텍스트 분류를 위한 자질 순위화 기법에 관한 연구

An Experimental Study on Feature Ranking Schemes for Text Classification

김판준 (Pan Jun Kim)*

초 록

본 연구는 텍스트 분류를 위한 효율적인 자질선정 방법으로 자질 순위화 기법의 성능을 구체적으로 검토하였다. 지금까지 자질 순위화 기법은 주로 문헌빈도에 기초한 경우가 대부분이며, 상대적으로 용어빈도를 사용한 경우는 많지 않았다. 따라서 텍스트 분류를 위한 자질선정 방법으로 용어빈도와 문헌빈도를 개별적으로 적용한 단일 순위화 기법들의 성능을 살펴본 다음, 양자를 함께 사용하는 조합 순위화 기법의 성능을 검토하였다. 구체적으로 두 개의 실험 문헌집단(Reuters-21578, 20NG)과 5개 분류기(SVM, NB, ROC, TRA, RNN)를 사용하는 환경에서 분류 실험을 진행하였고, 결과의 신뢰성 확보를 위해 5-fold cross validation과 t-test를 적용하였다. 결과적으로, 단일 순위화 기법으로는 문헌빈도 기반의 단일 순위화 기법(chi)이 전반적으로 좋은 성능을 보였다. 또한, 최고 성능의 단일 순위화 기법과 조합 순위화 기법 간에는 유의한 성능 차이가 없는 것으로 나타났다. 따라서 충분한 학습문헌을 확보할 수 있는 환경에서는 텍스트 분류의 자질선정 방법으로 문헌빈도 기반의 단일 순위화 기법(chi)을 사용하는 것이 보다 효율적이라 할 수 있다.

ABSTRACT

This study specifically reviewed the performance of the ranking schemes as an efficient feature selection method for text classification. Until now, feature ranking schemes are mostly based on document frequency, and relatively few cases have used the term frequency. Therefore, the performance of single ranking metrics using term frequency and document frequency individually was examined as a feature selection method for text classification, and then the performance of combination ranking schemes using both was reviewed. Specifically, a classification experiment was conducted in an environment using two data sets (Reuters-21578, 20NG) and five classifiers (SVM, NB, ROC, TRA, RNN), and to secure the reliability of the results, 5-Fold cross-validation and t-test were applied. As a result, as a single ranking scheme, the document frequency-based single ranking metric (chi) showed good performance overall. In addition, it was found that there was no significant difference between the highest-performance single ranking and the combination ranking schemes. Therefore, in an environment where sufficient learning documents can be secured in text classification, it is more efficient to use a single ranking metric (chi) based on document frequency as a feature selection method.

키워드: 텍스트 분류, 텍스트 범주화, 자질선정, 자질 순위화 기법, 문헌빈도, 용어빈도
text classification, text categorization, feature selection, feature ranking schemes, document frequency, term frequency

* 신라대학교 문헌정보학과 교수(pjkim@silla.ac.kr)

■ 논문접수일자 : 2023년 2월 1일 ■ 최초심사일자 : 2023년 3월 8일 ■ 게재확정일자 : 2023년 3월 17일
■ 정보관리학회지, 40(1), 1-21, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.1.001

1. 서 론

1.1 연구의 필요성 및 목적

전 세계적으로 디지털 텍스트의 양이 급속히 증가하면서 이러한 텍스트의 분류에 대한 중요성이 커지고 있다. 효율적인 색인 및 검색을 위해 텍스트는 사전 정의된 주제 또는 범주별로 분류되어야 할 필요가 있다. 그러나 엄청난 양의 텍스트 데이터를 수작업으로 처리하는 것은 불가능하며, 이를 자동으로 분류하는 많은 기술과 방법이 제안되었다(Deng et al., 2019). 텍스트 자동 분류 또는 텍스트 범주화는 미분류된 입력 문헌(자연어 텍스트)에 사전 정의된 범주(class)나 범주명(label)을 자동 할당하는 것이다(Sebastiani, 2002; Yang & Pedersen, 1997). 텍스트 분류의 대상이 되는 문헌집합은 대부분 수만 개의 고유한 단어를 포함하고 있다(Joachims, 2002; Wang et al., 2014). 이러한 고차원 데이터는 분류기의 실행 시간과 정확도 측면에서 성능을 저하시키는 근본적인 원인이 되고 있으며(Wang et al., 2016; Wu & Zhang, 2004), 정보를 제공하지 않는 0의 값을 가지는 경우가 많다(Su, Shirab, & Matwin, 2011). 또한, 텍스트 분류에 문헌집합에 출현한 모든 용어를 사용하면 분류에 도움이 되지 않는 용어들로 인해 좋지 않은 결과를 얻을 수 있는 것으로 알려져 있다(Liu & Yu, 2005; Sebastiani, 2002). 이로 인해 텍스트 분류에서 자질선정 및 추출과 같은 차원축소 기법은 학습의 정확도를 향상시킬 수 있어 유망한 데이터 사전처리 단계로 간주되고 있으며(Cai & Zhu, 2018), 특히 정확한 분류를 위해서는 자질선정이 가장 중요한 과제가 되고 있다(Iqbal et al., 2020).

텍스트 분류를 위한 목적으로는 최근까지 filter 기반의 자질선정에 관한 연구가 가장 활발하게 이루어졌다(Pintas, Fernandes, & Garcia, 2021). Filter 기반의 자질선정이 가장 선호되고 있는 이유는 분류기와 독립적으로 데이터의 속성에만 의존하는 관계로 단순성과 효율성 측면에서 다른 방법들에 비해 장점을 갖고 있기 때문이다(Agnihotri, Verma, & Tripathi, 2017; Cai et al., 2018; Kumar & Minz, 2014; Liu & Yu, 2005; Uysal, 2016). 기본적으로 filter 기반의 자질선정 기법은 출현빈도에 기초하여 각 자질을 순위화한 다음 상위의 자질을 선택한다(Abiodun et al., 2021). 이러한 과정에서 기존의 filter 기반 자질선정 기법은 주로 문헌빈도(df)를 사용하는 경우가 많았다(Forman, 2003; Rehman et al., 2018). 이들 기법은 저빈도어에 대한 신뢰성이 떨어짐은 물론 특정 용어의 한 문헌 내 출현여부 정보(df)만을 사용하면서 특정 문헌 내 용어의 출현정보(tf)는 무시한다는 문제가 있다(Wang et al., 2012). 그러나 용어빈도(tf)는 각 문헌 내 자질의 중요성을 나타내기 때문에 자질선정에서 중요한 정보라고 할 수 있다(Wu & Xu, 2015). 이에 따라 텍스트 분류를 위한 자질 순위화 기법으로 용어빈도를 고려한 다수의 연구들이 수행되었다(Ávila-Argüelles et al., 2010; Baccianella, Esuli, & Sebastiani, 2013; Wang et al., 2012; Wang et al., 2014; Wu & Xu, 2015; Yao et al., 2017).

본 연구는 텍스트 분류를 위한 효율적인 자질선정 방법으로 자질 순위화 기법의 성능을 구체적으로 검토해 보고자 하였다. 이를 위해, 먼저 용어빈도와 문헌빈도를 개별적으로 적용한 단

일 순위화 기법들의 성능을 살펴보았다. 다음으로 이전 실험에서 가장 좋은 성능을 보인 용어 빈도와 문헌빈도를 함께 사용하는 조합 순위화 기법의 성능을 검토하였다. 구체적으로 두 개의 실험 문헌집단(Reuters-21578, 20NG)과 5개 분류기(SVM, NB, ROC, TRA, RNN)를 사용하는 환경에서 텍스트 분류 실험을 수행하였고, 실험 결과의 신뢰성을 확보하기 위하여 5-fold 교차검증과 t-검정 방법을 적용하였다.

1.2 텍스트 분류를 위한 자질 순위화 기법

텍스트 분류에서 고려해야 할 주요 문제는 고차원의 자질공간으로 자질 수가 크게 증가할 수 있다는 것이며, 이는 분류기의 성능에 상당한 영향을 미친다. 따라서 텍스트 분류에서 자질공간의 차원을 줄이고 분류기의 성능을 향상시킬 수 있는 자질선정이 널리 사용되고 있다(Chang et al., 2015). 자질선정의 주요 목적은 분류 작업을 위한 가장 적합하고 식별력이 가장 높은 자질을 선정하여 자질공간의 차원을 줄이는 것이다(Shang et al., 2007).

