KIPS Trans. Softw. and Data Eng.
Vol.12, No.4 pp.173~178
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.4.173

Ensemble Learning-Based Prediction of Good Sellers in Overseas Sales of Domestic Books and Keyword Analysis of Reviews of the Good Sellers

Do Young Kim† · Na Yeon Kim†† · Hyon Hee Kim†††

ABSTRACT

As Korean literature spreads around the world, its position in the overseas publishing market has become important. As demand in the overseas publishing market continues to grow, it is essential to predict future book sales and analyze the characteristics of books that have been highly favored by overseas readers in the past. In this study, we proposed ensemble learning based prediction model and analyzed characteristics of the cumulative sales of more than 5,000 copies classified as good sellers published overseas over the past 5 years. We applied the five ensemble learning models, i.e., XGBoost, Gradient Boosting, Adaboost, LightGBM, and Random Forest, and compared them with other machine learning algorithms, i.e., Support Vector Machine, Logistic Regression, and Deep Learning. Our experimental results showed that the ensemble algorithm outperforms other approaches in troubleshooting imbalanced data. In particular, the LightGBM model obtained an AUC value of 99.86% which is the best prediction performance. Among the features used for prediction, the most important feature is the author's number of overseas publications, and the second important feature is publication in countries with the largest publication market size. The number of evaluation participants is also an important feature. In addition, text mining was performed on the four book reviews that sold the most among good-selling books. Many reviews were interested in stories, characters, and writers and it seems that support for translation is needed as many of the keywords of "translation" appear in low-rated reviews.

Keywords : Ensemble Learning, Good Seller Prediction, Book Review Analysis, Text Mining, Keyword Analysis

앙상블 학습 기반 국내 도서의 해외 판매 굿셀러 예측 및 굿셀러 리뷰 키워드 분석

김 도 영† · 김 나 연†† · 김 현 희†††

요 약

한국 문학이 세계적으로 관심을 받게 됨에 따라 해외 출판시장에서의 수요가 지속적으로 증가하고 있다. 따라서 해외 출판시 도서 판매량의 예측과 과거 해외 독자들의 선호도가 높았던 도서들의 특징을 분석하는 것이 중요하다. 본 논문에서는 최근 5년간 해외 출간된 도서 중에서 굿셀러로 분류되는 누적 5천 부 이상 판매 여부 예측 모델을 제안하고 굿셀러의 요인이 되는 변수들을 분석하였다. 이를 위해, XGBoost, Gradient Boosting, Adaboost, LightGBM, Random Forest의 다섯 개 앙상블 학습 모델과 Support Vector Machine, Logistic Regression, Deep Learning을 적용한 결과, 불균형 데이터 문제 해결에 앙상블 알고리즘이 큰 효과를 보였음을 확인했으며, 그 중에서도 LightGMB 모델이 99.86%의 AUC 값을 얻어 가장 좋은 예측 성능을 보임을 검증하였다. 예측을 위해 사용된 변수 중 가장 중요한 변수는 작가의 해외 출간 횟수로 나타났으며, 평점 평균, 상위 출판 시장 규모를 가진 국가에서 출판 여부와 평점 참여자 수 등이 중요한 변수로 나타났다. 또한, 굿셀러 도서에 대한 독자들의 반응을 분석하기 위해서, 굿셀러 도서 중에서도 가장 많이 판매된 4권의 작품 리뷰에 대해 텍스트 마이닝을 실시하였다. 분석 결과 스토리, 등장인물, 작가 순으로 관심을 둔 리뷰가 많았음을 알 수 있었으며, 평점이 낮은 리뷰로부터 번역 키워드가 도출된 것으로 보아, 번역에 대한 지원을 확대하는 것이 필요할 것으로 보인다.

키워드 : 앙상블 학습, 해외 굿셀러 예측, 리뷰 분석, 텍스트 마이닝, 키워드 분석

1. 서 론

한국 문학이 세계적으로 관심을 받게 됨에 따라 다양한 문화권에서 주목받고 있다. 한국문학번역원에 따르면 한국 문학은 연평균 10% 정도로 해외 출판이 증가하고 있으며, 번역원 전체 지원 건수 중 해외 출판사가 한국 문학의 번역 및 출판을 신청하는 경우가 80%에 달하는 것으로 알려졌다[1]. 따라서 늘어가는 해외 시장의 수요를 고려하여 해외 출간 도서의 판매량 예측이 점차로 중요할 것으로 보인다. 본 연구는 한국문학번역원에서 실시한 해외 출간 도서 판매 현황 데이터를 기반으로 온라인 서적 사이트로부터 리뷰 정보, 국내 도서와 작가 정보를 크롤링하여 서적을 해외에 번역하여 판매하였을 경우 판매 부수를 예측할 수 있는 모델을 제시하고 키워드를 분석하였다.

