댓글에서 수집하였다. 웹툰의 처음 5편에서 임의로 댓글을 25개씩 수집한 후 8:2의 비율로 훈련 데이터와 테스트 데이터를 나누었다.

3. 딥러닝을 활용한 독자 반응 분석

relu 활성화 함수를 사용하여 완전 연결층을 쌓는 것은 입력 데이터가 벡터값이고 레이블이 1 또는 0인 문제를 해결하는 데에 가장 적합하다. Dense 층에 전달한 매개변수는 은닉 유닛의 개수로, 하나의 은닉 유닛은 층이 나타내는 표현 공간에서 하나의 차원이 된다. 은닉 유닛을 늘리면 신경망이 더욱 복잡한 표현을 학습할 수 있다는 장점이 있지만 동시에 계산 비용이 커지고 원하지 않는 패턴을 학습할 수도 있다. 적당한 은닉 유닛 개수를 찾기 위해 각각의 은닉 개수에 따른 훈련과 검증 손실 그래프를 그려보았다 <그림 1>. 그 결과 128개와 64개가 여섯 번의 훈련 만에 손실도가 최소화되는 것을 확인 할 수 있다. 따라서 64개의 은닉 개수를 갖도록 하고 여섯 번의 에포크 동안만 훈련하도록 설정하였다.

<그림 1> 은닉층 수에 따른 모형 정확도와 오차



훈련 결과 약 0.627의 손실과 0.819의 정확도를 갖는 모델을 생성하였다. 토큰화한 댓글을 예측하여 작품들의 각 댓글이 긍정 댓글인지 부정 댓글인지를 점수화하였다. 예측 점수를 기준으로 0.5보다 클 경우 긍정 점수에 1점을, 0.5보다 작을 경우 부정점수에 1점을 더하여 해당 작품이 긍정적인 반응을 갖는지 부정적인 반응을 갖는지를 나타내었다. 하지만 0.5를 기준으로 긍정과 부정을 나눌 경우 중립적인 반응의 댓글임에도 불구하고 긍정 확률, 혹은 부정 확률로 계산되는 댓글을 볼 수 있다. 그로 인해 긍정 점수와 부정 점수의 차가 비정상적으로 커지는 경우가 있었다. 이 문제를 해결하기 위해 점수가 0.7보다 클 경우는 긍정 점수에 1점을, 0.5보다 크고 0.7보다 작을 경우는 긍정 점수에 0.5점을 부여하는 방식으로 계산하였다.

$$\begin{cases} score > 0.7 \ then \ pos + 1 \\ 0.5 < score \ \leq 0.7 \ then \ pos + 0.5 \\ 0.3 < score \leq 0.5 \ then \ neg + 0.5 \\ score \ \leq 0.3 \ then \ neg + 1 \end{cases} \quad (1)$$

이 계산식을 요약하면 수식(1)과 같다. 이러한 과정을 통해 각 작품이 긍정적인 반응을 보일 확률을 계산하여 점수화하였다.

4. 예측 모델 생성

4.1. 종속 변수

네이버 웹툰의 메인 화면에서 조회 수를 기준으로 웹툰이 정렬된 것을 볼 수 있다. 또한, 미디어에서 성공한 웹툰 작가나 작품을 소개할 때, '1위 웹툰', 혹은 '성공한 웹툰 작가'라고 하는 것을 종종 볼 수 있는데, 여기서 1위를 했다는 것은 조회 수를 기준으로 1위라는 것을 의미한다. 이를 통해 일반적으로 독자들이 생각하는 웹툰의 성공 여부는 조회 수로 결정된다는 것을 알 수 있다. 또한, 웹툰의 성공 여부에 관한 선행 연구를 참조하여 조회 수를 종속 변수로 설정하였다.[1]

선행 연구에 따라 네이버 웹툰의 완결된 작품 602편을 조회 순으로 나열한 후 등수를 기준으로 4개의 집단으로 구성하였다. 연재 중인 작품 또한 각 요일별로 4개의 집단으로 구성하여 등급을 부여하였다. 이를 토대로 1등급과 2등급은 성공(1)을, 3등급과 4등급은 실패(0)라는 레이블을 부여하고 이를 종속 변수로 설정하였다.

