기업 부도 사전 예측 모형 연구

– 머신러닝 기법을 중심으로 –

연세대학교 정보대학원

비즈니스빅데이터분석 전공

이 준 기

기업 부도 사전 예측 모형 연구

– 머신러닝 기법을 중심으로 –

지도교수 이 상 우

이 논문을 석사 학위논문으로 제출함

2020년 12월

연세대학교 정보대학원

비즈니스빅데이터분석 전공

이 준 기

이준기의 석사 학위논문을 인준함

심사위원 이 상 우 인

심사위원 김 희 웅 인

심사위원 권 태 경 인

연세대학교 정보대학원

2020년 12월

목 차

표 목 차

그 림 목 차

국 문 요 약

기업 부도 사전 예측 모형 연구

- 머신러닝 기법을 중심으로 -

본 연구에서는 기업의 부도를 사전에 예측하기 위한 모형을 연구하였고 정량 정보인 재무 데이터와 비정형 정보인 뉴스 콘텐츠의 감성분석 정보를 변수로 선정하였다. 재무 정보는 회계와 재무분야의 문헌에서 잘 알려지고 오랜 기간을 통하여 검증된 3개의 재무모델 변수(Altman, 1968; Beaver, 1968; Horrigan, 1966)와 기업의 경영상태를 종합적으로 분석하는 방법인 기업경영분석 지표(한국은행)를 결합하여 총 3개년치의 재무변수를 선정하였다.

기업의 부도 징후를 나타내는 유의미한 재무적 요인을 도출하기 위해 t-test와 logistic regression방법으로 통계적 검증 작업을 진행하였고, 연구 결과 총자산 이익잉여금률, 총자산 이익률, 매출액 운전자본 비율, 자본 매출액 비율, 차입금 의존도가 유의미한 재무 변수로 확인되었다.

비정형 정보인 뉴스 콘텐츠가 기업 부도를 예측하는데 얼마나 효과적인지 검증하기 위해 재무 변수에 뉴스 감성 분석 점수를 추가하여 모델링을 적용하였다. 부도 기업인 경우 부도 직전 6개월치 뉴스를, 정상 기업인 경우 2019. 7 ~ 12월까지의 6개월치 뉴스 기사를 크롤링 하였고 실제 기업 뉴스와 상관없는 기사들은 정제 작업 등의 전처리를 진행하였다. 정제가 완료된

텍스트에서 명사를 추출하여 말뭉치 기반의 감성사전을 구축하고 뉴스의 수집 기간별 감성점수를 변수로 추가하였다.

기업 뉴스의 감성점수를 추가한 결과 재무 데이터만을 사용했을 때보다 훨씬 더 좋은 성능이 나타남을 알 수 있었다. 민감도 기준(실제 부도기업을 부도기업으로 예측)으로 가장 성능이 우수한 SVM(Support Vector Machine)의 경우 재무 변수만 사용했을 때 87.50%였는데 뉴스 감성점수를 추가했을 때 93.75%로 약 6% 정도의 성능 향상이 있었다. 그리고 뉴스 수집기간은 3,4개월치를 적용 했을 때 민감도의 성능이 가장 좋은 것으로 확인되었다.

금융기관에서는 전통적으로 부도 예측을 위해 재무 데이터를 주로 사용하고 있는데 해당 정보는 분기별로 업데이트가 되는 정보의 적시성에 문제가 있을 수 있다. 따라서 부도 예측시 본 연구에서 실증한 온라인 뉴스의 감성분석 정보인 비정형 데이터를 함께 사용한다면 효과적인 여신 의사결정 지원 체계를 수립하는데 많은 도움이 될 것으로 판단된다.

key words : 머신러닝, 감성분석, 부도, TF-IDF, 예측모델링

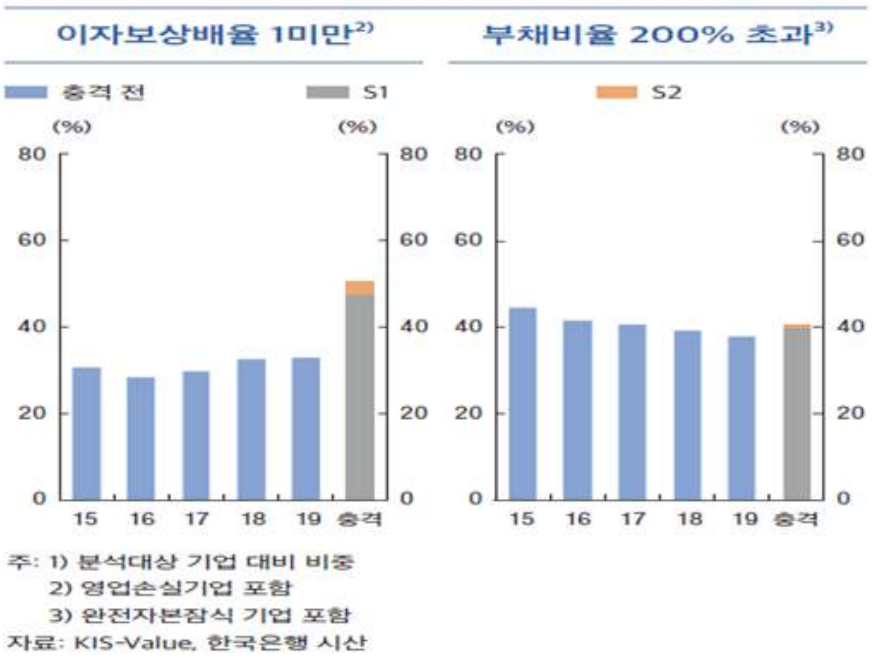
제 1 장 서론

부도 예측 모형은 여러 산업 분야중 특히 금융 산업 분야에서 연구가 활발히 진행되어 왔고 중요한 과제로 인식 되어 왔다. 현대경제연구원에서 발표한 2020 국내외 경제이슈 보고서에 따르면 2019년 경제성장률은 취약한 성장세 지속으로 글로벌 금융위기 이후 최저 수준에 머물고 2020년에도 반등 흐름이 미약할 것으로 예상 했다(현대경제연구원, 2019). 특히 올해는 예상치 못한 코로나(COVID-19) 충격의 장기화에 따라 국내외 경제활동이 위축되면서 기업들의 실적이 크게 악화 되고 있으며 이에 따라 기업의 재무건전성이 저하되고 유동성도 악화될 가능성이 큰 상황이다.

한국거래소 자료에 따르면 연결재무제표를 제출한 유가증권시장 상장법인(594사)의 ’ 20년 상반기 실적중 매출액, 영업이익 및 순이익은 전년 동기 대비 각 5.78%, 24.18%, 34.10% 감소하였으며 삼성전자(매출액 비중 11.48%) 제외시 매출액, 영업이익 및 순이익은 6.46%, 35.38%, 47.08% 감소하였다(한국거래소, 2020).

한국은행에서는 코로나 확산이 기업 매출 및 재무적 충격이 가해지는 스트레스 상황을 올해 3/4분기까지 이어지는 상황을 기본 베이스인(S1)으로, 동 충격이 연중 지속되는 상황을 심각한 시나리오(S2)로 설정하여 기업들의 재무건전성 취약기업 비중 변화를 살펴보았다. [그림 1]의 결과 이자보상비율(영업이익/이자비용) 1미만인 기업 비중은 2019년 32.9%에서 S1 47.7%, S2 50.5%로 크게 증가하였고, 부채비율(부채/자기자본)이 200% 초과하는 기업 비중은 2019년 37.9%에서 S1 39.9%, S2 40.5%로 상승하였다.

[그림 1] 재무건전성 취약 기업 비중 변화

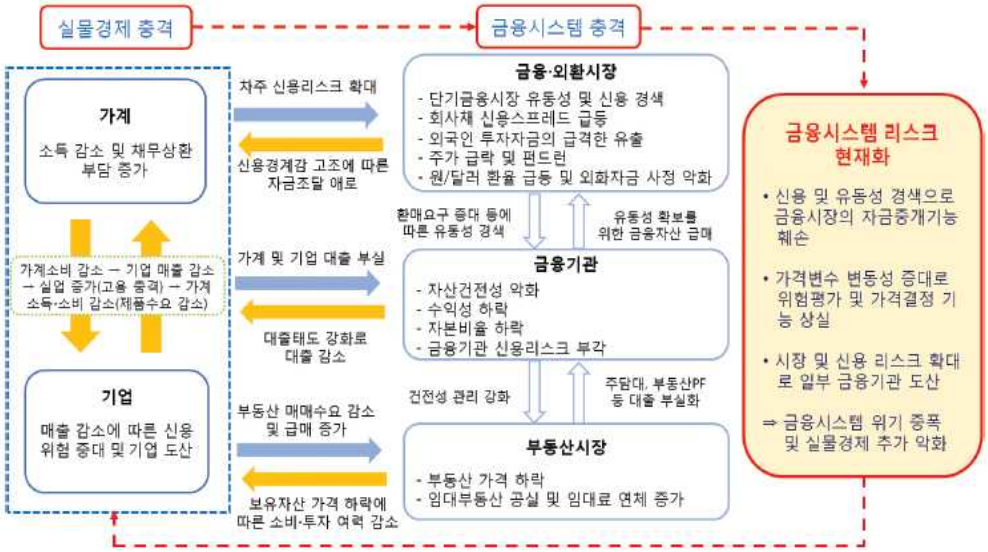


자료 : 한국은행 금융안정보고서(2020)

그리고 국내 기업 가운데 영업이익으로 이자도 못갚는 한계기업의 비중이 작년 14.8%로 사상 최고치를 기록하고 있으며 이는 저금리 기조에 기대어 한계기업들이 장기간 연명하면서 국내 경제에 부담이 되고 있다. 전 세계 부채 규모가 30경원을 돌파하였고 우리나라 기업부채 증가 속도는 경제협력개발기구(OECD) 국가 중 3위에 랭크되어 있다(금융감독원장, 2020.12.07).

한계기업들의 부실화가 심해질수록 [그림2] 와 같이 금융 · 외환시장 및 부동산 시장간의 연계 구조를 통해 금융 시스템 및 실물경제에 충격을 주게 되며 국가의 사회-경제적 구조에 큰 영향을 미치게 된다.

[그림 2] 금융시스템에 발생 가능한 리스크 및 파급경로



자료 : 한국은행 금융안정보고서(2020)

기업의 재무 위험도를 선제적으로 측정하고 모니터링하여 차입금의 회수 가능성을 평가하는 것은 금융기관의 중요한 경쟁력중 하나로 인식되며 기업 부실을 사전에 예측하여 국가 경제적으로 영향을 미칠 손실과 충격을 최소화 시킬 수 있는 대응 방안이 절실히 필요한 시점이다. 대내외 경제 불확실성이 확대된 현 시점이 효과적인 기업 구조조정 전략을 수립할 적기이며 조기 구조조정을 통한 골든타임을 확보하는 것이 중요할 것이다.

따라서 본 연구에서 진행하고자 하는 기업 부도 사전 예측 모형 연구는 기업 여신을 담당하는 금융기관에서 부실 차주를 효과적으로 관리할 수 있는 대응 기반을 제공한다는 점에서 매우 의의가 있다고 볼 수 있다. 본 연구에서 도출된 부도 예측 모델링 기법을 기업여신 실무 담당자에 적용한다면 기업 구조조정을 위한 의사결정에 많은 도움이 될 것으로 기대한다.

선행 연구에 따르면 기업 부도 예측은 꽤 오래전부터 다양한 통계기법을 이용하여 연구가 진행되어 왔다. Beaver(1968), Altman(1968)은 판별분석을 통하여 부도예측 모형을 제시하였고 Ohlson(1980)은 로짓모형을 이용하여 분석하였다. 오세경(2001)은 다변량 판별분석과 KMV로 시계열적인 변화 추이를 분석하였고 Nam et al.(2008)은 재무, 시장 및 거시경제 정보가 부도 예측력을 높이는데 도움이 될 수 있음을 나타내었다.

2010년대에 들어서면서 머신러닝, 딥러닝과 각종 온라인 뉴스를 분석하는 텍스트마이닝 기법을 이용하여 기업 부도 예측 모형의 성과를 측정하고 비교하는 연구가 대부분 진행되었다. 배재권(2006)은 기존 단일모형과 인공지능 기법을 결합한 통합모형의 성과를 비교 측정하였고 민성환(2014)은 배깅(Bagging)의 성능을 개선하는 연구를 진행하였다. Wang et al.(2014)은 재무비율을 활용하여 앙상블 기법의 우수성을 실증하였고 Antunes et al.(2017)은 SVM이 90% 전후의 정확도를 보였다.

최근에는 재무 데이터의 적시성의 한계를 극복하고자 온라인 뉴스의 감성분석 정보를 변수로 추가하여 활용한 연구가 많이 진행되었다. Jo & Shin(2016)은 뉴스에서 추출된 감성점수와 재무비율을 혼합하여 전통적 부도예측 모형보다 성과가 향상됨을 나타내었고 김찬송(2018)은 해설, 논평 관련 뉴스를 재무비율과 같이 인공신경망을 사용하면 가장 높은 예측력을 보였음을 확인하였다.

기존 연구들에서는 대부분 1~3년치의 재무비율을 사용하였으나 연도별로 각각 유용한 재무비율을 통계적인 방법으로 검증하고 추출 후 통합하여 실증한 사례가 없었고 재무비율을 선정하는 이론적 근거가 부족하였다.

이에 따라 본 연구에서는 회계와 재무분야의 문헌에서 잘 알려지고 오랜 기간을 통하여 검증된 3개의 재무모델 변수(Altman, 1968; Beaver, 1968; Horrigan, 1966)와 기업의 경영상태를 종합적으로 분석하는 방법인 기업경영분석 지표(한국은행)를 결합하여 재무변수를 선정하였다. 그리고 온라인 뉴스의 감성분석 정보를 추가적으로 활용하여 랜덤포레스트, SVM 등 다양한 머신러닝 기법을 중심으로 기업 부도 예측 모형을 연구하고자 한다.

