

## 2017년 한국재무학회 추계학술대회

### 빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구

최정원, 오세경, 장재원

**To cite this article :** 최정원, 오세경, 장재원 (2017) 빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구, 2017년 한국재무학회 추계학술대회, pp.396-435

① earticle에서 제공하는 모든 저작물의 저작권은 원저작자에게 있으며, 학술교육원은 각 저작물의 내용을 보증하거나 책임을 지지 않습니다.

② earticle에서 제공하는 콘텐츠를 무단 복제, 전송, 배포, 기타 저작권법에 위반되는 방법으로 이용할 경우, 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

[www.earticle.net](http://www.earticle.net)

# 빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구

최정원\* 오세경\*\* 장재원\*\*\*

## <요약>

본 연구는 기업 부도 예측 과정에서 새로운 정보 원천으로 비정형 데이터인 뉴스 텍스트 데이터를 계량화하여 활용할 수 있도록 인공지능 기법인 'Word2vec' 방법으로 측정하는 방법을 제시한다. 또한 인공지능 기반의 예측 방법론을 제시하고 기존의 방법론과 예측력을 비교 분석하였다. 연구 결과, 우선 연간 모형에서는 인공지능 기법인 Random forests 기법이 가장 우수한 예측력이 나타나는 것으로 분석되었다. 또한 인공지능을 이용한 다른 방법론들도 전반적으로 기존의 전통적인 예측 방법보다 예측력이 우수한 것으로 나타났다. 뉴스 텍스트를 추가적인 정보 원천으로 추가한 효과는 연간 예측 모형에서는 다소 미미하였다. 하지만 월간 예측 모형에서는 텍스트 정보 기반의 예측 모형이 시장 정보 기반의 예측 모형인 KMV 모형과 유사한 결론을 도출할 수 있어 기업 부도 예측 과정에서 조기 경보 모형으로 충분히 활용이 가능함을 실증하였다.

핵심단어 : 기업부도예측, 텍스트마이닝, Word2vec, 인공지능, 머신러닝

\* 주저자, 건국대학교 경영대학 박사과정(Email: garden31@gmail.com)

\*\* 교신저자, 건국대학교 경영대학 교수(Email: skoh@konkuk.ac.kr)

\*\*\* 고려대학교 의학통계학과 석사과정(Email: jeawonlll@naver.com)

## 1. 연구배경 및 목적

과거 금융위기 등의 주요 발생 원인으로 신용 위험에 지목되면서, 대출 차주의 부실을 선제적으로 예측하는 방법론은 그간 꾸준히 발전되어 왔다. 하지만 글로벌화, 기술의 발전, 산업 구조 변화 등의 원인으로 기업 경영환경이 급속하게 변화함에 따라 신용 부실(부도) 사건은 여전히 자주 발생하고 있다. 특히 기업은 부실이 발생할 경우 관련 기업의 근무자, 유관기업 및 개인, 금융기관을 비롯한 주요 투자자 및 채권자 등 모든 연관된 경제 주체에 연쇄적인 피해가 발생할 수 있기 때문에 기업 부도 위험을 보다 정확히 예측하고 평가하는 방법을 개발하는 것은 여전히 중요한 연구 과제이다.

과거 기업 부실을 예측하는 많은 연구는 주로 재무(회계) 정보와 (주식)시장 정보를 기반으로 예측을 수행하였다. 재무 정보는 공시된 정보를 활용하므로 기업의 현황을 가장 객관적이고 표준화된 형태의 데이터를 활용할 수 있다는 장점이 있다. 하지만 재무정보는 분기 혹은 연단위로 작성되고, 각 기업의 결산 시점 이후 공시되는데 까지 일정 기간이 소요되어 적시성이 떨어지는 근본적인 한계점이 있다.

이러한 단점을 보완하기 위하여 'KMV 모형'으로 대표되는 시장 정보를 활용할 수 있는 방법론이 제시되었다. 기업의 주가는 투자자들에 의하여 실시간으로 평가된 결과가 시장가격으로 형성되므로, 기업의 현황 수준을 가장 빠르게 반영하여 주는 정보이기 때문이다. 하지만 시장 정보는 유가증권 시장에서 거래가 활발한 상장기업만을 대상으로 적용할 수 있으며, 주가에 영향을 주는 거시경제 요인이나 산업 요인, 각종 뉴스에 의한 일시적인 요인 등의 영향을 통제하지 못한다는 단점도 가지고 있다.

본 연구에서는 과거 활용이 어려웠던 텍스트 형태의 비정형 정보인 뉴스 정보를 활용하여 부도 예측 수준이 향상될 수 있는지 연구하고자 한다. 텍스트 정보는 뉴스 정보 뿐만 아니라 기업 공시 데이터, 웹 게시판, 기업관련 SNS 등 매우 광범위한 매체로 확장이 가능하며, 기업에 관한 뉴스는 해당 기업에 대한 가장 빠른 정보 중 하나이다. 따라서 기업의 부실을 선제적으로 예측하기 위한 추가적인(additional) 혹은 대체(alternative)할 수 있는 정보 원천으로서의 충분한 가치가 있다. 분석을 위하여 먼저 빅데이터 연구 분야에서 많이 활용되는 텍스트마이닝과 인공지능 기법을 이용하여 텍스트 정보를 측정 가능한 변수로 계량화하는 방법을 제시하고, 앞서 언급한 전통적인 정보 원천과 함께 텍스트 정보의 유용성을 실증 분석하였다.

텍스트마이닝은 문서, 웹 등의 텍스트 정보를 데이터베이스로 수집하고 데이터로 정제하는 과정을 포괄하는 개념으로서 정보처리 기술과 관련 기반(infra)의 발전에 따라 최근 급격하게

활용도가 높아지고 있다. 우선 정확한 예측 모델을 구축을 위하여 되도록이면 많은 텍스트 정보의 분석을 필요로 하게 되는데, 텍스트 데이터를 수기(scrap)로 취합할 경우 분석 표본(sampling) 데이터 수집 범위의 한계가 발생한다. 따라서 텍스트마이닝 과정에서는 광범위한 데이터의 보다 효율적인 확보 및 관리를 위하여 웹 데이터베이스(DB)에 직접 접근하여 데이터를 확보하는 웹 크롤링(web crawling) 방법을 주로 활용한다. 이는 기존의 수작업에 의존하여 텍스트 데이터를 수집하는 방법에 비하여 가용한 데이터의 범위를 크게 증진시킬 수 있고 분석자의 편의(bias) 혹은 실수 등의 오류(error) 또한 감소시킬 수 있다. 또한 자연어 처리(Natural Language Process, 이하 NLP), 단어(키워드) 빈도 분석, 감성 분석 등 텍스트 정보를 기반으로 계량화된 지표를 산출하는 방법론을 연구에 적용하였다.

또한 본 연구는 인공지능(A.I.) 분야의 여러 방법론을 적용하여 기존 방법론과 예측력을 비교 분석 하였다. 머신-러닝(Machine-Learning), 딥-러닝(Deep-Learning) 등의 용어로 대표되는 인공지능 분야는 컴퓨터 공학을 이용하여 인간의 두뇌와 같이 컴퓨터가 학습 과정을 거쳐 예측 프로세스 등의 의사결정을 수행하는 체계를 의미한다. 과거에는 다양하고 동시 다발적인 경우의 수를 처리하는데 있어 물리적으로 발생하는 한계로 인하여 주목 받지 못하였으나, 최근 'Google'사의 'AlphaGo'로 대표되는 딥-러닝 체계가 실제 인간의 판단 수준과 속도가 대등하거나 오히려 능가할 수 있다는 것을 증명 함으로서 전 세계적으로 큰 관심을 받고 있다. 인공지능은 학습 데이터가 많을수록 예측력이 우수해지는 특성이 있으므로, 텍스트 데이터 등의 빅데이터를 원천으로 활용하는 본 연구와 같은 예측 과정에서 더욱 우수한 효과를 기대할 수 있다.

빅데이터 및 인공지능은 제 4 차 산업혁명의 핵심 기술로서 여러 분야에서 많은 관심을 받고 있지만, 금융, 재무 영역의 연구에 적용된 사례는 아직은 많지 않다. 따라서 본 연구는 기업 부도예측 과정에 이러한 새로운 방법론 적용을 시도하고 예측 결과의 정확도를 평가 함으로서 기존의 방법론 대비 유용성을 실증해보고자 하였다.

## 2. 연구 방법론

### 2.1. 선행연구

#### 2.1.1. 기업 부도예측 연구

Altman(1968)의 다변량 판별분석과 Ohlson(1980)의 로짓 모형으로 대표되는 전통적인 재무 정보 기반의 기업부도예측 연구는 이후 재무 정보에만 국한하지 않고 시장 정보를 기반으로 기업 부도예측의 우수성을 연구하는 방향과 다양한 방법론은 적용하여 예측 성과를 높이는 방향으로 발전되기 시작하였다.