4.2. 독립 변수

본 논문은 선행 연구를 기반으로 하여 좋아요 수, 작품의 전체 평점, 장르, 첫 화 댓글 수, 5화까지의 댓글 수 평균, 독자 반응의 긍정 점수 등 6가지 요소를 웹툰의 성공 요소로 두었다.

장르의 경우 스릴러, 액션, 일상, 개그, 판타지, 순정 등 6가지로 재분류하였다. 한 웹툰이 다수의 장르에 해당할 경우, 웹툰에 흥행을 미치는 영향력이 높은 순으로 재설정하였다. 이 순서는 앞서 언급한 순서와 같다.[1]

대부분의 웹툰 댓글이 1화에서 2화로 넘어갈 때 급격히 줄어드는 것을 확인할 수 있었다. 그 이후의 댓글 수는 비슷한 숫자를 유지하고 있었다. 따라서 1화의 댓글 수가 유의미하다고 판단했기 때문에 1화 댓글 수를 따로 수집하

였고, 댓글 수의 전체 평균이 아닌 5화까지의 댓글 수를 수집하였다.

4.3. 댓글 반응을 활용하지 않은 기존 모델

댓글 반응을 활용한 모델과 그렇지 않은 모델을 비교하기 위해 먼저 독자 반응을 독립변수로 사용하지 않은 모델을 생성하였다. 각 모델은 5개의 독립 변수를 입력하였다.

로지스틱 회귀(Logistic Regression Model)는 샘플이 특정 클래스에 속할 확률을 추정하는 데에 사용된다. 로지스틱은 0과 1 사이의 값을 출력하는 시그모이드 함수이다.[2] 예측한 결과, 정확도는 약 0.653이었다.

두 번째 모델로는 아다 부스트(Ada Boost)를 사용하였다. 아다 부스트는 모든 예측기의 예측을 계산하고 가중치를 더해 예측 결과를 만든다. 가중치의 합이 가장 큰 클래스가 예측 결과가 된다. 사이킷런은 SAMME라는 아다 부스트의 다중 클래스 버전을 사용한다. 본 논문에서는 30개의 결정 트리를 기반으로 하는 아다 부스트 분류기를 훈련시켰다.[2] 결정 트리는 최대 depth를 1로, 결정 노드 하나와 리프 노드 두 개로 이루어져 있다. 예측 결과의 정확도는 약 0.713이다.

세 번째 모델인 서포트 벡터 머신(Support Vector Machine, SVM)은 선형, 비선형 분류, 회귀, 이상치 탐색에도 사용할 수 있는 다목적 모델이다. 그 중 커널 서포트 벡터 머신은 강력한 모델이며 다양한 데이터 셋에서 잘 작동하기 때문에 본 논문에서는 커널 서포트 벡터 머신을 이용하였다. 그중 모든 차수의 모든 다항식을 고려하는 가우시안 커널을 사용하였다.[2] 서포트 벡터 머신을 이용한 모델의 정확도는 약 0.653이다.

<표 1> 댓글 반응을 활용하지 않은 예측 모델 평가

Model	Accuracy	Precision	Recall
로지스틱 회귀모형	0.653	0.589	0.896
아다 부스트	0.713	0.569	0.854
서포트 벡터 머신	0.653	0.589	0.896

가장 높은 정확도를 갖는 모델은 아다 부스트 모델이다.

4.4. 댓글 반응을 활용한 모델

댓글 반응을 포함한 6개의 독립 변수를 사용하여 모델을 생성하였다.

첫 번째 모델인 로지스틱 회귀의 정확도는 약 0.693이며 아다 부스트를 이용한 모델의 정확도는 약 0.733이다. 마지막으로 서포트 벡터 머신의 정확도는 약 0.673으로 댓글 반응을 활용하지 않은 기존 모델보다 정확도가 높은 것을 알 수 있다.

<표 2> 댓글 반응을 활용한 예측 모델 평가

Model	Accuracy	Precision	Recall
로지스틱 회귀모형	0.693	0.620	0.917
아다 부스트	0.733	0.672	0.854
서포트 벡터 머신	0.673	0.606	0.896

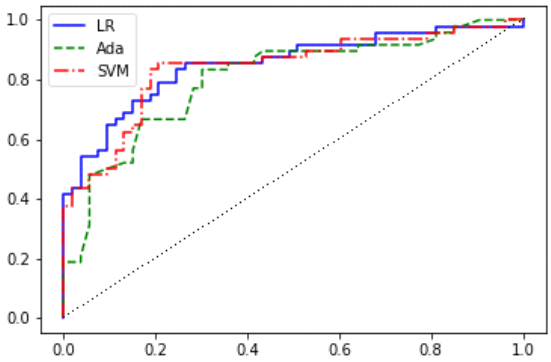
딥러닝을 활용하여 댓글 반응을 새로운 독립 변수로 사용한 모델 중 가장 정확도가 높은 모델은 아다 부스트 모델이다.

