공공도서관 도서 분류를 위한 머신러닝 적용 가능성 연구*

- 사회과학과 예술분야를 중심으로 -

A Study on Applicability of Machine Learning for Book Classification of Public Libraries: Focusing on Social Science and Arts

곽 철 완 (Chul Wan Kwak)**

초 록

이 연구의 목적은 공공도서관의 도서 분류를 위해 표제를 대상으로 머신러닝 기법의 적용 가능성을 조사하는데 있다. 데이터 분석은 아나콘다 플랫폼의 쥬피터 노트북을 통하여 파이썬의 싸이킷런 라이브러리를 이용하였다. 한글 형태소 분석을 위해 KoNLPy 분석기와 Okt 클래스를 사용하였다. 분석 대상은 공공도서관의 KORMARC 레코드에서 추출된 2,000건의 표제 필드와 KDC 분류기호(300대와 600대)이었다. 6가지 머신러닝 모델을 이용하여 데이터를 분석한 결과, 도서 분류에 머신러닝 적용 가능성이 있다고 판단되었다. 사용된 모델 중 표제 분류의 정확도는 신경망 모델이 가장 높았다. 표제 분류의 정확도 향상을 위해 도서 표제에 대한 조사와 표제의 토큰화 및 불용어에 대한 연구 필요성을 제안하였다.

ABSTRACT

The purpose of this study is to identify the applicability of machine learning targeting titles in the classification of books in public libraries. Data analysis was performed using Python's scikit-learn library through the Jupiter notebook of the Anaconda platform. KoNLPy analyzer and Okt class were used for Hangul morpheme analysis. The units of analysis were 2,000 title fields and KDC classification class numbers (300 and 600) extracted from the KORMARC records of public libraries. As a result of analyzing the data using six machine learning models, it showed a possibility of applying machine learning to book classification. Among the models used, the neural network model has the highest accuracy of title classification. The study suggested the need for improving the accuracy of title classification, the need for research on book titles, tokenization of titles, and stop words.

키워드: 머신러닝, 표제 분류, 도서관 분류, 파이썬, 사이킷런 라이브러리
Machine Learning, Title Classification, Library Classification, Python, Scikit-learn Library

* 본 연구는 (2019학년도) 강남대학교 교내연구비 지원에 의해 수행되었음.
** 강남대학교 산업데이터사이언스학부 데이터사이언스전공 교수(ckwak@kangnam.ac.kr)
논문접수일자 : 2021년 2월 22일 논문심사일자 : 2021년 3월 1일 게재확정일자 : 2021년 3월 19일
한국비블리아학회지, 32(1): 133-150, 2021. http://dx.doi.org/10.14699/kbiblia.2021.32.1.133

1. 서 론

1.1 연구의 필요성과 목적

다양한 분야에 머신러닝 기법이 적용되기 시작하면서 문헌정보학 분야에서도 이를 활용하는 분야가 확대되고 있다. 과거에는 정보학 분야에서 정보검색 혹은 자동 색인과 초록 작성에 머신러닝 기법이 활용되었으나, 오늘날에는 도서관 서비스 분야에도 머신러닝 기법의 활용이 확대되고 있다. 이러한 변화 속에서 도서관 자료 분류 분야에도 머신러닝 적용 필요성이 증가하고 있다.

도서관 자료 분류는 도서관 분류표를 기준으로 자관의 장서 구조와 이용자 요구를 잘 반영하여 수행되어야 한다. 일반적으로 개별 도서관에서 이루어지는 신착 도서에 대한 분류 단계는 몇 단계로 구분할 수 있다. 이중 첫 번째 단계는 자관에 해당 도서를 소장하고 있는지 혹은 유사한 도서가 있는지 조사하는 단계이다. 이때 분류 담당 사서가 자관의 장서 구조에 대하여 잘 알고, 많은 분류 경험이 있다면, 이용자에게 효과적인 도서 분류가 가능할 것이다. 하지만, 공공도서관의 경우 사서가 동일한 도서관에서 동일한 업무를 오랜 기간 담당하는 것이 현실적으로 쉽지 않다. 특히, 최근에 많은 공공도서관의 경우, 분류 및 목록 업무를 외부에 위탁하는 사례가 많아, 자관의 장서 구조와 이용자 요구를 반영한 효과적인 도서 분류가 어렵다.

이러한 상황에서 공공도서관의 도서 분류에 머신러닝 기법을 적용할 수 있을지 질문할 수 있다. 머신러닝의 기본 개념은 사전에 많은 데이터를 학습한 후, 새로운 데이터의 값을 스스로 예측하거나, 데이터를 분류 혹은 여러 범주로 구분하는 것이다. 머신러닝 개념을 도서관의 도서 분류에 적용한다면, 해당 도서관에 소장하고 있는 도서의 분류 기호를 학습한 후, 새로운 도서에 가장 유사한 분류 기호를 부여하는 것이다.

머신러닝을 이용한 도서 분류에 있어서 연구의 대상은 전통적인 도서 분류 절차에서 찾을 수 있다. 한국십진분류법(KDC)에 도서 분류를 위한 주제 파악 단계에 대해, 우선 서명을 확인하고 목차를 살펴보며, 필요하다면 서문 등을 참고하라고 안내하고 있다(한국도서관협회, 2013). KDC에서도 언급하고 있듯이 도서 분류에 있어서 필수적인 정보는 해당 도서의 표제이므로, 표제를 연구의 대상으로 선정하는 것이 적절하다. 이에 이 연구의 목적은 공공도서관의 도서 분류를 위해 표제를 대상으로 한 머신러닝 기법의 적용 가능성을 조사하는데 있다.

1.2 연구의 제한점

이 연구는 몇 가지 제한점을 가지고 있다. 첫째, 도서 표제를 대상으로 머신러닝 기법을 이용하여 분류가 가능한지 여부를 실험하기 위해 제한된 양의 데이터를 대상으로 적용하였기 때문에 실제 현장에 적용할 때 약간의 차이가 있을 수 있다. 특히, 기존의 도서에 대한 정보는 KORMARC 레코드에서 수집할 수 있는데, 여기에서 키워드 정보가 거의 포함되어 있지 않아 표제만을 분석의 대상으로 삼았다. 둘째, 실제 도서관의 현장과 유사하게 만들기 위해 일정기간에 수입된 도서를 대상으로 하였기 때문

에 분류 범주별로 도서의 양이 차이가 많아 분류 작업에 영향을 미칠 수 있다. 셋째, 표제 분류 정확도 향상을 위해 연구자가 임의로 불용어 리스트를 만들어 사용하였기 때문에 분류에 있어서 실제와 차이가 있을 수 있다. 넷째, 파이썬의 사이킷런 라이브러리의 머신러닝 모듈을 사용하였기 때문에 한글 텍스트 분석을 위한 다른 알고리즘과 차이가 있을 수 있다.