제 2 장 선행 연구 및 연구 문제

2.1 머신러닝을 활용한 부도예측

전통적으로 기업 부도 예측에 관한 선행연구를 살펴보면 통계적 모형을 활용한 방법과 머신러닝 기법을 활용한 방법으로 크게 구분하고(김형준 등, 2019), 텍스트마이닝을 활용하여 온라인 감성분석을 적용한 사례로 나누어 볼 수 있다.

초기 부도예측 모형은 1960년대부터 본격적으로 진행되어 Beaver(1968)는 단변량 판별분석을 사용하였고 부도예측을 위한 5가지 재무비율을 현금성자산/총부채, 손이익/총자산, 총부채/총자산, (유동자산-유동부채)/총자산, 유동비율로 선정하였다. Beaver는 부도예측 뿐만 아니라 회계 평가 체계의 틀을 마련한 점이 중요한 의의라고 볼 수 있다. 이후 Altman(1968)은 단변량 판별분석을 보완하고자 다변량 판별분석을 제시하였다. Altman은 22개의 재무 비율중 순운전자본/총자산, 이익잉여금/총자산, 영업이익/총자산, 시가총액/총부채, 매출액/총자산 5개의 재무 비율을 선정하여 판별분석을 통해 기업부도를 예측하는 모델을 최초로 개발하였다. 이러한 판별분석은 기업의 부도를 가장 잘 나타내는 요인들을 추출하고 선형결합하여 판별함수를 계산하는 방식이다. Horrigan(1966)은 사채의 신용등급평가에 대한 모델을 최초로 개발하였고 영업이익/매출액, (자산-부채)/총부채, (유동자산-유동부채)/매출액, 매출액/(자산-부채) 변수를 사용하여 재무 평점을 산출하였다.

Beaver의 모델은 회계학 분야에서 널리 알려져 있으며 Altman의 판별분석 모델은 현재까지도 업종에 상관없이 모든 기업의 재무 건전성을 평가하는 도구로 널리 쓰이고 있으며 퀀트 투자자들 사이에도 많이 알려져 있는 모형이다. Horrigan의 신용평가등급 모델도 오랜 기간동안 널리 알려진 모델이며 Beaver, Altman과 더불어 재무 회계 분야에서 검증되어 왔고 관련 분야에서의 공헌도가 높다고 볼 수 있다.

그 이후 Ohlson(1980)의 로짓모형은 대표적인 이진반응모형으로 발전되었으며 정상기업은 0, 부도기업은 1로 설정하여 변수정보로부터 부도기업이 될 확률을 추정하는 방식이다. 그리고 Cox(1972)의 위험 회귀분석을 사용한 위험모형(Hazard Model)을 Shumway(2001)가 대표적으로 사용했고 이인로 · 김동철(2015)은 회계정보와 시장정보를 이용하여 헤저드 모형을 국내 기업에 적합하도록 수정하여 부도 예측 모형으로 활용하였다.

이재식 · 한재홍(1995)은 비재무정보를 활용하여 인공신경망 기반의 부도예측 모형을 제시하였고 김명종(2009)은 머신러닝 기법에 대하여 앙상블 학습을 적용한 성과를 비교 연구 하였다. 민성환(2014)는 사례 선택(Instance Selection)과 배깅(Bagging)을 연결하는 새로운 모형을 제시하였고 민성환(2016)은 KNN 앙상블 모형의 성능 개선에 관한 연구를 진행하였다. 수익성, 안정성, 성장성 등 24개의 재무비율을 선정하였는데 이에 대한 근거는 없었고 변수 검증은 진행하지 않았다. Alaminos et al.(2016)도 마찬가지로 재무비율 11개를 선정하여 부도 예측을 진행하였지만 변수 검증 작업은 없었다. 오우석 · 김진화(2017)는 인공지능을 이용한 기업부도 예측시 단순히 선행연구에 기초하여 18개의 재무비율을 선정하였고 t-test를 통해 변수 검증을 진행하였다. 김찬송(2018)은 Altman 모델에서 사용한 재무 변수

5개을 사용하여 t-test 검증을 실시하였지만 부도를 예측하는데 재무적 요인이 부족해보였다. 이처럼 대부분의 선행연구에서는 부도 징후를 나타내는 유의미한 재무적 요인 검증에 대해서는 미흡했고 재무 비율도 단순히 KIS-VALUE(나이스 신용평가)에서 제공하는 데이터를 사용하는 등 변수 선정에 대한 명확한 근거를 찾기가 힘들었다.

따라서 본 연구에서는 기존 선행연구들을 보완하고자 학술적 관점으로 재무회계 분야의 검증된 Altman, Beaver, Horrigan 모델에서 사용된 변수를 사용하였다. 재무 변수의 유의미함을 실무적 관점에서도 보완하기 위해 1962년부터 매년 작성해오고 있는 지표인 한국은행의 기업경영분석지표 변수(자기자본비율, 부채비율, 차입금의존도, 유동비율, 비유동비율, 비유동장기적합률, 매출액영업이익률, 금융비용부담률, 이자보상비율, 매출액증가율, 유형자산증가율, 현금흐름보상비율, 현금흐름이자보상비율, 투자안정성비율)도 통합 사용하여 재무 변수 선정시 체계적인 근거를 확보하였다. 그리고 재무비율 3개년치를 통합하여 각 연도별로 어떤 변수가 부도 예측에 유의미한지를 실증하는 부분이 기존 연구와 차별화된 부분이라고 할 수 있겠다. 이를 바탕으로 도출된 연구문제 1은 다음과 같다.

연구문제 1. 연도별 기업의 부도 징후를 나타내는 유의미한 재무적 요인은 무엇인가?

최근에는 머신러닝, 딥러닝 예측 모델링 기법이 발전하면서 전통적인 통계모형 보다는 이와 관련된 알고리즘을 사용한 선행연구가 대다수를 차지하였다. 최소윤·안현철(2015)은 퍼지 이론과 SVM 결합을 통하여 기업부도예측을 최적화 하였으며 오우석·김진화(2017)는 유동비율 등 6개의

유의미한 재무변수를 바탕으로 예측한 결과 의사결정나무모형의 판별 능력이 67.1%로 가장 우수한 것으로 실증하였다. Wang et al.(2014)은 30개의 재무변수를 바탕으로 앙상블 기법을 적용했을 경우 성능이 81.65%로 가장 좋았다. 차성재·강정석(2018)은 다중판별분석, 로짓, Lasso회귀분석을 통해 최적의 변수군 3개를 생성하고 딥러닝 알고리즘(LSTM) 기반의 예측 모형이 97.9%로 성능의 우수함을 확인하였다. 안철휘·안현철(2018)은 ROSE기법과 SVM 알고리즘을 결합시 정확도가 가장 높았음을 확인하였고 Barboza et al.(2017)는 Altman(1968)의 재무비율과 영업이익률 등 6개의 변수를 추가하여 Boosting기법이 86.31%로 가장 정확도가 높았음을 나타내었다. Alaminos et al.(2016)은 부도 직전 3개년치의 재무변수 총 11개를 이용하여 아시아, 유럽, 아메리카 등 글로벌 기업에 대해 부도 예측을 실시하였고 Logit 분석을 통해 직전 1년 재무는 90.11%, 2년 재무는 84.35%, 3년 재무는 78.85%의 정확도를 보였다.

하지만 기존 선행 연구들은 머신러닝 예측 모델링 기법을 대부분 사용하다보니 모델의 성능 이슈에 초점을 많이 맞추고 있으며 모델의 성능중에서도 부도, 정상 기업을 모두 예측하는 정확도를 기준으로 측정하고 있는 연구가 다수를 이루고 있었다. 하지만 부도기업 예측 연구에서는 실제 부도기업을 부도 기업으로 예측하는 민감도가 가장 중요한 요소로 판단된다. 즉, 민감도를 기준으로 성능이 좋은 예측 모형을 찾는 연구가 중요하다고 할 수 있으며 이를 바탕으로 도출된 연구문제 2는 다음과 같다.

연구문제 2. 기업의 부도를 예측하는 여러 머신러닝 기법중 민감도 기준으로 가장 효과적인 성능을 나타내는 기법은 무엇인가?

2.2 비정형 감성 분석 정보를 활용한 부도예측

재무 데이터는 부도 예측 모형에서 전통적으로 널리 사용되어온 변수였으나 재무제표의 결산 시점에 대한 시차 발생과 이로 인한 정보 획득의 적시성 문제 때문에 최근 연구에서는 온라인 뉴스 콘텐츠의 감성 분석을 추가한 연구가 활발히 진행되고 있다.

최정원 등(2015)은 기업의 뉴스 콘텐츠를 중심으로 텍스트마이닝 기법을 통해 기업 부도예측의 가능성을 시도하였다. 부도 발생 2개월전에 워크아웃, 증자, 횡령, 채권단 등과 같은 키워드가 많이 나타났으며 이러한 키워드는 연관성 분석을 통해 상장폐지, 부도 등과 같은 부도 이벤트를 나타내는 키워드와 함께 도출되었다. Gupta et al.(2016)은 텍스트 감정분석을 통해 글로벌 금융 위기 동안 은행의 건전성을 평가하기 위한 새로운 프레임워크를 제시했고 긍정적인 감정이 부정적인 감정보다 더 강한 예측력을 가지고 있음을 발견했다. Jo & Shin(2016)은 빅데이터 기반의 정성 정보를 추가적인 입력 변수로 활용하여 부도 예측 모형을 제안하였고 뉴스 데이터로 구축된 어휘 사전을 기반으로 감성 점수를 부여시 모형의 성과를 개선하는 것으로 나타내었다. 오세경 등(2017)은 재무, 시장, 거시경제 지표에 비정형 정보를 추가하여 부도를 예측하였다. 비정형 정보는 기업의 뉴스 콘텐츠를 크롤링하여 단어들 간의 연관된 규칙을 도출하였고 각 단어의 선후 관계를 벡터 형태로 계산하는 word2vec 알고리즘을 사용하였다. 부도와 연관된 기사 횟수, 비율, 부도 유사도를 변수로 활용하는 비정형 정보를 추가한 결과 연간 예측 모형에서는 부도 예측의 효과가 미미했으나 월간 예측 모형에서는 조기 경보 모형으로 활용될 가능성이 있음을 실증하였다.

한주동(2017)은 기업 뉴스 콘텐츠를 대상으로 문맥으로부터 특정한 단어가 등장할 예측 확률을 최대화 하는 Paragraph Vector 방법론을 활용하여 키워드 가중치를 부여한 방식과 비교하였다. 기업의 신용을 판단하는 분류 성능에서는 유사한 수준이었지만 Paragraph Vector가 뉴스를 수치화하는데 연구자의 주관이 개입될 여지가 적어서 더 유연하게 활용될 수 있음을 시사하였다.

김찬송(2018)은 기업 뉴스 콘텐츠를 크롤링 후 말뭉치 기반의 감성사전을 구축하고 단어들의 극성을 수치화 하여 점수를 할당하였고 기업 부도 4개월전의 뉴스 감성분석 결과를 바탕으로 부도예측한 모형이 가장 효과적인 것으로 입증하였다. 추가적으로 뉴스의 유형은 기업에 대한 의견이 포함된 해설 및 논평, 컬럼이 객관적으로 기술한 스트레이트 뉴스보다 부도예측에 더 효과적임을 밝혔다.

부도 예측 이외에도 비정형 감성 분석 정보를 이용한 연구에는 주가지수 예측, SW교육, 정치 토론, 영화 평점 예측 등 다양한 분야에서 활용되고 있으며 그 효과에 대해서도 입증되어 왔다. 이상훈 등(2016)은 영화 장르별 평점 예측시 범용적인 감성사전 대비 말뭉치 기반의 맞춤형 감성사전이 더 효과적임을 입증하였다. 유은지 등(2013)도 주가지수의 등락을 예측하기 위해 주식 도메인에 특화된 주제지향 감성사전을 사용한 모형이 범용 감성사전을 사용한 모형보다 예측력이 우수한 것으로 나타났다.

기존 선행연구에서 입증되었듯이 본 연구에서도 기업 부도 예측의 성능을 향상시키기 위해 감성 분석 정보를 활용한 변수를 추가하였고 일반적인 범용 사전보다는 성능이 우수한 말뭉치 기반의 감성 사전을 구축하는 방법론을 사용하였다. 따라서 본 연구에서도 연구문제1에서 추출한 변수에 비정형 감성분석 정보를 추가하였을 경우 기업의 부도를 예측하는데 얼마나 효과적인지 실증해보고자 한다.

연구문제 3. 기업의 부도를 예측하는데 비정형 데이터(온라인 뉴스 콘텐츠)는 얼마나 효과적인가?

제 3 장 연구 방법

본 연구에서는 <연구문제 1>에 대한 결과를 확인하기 위해 정상(2019년말 기준) 및 부도 기업(부도 직전 년도)의 3개년치 재무 데이터를 가져와서 재무 비율을 직접 계산하였다. 재무 데이터는 나이스 신용평가에서 제공하고 있는 기업 분석 솔루션인 KIS-VAlUE를 이용하였고 재무 비율은 3가지 재무모델 변수(Altman, 1968; Beaver, 1968; Horrigan, 1966)와 한국은행 기업경영분석지표의 재무 비율로 사용하였다. 수집된 재무 비율을 바탕으로 통계적인 검증작업을 통해 연도별로 유의미한 재무 비율에 대해 검증하였다.