4.5. 모델 성능 평가

독자들의 반응을 알 수 있는 댓글 반응을 딥러닝을 활용하여 점수화하였다. 그리고 해당 변수를 독립 변수로 사용한 모델과 사용하지 않은 기존 모델의 정확도에서 차이가 발생하는 것을 알 수 있다. 세 모델 모두 댓글 반응을 활용할 경우에 정확도가 높은 것을 확인할 수 있다. 따라서 정확도 외에 모델의 성능을 평가할 수 있는 방법인 ROC 커브(Receiver Operating Characteristic Curve)와 AUC 면적(Area Under Curve)을 비교해보았다.

ROC 커브는 이진 판별모형의 성능을 컷오프에 맞춰 시각화한 그림이다. 최적 모형은 상단의 (0,1) 지점을 지나는 모형이다. <그림 2>를 보면 댓글 반응을 활용한 로지스틱 회귀모형이 가장 최적의 모형에 가까운 것을 알 수 있다.

<그림 2> 각 모델의 ROC 그래프



ROC 곡선 아래의 면적인 AUC로도 어떤 모형이 가장 적합한지 알 수 있다. AUC가 1인, 즉 정사각형 모양의 ROC 그래프를 그리는 모델이 최적의 모델이다. 로지스틱 회귀모형은 AUC 면적이 약 0.851이다. 아다 부스트 모형은 약 0.803이며 서포트 벡터 머신의 AUC 면적은 약 0.839이다. 로지스틱 회귀모형의 AUC 면적이 가장 큰 것을 알 수 있다.

댓글 반응을 활용한 모델이 활용하지 않은 모델보다 정확도가 높은 것을 알 수 있었으며, ROC 커브 그래프, AUC 면적에 의해 그 중 로지스틱 회귀모형이 가장 최적의 모델이라고 할 수 있다.

5. 성공 요인 분석

로지스틱 회귀모형에서 6가지 독립변수의 변수 중요도를 알아보았다. 각 변수는 순서대로 좋아요 수(x1), 평점(x2), 5화까지의 평균 댓글 수(x3), 1화 댓글 수(x4), 댓글 반응(x5)이며 변수 x6부터는 범주형 변수인 장르를 나타낸다. 개그(x6), 순정(x7), 스릴러(x8), 액션(x9), 일상(x10), 판타

지(x11) 순서이다. Z-값은 모델의 통계적 유의성에 대한 결정을 내릴 때 p-값을 계산하기 위해 사용하는 값이다. Z-값이 0과 가까울수록 해당 변수는 영향을 거의 미치지 않는다고 판단하며 양의 값을 가질 경우 양의 상관관계를, 음의 값을 가질 경우 음의 상관관계를 갖는다는 것을 의미한다. Z-값을 기준으로 했을 때, 가장 높은 영향력을 가지는 변수는 6.5796으로 좋아요 수(x1)이며, 두 번째로 가장 영향력이 높은 변수는 댓글 반응(x5)이다. Z-값은 1.9003이다. 1화 댓글 수(x4)는 모델에 세 번째로 높은 영향력을 미치는 변수로 1.7493의 Z-값을 가진다.

좋아요 수가 높다는 것은 웹툰이 연재를 시작한 시점부터 꾸준하게 긍정적인 반응을 끌어온 것을 의미한다고 할 수 있다. 또, 댓글의 반응이 긍정적이면서 특히 첫 화의 댓글 수가 많은 웹툰은 분량이 많거나 재미있고 신선한 소재를 사용한 웹툰인 경우가 많았다.