본 연구에서는 국내 도서가 해외에 출간될 경우 5천 부 이

※ 이 논문은 2022년 한국정보처리학회 ASK 2022의 "빅데이터를 활용한 국내 도서의 해외 판매 시 굿셀러 예측"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 동덕여자대학교 정보통계학과 학사과정
†† 준 회 원 : 동덕여자대학교 정보통계학과 학사
††† 종신회원 : 동덕여자대학교 정보통계학과 부교수
Manuscript Received : August 2, 2022
First Revision : October 4, 2022
Accepted : October 20, 2022
* Corresponding Author : Hyon Hee Kim(heekim@dongduk.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

상 판매 가능한지 예측 모델을 생성하였으며, 또한 가장 많이 판매된 해외 판매 서적들의 독자 리뷰를 분석하여 그 특성을 파악하고자 하였다. 먼저 해외 출간된 도서 중에서 누적 5천 부 이상 판매된 서적들 Digital Library of Korean Literature (DLKL)[2] 사이트를 통해 2016~2020년 한국문학번역원의 지원으로 해외 출간된 도서 총 578종을 대상으로 누적 5천 부 이상 판매된 해외 출간 도서 데이터를 종속변수로 활용해 누적 5천 부 판매 여부 예측 모델을 제작했다. 이를 위해 한국문학번역원으로부터 누적 총 5천 부 이상 판매된 23권의 한국 문학 목록을 제공받았으며, 국내 도서 초판 부수가 2천 부에서 3천 부 내외이므로 판매 부수 5천부는 해당 도서가 평균적인 초판 부수 이상 현지에서 지속적으로 판매되었다고 볼 수 있어 5천 부 이상 판매 도서를 굿셀러로 정의하였다.

다음으로 굿셀러 중에서 가장 많이 판매된 상위, 네 개의 작품인 조남주의 〈82년생 김지영〉, 한강의 〈채식주의자〉, 손원평의 〈아몬드〉, 정유정의 〈종의 기원〉에 대해 해외 독자들의 반응을 파악하기 위해 아마존과 굿리즈 리뷰 데이터를 수집하여 텍스트 마이닝을 진행했다. 해당 서적에 대한 긍정적 리뷰와 부정적 리뷰에 따라 어떤 키워드가 담겨있는지 파악하기 위해서 5점 척도의 평점에서 1~2점대의 리뷰와 4~5점대의 리뷰로 나누어 분석하였다.

분석에 사용된 굿셀러는 총 23권이었고 일반 판매 서적은 총 555권으로 그 비율이 약 1:24.1로 불균형이 심하였으므로 좋은 성능의 모델을 생성하기 위해서 Synthetic Minority Over-Sampling Technique(SMOTE) [3] 알고리즘과 앙상블 학습 알고리즘[4]을 적용하였다. 앙상블 알고리즘으로 Xgboost [5], GradientBoosting [6], Adaboost [7], Lightgbm [8], Random Forest [9]를 적용하였다. 추가로 Support Vector Machine [10], Logistic Regression [11], Deep Learning [12] 도 적용해서 앙상블 알고리즘을 적용한 모델과의 성능 비교를 진행하였다. 각 모델을 생성해 예측해 본 결과, SMOTE를 적용한 모델의 성능이 향상되었으며 앙상블 알고리즘이 다른 알고리즘보다 성능이 뛰어남을 알 수 있었다.

최종 선정 모델은 LightGBM에 SMOTE를 적용한 모델로, AUC 스코어 99.86%이었다. 예측에 영향을 미친 주요 요인들을 찾기 위한 변수 중요도 분석 결과, 작가별 해외 출간 횟수, 평점 평균, 대규모 출판 시장에 속하는 국가 출간과 평점 참여자 수가 주요한 요인으로 작용했음을 알 수 있었다. 그리고 평점 1점~2점 사이 리뷰에선 'translation' 키워드가 도출된 것으로 보아, 작품 번역 시 더 많은 시간을 투자해야 할 것으로 확인되었다. 평점 4점~5점 사이 리뷰에선 'story', 'character', 'author' 키워드가 도출된 것으로 보아, 스토리와 등장인물, 작가에 관심을 두고 주목한 리뷰가 많음을 알 수 있었다.

본 연구의 굿셀러 예측 및 키워드 분석 결과는 향후 국내 도서의 해외 출간 시 고려해야 할 점을 제시하여 한국 문학의 국제적 성장에 한층 기여 할 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 2장은 데이터 수집과 전처리에 대해 설명한다. 3장에서는 굿셀러 판매 부수 예측 모델 제안하고 성능 평가 결과를 보여준다. 4장에서는 분석 기법과 결과를 설명하고. 5장에서는 결론 및 향후 연구를 제시한다.