Mcquown(1993)은 자본시장의 시장 가격을 바탕으로 옵션가격 평가모형을 적용하여 기업의 부도 위험 수준인 EDF(Expected default frequency)를 측정하는 모형(KMV 모형)을 제시하였다. 이 연구는 EDF 를 이용하여 채무불이행 확률이 발생할 추정치를 도출하는 것은 단순한 기대손실과는 다르며, 채무불이행 예측에 보다 효율적임을 실증하였다. 오세경(2001)은 국내 기업을 대상으로 로짓(Logit) 모형을 이용한 다변량 판별분석과 함께 옵션가격 평가모형을 이용하여 EDF 의 시간별 변화 추이를 분석하였다. 연구 결과 부실 기업들의 EDF 가 부도가 발생하기 수개월 또는 1 년 이상 전부터 급격히 올라가는 것을 실증 함으로서 시장 정보에 의한 기업 부도예측이 국내 기업의 예측 추정에도 유용함을 증명하였다.

이와 같이 각각 부도예측 과정에 활용되던 재무 정보와 시장 정보는 Shumway(2001)가 회계 정보와 시장 정보를 헤저드 모형으로 통합하여 부도예측력을 높일 수 있는 방법을 제안하면서 본격적으로 두 정보 원천을 통합한 연구가 진행되었다. 이 연구는 재무 정보와 시장 정보가 상호 보완적으로 부도예측 성과를 높일 수 있다고 주장하였다. Campbell, Hilscher, and Szilagyi(2008) 또한 후속 연구에서 회계모형과 시장 정보를 결합한 헤저드 모형이 기존의 각각의 모형 보다 부도예측력이 우수하다는 것을 실증하였다. 이 연구는 기존에 활용되지 않았던 시장 정보 기반의 변수를 모형에 적용하여 Shumway(2001) 보다 부도예측력이 개선된 헤저드 모형을 제시한 것이 특징이다.

이인로 · 김동철(2015)은 Campbell et al(2008)의 연구 결과를 활용하여 회계정보와 시장 정보를 통합한 헤저드 모형으로 부도예측을 수행하였다. 이 때, 기존의 연구에서 미국기업에 적용하여 선정된 변수를 국내기업에 그대로 적용하여 변수의 계수만을 재 추정한 기존 모형과 국내기업에 적합하도록 모형을 수정한 새로운 헤저드 모형을 별도로 추정한 모형의 부도예측력을 비교 분석한 결과 국내기업에 적합한 변형된 모형이 최종적으로 가장 우수한 예측 성능을 나타내는 것을 실증하였다. 최정원·오세경(2016) 또한 비례 헤저드 모형을 활용하는 생존분석과 KMV 모형을 활용하여 재무 정보와 주가정보를 결합하는 방법을 연구하였다.

한편, 기업 부도에 또 다른 요인인 거시 경제 변수의 영향을 반영한 통합 모형 연구도 수행되었다. Nam, C., T. Kim, N. Park, and H. Lee(2008) 시간 가변적인(Time-varying) 헤저드 모형을 사용하여 거시경제 변동이 기업의 부도(헤저드) 확률을 상승시킬 수 있음을 실증분석하였다. Tinoco and Wilson(2013)은 재무 지표, 시장 정보와 함께 거시경제 변동 수준을 설명변수로 포괄하는 Panel Logit 기반의 다중회귀분석을 활용한 부도예측 모형을 연구하였다. 연구결과 거시경제 변동은 부도에 매우 결정적인(conclusive)영향을 미치지지는 않았지만 시장 정보 등 타 요인의 한계적(marginal)으로 영향을 미칠 수 있음을 연구하였다. 국내 연구로서는 김성규(2010)이 거시경제 상승/하락을 더미변수로 활용하여 기업 부도예측 모형에 활용할 수 있음을 제시하였다.

### 2.1.2. 빅데이터 기법을 활용한 관련 분야 연구

빅데이터를 활용한 예측 모형 연구는 최근 관련 분야의 대내외적인 관심 증가로 인하여 폭발적으로 증가하고 있다. 특히 텍스트 마이닝은 적용할 수 있는 영역이 광범위하여 선행 연구들은 다양한 분석 방법을 제시하고 있다. 배상진·박철균(2003)은 텍스트-마이닝 과정을 4 단계로 나누었는데 각각 문서 수집, 문서 전처리, 텍스트 분석, 그리고 결과 해석 및 정제 과정으로 설명하였다. 특히 기존의 데이터 수집 방법에 비하여 강조되는 부분은 전처리 과정으로서 텍스트-마이닝에 필요 없는 단어 또는 기호를 정제하는 과정과 문장의 정확한 의미 파악을 위해서 각 단어의 어간을 파악하고 동의어를 할당하는 정규화 과정을 필요로 한다고 하였다. 또한 한글의 경우 동의어, 유사어 처리를 위해서 문장에서 최소의 의미단위를 추출해 내는 형태소 분석 (morphological analysis) 단계와 통사구조를 파악하는 구문 구조 분석 (syntactic analysis) 단계, 의미 구조를 추출하는 의미 분석 (semantic analysis) 단계를 나누어 분석하여야 함을 언급하였다. 김근형·오성렬(2009) 도 전처리 과정과 텍스트 분석 과정으로 나누어 설명하였는데, 전처리 과정을 일반적인 텍스트 데이터들을 컴퓨터가 처리하기 쉽도록 변화하는 작업으로 논하였다. 특정단어와 관련된 문서들을 신속하게 검색할 수 있도록 FB (Frequency-Based), IDF (Inverse Document Frequency), LSI (Latent Semantic Indexing) 등의 계량화된 지표(index)를 만드는 과정을 것이라고 설명하고 있다. 또한 문서와 단어간의 연관성 분석 등 계량화 방법도 제시하였다.

텍스트 정보를 이용한 분석과정에서 유의할 점은 단순히 텍스트의 횟수를 분석하는 방법은 텍스트가 담고 있는 감성(Opinion)을 분석 결과에 반영하기 어렵기 때문에 별도의 감성 분석을 필요로 한다는 점이다. 김유신·김남규·정승렬(2012)은 뉴스 키워드의 감성 분석을 이용하여 투자 의사결정 모형을 구축하고, 이 모형이 시장대비 초과 수익률을 얻을 수 있는 투자 전략임을 실증하였다. Martinez, Garcia, and Sanchez(2012)도 금융 경제 관련 뉴스 텍스트를 추출하여

의미와 감성을 분석하는 방법을 제시하였다.

기업 부도예측에 텍스트 정보가 활용된 사례는 비교적 최근의 연구들이 많은 편이다. 이광석(2014)은 기존의 재무 정보와 시장 정보 기반으로 중소기업 대상의 부도예측은 한계가 있음을 지적하고 해당 기업의 신용 거래, 연체 정보 등을 실시간으로 이용하여 부도예측을 수행하는 방법을 제시하였다. 이 연구는 기존의 부도예측의 사각 지대인 중소기업 및 개인기업을 대상으로 하는 매우 유용한 연구이기는 하나 제시된 주요 분석 정보 데이터가 아직은 공공재로 공개되지 않은 공공기관 내부 데이터로서 타 연구에 적용하기 어려운 한계가 있다. 최정원·한호선·이미영·안준모(2015)는 부도 기업과 정상 기업의 인터넷 뉴스 텍스트를 각각 수집하여 부도 기업 뉴스에서 주로 나타나는 키워드를 분석하고 해당 키워드가 포함된 기사가 발생하는 경우를 부도로 예측하였을 때 실질적으로 부도예측이 가능함을 검증하였다. 조남옥·신경식(2016)도 뉴스 텍스트에 대한 감성분석 기반의 multiple discriminant analysis(MDA)과 로짓 분석, 인공신경망, support vector machines (SVM) 등의 방법을 적용한 부도예측 모형을 제시하고 예측 모형으로서 유용성을 실증하였다.

부도예측 이외에도 금융 재무 분야의 뉴스 텍스트를 이용한 많은 연구는 주로 주식 등의 투자 자산 가격 예측에 관한 연구를 중심으로 진행되어 왔다. Chen, De, Hu, and Hwang(2014)은 인터넷과 SNS 상의 게시물을 '집단(군중)의 지성(wisdom of crowd)'으로 명명하고 텍스트 분석을 통하여 주가 예측이 가능함을 실증분석 하였다. 국내에서도 김민수·구평희(2013)가 검색엔진이 제공하는 검색어 추세를 기반으로 주가 예측을 수행하는 연구를 수행 함으로서, 전통적인 정보 외에 다른 정보 원천들도 빅데이터 분석으로 기존의 정보 원천을 대체할 수 있음을 연구하였다. 안성원·조성배(2010)도 뉴스 텍스트마이닝 기법을 시계열 분석 과정에 적용하여 주가예측 모형에 활용이 가능함을 실증 분석하였다.

### 2.1.3. 인공지능 기법을 활용한 관련 분야 연구

머신러닝, 딥러닝으로 대표되는 인공지능 기법은 비교적 최신 기술로서 금융 및 재무 분야에서는 전통적인 방법론에 의한 예측 방법론에 비하여 연구의 양과 질 모두 부족한 상황이다. 하지만 최근 기술의 발전 및 전 세계적인 관심 증가와 함께 관련 연구가 매우 급격하게 늘어나고 있으며, 부도 예측 분야도 몇몇 선도적인 연구가 진행 되었다.