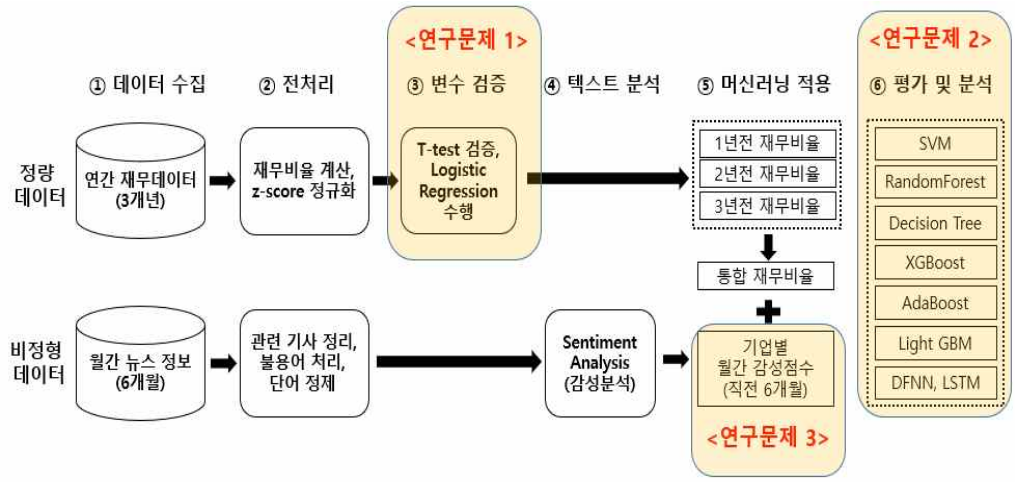
<연구문제 3>에 대한 결과를 얻기 위해 네이버 포털에서 검색된 정상 및 부도 기업 관련 뉴스 콘텐츠를 크롤링하여 말뭉치 기반의 감성 사전을 구축하고 기업별 감성 점수를 재무 변수에 추가하였다. 수집된 뉴스 콘텐츠로부터 불용어 처리, 정제 작업 등을 거쳐 명사를 추출하였고 기업 뉴스에서 도출된 단어의 빈도수 등을 계산하여 유의미한 단어 321개를 추출하여 감성 사전을 구축하였다. 감성 사전의 단어 점수는 -1에서 1사이로 나타나며 기업 뉴스별 감성 점수를 부여하고 기업별 평균 점수를 사용하여 감성 변수를 생성하였다. 6개월 치의 뉴스 데이터를 가지고 각 개월수별로 어느 시점의 뉴스가 부도 예측에 효과적인지, 재무 비율 변수만 사용하였을 때보다 모델의 성능 향상에 얼마나 도움을 주는지 등을 분석하였다.

<연구문제 2>를 확인하기 위해서 전통적 이진분류 모형인 Logistic Regression에서 딥러닝의 대표적인 모델인 LSTM(Long Short-Term Memory)까지 머신러닝 및 딥러닝의 다양한 알고리즘(10개 모형)을 적용하고 실제

부도기업을 부도기업으로 예측하는 민감도를 중심으로 성능에 대한 비교 평가를 진행하였다.

본 연구의 전체적인 절차는 [그림 3] 과 같다. t-test 검증 및 Logistic Regression 수행은 IBM Statistics SPSS 25를 사용하였고 나머지 뉴스 크롤링, 머신러닝 적용, 평가 및 분석 등은 Python을 사용하였다. 정량 데이터인 재무 비율의 유의성을 통계적인 방법으로 검증하여 <연구문제 1>을 해결하고 <연구문제 3>을 위해 비정형 데이터인 뉴스 콘텐츠를 부도 직전 6개월간 수집였다. 전처리가 완료된 뉴스 콘텐츠로 감성분석을 실시하고 기업별 월간 감성점수를 계산하여 재무 변수에 추가하였다. <연구문제 2>는 <연구문제 1>에서 확인된 통계적으로 유의한 변수만을 대상으로 <연구문제 3>에서 계산된 기업별 월별 감성 점수를 통합하여 다양한 머신러닝 예측 모델링 기법을 사용하여 성능을 비교하였다.

[그림 3] 연구 진행 절차



3.1 데이터 수집 및 전처리

3.1.1 정상 및 부도 기업 표본 선정

부도 기업 예측의 유용한 결과를 도출하기 위해 한국 코스피, 코스닥 시장에서 상장폐지가 결정된 기업들을 부도기업으로 한정하였다. 상장폐지 사건이 부도와 100% 일치하는 개념은 아니지만 보수적인 관점에서 상장폐지를 부도로 인식하여 연구를 진행하였다.

부도 기업에 대한 표본은 한국거래소(KRX)에서 제공되는 정보를 바탕으로 2010년부터 2019년까지 상장폐지된 기업 리스트를 총 593건을 수집하였다. 피흡수합병, 이전상장, 존속기간 만료 등 부도와 관련없는 기업을 삭제하였고 최종적으로 265개의 기업을 선정하였다.

정상 기업에 대한 표본은 2020년 7월말 기준 코스피 및 코스닥에 상장된 기업중 기업인수목적회사(SPAC)인 페이퍼 컴퍼니, 금융지주 회사 등을 제외하고 최종적으로 2,071개 기업을 선정하였다.

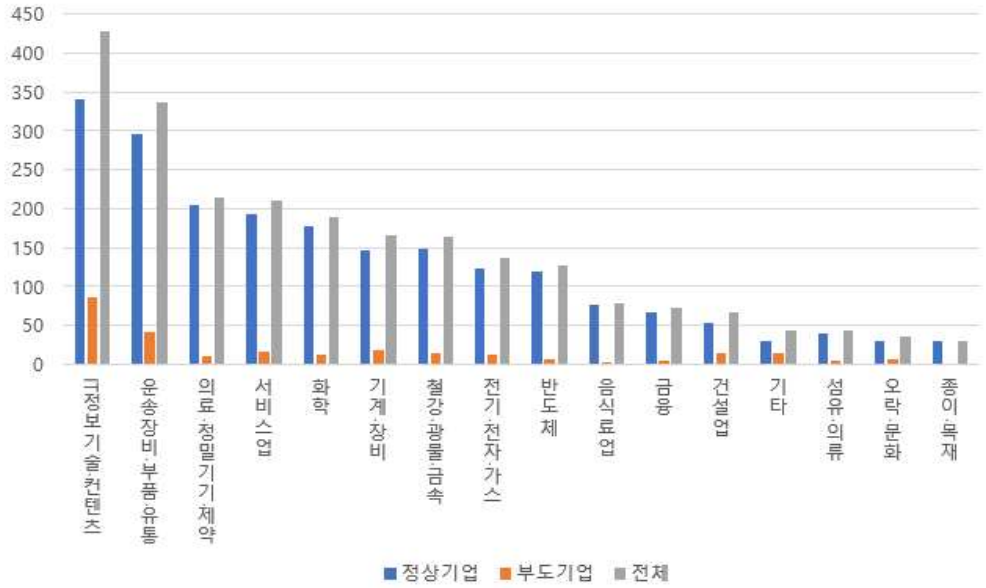
아래 [그림 4]와 같이 연도별 기업 규모별 부도 기업 현황을 살펴보면 대기업은 총 21개 기업이고 중소기업이 244개 기업이었다. 대기업의 경우 2018년 이후부터는 부도 기업이 없었으며 중소기업의 경우 2010년에 가장 많은 부도 기업이 발생했다.

[그림 4] 연도별 기업 규모별 부도 기업 현황

자료 : 한국거래소(KRX Marketdata) 상장폐지현황 자료 재구성

업종별 정상 및 부도 기업 현황을 살펴보면 IT정보기술·콘텐츠 업종의 부도 기업이 86개로 가장 많았으며 그 뒤로 운송장비·부품 업종 41개, 기계·장비 업종 18개 순으로 높았다.

[그림 5] 업종별 정상, 부도 기업 현황



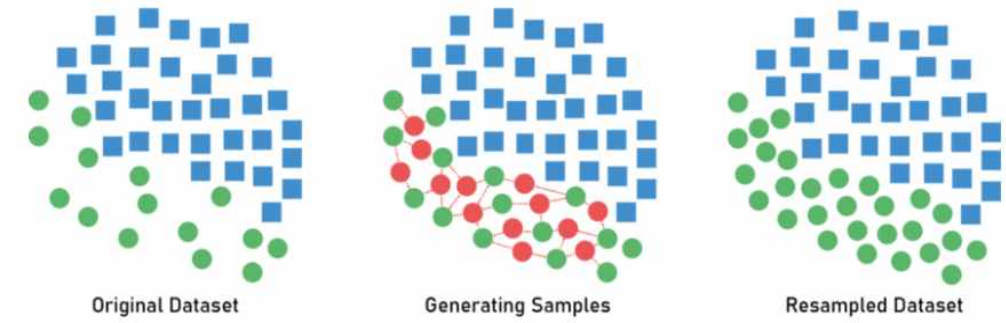
자료 : 한국거래소(KRX Marketdata) 상장폐지현황, 나이스평가정보(KIS-VALUE) 자료 재구성

정상 기업(2,071개)과 부도 기업(265개) 데이터의 비율이 88.7%(정상 기업), 11.3%(부도 기업)로 데이터의 비율이 어느 한쪽으로 과도하게 치우친 데이터 불균형의 문제가 발생하였다. 특정 범주의 빈도가 지나치게 높을 경우 머신러닝 기반 예측 모델링 적용시 의도치 않은 결과가 나올 수 있다. 이러한 데이터 불균형 문제를 해결하기 위해서 샘플링 방법을 적용하고자 한다. 샘플링 방법에는 언더샘플링(undersampling) 방법과 오버샘플링(oversampling) 방법이 있는데 먼저 언더샘플링의 경우 다수 범주의 데이터 크기를 소수 범주의 데이터 크기 만큼 줄이는 방법이다. 이 방법을 사용하면 전체 데이터 크기가 줄어들게 되어 유용한 정보가 그만큼 없어지게 되는 단점이 있다.

반면 오버샘플링의 경우 소수 범주의 데이터를 복원 추출하여 다수 범주의 데이터 비율과 동일하게 늘리는 방법인데 소수 범주의 데이터 크기가 작을 때 효과적이다. 본 연구에서는 오버샘플링을 적용하였고 해당 기법중 가장 널리 사용되고 있는 SMOTE(Synthetic Minority Oversampling Technique) 기법을 사용하였다(Chawla et al. 2002).

SMOTE 기법은 [그림 6]과 같이 단순하게 데이터를 무작위로 복원하여 추출한 것이 아니라 기존 샘플을 주변의 이웃을 고려해서 k-neighbors를 기반으로 하는 방식이다. 현재 데이터 샘플과 찾은 k개 이웃 사이의 거리를 측정하고 0과 1사이의 임이의 값을 곱하여 기존 데이터 샘플을 더하면서 데이터가 생성되는 방식으로 과적화 문제를 해결할 수 있는 장점이 있다.

[그림 6] SMOTE 기법



자료 : https://blog.naver.com/zeroalgorithm/222154416814

3.1.2 재무 데이터

재무 변수는 전통적 통계모형에서 오랫동안 검증되었고 부도예측 연구에서 표준으로 사용되고 있는 3가지 재무모델 변수(Altman, 1968; Beaver, 1968; Horrigan, 1966)를 사용하였다. 조정만(2005)은 3가지 재무모델의 유용성에 대해서 검증을 하였고 Barboza(2017), 김찬송(2018)에서는 부도 예측시 Altman 모델에서 사용한 재무 비율을 사용하여 해당 변수에 대한 유용성을 검증한 바 있다. 그리고 실무적인 시사점을 얻기 위해 한국은행에서 1962년부터 매년 작성해오고 있는 기업경영분석지표에서 사용하고 있는 재무비율을 추가로 사용하였다. 한국은행의 기업경영분석은 은행과 같은 금융기관들이 거래기업의 신용상태를 파악하기 위해 비롯되었으나 최근에는 경영자나 투자자의 의사결정, 신용평가기관의 기업평가, 정부의 기업 정책 수립 등 다방면으로 활용되고 있다. 본 논문에서 사용한 재무비율 변수는 총 22개이며 아래 [표 1]과 같다.

[표 1] 재무비율 변수

참조	재무비율	비율 공식	변수명
A	총자산 이익잉여금률	이익잉여금/총자산	alt001
A	총자산 영업이익률	영업이익/총자산	alt002
A	총자산 매출액 비율	매출액/총자산	alt003
B	총자산 이익률	순이익/총자산	bea001
B	총자산 총부채 비율	총부채/총자산	bea002
A,B	총자산 운전자본 비율	(유동자산-유동부채)/총자산	bea003
B,한	유동비율	유동자산/유동부채	bea004

H,한	매출액 영업이익률	영업이익/매출액	hor001
H	총부채 자본비율	(자산-부채)/총부채	hor002
H	매출액 운전자본 비율	(유동자산-유동부채)/매출액	hor003
H	자본 매출액 비율	매출액/(자산-부채)	hor004
한	자기자본비율	자기자본/총자본	bok001
한	차입금의존도	차입금/총자본	bok002
한	비유동비율	비유동자산/자기자본	bok003
한	비유동장기적합률	비유동자산/(자기자본+비유동부채)	bok004
한	금융부담률	금융비용/매출액	bok005
한	이자보상비율	영업손익/이자비용	bok006
한	매출액증가율	당기매출액/전년동기매출액	bok007
한	유형자산증가율	당기말유형자산/전기말유형자산	bok008
한	현금흐름보상비율	(영업활동으로인한 현금흐름+이자비용)/(단기차입금+이자비용)	bok009
한	현금흐름이자보상비율	(영업활동으로인한 현금흐름+이자비용)/이자비용	bok010
한	투자안정성비율	영업활동으로인한 현금흐름/유형자산 투자지출	bok011

A : Altman(1968), B : Beaver(1968), H : Horrian(1966), 한 : 한국은행

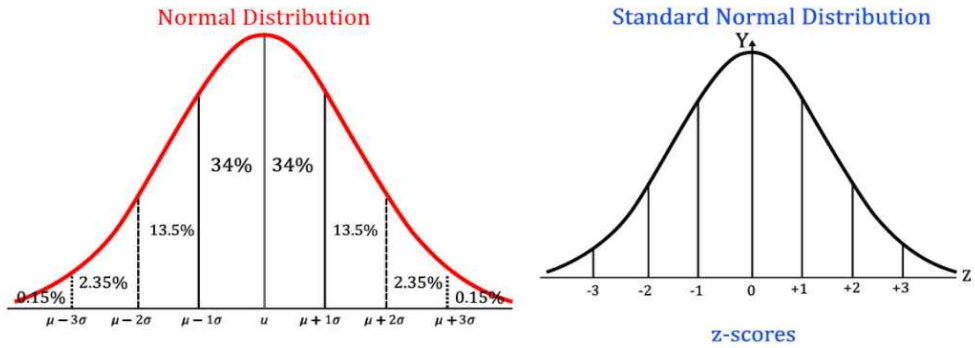
기업의 재무 데이터는 금융감독원의 전자공시시스템(DART)를 통해서 쉽게 조회가 가능하지만 개별적으로 접근해야 되는 한계점이 존재하여 본 연구에서는 나이스 신용평가에서 제공하고 있는 기업분석 솔루션인 KIS-VALUE를 이용하여 개별 재무 데이터를 수집 후 [표 1]의 비율 공식으로 재무비율 변수를 직접 계산하였다.