자질선정은 분류기와의 관계에 따라 세 가지 유형으로 구분할 수 있다(Javed, Babri, & Saeed, 2010; Li, Li, & Liu, 2017; Sebastiani, 2002). 첫째, filter 방법은 데이터의 일반적인 특성에 의존하는 것으로 분류 알고리즘과 독립적인 자질선정 프로세스를 수행한다(Dash & Liu, 1997). 둘째, wrapper 방법은 자질 하위집합의 상대적 유용성을 평가하기 위한 분류기를 포함하며(Kohavi & John, 1997), wrapper가 최적의 자질 하위집합을 탐색하도록 할 수 있도록 한다. 셋째, embedded 방법은 분류기를 학습하는 과정에서 자질선정을 수행하는 것으로, 자질의 최적 하위집합에 대한 탐색이 분류기의 학습 과정에 포함된다(Guyon et al., 2002). 또한, 자질선정의 결과에 따라 자질 순위화 기법(feature ranking metrics)과 자질 하위집합 선정 알고리즘(feature subset selection algorithms)으로 구분할 수도 있다(Bolón-Canedo & Alonso-Betanzos, 2019). 이 중에서 자질 순위화 기법은 문헌들의 범주를 식별하기 위한 개별 자질의 가치를 추정하고 점수를 할당한다(Chen et al., 2009; Van Hulse, Koshgoftaar, & Napolitano, 2011). 다음에 이러한 점수의 내림차순으로 자질을 정렬하고 상위 k개 순위의 자질을 선정한다(Guyon & Elisseeff, 2003). 따라서 분류기와 독립적으로 자질선정을 수행하는 filter와 자질 순위화 기법(feature ranking scheme)은 거의 유사한 것으로 상호교환적으로 사용되는 경우가 많으며(Rehman et al., 2018), filter 방법의 하위 범주로 자질 순위화 기법과 하위집합 탐색 알고리즘을 세분하는 견해도 있었다(Lazar et al., 2012).

Filter 방법은 분류기와는 독립적으로 데이터의 속성에만 의존하기 때문에 컴퓨터 처리의 효율성(computational efficiency)과 과적합(overfitting) 문제를 줄일 수 있다는 장점이 있다(Abiodun et al., 2021; Venkatesh & Anuradha, 2019). 따라서 다른 두 가지 방법보다 선호되는 경향이 있어(Agnihotri, Verma, & Tripathi, 2017; Cai et al., 2018; Günal, 2012; Talavera, 2005; Uysal, 2016), 최근 가장 많은 연구가 이루어졌다(Pintas, Fernandes, & Garcia, 2021). 마찬가지로 자질 순위화 기법은 컴퓨터 처리의 효율성이라는 두드러진 장점으로 인해 텍스트 분류를 위한 자질선정 방법으

로 선호되고 있어(Joachims, 2002; Sebastiani, 2002), 최근까지 관련 연구가 활발하게 진행되었다(Chen et al., 2009; Forman, 2003; Parlak & Uysal, 2021; Rehman et al., 2018; Rehman, Javed, & Babri, 2017; Uysal & Gunal, 2012; Van Hulse, Hoshgoftaar, & Napolitano, 2011).

자질 순위화 기법은 데이터의 일반적인 특성에 기반하고 있으며 각 자질에 평가 함수를 적용하여 산출한 점수에 따라 가장 높은 순위의 자질을 선택한다(Abiodun et al., 2021). 여기서 평가함수는 단어의 용어빈도 또는 문헌빈도에 기초하여 개별 자질의 적합성(relevance)를 산출하기 위한 공식으로 표현될 수 있는데(Wang & Hong, 2019), 지금까지 텍스트 분류를 위한 자질 순위화 기법에서는 주로 문헌빈도를 사용하는 경우가 많았다(Baccianella, Esuli, & Sebastiani, 2013). 그러나 이들 기법은 저빈도어에 대해 신뢰성이 떨어짐은 물론 특정 용어의 한 문헌 내 출현정보만을 사용하면서 특정 문헌 내 용어의 출현정보는 무시한다는 문제가 있는 것으로 지적되었다(Wang et al., 2012). 또한, 텍스트 분류를 위한 자질 순위화 기법에서 용어빈도와 문헌빈도의 성능을 비교한 실험에서 용어빈도에 기반한 공식은 특히 더 작은 자질집합을 사용하는 경우에 유용한 것으로 보고된 바 있다(Azam & Yao, 2012). 이에 따라 본 연구는 텍스트 분류를 위한 효율적인 자질선정 방법으로 용어빈도와 문헌빈도에 기초한 자질 순위화 기법의 성능에 대하여 구체적으로 검토해 보았다.

1.3 연구 문제

텍스트 분류를 위한 효율적인 자질선정 방법으로서 용어빈도와 문헌빈도에 기초한 자질 순위화 기법에 대하여, 다음과 같이 두 가지 연구문제를 설정하고 실험을 수행하였다.

- 연구문제 1. 텍스트 분류를 위한 자질선정 방법으로 용어빈도와 문헌빈도에 기초한 단일 자질 순위화 기법들 간에 분류 성능의 차이가 있는가?
- 연구문제 2. 텍스트 분류를 위한 자질선정 방법으로 단일 자질 순위화 기법과 조합 자질 순위화 기법 간에 분류 성능의 차이가 있는가?

2. 실험 설계

2.1 실험 단계

본 연구에서 텍스트 분류를 위한 자질 순위화 기법의 성능을 검토하기 위한 실험 단계는 다음과 같다.

첫째, 연구문제1에 대한 실험은 2개의 문헌집합(Reuters-21578, 20NG)과 5개 분류기(NB, Roccihio, SVM, Transformer, RNN)를 사용하는 환경에서, 자질선정 방법으로 단일 순위화 기법 15개를 적용한 분류 성능을 검토하였다. 여기서 단일 순위화 기법은 각각 용어빈도에 기초한 4개(atf, itf, ltf, otf)와 문헌빈도에 기반한 11개 기법(acc, acc2, chi, df, gss, idf, jac, lor, mi, pcc, rf)으로 구분하여 실험을 수행하였다. 또한, 각 자질 순위화 기법으로 선정된 자질 수를 최소 100개부터 최대 2000개까지 100개씩 순차적으로 증가시킨 결과를 평균

하여 분류 성능을 산출하였다.

둘째, 연구문제2에 대한 실험은 2개의 문헌집합(Reuters-21578, 20NG)과 3개 분류기(NB, Roccihio, SVM)를 사용하는 환경에서, 이전 실험에서 가장 좋은 성능을 보인 용어빈도(atf)와 문헌빈도(chi, jac)를 함께 사용한 조합 순위화 기법(atf*chi, atf*jac)을 적용한 분류 성능을 검토하였다. 즉, 단일 순위화 기법 중에서 용어빈도(atf)와 문헌빈도(chi, jac)의 성능과 양자를 함께 사용한 조합 순위화 기법(atf*chi, atf*jac)의 성능을 비교하였다.

셋째, 연구문제2의 실험 결과에 대한 신뢰성을 확보하기 위하여 5-fold 교차검증과 t-검정을 수행하였다. 먼저, 단일 순위화 기법(chi, jac)과 조합 순위화 기법(atf*chi, atf*jac) 간의 성능 차이를 확인하기 위한 5-fold 교차검증을 수행하였다. 다음으로 단일 순위화 기법(chi. jac)와 조합 순위화 기법(atf*chi)의 성능 차이에 대한 t-검정을 수행하였다.