이재식·한재홍(1995)은 기존의 재무정보만 활용한 부도예측의 한계가 있음을 지적하고 이를 보완하기 위하여 비재무정보를 활용한 인공신경망 기반의 부도예측 모형을 제시하였다. 연구 결과 재무 정보가 불투명한 중소기업의 경우 이러한 예측 모형이 더욱 효과적임을 분석하였다. Kim and So(2010)는 support vector machines (SVM)을 이용하여 부도 예측을 수행하였다. 이 연구

역시 정보가 상대적으로 부족한 중소기업(SME)의 경우 기존의 방법론에 비하여 인공지능 기법이 예측 성능이 더 우수함을 연구하였다.

김성진·안현철(2016)은 금융기관의 신용위험관리의 중요한 도구인 기업신용등급 예측 과정에 인공지능 기법 중 랜덤 포레스트(Random Forests) 방법을 적용하였다. 이 연구는 다중판별분석, 인공신경망, 다분류 SVM 등 기존 연구에서 전통적으로 기업 부도 예측과정에 사용되어 온 방법론과 비교에 랜덤 포레스트 방법론이 예측 성능이 우수함을 실증 분석하였다.

국외에서는 Yeh, Wang, and Tsai (2015)은 딥러닝 개념의 인공신경망 기법 중 하나인 Deep Belief Networks (DBN)이 기존의 머신러닝 중 대표적 기법인 SVM 보다 기업 부도예측 성능이 더 우수함을 연구하였다. 또한 Addal(2016)은 인공신경망(Artificial Neural network), K 근접 군집분석(k-Nearest Neighborhood) 등의 방법론을 이용하여 기업부도예측 모형이 우수한 예측력을 보이는 것을 실증하였다.

한편, 부도 예측은 아니지만 Vahala(2016)는 외환시장의 환율에 대하여 인공신경망(Neural network) 기반의 예측 모형 구축이 가능함을 보였다. 또한 Kim(2003)은 SVM 이 금융시장의 time series 속성의 데이터를 예측하는데 더욱 효과적임을 실증 분석하는 등 최근 인공지능 관련 기술의 진보와 함께 금융, 재무 분야의 관련 연구도 역시 급증하고 있다.

## 2.2 연구방법론

### 2.2.1. 분석 데이터 정의

본 연구는 선행연구를 참고하여 확보 가능한 다양한 정보 원천을 모두 포괄하여 예측 모형에 활용하였다. 분석에 활용한 데이터의 종류 및 주요 특징은 <표 1>과 같다<sup>1</sup>.

재무 정보의 경우 기업에 대한 가장 기본적인 실적 지표로서 기업 부도예측에 반드시 활용되는 정보이다. 재무 정보는 손익 성과를 측정하는 수익성지표, 자본구조를 나타내는 건전성 지표, 성장성 지표, 활동성 지표 등으로 구분할 수 있다. '주가' 등의 기업에 대한 시장 정보는 분석 시점의 기업에 대한 최신 정보를 반영하고 있다는 특성이 있으므로 재무 정보의 적시성을 보완할 수 있다. 다만, 시장 정보는 유가증권 시장에 상장되어 주식이 거래되고 있는 기업만이 정보를 이용할 수 있다는 한계점이 있다.

<sup>1</sup> 후보변수 중 모형에 선정된 각 변수의 구체적인 정의 및 산출 방법은 Appendix 에서 확인할 수 있다. 재무 정보와 시장정보는 공시 데이터를 정리한 재무 DB Source (Data Guide Pro 5.0)를 활용하여 수집하였다. 거시경제 정보는 한국은행 통계시스템(ecos)를 이용하여 연 단위 데이터를 수집하였다.



구분	의의	활용가능 데이터
1. 재무 정보	기업 공시(재무제표) 정보 결산(연/분기)기준 재무비율	- 수익성: 자산(자본)대비 수익률 등 - 성장성: 매출증가율, 자산증가율 등 - 건정성: 부채비율, 이자보상배율 등 - 기타재무지표, 주주비율 등 기업정보
2. 시장 정보	상장 기업의 주식 거래 관련 정보	- 시장지표: 주가, 시가총액, 주가수익률, 거래량 - 재무비율 혼합지표 시장이 대비 장부가 비율 시장조정부채비율, 시장조정 등
3. 거시경제지표	주요 기관에서 집계&발표하는 거시경제지표	- 거시경제지표: 국가총생산(GDP), 통화량, 물가지수 (PPI, CPI), 기업경기실사지수(BSI) 등 - 금융시장지표: 금리, 종합주가지수, 변동성 지수 등
4. 비정형 정보	전통적인 방법으로 활용하기 어려웠던 비정형(텍스트) 데이터	- 뉴스 정보뉴스 및 방송잡지 등), - 공시자료, SNS(인터넷 사이트) 등의 정보 - 주로 텍스트 형태의 데이터로 확보

<표 1> 기업 부도예측을 위한 원천 정보 구분 및 특성

재무 정보와 시장 정보는 각각 활용할 수 있지만, 두 정보를 결합하여 모형에 반영할 수 도 있다. 이인로·김동철(2015)의 연구는 국내기업에는 단순히 장부가액 기준의 재무지표 보다 시장 조정된 재무지표를 사용하는 것이 보다 우수한 예측력을 산출되는 것으로 나타났다. 본 연구 또한 재무지표 중 '총자산'을 '장부가 기준 총자산' 과 자본가격을 시장 가격으로 조정한 '시장조정 총자산'으로 나누어 설명 변수로 활용하였다.

거시경제 지표의 경우 과거 일부 부도 예측 연구에서 설명 변수로 활용은 되고 있으나, 그 빈도가 재무지표나 시장지표에 비하여 떨어지는 편이다. 이는 거시경제 지표는 각 기업의 특성이나 현황을 정확히 반영할 수 없기 때문이다. 하지만 기업의 부도 발생은 거시경제 수준이나 산업의 경기 수준에 독립적일 수 없기 때문에 재무지표 혹은 시장 지표와 함께 모형에 반영할 경우 보다 정확한 예측을 수행할 수 있을 것으로 기대할 수 있다. 특히 경기에 민감한 업종의 경우 금융위기 기간에 집중적으로 부실이 발생하는 특징이 있기 때문에 거시경제 및 산업 변수를 보다 적극적으로 활용할 경우 과거 연구에 비하여 우수한 예측 모형이 추정될 것이다.

비정형 정보는 그간에 연구들이 주로 사용하지 못하였던 정보 원천인 뉴스 및 인터넷 등의 미디어 데이터를 주로 포함한다. 뉴스 정보 획득을 위하여 인터넷 뉴스 포털의 기업명을 키워드로 검색한 기사 결과를 기반으로 뉴스 정보를 수집하였다. 신뢰도 있는 정보를 위하여 <표 2> 에 해당되는 언론사 기사만 선택하여 취합하였으며, 텍스트 수가 매우 적은 단순 사실 보도자료, 스포츠 기사, 중복 기사 등은 제외하여 분석 대상 텍스트 데이터를 구성하였다.

구 분	언론 매체
종합	경향신문, 국민일보, 뉴시스, 동아일보, 로이터, 문화일보, 서울신문, 세계일보, 연합뉴스, 조선일보, 중앙일보, 한겨레, 한국일보, JTBC, KBS, MBC, SBS, YTN
경제	뉴스토마토, 매일경제, 머니투데이, 서울경제, 아시아경제, 이데일리, 조선비즈, 파이낸셜뉴스, 한국경제, 한국경제TV, 헤럴드경제, MBN, SBSCNBC
온라인/인터넷	데일리안, 오마이뉴스, 쿠키뉴스

&lt;표 2&gt; 뉴스 텍스트 수집 대상 언론 매체

### 2.2.2. 예측 모형

본 연구 분석 과정에서 활용한 예측 모형의 종류와 각 방법론의 특징은 <표 3>과 같다.

예측모형		
분류	방법론	특징
이진분류 방법	로지스틱 회귀분석 (Logit)	전통적(대표적) 이진분류 모형
	Decision Tree	대표적인 Data mining 기반 이진 분류 방법론
생존분석	Cox-PH Hazard	공변량의 특성에 따른 생존기간 예측 모형
인공지능 (머신러닝) (딥러닝)	Random-Forest (RF)	Random-Forest 여러 개의 Decision Tree 들을 임의적으로 반복 학습하여 추정하는 앙상블 기법을 활용한 예측 방법론
	SVM	데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모형을 만들어 예측하는 방법론
	Deep Neural Network (DNN)	인공신경망의 Hidden Layer 층을 겹겹(Deep)하게 설계한 방법론
	Recurrent Neural Network(RNN)	DNN 의 Hidden Layer 설계 시 변수간의 시간 순서(Sequence)를 고려하여 설계하여 학습 과정에 활용한 딥러닝 방법론
시장정보 <sup>2</sup>	KMV 모형	옵션 가격 결정모형을 기반으로 주가 변동에 따른 부도 확률을 산출하는 방법론

&lt;표 3&gt; 기업 부도예측 방법론 요약

<sup>2</sup> 다른 방법론은 모두 연간 부도예측 과정에 활용하지만, 시장정보를 활용한 KMV 모형은 텍스트 정보를 이용한 월 단위 부도예측 과정에서만 비교 분석 모형으로 활용하였다.