부도 기업의 경우 부도 직전 3개년도(해당년도 연말 기준)의 재무 데이터를 수집하였고 정상 기업의 경우 일괄적으로 2017년말 ~ 2019년말 기준의 3년치의 재무 데이터를 수집하였다.

특정 기업에서는 일부 재무 데이터에서 결측치가 발생한 경우가 있었는데 기업규모별(대기업, 중소기업) KIS-Sector별 해당 변수의 평균값으로 대체하였고 모든 재무변수는 정규화를 실시하였다.

다양한 범위의 값으로 이루어진 데이터를 동일한 모델에 적용하기 위해서는 데이터 Scale을 일정하게 만드는 작업이 필요하고 이를 데이터 정규화(Normalization)라고 한다. 본 연구에서는 각 변수들의 값에서 평균값을 뺀 후 표준편차로 나누어준 값을 사용하는 z-score 정규화 방식을 사용하였다 [그림 7].

[그림 7] z-score 정규화



자료 : https://calcworkshop.com/functions-statistics/z-score/

3.1.3 뉴스 데이터

뉴스 데이터는 국내 최대의 포털인 네이버 뉴스에서 기업명 검색어 기반으로 크롤링 하였으며 검색된 뉴스중 네이버 뉴스만 파싱하여 크롤링을 수행하였다 [그림 8]. 네이버뉴스가 아닌 일반 뉴스들의 경우 HTML 구조가 인터넷 언론사 마다 제각각이라 크롤링을 수행하는데 큰 어려움이 있었고 HTML 구조가 표준화 된 네이버 뉴스를 기반으로 데이터를 크롤링하여 수집하였다.

[그림 8] 네이버 뉴스 크롤링

자료 : https://search.naver.com/search.naver?where=news&sm=tab_jum&query=삼성전자

부도 기업의 경우 부도 직전 1개월~6개월치의 뉴스 기사를 누적하여 크롤링을 수행하였고 정상 기업의 경우 재무 데이터에서 수집한 2019년말의 기준으로 2019. 7.1 ~ 12.31까지 총 6개월치의 뉴스 기사를 크롤링 하였다.

부도 기업 265개의 6개월치 뉴스는 총 58,907건이 수집되었고 정상 기업 2,071개의 6개월치 뉴스는 총 423,867건이 수집되었다. “우영” 이라는 기업은 연예인 기사가 다수를 차지하였고 “대국” 이라는 기업은 무역 강대국, 북한 관련 기사 등 실제 기업과 상관없는 기사들이 많이 포함된 경우는 직접 확인하여 삭제를 하였다. 그리고 자주 반복되는 기자 이름, 언론사 이름 등 불용어를 처리하고 기타 필요한 전처리 작업을 수행한 결과 부도 기업은 32,921건, 정상 기업은 206,235건, 총 239,156건의 기업뉴스가 분석 대상으로 선정 되었다.

[표 2] 정상, 부도 기업별 뉴스 크롤링 건수

구 분	정제 前	정제 後
정상 기업	423,867	206,235
부도 기업	58,907	32,921
합 계	482,774	239,156

3.2 연구 방법

3.2.1 재무 변수 유의성 검증

부도 예측에 유용한 재무 비율 변수를 검증하기 위해 정상 기업과 부도 기업간 평균 차이를 검증하는 독립표본 t-test(independent two sample t-test)를 실시하여 유의미한 변수를 추출하였다. t-test는 모집단의 분산이나 표준편차를 알지 못할 때 표본으로부터 측정된 분산이나 표준편차를 이용하여 두 모집단의 평균 차이를 알아보는 검정 방법이다.

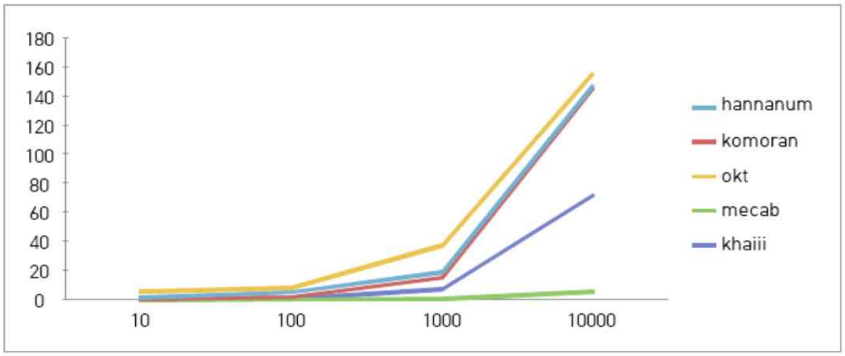
추가적으로 Logistic Regression을 이용하여 정상 기업과 부도 기업을 구분짓는데 영향을 미치는 유의미한 변수를 추출하였다. Logistic Regression은 분석을 위한 회귀분석 중에서 특히 종속 변수가 이분형(정상, 부도)일 때 수행할 수 있는 회귀 분석 기법의 한 종류로 이항형 로지스틱의 회귀 분석에서 2개의 카테고리는 0과 1로 나타내어지고 각각의 카테고리로 분류될 확률의 합은 1이 된다. 이렇게 t-test와 Logistic Regression을 수행하여 둘 다 유의미하게 만족하는 재무 변수를 최종 선정하여 부도 예측에 유용한 변수로 선정하였다.

3.2.2 뉴스 콘텐츠 감성 분석

재무 데이터의 업데이트 시점에 대한 적시성을 보완하기 위해 실시간으로 정보 획득이 가능한 기업 뉴스 콘텐츠의 감성 분석을 진행하였다. 분석을 위해 활용된 Tool은 Python의 Konlpy를 사용하였다. Konlpy는 한글 형태소를

분석하는 대표적인 패키지이며 Hannanum, Komoran, okt, mecab 등 다양한 형태소 분석기가 존재한다. 본 연구에서는 약 24만여건이나 되는 대량의 뉴스 데이터를 분석해야 했기 때문에 연산속도 측면에서 훨씬 유리하고 비교적 성능도 좋은 mecab을 사용하여 뉴스 콘텐츠의 명사를 추출하였다.

[그림 9] Konlpy 형태소 분석기별 연산 속도 비교



자료 : https://passerby14.tistory.com/3

추출된 명사를 바탕으로 문서 내의 가중치를 계산하여 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치인 TF-IDF(Term Frequency - Inverse Document Frequency)를 사용하였다. 단어 빈도 또는 등장 여부를 그대로 쓰면 특정 단어가 많이 나타났다 하더라도 어떤 문서에든 쓰이기 때문에 문서의 주제를 가늠하기 어려울 수가 있어서 TF-IDF는 이러한 단점을 보완하기 위해 제안된 기법이다.

TF(단어 빈도, Term Frequency)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값이며 IDF는 역문서 빈도(Inverse Document

Frequency)라고 하며 TF-IDF는 TF와 IDF를 곱한 값이다. 특정 문서 내에서 단어 빈도가 높을 수록, 그리고 전체 문서들 중 그 단어를 포함한 문서가 적을 수록 TF-IDF값이 높아지며 해당 값을 이용하여 정상 및 부도 기업 뉴스에서 주로 나타나는 단어들의 빈도를 계산하였다.

$$TF-IDF(w) = TF(w) \times \log(\frac{N}{DF(w)}) \quad \text{식(1)}$$

감성 분석을 하기 위한 첫 단계로 단어의 극성 판별을 통해 감성사전을 구축하였다. 일반적인 범용감성사전에 비해 실제 수집된 텍스트에서 문장 분석을 통해 구축되는 말뭉치 기반의 주제별 맞춤형 감성 사전이 연구에 더 효과적이라는 연구(이상훈 등 2016)에 따라 말뭉치 기반으로 감성 사전을 구축하였다.

전처리를 통해 만들어진 단어별 TF-IDF값을 통해 정상 기업 뉴스에서 추출한 단어의 빈도와 부도 기업 뉴스에서 추출한 단어의 빈도를 집계하여 단어 사전의 목록을 만들었다(김찬송, 2018).

$$\text{단어감성점수}(t) = \frac{\text{정상}(t) - \text{부도}(t)}{\text{전체}(w)} \quad \text{식(2)}$$

이렇게 만들어진 단어 목록을 바탕으로 각 정상 기업, 부도 기업 뉴스 전체에서 단어 w가 사용된 수를 전체(w)라고 하며 정상 기업 뉴스에서 사용된 단어 수인 정상(t)에서 부도 기업 뉴스에서 사용된 단어 수인 부도(t)를 뺀 값을 전체(w)로 나누어 식(2)와 같이 감성점수를 계산한다.

단어의 감성 점수는 -1에서 1 사이로 나타나며 부정적인 단어는 -1에 가깝고 긍정적인 단어는 +1에 가까워 진다. 이렇게 구축한 감성사전의 점수를 개별 기업의 뉴스에 적용하여 긍정 점수의 평균값과 부정 점수의 평균값을 사용하여 감성 변수를 생성하였다.

3.2.3 예측 모델링 기법

본 연구에서 부도예측을 위한 방법론으로 머신러닝과 딥러닝 방법을 적용하고자 한다. 전통적 이진분류 모형인 Logistic Regression에서 딥러닝의 대표적인 모델인 LSTM(Long Short-Term Memory)까지 다양한 분석 알고리즘을 적용하고 정확도에 대해서 비교 평가를 진행하였다. 예측 모델링에는 Python의 scikit-learn 및 Keras 라이브러리를 활용하였다.

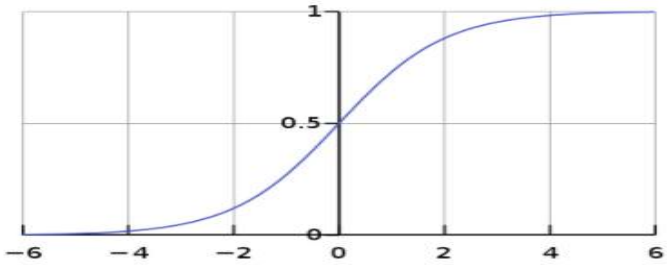
Decision Tree는 기준값을 노드로 설정하여 Tree 구조의 분류 나무를 만들고 그 결과들을 도식화 한 의사 결정 지원 도구의 일종이다. 의사 결정 분서에서 목표에 가장 가까운 결과를 낼 수 있는 전략을 찾기 위해 주로 사용되며 본 연구에서도 부도 결과를 예측하기 위한 방법으로 사용 하였다.

분석 방법에는 CART(Classification And Regression Tree), CHAID, C4.5, C5.0 등이 있으며 이중에서 CART가 가장 많이 사용되며 지니지수를 이용하여 최적의 설명변수를 찾아내는 방식이다(Lewis, 2000).

Logistic Regression은 선형 함수 대신 최적의 시그모이드 함수를 도출하고 독립변수를 이 시그모이드 함수에 입력해 반환된 결과를 확률값으로 반환해 예측 레이블을 결정한다. 독립변수의 선형 결합으로 종속 변수를

설명한다는 관점에서는 선형 회귀분석과 유사하지만 로지스틱 회귀는 종속변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류(classification) 기법으로 볼 수 있다. 로지스틱 함수의 그래프는 아래 [그림 10]과 같고 독립변수 x가 주어졌을 때 종속변수가 1의 범주에 속할 확률을 의미한다.

[그림 10] Logistic 함수



자료 : https://ko.wikipedia.org/wiki/로지스틱 회귀

즉, p(y=1|x)를 의미하며 로짓 변환을 통해 만들어진 로지스틱 함수는 다음과 같다.

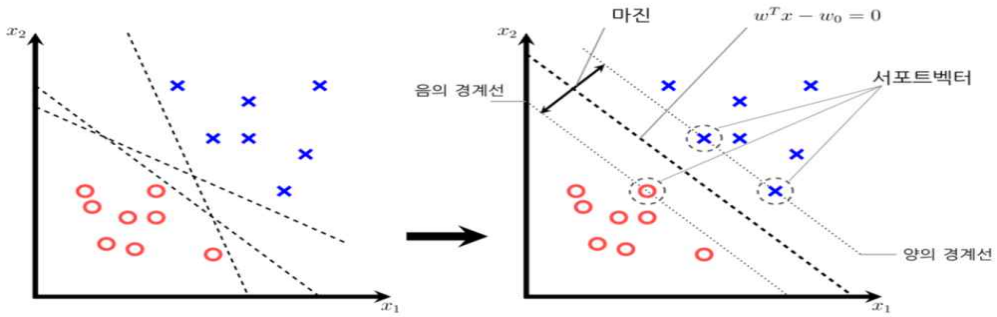
$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} \qquad \text{식(3)}$$

k-최근접 알고리즘(KNN)은 모두 입력이 특징 공간 내 k개의 가장 가까운 훈련 데이터로 구성되어 k개의 최근접 이웃이 가진 값을 찾는 방식이다. KNN의 하이퍼파라미터는 탐색할 이웃 수(k), 거리측정 두가지 방법이며 거리 측정시 흔히 사용하는 Euclidean Distance는 두 관측시 사이의 직선 최단 거리를 의미한다.