1.3 선행연구

머신러닝 기법을 이용하여 도서관 분류법에 기반한 도서 분류 연구가 스웨덴에서 수행되었다. Golub, Hagelback, Ardo(2018)는 MARC 레코드에 포함된 표제와 키워드를 활용하여 분류의 세분화에 따른 분류 정확도를 연구하였다. 이들은 서포트 벡터 머신 모델(SVM)과 다항 나이브 베이즈 모델을 이용하여 DDC로 분류된 도서를 분류(3자리)한 후, 분류의 정확도를 측정하였다. 표제만 사용하였을 경우, SVM 모델의 정확도는 40.91%, 다항 나이브 베이즈 모델은 34.89%를 나타내었다. 하지만 표제를 기준으로 강목(2자리)까지만 분류를 하였을 경우는 분류 정확도가 SVM 모델은 57.99%, 다항 나이브 베이즈 모델은 54.21%로 증가하였다고 보고하였다. 반면에 키워드를 사용하여 강목까지 분류한 경우는 SVM 모델의 정확도가 79.56%, 다항 나이브 베이즈 모델이 77.25%로 증가하고 있음을 보여주었다.

국내 문헌정보학분야에서 텍스트 분류를 위해 머신러닝 기법을 적용한 연구가 있다. 조현양(2017)은 서평을 이용한 도서추천시스템 구축을 위해 머신러닝 기법인 LIBLINEAR 모델과 LibSVM 모델을 적용하여 성능을 측정하였다. 성능 측정 결과 tf-idf 가중치를 적용한 LIBLINEAR 모델이 역문헌빈도를 적용한 LibSVM 모델보다 더 나은 성능을 보여주었다고 발표하였다. 김성희, 엄재은(2008)은 미국 국립의학도서관 자료의 표제를 대상으로 MeSH의 8개 주제를 기준으로 tf-idf 가중치를 적용하여 신경망 모델, C5.0 및 CHAID 결정트리 모델, k-NN 모델의 4가지 모델을 이용하여 각각의 정확도를 비교 분석하였다. 분석 결과 신경망과 C5.0 결정트리 알고리즘을 병행한 모델이 가장 높은 정확도를 보였다고 발표하였다.

컴퓨터과학 관련 분야에서 한글 텍스트 분류에 관한 연구가 있다. 이철성 외(2013)는 트위트의 한글 텍스트를 다항 나이브 베이즈 모델과 SVM 모델을 이용하여 7가지 감정으로 분류하였다. 분류 결과의 정확도는 SVM 모델이 더 높은 것으로 파악되었다. 이와 유사하게 한명호, 류주현, 서수영(2014)은 SVM 모델을 이용하여 인터넷 포털 사이트의 영화평을 7가지 감정으로 분류하는 작업을 수행하였다.

2. 연구 방법

2.1 분석 대상

분석 대상 데이터는 서울에 위치한 A 공공도서관에서 소장하고 있는 KORMARC 레코드에서 추출한 2,000건의 표제 필드와 분류기호이다. 입수된 KORMARC 레코드는 KDC 300대(사회과학) 1,000건과 600대(예술) 1,000건이며, 표제, 저자명, 출판연도, 청구기호가 포함되어

있었다. 이 레코드에서 표제(표제관련정보 포함)와 분류기호만 추출한 후, 분류기호를 강목(예, 300, 310, ... 690)과 요목(예, 300, 304, ... 699)으로 변환하여 csv 파일 형식으로 저장한 후 분석에 사용하였다.

2.2 분석 도구 및 절차

데이터 분석은 아나콘다(Anaconda Navigator 4.9.2) 플랫폼의 쥬피터 노트북(Jupyter Notebook 6.1.4)을 통하여 파이썬(Python 3.8.5)의 싸이킷런(scikit-learn 0.23.2) 라이브러리를 이용하여 실행하였다(Muller & Guido, 2017). 분석 절차는 크게 여섯 단계로 구분할 수 있다.

첫 번째 단계에서는 분류 범주 수와 범주에 포함된 레코드의 수가 머신러닝 기반의 분류 정확도에 미치는 영향을 조사하기 위해 분석 대상 파일을 3가지로 구분하였다. 첫째 파일은 분류 기호를 KDC 요목별(3자리 분류기호)로 구분하여 분류 범주 수를 최대로 하고, 둘째 파일은 KDC 강목별(2자리 분류기호)로 표제를 구분하여 분류 범주 수를 중간으로 하고, 셋째 파일은 강목별로 구분한 파일에서 분류 범주에 포함된 레코드 수가 적은 범주를 제거하여 10가지 분류 범주로 구분하여, 분류 범주의 수를 최소화하였다.

두 번째 단계에서는 첫 번째 단계를 기반으로 전체 표제의 수와 각 표제에 포함된 단어의 수를 분석하였다.

세 번째 단계에서는 표제 분류에 의미가 있는 단어를 파악하기 위해 KoNLPy 한국어 형태소 분석기와 Okt 클래스를 사용하여 표제를 한글 형태소로 분석하였다. 한글 형태소 분석 결과를 통해 명사형과 동사형 단어를 남기고, 의미를 가지지 않는 단어는 모두 불용어 처리하여 표제에서 제거하였다.

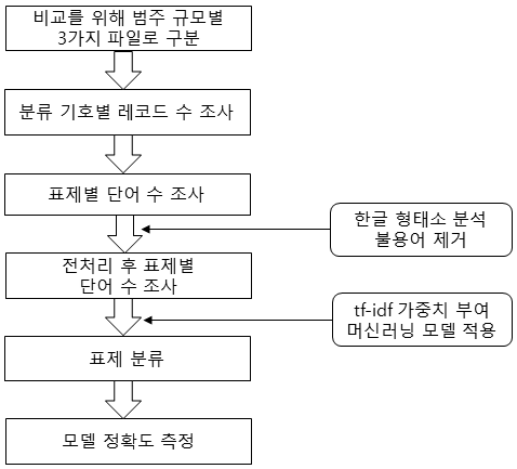
네 번째 단계에서는 불용어([부록] 참조)가 제거된 단어를 대상으로 다시 표제별 단어 수를 파악하였다. 그리고 단어별 tf-idf 가중치를 부여하기 위해, feature_extraction.text 모듈에 포함된 TfidfVectorizer 함수를 이용하여 표제의 단어를 수치 벡터로 변환시킨 후, 머신러닝 모델을 적용하였다.