### 2.2.2.1. 전통적 이진분류 방법론

기업 부도예측과 같은 이진(binary) 변수의 분류 방법 중 가장 대표적인 방법은 로지스틱 회귀모형과 의사결정나무(Decision Tree)를 들 수 있다. 로지스틱 회귀분석은 재무지표, 시장지표 등의 정보를 설명변수로 활용하여 기업의 부도 여부(1 or 0)를 추정할 수 있는 방법이다. 의사결정나무 역시 이진 분류에 많이 활용되는 방법으로서 부도 여부를 결정하는 중요한 요인 및 기준 값을 노드(분류 기점)로 설정하여 분류 나무(tree) 구조를 설계 함으로서 부도 여부를 판단할 수 있는 모형이다.

이 두 모형은 그 동안의 연구에서 지속적으로 활용되어 왔으므로 새로운 분류(예측) 기법을 평가하는 기준 모형으로 많이 활용된다. 본 연구 또한 예측 모형에 적용되는 동일한 분석 데이터를 로지스틱 회귀모형과 의사결정나무에 적용하여 기존 방법론과 새로운 방법론의 예측 성능을 비교한다.

### 2.2.2.2. Cox 비례(PH) 헤저드 모형

생존분석 방법론 중 하나인 헤저드 모형(hazard model)은 회계정보와 시장 정보를 통합하여 부도를 예측하는 모형으로 부도 발생시점까지의 시간을 고려하는 방법론이다. 특히 공변량을 모형에 적용할 수 있는 Cox 비례위험 모형(Cox PH Regression)은 종속 변수가 부도 여부를 판별하는 이진 분석 방법론에 비하여 기업 생존 주기에 따른 부도 발생 확률이라는 추가적인 정보를 적용할 수 있다는 장점이 있다(최정원 외, 2016).

이인로 외(2015)는 헤저드 모형을 기반으로 회계정보와 시장 정보를 결합하여 부도예측을 수행하였을 때 기존의 방법론에 비하여 우수한 예측력을 얻을 수 있음을 언급하였다. 이 연구는 Campbell et al.(2008)이 제시한 헤저드 모형의 경우 미국 기업에 맞도록 변수가 설계되어 있어 수정이 필요함을 주장하고, 국내 현황에 맞도록 수정한 변수를 적용한 새로운 헤저드 모형이 보다 더 우수한 예측력이 나타나는 것을 실증 분석하였다. 본 연구는 이밖에 많은 선행 연구들이 제시한 유의한 변수와 방법론을 적용한 헤저드 모형을 구축함으로서 부도예측 최적의 헤저드 모형 추정을 시도하였다.

Cox 비례헤저드 모형을 추정하는 과정은 다음과 같다. 어떠한 개체(기업)의 사망(부도)가 발생하는 시점을  $T$  라고 가정하면, 현재( $t_0$ ) 시점에서의 추정 생존기간은  $T - t_0$  가 된다. 이와 같은 가정하에 생존기간은 식 (1) 과 같은 확률밀도 함수를 가지는 확률변수로 표현할 수 있다.

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t}$$

이러한 확률변수를  $F(t) = P(T < t)$  의 누적함수 형태로 정의하면,  $t$  시점 이전에 사망하지 않을 확률을 식(2)와 같은 생존함수 형태로 표현 할 수 있다. 또한 이 생존함수를 식(3)과 같이 역함수 형태로 변환하면 헤저드(위험) 함수를 얻을 수 있다.

$$S(t) = P(T > t) = 1 - F(t)$$

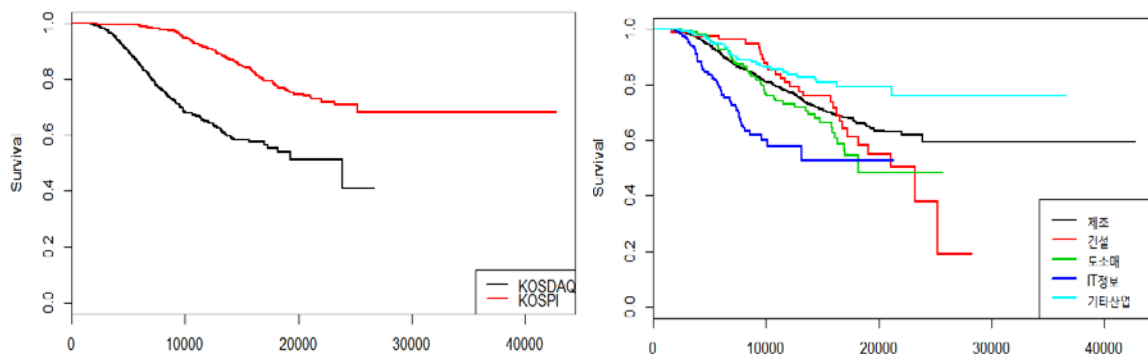
$$h(t) = \frac{f(t)}{S(t)}$$

$$\frac{h_1(t)}{h_0(t)} = \exp(b_1 * X_i)$$

도출된 헤저드 함수를 기저 함수로 공변량(설명변수)의 영향을 반영하는 Cox 비례위험 모형은 식 (4) 와 같이 정의된다.  $h_1(t)$ 는 사망(부도) 발생 기업의 헤저드 함수이고  $h_0(t)$ 는 정상기업의 헤저드 함수이다. 따라서 이 모형은 재무 정보, 시장 정보 등의 설명변수가 기업의 부도(헤저드) 확률에 비례적으로 어떠한 영향을 주는지 도출하는 모형이다.

본 연구에서는 기저함수를 추정하는 데 있어 대표적인 비모수적 함수 추정 방법으로 Kaplan-Meier 법(K-M 법)을 이용한다. K-M 법은 해당 기간에 누적으로 생존한 개체 수를 비율로 표시하여 주는 방법인 단순 누적 생존확률과 크게 다르지 않다. 하지만 확률론에 입각하여 모수 분포의 가정이 필요 없고 중도 절단이 있는 생존 자료의 특성을 반영할 수 있기 때문에 적은 양의 표본을 통해서도 생존함수를 추정할 수 있는 장점이 있는 방법론이다.

본 연구는 기저 헤저드 함수 추정에 있어서 산업별 층화 헤저드 함수를 적용하였다. 각 산업은 특징에 따라 생존 주기에 차이가 날 수 있다. 이를 그래프로 도식하면 <그림 1> 과 같다.



<그림 1> 시장별 산업별 헤저드(생존) 함수 산출 결과

우선 <그림 1>의 좌측 그래프 시장별 생존 함수의 차이이다. KOSDAQ 에 속한 기업이 상대적으로 생존확률이 상당히 떨어지는 것을 확인할 수 있다. 우측 그래프는 주요 산업(그룹)<sup>3</sup> 별 생존함수 이다. '건설' 산업에 속한 기업은 타 산업에 비하여 초기에는 오히려 생존 확률이 높다가 일정기간 이후 급격하게 부도가 많이 발생하는 것을 확인할 수 있다. 'IT 정보서비스' 산업은 오히려 사업 초기에 부도가 많이 발생하지만 일정 기간 이후에는 부도 기업이 증가하지 않는 것을 볼 수 있다. 이처럼 산업별로 생존 함수는 약간의 차이가 나타나게 되므로 각 기업이 속한 산업별로 생존함수를 적용하여 층화(strata) Cox 비례위험 모형을 적용하면 보다 우수한 예측 성과를 기대할 수 있다. 더욱이 층화 모형은 <그림 1>의 산업별 생존함수처럼 함수가 교차하는 경우 보다 우수한 예측 성능을 나타낸다(박재빈, 2006).

### 2.2.2.3. 인공지능 기법

인공지능 기법 중 Support vector machine(이후 SVM)은 최근 가장 빈번하게 기업 부도예측 연구에서 활용되는 방법론 중 하나이다. SVM 은 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘이 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모형이다. 만들어진 모형은 데이터 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. SVM 은 선형 분류와 더불어 비선형 분류에서도 사용될 수 있다는 점과 기존의 머신러닝(데이터마이닝) 방법론에 비하여 과적합(over-fitting)이 발생할 가능성이 낮은 장점이 있다고 알려져 있다(김경재, 2002). 하지만 SVM 은 효과적인 입력변수 선정에 대한 과정이 알고리즘 내부에 포함되어 있지 않다는 점, 많지는 않지만 커널함수 및 커널파라미터 등과 같은 직관에 의해 설정되어야 할 모수들이 있다는 점에서 다소 한계가 있다.