2.2 문헌집합

문헌집합은 텍스트 분류에서 가장 많이 사용되어 온 것으로 Reuters-21578과 20NG(20 Newsgroups, 이후 20NG로 표기)를 사용하였다(Pintas, Fernandes, & Garcia, 2021; Rehman et al., 2018; Uysal, 2016; Wang et al., 2012). 첫째, 텍스트 분류에 사용할 자질선정을 위한 사전처리 과정으로 불용어 제거(stop word removal), 어간 추출(porter stemming), 가지치기(pruning)를 수행하였다. 특히 가지치기(pruning)를 통해 저빈도어(df<=3)를 제거하였다(Forman, 2003). 둘째, 텍스트 분류를 위한 자질선정 연구에서 가장 많이 사용된 문헌표현(Bag of Words)과 가중치(tf*idf)를 적용하였다(Aggarwal & Zhai, 2012; Harish & Revanasiddappa, 2017; Joachims, 1996; Pintas, Fernandes, & Garcia, 2021). 그 결과로 생성된 문헌집합의 통계는 〈표 1〉과 같다. 여기서 Reuters-21578의 범주는 상위 10개, 20NG는 전체 20개 범주를 사용하였기 때문에 본 연구의 범주 집합은 비교적 충분한 수의 학습문헌(최소 224개 이상)을 갖는다.

2.3 분류기

텍스트 분류에서 자질선정 방법의 평가에 가장 많이 사용되어 온 3개 분류기(지지벡터기계

〈표 1〉 문헌집합 통계 -> data정보.xlsx/all

항목	Reuters-21578*	20NG**
문헌 수	8,599	19,997
자질 수(사전처리 이후)	8,375(7,948)	30,468(29,636)
범주 수	top10	20
범주당 문헌 수(최대/최소/평균)	3,776/224/955.5	1,000/997/999.9
범주 특성	Imbalanced	Balanced

* http://www.daviddlewis.com/resources/testcollections/reuters21578

** http://disi.unitn.it/moschitti/corpora.htm

/SVM, 나이브 베이즈/NB, 로치오/Rocchio)와 딥러닝 기반의 2개 분류기(순환신경망/RNN, 트랜스포머/Transformer)를 사용하였다. 먼저, 지지벡터기계(이후 SVM으로 표기)와 나이브 베이즈(이후 NB로 표기)는 텍스트 분류에서 우수한 성능을 보이는 것으로 가장 많이 사용되고 있으며(김판준, 2022; 육지희, 송민, 2018; Agnihotri, Verma, & Tripathi, 2017; Cai et al., 2018; Pintas, Fernandes, & Garcia, 2021), 로치오(이후 ROC로 표기)는 단순하면서도 비교적 좋은 성능을 보이는 것으로 알려져 있다(김판준, 2016; 2018; Han & Karypis, 2000). 본 연구에서 SVM은 scikit-learn 모듈의 Linear SVC 분류기를 적용하였고, 스케일러(scaler)는 최대절대값(maximum absolute value)을 적용하였다. NB는 scikit-learn 모듈의 다항분포 나이브 베이즈(Multinomial Naive Bayes) 분류기를, 스무딩(smoothing)은 라플라스(Laplace) 방식으로 α=1을 적용하였다. 또한 ROC는 파이썬 언어로 직접 구현한 것으로, 긍정예제만 학습에 사용하는 방식을 적용하였다. 다음으로 순환신경망(RNN)과 트랜스포머(Transformer)는 텍스트 분류에 많이 사용되는 딥러닝 기반의 분류기이다(김인후, 김성희, 2022; 한지영, 허고은, 2021; Cunha et al., 2021; Devlin et al., 2018). 순환신경망(이후 RNN으로 표기)은 케라스(keras) 모듈을 사용하여 구성하였는데, 단어 임베딩(embedding)은 64차원으로 구성하였고, 이를 LSTM 레이어를 적용하여 처리하도록 하였다. 또한, 트랜스포머(이후 TRA로 표기)는 케라스(keras) 모듈을 사용하여 구성하였고, 임베딩 사이즈는 32, 어텐션 헤드의 개수는 2개, 히든 레이어 사이즈는 32개로 설정하였다. 그리고 딥러닝 기반 2개 분류기의 최종 출력 값은 소프트맥스(softmax) 함수를 통해 1개의 범주만 할당하도록 하였다.

2.4 자질 순위화 기법

텍스트 분류를 위한 자질선정 방법으로 자질 순위화 기법은 크게 단일 순위화 기법과 조합 순위화 기법으로 구분할 수 있다. 먼저, 용어빈도에 기초한 단일 순위화 기법들은 〈표 2〉와 같다(김판준, 2008). 여기서 용어빈도는 한 문헌 내 특정 용어의 출현빈도를 말하며, 전역적인 사용을 위해 특정 용어가 출현한 전체 문헌에 대한 평균값을 사용하였다.

대부분의 자질 순위화 기법들은 문헌빈도에 기반하고 있으며(Forman, 2003), 특정 자질(t_i)의 출현과 범주(c_j)의 부여 정보에 따라 다음의

〈표 2〉 용어빈도에 기초한 단일 순위화 기법

번호	단일 기법(약어)	공식	비고
1	atf	$0.5+(0.5\times(\frac{tf}{\max tf}))$	augmented tf
2	ltf	$\log(1 + tf)$	log tf
3	itf	$1-(\frac{1}{(1+tf)})$	Inverse tf
4	otf	$\frac{tf}{(2+tf)}$	okapi tf

네 가지 문헌빈도를 정의할 수 있다(Rehman, Javed, & Babri, 2017). (a) true positive(tp)는 범주 c_j에 속하고 자질 t_i가 출현한 문헌 수, (b) false positive(fp)는 범주 c_j에 속하지 않고 자질 t_i가 출현한 문헌 수, (c) false negative(fn)은 범주 c_j에 속하고 자질 t_i가 출현하지 않은 문헌 수, (d) true negative(tn)은 범주 c_j에 속하지 않고 자질 t_i가 출현하지 않은 문헌 수. 여기서 전체 문헌 수(N)은 이들을 모두 합한 값(a+b+c+d)가 된다. 이러한 문헌빈도에 기초하여 본 연구에서 사용한 11개 단일 순위화 기법은 〈표 3〉과 같다.

다음으로, 조합 순위화 기법(combination ranking techniques)은 용어빈도와 문헌빈도를 함께 사용하는 것으로 양자를 곱하는 방식이다. 문헌빈도에 기초한 단일 순위화 기법들의 문제를 해소하기 위한 목적으로 용어빈도를 함께 사용하는 조합 순위화 기법은 대부분 이러한 방법을 사용하고 있다(Harish & Revanasiddappa, 2017; Wang et al., 2012; Wang et al., 2014; Wu & Xu, 2015; Yao et al., 2017). 이에 따라 앞서 제시한 '용어빈도에 기초한 단일 순위화 기법 × 문헌빈도에 기초한 단일 순위화 기법'의 형식으로 총 44개의 조합이 가능하다. 이 중에서 단일 순위화 기법에 대한 사전 실험 결과에서 가장 좋은 성능을 보인 것으로 〈표 4〉의 조합 순위화 기법을 사용하였다.

〈표 3〉 문헌빈도에 기초한 단일 순위화 기법

번호	단일 기법(약어)	공식	출처
1	acc	$a-b$	(Forman, 2003; Rehman et al., 2018)
2	acc2	$\lvert\frac{a}{(a+c)}-\frac{d}{(b+d)}\rvert$	(Forman, 2003; Rehman et al., 2018)
3	chi	$\frac{N(ad-bc)^2}{((a+c)(b+d)(a+b)(c+d))}$	(이재윤, 2005; 김판준, 2022; Chang et al, 2015; Forman, 2003)
4	df	$a+b$	(Forman, 2003; Rehman, Javed, & Babri, 2017; Yang & Pedersen, 1997)
5	gss	$\frac{(ad-bc)}{N^2}$	(김판준, 2022; 이재윤, 2005)
6	idf	$\log(1+(\frac{N}{a+b}))$	(이재윤, 2005; Yao et al., 2017)
7	jac	$\frac{a}{a+b+c}$	(김판준, 2008; 이재윤, 2005)
8	lor	$\log(\frac{ad}{bc})$	(김판준, 2022; 이재윤, 2005; Agnihotri, Verma, & Tripathi, 2017)
9	mi	$\log(\frac{Na}{(a+b)(a+c)})$	(김판준, 2008; Cai et al., 2018; Chang et al, 2015)
10	pea	$\frac{(ad-bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	(김판준, 2022; Cai et al., 2018; Venkatesh & Anuradha, 2019)
11	rf	$\log(1+(\frac{a+b}{b}))$	(Harish & Revanasiddappa, 2017; Lan et al., 2008)

