韓國IT서비스學會誌
第18卷 第1號
2019年 3月, pp.187-200

Journal of Information Technology Services
https://doi.org/10.9716/KITS.2019.18.1.187

부도예측 모형에서 뉴스 분류를 통한 효과적인 감성분석에 관한 연구

김찬송* · 신민수**

A Study on Effective Sentiment Analysis through News Classification in Bankruptcy Prediction Model

Chansong Kim* · Minsoo Shin**

■ Abstract ■

Bankruptcy prediction model is an issue that has consistently interested in various fields. Recently, as technology for dealing with unstructured data has been developed, researches applied to business model prediction through text mining have been activated, and studies using this method are also increasing in bankruptcy prediction. Especially, it is actively trying to improve bankruptcy prediction by analyzing news data dealing with the external environment of the corporation. However, there has been a lack of study on which news is effective in bankruptcy prediction in real-time mass-produced news.

The purpose of this study was to evaluate the high impact news on bankruptcy prediction. Therefore, we classify news according to type, collection period, and analyzed the impact on bankruptcy prediction based on sentiment analysis. As a result, artificial neural network was most effective among the algorithms used, and commentary news type was most effective in bankruptcy prediction. Column and straight type news were also significant, but photo type news was not significant. In the news by collection period, news for 4 months before the bankruptcy was most effective in bankruptcy prediction. In this study, we propose a news classification methods for sentiment analysis that is effective for bankruptcy prediction model.

Keyword : Sentiment Analysis, Text Mining, Bankruptcy Prediction

Submitted : January 31, 2018 1st Revision : November 2, 2018 Accepted : January 10, 2019

* 한양대학교 일반대학원 비즈니스 인포매틱스학과
** 한양대학교 경영학부 교수, 교신저자

1. 서 론

부도예측 연구는 오랫동안 다양한 분야에서 중요하게 다루어졌다. 기업의 부도는 개인투자자가 안전한 투자대상인지 선택하고, 금융 기관의 경우 기업의 대출 여부를 결정하고, 국가에서는 기업 부도가 경제 위기를 초래할 수 있는 위험을 인지하는 등 다양한 이유로 부도예측에 많은 관심을 가졌다. 이에 따라 연구자들이 부도 위험을 측정하고 평가하는 다양한 방법을 연구하였고, 최근까지도 지속적으로 이루어지고 있다.

부도예측의 초기 연구는 Beaver(1966)의 통계방법론을 사용한 이후로 과학적이고 통계적인 방법론들을 적용하면서 정교한 부도예측을 위해 노력하였다. 1968년 Altman의 연구에서 다변량 판별분석(Multi-variate Discriminant Analysis, MDA)을 사용해 개별적으로 관찰하던 재무비율 변수를 과학적으로 종합하고 명확하게 하였다. 이후 Ohlson(1980)의 로짓모형과 Zmijewski(1984)의 프로빗분석과 같은 이산확률모형이 발전하였으며, 1980년대 후반부터 귀납적 학습방법과 기계학습 방법인 인공신경망, SVM 등의 다양한 기법이 개발되면서 부도예측에 이를 적용하는 연구들이 진행되었다. 특히 기계학습 방법은 정규분포를 가정하지 않고, 비선형성을 반영한다는 점에서 부도예측 분야에 적합한 기법으로 인식되고 있고 현재까지도 계속 사용되고 있다.

선행연구들은 부도예측 정확도를 높이기 위해 재무변수를 차별화하거나 여러 가지 통계적 방법론은 적용하여 부도의 예측 정확도를 향상시키고자 하였다. 재무정보는 기업성과의 현황과 상태 등 전반적인 부분을 반영하기 때문에 기업 부도의 가능성을 측정할 수 있는 객관적인 지표라고 할 수 있다. 하지만 재무정보는 기업의 과거 실적을 바탕으로 결산 이후 정기적으로 작성하기 때문에, 재무비율만 가지고 급변하는 경영 환경 속에서 부도를 정확하게 예측하기에는 어려움이 존재한다. 이러한 재무정보를 이용한 연구는 대개 부도 1년 전 데이터를 사용하기 때문에 정적 모델에 해당한다고 말한 바 있다(Altman et al., 2010). 따라서 재무정보를 보완하기 위해 기업의 다양한 이슈를 부도예측 모형에 결합하기 위한 노력이 생기게 되었다.

텍스트 마이닝의 발달은 기업의 실시간으로 변화하는 외부 환경을 설명할 수 있는 새로운 대안으로 떠오르고 있다. 텍스트 마이닝은 자연어 처리 기술을 통해 반정형 또는 비정형 텍스트 데이터를 정형화하여 특징을 추출하고, 추출된 특징으로부터 의미 있는 정보를 발견하는 기술이다. 뉴스, 소셜 네트워크 서비스(SNS) 및 블로그 게시물, 트위터 등에서 생성되는 다양한 형식과 방대한 양의 구조화 되지 않은 데이터를 분석 할 수 있게 되면서 이를 사용해 비즈니스 예측 모델에 결합하는 연구들이 시행되었다. 텍스트 형태의 데이터는 형식적이지 않은 다양한 정보들을 내포하기 때문에, 기업의 외부 환경 또는 고객의 감정 등을 분석하여 이를 마케팅, 금융 분야 등 다양한 분야에 적용하는 연구들이 진행되고 있다.

텍스트 마이닝 기술을 부도예측에 활용한 방법 중 많이 사용되는 방법은 감성분석(Sentiment Analysis)이 있다. 감성분석이란 구조화 되지 않은 텍스트 데이터의 표출된 의견을 추출, 분류, 이해, 자산화하는 과정을 의미한다(Dave et al., 2003). 감성분석을 통해 언어에 나타난 긍정과 부정의 요소 및 정도성을 판별하여 정량화할 수 있다. 따라서 기업의 관한 소식을 대중에게 전달하는 뉴스 텍스트를 감성분석을 실시해 정서적 태도를 부도예측에 적용할 수 있다.

감성분석을 실시할 때 필요한 감성사전은 문서에 감성어 또는 감성어구가 갖는 감성 값을 의미한다. 감성사전은 구축 방법론에 따라 크게 사전기반 접근법(Dictionary-based method)과 말뭉치기반 접근법(Corpus-based method)으로 분류된다(Aggarwal et al., 2012; 김승우, 김남규, 2014; 조남옥, 2016).