Breiman(2001)은 의사결정나무(Decision tree) 보다 강건한 예측 방법론으로 Random forests 방법론을 제안하였다. Random forests 는 독립적인 난수 sample vector 로 개별적인 의사결정나무 구조를 반복적으로 구성하고 이를 통합적(앙상블, ensemble)으로 대표할 수 있는 모형을 찾아내는 방법이다. 대수의 법칙에 의해 숲(Forests)의 크기(나무의 수)가 커질수록 모형의 정확도가 상승하고, 일반화 오류가 특정 값으로 수렴하게 되어 과적합화를 피할 수 있다. 또한 각 개별 의사결정나무들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 잡음(Noise) 및 이상값(Outlier) 으로부터 크게 영향을 받지 않는다는 장점이 있다. Random forests 가 갖는 또 다른 큰 장점은 모형의 설계자가 입력변수 선정으로부터 자유로울 수 있다는 점이다. 때문에, 많은 수의 독립변수와 방대한 양의 학습 사례로부터 분류·예측을 수행하여야

<sup>3</sup> 산업 그룹은 표준산업대분류를 기준으로 각 산업에 속한 기업 수를 기준으로 특정 개수 이상인 그룹을 별도 구분하였다. 기타 그룹은 산업 간 유사한 속성으로 보기는 어렵지만 Sample 수가 적어 통합하지 않으면 생존함수 추정이 불가능하다.

하는 본 연구에 매우 적합한 방법론이 될 수 있다. 또한 Random forests 는 빈도가 불균형한(imbalanced) 이항분류의 예측에 있어 가장 우수한 예측력을 보인 것으로 보고되고 있다(김성진 외, 2016).

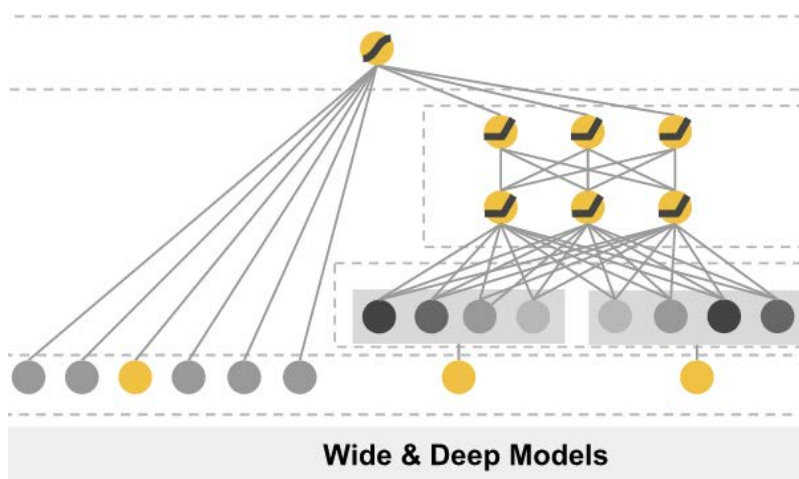
딥러닝 기법은 머신러닝의 한 종류로서, 1980 년대 등장한 인공신경망(ANN, Artificial neural network)를 기반으로 설계된 개념이다. 인공지능은 IT 기술 및 각종 분석 기법의 발전과 함께 단점들을 보완하며 점차 그 한계를 극복하여 왔는데, 최근 'AlphaGo' 로 대변되는 Google 사의 'DeepMind' 체계가 경우 '비지도 학습(unsupervised learning)'을 통한 최적화로 경우의 수가 무한에 가까운 바둑 분야에서도 인간을 넘어서는 능력을 보여 줄 수 있음을 증명 함으로서 '딥러닝'이라는 체계가 전 세계적으로 조명을 받고 있다.

딥러닝의 구조적인 특징은 기존의 인공신경망(neural network)에서 활용되는 은닉층(hidden layer)을 겹겹히(deep) 쌓아 특정한 조건에서 컴퓨터가 스스로 최적의 모형을 도출하도록 유도한다는 점이다. 과거에는 이러한 다중 구조의 최적화 자체가 쉬운 일이 아니었으나, 컴퓨터 처리 속도의 향상, 데이터 처리 기술의 발달, Back Propagation 등의 연산방법 개발 등이 이루어지며 직접 구현이 가능한 수준으로 발전하였다. 더욱이 최근에는 'TensorFlow' 등 간단한 딥러닝 엔진은 Python 등 open source 로 비교적 손쉽게 개인 컴퓨터로 개발하여 활용할 수 있기 때문에 더욱 관련 분야가 발전되고 있다.

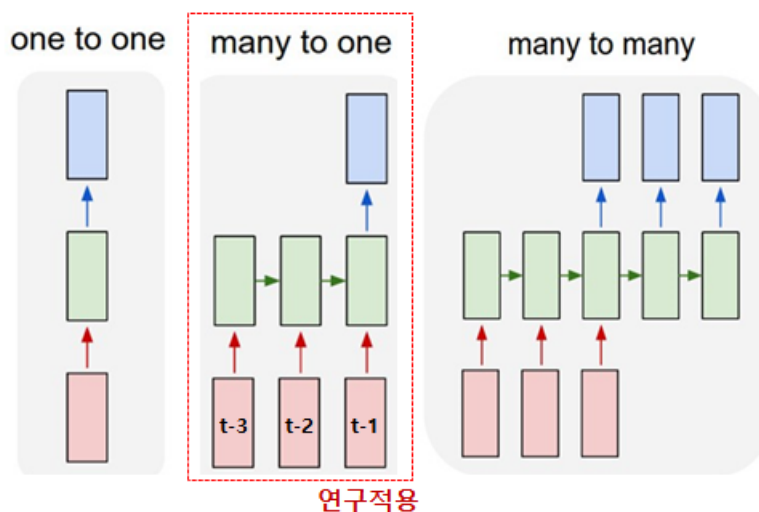
딥러닝은 은닉층(Hidden Layer)을 어떻게 구성하는가에 따라 여러가지 구조로 모형을 구성할 수 있다(그림 2). 은닉층을 넓게(wide) 혹은 깊게(Deep) 이론적으로는 모형의 정확도가 상승한다. 반면, 은닉층이 넓거나 깊게 설계할 경우 학습 및 추정하기 위하여 컴퓨터가 활용하여야 하는 Resource 의 물리적인 양이 증가하므로, 추정 시간이 오래 걸리거나 컴퓨터의 CPU, 메모리 등의 고성능 하드웨어가 필요할 수 있다는 단점이 있다.<sup>4</sup>

딥러닝 체계 설계 시 변수 간의 시간(순서, sequence)를 반영한 모형을 적용하기 위해서는 RNN 체계를 설계하는 것을 고려할 수 있다(Gu, Zhang, Zhang and Kim, 2016). 이는 Panel data analysis 혹은 VAR(vector auto-regression) 모형과 같이, 모형에 투입되는 변수의 선후 관계 혹은 시간 등 순서를 지정하여 추정하는 방법이다. 본 연구에서 부도예측의 설명변수로 활용하는 재무/시장/거시경제 정보는 전기( $t-1$ ) 뿐만 아니라 그 전의 기간 ( $t-2$ ,  $t-3$ , ...)에도 영향 받을 수 있다(그림 3). 따라서 RNN 체계를 활용할 경우 좀 더 정확한 예측이 가능할 것으로 기대하였다. RNN 체계는 구조가 복잡하여 학습과 예측에 투입되는 Data 양이 많아야 하고 역시 계산에 소요되는 Resource(시간, 하드웨어 성능)이 추가적으로 필요하다는 단점이 있다.

<sup>4</sup> DNN 최적의 예측 모형을 산출하기 위해서는 Cost 함수기준, 층별 가중치(LSTM), 시작값 등 선택하여 딥러닝 체계를 설계한 후에 학습 및 예측 과정을 수행하여야 한다.



<그림 2> DNN 체계 구성 개념



<그림 3> RNN 체계 구성 개념

#### 2.2.2.4. KMV 모형

재무정보가 포함된 부도 예측 모형은 재무 정보의 생성 주기가 연간<sup>5</sup>으로서 재무지표 공시 기간 동안에는 기업 재무 현황이 변동되어도 재무지표에 반영되지 않는다. 따라서 재무정보 기반의 부도 예측 모형은 적시성이 떨어지는 단점을 필연적으로 가지고 있다.

<sup>5</sup> 상장기업의 경우 분기 재무제표 공시가 의무화되어 있지만, 기업의 현황을 정확히 반영하는 정보는 여전히 연 정기 감사보고서 기준의 재무제표 정보를 대상으로 분석하여야 한다. 분기 재무지표는 연간 재무지표에 비하여 전기 대비 변동이 매우 적고 세부 계정단위로 정확한 데이터를 수집하는 것도 상대적으로 어렵다.

이와 같은 단점을 보완하기 위하여 제시된 개념이 Merton(1973)의 옵션 가격 결정 모형 기반의 시장 정보를 활용한 부도예측 모형(이하 KMV 모형)이다. KMV 모형은 기업의 정보가 즉각 반영되는 시장 정보(주가)를 기반으로 부도 확률을 예측하는 체계이기 때문에 앞서 언급한 재무 정보 변동 공백 기간의 적시성 문제를 보완할 수 있는 특성을 가지고 있다. 다만, 주가를 얻을 수 있는 상장 기업 만을 대상으로 분석이 가능하다는 한계점이 있다.

본 연구는 미디어 기사를 대상으로 텍스트 분석을 통하여 얻어진 정보를 기반으로 KMV 모형과 유사한 형태의 부도 예측 모형을 설계하여 기존의 KMV 모형과 예측 성과를 비교하는 연구를 수행하였다.