Support Vector Machine(SVM)은 새로운 데이터가 어느 카테고리에 속할지 판단하는데 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. SVM은 고차원의 특징 공간에서 클래스를 잘 나눌 수 있도록 초평면(hyper-plane)을 학습하여 결정하는데 RBF 커널 SVM에서 Polynomial 커널, Sigmoid 커널, 가우시안 RBF 커널중 가장 많이 사용하는 방식이 RBF(Radial Basis Function) 커널이다. C, gamma 라는 두 개의 매개변수가 사용자에 의해 지정되어야 하며 여러 조합들을 테스트해서 가장 좋은 성능을 내는 매개변수를 찾아낼 수 있다. C는 데이터들이 다른 클래스에 놓이는 것을 허용하는 정도이며 gamma는 결정 경계의 곡률을 결정한다.

선형 SVM의 경우 [그림 11]와 같이 데이터를 선형으로 분리하는 최적의 선형 결정 경계를 찾는 방식이며 클래스가 다른 데이터들을 가장 큰 마진(margin)으로 분리해 내는 선을 찾는다. 이 때 마진이란 두 클래스군과 결정 경계와 얼마나 떨어져있는지 정도를 의미하고 마진이 가장 큰 결정 경계를 찾는 것이 목표이다. 서포트 벡터는 두 클래스 사이의 경계에 위치한 데이터들을 말하며 해당 벡터들이 결정 경계를 만드는데 영향을 주게 된다.

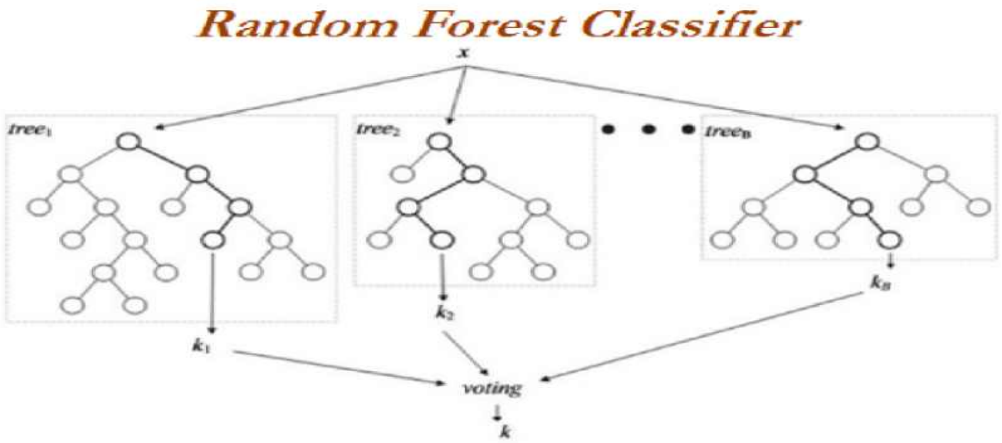
[그림 11] Support Vector Machine



자료 : https://blog.naver.com/daily__record/222013754730

RandomForest는 배깅의 대표적인 알고리즘으로 여러개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측을 결정하게 된다(Breiman, 2001). 배깅(Bagging)은 Bootstrap Aggregating의 약자로 부트스트랩(bootstrap)을 통해 다른 훈련 데이터에 대해 훈련된 기초 분류기들을 결합(aggregating) 시키는 방식으로 주어진 훈련 데이터에서 중복을 허용하며 원 데이터셋과 같은 크기의 데이터셋을 만드는 과정을 말한다(Breiman, 2001).

[그림 12] RandomForest



자료 : https://www.mygreatlearning.com/blog/random-forest-algorithm/

최정원 등(2017)의 연구에서 RandomForest 알고리즘 방식이 부도예측에 있어서 가장 높은 정확도를 보였고 차성재·강정석(2018) 연구에서도 좋은 성과를 보였다.

AdaBoost(Adaptive boosting)는 여러개의 학습기를 순차적으로 학습-예측하면서 잘못 예측한 데이터에 가중치 부여를 통해 오류를 개선해 나가는 학습방식이며 Light GBM은 AdaBoost와 유사하나 가중치 업데이트를 경사하강법을 이용하는 것이 큰 차이이다. Boosting 분류기의 일종으로서 해당 기법은 Adaboost, Xgboost 등으로 다양한 방식으로 발전되었고 Bagging과 유사하게 초기 샘플 데이터로 다수의 분류기를 생성하지만 가장 큰 차이는 순차적인 방법이라는 것이다. 즉 이전 분류기의 학습 결과를 다음 분류기의 학습 데이터의 샘플 가중치를 조정해 학습을 진행하는 방법이다.

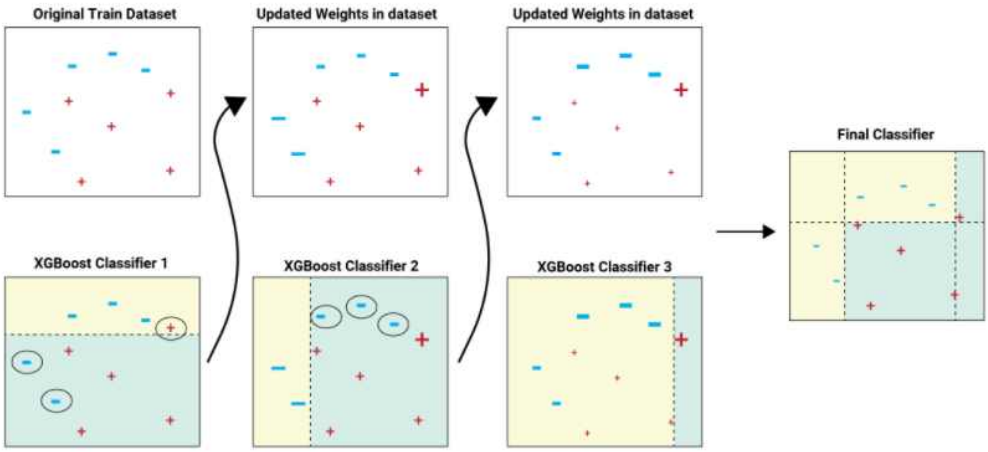
기존 알고리즘의 경우 각 개별 분류기에서 의사결정을 위한 가중치가 동일하게 부여되었다면 Adaboost에서는 분류기의 가중치가 서로 다르게 반영되는 알고리즘이다.

Light GBM은 Gradient Boosting의 프레임워크로 Tree 기반 학습 알고리즘이다. 다른 알고리즘 Tree는 수평적으로 확장되는 반면에 Light GBM은 Tree가 수직적으로 확장되며 동일한 leaf를 확장할 때 leaf-wise 알고리즘은 loss를 줄일 수 있게 된다. 큰 사이즈의 데이터를 다룰 수 있고 적은 메모리를 차지하며 GPU 학습을 지원하기 때문에 폭넓게 사용되고 있는 알고리즘이다.

XGBoost(eXtra Gradient Boost)는 트리 기반의 앙상블 학습에서 가장 각광받고 있는 알고리즘으로 GBM의 단점인 느린 수행시간 및 과적합 규제 부재 등의 문제를 해결해서 매우 각광을 받고 있으며 병렬 CPU 환경에서 병렬 학습이 가능해 기존 GBM보다 빠르게 학습을 완료할 수 있다는 장점이 있다.

XGboost는 여러개의 Decision Tree를 조합해서 사용하는 Ensemble 알고리즘이며 성능이 좋고 컴퓨팅 자원 활용률이 좋아서 근래에 많이 사용하고 있으며 최근 Kaggle 상위 랭커들이 주로 사용하면서 유명해졌다.

[그림 13] XGBoost



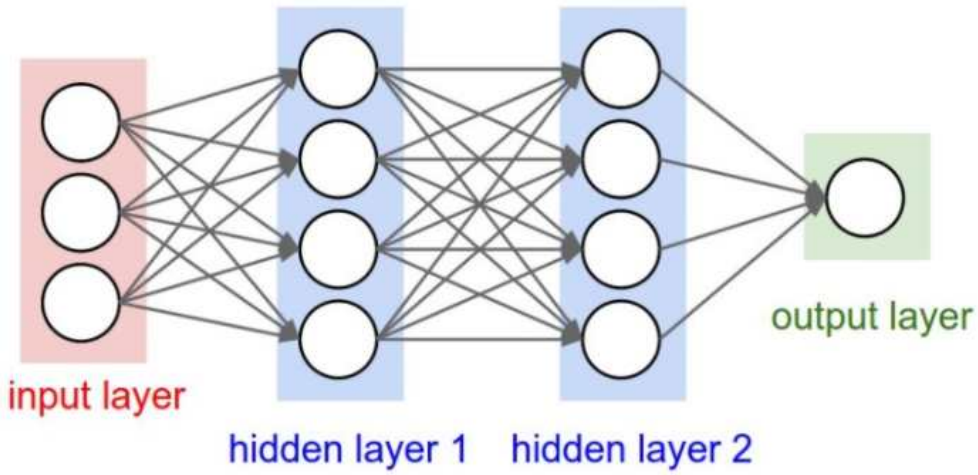
자료 : https://blog.quantinsti.com/xgboost-python/.

딥러닝 알고리즘 DFNN은 Deep FeadForward Neural Network이며 은닉층(hidden layer)을 겹겹이(deep) 쌓아 특정한 조건에서 컴퓨터가 스스로 최적의 모형을 찾는 기법이다. 흔히 Multi-layer perceptron이라고 하며 입력층(input layer), 은닉층(hidden layer), 출력층(output layer) 방향으로 가중치(weight) 값으로 연결되어 있다.

활성화 함수(Activation Function)로는 Non-Linear한 Sigmoid, Relu, Tanh 등을 사용하며 역전파(Backpropagation)를 통해 네트워크 가중치를 업데이트 하는 방식으로 학습한다. Forward Propagation을 통과하면서 실제값과

예측값의 차이인 오차를 계산하는 손실함수(Loss function)로는 cross-entropy, softmax 등의 방식이 존재한다.

[그림 14] Deep Feed Forward Network 구조



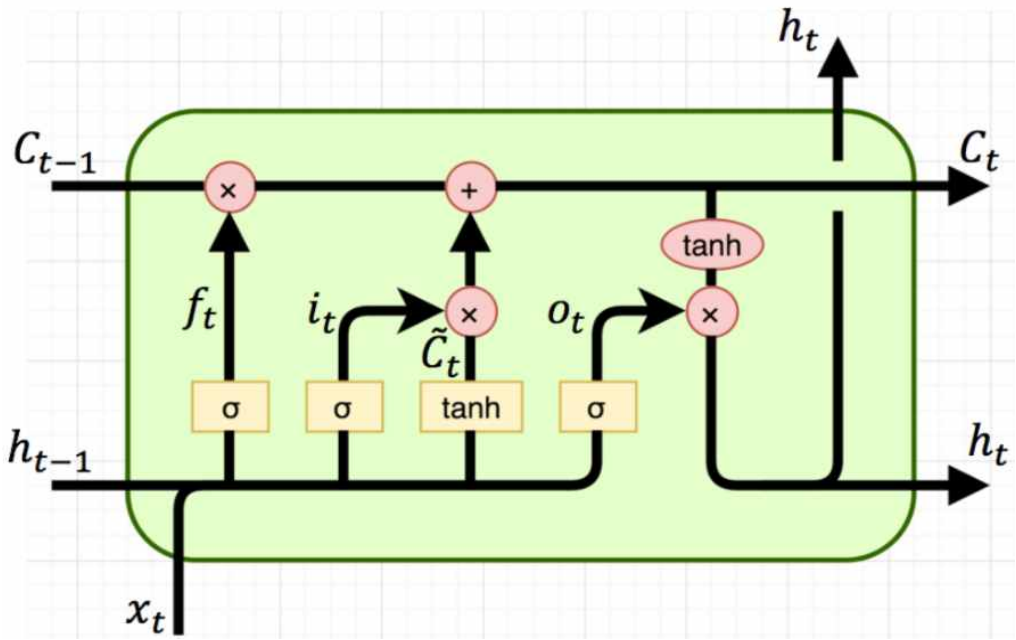
자료:https://www.pyimagesearch.com/2016/09/26/a-simple-neural-network-with-python-and-keras/

RNN(Recurrent Neural Network)은 과거의 이벤트가 미래의 결과에 영향을 줄 수 있는 순환신경망으로 시간을 많이 거슬러 올라갈수록 경사를 소실하는 장기 의존성(Long-Term Dependency)의 문제점이 있다. 이와 같은 문제점을 개선하기 위해 LSTM(Long Short-Term Memory)이 만들어졌다.

[그림 15] LSTM 구조를 먼저 살펴보면 cell state로부터 어떤 정보를 버릴 것인지를 정하는 것으로 Forget gate layer에서는 h_{t-1}과 x_t를 받아서 0과 1사이의 값을 C_{t-1}에 보내준다. Input gate layer에서는 새로운 정보중 어떤 것을 cell state에 저장할 것인지를 정하는데 C_{t-1}를 업데이트해서 새로운

cell state인 C_t를 만든다. 마지막으로 output gate layer에서는 어떤 값을 출력할지 결정하며 최종적으로 얻어진 cell state값을 반영하여 output(h_t)가 출력되는 구조이다.

[그림 15] Long Short-Term Memory 구조



자료 : http://docs.likejazz.com/lstm/

[표 3] 예측 모델링 방법론 요약

분류		방법론	특징
머신 러닝	개별모형	Decision Tree	데이터에 있는 규칙을 학습하여 트리(tree)기반의 분류 규칙 생성
		Logistic Regression	독립변수와 종속변수의 선형 관계성에 기반(전통적 이진분류 모형)

		KNN	근접 거리를 기준으로 하는 알고리즘
		SVM	개별 클래스 간의 최대 분류 마진을 효과적으로 찾아주는 방법론
	Bagging	RandomForest	결정트리기반으로 여러 개의 분류기를 만들어서 보팅하는 배깅알고리즘
	Boosting	Adaboost	오류 데이터에 가중치를 부여하면서 학습하는 부스팅알고리즘
		Light GBM	리프 중심 트리분할 방식(Leaf Wise)을 사용하는 부스팅알고리즘
		XGBoost	GridSearch방식의 하이퍼 파라미터 튜닝을 수행하는 부스팅알고리즘
딥러닝	DFNN	DFNN	입력층, 은닉층, 출력층 방향으로 진행하며 역전파 방식으로 가중치를 업데이트하고 학습함
	LSTM	LSTM	Forget, Input, Ouput 3개의 게이트를 사용하여 Cell State를 업데이트하여 장단기 기억을 보존함

3.2.4 예측 모델링의 성능 평가

머신러닝의 예측 또는 분류 모델에서 알고리즘의 성능을 평가하기 위한 방법으로 [표 4] 와 같이 혼동행렬(Confusion Matrix)를 일반적으로 사용한다.