사전기반 접근법으로 감성사전을 구축하는 방법은 먼저 확실한 감성을 갖는 시드 어휘에 감성 값을 부여한 다음, 시드 어휘를 바탕으로 다른 어휘의

감성 값을 추가하며 확장시키는 방법이다. 사전기반 접근법으로 구축되어 각종 연구에 널리 사용되고 있는 감성사전은 WordNet(Miller et al., 1990)과 SentiWordNet(Esuli and Sebastiani, 2005)이 있다. 우리나라에서는 지속적인 감성사전에 관한 연구가 필요한 실정이며, 현재 개발된 감성사전으로는 오픈한글(안정국, 김희웅, 2014)과 KOSAC(김문형 외 3인, 2013)이 있다. 감성사전은 시드 어휘와 다른 어휘 간의 거리를 계산하여 감성사전을 구축하는 데, 감성사전 안에 있는 감성 값의 정확도에 따라 연구결과에 상당한 영향을 미치기 때문에 감성분석을 위해 명확한 감성사전이 필요하다.

말뭉치기반 접근법은 실제 수집된 문장들을 구문 분석하여 감성사전을 구축하는 방법이다. 말뭉치기반 접근방법은 수집된 문장을 이용하여 표본을 추출하고 감성사전을 구축했기 때문에, 단어들의 감성 값이 일반적으로 사용되는 감성 값이라기보다 특정한 상황과 목적에 특화된 목적지향 감성사전이라고 할 수 있다. 말뭉치기반 접근법으로 구축된 감성사전은 다른 분야에 적용하기에는 어려우나, 특정 주제를 분석하고자 할 때 주제에 특화된 감성 값을 가진 감성사전을 구축할 수 있어 많이 사용되고 있다. 말뭉치기반 접근법으로 감성분석을 실시한 선행연구에는 유은지 외 3인(2013)이 뉴스를 이용하여 주가지수의 방향을 예측하기 위해 도메인별 감성사전을 실시한 사례가 있고, Salah et al.(2013)이 정치 토론 데이터를 사용하여 범용 감성 사전과 전문 감성사전을 비교하여 전문 감성사전 성능이 우수함을 입증한 연구 등이 있다.

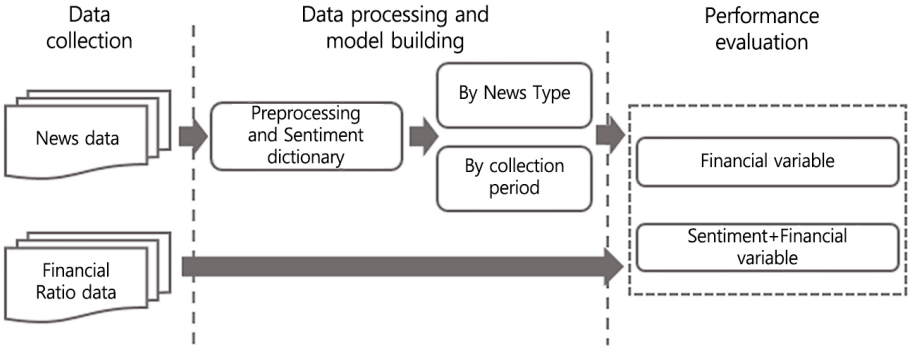
부도예측 연구에서도 뉴스를 텍스트 마이닝 실시하여 부도예측에 적용한 선행 연구들이 진행되었다. Lu et al.(2013)은 중국 금융 뉴스를 이용하여 기업의 조기경보 모델을 제시하였고, 최정원 외 3인(2015)은 뉴스를 텍스트 마이닝 하여 부도와 가장 연관된 단어를 추출하여 부도예측에 사용하였다. 조남옥(2016)은 건설기업 경기실사지수(CBSI)를 기준으로 감성사전을 구축하고 뉴스의 감성분석을 진행하였다. 뉴스에서 추출된 감성점수와 재무비율을 사용해 전통적 부도예측 모형보다 성과가 향상됨을 보였다. Gupta et al.(2016)은 미국 은행의 2000년부터 2014년 데이터를 가지고 금융 관련 감성사전을 활용하여 부정적인 감정과 긍정적인 감정으로 텍스트 감정을 추출하였다. 재무데이터와 감성분석 결과를 사용하여 건전성 평가 및 부도예측을 진행한 결과 긍정적인 감정이 부정적인 감정보다 강한 예측력을 가지고 있음을 보였다.

선행 연구들을 통해 부도 시점 이전 뉴스의 감성분석이 부도예측 연구에서 예측 성능이 향상되는 것은 알 수 있었지만, 어떠한 뉴스를 선별하는 것이 부도예측 연구에 효과적인지 찾는 연구는 부족하였다. 따라서 본 연구에서는 뉴스의 분류에 따라 부도예측에 미치는 영향이 다르다면, 실시간 대량으로 생성되는 뉴스 중 부도예측에 유용한 뉴스를 선택하여 사용하는 것이 효과적인지에 대한 연구를 진행하였다. 이를 위해 일괄적으로 뉴스를 수집하던 방식을 지양하고, 부도예측에 미치는 영향이 높은 뉴스를 연구하기 위해 유형에 따른 분류와 수집 기간에 따른 분류로 뉴스를 분류하였고, 이후 감성분석 시행하여 부도예측에 미치는 영향을 분석하였다.

2. 연구 방법

2.1 연구모형

본 연구에서는 기업 관련 뉴스를 전체 감성분석하여 부도예측에 적용하는 방식에서 부도예측에 미치는 영향이 높은 뉴스를 선정해 부도예측에 적용하고, 이에 따른 효과를 연구하고자 하였다. 이를 위해 유형에 따른 뉴스 분류와 수집 기간에 따른 뉴스 분류를 통해 감성분석 기반으로 부도에 미치는 영향을 분석하고자 하였다. 이러한 연구목적을 달성하기 위해 <Figure 1>과 같은 연구모델을 설정하였다.



〈Figure 1〉 Research Model for Sentiment Analysis that is Effective in Bankruptcy Prediction

첫 번째 단계는 부도 시점 이전의 뉴스데이터와 재무비율 데이터를 수집하는 데이터 수집 단계이다. 데이터 수집 이후엔 데이터를 전처리 후 모델을 구축한다. 데이터 처리 및 모델 구축 단계에서는 수집된 뉴스로 감성사전을 구축하고 뉴스를 유형 별, 수집 기간 별로 분류한다. 구축한 감성사전을 바탕으로 뉴스데이터에 감성점수를 적용하여 기업의 감성 점수를 추출한다. 마지막 단계는 기업의 감성 점수로 나타난 변수와 재무 변수, 그리고 이를 모두 사용하여 부도예측성을 평가하는 성능평가 단계이다. 성능을 측정하기 위해 부도예측에 효과적이고 많이 사용되고 있는 로지스틱 회귀분석, SVM, 인공신경망 알고리즘을 사용하였다. 각 알고리즘 별 성능평가를 실시한 후 최적의 알고리즘을 선택하였다.

