뉴스 데이터를 활용한 한계기업 탐지에 관한 연구

정한성[1], 임희석[1]
[1]고려대학교 컴퓨터정보통신대학원
jhs5001@korea.ac.kr, limhseok@korea.ac.kr

A Study on The Detection of Marginal Firms Using News Data

Han-Sung Jung[1], HeuiSeok Lim[1]
[1]Graduate School of Coputer & Information Technology, Korea University

요 약

한계기업은 성장가능성이 있는 기업들에게 돌아가야 할 자금 및 지원정책을 기업의 연명수단으로 전략하게 될 가능성이 있어 비효율적 자원배분을 초래하게 되며 이는 궁극적으로는 경제성장의 제약을 유발하게 된다. 따라서 본 연구에서는 뉴스 데이터를 활용하여 이러한 한계기업을 초기에 탐지할 수 있는 방법을 제안하고자 한다. 연구결과, 뉴스 데이터를 활용하였을 경우, 그렇지 않은 경우보다 모든 지표가 우수한 것으로 나타나 실제적인 문제에서의 적용 타당성과 가능성을 보였다. 이를 통해 기업은 부실화된 정도를 사전에 예측하여 경영 전략 재수립을 위한 지표로 활용할 수 있을 것이며, 투자자는 리스크를 관리할 수 있는 수단으로 활용될 수 있다.

1. 서론

1997년 외환위기와 2008년 금융위기를 겪으면서 한국 경제는 많은 변화를 겪었다. 제조업 위주의 산업은 서비스업을 중심으로 개편되었으며, 기업의 고용 형태도 크게 변화하게 되었다. 금융 및 자본시장을 비롯한 거의 모든 분야가 개방된 한국 경제시장은 미국의 서브프라임사태, 중국의 경제 경착륙 우려, 유럽의 재정위기 등에 직·간접적인 영향을 받고 있는 실정이다. 잠재성장률 하락과 청년실업문제 등 한국의 경제는 어려운 상황에 처해 있다.

이러한 상황을 돌파하기 위해서는 새로운 아이디어와 기술을 확보한 신생 기업의 시장 진입도 중요하지만 경쟁력을 상실한 기존 기업의 소멸 또한 경제발전에 중요한 요소이다. 하지만 2008년 금융위기 발생 이후 이러한 시장의 선택 체제가 원활하게 작동하지 않아 시장 퇴출의 경계선에 있는 많은 부실기업들이 퇴출되지 못하고 있다[1]. 이러한 퇴출의 경계선에 있는 기업들을 좀비기업 혹은 한계기업이라고 하는데, 한계기업은 고용창출과 설비투자가 미흡하고 생산성도 상대적으로 낮은 기업들을 의미한다. 또한 한경 경제용어사전에서 한계기업은 3년 연속 이자보상비율(영업이익/이자비용)100% 미만이거나 영업활동 현금흐름이 마이너스를 기록하고 있는 기업이라고 정의하였다.

이러한 한계기업들이 많아지게 되면 성장성 있는 다른 기업들에 돌아가야 할 자금과 지원정책이 성장가능성 없는 기업의 연명수단으로 전락하게 되어 비효율적 자원 배분을 초래하게 된다. 이는 궁극적으로 경제성장을 제약하는 요인으로 작용하며 경기부진이 지속될 경우 금융기관의 동반 부실화를 초래하게 된다. 따라서 사전에 한계기업의 징후를 감지할 수 있는 수단이 제공되면 전 다각적 한계기업 관리 체계 및 투자자 보호수단 구축이 가능하다.

빅데이터 기법 중 텍스트 마이닝 기법은 언어학, 통계학, 전산학 등이 융합된 분석방법이다[2]. 세상에 존재하는 데이터의 80% 이상이 비정형 데이터로 추산되는데, 그 중에서도 텍스트는 가장 기본적이고 광범위한 비중을 차지하는 비정형 데이터이다[3]. 텍스트 마이닝을 통해 텍스트 내 단어를 중심으로 유의미한 결과를 발견하고, 연구자에 따라 다양한 방향으로 해석하는 데 사용된다. 최근에는 경영학, 경제학, 심리학 등 다양한 분야에서 온라인 리뷰, 뉴스 데이터 등을 바탕으로 통찰을 이끌어내는데 사용되고 있다.

이러한 방대한 텍스트 데이터를 생산하고 있는

언론은 위험의 경고자이자 의제 설정자, 의견의 전달자 등의 역할을 수행한다. 한편, 언론은 지나친 속보 경쟁으로 인한 부정확한 보도, 공포를 조장하는 자극적인 보도를 한다는 지적을 받기도 한다. 따라서, 언론이 기업과 관련하여 어떠한 의제를 설정하여 어떻게 보도하는지 객관적인 데이터를 바탕으로 분석하는 연구가 필요하다.