〈표 4〉 용어빈도와 문헌빈도에 기초한 조합 순위화 기법

번호	조합 기법(약어)	공식
1	atf*chi	$\left(0.5+(0.5\times(\frac{tf}{\max tf}))\right) \times \left(\frac{N(ad-bc)^2}{((a+c)(b+d)(a+b)(c+d))}\right)$
2	atf*jac	$\left(0.5+(0.5\times(\frac{tf}{\max tf}))\right) \times \left(\frac{a}{a+b+c}\right)$

2.5 성능 척도 및 검증

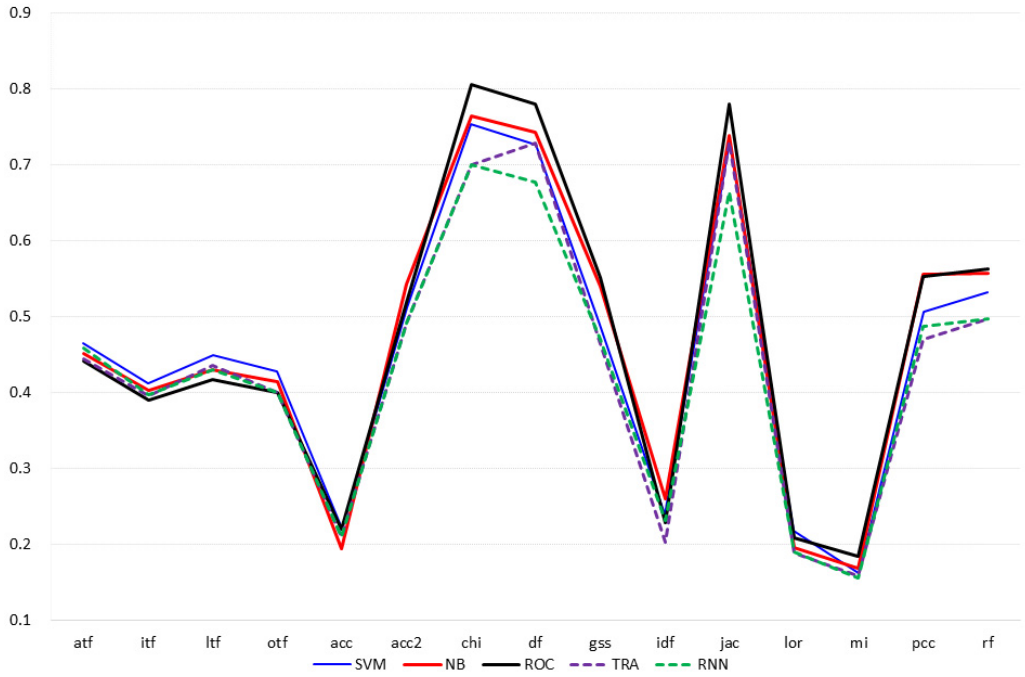
텍스트 분류를 위한 자질 순위화 기법의 성능 평가를 위한 척도로 매크로 F1(이하 mac_F1로 표기)과 마이크로 F1(이하 mic_F1로 표기)이 많이 사용되고 있다. mac_F1은 개별 범주의 F1값을 계산한 후에 모든 범주에 대한 평균을 구하는 것으로 소범주(rare classes)의 영향이 더 큰 반면, mic_F1은 모든 범주에 대하여 통합된 F1 값을 구하는 것으로 대범주(common classes)의 영향이 크게 작용하는 것으로 알려져 있다. 본 연구에서는 서로 다른 특성을 갖는 이러한 두 개의 척도를 사용한 결과를 종합적으로 검토하였다(김판준, 2022; Cunha et al., 2021; Rehman et al., 2018).

텍스트 분류를 위한 자질 순위화 기법들의 성능을 보다 신뢰성 있게 평가하기 위해 교차검증(cross validation)과 대응표본 t-검정(이후 t-test로 표기)을 수행하였다. 먼저, 교차검증은 5개로 분할된 문헌집합을 사용하는 5-폴더 교차검증(이후 5-fold cross validation으로 표기) 방법을 사용하였다(Agnihotri, Verma, & Tripathi, 2017; Rehman, Javed, & Babri, 2017). 다음으로 이러한 교차검증의 결과를 확인하기 위하여 t-test(95% 신뢰도)를 수행하였다(Cunha et al., 2021; Forman, 2003; Pinheiro, Cavalcanti, & Ren, 2015; Wang & Hong, 2019).

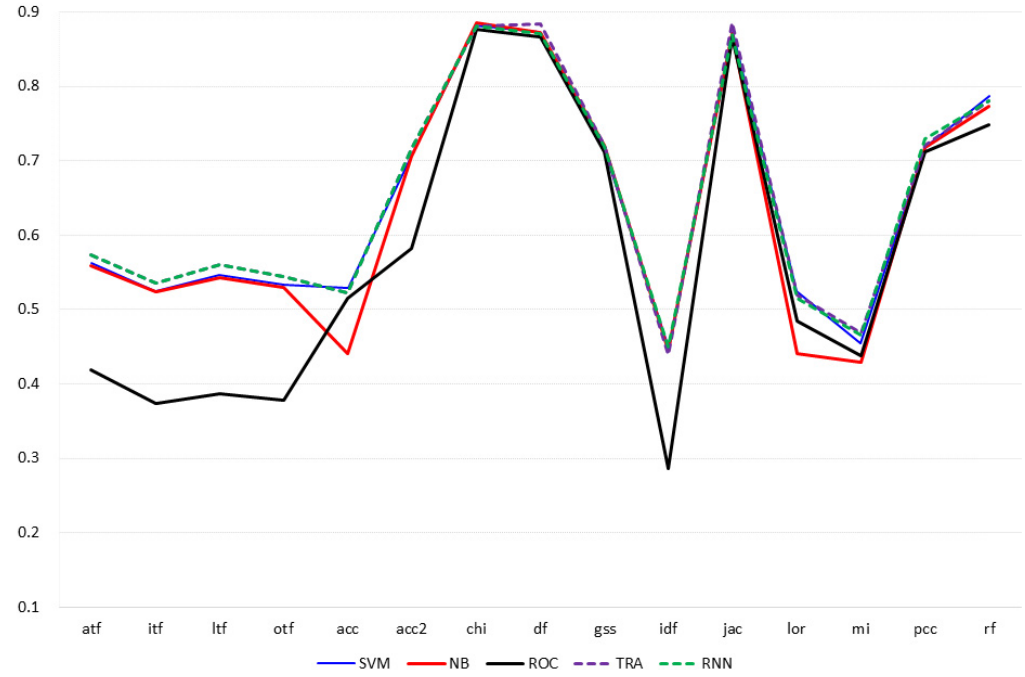
3. 실험 결과 및 분석

3.1 단일 순위화 기법

연구문제1에 대한 실험으로 2개의 문헌집합(Reuters-21578, 20NG)과 5개 분류기(NB, Roccihio, SVM, RNN, TRA)를 사용하는 환경에서, 용어빈도와 문헌빈도에 기초한 단일 순위화 기법 15개를 적용한 분류 성능을 살펴보았다. 먼저, Reuters-21578에 대하여 5개 분류기를 사용하면서 각 단일 순위화 기법을 통해 자질 수를 100개에서 2000개까지 변화시킨 평균 mac_F1 성능은 〈그림 1〉과 같다. 여기서 ROC 분류기와 문헌빈도에 기초한 단일 기법인 chi를 적용한 성능이 가장 좋았고(ROC+chi/0.8048), 그 다음이 NB(NB+chi/0.7642), SVM(SVM+chi/0.7534)의 순이었다. 문헌빈도에 기초한 단일 기법들이 대체로 용어빈도 기반의 단일 기법보다 높은 성능을 보였지만, 일부 문헌빈도(acc, idf, lor, mi)는 이보다 낮은 성능 수준을 보였다. 〈그림 2〉는 Reuters-21578에 대한 평균 mic_F1 성능으로, TRA와 NB 분류기에 각각 문헌빈도(jac, chi)를 적용한 경우가 거의 동등한 수준으로 가장 좋았고(TRA+jac/0.8851,