2. 관련 연구

최근에는 기본적인 서지 정보 혹은 리뷰 데이터를 이용하여 도서 흥행을 예측하는 연구가 주목받고 있다. 저자 정보, 장르 등 기본적인 서지 정보를 활용하여 책 판매에 보편적인 패턴을 파악하는 연구[13]와 저자 정보, 출판 연도 등 기본적인 서지 정보와 평점 수, 평균 평점 등 리뷰 데이터를 활용하여 베스트셀러가 될 것인지 예측하는 연구[14], 그리고 다양한 속성과 과거 판매 순위 데이터 기반으로 도서 판매 순위를 예측하는 연구[15] 등 도서 정보 및 리뷰를 기반으로 다양한 분석 기법을 적용하는 연구가 점차 증가하고 있다. [13]에서는 뉴욕 타임스 베스트셀러 목록으로부터 책 판매에 보편적인 패턴 파악 후 책의 전체 판매 방향을 재현할 뿐만 아니라 초기 판매량을 기반으로 평생 판매될 총 부수를 예측하였다. [14]의 연구는 아마존과 굿리즈 데이터로부터 아마존 베스트셀러와 베스트셀러가 아닌 도서의 차이점을 식별하고 로지스틱 회귀분석과 서포트 벡터 머신을 실시하여 책이 출판된 지 15일, 한 달 후 베스트셀러가 될 것인지 예측하였다. [15]의 연구는 아마존 사이트에서 수집한 데이터를 기반으로 도서 판매 순위 예측을 위한 GAN(Generative Adversarial Network) 프레임워크를 개발하고, 그 결과를 MLP, DBN, CNN과 같은 다른 신경망 모델들과 비교했다.

이러한 연구들을 통해서 도서 정보 및 리뷰 데이터를 분석하여 도서의 흥행을 예측하고 특징을 분석하는 것이 효과적인 접근 방법임을 알 수 있다. 따라서 본 연구에서는 국내 도서의 해외 판매 굿셀러 예측을 진행하고 텍스트 마이닝을 통해 해외 독자들이 도서의 어떤 속성에 관심도가 높은지 분석해 국내 문학의 성장에 기여하고자 한다.

3. 데이터 요약

아마존 [16]과 굿리즈 [17]에 등록된 한국문학번역원 지원의 해외 출간 도서 중에서 2016년부터 2020년 사이에 출간된 578종을 분석하였다. Dgital Library of Korean Literature (DLKL) 사이트에서 평점 평균, 평점 참여자 수,작성된 리뷰 총 26,105개를 크롤링하였다. 웹 크롤링은 웹 사이트의 내용을 자동화된 방법으로 수집하는 기술이다. 많은 양의 정보를 쉽게 수집할 때 적합한 기법으로, 본 연구에서는 셀레늄(Selenium) 라이브러리를 사용하여 데이터를 수집하였다[18].

도서별 평균 평점 및 분산, 평점을 준 독자 수, 리뷰 수, 리뷰 추천수를 변수로 사용하였다. 글자 수 분포 중 3분위 수가 넘는 리뷰들을 장문 리뷰로 정의하고 작품별 장문 리뷰가 차지하는 비율도 변수로 활용하였다. 번역 점수는 번역 관련 키워드가 포함된 문장들에 대해 Vader NLTK[19]를 활용하여 작품별 평균값을 산출하여 사용하였다. 또한, 번역 점수가 부여되지 않은 경우, 번역에 대한 언급이 없는 것이므로 평균값으로 대체하였다. 도서에 대한 긍정 및 부정 리뷰의 비율도 변수로 활용하였으며, 리뷰의 긍정 및 부정 평가를 위해 Bidirectional Encoder Representations from Transformers (BERT)[20]를 활용하였다. 1점을 받은 리뷰를 부정 리뷰로, 5점을 받은 리

뷰를 긍정 리뷰로 라벨링하여 학습 데이터를 생성하였고, 미세 조정을 실시하여 0.9644의 정확도를 얻었다. 작품별 번역 출간 횟수 및 작가의 해외 출간 횟수도 변수로서 고려하였다. 또한 대규모 출판 시장 상위 10개국에 속하는 국가 출간 횟수도 변수로 사용하였으며 이를 산출하기 위해 해당 국가에서 출간될 때마다 점유율을 계산하여 점수화하였다.

수집한 리뷰가 다양한 언어였으므로 구글 시트 GOOGLE TRANSLATE 함수를 사용하여 영어로 변환 후 대문자를 소문자로 변환, 불용어 제거, 표제어 추출, 명사 추출 전처리 과정을 거쳐 작품 전체 리뷰 와 작품별 리뷰로 나누었다. 그 후 평점을 기준으로, 1점에서 2점 사이 리뷰를 부정 리뷰로, 4점에서 5점 사이를 긍정 리뷰로 나누었다.