다섯 번째 단계에서는 6가지 머신러닝 모델을 이용하여 표제를 분류하였다. 표제 분류를 위해 학습용 데이터 세트와 평가용 데이터 세트를 model_selection 모듈의 train_test_split 함수를 이용하여 70%와 30% 비율로 분배하였다. 분석에 사용된 모델은 다음과 같다. 첫째, 다항 로지스틱 회귀 모델은 linear_model 모듈의 LogisticRegression함수를 이용(multi_class ='multinomial' 적용)하였고, 둘째, 다항 나이브 베이즈 모델은 naive_bayes 모듈의 MultinomialNB 함수를 이용하였고, 셋째, 의사결정트리 모델은 tree 모듈의 DecisionTreeClassifier 함수를 이용하였고, 넷째, 랜덤 포레스트 모델은 ensemble 모듈의 RandomForestClassifier 함수를 이용하였고, 다섯째, SVM 모델은 svm 모듈의 SVC 함수를 이용하였고, 여섯째, 신경망 모델은 neural_network 모듈의 MLPClassifier 함수를 이용하였다(Aggarwal, 2018)(〈표 1〉 참조).

그리고 여섯 번째 단계에서는 각 머신러닝 모델별 표제 분류의 정확도를 측정하였다. 정확도는 학습용 데이터 세트의 정확도와 평가용 데이터 세트의 정확도로 구분하여 측정한 후, 각 모델별로 비교하였다(〈그림 1〉 참조).

〈표 1〉 사용된 모듈 및 함수

모델	사용된 모듈 및 함수
다항 로지스틱 회귀	from sklearn.linear_model import LogisticRegression LRclassifier = LogisticRegression(multi_class='multinomial') LRclassifier.fit(x_train, y_train.ravel()) LRclassifier.score(x_train, y_train) LRclassifier.score(x_test, y_test)
다항 나이브 베이즈	from sklearn.naive_bayes import MultinomialNB NBmodel = MultinomialNB() NBmodel.fit(x_train, y_train.ravel()) NBmodel.score(x_tran, y_train) NBmodel.score(x_test, y_test)
의사결정트리	from sklearn.tree import DecisionTreeClassifier Dtree = DecisionTreeClassifier(random_state = 0) Dtree.fit(x_train, y_train) Dtree.score(x_train, y_train) Dtree.score(x_test, y_test)
랜덤 포레스트	from sklearn.ensemble import RandomForestClassifier RForest = RandomForestClassifier(n_estimators = 5, random_state = 2) RForest.fit(x_train, y_train.ravel()) RForest.score(x_train, y_train) RForest.score(x_test, y_test)
서포트 벡터 머신(SVM)	from sklearn.svm import SVC svm = SVC(kernel = 'rbf', C = 10, gamma = 0.1) svm.fit(x_train, y_train) svm.score(x_train, y_train) svm.score(x_test, y_test)
신경망	from sklearn.neural_network import MLPClassifier MLP = MLPClassifier(hidden_layer_sizes=(150,), activation = 'logistic', solver = 'lbfgs', alpha = 0.01, batch_size = 32, max_iter = 500) MLP.fit(x_train, y_train.ravel()) MLP.score(x_train, y_train) MLP.score(x_test, y_test)



〈그림 1〉 분석 절차

3. 분석 결과

3.1 기본 정보

3.1.1 분류 범주별 레코드 수

분류 범주별 표제 수를 조사한 결과 KDC 요목별로 구분할 경우, 분류 범주 당 5건 이상의 표제가 포함된 범주는 68개이며, 전체 표제 수는 총 1,917건이었다(〈표 2〉 참조). 분류 범주별로 비교할 때, 분류기호 325에 포함된 표제가 379건으로 가장 많았고, 다음으로 327에 90건, 670에 72건 순으로 분포되어 있었다. 300대에 분류 범주가 27개, 600대에 분류 범주가 41개로 600대에 분류 범주가 더 넓게 분포되어 있었다. 이는 600대와 비교했을 때, 300대에는 특정 분류 범주에 표제가 상대적으로 많이 포함되어 있다는 의미이다.

〈표 2〉 KDC 요목별 레코드 수

분류기호	300	304	309							계
레코드 수		10	23							33
분류기호	320	320	321	322	324	325	326	327	329	계
레코드 수		60	32	12	9	379	12	90	5	599
분류기호	330	331	332	334	337	338				계
레코드 수		79	13	16	11	21				140
분류기호	340	340	342	349						계
레코드 수		67	9	11						87
분류기호	360	360	367							계
레코드 수		10	6							16
분류기호	370	370	373	375	377	378				계
레코드 수		9	31	7	7	20				74
분류기호	380	381								계
레코드 수		5								5
분류기호	390	392								계
레코드 수		5								5
분류기호	600	600	601	606	607	609				계
레코드 수		56	31	13	7	54				161
분류기호	610	610	617	619						계
레코드 수		42	24	9						75
분류기호	630	630	631	634	636	637	638	639		계
레코드 수		8	6	14	14	5	10	17		74
분류기호	650	650	651	652	653	654	656	657	658	계
레코드 수		62	8	5	33	10	27	66	52	263
분류기호	660	660	662	668						계
레코드 수		23	19	9						51
분류기호	670	670	671	673	674	675	676	677		계
레코드 수		72	15	27	17	5	5	13		154
분류기호	680	681	688							계
레코드 수		5	45							50
분류기호	690	691	692	694	695	698	699			계
레코드 수		36	8	6	55	7	18			130
합계										1,917

여기에서 분석 대상이 1,917건으로 줄어든 이유는 입수한 KORMARC 레코드 2,000건 중 각 분류 범주에 포함된 표제의 수가 4건 이하인 분류 범주는 분석에서 제외하였기 때문이다. 머신러닝 기법을 이용하여 표제를 분류할 때, 데이터 세트를 학습용 데이터 세트와 평가용 데이터 세트로 구분한 후 학습용 데이터 세트를 학습시키는데, 이를 위해 최소 표제 수가 2건 이상이 되어야 한다. 그러므로 보다 명확하게 정확도 측정을 위해 분류 범주 당 표제의 수가 4건 이하는 제거하였다.

다음으로 분류 범주의 수를 줄이고 범주별 표제 수를 증가시키기 위해 표제를 KDC 강목별로 구분하였다. KDC 강목을 기준으로 표제 분류에 사용된 표제 수는 총 1,992건이었다. 표제를 강목별로 구분한 결과, 300대에 있어서는 320에 603건, 330에 148건의 표제가 포함되어 있었고, 380과 390에는 각각 6건의 표제만 포함되어 있어 분류 범주별 표제 수의 차이가 컸다. 600대에는 650에 263건, 600[1]에 165건, 690에 136건의 표제가 포함되어 있었다. 또한 300대에는 310과 350에 표제가 각각 한 건밖에 포함하고 있지 않았고, 600대의 610에는 2건 모두 영어 표제이었고, 640은 표제가 4건밖에 포함되어 있지 않아 이들을 제거하였다(〈표 3〉 참조).