KMV 모형은 시장정보인 주가로서 기업의 일정기간 동안의 부도확률을 예측하는 모형이다. 이 모형은 Merton(1973)의 옵션 가격 결정 모형을 사용하여 기업의 부채 구조를 분석하였다. KMV 모형의 가장 큰 의미는 시시각각 시장 정보에 따라 변화하는 기업 주가로서 일정기간 동안의 부도 확률을 구할 수 있다는 점이다. 기존의 재무제표 변수는 회계정보의 기간 단위 보고의 특성상 즉각적인 정보의 적용이 어렵다는 단점이 있으나 KMV 모형은 매 시점에서 움직이는 주가 정보로서 EDF 를 도출함으로써 이를 보완하여 보다 빠르게 기업 부도 위험을 인지할 수 있다는 것이 최대 장점이다. KMV 모형은 또한 EDF 를 구하기 위한 과정이 매우 간단하면서도, 블랙-숄즈-머튼 옵션 가격 모형을 사용하였기 때문에 이론적으로 기반이 확실하다는 장점을 가지고 있다(최정원 외, 2016).

Merton(1973) 기업의 자산가치, 자기자본가치, 부채가치 사이에는 다음과 같은 관계식이 성립한다고 하였다.

$$V_E = V_A * N(d_1) - e^{-r_T T} * X_T * N(d_2)$$

$$d_1 = \frac{\ln\left(\frac{V_A}{X_T}\right) + \left(r_T + \frac{\sigma_A^2}{2}\right) * T}{\sigma_A * \sqrt{T}}$$

$$d_2 = d_1 - \sigma_A * \sqrt{T}$$

$V_E$  : 해당시점 시가총액 (주가 \* 발행주식수)

$V_A$  : 해당시점 자산의 가치

$X_T$  : T 시점에서 만기가 되는 부채의 장부가치

$\sigma_A$  : 자산가치의 변동성

$r_t$  : 만기t 인 무위험이자율

T : 추정 기간( 해당 기간 안에 부도 확률 추정)

$N(\bullet)$  : 표준누적정규분포의 값



산식의 추정을 위해서는 자산의 변동성이 필요하지만 이것을 직접 구할 수 없다. 따라서 주식의 변동성은 시장 정보를 통하여 알 수 있으므로, KMV 모형의 정의에 따라 주식의 변동성과 자산의 변동성 사이에 다음과 같은 관계가 성립하는 것을 이용하여 자산의 변동성을 도출한다. 두 산식을 이용한 연립 방정식을 풀기 위해서는 수치적인 해를 반복적 시행착오의 조정 과정을 거쳐서 최적화 값을 찾아내야 한다.

$$\sigma_E = d_1 - \sigma_A \cdot \sqrt{T}$$

이와 같은 과정을 통하여 산출된 자산 변동성을 활용하여 부도 확률을 예측하기 위해서는 부도 거리 (Default to distance, 이후 D.D.)를 산출하여야 한다. D.D.를 추정하기 위한 식과 가정은 다음과 같다.

$$D.D.T = \frac{\ln\left(\frac{V_A}{X_T}\right) + \left(\mu - \frac{\sigma_A^2}{2}\right) \cdot T}{\sigma_A \cdot \sqrt{T}}$$

$\sigma_A$  : 해당기업 연간 자산의 변동성

$\mu$  : 연 평균 성장률<sup>9)</sup>

$V_A$  :  $V_E + V_D$  (자산 가치 = 자본 가치 + 총 부채)

$V_E$  : (자본가치 = 발행주식수 \* 해당시점 주가)

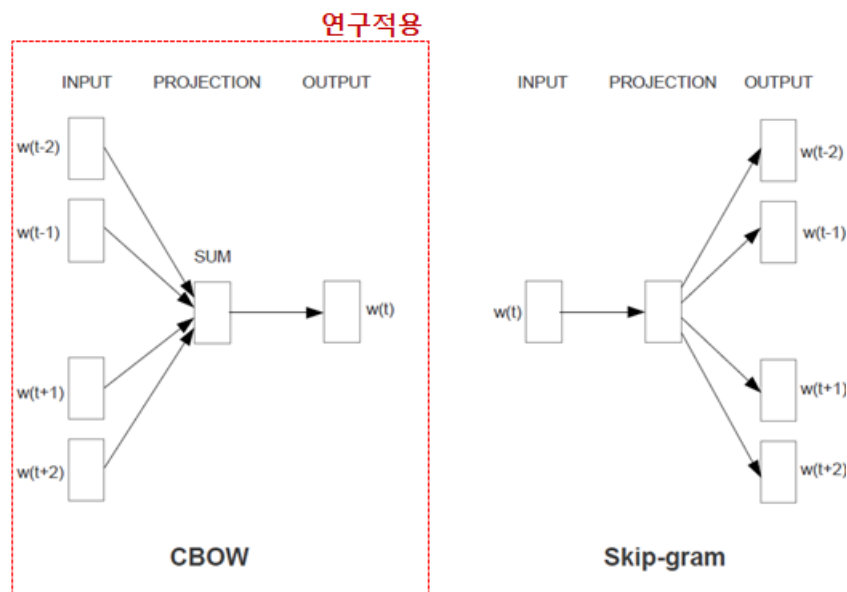
$X_T$  : T기간안에 만료되는 유동부채 잔액

### 2.2.3. 텍스트 계량화 방법론

텍스트 정보는 가장 대표적인 비정형 데이터로서 문서, 출판물, 웹 페이지, 메일 & 메시지 등 여러 가지 원천에서 확보할 수 있다. 또한 최근 발전하고 있는 음성 인식이나 영상 인식 기술과 결합할 경우 이러한 정보 원천의 범위는 더욱 확장 될 수 있다. 텍스트 데이터를 예측 모형 등에 활용하기 위해서는 계량화된 변수로 측정하는 과정을 필요로 한다. 본 연구에서 활용한 계량화 방법론은 다음과 같다.

### 2.2.3.1. Word2vec 활용

'Word2vec'은 단어들 간의 연관된 규칙을 찾아서 각 단어의 관계를 계량적으로 산출하는 방법론으로서, 각 단어 간의 앞 뒤 관계를 보고 근접도를 벡터의 형태로 계산하는 알고리즘이다. 'Word2vec'은 사전적으로 학습시키는 단계를 수행하지 않으므로 '비지도 학습' 기반의 인공지능(머신러닝)의 일종으로 볼 수 있다. 단어간의 관계에 대한 정확한 벡터를 산출하기 위해서는 분석 대상이 되는 대규모의 텍스트 데이터 문서(corpus) 데이터베이스를 필요로 한다.



'Word2vec'은 <그림 4> 와 같이 continuous bag-of-words (이하 CBOW) or continuous skip-gram(이하 Skip-gram) 두 가지 방법론이 있다. CBOW 는 여러 단어로부터 한 단어를 추정하는 방법으로서, 주로 주변 단어로부터 목적이 되는 한 개의 단어를 찾는 과정에 활용된다. CBOW 는 상대적으로 작은 Data-set 일 때도 효과적으로 동작하고 추정 속도도 빠른 것으로 알려져 있다. Skip-gram 은 한 개의 단어로 연관되는 여러 단어를 예측할 경우 활용한다. 예를 들어, 어떠한 단어가 현재 나타났을 때 향후 어떤 단어가 나타날 지를 추정하는 것을 목적으로 하는 경우 사용하게 된다.

본 연구에서는 'Word2vec' 방법론을 활용하여 뉴스 기사 내에 언급된 단어 간의 관계를 계량적으로 분석하여 연구 과정에 활용 하고자 한다. 기업의 부도 예측을 위해서는 부도와 연관된 기사가 보도되는 횟수, 비율 등을 파악하여야 하는데, 부도와 연관된 기사라고 해서 모든 기사에 반드시 '부도' (혹은 '상장폐지'. 이하 동일) 라는 단어가 포함되지는 않는다. 내용은 부도와 연관되어 있지만 '부도' 라는 단어 대신 다른 어휘를 사용한다거나 비슷한 느낌을 전달하는

단어를 선택할 수 있기 때문이다. 이 때 'Word2vec'을 활용하면 기사 중에 '부도' 단어와 유사한 의미로 사용되는 단어들이 유사도가 높게 산출되므로 다른 단어로 표현된 '부도' 기사를 판단할 수 있다.

### 2.2.3.2. 부도 관련 기사 비율 측정

기업이 부도가 실제로 발생하기 전부터 여러 가지 징후가 부도시점 이전부터 나타나게 된다. 이 때 기자들은 이러한 징후를 파악하여 부정적인 의견의 뉴스 기사를 작성하게 된다. 본 연구에서는 이러한 현상을 계량적으로 분석하기 위하여 '부도 관련 기사 비율'을 측정하고자 한다. 기간 별로 전체 기사 중 부도와 관련된 기사의 비중을 산출하고, 이 비율이 높게 나타날 경우 이를 사전적인 '부도'의 징후로 판단하여 부도 예측에 활용하는 것이다.

$$\text{부도 기사 비율}_{it} = \frac{\text{부도 관련 기사 수}}{\text{총 정상 기사 수}}, \quad i = \text{기업}, t = \text{분석 기간(월간, 부도발생 기준 직전 각 12 개월)}$$

부도 기사 비율 산출을 위해서는 부도 기사에 대한 정의를 필요로 한다. 이 과정에서 앞서 산출한 'Word2vec' 유사도를 측정하여 부도와 연관된 기사를 판별하는 과정에 활용할 수 있다. 예를 들어, '부도'와 특정 기준이상의 유사도를 나타내거나, 유사도 기준으로 순위(rank)를 부여하여 상위 단어들을 '부도 유사 단어'로 선정할 수 있다. 이후 선정된 부도 유사 단어 중 1 개라도 포함된 기사를 '부도 관련 기사'로 판별할 수 있다<sup>6</sup>.