본 연구에 쓰인 데이터 셋과 소스 코드는 https://github.com/nayeonkim1/Good-Seller-predictive-model-and-Keyword-Analysis- 에서 확인이 가능하다.

4. 예측 모델 및 성능 평가

4.1 굿셀러 예측 모델

해외 출간 도서 데이터는 굿셀러 서적이 23개, 일반 도서가 555개로 심각한 불균형을 이루고 있으며, 이를 고려하여 SMOTE를 적용한 오버샘플링과 앙상블 학습을 적용하였다 [21-22]. 클래스 비율을 각각 1:10, 1:5, 1:4, 1:2, 1:1로 다양하게 구성하여 실험을 진행하였으며, 앙상블 알고리즘으로는 Adboost, GradientBoosting, LightGBM, XGboost, 그리고 random forest를 적용하였다.

부스팅은 각 단계의 잔차를 모델링하여 이전 예측의 실수를 수정하는 알고리즘으로, 모델을 반복적으로 개선하여 예측 오류를 줄인다. 배깅은 일종의 병렬 앙상블 방법으로, 부트 스트랩을 통해 모델의 높은 분산을 줄임으로써 예측력을 향상시키는 알고리즘이다. 앙상블 알고리즘 외에도 Support Vector Machine, Logistic Regression, Deep Learning 알고리즘도 적용해 앙상블 알고리즘을 적용한 모델과의 성능을 비교하고자 했다.

4.2 실험 결과

Fig. 1은 클래스 구성 비율에 따른 모델별 교차 검증 성능 평가 그래프이다. SMOTE를 적용하여 두 클래스간 비율을 동일하게 하였을 때 모든 알고리즘에서 성능이 향상됨을 알 수 있다. 또한, 비율이 1:5 비율로 구성했을 때부터 앙상블 알고리즘을 적용한 모델들의 성능이 Support Vector Machine, Logistic Regression, Deep Learning 알고리즘과 비교했을 때 월등히 우수한 성능을 보이는 것으로 확인됐다. 그중에서도 LightGBM의 성능이 가장 높게 나타났다. 아마존 베스트셀러 여부를 예측했던 [14]의 연구와 비교하면, Support Vector Machine (AUC=0.9078)과 Logistic Regression (AUC=0.8932)의 성능은 [14]의 연구 (SVM=0.919, LR= 0.963) 보다 낮았으나, 앙상블 알고리즘을 적용한 5개의 모델의 경우 모두 [14]의 연구의 성능보다 높은 것을 확인할 수 있었다.

Fig. 2는 비율을 조정하지 않은 경우와 SMOTE를 이용하여 동일하게 조정한 경우의 성능 평가 결과를 보여주고 있다. SMOTE를 적용한 경우의 성능이 크게 향상됨을 알 수 있다.

Fig. 3은 SMOTE를 적용하여 클래스간 비율을 동등하게 조절한 다음 알고리즘 별 성능 평가 그래프이다. Support Vector Machine, Logistic Regression, Deep Learning 모델과 비교했을 때, 앙상블 알고리즘인 XGboost, Gradient-Boosting, Adaboost, LightGBM, Random Forest를 적용했을 때 더 좋은 성능을 얻은 것을 확인할 수 있었다. 앙상블 알고리즘을 SMOTE 기법으로 데이터 범주의 균형을 맞추는 전처리 과정과 함께한 경우, 그렇지 않은 경우보다 더 큰 효과를 볼 수 있음을 알 수 있다. 가장 좋은 성능은 SMOTE를 적용한 LightGBM 알고리즘으로 AUC 스코어 0.9986을 나타냈다.

최종적으로 선택된 LightGBM 모델로부터 변수 중요도를 산출한 결과를 Fig. 4에서 볼 수 있다. 굿셀러 예측에 가장 중요한 요인은 작가의 해외 출간 횟수로 나타났으며, 평균 평점, 대규모 출판 시작에 속하는 국가에서 출간 횟수, 평점 참여자수 등이 뒤를 이어 중요한 요인으로 나타났다. 마지막으로 번역에 대한 만족도와 장문 리뷰 비율 역시 판매 예측에 적지 않은 영향을 미친 것으로 나타났다. 작품의 국제 수상 여부는 예상과 달리 영향이 미미한 것으로 나타났다.