[표 4] 혼동행렬(Confusion Matrix)

구 분		예측	
		T(부도=1)	F(정상=0)
실제	T(부도=1)	TP(True Positive)	FN(False Negative)
	F(정상=0)	FP(False Positive)	TN(True Negative)

정확도(Accuracy)는 식으로 표현하면 TP + TN / TP + FN + FP + TN 이렇게 되며 모델이 입력된 데이터에 대해 얼마나 정확하게 예측하는지 나타낸다. 민감도(Sensitivity)는 TP / TP + FN으로 표현하며 실제 T값중에 모델이 예측한 실제 T값의 비율을 나타내는 지표로 FP보다는 FN을 줄이는 것이 중요한 경우 사용하는 지표이다. 본 연구에 적용하면 실제 부도 기업을 부도 기업으로 분류하는 비율이라고 할 수 있겠다. 그리고 특이도(Specificity)는 TN / TN + FP로 표현하며 실제 F값중에 모델이 예측한 실제 F값의 비율을 나타내는 지표이다. 본 연구에 적용하면 실제 정상 기업을 정상 기업으로 분류하는 비율이라고 할 수 있겠다.

마지막으로 ROC Curve(Receiver Operating Characteristic Curve)는 민감도(Sensitivity) / 특이도(Specificity) 도표이며 완벽한 분류기는 민감도 = 1 / 특이도 = 0인 직선이며 이에 가까울수록 성능이 좋다고 할 수

있다. ROC Curve의 아래 면적을 통계에 의해 측정하는 AUC(Area Under Curve)도 있다.

본 연구에서 수집된 부도 기업의 표본은 265개, 정상 기업의 표본은 2,071개인데 SMOTE 기법을 통해 데이터의 불균형 문제를 해결하였기 때문에 클래스의 분포가 동일하다고 판단하여 정확도(Accuracy)를 일반적인 지표로 사용하였다.

하지만 본 연구의 목적에 부합하기 위해 부도기업을 부도로 예측하는 것이 정상 기업을 정상 기업으로 예측하는 것보다 중요하다고 생각하여 정확도보다는 민감도(Sensitivity)를 우선하여 평가 기준으로 선정하였다.

제 4 장 연구 결과

4.1 재무 변수 유의성 검증 결과

<연구문제 1>의 재무 변수의 유의성을 통계적으로 검증하기 위해 1단계로 독립표본 t-test를 먼저 수행하고 2단계로 Logistic Regression 검증을 수행하였다. 먼저 정상기업과 부도기업 집단간의 변수들의 평균 차이를 검증하기 위해 수행한 t-test는 두 집단 모두 독립적인 개별 데이터를 가지고 있고 각 집단의 개별적 표준오차를 구할 수 있기 때문에 t-통계량을 산출할 수 있다.

본 연구에서는 부도기업의 경우 부도 직전 1년전, 2년전, 3년전의 재무 데이터와 정상기업의 경우 2019년말, 2018년말, 2017년말 기준의 재무 데이터를 가지고 t-test를 실시하였다.

먼저 재무 데이터의 1년전 자료를 바탕으로 재무 변수 검증을 [표 5]와 같이 진행하였다. 1단계 t-test 결과를 만족하면서 2단계 Logistic Regression 결과를 만족하는 변수를 확인한 결과 alt001;총자산 이익잉여금률, bea001;총자산 이익률 , hor003;매출액 운전자본 비율, bok002;차입금 의존도, bok011;투자안정성 비율 5개 변수가 통계적으로 유의미한 결과로 도출되었다.

[표 5] 1년전 재무 데이터의 t-test 및 Logistic Regression 결과

변수명	구분	N	t-test 결과		logit 결과	
			t	유의확률	Exp(B)	유의확률
alt001	부도	265	**-11.555*** **	**0.000**	**0.483*** **	**0.000**
	정상	2,071				
alt002	부도	265	-3.471*	0.001	0.616	0.348
	정상	2,071				
alt003	부도	265	-1.459	0.145	0.815	0.065
	정상	2,071				
bea001	부도	265	**-6.922*** **	**0.000**	**0.194*** **	**0.000**
	정상	2,071				
bea002	부도	265	8.029***	0.000	0.541	0.364
	정상	2,071				
bea003	부도	265	-7.224***	0.000	1.304	0.104
	정상	2,071				
bea004	부도	265	-1.047	0.296	1.277*	0.035*
	정상	2,071				
hor001	부도	265	-2.002*	0.046	1.071	0.465
	정상	2,071				
hor002	부도	265	-1.064	0.287	0.809	0.310
	정상	2,071				
hor003	부도	265	**-2.015***	**0.044**	**0.624***	**0.031**
	정상	2,071				
hor004	부도	265	1.206	0.229	1.044	0.745
	정상	2,071				
bok001	부도	265	-8.092***	0.000	0.330	0.099
	정상	2,071				

bok002	부도	265	**9.962*****	**0.000**	**1.550****	**0.005**
	정상	2,071				
bok003	부도	265	1.290	0.198	0.962	0.721
	정상	2,071				
bok004	부도	265	0.273	0.785	1.080	0.145
	정상	2,071				
bok005	부도	265	1.872	0.062	1.100	0.348
	정상	2,071				
bok006	부도	265	-0.331	0.741	0.000	0.541
	정상	2,071				
bok007	부도	265	-0.918	0.359	0.829**	0.004
	정상	2,071				
bok008	부도	265	-0.285	0.776	0.990	0.943
	정상	2,071				
bok009	부도	265	-1.842	0.066	1.091	0.426
	정상	2,071				
bok010	부도	265	-0.311	0.756	10212.394	0.541
	정상	2,071				
bok011	부도	265	**-1.980***	**0.049**	**0.854***	**0.021**
	정상	2,071				

t>=1.96, p<0.06 *, t>=2.58 p<0.01**, t>=3.30, p<0.001***

그 다음으로 재무 데이터의 2년전 자료를 바탕으로 재무 변수 검증을 [표 6]과 같이 진행하였다. 1단계 t-test 결과를 만족하면서 2단계 Logistic Regression 결과를 만족하는 변수를 확인한 결과 alt001;총자산 이익잉여금률, alt002;총자산 영업이익률 , alt003;총자산 매출액 비율, hor004;자본 매출액 비율, 4개 변수가 통계적으로 유의미한 결과로 도출되었다.

[표 6] 2년전 재무 데이터의 t-test 및 Logistic Regression 결과

변수명	구분	N	t-test 결과		logit 결과	
			t	유의확률	Exp(B)	유의확률
alt001	부도	265	**-9.518*** **	**0.000**	**0.380*** **	**0.000**
	정상	2,071				
alt002	부도	265	**-5.766*** **	**0.000**	**0.214*** **	**0.000**
	정상	2,071				
alt003	부도	265	**-4.387*** **	**0.000**	**0.762* **	**0.021**
	정상	2,071				
bea001	부도	265	-7.778***	0.000	1.037	0.808
	정상	2,071				
bea002	부도	265	5.415***	0.000	0.999	0.999
	정상	2,071				
bea003	부도	265	-5.099***	0.000	1.068	0.665
	정상	2,071				
bea004	부도	265	-1.884	0.060	0.781	0.531
	정상	2,071				
hor001	부도	265	0.345	0.730	21.493***	0.000
	정상	2,071				
hor002	부도	265	-1.176	0.240	0.045	0.089
	정상	2,071				
hor003	부도	265	-0.385	0.701	0.004	0.551
	정상	2,071				
hor004	부도	265	**2.668** **	**0.008**	**1.276* **	**0.046**
	정상	2,071				
bok001	부도	265	-5.409***	0.000	1.215	0.918
	정상	2,071				

bok002	부도	265	8.497***	0.000	1.152	0.294
	정상	2,071				
bok003	부도	265	4.443***	0.000	1.034	0.770
	정상	2,071				
bok004	부도	265	2.091*	0.037	1.068	0.157
	정상	2,071				
bok005	부도	265	-0.026	0.979	14.480**	0.001
	정상	2,071				
bok006	부도	265	-0.485	0.628	5.973	0.128
	정상	2,071				
bok007	부도	265	0.164	0.870	1.173*	0.011
	정상	2,071				
bok008	부도	265	-0.797	0.426	0.414	0.167
	정상	2,071				
bok009	부도	265	-0.843	0.399	14.072	0.900
	정상	2,071				
bok010	부도	265	-0.851	0.395	0.009	0.822
	정상	2,071				
bok011	부도	265	-2.666**	0.008	0.912	0.247
	정상	2,071				

t>=1.96, p<0.06 *, t>=2.58 p<0.01**, t>=3.30, p<0.001***

그 다음으로 재무 데이터의 3년전 자료를 바탕으로 재무 변수 검증을 [표 7]과 같이 진행하였다. 1단계 t-test 결과를 만족하면서 2단계 Logistic Regression 결과를 만족하는 변수를 확인한 결과 bok002;차입금 의존도 1개 변수가 통계적으로 유의미한 결과로 도출되었다.

[표 7] 3년전 재무 데이터의 t-test 및 Logistic Regression 결과

변수명	구분	N	t-test 결과		logit 결과	
			t	유의확률	Exp(B)	유의확률
alt001	부도	265	0.260	0.795	0.380***	0.000
	정상	2,071				
alt002	부도	265	0.364	0.716	0.214***	0.000
	정상	2,071				
alt003	부도	265	0.511	0.610	0.762*	0.021
	정상	2,071				
bea001	부도	265	1.372	0.171	1.037	0.808
	정상	2,071				
bea002	부도	265	-0.503	0.615	0.999	0.999
	정상	2,071				
bea003	부도	265	-0.915	0.361	1.068	0.665
	정상	2,071				
bea004	부도	265	-0.804	0.421	0.781	0.531
	정상	2,071				
hor001	부도	265	-0.888	0.375	21.493***	0.000
	정상	2,071				
hor002	부도	265	-1.370	0.171	0.045	0.089
	정상	2,071				
hor003	부도	265	-0.192	0.847	0.004	0.551
	정상	2,071				
hor004	부도	265	1.040	0.299	1.276*	0.046
	정상	2,071				
bok001	부도	265	-7.426***	0.000	1.215	0.918
	정상	2,071				

bok002	부도	265	**8.287*** **	**0.000**	**1.659*** **	**0.000**
	정상	2,071				
bok003	부도	265	2.172*	0.031	1.034	0.770
	정상	2,071				
bok004	부도	265	0.367	0.714	1.068	0.157
	정상	2,071				
bok005	부도	265	0.178	0.859	14.480**	0.001
	정상	2,071				
bok006	부도	265	-0.449	0.654	5.973	0.128
	정상	2,071				
bok007	부도	265	-0.479	0.632	1.173*	0.011
	정상	2,071				
bok008	부도	265	0.271	0.786	0.414	0.167
	정상	2,071				
bok009	부도	265	-0.815	0.415	14.072	0.900
	정상	2,071				
bok010	부도	265	-0.931	0.352	0.009	0.822
	정상	2,071				
bok011	부도	265	-1.450	0.147	0.912	0.247
	정상	2,071				

t>=1.96, p<0.06 *, t>=2.58 p<0.01**, t>=3.30, p<0.001***

마지막으로 위에서 도출한 재무 1년전, 2년전, 3년전 데이터의 유용한 변수들을 통합하여 t-test와 Logistic Regression을 다시 수행하였다. t-test 결과는 모두 유의미한 결과가 도출되었지만 Logistic Regression 수행시 t2_alt002(2년전), t2_alt003(2년전) 변수가 유의미한 결과가 도출되지 않아 2개의 변수는 제거하였고 최종적으로 변수 검증한 결과는 [표 8] 과 같다.

[표 8] 최종 변수 검증 결과(재무변수 연도별 통합)

변수명	구분	N	t-test 결과		logit 결과	
			t	유의확률	Exp(B)	유의확률
t1_ alt001	부도	265	-11.555***	0.000	1.964**	0.006
	정상	2,071				
t1_ bea001	부도	265	-6.922***	0.000	0.057***	0.000
	정상	2,071				
t1_ hor003	부도	265	-2.014*	0.044	0.660*	0.023
	정상	2,071				
t1_ bok002	부도	265	9.958***	0.000	1.550***	0.000
	정상	2,071				
t1_ bok011	부도	265	-1.980*	0.049	0.866**	0.003
	정상	2,071				
t2_ alt001	부도	265	-9.518***	0.000	0.277***	0.000
	정상	2,071				
t2_ hor004	부도	265	2.668**	0.008	1.213**	0.006
	정상	2,071				
t3_ bok002	부도	265	8.287***	0.000	1.211*	0.017
	정상	2,071				

t>=1.96, p<0.06 *, t>=2.58 p<0.01**, t>=3.30, p<0.001***, t1: 1년전, t2: 2년전, t3: 3년전

4.2 뉴스 콘텐츠 감성 분석 결과

<연구문제 3>의 결과를 얻기 위해 먼저 정제 완료된 정상 및 부도 기업의 뉴스 기사 239,156건으로 형태소 분석을 실시하여 명사를 추출하였다. 그리고 단어를 10,000개로 설정하여 벡터화를 진행하였고 단어 벡터를 더하여 해당 단어가 전체 문장에서 몇 번 등장하는지 빈도수를 구하고 TF-IDF 가중치를 계산하였다. TF-IDF값 기준으로 상위 15개의 단어 목록을 아래 [표 9]에 나열하였다.