〈그림 1〉 단일 순위화 기법을 적용한 평균 mac_F1: Reuters-21578

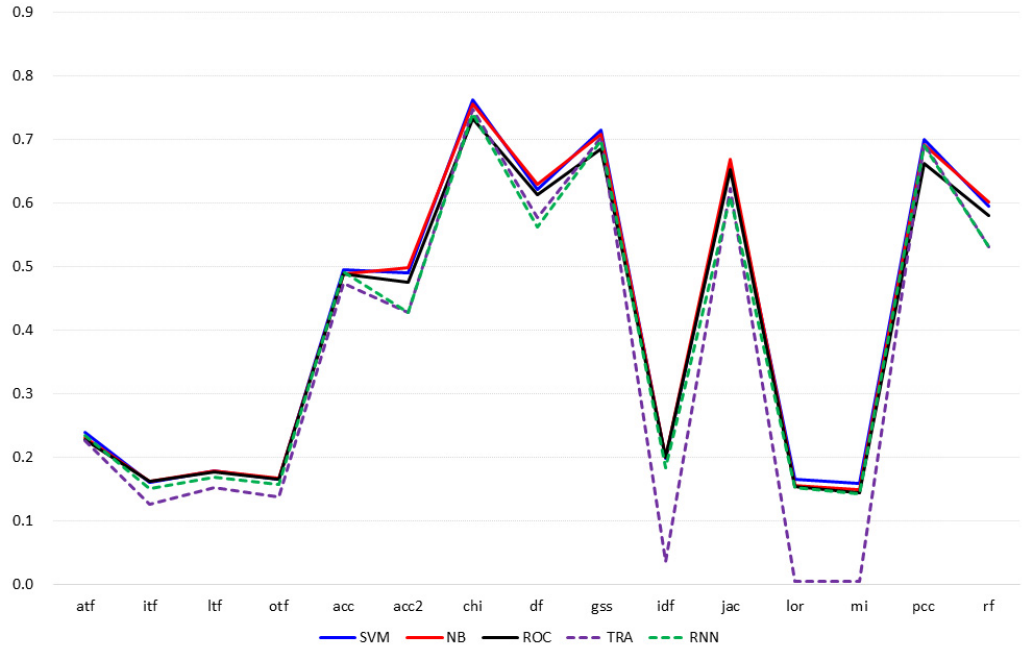


〈그림 2〉 단일 순위화 기법을 적용한 평균 mic_F1: Reuters-21578

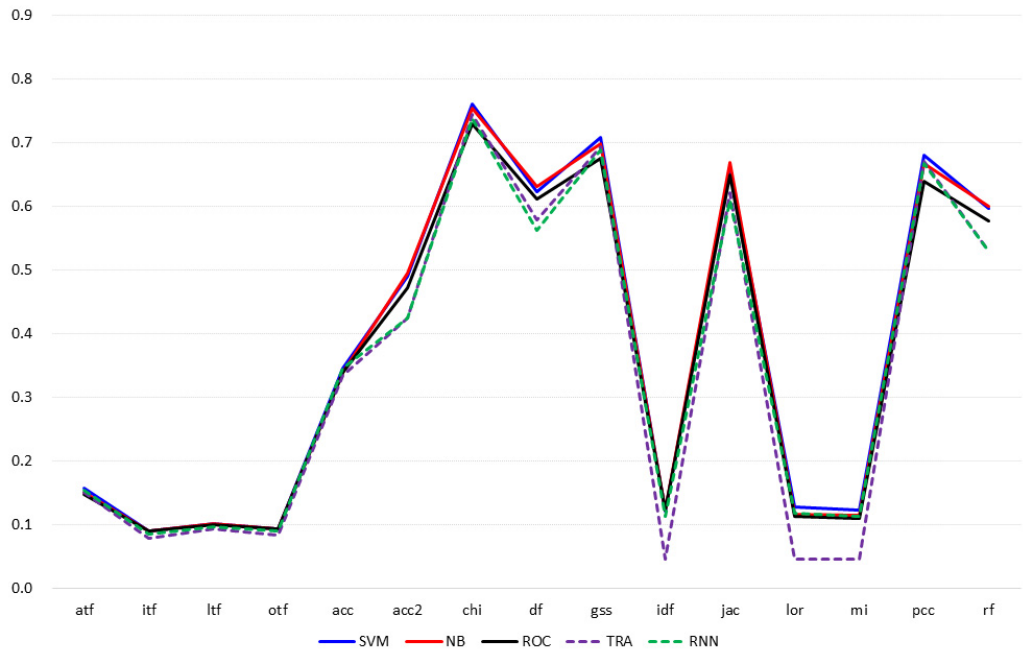
NB+chi/0.8849), 그 다음은 SVM과 RNN 분류기에 문헌빈도(chi)를 적용한 것이었다(SVM+chi/0.8817, RNN+chi/0.8806). 또한, 〈그림 1〉과 마찬가지로 평균 mic_F1 측면에서도 용어빈도보다는 문헌빈도에 기초한 단일 기법들의 성능이 전반적으로 더 좋은 것으로 나타났다. 그러나 일부 문헌빈도(idf, lor. mi)는 용어빈도에 기초한 단일 기법과 유사하거나 오히려 더 낮은 수준의 성능을 보였다.

다음으로, 20NG에 대하여 5개 분류기를 사용하면서 15개 단일 순위화 기법을 통해 자질 수를 100개에서 2000개까지 변화시킨 평균 mac_F1 성능은 〈그림 3〉과 같다. NB와 SVM 분류기에 각각 문헌빈도(chi)를 적용한 경우에 상호간에 거의 차이가 없이 가장 좋은 성능이었다. 또한, 대부분 문헌빈도 기반의 단일 기법들이 용어빈도보다 높은 성능을 보였지만, 일부 문헌빈도(idf, lor, mi)는 용어빈도와 유사하거나 더 낮은 성능을 보였다.

〈그림 4〉는 20NG에 대한 평균 mic_F1 성능으로 SVM 분류기에 문헌빈도(chi)를 적용한 성능이 가장 좋았고(SVM+chi/0.7619), 다음은 NB(NB+chi/0.7544), TRA(TRA+chi/0.7449) 등의 순이었다. 또한, 〈그림 3〉과 마찬가지로 대체로 용어빈도보다는 문헌빈도에 기초한 단일 기법들의 성능이 더 높은 수준인 것으로 나타났다. 그러나 여기서도 3개 문헌빈도(idf, lor. mi)는 용어빈도들과 유사하거나 더 낮은 수준의 성능을 보였다. 한편, 용어빈도에 기초한 단일 기법 중에서는 모든 경우에 atf가 가장 좋은 성능을 보였다. 결과적으로, 두 개의 문헌집합과 5개 분류기를 사용한 단일 자질 순



〈그림 3〉 단일 순위화 기법을 적용한 평균 mac_F1: 20NG



〈그림 4〉 단일 순위화 기법을 적용한 평균 mic_F1: 20NG

위화 기법의 평균 성능 측면에서는 대체로 용어빈도보다 문헌빈도에 기초한 기법들이 좋은 성능을 보이는 것으로 나타났다. 특히, 기존의 텍스트 분류 연구의 결과에서와 같이, 전반적으로 문헌빈도 기반의 단일 순위화 기법(chi)이 가장 우수한 성능을 보였다(Forman, 2003; Mesleh, 2011; Yang & Pedersen, 1997). 한편, 분류기 측면에서는 대체로 기계학습 기반의 분류기(SVM, NB, ROC)가 신경망에 기초한 분류기(TRA, RNN)보다 나은 성능을 보이는 것으로 나타났다. 〈그림 2〉에서 최고 성능(mic_F1)을 보인 신경망 기반의 분류기(TRA)는 NB와 거의 유사한 성능을 보였지만(TRA+jac/0.8851 ≒NB+chi/0.8849), 상대적으로 학습에 필요한 시간이 지나치게 많이 소요되는 문제가 있었다. 이러한 문제는 텍스트 분류의 효율성 측면에서 신경망 분류기와 비신경망 분류기를 비교한 최근의 연구에서도 지적된 바 있다(Cunha et al., 2021). 따라서 후속 실험에서는 효율성 측면에서 기계학습 기반의 분류기 3개(SVM, NB, ROC)를 사용하였다.