**부도 유사 단어(1):** '부도' 단어와 'Word2vec' 유사도 상위 20 개 단어

**부도 유사 단어(2):** '부도'와 '상장폐지' 단어와 동시 'Word2vec' 유사도 상위 20 개 단어

최정원·한호선·이미영·안준모(2015)는 부도가 발생한 기업의 뉴스 텍스트 데이터를 텍스트마이닝 기법으로 분석하여 기업 부도예측의 가능성을 시도하였다. 부도발생 기업의 뉴스에서는 빈도 분석 결과, 정상 기업의 뉴스보다 유의적으로 많이 나타나는 주요 단어들을 도출할 수 있었으며, 이러한 단어가 부도와 연관성이 있음을 분석하였다. 본 연구는 이 연구가 제시하는 단어도

<sup>6</sup> 이러한 방식으로 산출할 경우 부도 유사단어와 부정 서술문이 결합된 경우(예: '부도가 발생하지 않았다')를 별도로 구별하기 어렵다는 단점이 있다. 하지만 부도를 부정하는 경우도 일단 부도와 연관성이 아주 낮은 상황이라 단정하기 어렵고, 분석 기사 수가 증가함에 따라 이러한 현상은 희석되므로 일단은 상관없이 분석을 진행하였다.

연구에 활용하여 부도 예측 여부를 판단하여 보았다.

**부도 유사 단어(3):** '회생', '공시', '자금', '횡령', '증자', '채권단', '워크아웃'

이러한 방식으로 각각의 도출된 [부도 유사 단어 (1)~(3)]을 기준으로 [부도 기사 비율 (1)~(3)]을 각각 추정할 수 있다.

### 2.2.3.3. 기사/기업/기간 단위 유사도 수준 측정

Word2vec<sup>7</sup>을 이용하면 기사를 구성하는 모든 단어<sup>7</sup>에 대하여 '부도' 단어와 유사도를 측정할 수 있으므로, 기사를 구성하고 있는 해당 단어들의 유사도 평균 값을 산출하면 해당 기사의 '부도' 단어와의 유사도 수준을 측정할 수 있다. 또한 기사 단위 유사도는 기업별, 기간별로 다시 평균 값을 산출 함으로서 특정 기간의 해당 기업에 대한 기사를 구성하고 있는 단어들의 '부도'와의 유사도 평균 수준을 산출할 수 있다. 이러한 방식으로 분석 대상이 되는 기업과 해당 기간에 대한 뉴스의 '부도' 와 유사도 수준을 측정하여 계량화된 변수를 산출할 수 있다.

**부도 유사도(1):** 특정월의 해당 기업의 기사를 구성하고 있는 모든 단어의 유사도 평균 수준

**부도 유사도(2):** 특정월의 해당 기업의 기사단위 유사도 평균 (단어 유사도 총합 / 기사 수)

<sup>7</sup> 모든 단어에 유사도를 부여하는 것이 가능하기는 하지만, 분석 resource(시간, 데이터 량 등)가 소모되는 수준에 비하여 분석의 실효성은 떨어진다. 따라서 모든 뉴스 기사를 취합한 기준으로 최소 200 회 이상 언급된 단어 5,335 개에 대해서만 유사도를 측정하여 분석에 활용하였다.

## 2.3. 모형의 예측력 평가 방안

### 2.3.1. 모형 예측력 평가 지표

앞서 설명한 여러 가지 방법론을 적용하여 기업 부도예측을 수행할 경우 모형의 성능을 비교하기 위해서는 동일한 개념으로 적용이 가능한 객관적인 모형 평가 방법이 필요하다. 예측 모형의 성능은 '구축된 모형이 얼마나 예측 분류에서 실제 분류와 똑같이 분류하는가?'가 모형의 평가의 핵심이 될 것이다. 즉, 본 연구의 기업 부도예측과 같은 이진 분류 예측의 상황은 두 범주(부도, 건실)간의 정확한 분류가 가능한지를 여러 모형 간에 비교하여 봄으로서 모형 평가를 수행할 수 있다(최정원 외, 2016). 예측 값과 실제 값 기준의 정확도의 산출 방법은 <표 4> 와 같다.

		예측 범주		합 계
		1	0	
실제 범주	1	$n_{11}$	$n_{10}$	$n_{1+}$
	0	$n_{01}$	$n_{00}$	$n_{0+}$
합 계		$n_{+1}$	$n_{+0}$	$n_{++}$

$$\text{정확도(Accuracy, 정분류율)} = (n_{11} + n_{00}) / n_{++}$$

$$\text{민감도(Sensitivity)} = n_{11} / n_{1+}$$

$$\text{특이도(Specificity)} = n_{00} / n_{0+}$$

<표 4> 이진분류 모형의 예측 정확도 지표 산출방법

기업예측 모형과 같은 이진 판별 예측은 할 때, 0 에서 1 사이까지 나타나는 추정 값들 사이에서 판별 값(Threshold)이 변함에 따라 민감도와 특이도를 포함한 정확도가 변동하게 된다. 이러한 판별 값 별로 변하는 민감도와 특이도 간의 관계를 그래프로 나타낸 것이 ROC(Receiver Operation Characteristic)곡선 그래프이다. ROC 곡선의 특성은 민감도와 특이도가 크면 클수록 좌상향으로 치우칠 것이며, 이와 같은 경우가 가장 정확도가 높은 수준으로 추정할 수 있다.<sup>8</sup> 본 연구에서는 각 예측 모형 추정결과의 ROC 를 모두 도출하여 판별 값과 상관없이 가장 정확도가

<sup>8</sup> 예측 목적에 따라 정확도가 아닌 민감도 혹은 특이도를 예측 모형 평가 지표로 활용하는 경우가 있다. 예를 들어 부도 기업예측 시, 부도(1) 인 기업을 부도(1)로 예측하는 것이 건실(0)기업을 건실(0) 기업으로 예측하는 것보다 중요하다고 생각한다면 정확도 보다는 민감도를 평가 기준으로 삼아야 한다. 이러한 가정은 보통 부실기업의 sample 수가 현저하게 작아서 정확도로 예측 모형의 성능을 정확하게 평가하기 어려운 경우 사용한다. 본 연구는 Test set 구성 시, 부도(1)와 건실(1) 비중을 50%:50% 균형 sample 로 설정하여 분석하므로 정확도를 모형 예측의 평가 지표로 설정하였다.

높은 수준을 각 모형의 예측 수준으로 평가 하였다.<sup>9</sup>

### 2.3.2. 모형 평가 강건성 증대 방안

만약 예측 모형 도출하여 모형의 예측력을 평가하는 과정에서 모형 도출 과정에서 활용한 학습(training) 데이터를 상기 평가 방법과 같은 예측력 평가로 적용하면 상당히 우수한 예측력이 나올 가능성이 높다. 이는 과잉 적합과 함께 대표적으로 인공지능과 같은 귀납적 추론 과정에서 흔히 나타나는 오류이다.

이를 방지하기 위해서는 Sample data 를 학습 세트(training set)와 평가 세트(test set)으로 나누어 예측 정확도(Accuracy)를 산출하고 이를 근거로 모형의 성능을 평가하여야 한다. 본 연구도 학습 세트와 평가 세트를 전체 표본 중 중복되지 않도록 70% 대 30%의 비중으로 배분하여 모형의 추정과 예측력 평가 과정에 각각 사용하여 이와 같은 오류를 최소화 하고자 하였다.

한편, 그 동안의 연구에서는 부도 기업의 표본(sample) 수가 정상 기업에 비하여 매우 작은 경우가 많이 나타나기 때문에 꾸준히 표본의 불균형에 의한 모형 예측력 평가의 어려움이 있음을 한계로 지적하여 왔다<sup>10</sup>. 이에 본 연구는 부도 기업의 표본은 고정하고 건설 기업의 표본을 부도 기업 수만큼만 Random 형태로 Sampling 하여 균형(equal-weighted, 50% 대 50%) 표본을 구성하여 모형의 추정과 평가에 활용하는 방안을 적용하였다. 다만 이러한 방식을 사용할 경우 정상 기업 표본에서 표본 선택에 따른 편의(bias)가 발생할 수 있으므로, 평가 과정의 강건성을 얻기 위하여 정상 기업 표본을 반복적으로 총 100 세트(set)를 임의 확률(random)로 구성하여 모형 평가 과정에 활용하였다. 따라서 각 방법론의 예측 수준 평가를 위한 정확도 값은 모든 평가 세트(100 set)에서 산출된 정확도의 평균 수준으로 산출하였다.