다음으로 KDC 강목별로 구분한 파일에서 각 분류 범주에 80건 이상의 표제가 포함된 새로운 파일을 만들었다. 이 결과 전체 분류 범주는 10가지로 감소되었고, 전체 표제 수는 1,807건이었다(〈표 4〉 참조).

표제 분석에 사용된 표제와 분류기호의 예는 〈표 5〉와 같다.

3.1.2 표제별 단어 수

각 표제에 포함된 단어 수를 분석한 결과, 한 표제에 포함된 가장 많은 단어의 수는 19개이며, 가장 적은 단어의 수는 1개이었다. 각 표제의 단어 수에 대한 평균 값은 4.98개이며, 중간 값은 4.0개이었다. 또한 표제의 단어 수를 나열하였을 때, 상위 25%에 해당되는 표제의 단어 수는 7.0개이었고, 75% 이상의 표제가 3.0개 이상의 단어를 포함하고 있었다. 이러한 단어 수의 다양성은 표제에서 본표제 외에 표제관련

〈표 3〉 KDC 강목별 레코드 수

분류기호	300	320	330	340	360	370	380	390	계
레코드 수	35	603	148	91	26	83	6	6	998
분류기호	600	610	630	650	660	670	680	690	계
레코드 수	165	80	80	263	52	158	60	136	994

〈표 4〉 10개 분류 범주별 레코드 수

분류기호	320	330	340	370	600	610	630	650	670	690	계
레코드 수	603	148	91	83	165	80	80	263	158	136	1,807

1) 분류기호가 600, 601, 606, 607, 609에 해당하는 표제임.

〈표 5〉 표제 사례

id	표제	분류기호
0	권력이란 무엇인가 =Was ist macht?	300.1
1	복잡성 사고 입문	301
2	광기와 천재: 루소에서 히틀러까지 문제적 열정의 내면 풍경	304
3	욕망해도 괜찮아: 나와 세상을 바꾸는 유쾌한 탈선 프로젝트	304
...	...	...
1993	수집의 즐거움: 평범한 사람들의 특별한 수집 이야기	699.93
1994	쁘띠 플라워: 내 방에 작은 정원	699.95
1995	작은 꽃다발 책: 간단하고 세련된 꽃다발 만들기 레슨	699.95
1996	블레스유 플라워 레슨 =Bless you flower lesson	699.95
1997	플라워 레시피 북	699.95

정보를 가지고 있는지 여부에 따라 나타난다고 판단할 수 있다(〈표 6〉 참조).

표제를 워드 클라우드 기법을 이용하여 그림을 그리면 다음과 같다. 표제에 포함된 단어는 '위한', '그림', '디자인', '그림', '왜', '읽는' 등이 많이 포함되어 있는 것을 알 수 있다. 이들 단어 중 일부는 KDC 분류와 직접적으로 상관이 없으므로 한글 형태소 분석기를 사용하여 표제에서 제거하였다(〈그림 2〉 참조).

〈표 6〉 표제별 단어 수의 분석

최대 값	최소 값	평균 값	표준편차	중간 값	제1사분위	제3사분위
19	1	4.98	2.77	4.0	3.0	7.0

〈그림 2〉 표제의 워드 클라우드

3.2 표제 전처리

3.2.1 전처리 후 표제 사례

한글 형태소 분석기를 사용하여 단어를 구분하고, 주제와 밀접한 관련이 없는 단어를 불용어 처리하였다. 전처리 후 다시 표제를 나열하면, 전처리 전 표제에 있던 일부 형용사형 단어는 동사형으로 변형되었고, 복합어가 분리되었으며, 불용어에 포함된 단어들과 영어, 기호들이 제거되었다. 하지만, 전처리 과정에서 일부 오류도 보인다. 예를 들면, 표제가 '욕망해도 괜찮아: 나와 세상을 바꾸는 유쾌한 탈선 프로젝트'는 '욕망 괜찮다 나오다 세상 바꾸다 유쾌하다 탈선 프로젝트'로 변환되었다. 여기에서 '나와'가 '나오다'로 잘못 변환된 것을 볼 수 있다(〈표 7〉 참조).

3.2.2 요목별 표제 단어 수

요목별 파일의 각 표제의 단어 수를 분석한 결과, 표제별 단어의 수는 20개가 가장 많았고, 한 단어로 되어 있는 표제도 있었다. 표제별 단어 수의 최대 값이 전처리 전에는 19개이었는데, 전처리 후에는 20개로 증가하였다. 이는 한글 형태소 분석기가 단어를 처리하는 과정에서 복합어가 2개 단어로 나뉜 결과라 추정한다. 표제의 단어 수의 중간 값은 4.0개이며, 단어 수를 기준으로 상위 25%는 표제에 단어가 6.0개 포함되어 있었고, 75% 이상의 표제에 3.0개 이상의 단어를 포함하고 있었다. 표제의 전처리 전과 비교하면 상위 25%의 표제의 단어 수가 7.0개에서 6.0개로 감소된 것을 알 수 있다(〈표 8〉 참조).

워드 클라우드 기법을 이용하여 전처리 한 표제의 단어를 출현 빈도 순으로 살펴보면, '읽다', '그림', '미술' 등의 출현 빈도가 높은 것을 알 수 있다. 이는 전처리 이전에 많이 나타났던 '위한', '왜' 등이 삭제되고 '읽다'를 제외하고는 명사형 단어가 많이 나타나고 있다(〈그림 3〉 참조).

〈표 7〉 전처리 후 표제 사례

id	title	class
0	권력	300
1	복잡 사고 입문	301
2	광기 천재 루소 히틀러 문제 열정 내면 풍경	304
3	욕망 괜찮다 나오다 세상 바꾸다 유쾌하다 탈선 프로젝트	304
...	...	...
1989	수집 즐거움 평범하다 사람 특별하다 수집 이야기	699
1990	쁘띠 플라워 방 작다 정원	699
1991	작다 꽃다발 책 간단하다 세련되다 꽃다발 만들기 레슨	699
1992	블레스 유 플라워 레슨	699
1993	플라워 레시피 북	699

〈표 8〉 전처리 후 요목별 표제 단어 수 분석

최대 값	최소 값	평균 값	표준편차	중간 값	제1사분위	제3사분위
20	1	4.39	2.43	4.0	3.0	6.0

〈그림 3〉 전처리 후 표제의 워드클라우드(요목별)

3.2.3 강목별 표제 단어 수

강목별 표제 단어 수는 요목별 표제 단어 수와 약간씩 차이가 있었다. 표제의 단어 수의 중간 값은 4.0개이며, 단어 수를 기준으로 상위 25%는 표제에 단어가 6.0개 포함되어 있었고, 75% 이상의 표제에 2.25개 이상의 단어를 포함하고 있었다. 요목별 표제 단어 수와 비교하면 강목별 표제 단어 수의 평균 값이 4.39개에서 4.38개로, 표준편차는 2.43개에서 2.42개로 감소되었다. 동시에 제1사분위 값도 3.0개에서 2.25개로 줄어들었다(〈표 9〉 참조).