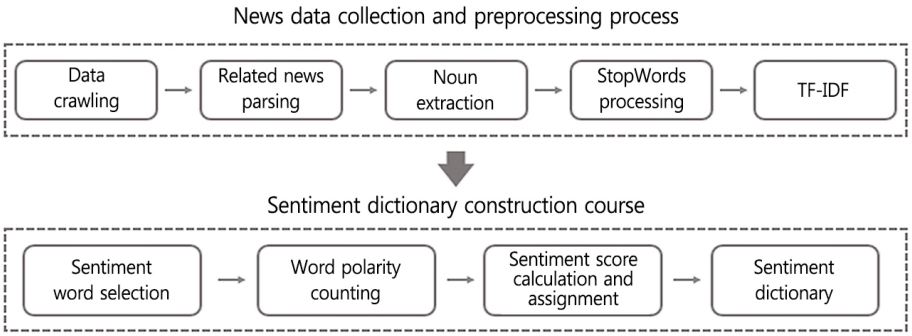
기업 관련 뉴스데이터는 기업과 관련된 현실에서 일어나는 각종 현상에 대한 객관적, 주관적인 설명 등을 포함한다. 또한, 보도의 대상에 대한 제약이 없어 다양한 시각에서 나타나는 방대한 정보원천으로서 활용가치가 있다. 뉴스데이터는 뉴스 이용자에게 적시성이 있는 정보를 제공하기 위해, 전달하고자 하는 내용을 정확하고 빠르게 전달하기 때문에 변화하는 기업의 환경을 분석하기에 적합한 데이터로 사용되고 있다.

재무비율 데이터는 Altman(1968)이 부도예측을 위해 사용한 이후로 부도예측 연구에서 비교 표준으로 사용되고 있다. 후속 연구들을 통해 재무비율 변수가 부도예측에 효과가 있다는 것은 이미 입증이 되었다. 본 연구에서는 뉴스의 분류에 따른 감성분석이 부도예측 성능 향상에 효과가 있는지 비교하기 위해 재무비율 변수를 사용하였다.

2.2 말뭉치기반 감성사전 구축

뉴스 데이터에서 유용한 감성 점수를 추출하기 위해 주제 지향적인 말뭉치기반 감성사전을 구축하고자 하였다. 기업 관련 뉴스는 기업에 일어난 사실들과 관련해서 좋고, 좋지 않은 감정들을 내포하고 있다(Zhang et al., 2011). 뉴스 안에 내포된 뉴스의 감성을 추출할 때 모든 상황에서 사용되는 사전기반 접근법으로 구축된 범용 감성사전은 기업만의 특수한 용어와 그 의미 방향에 맞는 점수를 구하기 어렵다. 따라서 단어의 고유한 의미를 고려한 사전기반 접근법보다 문맥 중심의 감성을 측정한 말뭉치기반 접근법으로 감성사전을 구축하였다. 다음 <Figure 2>는 뉴스데이터를 수집하여 데이터 전처리 이후 감성사전을 구축하는 과정을 설명하는 그림이다.

말뭉치기반 접근법으로 감성사전을 구축하기 위해선 실제 수집한 데이터에서 전처리 과정을 거친 후 구축해야 한다. 데이터를 처리하는 과정에서 부도예측과 관련한 단어를 정제한 다음 단어들의 극성을 점수화하여 단어마다 점수를 구하는 과정이 필요하다.



〈Figure 2〉 Data Pre-Processing and Sentiment Dictionary Building Process

2.3 뉴스 유형

뉴스 유형들 가운데 부도예측을 설명하는 뉴스를 검증하고, 유형에 따라 뉴스를 선별하는 것이 부도예측 능력에 영향을 미치는지 분석하고자 하였다. 뉴스의 유형별 분석의 필요성은 새로운 뉴스들이 지속해서 유입되는 기업 환경에서 텍스트를 분류하고 분석하기 용이하며, 기존의 정의된 개념을 사용하기 때문에 뉴스의 분류가 쉽게 이루어진다. 또한, 뉴스 유형에 따라 전달하는 방식이 다르기 때문에, 부도예측에 있어 어떤 뉴스를 분석하는 것이 효과적인지 알 수 있다. 뉴스 유형에 관한 선행연구와 선행 연구가 사용한 뉴스 유형은 다음 <Table 1>과 같다.

〈Table 1〉 Types of News Previous Research

Author	News Types
Mandel (1962)	spot news, interview, picture, speech press, personnel identification, notification, comments and columns, editorial
Seol (2012)	straight, interviews and speeches, commentary, picture, notification, trot, photos and videos et al.
Bang et al. (2012)	straight, opinions, comments and columns, interviews, report, short article
Kuhn et al. (2013)	straight, columns, comment, interview, picture
Jo(2015)	straight, columns, comment, editorial, report
Salgado et al. (2016)	straight, commentary, columns, interview, picture

뉴스의 유형은 다양하여 서로 엄격하게 구별하기는 어렵다. 연구의 목적에 따라 세분화하여 나누거나, 의미적으로 비슷한 유형을 통합하여 사용하기도 한다. 선행연구들이 사용한 뉴스들의 유형 중 기업 환경을 고려하여 다음과 같이 5가지 유형으로 분류하였다.

1) 스트레이트 뉴스 : 개인적 의견을 배제한 사실 보도 기사
2) 해설 및 논평 : 어떤 문제가 발생했을 때 그에 대한 배경 및 분석과 전망 등을 서술한 기사
3) 칼럼 : 의견이나 주장을 통해 독자를 설득하고자 하는 뉴스
4) 사진 기사 : 사진을 중심으로 내용을 전달하는 뉴스
5) 인터뷰 : 취재원과 기자가 접촉하는 여러 대화 방식을 기사로 작성한 뉴스

스트레이트 뉴스는 개인적 의견보다 사실 위주로 쓴 기사이다. 가장 많은 유형의 뉴스가 스트레이트 뉴스이며, 기업에서 공시한 내용 또는 발생한 사실을 전달하는 내용으로 작성되었다. 해설 및 논평은 기업 관련한 일들에 배경을 전달하거나, 분석하고 전망하는 뉴스를 말한다. 해설 및 논평 유형의 뉴스에는 기업이 발생한 사실의 분석하거나 일어날 일을 전망하는 리포트 형식의 뉴스, 기업이 속한 산업에 대한 현황을 전달하는 뉴스 등이 포함되었다. 칼럼은 의견이나 주장하는 바를 전달하여 설득하는 유형의 뉴스이다. 이는 주로 기업의 상태에 따라 기업의

투자자에게 전달하는 뉴스로 기획된 연재 등이 포함된다. 사진 뉴스는 기업이 활동한 바를 사진과 내용을 전달하지만, 주된 중심내용이 사진에 있는 뉴스이다. 사진 뉴스에는 신제품 출시나 기업의 봉사 활동, 기업의 확장 등 기업 이미지를 위한 뉴스가 포함된다. 마지막으로 인터뷰는 기자가 취재한 내용을 기사로 작성한 뉴스이다. 기업을 취재할 소재가 발생할 때 작성되기 때문에 인터뷰 형식의 뉴스는 다른 유형보다 적게 발생한다.