3.2 조합 순위화 기법

연구문제2에 대한 텍스트 자동분류 실험은 2개의 문헌집합(Reuters-21578, 20NG)과 3개 분류기(NB, Roccihio, SVM)를 사용하는 환경에서, 이전 실험에서 가장 좋은 성능을 보인 용어빈도(atf)와 문헌빈도(chi, jac)에 기초한 단일 순위화 기법과 이들을 함께 사용한 조합 순위화 기법(atf*chi, atf*jac)의 성능을 비교하였다. 즉, 단일 순위화 기법과 조합 순위화 기법으

로 선정한 20개 자질집합(100~2000)을 적용하여 산출한 평균 성능을 검토하였다. 〈표 5〉는 Reuters-21578에 대하여 자질선정 방법으로 각각 단일 순위화 기법과 조합 순위화 기법을 적용하여 산출한 3개 분류기의 성능이다. 먼저, mac_F1 측면에서는 ROC 분류기와 조합 순위화 기법(atf*chi)을 사용한 경우에 가장 좋은 성능을 보였다(0.8073). 또한, 3개 분류기 모두 문헌빈도 기반의 chi는 용어빈도(atf)와 함께 사용하는 경우에 성능 변화가 거의 없었지만, jac는 어느 정도 성능 향상이 있는 것으로 나타났다. 다음으로, mic_F1으로는 NB 분류기와 단일 순위화 기법(chi)를 사용한 경우에 가장 좋은 성능이었다(0.8849). 또한 3개 분류기 모두에서 문헌빈도 기반의 chi는 용어빈도(atf)를 함께 사용하는 조합 순위화 기법(atf*chi)과 거의 성능 차이가 없는 반면, jac는 조합 순위화 기법(atf*jac)을 사용하는 경우에는 어느 정도 성능 향상이 있는 것으로 나타났다.

〈표 6〉은 20NG에 대하여 자질선정 방법으로 각각 단일 순위화 기법과 조합 순위화 기법을 적용한 3개 분류기의 성능이다. 먼저, mac_F1 측면에서는 SVM 분류기와 단일 순위화 기법(chi)을 사용한 경우에 가장 좋은 성능을 보였다(0.7629). 또한, 3개 분류기 모두 문헌빈도 기반의 chi는 용어빈도와 함께 조합 순위화 기법(atf*chi)을 사용하는 경우와 비교하여 거의 성능 차이가 없지만, jac는 조합 순위화 기법(atf*jac)을 적용하면 어느 정도의 성능 향상이 있는 것으로 나타났다. 다음으로, mic_F1으로는 SVM 분류기와 단일 순위화 기법(chi)를 사용한 경우에 가장 좋은 성능이었다(0.7619). 또한 3개 분류기 모두에서 chi는 조합 순위화 기법(atf*chi)과 비교하여 거의 성능 차이가 없는 반면, jac는 조합 순위화 기법(atf*jac)을 적용하는 경우에 약간의 성능 향상이 있는 것으로 나타났다.

3.3 5-fold cross validation & t-test

본 연구의 실험 결과에 대한 신뢰성을 확인하기 위해 5-fold cross validation과 t-test

〈표 5〉 단일 vs. 조합 순위화 기법의 평균 성능: Reuters-21578, 자질 수(100~2000)

성능 척도	순위화 기법		분류기		
			SVM	NB	ROC
mac_F1	단일	atf	0.4644	0.4518	0.4408
		chi	**0.7534**	0.7642	0.8048
		jac	0.7318	0.7378	0.7793
	조합	**atf*chi**	0.7522	**0.7667**	0.8073
		atf*jac	0.7504	0.7588	0.8009
mic_F1	단일	atf	0.5626	0.5592	0.4185
		chi	0.8817	0.8849	0.8766
		jac	0.8731	0.8708	0.8669
	조합	atf*chi	**0.8823**	0.8846	**0.8776**
		atf*jac	0.8808	0.8795	0.8775

〈표 6〉 단일 vs. 조합 순위화 기법의 평균 성능: 20NG, 자질 수(100~2000)

성능 척도	순위화 기법		분류기		
			SVM	NB	ROC
mac_F1	단일	atf	0.2399	0.2309	0.2282
		chi	0.7629	**0.7560**	**0.7325**
		jac	0.6592	0.6684	0.6536
	조합	atf*chi	0.7597	0.7472	0.7286
		atf*jac	0.6983	0.7002	0.6888
mic_F1	단일	atf	0.1577	0.1512	0.1483
		chi	0.7619	**0.7544**	**0.7293**
		jac	0.6616	0.6693	0.6503
	조합	atf*chi	0.7584	0.742	0.7242
		atf*jac	0.7002	0.7005	0.6843

를 수행하였다. 먼저, 이전 실험에서 좋은 성능을 보인 문헌빈도 기반의 단일 순위화 기법 2개(chi, jac)와 조합 순위화 기법 2개(atf*chi, atf*jac)의 성능을 확인하기 위하여, 5-fold cross validation을 수행하였다. 다음으로 가장 좋은 성능을 보인 단일 순위화 기법과 조합 순위화 기법의 성능 차이에 대한 paired t-test를 수행하였다.

〈표 7〉은 Reuters-21578 문헌집합에 대하여 자질선정 방법으로 이전 실험에서 최고 성능을 보인 문헌빈도 기반의 단일 순위화 기법(chi, jac)과 조합 순위화 기법(atf*chi, atf*jac)에 대하여 5-fold cross validation을 수행한 결과이다. 여기서 mac_F1 측면에서 가장 좋은 성능을 보인 것은 ROC 분류기와 조합 순위화 기법(atf*chi/0.7947)이었고, mic_F1으로는 SVM 분류기와 조합 순위화 기법(atf*jac/0.8921)이었다. 또한 〈표 8〉은 20NG 문헌집합에 대하여 자질선정 방법으로 문헌빈도 기반의 단일 순위화 기법(chi, jac)과 조합 순위화 기법(atf*chi, atf*jac)을 적

〈표 7〉 단일 vs. 조합 순위화 기법의 평균 성능: Reuters-21578, 5-fold cross-validation

성능 척도	순위화 기법		분류기		
			SVM	NB	ROC
mac_F1	단일	**chi**	0.7777	0.7759	**0.7945**
		jac	0.7755	0.7747	0.7932
	조합	**atf*chi**	**0.7778**	**0.7768**	0.7947
		atf*jac	0.7771	0.7747	0.7936
mic_F1	단일	chi	0.8914	0.8705	0.8620
		jac	**0.8911**	0.8697	0.8620
	조합	atf*chi	0.8912	**0.8707**	0.8621
		atf*jac	0.8921	0.8703	**0.8625**

〈표 8〉 단일 vs. 조합 순위화 기법의 평균 성능: 20NG, 5-fold cross-validation

성능 척도	순위화 기법		분류기		
			SVM	NB	ROC
mac_F1	단일	**chi**	0.8003	**0.7973**	**0.7674**
		jac	0.7473	0.7444	0.7286
	조합	**atf*chi**	**0.7988**	0.7865	0.7634
		atf*jac	0.7624	0.7564	0.7417
mic_F1	단일	**chi**	0.8003	**0.7974**	**0.7655**
		jac	0.7473	0.7446	0.7260
	조합	**atf*chi**	**0.7989**	0.7846	0.7620
		atf*jac	0.7625	0.7566	0.7391

용하여 5-fold cross validation을 수행한 결과이다. 여기서 mac_F1과 mic_F1 모두 SVM 분류기와 문헌빈도 기반의 단일 순위화 기법(chi)을 적용한 경우에 가장 좋은 성능을 보였다(0.8003).