워드 클라우드 기법을 이용하여 단어의 수를 시각화한 결과를 보면, '읽다'를 제외하고는 예술 분야의 '그림', '미술', '디자인' 등의 단어가 많이 출현하고 있지만, 사회과학 분야의 단어 중 출현빈도가 높은 단어는 거의 보이지 않는다. 즉, 예술 분야의 경우 많은 표제에서 동일한 단어가 사용되지만, 사회과학 분야에서는 표제에 동일한 단어가 사용되지 않고 다양한 단어가 사용되고 있음을 알 수 있다(〈그림 4〉 참조).

3.2.4 10개 범주별 표제 단어 수

10개 범주별 표제 단어 수는 강목별 범주의 표제에 포함하고 있는 단어 수와 비교할 때, 평균값은 4.38개에서 4.40개, 표준편차는 2.42개에서 2.43개, 제1사분위는 2.25개에서 3.0개로 약간씩 단어 수가 증가하는 양상을 보여주고 있었다(〈표 10〉 참조).

〈표 9〉 전처리 후 강목별 표제 단어 수 분석

최대 값	최소 값	평균 값	표준편차	중간 값	제1사분위	제3사분위
20	1	4.38	2.42	4.0	2.25	6.0

〈그림 4〉 전처리 후 표제의 워드 클라우드(강목별)

〈표 10〉 전처리 후 10개 범주별 표제 단어 수 분석

최대 값	최소 값	평균 값	표준편차	중간 값	제1사분위	제3사분위
20	1	4.40	2.43	4.0	3.0	6.0

〈그림 5〉 전처리 후 표제의 워드 클라우드(10개 범주)

워드 클라우드 기법을 이용하여 단어 출현빈도를 시각화한 결과, '그림', '디자인', '미술' 등의 단어의 출현빈도가 높아졌다. 강목별 분류기호 범주와 비교하면, '읽다' 단어의 수가 상대적으로 감소하였고, '디자인'과 '그림'의 단어 수가 비슷해진 것을 알 수 있다(〈그림 5〉 참조).

3.3 표제 분류의 정확도

3.3.1 요목별 분류

각각의 머신러닝 모델의 표제 분류 정확도를 측정을 위해 score 함수를 이용하였다. 학습용 데이터 세트와 평가용 데이터 세트의 정확도를 각각 측정하였는데, 학습용 데이터 세트의 정확도는 해당 모델이 학습용 데이터 세트를 얼마나 정확하게 학습하는지 여부를 평가한다. 평가용 데이터 세트 정확도는 학습용 데이터 세트에서 학습한 결과 만들어진 모델을 평가용 데이터 세트에 적용할 때, 표제가 실제 분류기호와 얼마나 일치하는지 측정한다. 일반적으로 학습용 정확도와 평가용 정확도가 차이가 크면 클수록 해당 모델이 학습용 데이터 세트에 과대적합하다고 한다. 즉 모델이 너무 학습용 데이터 세트 위주로 만들어져 이 모델을 실제에 적용하여 분류를 할 경우에 정확도가 낮다는 것이다.

6가지 머신러닝 모델을 이용하여 표제를 분류한 결과, 평가용 정확도가 높은 모델은 신경망 모델(50%)과 SVM 모델(49%)이었고, 다음으로 의사결정트리 모델(45%), 랜덤 포레스트 모델(44%), 다항 로지스틱 회귀 모델(33%), 다항 나이브 베이즈 모델(22%) 순이었다. 학습용 정확도와 평가용 정확도가 차이가 적은 모델은 다항 나이브 베이즈 모델(4%)이었고, 가장 큰 모델은 의사결정트리 모델(54%)이었다. 평가용 정확도가 높은 모델들의 경우 학습용 정확도와 평가용 정확도 차이는 44~54% 수준이었다. 비록 이들 머신러닝 모델들에 있어서 평가용 정확도가 높다 하더라도 정확도가 50% 이하에 불과하였다. 다른 한편으로 다항 로지스틱 회귀 모델과 다항 나이브 베이즈 모델의 학습용 정확도는 44%와 26%에 불과하여 표제 분류에 대한 학습이 거의 이루어지 않았다고 볼 수 있다(〈표 11〉 참조).

3.3.2 강목별 분류

강목별 분류에 있어서 평가용 정확도가 가장 높은 모델은 요목별 분류와 마찬가지로 신경망 모델[2](64%)이었고, 다음으로 SVM 모델(63%)이었다. 가장 낮은 모델은 다항 나이브 베이즈 모델로 정확도가 43%이었지만, 요목별 분류와 비교해서 정확도가 약 20% 이상 증가하였다. 6가지 모델 중에서 의사결정트리, SVM, 신경망 모델의 학습용 정확도는 거의 100%에 가까워 잘 학습되었다고 판단된다. 하지만, 평가용 정확도와 차이가 커서 학습 모델이 학습용 데이터 세트에 과대적합 되었다고 판단할 수 있다.

〈표 11〉 머신러닝 모델의 표제 분류 정확도(요목별 적용)

모델	다항 로지스틱 회귀	다항 나이브 베이즈	의사결정 트리	랜덤 포레스트	서포트 벡터 머신(SVM)	신경망
학습용	0.44	0.26	0.99	0.93	0.98	0.99
평가용	0.32	0.22	0.45	0.44	0.49	0.50

2) 사용된 코드: MLPClassifier(hidden_layer_sizes=(200,), activation = 'logistic', alpha = 0.01, batch_size = 32, max_iter = 500)

이 결과를 선행연구에서 인용한 스웨덴의 사례와 비교하면, 스웨덴 연구에서 SVM 모델의 정확도는 57.99%이었지만, 이 연구에서는 63%로 다소 높았고, 나이브 베이즈 모델은 54.21%이었지만, 이 연구에서는 43%로 차이가 있었다. 이러한 차이는 사용된 데이터의 특징으로 인한 결과라 판단된다. 즉, 사용된 언어의 차이점, 주제 분야의 차이점, 그리고 사용된 데이터의 양의 차이가 다소 다른 결과를 보여주었다고 판단된다(〈표 12〉 참조).