[표 9] TF-IDF값 기준 상위 단어 목록(15개)

순위	정상 기사			부도 기사		
	단어	TF-IDF	빈도	단어	TF-IDF	빈도
1	공시	9,189	56,413	폐지	3,614	56,068
2	기술	8,743	144,984	감사	2,914	53,997
3	투자	8,364	138,076	보고서	1,970	27,027
4	서비스	6,723	91,582	공시	1,817	18,698
5	매출	6,715	89,385	의견	1,804	29,180
6	제공	6,483	98,279	결정	1,470	17,713
7	증가	6,116	74,294	투자	1,424	27,548
8	실적	5,855	74,826	관리	1,186	19,028
9	지원	5,677	77,603	신청	1,116	13,884
10	성장	5,324	78,476	발생	1,094	14,852
11	지분	5,013	56,268	자본	1,027	15,337
12	매출액	5,010	39,480	인수	923	12,526
13	생산	4,966	67,603	지정	838	10,797
14	출시	4,574	54,599	절차	833	10,052
15	계획	4,567	70,328	하락	818	7,518

단어의 감성사전은 정상과 부도 기업 뉴스 전체에서 사용된 빈도에서 정상 기업 뉴스에서 사용된 빈도와 부도 기업 뉴스에서 사용된 빈도의 차이를 나누어 각각 계산하였다. 감성 점수는 -1부터 1사이 값으로 나타나며 양수값은 긍정 단어, 음수값은 부정 단어라고 할 수 있겠다. TF-IDF값 기준으로 의미있는 단어를 총 321개를 선정하여 말뭉치 기반 감성사전을 아래 [표 10] 과 같이 구축하였다. 이렇게 구축된 감성 사전을 바탕으로 뉴스별 감성점수의 평균값을 계산하고 기업별 평균점수를 계산한 감성점수를 최종적인 변수값으로 활용하였다.

[표 10] 말뭉치 기반 감성사전 구축 예시

순위	긍정 단어 목록		부정 단어 목록	
	단어	감성 점수	단어	감성 점수
1	상업화	0.985	워크아웃	-0.785
2	증원	0.975	폐지	-0.714
3	격려금	0.962	동전주	-0.686
4	수출국	0.950	깡통	-0.661
5	혁신	0.948	후순위채	-0.596
6	효능	0.948	감사	-0.570
7	앞장	0.943	탕감	-0.523
8	청신호	0.930	보증채무	-0.516
9	가속도	0.920	경영권	-0.496
10	기술	0.918	어음	-0.480
11	생산	0.916	미지급금	-0.456
12	마중물	0.915	망연자실	-0.450
13	성장	0.914	보고서	-0.338
14	출시	0.911	사채권자	-0.263
15	서비스	0.910	미납	-0.229

4.3 예측 모델링 적용 결과

<연구문제2>를 해결하기 위한 예측 모델링은 일반적으로 많이 사용하고 있는 머신러닝과 딥러닝 알고리즘을 각각 적용하였다. 머신러닝은 classification 분류 기법에 대부분 사용하는 DT(Decision Tree), LR(Logistic Regression), KNN(Kneighbors), SVM(Support Vector Machine), RF(RandomForest), ADA(Adaboost), LGBM(Light GBM), XGB(XGBoost) 총 8가지 방법론을 적용하였고 딥러닝은 DFNN(Deep Feed Forward Network), LSTM(Long Short-Term Memory) 총 2가지 방법론을 적용하였다.

재무 변수는 먼저 t-test와 Logistic Regression 수행을 통해 추출된 유의미한 재무 비율만을 대상으로 3개년을 통합한 데이터로 모델링을 적용하였다.

그리고 3개년을 통합한 재무 데이터와 함께 감성 분석한 결과 값을 추가하여 모델링을 적용하였으며 뉴스 수집 기간에 따른 결과를 확인하기 위하여 부도 직전(정상 기업의 경우 19년말 기준) 1개월전(+1M), 1+2개월전(+2M), 1+2+3개월전(+3M), 1+2+3...+6개월전(+6M) 등 총 6개의 감성 분석 결과 값을 사용하였다.

재무 및 감성분석 결과를 반영한 총 7개의 데이터셋을 구성하여 위에서 설명한 10가지 방법론을 적용하여 민감도, 정확도값을 산출하였다.

머신러닝 방법론의 경우 기본적으로 제공되는 파라미터값으로 설정하였고 DFNN은 hidden layer 10개, 32개의 뉴런, activation function은 relu, Optimizer는 rmsprop을 사용하였고 epoch 20, batchsize 10으로 설정하였다. 그리고 LSTM의 경우 16 메모리 셀을 가진 LSTM layer 5개로 구성하고 나머지 조건은 DFNN과 동일하게 설정하였다.

다음 [표 11]의 예측 모델링 민감도 결과를 확인하면 머신러닝의 경우 재무 데이터만 적용했을 때 LR과 SVM이 87.50%로 성능이 가장 좋았으며 재무 데이터에 뉴스 1개월전의 감성분석 결과를 적용하면 SVM과 RF가 92.50%로 성능이 가장 좋았다. 그리고 여기에 뉴스 2개월전의 감성분석 결과를 적용하면 RF가 93.75%로, 3,4개월전의 감성분석 결과를 적용하면 SVM이 각 93.75%로, 5개월전의 감성분석 결과를 적용하면 SVM이 92.5%, 6개월전의 감성분석 결과를 적용하면 SVM과 ADA가 92.50%로 성능이 가장 좋았다.

평균적으로 SVM이 가장 좋은 성능을 보였고 뉴스 수집 기간에 따른 민감도를 확인하면 평균적으로 부도직전 4개월의 뉴스 감성점수가 가장 효과적이었다고 볼 수 있겠다.

그리고 뉴스 데이터의 감성분석을 적용하여 예측한 민감도 결과값이 재무 데이터만을 사용하여 예측한 결과값보다 성능이 다소 높은 것을 알 수 있고 가장 성능이 좋은 SVM을 기준으로 본다면 평균적으로 약 5% 정도의 성능 향상을 확인할 수 있었다.

[표 11] 예측 모델링 민감도(Sensitivity) 결과

(단위 : %)

방법론	재무	부도직전 월별 감성분석 민감도 결과						평균
		+1M	+2M	+3M	+4M	+5M	+6M	
DT	73.75	82.50	85.00	85.00	85.00	82.50	85.00	82.68
LR	**87.50**	90.00	91.25	91.25	91.25	90.00	91.25	90.36
KNN	75.00	83.75	86.25	82.50	82.50	82.50	80.00	81.79
SVM	**87.50**	**92.50**	91.25	**93.75**	**93.75**	**92.50**	**92.50**	**91.96**
RF	81.25	**92.50**	**93.75**	91.25	90.00	88.75	90.00	89.64
ADA	83.75	90.00	92.50	90.00	88.75	88.75	**92.50**	89.46
LGBM	85.00	88.75	87.50	87.50	88.75	87.50	88.75	87.68
XGB	56.25	87.50	87.50	86.25	83.75	82.50	82.50	80.89
DFNN	79.43	82.53	85.83	87.31	89.47	83.41	82.53	84.36
LSTM	84.17	**92.50**	93.41	85.30	93.41	82.27	85.91	88.14
평균	79.36	88.25	89.42	88.01	**88.67**	86.01	87.09	

다음 [표 12]는 예측 모델링의 정확도 결과를 나타내었다. 머신러닝의 경우 재무 데이터만 적용했을 때 LGBM이 93.72%로 성능이 가장 좋았으며 재무 데이터에 뉴스 1개월전의 감성분석 결과를 적용하면 RF와 LGBM이 97.86%로 성능이 가장 좋았다. 그리고 뉴스 2개월전 ~ 6개월전까지의 감성분석 결과를 적용하면 XGB가 가장 좋은 성능을 나타내었고 2개월전 98%, 3개월전 97.86%, 4개월전 97.29%, 5개월전 96.29%, 6개월전 96.15%의 정확도를 보였다.

[표 12] 예측 모델링 정확도(Accuracy) 결과

(단위 : %)

방법론	재무	부도직전 월별 감성분석 정확도 결과						평균
		+1M	+2M	+3M	+4M	+5M	+6M	
DT	88.45	95.44	95.29	94.72	93.72	93.72	93.30	93.52
LR	89.16	95.86	95.01	94.58	93.72	92.87	93.15	93.48
KNN	88.30	93.01	93.30	92.01	91.73	91.58	91.44	91.62
SVM	89.02	95.01	95.01	94.01	93.72	92.72	93.15	93.23
RF	91.01	**97.86**	97.00	96.58	96.01	94.58	95.15	95.46
ADA	92.58	97.72	97.15	96.43	95.15	94.44	95.58	95.58
LGBM	**93.72**	**97.86**	96.86	96.29	96.15	94.72	95.29	95.84
XGB	89.87	97.66	**98.00**	**97.86**	**97.29**	**96.29**	**96.15**	**96.16**
DFNN	89.44	96.29	95.58	94.58	93.72	93.44	96.29	94.19
LSTM	84.59	76.46	72.75	84.31	75.89	85.02	77.46	79.50
평균	89.61	94.32	93.60	**94.14**	92.71	92.94	92.70	

제 5 장 결론

5.1 연구 결과 요약

본 연구는 부도 예측에 유의미한 변수를 확인하고 말뭉치 기반의 감성사전을 구축하여 뉴스의 수집 기간별 감성점수가 부도 예측에 얼마나 효과적인지 다양한 머신러닝 및 딥러닝 모델링 기법을 통해 연구를 진행하였다. 먼저 유의미한 재무 변수를 확인하기 위하여 3가지의 재무모델(Altman, 1968; Beaver, 1968; Horrigan, 1966) 및 한국은행의 기업경영분석지표에서 변수를 추출하였고 t-test와 Logistic Regression을 순차적으로 수행하였다.

각 연도별로 유의미한 변수를 추출한 결과 1년전의 총자산 이익잉여금률(t1_alt001), 1년전의 총자산 이익률(t1_bea001), 1년전의 매출액 운전자본 비율(t1_hor003), 1년전의 차입금 의존도(t1_bok002), 1년전의 투자 안정성 비율(t1_bok011), 2년전의 총자산 이익잉여금률(t2_alt001), 2년전의 자본 매출액 비율(t2_hor004), 3년전의 차입금 의존도(t3_bok002) 총 8개의 변수가 유의미한 것으로 나타났다.

재무 변수 기준으로는 총자산 이익잉여금률(alt001), 총자산 이익률(bea001), 매출액 운전자본 비율(hor003), 자본 매출액 비율(hor004), 차입금 의존도(bok002), 투자 안정성 비율(bok011) 총 5개의 변수가 유의미한 것으로 나타났다.

기업의 건전성을 파악하기 위해서는 매출액과 순이익을 잘 내는 회사인지, 총 자본 대비 차입을 통하여 조달한 자금의 비중이 적절한지(차입금 의존도가 낮은지), 그리고 외부 자금에 의존하지 않고 순수 영업활동에 의한 현금흐름으로 유형자산투자를 충분히 실행할 수 있는지를 잘 확인해야 할 것이다.

다음으로 뉴스의 감성점수를 사용하여 부도 예측에 효과적인 알고리즘을 확인하였다. 머신러닝과 딥러닝 등 총 10가지의 알고리즘을 적용하였고 실제 부도 기업을 부도 기업으로 예측하는 민감도 기준으로 SVM(Support Vector Machine)이 가장 우수한 예측력을 보였다. 재무 데이터만을 사용했을 때보다 뉴스 감성분석 정보까지 추가로 반영된 데이터셋을 활용하면 부도예측 성능이 평균적으로 약 5% 정도 향상되는 효과가 있었음을 실증하였다.

[그림 16]은 민감도 및 정확도에서 가장 성능이 우수한 SVM, XGBoost를 대상으로 데이터Set의 민감도 및 정확도 그래프를 나타낸 것이다. 재무 데이터만을 이용했을 때와 재무데이터에 뉴스 감성정보를 1개월~6개월 각각 추가하였을 경우를 아래 그래프로 확인할 수 있다. 정확도가 우수한 XGboost는 뉴스 수집 기간이 2개월일 때, 민감도가 좋은 SVM의 경우 뉴스 수집기간이 3,4개월일 때 부도예측에 가장 효과적인 것으로 나타났다. 김찬송(2018)에서도 민감도 기준으로 본 연구와 동일한 결과가 나타났다.

[그림 16] 데이터Set의 민감도 및 정확도 그래프

정확도 측면에서는 GridSearch방식의 하이퍼 파라미터 튜닝을 수행하는 부스팅알고리즘인 XGBoost 기법이 가장 우수한 성능을 가진 것으로 나타났다. XGBoost는 유명한 캐글 경연대회(Kaggle Contest)에서 상위를 차지한 많은 데이터 과학자가 이 기법을 이용하면서 널리 알려졌고 분류에 있어서는 일반적으로 다른 머신러닝보다 뛰어난 예측 성능을 나타낸다고 하는데 본 연구에서도 정확도 측면에서는 해당 모형의 우수성에 대해서 확인할 수 있었다.

5.2 연구 시사점

5.2.1 학술적 시사점

본 연구에서의 학술적 시사점은 기존에 회계 및 재무분야의 문헌에서 잘 알려진 3개의 재무모델 변수(Altman, 1968; Beaver, 1968; Horrigan, 1966)를 검증함과 동시에 한국은행에서 1962년부터 매년 작성해오고 있는 지표인 기업경영분석 지표의 재무변수를 같이 사용하여 학술적인 관점에서 검증을 분석해 보았다는 점이다. 검증을 위해 t-test와 Logistic Regression 2단계를 진행하였고 1,2,3개년의 재무 데이터를 통합하여 연도별로 어떤 재무 변수가 유의미한지 검증을 진행하였다. 이처럼 연도별로 유용한 재무 변수를 통계적인 방법으로 2단계 검증 후 통합하여 실증한 사례가 없었기 때문에 이것은 기존 연구와의 차별화된 부분이라고 할 수 있겠다. 그리고 앞서 선행연구에서 살펴본 것처럼 재무 변수 선정시 단순히 KIS-VALUE(나이스 신용평가)에서 제공하는 기본적인 재무 데이터를 사용하거나, 기존에 사용되어왔던 변수를 선택하여 정확도에 초점을 맞춘 연구들이 많았다. 하지만 본 연구에서는 전통적으로 알려진 재무모델 변수와 국내 금융기관중 가장 영향력이 있는 한국은행에서 사용하고 있는 변수를 통합하여 사용했기에 재무 변수 선정에 대한 체계적인 근거를 확보하였다.