Reuters-21578과 20NG 문헌집합에 대하여 3개 분류기를 사용한 5-fold cross-validation에서 가장 좋은 성능을 보인 단일 순위화 기법과 조합 순위화 기법의 성능 차이에 대한 t-test를 수행한 결과는 다음과 같다. 첫째, Reuters-21578 문헌집합을 대상으로 mac_F1 측면에서 최고 성능을 보인 ROC 분류기에 조합 순위화 기법(atf*chi)과 단일 순위화 기법(chi)을 적용한 경우의 성능에 대한 t-test 결과는 95% 유의 수준($p=0.77$)에서 차이가 없었다. 또한, mic_F1으로는 SVM 분류기에 조합 순위화 기법(atf*jac)과 단일 순위화 기법(jac)을 적용한 t-test 결과에서도 95% 유의 수준(0.086)에서 차이가 없었다. 둘째, 20NG 문헌집합을 대상으로 mac_F1과 mic_F1 모두에서 최고 성능을 보인 SVM 분류기에 조합 순위화 기법(atf*chi)과 단일 순위화 기법(chi)을 적용한 경우의 성능에 대한 t-test 결과는 각각 95% 유의 수준($p=0.62$, 0.65)에서 차이가 없었다. 결과적으로 텍스트 분류를 위한 자질선정 방법으로 단일 순위화 기법과 조합 순위화 기법 간에 유의한 성능 차이는 없는 것으로 나타났다.

4. 결 론

본 연구는 텍스트 분류를 위한 효율적인 자질선정 방법으로 자질 순위화 기법의 성능을 구체적으로 검토하였다. 첫째, 텍스트 분류에서 가장 많이 사용되어 온 2개의 문헌집합(Reuters-21578, 20NG)과 5개 분류기(SVM, NB, Roccihio, TRA, RNN)를 사용하는 환경에서, 자질선정 방법으로 각각 용어빈도와 문헌빈도에 기초한 단일 순위화 기법 15개를 적용한 분류 성능을 살펴보았다. 둘째, 2개의 문헌집합(Reuters-21578, 20NG)과 3개 분류기(NB, Roccihio, SVM)를 사용하는 환경에서, 이전 실험에서 가장 좋은 성능을 보인 단일 순위화 기법(용어빈도, 문헌빈도)과 양자를 함께 사용하는 조합 순위화 기법을 적용한 분류 성능을 비교하였다. 셋째, 서로 다른 특성을 지닌 mac_F1과 mic_F1을 텍

스트 분류의 성능 평가를 위한 척도로 사용하였고, 실험 결과의 신뢰성을 확인하기 위하여 5-fold cross validation과 t-test를 수행하였다.

텍스트 분류를 위한 자질선정 방법으로 단일 순위화 기법과 조합 순위화 기법의 성능을 검토한 실험 결과는 다음과 같다. 첫째, 두 개의 문헌집합(Reuters-21578, 20NG)과 5개 분류기(SVM, NB, Roccihio, TRA, RNN)를 사용한 단일 자질 순위화 기법들의 평균 성능(mac_F1, mic_F1) 측면에서 대체로 용어빈도보다 문헌빈도에 기초한 기법들이 더 좋은 성능을 보이는 것으로 나타났다. 특히, 전반적으로 문헌빈도 기반의 단일 순위화 기법(chi, jac)이 가장 우수한 성능을 보였다(연구문제1). 둘째, 두 개의 문헌집합(Reuters-21578, 20NG)과 3개 분류기(SVM, NB, Roccihio)를 사용한 평균 성능 측면에서도 문헌빈도에 기초한 단일 순위화 기법(chi)이 대체로 가장 좋은 성능을 보였다. 예외적으로 Reuters-21578과 ROC를 사용한 mac_F1 성능에서는 조합 순위화 기법(atf*chi/0.8073)이 단일 순위화 기법(chi/0.8048)보다 나은 결과를 보였지만 그 차이는 크지 않았다(연구문제2). 셋째, 이전 실험에서 좋은 성능을 보인 단일 순위화 기법(chi, jac)과 조합 순위화 기법(atf*chi, atf*jac)의 결과에 대한 신뢰성을 확보하기 위하여, 5-fold cross validation과 t-test를 수행한 결과에서도 양자 간에 유의한 성능 차이는 없는 것으로 나타났다(연구문제2). 따라서 학습문헌이 비교적 충분한 환경에서 텍스트 분류를 수행하기 위한 자질 순위화 기법으로는 문헌빈도에 기초한 단일 순위화 기법(chi)을 사용하는 것이 가장 효율적이라 할 수 있다.

본 연구의 의의는 기존의 텍스트 분류를 위한 자질 순위화 기법에서 주로 사용되어 온 문헌빈도와 함께 자질의 또 다른 주요 정보로서 용어빈도의 적용에 대한 가능성을 모색한 것이라 할 수 있다. 그러나 이는 특정 문헌집합(Reuters-21578, 20NG)과 비교적 학습문헌이 충분히 존재하는 환경에서 실험한 결과이므로, 다양한 특성을 가진 실제 환경의 텍스트 문헌집합으로 일반화하기에는 어려움이 있다. 따라서 보다 다양한 특성을 갖는 문헌집합과 범주집합을 대상으로 하는 후속 연구가 필요할 것이다.

참 고 문 헌

김인후, 김성희 (2022). 딥러닝 기반의 BERT 모델을 활용한 학술 문헌 자동분류. 정보관리학회지, 39(3), 293-310. http://dx.doi.org/10.3743/KOSIM.2022.39.3.293

김판준 (2008). 용어 가중치부여 기법을 이용한 로치오 분류기의 성능 향상에 관한 연구. 정보관리학회지, 25(1), 211-233. http://dx.doi.org/10.3743/KOSIM.2008.25.1.211

김판준 (2016). 기계학습에 기초한 자동분류의 성능 요소에 관한 연구. 정보관리학회지, 33(2), 33-59. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

김판준 (2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

김판준 (2022). 자질선정을 통한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 39(1), 69-90. http://dx.doi.org/10.3743/KOSIM.2022.39.1.069

육지희, 송민 (2018). 토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구. 정보관리학회지, 35(2), 63-88. http://dx.doi.org/10.3743/KOSIM.2018.35.2.063

이재윤 (2005). 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구. 한국문헌정보학회지, 39(2), 123-146. http://dx.doi.org/10.4275/kslis.2005.39.2.123

한지영, 허고은 (2021). 토픽 모델링 기반 비대면 강의평 분석 및 딥러닝 분류 모델 개발. 한국문헌정보학회지, 55(4), 267-291. http://dx.doi.org/10.4275/KSLIS.2021.55.4.267

Abiodun, E. O., Alabdulatif, A., Abiodun, O. I., Alawida, M., Alabdulatif, A., & Alkhawaldeh, R. S. (2021). A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Computing & Applications, 33(4), 1-28. https://doi.org/10.1007/s00521-021-06406-8

Aggarwal, C. C. & Zhai, C. (2012). A Survey of Text Classification Algorithms. In: Aggarwal, C., Zhai, C. (eds) Mining Text Data. https://doi.org/10.1007/978-1-4614-3223-4_6

Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. Expert Systems with Applications, 81, 268-281. https://doi.org/10.1016/j.eswa.2017.03.057

Ávila-Argüelles, R., Calvo, H., Gelbukh, A., & Godoy-Calderón, S. (2010). Assigning Library of Congress Classification codes to books based only on their titles. Informatica, 34(1), 77-84.

Azam, N. & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications, 39(5), 4760-4768. https://doi.org/10.1016/j.eswa.2011.09.160

Baccianella, S., Esuli, A., & Sebastiani, F. (2013). Using micro-documents for feature selection: The case of ordinal text classification. Expert Systems with Applications, 40(11), 4687-4696. https://doi.org/10.1016/j.eswa.2013.02.010

Bolón-Canedo, V. & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. Information Fusion, 52, 1-12. https://doi.org/10.1016/j.inffus.2018.11.008

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79. https://doi.org/10.1016/j.neucom.2017.11.077

Cai, Z. & Zhu, W. (2018). Multi-label feature selection via feature manifold learning and sparsity

regularization. International journal of machine learning and cybernetics, 9(8), 1321-1334. https://doi.org/10.1007/s13042-017-0647-y

Chang, F., Guo, J., Xu, W., & Yao, K. (2015). A Feature Selection Method to Handle Imbalanced Data in Text Classification. Journal of Digital Information Management, 13, 169-175.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36(3), 5432-5435. https://doi.org/10.1016/j.eswa.2008.06.054

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., & Gonçalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. Information Processing & Management, 58(3), 102481. https://doi.org/10.1016/j.ipm.2020.102481

Dash, M. & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1, 131-156. https://doi.org/10.1016/S1088-467X(97)00008-5

Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. Multimedia Tools and Applications, 78, 3797-3816. https://doi.org/10.1007/s11042-018-6083-5

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289-1305.