3. 연구 결과

3.1 자료 수집 및 감성사전 구축

부도예측 연구에서 유용한 결과를 얻기 위해 기업의 부도에 대한 명확한 기준을 설정하는 것은 중요하다. 부도 기업에 대한 정의는(Campbell et al., 2008)과 같이 실적 부진의 이유로 상장폐지 한 기업을 부도 기업으로 정의하였다. 한국거래소 전자공시(KIND)에서 상장폐지 사유를 확인하여 합병, 자진등록취소, 주식분산기준 미달, 지주회사 신규상장, 기타 등록 취소 등 실적 부진과 관계없이 상장폐지 한 기업은 정상 기업으로 간주했다.

분석에 사용한 기업별 재무비율 정보는 KIS Value에서 수집하였다. 2011년부터 2016년까지의 비금융기업 중 부도난 기업을 분석 대상으로 선정하였으며, 부도 관측 기간은 부도가 발생한 시점으로부터 1년 전까지로 설정하였다. 또한, 기업의 재무상태가 이미 악화된 재무정보를 사용하여 부도예측의 오류를 발생시킬 수 있기 때문에, 2년의 시차를 두고 재무정보를 수집하였다.

본 연구는 정상 기업을 언더샘플링 하여 부도예측 연구에 적용한 결과가 예측력에 있어 다른 샘플링 결과와 큰 차이가 없다는(Zhou, 2013)의 연구에 근거하여 부도와 정상 기업의 비율을 1:1로 선정하였다. 또한, 기업의 규모에 따른 부도의 정도가 편향되지 않도록 비례층화추출법을 사용하였다. 비례층화추출법은 모집단에 속한 각층에서 차지하고 있는 비례에 따라 추출하는 방법이다. 본 연구에서는 부도 기업을 먼저 추출한 후, KIS Value에서 제공하는 부도 기업의 크기와 산업을 고려하여 정상 기업을 선정하였다.

재무비율 변수는 Altman이 부도예측에 사용한 다섯 가지 재무 비율 변수를 사용하였다. 다섯 가지 재무 비율 변수는 기업의 안정성, 수익성, 활동성, 유동성을 나타내는 변수로 자리 잡았으며, 부도예측의 비교 표준으로 사용되고 있다. 변수에 대한 설명은 다음 <Table 2>와 같다.

〈Table 2〉 Altman's Financial Ratio Variable

Variable	Variable formula
WCTA	working capital/total assets
RETA	retained earnings/total assets
NITA	operating profit/total assets
METL	equity capital/total liabilities
SLTA	sales/total assets

3.2 뉴스데이터 수집 및 감성사전 구축

뉴스데이터는 우리나라의 대형 포털사이트인 네이버에서 경제 기사를 크롤링 한 후 기업 관련 뉴스 데이터를 파싱하였다. 대형 포털사이트에서 뉴스를 크롤링한 이유는 특정 신문사에서 수집할 경우 특정 신문사가 나타내는 감성에 크게 영향을 받기 때문에 여러 신문사의 데이터를 모두 수집할 수 있는 대형 포털에서 크롤링하였다. 뉴스 데이터의 수집기간은 재무데이터를 수집한 기간과 동일하게 부도시점 이전의 뉴스 데이터를 1년 기간에 걸쳐 수집하였다. 파싱한 뉴스 중 중복된 뉴스 및 불필요한 광고 뉴스, 해당 기업의 내용이 한 문장이 되지 않는 뉴스 등을 제외하였다. 정상 기업과 부도 기업으로 선정한 기업들 중 수집기간인 1년 동안 뉴스의 수가 20건이 되지 않는 기업은 감성분석 결과 편향을 줄 수 있어 제외하였다. 정상 기업과 부도 기업은 206개로, 정상 기업의 수는 103개, 부도 기업의 수는 103개 기업으로 구성되었다. 선정된 206개 기업에 해당하는 뉴스는 17,690개로 한 기업 당 평균 85.8개의 뉴스가 연결되었다.

〈Table 3〉 Top 20 Words by Frequency

Ranking	Positive word(+)		Negative word(-)	
	word	score	word	score
1	cooperation	0.574	limit	-0.517
2	positive	0.461	overdue	-0.515
3	predominance	0.414	clearing	-0.500
4	undervalued	0.393	penalty	-0.484
5	appropriate	0.361	respite	-0.455
6	activity	0.345	suspicion	-0.452
7	availability	0.333	spoliation	-0.450
8	Inflation	0.325	dead	-0.444
9	listed stock	0.321	illegal	-0.444
10	superiority	0.317	argument	-0.441
11	new business	0.314	encroachment	-0.440
12	good	0.313	serious	-0.438
13	upward	0.308	under	-0.436
14	diversification	0.304	incident	-0.429
15	trends	0.304	anxiety	-0.419

〈Table 4〉 Examples of Corpus-based Sentimental Dictionary Words

Ranking	Word	TF-IDF	Frequency
1	scale	3.682	5,212
2	decision	2.194	4,695
3	prepare	2.276	3,783
4	disclosure	3.174	3,233
5	event	2.853	3,170
6	record	3.389	3,007
7	invest	3.952	2,986
8	increase	3.524	2,724
9	operating profit	2.937	2,607
10	transaction	3.631	2,566
11	rising	3.762	2,548
12	degradation	3.813	2,192
13	sales	4.021	2,147
14	stock	3.230	2,035
15	decrease	3.475	1,961
16	acquisition	4.239	1,801
17	conclusion	4.683	1,763
18	contract	3.885	1,727
19	shareholder	4.167	1,656
20	sale	4.571	1,592

뉴스 데이터에서 숫자, 특수 문자 및 구두점을 제거한 후 명사를 추출하였다. 추출된 단어 중 동일한 단어에서 다른 형태로 파생된 단어들을 형태소 분석을 실시해 동일 단어로 취급하였다. 뉴스를 통해 얻은 초기의 용어는 67,000개 정도의 단어 목록이 생성되었다. 이후 TF-IDF를 통해 단어가 사용된 뉴스의 수가 최소 20개(전체 어휘의 약 1%) 이하에서 나타난 용어들은 제거되었다. '부도', '파산', '정리매매', '상장폐지', '상폐', '도산', '폐업'과 같이 부도의 의미를 뜻하거나, 부도 이후 발생하는 사건을 나타내는 단어를 포함한 뉴스는 부도 발생 시점부터 부도 시점 이후의 뉴스들이 대부분이기 때문에 예측에 영향을 줄 수 있어 이를 제거하였다. 또한, 기업명 또는 산업 등의 단어들은 기사와 도메인의 감성을 설명하는 데 적절하지 않은 단어들은 제거하였고, 이후 647개의 단어들이 수집되었다. 수집된 단어들의 빈도 측면에서 가장 높은 순위의 단어들을 <Table 3>에 나타내었다.