<그림 3> 로지스틱 회귀 모형의 변수 중요도

Model:	Logit	Pseudo R-squared:	0.344
Dependent Variable:	성공	AIC:	214.8236
Date:	2019-08-30 17:10	BIC:	251.7460
No. Observations:	212	Log-Likelihood:	-96.412
Df Model:	10	LL-Null:	-146.86
Df Residuals:	201	LLR p-value:	3.5974e-17
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
x1	3.5950	0.5464	6.5796	0.0000	2.5241	4.6659
x2	-2.5529	2.0977	-1.2170	0.2236	-6.6644	1.5585
x3	-3.3388	4.8404	-0.6898	0.4903	-12.8258	6.1483
x4	8.7040	4.9756	1.7493	0.0802	-1.0481	18.4560
x5	1.7962	0.9452	1.9003	0.0574	-0.0564	3.6487
x6	-1.1285	1.9671	-0.5737	0.5662	-4.9839	2.7269
x7	-0.6061	1.9923	-0.3042	0.7610	-4.5108	3.2987
x8	-0.4681	2.0373	-0.2298	0.8183	-4.4611	3.5249
x9	-0.5565	2.0464	-0.2720	0.7857	-4.5673	3.4543
x10	-0.9640	1.9581	-0.4923	0.6225	-4.8017	2.8737
x11	-0.7114	1.9426	-0.3662	0.7142	-4.5188	3.0959

6. 결론

모델 검증 결과, 댓글 반응을 변수로 사용할 때 정확도가 높아지는 것을 볼 수 있다. 따라서 신인 작가들의 웹툰 성공을 예측할 때에 독자들의 반응을 살펴보는 것이 필요함을 알 수 있다. 댓글 반응을 활용한 모델 중 가장 우수한 알고리즘으로 로지스틱 회귀모형이 선정되었다. 해당 알고리즘으로 모델을 구현한 결과, 웹툰의 성공에 가장 큰 영향을 미치는 요소로 좋아요 수, 독자 반응, 1화 댓글 수를 꼽을 수 있었다. 세 가지 변수들이 높은 값을 갖는 웹툰은 초반에 긍정적인 이슈를 얻고 꾸준한 인기를 얻어 성공한 웹툰이 되리라 판단할 수 있다.

제안하는 모델을 실제 훈련 데이터나 테스트 데이터로 사용하지 않은 5화 미만 신작 웹툰에 적용해 보았다. 7월 30일 17:00 기준 11개의 작품 중 3개의 작품이 성공할 것으로 예측하였다. 0.90의 가장 높은 확률로 성공할 것이라 예측한 웹툰은 '칼 가는 소녀(오리)' 작품이었으며 두 번째는 0.66의 확률로 성공할 것이라고 한 '친애하는 X(반지운)', 세 번째는 '판타지 여동생!(우뉴)'로 0.60의 확률로 예측했다. 해당 작품들은 다른 작품에 비해 좋아요 수가 높고 1화 댓글 수가 높은 것을 알 수 있다. 실제로 8월 31일 19:00 기준 조회 수 1등급, 2등급에 위치하는 것을 확인할 수 있다. 좋아요 수가 높은 작품이 있었지만, 해당 작품은 1화 댓글 수가 상대적으로 적었기 때문에 실패라고 예측한 것으로 보인다.

본 논문에서는 웹툰이 성공하는 요소 중 하나인 작가파워가 없는 신인 작가들이 어떻게 하면 성공할 수 있을지에 대한 성공 지표를 제공하였다. 특히, 이전 선행 연구에서는 다루지 않았던 독자들의 직접적인 반응이라고 할 수 있는 댓글 반응을 딥러닝을 통해 성공 요인 중 하나로 다루었다는 점에서 의의를 가진다. 향후 연구에서 더 많은 데이터를 수집하고 구체적인 지표를 제시함으로써 신인 웹툰 작가들에게는 성공의 방향을, 독자들에게는 다양한 작품을 볼 기회를 제공할 수 있을 것이다.

참고문헌

[1] 양지훈, 이지영, 이상우, "웹툰(Webtoon)의 흥행 결정 요인 연구", Vol. 16 No. 5, 한국콘텐츠학회논문지, 2016

[2] 프랑소와 숄레, "케라스 창시자에게 배우는 딥러닝", 박해선 옮김, 길벗, 2018

[3] 한국콘텐츠진흥원, "2018 웹툰 사업체 실태조사", 2018

[4] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31: 249-268, 2007

[5] 함효민, 남다은, 박현종, 안성민, 홍성준, 임경식, "로지스틱 판별 분석을 통한 한국어 구어체 문장 감정 분류 정확도 분석", 한국정보과학회 학술발표논문집: 1870-1872, 2017