3.3.3 10개 범주별 분류

KDC 강목별 분류 범주를 10개 분류 범주로 축소한 후, 6가지 머신러닝 분류 모델을 적용한 결과, 분류 정확도가 이전과 약간의 차이가 있었다. 학습용 정확도는 전반적으로 상승하거나 일정하였지만, 평가용 정확도는 일정하지 않았다. 다항 로지스틱 회귀 모델의 경우 학습용 정확도는 분류 범주가 축소되면서 6% 증가했고, 평가용 정확도도 1% 증가했다. 다항 나이브 베이즈 모델 역시 학습용 정확도가 8% 증가했고, 평가용 정확도도 1% 증가하였다. 하지만, 의사결정트리 모델은 변화가 없었다. 랜덤 포레스트 모델은 학습용 정확도가 1% 증가했고, 평가용 정확도는 3% 증가하였다. 반면에 SVM은 학습용 정확도는 1% 증가했지만, 평가용 정확도는 3% 감소했다. 신경망 모델은 학습용 정확도는 동일하였지만, 평가용 정확도는 1% 감소하였다. 그럼에도 불구하고 신경망 모델의 평가용 정확도가 63%이었다(〈표 13〉 참조).

3.3.4 분류 범주 크기 별 평가용 정확도 비교

분류 범주의 수를 기준으로 머신러닝 모델들의 정확도를 비교한 결과, 모델별로 다양한 차이점을 보여주고 있다. 분류 범주의 수가 적어지고 범주에 포함된 표제 수가 많아지면서, 평가용 정확도는 10~20% 증가하였다. 특히, 분류 범주가 요목별 68개에서 강목별 16개로 감소하면서 6가지 머신러닝 모델 모두의 평가용 정확도가 증가하였다. 하지만, 분류 범주가 강목별 16개에서 10개로 감소할 때는 일부 모델의 정확도는 다양하게 변화되었다.

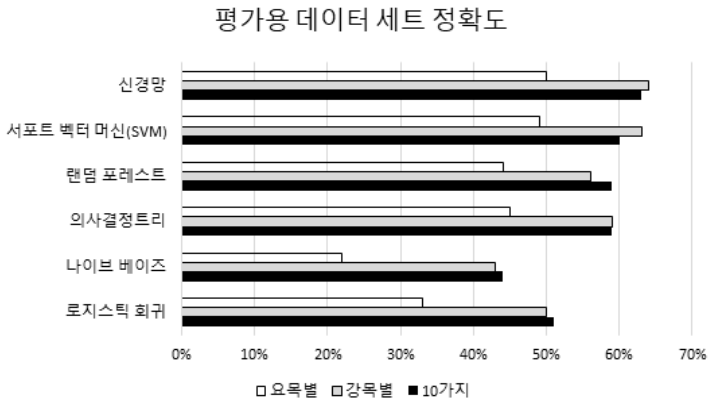
머신 러닝 모델별로 살펴보면, 다항 로지스틱 회귀 모델, 다항 나이브 베이즈 모델, 랜덤

〈표 12〉 머신러닝 모델의 표제 분류 정확도(강목별 분류)

모델	다항 로지스틱 회귀	다항 나이브 베이즈	의사결정 트리	랜덤 포레스트	서포트 벡터 머신(SVM)	신경망
학습용	0.66	0.54	0.99	0.94	0.98	0.99
평가용	0.50	0.43	0.59	0.56	0.63	0.64

〈표 13〉 머신러닝 모델의 표제 분류 정확도(10개 범주 적용)

모델	다항 로지스틱 회귀	다항 나이브 베이즈	의사결정 트리	랜덤 포레스트	서포트 벡터 머신(SVM)	신경망
학습용	0.72	0.62	0.99	0.95	0.99	0.99
평가용	0.51	0.44	0.59	0.59	0.60	0.63



〈그림 6〉 분류 범주 크기 변화에 따른 모델별 평가용 정확도 변화

포레스트 모델은 분류 범주의 수가 감소하면서 정확도는 증가하였다. 의사결정트리 모델은 분류 범주 수가 감소하면서 정확도가 증가했지만, 범주 수가 강목별 16개에서 10개로 줄어들 때는 정확도가 동일하였다. 하지만, SVM과 신경망 모델은 분류 범주의 수가 요목별 68개에서 강목별 10개로 변화할 때는 정확도가 증가하였으나, 강목별 16개에서 10개로 축소할 때는 정확도가 오히려 감소하였다(〈그림 6〉 참조).

4. 시사점

연구 결과를 분석한 결과 공공도서관의 도서 분류를 위해 표제를 대상으로 머신러닝 적용 가능성은 있다고 판단된다. 하지만, 실제 공공도서관 현장에 도서 분류를 위해 머신러닝을 적용하기 위해서는 다음과 같은 시사점들을 고려할 필요가 있다.

첫째, 머신러닝 모델을 이용하여 도서 표제를 분류할 때, 의사결정트리, 랜덤 포레스트, 서포트 벡터 머신, 신경망 모델이 다항 로지스틱 회귀 모델, 다항 나이브 베이즈 모델에 비해 분류 정확도가 더 높았다. 이들 머신러닝 4가지 모델들은 요목별(68가지) 분류 범주의 데이터 세트를 사용하였을 때 대략 44~50%의 정확도를 보여주고, 강목별(16가지) 분류 범주에서는 정확도가 56~64%, 10가지의 분류 범주에서는 정확도가 59~63%이었다. 실제 공공도서관에 이 모델을 적용한다면 분류 범주 당 데이터의 수도 많아져서 표제 분류의 정확도가 증가할 수 있을 것으로 추정된다.

일반적으로 도서관 장서 수에 따라 차이가 있겠지만 일부 도서관에서는 분류 범주의 수가 수백 개 이상이 될 수 있다. 이 경우는 머신러닝 모델을 이용하여 도서를 분류하는 것이 쉽지 않을 수 있다. 하지만, 대규모 장서를 소장하고 있지 않는 중소규모의 도서관에서는 KDC 사용법에서 제시된 '본표를 그대로 사용하지 않고 간략하게 사용할 수 있다'를 적용할 수 있다(한국도서관협회, 2013). 이 방법을 적용한 도서관의 경우, 분류 범주 수가 많이 줄어들 수 있기 때문에 머신러닝 모델을 이용한 도서 분류 방법이 효과가 있을 것으로 판단된다.