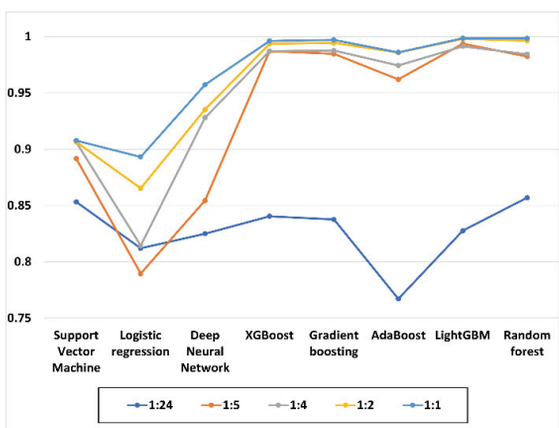


Fig. 1. Performance Evaluation by the Class Ratio

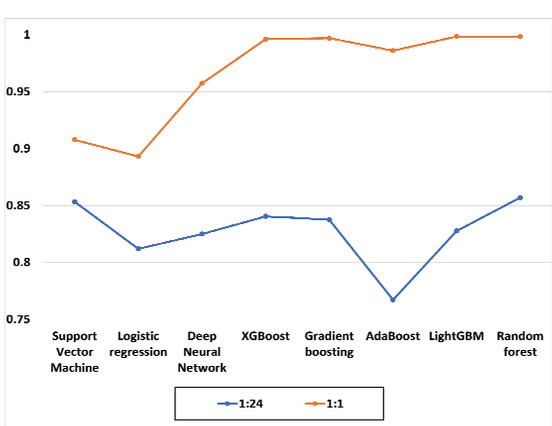


Fig. 2. Performance Evaluation by SMOTE

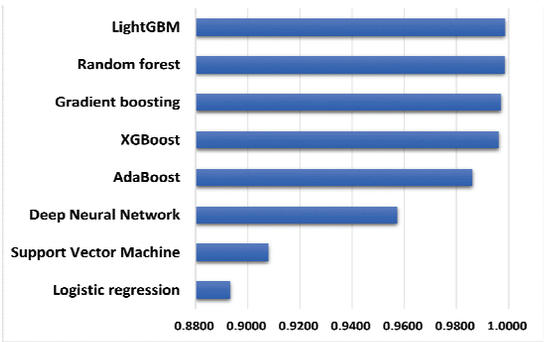


Fig. 3. Performance Evaluation by Prediction Models

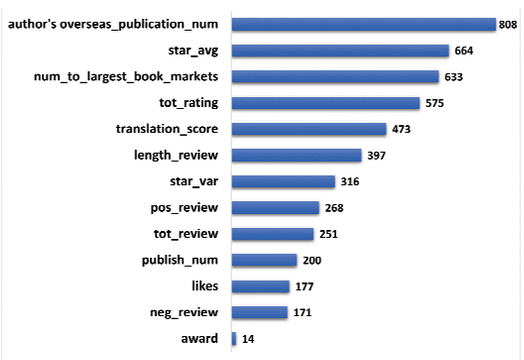


Fig. 4. Feature Importance in the Final Model

5. 리뷰 데이터의 키워드 분석

5.1 자료 분석 및 분석 방법

본 연구에서는 아마존과 굿리즈 사이트에서 2022년 4월 16일까지 등록된 리뷰를 크롤링하였다. 조남주의 〈82년생 김지영〉은 734개, 손원평의 〈아몬드〉는 572개, 정유정의 〈종의 기원〉은 488개, 한강의 〈채식주의자〉는 1,789개 리뷰를 수집했다. 수집한 데이터는 구글 시트 GOOGLETRANSLATE 함수를 사용하여 영어로 변환 후 대문자를 소문자로 변환, 불용어 제거, 표제어 추출, 명사 추출 전처리 과정을 거쳐 작품, 네 개의 리뷰 분석 및 작품별 리뷰 분석을 실시하였다. 작품 전체 리뷰 진행 시 작품별 450개씩 추출하여 총 1,800개 리뷰를 사용했다. 그 후 독자들의 반응을 파악하고자 4점에서 5점 사이는 긍정 리뷰, 1점에서 2점 사이는 부정 리뷰로 데이터를 나누었다.

전처리 과정 진행 후 독자의 반응을 알기 위하여 단어 빈도와 TF-IDF 가중치를 적용하여 중요 키워드를 추출하였다.

5.2 분석 결과

Table 1은 네 개 작품 전체 리뷰의 TF-IDF 가중치를 적용한 주요 키워드를 나타낸 것으로 Table 1의 positive는 긍정 리뷰 분석 결과이고 Table 1의 negative는 부정 리뷰 분석 결과이다. 긍정 리뷰의 상위 3개의 키워드 'woman', 'story', 'life'를 볼 때, 한강의 〈채식주의자〉, 조남주의 〈82년생 김지영〉에 내재한 페미니즘 메시지에 주목한 평이 많았음을 알 수 있다. 또한 'story', 'character', 'author' 등의 키워드가 등장한 것으로 보아 '줄거리', '등장인물', '작가' 순으로 관련 평이 들어간 리뷰가 많이 쓰이고 있는 것을 확인할 수 있다. 특히 부정적인 리뷰에서 'translation' 키워드가 도출된 것으로 보아, 작품 번역 시 더 많은 투자가 필요한 것으로 보인다.