기존 선행연구에서 사용한 감성사전을 점수화하는 방법을 사용하여 단어의 감성을 다음과 같은 방법으로 계산한다.

$$Sentiment\,Score(t) = \frac{nomal(t) - bakruptcy(t)}{N(t)}$$

N(t) = 단어 t가 사용된 전체상황의 수
nomal(t) = 정상상황에서 사용된 단어 t의 수
bankruptcy(t) = 부도상황에서 사용된 단어 t의 수

단어의 극성은 단어가 나타난 상황에 따라 결정된다. 단어가 나타난 상황이 긍정적인 상황에서 자주 사용되면 긍정적인 단어, 부정적인 상황에서 자주 사용되면 부정적인 단어로 할당되며 긍정, 부정적인 상황에서 비슷하게 발생하면 중립 단어에 해당한다. 하나의 단어가 수집한 전체 뉴스에서 부도 기업에서 더 많이 사용될수록 단어의 부정 점수가 높아진다.

단어들의 감성점수를 계산하면 단어의 감성 점수는 -1부터 1사이로 나타나며, 0 이하는 부정단어, 0 이상은 긍정단어라 할 수 있다. 이 중 -0.1~ 0.1사이의 단어들은 중립적인 단어들로 단어의 극성이 존재하기보다 긍정, 부정적인 상황에서 모두 사용되는 흔한 단어로 이를 제거하였다. 말뭉치 기반 감성사전 구축 결과 421개의 단어로 이루어진 사전이 구축되었고, 감성의 극성이 높은 15개의 단어 예시를 <Table 4>에 나타내었다.

3.3 분석 결과

본 연구에서는 다양한 모형을 적용하여 부도예측에 적합한 알고리즘을 찾고자 하였으며, 이에 따라 연구에 사용한 모형은 부도예측에서 많이 사용되고 있는 로지스틱 회귀분석, SVM, 인공신경망을 사용하였다. 전체 데이터가 많지 않아 일반화된 모형을 통한 연구결과를 얻기 위하여 교차타당성(cross validation)을 이용하였다. 교차검증은 5-fold를 사용하였다.

로지스틱 회귀분석의 식은 아래와 같으며, 집단에 속할 확률 p를 구하는 공식으로 결과값은 0에서 1사이로 나타난다. 로지스틱에 사용한 변수로 각 기업들의 재무비율 변수 5개와 각 기업에 해당하는 뉴스의 유형별 긍정 점수 평균과 부정 점수 평균 변수 2개를 사용하였다.

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1\chi_1+\beta_2\chi_2\cdots+\beta_q\chi_q)}}$$

SVM의 식은 아래와 같으며, 데이터들이 평면상에 있을 때 집단의 거리를 최대한 멀리 분리해 서로 다른 집단으로 분류시키는 최적의 초평면(hyperplane)을 찾는 알고리즘이다. SVM도 로지스틱과 마찬가지로 재무비율 변수 5개와 긍정, 부정 점수 2개를 사용하였다.

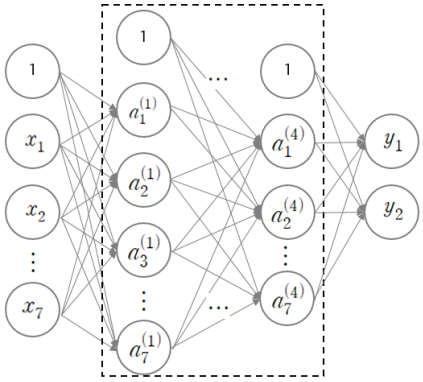
$$D_j = Y \times (\beta_0 + \beta_1 x_{1j} + \beta_2 x_2 \cdots \beta_n x_{nj}),$$
$$j = 1, 2, \cdots, m$$

$$\sqrt{\sum_{i=1}^{n} \beta_i^2} = 1$$

$$D_i \geq M, i = 1, 2, \cdots, m$$

인공신경망은 사람의 두뇌에서 뉴런들이 학습하여 의사결정을 내리는 과정을 본떠 만든 알고리즘이다. 인공신경망의 데이터는 입력층과 출력층 사이의 다층의 은닉층 사이에서 최적의 가중치를 훈련할 수 있다. 인공신경망의 모형도는 <Figure 3>과 같다.

반복적인 실험을 통한 결과 SVM은 Radial 커널, 인공신경망은 Dropout을 사용하며 은닉층의 수가 4개일 때 부도예측의 성능이 가장 높게 나타났다. 본 연구에서 사용한 분석 프로그램은 R 3.3.0 버전의 'e1071', 'caret' 패키지와 텐서플로를 사용하여 분석을 진행하였다.



<Figure 3> Artificial Neural Network Model

기업의 뉴스 데이터는 기업의 크기나 환경에 따라 뉴스 양의 차이가 발생한다. 기업의 감성점수를 뉴스의 양에 따라 뉴스가 많은 기업이 높은 점수가 되지 않도록 기업마다 긍정, 부정 감성의 평균 점수를 사용했다. 긍정 점수와 부정 점수를 하나의 변수로 더하게 되면 긍정과 부정의 의미를 잃어버리며, 또한 감성 값의 차이는 명확해지지 않으므로 긍정 점수와 부정 점수를 각각의 변수로 택하였다.

뉴스 유형은 스트레이트, 해설 및 논평, 칼럼, 사진 뉴스, 인터뷰가 있으며 17,690개의 뉴스에서 유형별 분류했을 때 뉴스의 수와 그 비율은 아래 <Table 5>와 같다.

<Table 5> The Number and Percentage of News Types

Nesw types	Frequency	Ratio(%)
straight	8,633	48.8
commentary	6,970	39.4
column	1,124	6.3
picture	731	4.1
interview	232	1.3

뉴스 유형은 스트레이트, 해설 및 논평, 칼럼, 사진 뉴스, 인터뷰 순으로 많이 나타났으나, 인터뷰 뉴스의 경우 결측값이 60% 이상인 것으로 나타나 분석 대상에서 제외하였다.