둘째, 이 연구에서는 제한된 양의 데이터를 사용하였기 때문에 일반화할 수는 없지만, 분류 범주 수와 머신러닝 모델의 정확도는 반비례 관계를 보여주고 있었다. 요목별 68개, 강목별 16개, 10개 분류 범주로 구분하여 각 머신러닝 모델의 정확도를 비교한 결과, 요목별 68개로 구분된 분류 범주와 강목별 16개로 구분된 분류 범주를 비교하면 분류 범주 수가 줄어들면서 정확도가 증가하는 것을 보여주었다. 하지만 강목별 16개과 10개 범주 사이에서는 정확도 차이가 명확하지 않아, KDC 요목별 파일을 기준으로 300대(27개 범주)와 600대(41개 범주)로 구분하여 두 집단 간 정확도를 비교하였다. 비교 결과 분류 정확도는 분류 범주의 수가 적은 300대에서 약간 높게 나타났다. 300대에서 정확도는 6가지 머신러닝 모델 전체에서 40%~55%이었지만, 600대에서 정확도는 36~52%이었다. 이 결과는 강목별 16개 범주와 10개 분류 범주처럼 범주 수 차이가 적을 경우 정확도 차이가 명확하게 나타나지 않았지만, 분류 범주 수 차이가 커지면서 정확도의 차이를 명확하게 보여주고 있다.

그러나 여기에서 한 가지 고려해야 할 사항이 있다. 300대와 600대의 차이점은 범주의 수 외에도 각 분류 범주에 포함된 표제 수가 특정 분류 범주에 치우쳐져 있는 점이다. 300대의 325에는 379건의 표제가 포함되어 있었지만, 나머지 범주는 모두 100건 미만의 표제가 포함되어 있었다. 반면에 600대는 670에 가장 많은 72건의 표제를 포함하고 있었고, 나머지 범주에는 표제가 전반적으로 골고루 분포되어 있었다. 다시 표현하면 300대에서는 320에 포함된 표제가 전체 300대 표제의 60%를 차지하였고, 320과 330을 합하면 전체 300대의 75%를 차지하는 등 매우 한쪽으로 편향된 모습을 보여주고 있었다. 600대는 300대에 비교해서 다소 분산되어 있지만, 600대, 650대, 670대, 690대 4가지 범주가 600대 전체의 72%를 차지하고 있었다. 이처럼 전 분류 범주에 표제가 골고루 분산되지 않고 일부 범주에 치우치는 상황이 분류 정확도에 미치는 영향을 연구할 필요하다고 생각된다.

셋째, 공공도서관에서 도서 분류에 머신러닝 기법을 적용할 때, 여러 번 반복적으로 분류하는 방안을 연구할 필요가 있다. 대규모 공공도서관의 경우에 1차로 KDC 주류 혹은 강목(綱目, division)을 기준으로 분류한 후, 해당 주류 혹은 강목별로 세부적인 분류를 실시한다면 분류 범주가 많아져서 정확도가 감소하는 문제를 해결할 수 있을 것으로 추정한다.

또한 머신러닝 기법을 이용하여 주제별 분류 외에도 전체 주제의 도서를 대상으로 이용자 대상에 대한 분류, 예를 들면 아동용, 청소년용, 성인용 등으로 구분하거나 문학류의 장르를 구분할 수 있을 것이다. 특히 다항 로지스틱 회귀 모델이나 다항 나이브 베이즈 모델 등은 분류 범주가 많은 데이터 세트에서 상대적으로 정확도가 낮았지만, 분류 범주 수가 적은 경우에 정확도가 상당히 높기 때문에 이들 모델을 활용하는 방안을 연구할 필요가 있다.

넷째, 효과적인 도서 분류를 위해 표제에 사용되는 단어에 대한 조사가 필요하다. 도서 표제에 단어가 많이 포함되어 있다면 머신러닝 기법을 이용하여 표제를 분류하는데 이점이 될 수 있다. 이때는 분류 기준이 될 수 있는 범주 내에서만 동일한 단어들이 사용되어야 보다 정확한 분류가 가능할 것이다. 그렇지 않고 전 분류 범

주에서 골고루 사용된다면 분류의 정확도가 감소할 수 밖에 없을 것이다.

예를 들어, 워드 클라우드 결과에서, 그림, 한국, 미술 등의 단어가 상대적으로 출현빈도가 높은 것을 알 수 있다. '그림'이나 '미술'은 특정 범주에 집중적으로 나타나지만, '한국'은 전 주제 분야에 골고루 분포하고 있다. 즉, '그림'이나 '미술'은 표제 분류를 위한 중요한 역할을 할 수 있지만, '한국'은 표제 분류의 정확도를 감소시킬 수 있다.

표제에 사용되는 단어는 도서관 유형에 따라 차이가 있을 수 있다. 공공도서관에서 소장하고 있는 도서는 대학도서관이나 전문도서관과 달리 전문 도서보다는 교양 도서가 많다. 교양 도서는 도서의 특성상 일반 이용자가 쉽게 내용을 이해할 수 있도록 표제에 많은 단어를 사용하는 경향이 있다. 다시 말하면 본표제 외에도 표제관련정보를 통하여 해당 도서에 대한 부연 설명 혹은 도서에 관련된 다른 정보를 제공하고 있다. 반면에 전문 도서는 해당 주제를 명확하게 보여줄 수 있는 소수의 단어로 표제를 간략하게 표현하고 있다. 이러한 측면에서 도서관 유형별로 도서 표제에 포함된 단어의 특징에 대한 연구가 필요하다.

다섯째, 표제 토큰화와 불용어에 대한 연구가 필요하다. 한글은 영어와 달리 토큰화와 불용어 선택이 쉽지 않다. 이 연구에서는 한글 형태소 분석을 위해 KoNLPy 분석기와 Okt 클래스를 사용하였다. 하지만 표제 단어의 품사를 명확하게 구분하지는 못하였다. 이는 한글의 특징 때문에 나타난 결과라 판단되며, 이 분야에 대한 학제간 연구가 요구된다.

불용어에 대해 이 연구에서 연구자는 임의로 명확한 주제를 나타내지 못한다고 판단된 단어들을 불용어로 처리하였다. 하지만, 앞에서 본 '한국'의 예처럼 분류표에서 지역구분과 같이 보조표 영역에 속하는 단어들을 불용어로 처리한다면, 언어학의 한국어나 문학의 한국문학에 속하는 도서, 혹은 역사와 지리에서 한국에 관한 도서를 분류할 수 없는 상황이 될 수 있다. 반면에 불용어로 처리하지 않는다면, 또한 분류의 정확도에 영향을 미칠 수 있을 것이다. 그러므로 도서 표제를 대상으로 불용어 리스트에 대한 연구가 필요하다고 판단된다.