김찬송(2018)에서는 t-test를 통해 Altman 모델의 변수가 모두 유의함을 확인하였지만 본 연구에서는 t-test뿐만 아니라 Logistic Regression을 통해 한번더 검증을 진행하였고 2년전 재무 데이터에서만 모두 유의함을 확인하였다. 이를 통해 전통적으로 검증된 모델에서 사용된 변수라고 하더라도 모든

변수가 통계적으로 유의미하지는 않았음을 실증하였고 한국은행의 기업경영분석지표의 변수도 모든 변수가 유의미하지는 않았다.

전통적으로 사용되어 온 3가지 재무 모델과 현재에도 계속 사용되어지고 있는 한국은행의 변수를 통합하고 3개년치의 재무 데이터를 활용하여 유의미한 변수를 실증한 부분에 대해서는 학술적 의의가 있다고 볼 수 있겠다.

5.2.2 실무적 시사점

전통적으로 금융기관에서는 부실차주를 식별하기 위해 주로 재무비율 등 정량적인 데이터를 바탕으로 조기경보모형을 설정하고 관리하여 왔다. 하지만 재무비율 데이터는 정보가 분기별로 업데이트 되기 때문에 정보 제공 주기에 대한 적시성에 문제가 있을 수 있다.

기업여신을 담당하는 금융기관에서는 본 연구에서 실증한 온라인 뉴스 콘텐츠의 감성분석 정보인 비정형 데이터를, 특히 부도 직전 3,4개월치의 뉴스 정보를 함께 부도예측에 사용한다면 효과적인 여신의사 결정 지원 체계의 기반을 마련하는데 많은 도움이 될 것이라고 판단한다.

매일 생산되는 기업과 관련된 뉴스 데이터를 분석하여 기업별 뉴스별 감성점수를 부여하고 긍/부정 여부를 판단하는 시그널을 시각화된 화면을 통해 제공한다면 즉각적인 이상징후에 대해 담당자가 실시간으로 파악할 수 있게 되며 의사결정시 보다 나은 도움을 줄 수 있을 것이다.

그리고 금융기관 뿐만 아니라 주식, 채권 등 기업에 투자하는 개인도 본 연구에서 실증한 기업 뉴스 콘텐츠의 감성 분석 정보를 충분히 활용할 수 있을 것이다. 본 연구에서 도출한 특정 연도의 유의미한 재무 비율을 바탕으로 실시간으로 쏟아지는 기업 뉴스의 감성 정보를 수치로 계산하고 SVM 알고리즘으로 부도 시그널 정보를 매일 확인한다면 개인의 소중한 자산을 지키는데 많은 도움이 되리라 판단된다.

결론적으로 재무 데이터의 경우 총자산 이익잉여금률, 총자산 이익률, 매출액 운전자본 비율, 자본 매출액 비율, 차입금 의존도 재무비율을 중점적으로 모니터링 하고 실제 부도기업을 부도기업으로 잘 예측하는 SVM 알고리즘을 사용하면 좀 더 예측력을 높일 수 있을 것이다.

5.3 연구 한계 및 향후 연구 방향

본 연구에서는 부도 기업보다 정상 기업의 표본수가 훨씬 많이 존재하는 데이터 불균형 문제가 있었다. 물론 SMOTE 방식을 통해 부도 기업의 표본을 정상기업의 표본과 1:1로 데이터를 복원 추출하는 방식을 사용하였지만 조금 더 많은 수의 부도 기업 표본을 추출하는데 한계점이 있었다. 부도 기업의 표본수를 정상기업의 표본에 맞추는 오버 샘플링 기법을 적용하였는데 정상 기업의 표본수를 부도 기업의 표본수에 맞추는 언더 샘플링 기법을 적용하여도 비슷한 결과가 나오는지는 후속 연구로 진행하면 좋을 것 같다. 그리고 뉴스 콘텐츠 수집을 위한 크롤링시 HTML 구조가 표준화된 네이버 뉴스만을 대상으로 한정하여 감성분석을 진행하였기에 감성사전 구축시 뉴스가 상대적으로 많은 기업들의 키워드에 영향을 받을 수 밖에 없었고

규모가 작은 중소기업의 경우 뉴스 노출이 상대적으로 적다 보니 뉴스 콘텐츠를 분석하기에 충분하지 않았다.

향후 연구에서는 이러한 한계점을 극복하기 위해 뉴스 기사 크롤링시 네이버 뿐만 아니라 다음, 구글 등 다른 포털 뉴스를 함께 사용하면 좋겠다. HTML 구조가 표준화된 대형 포털사 뿐만 아니라 HTML 구조가 제각각인 개별 인터넷 언론사 뉴스 사이트에서도 뉴스 데이터를 수집한다면 규모가 작은 중소기업의 데이터셋 크기를 보완할 수 있을 것이다. 온라인 뉴스 콘텐츠 뿐만 아니라 금감원 DART의 기업공시 정보, 기업 뉴스 관련 카페 게시글 및 댓글 등의 추가적인 데이터 정보 원천을 획득한다면 말뭉치 기반의 감성사전을 풍부하게 구축할 수 있을 것이다.

추가적으로 후속연구에서 감성사전 구축 방법론(지도 학습 및 비지도 학습)에 따라 부도 예측 결과는 어떻게 달라지는지, 비정형 데이터의 유형(기업 뉴스, 기업 관련 커뮤니티 게시글 및 댓글 등)에 따라 부도 예측 결과는 어떻게 달라지는지, 형태소 분석시 명사, 형용사, 동사 등 활용가능한 모든 형태소를 도출하여 어느 것이 부도 예측에 가장 효과적인지도 연구해보면 흥미로울 것 같다.

참 고 문 헌

Alaminos, D., del Castillo, A., & Fernández, M. Á. (2016). A global model for bankruptcy prediction. *Plos one*, 11(11), e0166693.

Antunes, F., Ribeiro, B., & Pereira, F. (2017). Probabilistic modeling and visualization for bankruptcy prediction. *Applied soft computing*, 60, 831-843.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert systems with applications*, 83, 405-417.

Beaver, W. H. (1968). Alternative accounting measures as predictors of failure. *The accounting review*, 43(1), 113-122.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

David, C. R. (1972). Regression models and life tables (with discussion). *Journal of the royal statistical society*, 34(2), 187-220.

Gupta, A., Simaan, M., & Zaki, M. J. (2016, December). Investigating bank failures using text mining. *In 2016 IEEE Symposium series on computational intelligence* (SSCI) (pp. 1-8). IEEE.

Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios. *Journal of accounting research*, 44-62.

Jo, N. O., & Shin, K. S. (2016). Bankruptcy prediction modeling using qualitative information based on big data analytics. 지능정보연구, 22(2), 33-56.

Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. *In Annual meeting of the society for academic emergency medicine in San Francisco*, California (Vol. 14).

Nam, C. W., Kim, T. S., Park, N. J., & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, 27(6), 493-506.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

Ratios, Altman El Financial. (1968). Discriminant analysis and the prediction of corporate bankruptcy. *Journal of Financt*, 589-609.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, 74(1), 101-124.

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert systems with applications*, 41(5), 2353-2361.

금융감독원장 (2020.12.07.). 기업부문 취약성:진단과 과제 심포지엄

김명종 (2009). 기업부실 예측 데이터의 불균형 문제 해결을 위한 앙상블 학습. 한국경영학회 통합학술발표논문집, 1-17.

김찬송 (2018). 부도예측 모형에서 효과적인 감성분석을 위한 뉴스 분류 방법에 관한 연구. 석사학위논문, 한양대학교.

김형준 · 류두진 · 조훈 (2019). 기업부도예측과 기계학습. 금융공학연구, 18(3), 131-152.

민성환 (2014). 개선된 배깅 앙상블을 활용한 기업부도예측. 지능정보연구, 20(4), 121-139.

민성환 (2016). 부도예측을 위한 KNN 앙상블 모형의 동시 최적화. 지능정보연구, 22(1), 139-157.

박선주 (2017). SW 교육 뉴스데이터의 감성분석. 정보교육학회논문지, 21(1), 89-96.

배재권 (2006). 재무비율을 이용한 기업부도예측 모형의 예측력 비교 연구. 석사학위논문, 서강대학교.

안철휘 · 안현철 (2018). 효과적인 기업부도 예측모형을 위한 ROSE 표본추출기법의 적용. 한국콘텐츠학회논문지, 18(8), 525-535.

오세경 · 최정원 · 장재원 (2017). 빅데이터를 이용한 딥러닝 기반의 기업부도예측 연구. KIF working paper, 2017(8), 1-113.

오세경 (2001). 다변량 판별분석모형과 주식옵션모형을 이용한 기업도산 예측.

오우석 · 김진화 (2017). 인공지능기법을 이용한 기업부도 예측. 산업융합연구 (구 대한산업경영학회지), 15(1), 17-32.

유은지 · 김유신 · 김남규 · 정승렬 (2013). 주가지수 방향성 예측을 위한 주제지향감성사전 구축 방안. 지능정보연구, 19(1), 95-110.

이상훈 · 최정 · 김종우 (2016). 영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석. 지능정보연구, 22(2), 97-113.

이인로 · 김동철 (2015). 회계정보와 시장정보를 이용한 부도예측모형의 평가 연구. 재무연구, 28(4), 625-665.

이재식 · 한재홍 (1995). 인공신경망을 이용한 중소기업 도산 예측에있어서의 비재무정보의 유용성 검증. 한국전문가시스템학회지, 1(1), 123-134.5

조정만 · 박세운 · 현영하 · 이계원 (2005). 도산예측모델의 유용성에 관한 실증적 연구-E. Altman, W. Beaver, J. Horrigan, 한국은행 및 종합금융회사 신용평가 모델적용결과 분석: 한국기업과 일본기업의 표본을 대상으로. 한국회계학회 학술연구발표회 논문집, 335-374.

차성재 · 강정석 (2018). 딥러닝 시계열 알고리즘 적용한 기업부도예측모형 유용성 검증. 지능정보연구, 24(4), 1-32.

최소윤 · 안현철 (2015). 퍼지이론과 SVM 결합을 통한 기업부도예측 최적화. 디지털융복합연구, 13(3), 155-165.

최정원 · 한호선 · 이미영 · 안준모 (2015). 텍스트마이닝 방법론을 활용한 기업 부도 예측 연구. 생산성논집 (구 생산성연구), 29(1), 201-228.

한주동 (2017). 기업의 뉴스정보를 이용한 신용위험 측정모델.
석사학위논문, 한국방송통신대학교.

한국거래소 (2020). 유가증권시장 상장기업(12월) 2020년 상반기 결산실적.

한국은행 (2020). 금융안정보고서(2020년 6월).

현대경제연구원 (2019). 2020년 국내외 경제 이슈.

네이버 검색 . https://search.naver.com/search.naver?where=news&sm=
tab_jum&query=삼성전자

데일리레코드 . (2020.06.27).
https://blog.naver.com/daily__record/222013754730.

위키백과. https://ko.wikipedia.org/wiki/로지스틱 회귀.

형태소 분석기 비교 . (2020.12.05). https://passerby14.tistory.com/3.

한국거래소 . http://marketdata.krx.co.kr/mdi#document=040606.

A simple neural network with Python and Keras . (2016.09.26).
https://www.pyimagesearch.com/2016/09/26/a-simple-neural-netw
ork-with-python-and-keras/.

Calcworkshop . (2020.01.20.).
https://calcworkshop.com/functions-statistics/z-score/.

Introduction to XGBoost in Python . (2020.02.13).
https://blog.quantinsti.com/xgboost-python/.

likejazz . (2018.05.30). http://docs.likejazz.com/lstm/.

Python oversampling 기법 . (2020.11.25).
https://blog.naver.com/zeroalgorithm/222154416814.
Random Forest Algorithm- An Overview . (20200.02.19).
https://www.mygreatlearning.com/blog/random-forest-algorithm.

Abstract

A Study on the Predictive Model for Corporate Bankruptcy with Machine Learning

Lee, Joon Ki
Business Big Data Analytics
The Graduate School of Information
Yonsei University

The purpose of this study is to predict corporate bankruptcies in advance and selected quantitatitive information which is financial information and selected sentiment analysis information of online news data which is unsctructured information, as vaiables. Financial information is a method that comprehensively analyzes the business status of a company and three financial model variables (Altman, 1968; Beaver, 1968; Horrigan, 1966), which are well known in accounting and finance literature and verified over a long period of time. A total of three-year financial variables were selected by combining the management analysis index (Bank of Korea). To derive meaningful financial factors that indicate signs of corporate default, Statistical

verification was performed using t-test and logistic regression methods, and as a result of the study, total asset retained earnings rate, total asset margin, sales working capital ratio, capital sales ratio, and dependence on borrowings were found to be significant financial variables.

In order to verify how effective news content, modeling was applied by adding news sentiment analysis scores to financial variables. A corpus-based sentiment dictionary was constructed by extracting nouns from the refined text, and sentiment scores for each news collection period were added as a variable. As a result of adding the sentiment score of corporate news, it was found that the performance was much better than when using only financial data. In the case of the SVM, which has the best performance in terms of sensitivity (predicting an actual bankrupt company as a bankrupt company), And it was confirmed that the sensitivity performance was the best when the news collection period was applied for 3 to 4 months.

Therefore, if unstructured data, which is the sentiment analysis information of online news verified in this study, is used together in predicting bankruptcy, it will be very helpful to establish an effective loan decision support system.

key words : Machine Learning, Sentiment Analysis, Bankruptcy, TF-IDF, Predictive Modeling