Günal, S. (2012). Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering and Computer Science, 20(Sup.2), 1296-1311.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1), 389-422. https://doi.org/10.1023/A:1012487302797

Han, E. H. & Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In European conference on principles of data mining and knowledge discovery, 421-431. https://doi.org/10.1007/3-540-45372-5_46

Harish, B. & Revanasiddappa, M. (2017). A comprehensive survey on various feature selection methods to categorize text documents. International Journal of Computer Applications, 164,

1-7. http://doi.org/10.5120/ijca2017913711

Iqbal, M., Abid, M. M., Khalid, M. N., & Manzoor, A. (2020). Review of feature selection methods for text classification. International Journal of Advanced Computer Research, 10(49), 138-152. http://dx.doi.org/10.19101/IJACR.2020.1048037

Javed, K., Babri, H. A., & Saeed, M. (2010). Feature selection based on class-dependent densities for high-dimensional binary data. IEEE Transactions on Knowledge and Data Engineering, 24(3), 465-477. http://dx.doi.org/10.1109/TKDE.2010.263

Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Carnegie-Mellon University Dept of Computer Science. Available: https://apps.dtic.mil/sti/citations/ADA307731

Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory and algorithms. Massachusetts: Kluwer Academic Publishers.

Kohavi, R. & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kumar, V. & Minz, S. (2014). Feature selection: a literature review. Smart Computing Review, 4(3), 211-229. htts://doi.org/10.6029/smartcr.2014.03.007

Lan, M., Tan, C. L., Su, J., & Lu, Y. (2008). Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4), 721-735. https://doi.org/10.1109/TPAMI.2008.110

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., De Schaetzen, V., Duque, R., Bersini, H., & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 1106-1119. https://doi.org/10.1109/TCBB.2012.33

Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. Knowledge and Information Systems, 53(3), 551-577. https://doi.org/10.1007/s10115-017-1059-8

Liu, H. & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491-502. https://doi.org/10.1109/TKDE.2005.66

Mesleh, A. M. (2011). Feature sub-set selection metrics for arabic text classification. Pattern Recognition Letters, 32(14), 1922-1929. https://doi.org/10.1016/j.patrec.2011.07.010

Parlak, B. & Uysal, A. K. (2021). A novel filter feature selection method for text classification: Extensive Feature Selector. Journal of Information Science, 49(1), 59-78.

https://doi.org/10.1177/0165551521991037

Pinheiro, R. H., Cavalcanti, G. D., & Ren, T. I. (2015). Data-driven global-ranking local feature selection methods for text categorization. Expert Systems with Applications, 42(4), 1941-1949. https://doi.org/10.1016/j.eswa.2014.10.011

Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review, 54, 6149-6200. https://doi.org/10.1007/s10462-021-09970-6

Rehman, A., Javed, K., & Babri, H. A. (2017). Feature selection based on a normalized difference measure for text classification. Information Processing & Management, 53(2), 473-489. https://doi.org/10.1016/j.ipm.2016.12.004.

Rehman, A., Javed, K., Babri, H. A., & Asim, N. (2018). Selection of the most relevant terms based on a max-min ratio metric for text classification. Expert Systems with Applications, 114, 78-96. https://doi.org/10.1016/j.eswa.2018.07.028

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47. https://doi.org/10.1145/505282.505283

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. Expert Systems with Applications, 33(1), 1-5. https://doi.org/10.1016/j.eswa.2006.04.001

Su, J., Shirab, J. S., & Matwin, S. (2011). Large scale text classification using semi-supervised multinomial naive bayes. In Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11), 97-104. Available: http://www.icml-2011.org/papers/93_icmlpaper.pdf

Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. In: Famili, A. F., Kok, J.N ., Peña, J. M., Siebes, A., Feelders, A. (eds) Advances in intelligent data analysis VI. IDA 2005. Lecture Notes in Computer Science, 3646. https://doi.org/10.1007/11552253_40

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. Expert Systems with Applications, 43(1), 82-92. https://doi.org/10.1016/j.eswa.2015.08.050

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2011). A comparative evaluation of feature ranking methods for high dimensional bioinformatics data. In 2011 IEEE International Conference on Information Reuse & Integration, 2011, 315-320. https://doi.org/10.1109/IRI.2011.6009566

Venkatesh, B. & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics

and Information Technologies, 19(1), 3-26. https://doi.org/10.2478//cait-2019-0001

Wang, D, Zhang, H., Liu, R., & Lv, W. (2012). Feature selection based on term frequency and T-test for text categorization. IProceedings of the 21st ACM International Conference on Information and Knowledge Management, 1482-1486. https://doi.org/10.1145/2396761.2398457

Wang, D., Zhang, H., Liu, R., Liu, X., & Wang, J. (2016). Unsupervised feature selection through gram-Schmidt orthogonalization-A word co-occurrence perspective. Neurocomputing, 173(P3), 845-854. https://doi.org/10.1016/j.neucom.2015.08.038

Wang, D., Zhang, H., Liu, R., Lv, W., & Wang, D. (2014). t-test feature selection approach based on term frequency for text categorization. Pattern Recognition Letters, 45, 1-10. https://doi.org/10.1016/j.patrec.2014.02.013

Wang, H. & Hong, M. (2019). Supervised Hebb rule based feature selection for text classification. Information Processing & Management, 56(1), 167-191. https://doi.org/10.1016/j.ipm.2018.09.004

Wu, G. & Xu, J. (2015). Optimized approach of feature selection based on information gain. In 2015 International Conference on Computer Science and Mechanical Automation, 157-161. https://doi.org/10.1109/CSMA.2015.38

Wu, Y. & Zhang, A. (2004). Feature selection for classifying high-dimensional numerical data. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, CVPR 2004, 2, 251-258. http://doi.org/10.1109/CVPR.2004.1315171

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning, 412-420.

Yao, H., Liu, C., Zhang, P., & Wang, L. (2017). A feature selection method based on synonym merging in text classification system. EURASIP Journal on Wireless Communications and Networking, 2017(1), 1-8. https://doi.org/10.1186/s13638-017-0950-z

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Han, Ji Yeong & Heo, Go Eun (2021). Analyzing students' non-face-to-face course evaluation by topic modeling and developing deep learning-based classification model. Journal of the Korean Society for Library and Information Science, 55(4), 267-291.

http://dx.doi.org/10.4275/KSLIS.2021.55.4.267

Kim, In Hu & Kim, Seong hee (2022). Automatic classification of academic articles using BERT model based on deep learning. Journal of the Korean Society for Information Management, 39(3), 293-310. http://dx.doi.org/10.3743/KOSIM.2022.39.3.293

Kim, Pan Jun (2008). A study on the performance improvement of rocchio classifier with term weighting methods. Journal of the Korean Society for Information Management, 25(1), 211-233. http://dx.doi.org/10.3743/KOSIM.2008.25.1.211

Kim, Pan Jun (2016). An analytical study on performance factors of automatic classification based on machine learning. Journal of the Korean Society for information Management, 33(2), 33-59. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

Kim, Pan Jun (2018). An analytical study on automatic classification of domestic journal articles based on machine learning. Journal of the Korean Society for Information Management, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

Kim, Pan Jun (2022). An experimental study on the automatic classification of korean journal articles through feature selection. Journal of the Korean Society for Information Management, 39(1), 69-90. http://dx.doi.org/10.3743/KOSIM.2022.39.1.069

Lee, Jae-Yun (2005). An empirical study on improving the performance of text categorization considering the relationships between feature selection criteria and weighting methods. Journal of the Korean Society for Library and Information Science, 39(2), 123-146. http://dx.doi.org/10.4275/kslis.2005.39.2.123

Yuk, JeeHee & Song, Min (2018). A study of research on methods of automated biomedical document classification using topic modeling and deep learning. Journal of the Korean Society for Information Management, 35(2), 63-88. http://dx.doi.org/10.3743/KOSIM.2018.35.2.063