Table 2는 〈82년생 김지영〉 리뷰의 TF-IDF 가중치를 적용한 중요 키워드로 Table 2의 positive는 긍정 리뷰 분석 결과이고, Table 2의 negative는 부정 리뷰 분석 결과이다. 긍정적인 리뷰에선 'woman', 'story', 'life', 'society', 'gender' 등의 키워드가 상당수 도출된 것으로 볼 때, 리뷰에서 여성이 사회에서 맞닥뜨린 차별과 불평등 문제에 관한 생각을 가장 많이 표현했음을 알 수 있다. 또한, 'korea', 'world', 'country', 'japan' 등의 키워드와, 'situation', 'feel', 'experience' 등의 키워드를 볼 때, '여성 혐오'가 전 세계적으로 관찰되는 보편적 현상으로서 해외에서도 공감을 얻고 있음을 알 수 있다. 특히, 'japan' 키워드는 우리나라는 제외하고 결과에 나타난 유일한 국가명으로, 〈82년생 김지영〉이 2020년 10월 기준 일본에서만 21만 부 이상을 판매하는 성과를 거뒀던 만큼, 일본에서의 반응이 특히 두드러지게 나타났음을 확인할 수 있다. 부정적인 리뷰에서 'translation' 키워드가 도출된 것으로 보아, 작품 번역 시 더 많은 투자가 필요한 것으로 보인다. 그리고 'woman', 'story', 'life', 'society', 'gender' 등의 키워드와 'discrimination', 'sexism' 등의 키워드를 볼 때, '여성 차별과 불평등'에 주목한 리뷰가 많음을 알 수 있다. 부정적인 리뷰 키워드와 긍정적인 리뷰 키워드가 동일하진 않지만 '여성 차별'에 주목한 리뷰가 상당수 존재함을 확인할 수 있다.

Table 3은 〈아몬드〉 리뷰의 TF-IDF 가중치를 적용한 중요 키워드이며 Table 3의 positive는 긍정 리뷰를 나타내고, Table 3의 negative는 부정 리뷰를 나타낸다. 〈아몬드〉의 등장인물인 'yunjae', 'gon', 감정표현 불능증을 앓고 있는 주인공을 나타내는 'emotion', 'feel', 'love', 'feeling', 그리고 'character', 'protagonist' 등의 키워드가 상당수 도출된 것으

Table 1. TF-IDF keywords of 4 books using TF-IDF

positive	woman, story, life, time, read, character, way, thing, people, author, world, book, novel, korea, reading, kim, family, emotion, feel, society, day, work, hye, yunjae, thriller, end, lot, mother, jiyoung, star
negative	story, character, woman, page, hye, time, book, life, novel, yeong, read, star, thing, author, feel, plot, day, person, people, way, translation, meat, message, korea, husband, lot, protagonist, end, dream, kang

Table 2. TF-IDF Keywords from 〈Kim Jiyoung, Born 1982〉

positive	woman, life, story, korea, kim, jiyoung, society, world, men, read, work, time, south, country, day, korean, way, gender, japan, young, fact, people, ji, child, reading, man, situation, mother, feel, experience
negative	woman, story, book, korea, character, life, time, word, style, way, review, problem, south, novel, society, kim, sexism, reputation, page, korean, translation, author, reason, point, buzz, discrimination, read, work, lot, thing

Table 3. TF-IDF Keywords from 〈Almond〉

positive	story, read, yunjae, emotion, time, character, life, reading, feel, people, book, lot, way, almond, author, love, heart, boy, star, day, thing, novel, gon, word, world, work, condition, protagonist, person, chapter
negative	feel, story, character, day, place, development, book, people, novel, protagonist, way, sympathize, quality, thing, life, lot, page, disease, recommend, condition, year, read, evaluation, end, production, buy, alexithymia, anger, yunjae, gon

Table 4. TF-IDF Keywords from 〈The Good Son〉

positive	story, thriller, character, jin, read, author, way, son, mother, time, page, yu, book, end, blood, reading, novel, mind, jeong, plot, star, thing, family, life, memory, reader, good, psychopath, night, great
negative	story, character, plot, mystery, page, time, spoiler, star, author, jin, novel, mother, thriller, translation, end, son, outcome, book, psychopath, cliche, epic, quite, yu, person, reader, murder, mind, korean, protagonist, jeong

로 비추어볼 때, 등장인물과 그 인물 간의 관계성에 주목한 평들이 많았음을 알 수 있다. 부정적인 리뷰에선 'disease', 'alexithymia' 등의 키워드를 볼 때, 주인공인 윤재가 앓고 있는 '감정표현 불능증'에 관심을 둔 리뷰가 존재함을 알 수 있다.