재무비율 데이터와 감성분석은 정량적인 수치로 구성되어 있지만, 실제 데이터가 가지는 값에는 차이가 있기 때문에 감성점수를 부도예측에 활용하기 위해선 적절한 방법이 필요하다. 이를 위해 4개 유형의 긍정, 부정 평균 감성점수 변수의 값을 MIN-MAX 정규화하였다. MIN-MAX 정규화는 수집된 변수의 값 중에서 가장 작은 값을 0으로, 가장 큰 값을 1로 만들어 모든 값들을 0에서 1사이에 값으로 변환시키는 것이다. 감성점수를 정규화 한 이후 이를 재무비율과 함께 로지스틱 회귀분석, SVM, ANN 알고리즘을 통해 결과를 도출하였다.

모형의 성능평가는 통계학에서 이진 분류를 측정할 때 많이 사용하는 정오분류표(Confusion Matrix) 중 정확도(accuracy)와 민감도(sensitivity)를 사용하였다. 정확도는 모형이 실제 값의 클래스를 얼마나 정확히 예측했는지 나타내는 값이며, 전체 기업 중에서 모형을 통해 실제 값을 맞힌 비율로 계산할 수 있다. 부도예측 연구의 경우 부도 기업을 정상 기업으로 판단하는 오분류와 정상 기업을 부도 기업으로 판단하는 오분류가 있다. 정확도는 두 오분류에 대한 위험이 동일하다고 가정하는 척도지만, 현실 문제에서는 오분류의 위험성을 다르게 두어야 한다. 연구의 목적은 부도를 정확히 예측하고자 하는 것이기 때문에 실제 부도난 기업 중 부도 기업으로 예측한 비율이 더욱 중요하다. 부도 기업을 모형을 통해 부도 기업으로 예측한 비율을 민감도라고 하며 부도예측 연구에서는 민감도 결과가 중요하다.

감성점수를 사용한 결과와 재무비율 변수만을 사용한 결과를 비교하기 위해 재무비율 변수의 결과값도 함께 제시하였다. 첫 번째 결과는 모든 분류에서 정확히 분류한 한 정도를 나타내는 정확도를 나타내며, 결과는 <Table 6>과 같다.

〈Table 6〉 Prediction Accuracy of the News Type to Bankruptcy Prediction

(Unit : %)

Variable	Logistic regression	SVM	ANN
Financial ratio	81.5	82.5	81.7
Financial ratio +straight	85.4	84.2	85.3
Financial ratio +commentary	85.9	**86.4**	86.2
Financial ratio +column	84.9	84.5	85.5
Financial ratio +picture	81.1	80.1	82.0

정확도는 여러 알고리즘에서 80% 이상의 결과가 나타났다. 부도예측의 정확도는 해설 및 논평 뉴스와 재무비율을 결합한 변수가 SVM 알고리즘을 사용했을 때 가장 높게 나타났다. 사진 뉴스는 비교변수인 재무비율만 사용하여 부도예측의 정확도를 구한 값과 비교했을 때 낮거나 비슷하게 나타났다.

하지만 부도예측은 정확도만 가지고 판단할 수 없다. 부도예측 모형에서 실제 부도 기업을 부도로 분류했는지를 나타내는 민감도를 통해 모형의 성과를 확인해야 한다. 다음 <Table 7>는 뉴스 유형의 감성점수를 재무비율과 결합하여 부도예측을 진행했을 때 민감도를 나타낸 것이다.

〈Table 7〉 Prediction Sensitivity of the News Type to Bankruptcy Prediction

(Unit : %)

Variable	Logistic regression	SVM	ANN
Financial ratio	77.6	78.6	78.4
Financial ratio +straight	82.3	81.9	82.7
Financial ratio +commentary	83.5	83.3	**84.6**
Financial ratio +column	83.2	82.9	84.1
Financial ratio +picture	77.4	78.1	78.8

민감도의 결과는 인공신경망 알고리즘에서 뉴스 유형과 재무비율을 결합한 결과가 가장 좋게 나타났다. 민감도를 기준으로 대부분의 결과가 인공신경망에서 가장 성능이 높게 나타난 것으로 볼 때, 인공신경망이 부도예측에 가장 효과적인 알고리즘으로 나타났다. 스트레이트, 해설 및 논평, 칼럼 유형 모두 재무비율만 사용한 변수보다 향상된 성능을 보였지만, 사진 뉴스와 재무비율을 결합한 변수는 오히려 재무비율만 사용했을 때보다 좋지 않게 나타났다. 뉴스 유형의 감성점수를 재무비율과 함께 사용했을 때 부도예측의 성능은 3~6% 정도 향상된 것을 볼 수 있다. 이로써 재무비율만 사용하여 부도예측 했을 때보다 감성분석을 실시하여 부도예측에 사용하는 것이 더 좋은 결과를 나타내는 것을 알 수 있다.

다음은 뉴스 유형에 사용한 감성변수가 부도 기업을 차별화하는데 중요한 지표인지 알아보기 위해 t-검정을 실시하였다. t-검정은 두 집단 간의 평균이 통계적으로 유의한 차이를 보이는지를 알아보는 검정 방법이다. 다음 <Table 8>에는 뉴스 유형 감성점수의 긍정, 부정 변수들의 부도 기업, 정상 기업의 평균과 t값과 유의수준을 표시하였다.

〈Table 8〉 t-Test Result of News Type Variable

Variable	Bankrupt company	Normal company	t
straight negative	0.303	0.248	-3.442***
straight positive	0.280	0.342	2.538***
commentary negative	0.571	0.441	-5.228***
commentary positive	0.443	0.556	8.142***
column negative	0.451	0.343	-4.968***
column positive	0.391	0.489	3.044***
picture negative	0.385	0.317	-1.521*
picture positive	0.347	0.394	2.171**

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

t-검정결과 뉴스 유형의 변수들은 모두 유의한 것으로 나타났다. 이에 따라 t-검정에서는 뉴스 유형의 긍정, 부정 감성점수 모두 부도예측에 유용한 변수인 것으로 나타났다. 재무비율 변수 또한 t-검정을 실시하여 재무비율 변수가 부도 기업을 차별화하는데 중요한 지표인지 알고자하였다. 재무비율 변수의 t-검정 결과는 <Table 9>와 같다.