본 연구에서는 데이터를 2012년도부터 2014년까지 코스피 상장사 데이터를 수집해서 수집된 데이터로부터 예측모형을 구축하여 한계기업을 예측하였다. 또한 기존 연구에서 쓰였던 변수 외에 뉴스 데이터에서 감성을 추출하여 알고리즘에 투입하였다. 이를 통해 기업은 부실화된 정도를 사전에 예측하여 경영 전략 재수립을 도모할 수 있으며, 투자자는 리스크를 관리할 수 있는 수단으로 활용될 수 있다.

2. 연구방법

2.1. 데이터 수집 및 처리

한계기업탐지를 위해 2012년도부터 2014년까지 연도별 코스피 상장사의 연도말 재무제표 데이터를 FnGuide에서 수집하였다. 수집된 데이터의 정상기업과 당해 한계기업 조건을 충족한 수는 <표 1>에 제시되어있다. 3개년도 평균적으로 총 712개의 기업데이터가 수집되었으며, 30.1%의 한계비율을 보였다.

<표 1> 수집 데이터의 연도별 기업현황

연도	정상기업 수	당해 한계기업 수	한계기업 비율
2012	491	220	30.9
2013	506	206	28.9
2014	495	218	30.6
평균	497	215	30.1

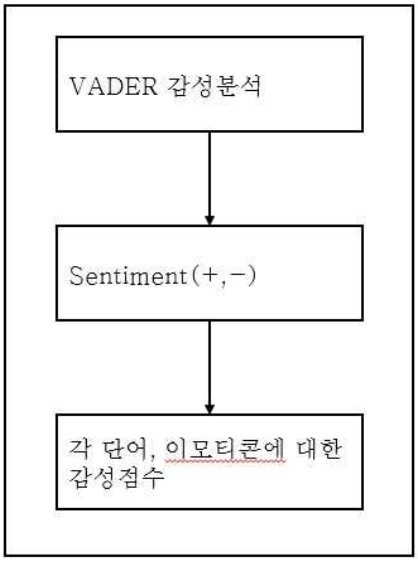
2012년도, 2013년도 재무제표 데이터에서 당해 이자보상비율(영업이익/이자비용)100% 미만이거나 영업활동 현금흐름이 마이너스 여부를 추출하였다. 조건을 충족하면 1, 충족하지 않으면 0으로 추출하였다.

2014년도 재무제표 데이터에서 타인자본비율(%), 자기자본비율(%), 차입금의존도(%), 현금및현금성자산구성비율(%), 유형자산구성비율(%), 무형자산구성비율(%), 부채비율(%), 차입금비율(%), 총자산/총자본(%), 차입부채조달금리(기말)(%), 차입부채/영업수익(%), 차입부채/영업이익(배), 차입부채/EBITDA(배), 현금흐름/총자산(%), 현금흐름/총자본(%), 현금흐름/총부채(%) 변수를 최종적으로 분석에 활용하였다. 또한 3년 연속 이자보상비율(영업이익/이자비용)100% 미만이거나 영업활동 현금흐름이 마이너스인 조건으로 한계기업은 1, 정상기업은 0, 판단 불가는 -1로 라벨링하였고, 현황은 <표 2>에 제시되어 있다. 한계기업을 분석할 때는 판단불가 데이터는 삭제하였다.

<표 2> 2014년 기업 현황

연도	정상기업 수	한계기업 수	판단불가 수
2014	569	124	20

뉴스 데이터 활용을 위해 각 기업명칭이 제목에 포함된 2014년도 온라인 뉴스를 BIG KINDS에서 크롤링하여 수집 후, Google Translate API를 사용하여 영문으로 변환하였으며, 변환된 데이터셋은 NLTK패키지의 VADER를 사용하여 긍정, 부정, 중립으로 분류하였다. 최종적으로 각 기업별 뉴스이미지의 긍정, 부정, 중립 뉴스를 비율로 계산하여 한계기업분류에 사용하였다.



(그림 1) VADER 감성분석

VADER는 감정의 강도 측정이 가능한 어휘 특징의 표준 목록과 문법 및 구문 규칙을 결합하여 감정을 평가하는 정성적 방법과 정량적 방법이 조합된 규칙 기반의 감성분석기법이다.

기업의 뉴스 특성상 경제용어가 많고, 경제용어의 감성분석을 하기 위해서 감성어 사전에 경제용어를 포함하여 구축되어야 한다. VADER 감성분석기법을 적용한 이유는 VADER가 학습한 감성어 사전

에 경제용어를 포함하여 잘 구축되어 소셜 미디어 스타일의 텍스트에 잘 작동하며, 감성을 단순 이분형(긍정, 부정)이 아닌 긍정, 중립, 부정을 수치화하여 구분할 수 있기 때문이다.