5. 결 론

이 연구의 목적은 공공도서관의 도서 분류를 위해 표제를 대상으로 한 머신러닝 기법의 적용 가능성을 조사하는데 있다. 이에 서울 지역에 위치한 A 공공도서관에 소장된 2,000건의 도서 표제를 분석 대상으로 하고, 다항 로지스틱 회귀 모델, 다항 나이브 베이즈 모델, 의사결정트리 모델, 랜덤 포레스트 모델, SVM 모델, 신경망 모델을 이용하여 각 모델의 분류 정확도를 측정하였다.

조사결과 공공도서관 도서 분류에 머신러닝 기법을 활용할 수 있는 가능성을 보여주었다. KDC의 요목별 68개 범주를 대상으로 분류한 결과, 의사결정트리, 랜덤 포레스트, SVM, 신경망 모델이 정확도가 44~50%로 높았고, 그중 신경망 모델이 50%로 다른 모델보다 약간 더 높았다. 하지만 다항 나이브 베이즈 모델은 정확도가 22%로 매우 낮았다. 강목별 16개 분류 범주로 표제를 분류할 경우, 6가지 모델의 정확도는

43~64%이며, 이중 정확도가 가장 높은 모델은 신경망 모델이었다. 10개 범주로 구분한 경우에 정확도는 44~63% 수준이었고, 정확도가 가장 높은 모델 역시 신경망 모델이었다.

머신러닝 기법을 이용하여 도서 표제 분류에 있어서 시사점은 다음과 같이 정리할 수 있다. 첫째, 의사결정트리, 랜덤 포레스트, SVM, 신경망 모델이 다항 로지스틱 회귀 모델과 다항 나이브 베이즈 모델을 비해 표제 분류 정확도가 더 높았다. 둘째, 분류 범주 수와 머신러닝 모델의 정확도는 반비례 관계를 보여주고 있었다. 셋째, 장서량이 많은 도서관의 경우, 여러 번 반복적으로 머신러닝을 적용하는 분류 방법의 연구가 필요하다. 넷째, 도서 표제의 단어에 대한 체계적인 조사가 필요하다. 다섯째, 표제 토큰화와 불용어에 대한 연구가 필요하다.

참 고 문 헌

김성희, 엄재은 (2008). 기계학습을 이용한 문서 자동분류에 관한 연구. 정보관리연구, 39(4), 47-66.

이철성, 최동회, 김성순, 강재우 (2013). 한글 마이크로블로그 텍스트의 감정분류 및 분석. 정보과학회 논문지: 데이타베이스, 40(3), 159-167.

조현양 (2017). 자동분류기반 성격 유형별 도서추천시스템 개발을 위한 실험적 연구. 한국도서관정보학회지, 48(2), 215-236.

한국도서관협회 (2013). 한국십진분류법 (제6판). 서울: 한국도서관협회.

한명호, 류주현, 서수영 (2014). 기계 학습을 이용한 한글 텍스트 감정 분류 및 분석. 한국정보과학회 학술발표논문집, 1722-1724.

Aggarwal, C. C. (2018). Machine Learning. Cham: Springer.

Golub, K., Hagelback, J., & Ardo, A. (2018). Automatic classification using DDC on the Swedish Union Catalogue. Proceedings of the 18th European Networked Knowledge Organization Systems (NKOS) Workshop co-located with the 22nd International Conference on Theory and Practice of Digital Libraries 2018 (TPDL 2018). Available: http://ceur-ws.org/Vol-2200/paper1.pdf

Muller, A. C. & Guido, S. (2017). Introduction to Machine Learning with Python. 박해선 옮김. (2019). 파이썬 라이브러리를 활용한 머신러닝 (번역개정판). 서울: 한빛미디어.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Cho, H. (2017). A experimental study on the development of a book recommendation system

using automatic classification, based on the personality type. Journal of Korean Library and Information Science Society, 48(2), 215-236.

Han, M, Rye, J., & Seo, S. (2014). Classification and analysis of emotion in Korean texts using machine learning. The Korean Institute of Information Scientists and Engineers, 1722-1724.

Kim, S. & Eom, J. (2008). A study on the documents's automatic classification using machine learning. Journal of Information Management, 39(4), 47-66.

Korean Library Association (2013). Korean Decimal Classification. 6th edition. Seoul: Korean Library Association.

Lee, C., Choi, D., Kim, S., & Kang, J. (2013). Classification and analysis of emotion in Korean microblog texts. Journal of KISS: Databases, 40(3), 159-167.

[부록] 불용어 리스트

'은', '는', '이', '가', '하', '아', '것', '들','의', '있', '되', '수', '보', '주', '등', '한', '을', '하다', '과', '를', '에', '으로', '와', '에서', '로', '나', '보다', '그', '내', '적', '가지', '왜', '년', '있다', '않다', '년', '까지', '다', '없다', '싶다', '이다', '라', '속', '인', '인가', '어떻다', '좋다', '서', '대', '꼭', '고', '하라', '인가', '어떻다', '서', '대', '꼭', '고', '하라', '부터', '무엇', '그리고', '되다', '에게', '제', '더', '도', '오다', '자다', '현', '처럼', '만', '따르다', '다시', '전', '누구', '크다', '알', '대다', '하고', '지금', '옛', '살', '과의', '세', '권', '쓰다', '사', '지', '관', '장', '분', '간', '담다', '론', '들다', '쉽다', '비다', '구','자', '밉다', '짓다', '만난', '빠지면', '부의', '너', '리', '초', '네', '즈', '나다', '리다', '움', '두', '이란', '되어다', '모르다', '아니다', '대한', '니', '때', '받다', '중', '인의', '작', '주년', '와의', '단', '치다', '경', '저', '파다', '타다', '기', '허다', '번째', '재', '풀다', '너무', '션', '뒤지다', '사로자다', '안', '자기', '에는', '후', '넘다', '란', '랑', '뜨다', '바로', '팅', '날', '오', '스', '년후', '워', '뒤', '빌다', '나은', '드럽다', '잘', '인데', '니까', '곳', '가다', '온', '누가', '면', '넌', '아직', '한번', '스스로', '내다', '깨우다', '프레', '테이', '좀', '사이', '간단', '존', '다른', '두다', '통해', '살기', '위해', '석', '드', '옆', '반드시', '록', '아주', '갖다', '타', '래', '써다', '뭐', '찍기', '파다', '아라', '길다', '늘다', '늦다', '무얼', '력', '주다', '이라는', '실로', '이번', '엔', '뿐', '걸다', '세우다', '말고', '주기', '체', '게', '어', '년차', '건', '위한', '그리고', '않는', '싶은', '우리', '위'