〈Table 9〉 t-Test Result of Financial Ratio Variable

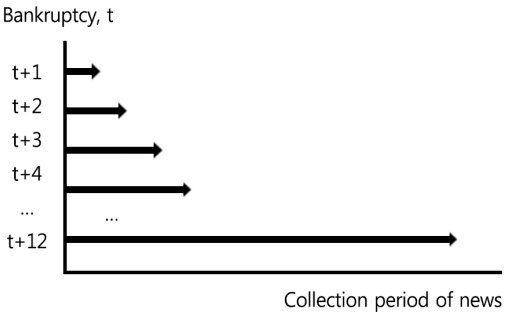
Variable	Bankrupt company	Normal company	t
WCTA	-0.251	0.186	4.517***
RETA	-1.451	0.195	6.336***
NITA	-0.308	0.219	2.716***
METL	1.988	0.634	-2.375***
SLTA	0.548	0.914	-2.543***

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

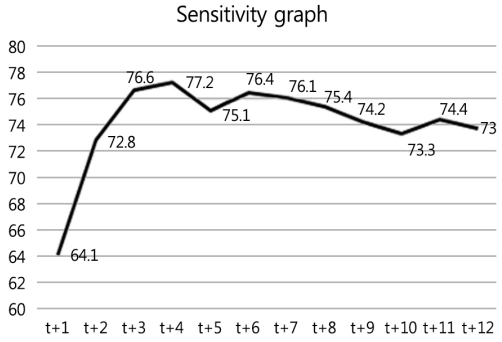
재무비율 변수의 t-검정 결과 5개 변수 모두 부도 기업을 차별화 할 수 있는 지표로 나타났다.

뉴스 수집 기간의 분류가 부도예측에 효과적인지 알고자 부도 발생 시점인 t로부터 1년까지의 뉴스를 수집하여 t+1부터 t+12까지로 분류하였다. 부도는 일정 시점의 감정이 아닌 지속적으로 이루어진 감정이기 때문에 뉴스의 수집 기간을 부도 발생 시점 전 1개월부터 12개월까지 늘려가며 기간에 따른 감성 점수가 부도예측 능력에 미치는 영향을 분석하였다. 수집 기간의 긍정, 부정 감성으로 부도를 예측한 이후 재무비율 변수와 수집 기간에 따른 부도예측을 진행하여 가장 효과적인 수집 기간이 언제인지를 분석하였다. 뉴스의 수집 기간은 다음 <Figure 4>과 같이 수집하였다.

뉴스 수집 기간의 긍정, 부정 감성점수를 가지고 부도예측을 실시하였다. 이때 사용한 알고리즘은 뉴스 유형 분석에서 가장 효과적으로 나타난 인공신경망을 사용하여 뉴스 감성의 민감도 분석을 실시하였다. 결과는 다음 <Figure 5>과 같다.



〈Figure 4〉 Collection Period of News



〈Figure 5〉 Sensitivity Graph According to News Collection Period

뉴스 수집 기간에 따른 민감도의 결과는 t+1을 제외한 t+2부터 t+12까지 부도예측에 유의하게 나타났다. 유의한 결과 중에서도 수집 기간이 t+3, t+4, t+6, t+7일 때, 부도를 부도로 가장 정확히 예측하는 것으로 나타났다. 수집 기간은 일정 기간까지 민감도를 높이다가, 수집 기간이 길어질수록 조금씩 떨어지는 경향을 보였다.

뉴스 수집 기간의 감성을 분석한 이후 재무비율 변수와 수집 기간으로 부도예측을 적용할 때 어느 기간이 부도예측에 가장 효과적인지를 분석하였다. 재무비율과 수집 기간을 사용하여 분석한 결과는 <Table 10>과 같다.

재무비율 변수와 뉴스 수집 기간의 감성점수를 사용하여 부도예측을 실시한 결과 정확도는 t+3, t+4, 민감도는 t+4일 때 예측 능력이 가장 좋은 것으로 나타났다. 이에 따라 뉴스 수집 기간이 t+4일 때 부도예측에 가장 효과적임을 알 수 있었다.

〈Table 10〉 Prediction Results Using News Gathering Period and Financial Ratios

(Unit : %)

Variable	Accuracy	Sensitivity	Specificity
Financial ratio+ (t+3)	87.4	83.5	91.3
Financial ratio+ (t+4)	87.4	85.4	89.3
Financial ratio+ (t+6)	86.4	82.5	90.3
Financial ratio+ (t+7)	86.1	81.1	91.0

4. 결 론

본 연구는 부도에 미치는 영향이 높은 뉴스를 선별하고자 뉴스의 유형, 수집 기간에 따라 분류하고 부도예측을 진행하였다. 이를 위해 206개의 기업에서 17,690건의 뉴스를 크롤링 및 전처리 후 말뭉치 기반 감성사전을 구축하였다. 구축한 감성사전을 뉴스 유형별, 수집 기간별 분류된 뉴스에 적용하여 뉴스의 긍정과 부정 감성을 도출하였다. 도출된 감성과 재무비율 변수를 로지스틱 회귀분석, SVM, 인공신경망 알고리즘에 적용하여 부도예측에 가장 효과적인 뉴스 분류 형태를 찾고자 하였다.

먼저 뉴스의 유형 중에선 해설 및 논평, 칼럼이 스트레이트 뉴스보다 효과적으로 나타났다. 이는 객관적으로 기술한 스트레이트 뉴스보다 해설 및 논평, 칼럼과 같이 의견과 설명, 전망 등을 기술한 뉴스 유형이 부도예측에 효과적인 감성을 더 많이 포함하고 있는 것으로 보인다. 뉴스의 유형 중 인터뷰는 수집된 뉴스에서 결측값이 많아 연구를 진행하지 못하였으며, 사진 뉴스는 주된 내용이 사진을 중심으로 한 뉴스로 기업의 부도예측에는 영향을 주지 못하는 뉴스 유형으로 분석되었다.

또한, 부도예측에 효과적인 알고리즘은 인공신경망인 것으로 나타났다. 뉴스 유형의 감성과 재무비율을 사용한 부도예측에서 실제 부도 기업을 부도라고 예측한 민감도 결과에서 대부분의 유형이 인공신경망에서 가장 높은 예측력을 보였다.

뉴스 수집 기간의 경우 기업의 부도 전 4개월에 해당하는 뉴스를 수집한 결과가 부도예측에 가장 효과적인 것으로 나타났다. 이는 기업의 외부 상황을 전달하는 뉴스가 부도나기 전 기업의 상황이 좋지 않을 때의 감성을 반영하고 있기 때문으로 보인다. 수집기간이 길어질수록 뉴스의 감성분석이 부도예측에 미치는 영향이 줄어듦을 통해, 부도예측에는 적절한 수집기간을 통한 감성분석이 필요함을 보였다.

본 연구는 뉴스가 부도예측에 미치는 영향을 감성분석을 통해 분석하면서, 실시간 대량으로 생성되는 뉴스 중에서 뉴스 분류를 통한 효과적인 감성분석을 제시하고 이를 검증하였다. 유형과 수집기간의 분류에 따라 부도예측에 다른 영향이 있음을 확인하였고, 이에 따라 부도예측을 위해 새롭게 수집되는 뉴스를 사용할 때 뉴스의 유형과 수집 기간설정에 도움을 줄 수 있다.