Table 4는 〈종의 기원〉 리뷰의 TF-IDF 가중치를 적용한 주요 키워드로서 Table 4의 positive는 긍정 리뷰 분석 결과이고, Table 4의 negative는 부정 리뷰 분석 결과를 나타낸다. 긍정적인 리뷰에선 'story', 'end', 'plot' 등의 키워드가 도출된 것으로 보아, 전반적인 이야기 구조에 대한 평이 많았음을 알 수 있다. 그리고 'thriller'가 'story'를 제외하면 최다 도출 키워드인 것으로 보아, 스릴러라는 장르에 주목한 평이 많았던 것을 알 수 있다. 부정적인 리뷰에서 'psychopath', 'murder', 'blood' 등의 키워드로 볼 때, '사이코패스', '살인', '피' 등 작품에 등장하는 잔인한 소재에 주목한 리뷰가 존재함을 알 수 있다. 그리고 'translation' 키워드가 도출된 것으로 보아, 작품 번역 시 더 많은 투자가 필요한 것으로 보인다.

Table 5는 〈채식주의자〉 리뷰의 TF-IDF 가중치를 적용한 중요 키워드로 Table 5의 positive는 긍정 리뷰 분석 결과를 나타내고, Table 5의 negative는 부정 리뷰 분석 결과를 나타낸다. 긍정적인 리뷰에선 키워드 'author' 말고도 'han', 'kang' 작가의 본명이 다수 도출된 것으로 보아 작가의 인지도를 확인할 수 있다. 특히 부정적인 리뷰에서 'translation' 키워드가 도출된 것으로 보아, 작품 번역 시 더 많은 투자가 필요한 것으로 보인다.

조남주의 〈82년생 김지영〉은 여성 차별과 불평등에 관한 사회 문제에 관심을 두고 있는 긍정적인 리뷰가 많았다. 부정적인 리뷰로는 국내 도서의 해외 번역 관련 리뷰가 많았다. 손원평의 〈아몬드〉는 등장인물들과 그 인물 간의 관계성에 주목한 긍정적인 리뷰가 많았다. 부정적인 리뷰로는 질병과 감정표현 불능증을 직접 언급하며 주인공의 질병에 관심을 둔 리뷰가 많았다. 정유정의 〈종의 기원〉은 스토리 전반적인

Table 5. TF-IDF Keywords from 〈Vegetarian〉

positive	story, hye, read, woman, life, time, character, novel, yeong, way, family, husband, book, meat, kang, sister, people, author, thing, han, reading, world, dream, body, lot, society, reader, law, violence, brother
negative	story, character, time, book, page, woman, read, reading, people, end, life, novel, point, hye, dream, review, did, thing, way, plot, author, yeong, translation, lot, husband, illness, waste, meat, language, kind

이야기 구조와 스릴러라는 장르, 작가 '정유정'에 주목한 긍정적인 리뷰가 많았다. 부정적인 리뷰로는 국내 도서의 해외 번역 관련 리뷰가 많았다. 한강의 〈채식주의자〉는 작가 '한강'에 주목한 긍정적인 리뷰가 많았다. 부정적인 리뷰로는 국내 도서의 해외 번역 관련 리뷰가 많았다.

6. 결 론

본 논문에서는 한국 문학이 해외에서 출간될 경우 판매 부수를 예측할 수 있는 예측 모델을 제안하고, 해외 판매에 영향을 미치는 요인들을 분석하였다. 또한 해외 출간 대표 서적 네 권의 고객 리뷰 분석을 통해 성공적인 해외 출간 도서의 특징을 분석하였다.

예측 모델은 성능 평가 결과 SMOTE를 적용하여 불균형을 해소하고 앙상블 알고리즘인 LightGBM을 적용한 모델의 성능이 가장 뛰어났으며, AUC 0.9986의 우수한 성능을 보였다. 변수 중요도 분석 결과 작가의 인지도가 가장 주요한 변수로 나타났다. 고객 평점 및 대규모 출판 시장에 속하는 국가 출간 횟수도 해외 판매를 예측하는 데 영향을 미치는 변수로 나타났다. 또한 번역의 질 역시 중요한 요인으로 밝혀졌다.