2.2. 분석방법

학습모형에 대한 성능을 평가하기 위해 전체 데이터를 7:3비율로 분할하여 훈련 및 시험데이터로 분류하였으며, 모델의 최적의 파라미터를 찾기 위한 파라미터 튜닝과 모델의 과적합 방지를 위해 Grid Search 5-fold Cross Validation를 수행하였다. 분석 알고리즘은 최종 목표변수가 이분형변수인 한계기업 여부인 관계로 Logistic Regerssion을 사용하였다. 또한 뉴스 데이터를 사용한 데이터셋과 사용하지 않은 데이터셋을 비교하여 최종 모델을 비교하였다.

3. 연구결과

뉴스 데이터를 활용하지 않은 경우와 활용한 경우 모두 Logistic Regression의 최적의 파라미터는 C:0.9, max_iter:10000, penalty:l2로 나왔다.

<표 3> 모형 검증 결과

지표	뉴스 데이터 미활용	뉴스 데이터 활용
Accuracy	.952	.957
roc_auc	.899	.912
Precision	.912	.914
Recall	.816	.842
F1-score	.861	.877

뉴스 데이터를 활용하지 않은 경우와 활용한 경우의 점수 지표는 <표 3>와 같다. 분석 결과, 뉴스 데이터를 활용하지 않았을 경우에는 Accuracy: .952, roc_auc: .899, Precision: .912, Recall: .816, F1-score: .861로 나타났으며, 뉴스 데이터의 감성을 활용한 데이터셋은 Accuracy: .957, Roc_auc: .912, Precision: .914, Recall: .842, F1-score: .877로 나타나 모든 지표에서 점수의 상승이 있는 것으로 나타났다.

<표 4> Logistic 분류결과

뉴스 데이터 미활용		예측(%)	
		정상기업	한계기업
실제	정상기업	80.4	1.4
	한계기업	3.3	14.8

뉴스 데이터 활용		예측	
		정상기업	한계기업
실제	정상기업	80.4	1.4
	한계기업	2.9	15.3

또한 Logistic 분류 결과, 뉴스 데이터를 활용하였을 경우, 한계기업의 탐지율이 늘어난 것을 확인할 수 있었다.

4. 논의 및 결론

한계기업은 본질적으로 시장에서 퇴출되어야 하는 기업임에도 불구하고 생존하여 정상기업과 신규 시장 진출 기업에 배분되어야할 자원을 점유하여 생산성과 경쟁력의 관점에서 부정적인 영향을 미치고 있다. 또한 잠재성장률이 계속해서 저하되는 상황에서 판데믹 상황까지 겹쳐지면서 한계기업이 점점 증가하고 있는 추세이며, 특히 만성적인 한계기업이 크게 증가하고 있는 실정이기에 이러한 문제에 대한 우려가 높다. 따라서 본 연구에서는 한계기업을 예측하는 정량적 모형을 구축함으로써 기업이 최종적으로 부도에 이르기 전 적절한 상시 기업구조조정 절차에 도움이 될 수 있는 방법을 조기에 시행 가능할 수 있게 됨에 그 의의가 있다. 또한 텍스트 마이닝 기법을 추가적으로 활용함으로써 비정형 데이터를 통한 추가적인 예측이 가능함을 시사하였다.

본 연구에서 제안된 알고리즘을 바탕으로 활용된 한계기업 예측모형은 기업의 부실한 정도를 예측가능한 정보로 제공해줌으로써 이해관계자들에게는 예측가능한 손실을 최소화 시키며, 투자자 입장에서는 감수해야할 리스트를 관리할 수 있는 지표로써 활용되며, 이는 궁극적으로 국내 금융시장 내의 안정성 공급에 기여할 수 있다. 향후 연구과제로서 추가적인 데이터 투입을 통한 예측 모형의 고도화 및 다양한 알고리즘 간 비교, 추가적인 feature발굴 등을 해볼 수 있을 것이라 기대된다.

참고문헌

[1] 김유진. “한계기업 회피를 위한 원가조정.”, 대한경영학회지, vol33 no10. pp1861-1876, 2020.

[2] 김성근, 조혁준, 강주영. “학술연구에서의 텍스트 마이닝 활용 현황과 주요분석기법.”, 정보기술아키텍처 연구, vol13 no2. pp317-329, 2016.

[3] Chakraborty, G., Pagolu, M., & Garla, S. “Text mining and analysis: practical methods, examples,

and case studies using SAS." SAS Institute, 2014.
[4] Hutto, C.J. & Gilbert, E.E. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014.