본 연구의 한계점으로는 규모가 작은 기업들의 뉴스 콘텐츠를 분석하기에 충분한 양을 확보하기 어렵다는 점이었다. 이런 한계점을 극복하기 위해 향후 연구에서는 뉴스 이외에도 기업 관련 댓글, 블로그, 카페 게시물 등을 활용한다면 다양한 콘텐츠 확보는 물론 여러 감성을 활용하여 보다 좋은 모델을 구성할 수 있을 것으로 보인다.

References

Aggarwal, C.C. and C.X. Zhai, *Mining Text Data*. Springer, New York, 2012.

Altman, E.I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The Journal of Finance*, Vol.23, No.4, 1968, 589-609.

Altman, E.I., G. Sabato, and N. Wilson, "The value of non-financial information in small and medium-sized enterprise risk management", *Journal of Credit Risk,* Vol.2, No.6, 2010, 95-127.

An, J.K. and H.W. Kim, "Building a Korean Sentiment Dictionary and Applications of Natural Language Processing", *Proceeding of Korea Intelligent Information Systems Society Conference*, No.11, 2014, 177-182.

(안정국, 김희웅, "한글 감성어 사전 API 구축 및 자연어 처리의 활용", *한국지능정보시스템학회 학술대회논문집*, 제11호, 2014, 177-182.)

Bang, E.J. and S.T. Kim, "A Study of SNS Sources in Korean Major Newspapers", *Journal of Cybercommunication Academic Society*, Vol.29, No.4, 2012, 145-189.

(방은주, 김성태, "국내 주요신문의 소셜미디어 정보원 뉴스보도 분석", *사이버커뮤니케이션 학보,* 제29권, 제4호, 2012, 145-189.)

Beaver, W.H., "Financial Ratios as Predictors of Failure", *Journal of Accounting Research*, Vol.4, 1966, 71-111.

Campbell, J., J. Hilscher, J. Szilagyi, "In Search of Distress Risk", *Journal of Finance*, Vol. 63, No.6, 2008, 2899-2939.

Cho, C.L., *Media Writing*, Communication Books, Seoul, 2015.

(조철래, *미디어 글쓰기*, 커뮤니케이션북스, 서울, 2015).

Dave, K., S. Lawrence, and D.M. Pennock, "Mining the peanut gallery : Opinion extraction and semantic classification of product reviews", *Proceedings of the 12th International Conference on World Wide Web*, 2003, 519-528.

Esuli, A. and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification", *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, 617-624.

Gupta, A., M. Simaan, and M.J. Zaki, "Investigating Bank Failures Using Text Mining", *Computational Intelligence (SSCI),* 2016.

Jo, N.O. and K.S. Shin, "Bankruptcy prediction modeling Using Qualitative Information Based on Big Data Analytics", *Journal of Intelligent and Information Systems*, Vol.22, No. 2, 2016, 33-56.

(조남옥, 신경식, "빅데이터 기반의 정성 정보를 활용한 부도예측모형 구축", *지능정보연구*, 제22권, 제2호, 2016, 33-56.)

Kim, M.H., H.-Y. Jang, Y.-M. Jo, and, H.P. Shin, "KOSAC : Korean Sentiment Analysis Corpus", *Proceeding of Korean Institute of Information Scientists and Engineers Conference*, 제6호, 2013, 650-652.

(김문형, 장하연, 조유미, 신효필, "KOSAC(Korean Sentiment Analysis Corpus) : 한국어 감정 및 의견 분석 코퍼스", *한국정보과학회 학술발표논문집*, 제6호, 2013, 650-652.)

Kim, S.W. and N.G. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification", *Journal of Intelligent and Information Systems*, Vol.20, No.1, 2014, 133-148.

(김승우, 김남규, "오피니언 분류의 감성사전 활용효과에 대한 연구", *지능정보연구*, 제20권, 제1호, 2014, 133-148.)

Kuhn, R. and E. Neveu, *Political Journalism : New challenges*, new practices, Routledg, 2013.

Lu, Y.C., C.H. Shen, and Y.C. Wei, "Revisiting early warning signals of corporate credit default using linguistic analysis", *Pacific-Basin Finance Journal*, Vol.24, 2013, 1-21.

Mandel, S., *Modern Journalism*. Pitman Publishing Corp, New York, 1962.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, "Introduction to wordnet : An on-line lexical database", *International Journal of Lexicography*, Vol.3, No.4, 1990, 235-244.

Ohlson, J.A., "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, Vol.18, No.1, 1980, 109-131.

Salah, Z., F. Coenen, and D. Grossi, *Generating Domain-Specific Sentiment Lexicons for Opinion Mining. Advanced Data Mining and Applications*, Springer, Berlin Heidelberg, 2013.

Salgado, S., J. Strömbäck, T. Aalberg, and F. Esser, *Comparing Political Journalism*. Routledge, London and New York, 2016.

Seol, J.A., "A Study on Attributes of News and Reports in Social News", *Korean Journal of Journalism & Communication Studies*, Vol.57, No.6, 2012, 149-175.

(설진아, "소셜 뉴스의 기사유형 및 뉴스특성에 관한 연구", *한국언론학보*, 제57권, 제6호, 2012, 149-175.)

Yu, E.J., Y.S. Kim, N.G. Kim, and S.R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary", *Journal of Intelligent and Information Systems*, Vol.19, No.10, 2013, 95-110.)

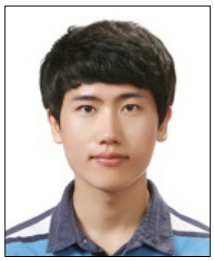
(유은지, 김유신, 김남규, 정승렬, "주가지수 방향성 예측을 위한 주제지향 감성사전 구축 방안", *지능정보연구*, 제19권, 제10호, 2013, 95-110.)

Zhang, Z., Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", *Expert Systems with Applications*, Vol.38, No.6, 2011, 7674-7682.

Zhou, L., "Performance of corporate bankruptcy prediction models on imbalanced dataset : The effect of sampling methods", *Knowledge-Based Systems*, Vol.41, No.1, 2013, 16-25.

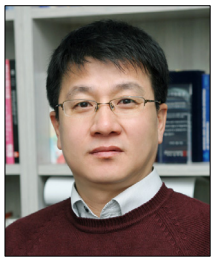
Zmijewski, M.E., "Methodological Issues Related to the Estimation of Financial Distress Prediction Models", *Journal of Accounting Research*. Vol.22, 1985, 55-82.

About the Authors



Chansong Kim (zldlr122@naver.com)

Chansong Kim received M.S. from the Graduate School of Business Informatics, Hanyang University. His interests is in combination of data and text mining and machine learning applications.



Minsoo Shin (minsooshin@hanyang.ac.kr)

Minsoo Shin received the Ph.D. degree in Management Information System from University of Cambridge. He is a professor in Department of Business Administration and Business Informatics of Hanyang University.