마지막으로 TF-IDF를 기반으로 한 워드 클라우드를 통해 굿셀러로 분류된 작품 중 상위 4개 판매를 기록한 서적의 긍정 및 부정 리뷰에 따른 핵심 키워드를 확인할 수 있었다. 부정 리뷰에서는 〈아몬드〉를 제외한 다른 세 개 작품에서 'translation' 번역 관련 키워드가 도출되었다. 또한 〈종의 기원〉에 등장하는 자극적인 소재가 키워드로 도출된 것을 확인할 수 있었다. 긍정 리뷰에서는 'story', 'character', 'author' 등의 키워드가 등장했다. '줄거리', '등장인물', '작가' 순으로 관련 리뷰가 많이 쓰였음을 알 수 있었다. 〈종의 기원〉과 〈채식주의자〉는 작가의 본명이 다수 도출된 것으로 보아 작가의 인지도가 중요한 요인임을 알 수 있었다.

제안하는 모델을 활용하면 한국 문학의 해외 출간 시 성공적 판매에 도움을 줄 수 있을 뿐 만 아니라 추가적인 분석을 통해 경쟁력을 향상시킬 수 있을 것으로 보인다.

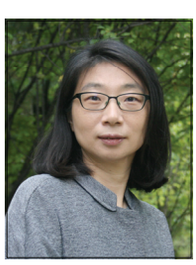
References

[1] M. Lee, "What is the best-selling Korean literature abroad? LTI Korea Research on the sales Korean literature published overseas in the last 5 years," Newspaper, 2022, http://www.news-paper.co.kr/news/articleView.html?idxno=76610

[2] Literature Translation Institute of Korea, [Internet] https://library.ltikorea.or.kr/
[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, Vol.16, No.1, pp.321-357, 2002.
[4] A. Geron, Hands-On Machine Learning with Skikit-Learn, Keras&TensorFlow, Orelly, 2019.
[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knoedge Discovery & Data Mining*, pp.785-794, 2016.
[6] L. Breiman, "Arcing The Edge," Technical Report 486, Statistics Department, University of California at Berkeley, Jun. 1997.
[7] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, Vol.55, No.1, pp.119-139, 1997.
[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T-Y Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp.3149-3157, Dec. 2017.
[9] L. Breiman, "Random Forests," *Machine Learning*, Vol.45, pp.5-32, Jan. 2001.
[10] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, Vol.13, No.4, pp.18-28, 1998.
[11] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society*, Series B, Vol.20, No.2, pp.215-242, 1958.
[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, Vol.521, pp.436-444, 2015.
[13] B. Yucesoy, X. Wang, J. Huang, and A-L Barabási, "Success in books: A big data approach to bestsellers," *EPJ Data Science*, Vol.7, 2018.
[14] S. K. Maity, A. Panigrahi, and A. Mukherjee, "Analyzing social book reading behavior on goodreads and how it predicts amazon best sellers," *Influence and Behavior Analysis in Social Networks and Social Media*, Sep. 2018.
[15] T. Q. Feng, M. Choy, and M. N. Laik, "Predicting book sales trend using deep learning framework," *International Journal of Advanced Computer Science and Applications*, Vol.11, No.2, pp.28-39, 2020.
[16] Amazon [Internet], https://www.amazon.com/
[17] Goodreads [Internet], https://www.goodreads.com/
[18] R. Lawsonl, Web Scraping with Python, Packt Publishing, 2015.
[19] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of International Conference on Weblogs and Social Media*, Vol.8, pp.216-225, Jan. 2015.
[20] J. Devlin, M-W Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of North American Chapter of the Association for Computational Linguistics*, pp.4171-4186, 2019.
[21] N. V. Chawla, A. Lazarevic, and O. Hall, "Smoteboost: Improving prediction of the minority class in boosting," in *Proceedings of Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.107-119, 2003.
[22] C. Seiffert, T. M. Khoshgoftaar, J. Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *Institute of Electrical and Electronics Engineers*, pp.185-197, 2010.

김 도 영
https://orcid.org/0000-0001-8214-2392
e-mail : rlaehdud159@gmail.com
2018년 ~ 현 재 동덕여자대학교 정보통계학과 학사과정
관심분야 : Big Data Analysis, Deep Learning & Unstructured

김 나 연
https://orcid.org/0000-0003-0084-5196
e-mail : 20181040@dongduk.ac.kr
2018년 ~ 현 재 동덕여자대학교 정보통계학과 학사
관심분야 : Big Data, Machine Learning, Text Mining



김 현 희
https://orcid.org/0000-0002-7507-8342
e-mail : heekim@dongduk.ac.kr
1996년 이화여자대학교 컴퓨터학과(학사)
1998년 이화여자대학교 컴퓨터학과(석사)
2005년 이화여자대학교 컴퓨터공학과(박사)
2005년 ~ 2006년 LG전자 디지털미디어 연구소 선임연구원
2006년 ~ 현 재 동덕여자대학교 정보통계학과 부교수
관심분야 : Machine Learning, Deep Learning, Big Data Analysis