<sup>9</sup> 이론적으로는 판별값은 0.5 수준을 설정하는 것이 맞으나 모형 및 데이터에 특성에 따라 판별 값이 많이 달라진다. 아직까지 확실하게 이론적으로 판별값을 지정하는 방법론이 확립되지 않아 대부분의 데이터 마이닝 Concept 의 연구는 본 연구와 같이 귀납적으로 판별값을 설정하여 예측 결과를 산출하고 있다.

<sup>10</sup> 예를 들어, 정상기업과 부도기업의 비중이 90%:10% 라면, 모두 정상 기업으로 판단하는 예측을 수행해도 예측 정확도가 0.9로 나타난다. 따라서 편중이 심한 표본은 항상 예측 모형의 정확도를 과대하게 평가할 수 있는 우려가 있다. 최정원 외(2016)은 이와 같은 문제점을 해결하는 방안으로 각 부도 기업별로 동일한 시장(코스피/코스닥), 유사한 산업, 유사한 재무 수준의 정상 기업을 1:1로 짝지어(mapping) 분석하는 방법을 적용하였다. 하지만 이 방법은 객관적인 기준으로 유사한 기업을 찾기가 쉽지 않아서 분석자가 임의적으로 대상을 선정하는 경우가 많이 발생한다. 이러한 편의(bias)를 줄이고자 본 연구에서는 임의확률(Random)을 이용하여 균등표본을 설계하는 방안을 적용하였다.

### 3. 실증분석

#### 3.1. 부도 사건의 정의

증권거래소, 법원 등에서 상거래상 기업의 현황을 정의하여야 하는 공적인 기관에서는 공식적인 부도를 정의하고 있다. 하지만 실제로 기업의 부도를 인식하는 기준은 분석하는 목적과 연구자에 따라 기준이 다를 수 있다. 또한 실제로 이미 기업의 실질적인 부실이 발생하고 상당한 기간이 소요된 후 부도가 공식적으로 인식되는 경우도 많이 발생하게 된다. 따라서 기업 부도예측 연구 과정에서 보다 유용한 결과를 얻기 위해서는 기업의 부도(부실)에 대한 명확한 정의를 하는 것이 매우 중요하다.

본 연구는 이인로·김동철(2015), 최정원·오세경(2016) 등의 선행연구와 같이 유가증권시장에서 '상장폐지'가 결정된 기업들 중 부도에 관련된 공시<sup>11</sup>가 발생한 기업들을 부도 발생기업으로 인식하고 분석을 진행하였다. 상장폐지 사건은 부도와 반드시 연결된다고 볼 수는 없으나 시장에서는 부도와 관련된 이유로 상장폐지가 발생한 대부분의 기업은 특수한 상황을 제외하고 부도가 발생하거나 부도에 준하는 재무상황이 발생하여 타 투자자에게 지분이 인수된다. 또한 부도가 발생하지 않더라도 상장폐지 사건은 거래 정지 및 주가 하락이 발생하여 투자자와 채권자가 큰 손실을 입을 수 있는 사건이므로 상장폐지를 부도로 인식하는 것은 보다 보수적인 기준에서 부도를 적절하게 평가하는 방법이라고 할 수 있다.

#### 3.2. 데이터 수집 및 정제

##### 3.2.1. 분석 대상 기업 정의

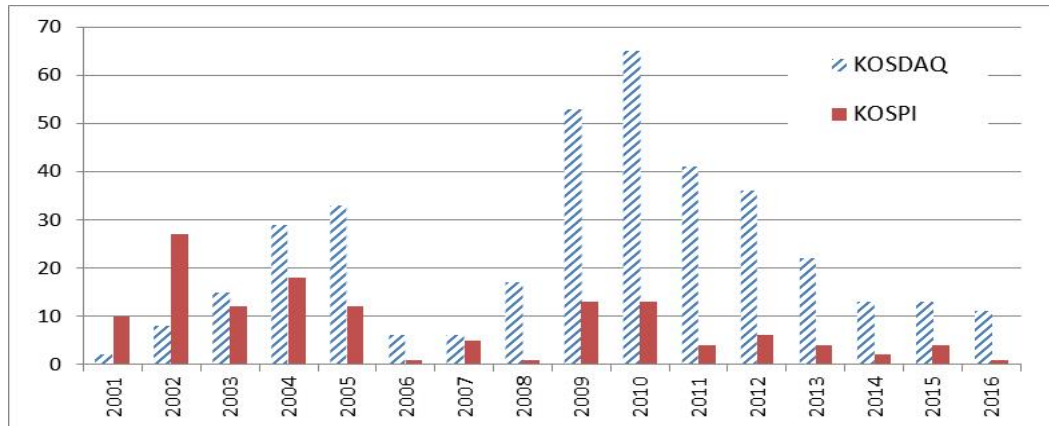
2001년부터 2015년 까지 상기 부도 정의에 따라 유가증권 시장에 상장된 기업을 대상으로 분석 대상이 되는 부도 기업과 건실 기업을 집계하면 <표 5>와 같다.

부도 기업은 상대적으로 KOSDAQ 시장에서 많이 발생하였다. KOSPI 시장의 경우 IMF 위기 이후 기간인 2002년 전후, KOSDAQ 시장의 경우 2008년 글로벌 경제위기 이후 기간인 2009년에서 2011년 사이에 집중적으로 부도기업이 발생한 것을 확인할 수 있다.

<sup>11</sup> '부도발생', '화의절차개시신청', '회사정리절차개시신청', '감사인의 의견 거절' 및 '은행거래정지' 등의 기업의 부실 및 지속 가능성이 심각하게 의심되는 사유로 발생한 상장폐지 사건을 부도로 정의하였다. 반면, '신규/변경 상장', '특수 목적에 의한 상장폐지', '기업 피인수' 등 원인의 상장폐지 공시는 부도 사건과 상관없는 공시로 정의하여 분석대상에서 제외하였다.

시장구분	건설기업	부도기업	Total
KOSPI	678	133	811
KOSDAQ	1108	370	1478
<b>Total</b>	<b>1786</b>	<b>503</b>	<b>2289</b>

&lt;표 5&gt; 분석대상 기업



&lt;그림 5&gt; 연도별 부도기업 추이

### 3.2.2 텍스트 데이터 수집

비정형 정보엔 뉴스 텍스트 데이터 수집을 위하여, 네이버 뉴스 검색 홈페이지를 활용하여, 분석 대상 기업들에 대한 2010년 1월부터 2016년 12월까지의 84기간의 뉴스 콘텐츠를 수집하였다.<sup>12</sup>

텍스트 DB를 구축하기 이전 분석 대상 기업의 전체 기사 수를 먼저 집계하여 다음과 같은 분석대상 제외 조건을 만족하는 총 650건(비부도기업 273건, 부도기업 377건)의 경우를 제외하고 진행하였다.

- (a) 2010년 전의 부도가 일어난 기업: 기사를 확보할 수 없음
- (b) Sample 수 부족: 분석대상기간(2010년 ~ 2016년) 동안 기사 수 100건 이하
- (c) 기업의 이름이 일상적인 용어와 같은 경우 (Ex: 전방, 청구, 부흥, 진도 등)
- (d) 기타 해당 기업의 기사인지 정확하게 확인할 수 없는 기업

<sup>12</sup> R 프로그램을 사용하였으며 N2H4 패키지를 사용하였다.



제외 후 텍스트 정보 수집 대상 기업은 총 1,788 개의 기업으로 총 2,506,080 건의 기사를 텍스트 DB 로 확보하였다. 기업 당 평균적으로 약 1,401 건의 기사를 수집하였고, 1 개월 당 평균적으로 약 16.6 건의 기사이다. 또한 년도 별로 기사 수 추이는 <표 6>과 같다.

구분	Total	2010	2011	2012	2013	2014	2015	2016
기사 수	2,506,080	110,213	339,040	390,764	394,128	402,792	426,991	442,152
기업당 평균	1,402	62	190	219	220	225	239	247

<표 6> 총 뉴스 기사 수 연간 추이 및 합계

텍스트 DB 는 이후 자연어 처리 과정(Natural Language Processing , NLP)을 진행하였다.<sup>13</sup> 기사 수와 마찬가지로 총 집계 200 개 이하의 키워드는 분석에서 제외하였고, 동의어는 의미상의 대표 단어로 변환하여 활용하였다. 또한 특정의미(회사명, 제품명, 인물명, 지명, 일자, 시간)명사는 제외하였다.

### 3.2.3. 부도 기사 비율 및 부도 유사도 산출 결과

앞서 3.2.3. 에서 설계한 방법론을 토대로 수집된 텍스트 데이터를 계량화하여 기업 부도 예측 모형에서 활용할 수 있도록 변수화 하는 과정을 수행하였다.

#### 3.2.3.1. Word2vec 산출 결과

수집된 텍스트 데이터베이스를 기반으로 '부도' 및 '상장폐지'와 기사 내에 언급된 단어 간의 유사도를 'Word2vec'을 이용하여 산출할 수 있다. 다음은 유사도 기준 상위 20 개 단어를 선별한 결과이다.

<sup>13</sup> 자연어 처리는 R program 의 KoNLP Package 를 사용하였으며, 자연어 처리의 성능 향상을 위해 한국정보화진흥원에서 개발한 형태소 사전을 이용하였다. 자연어 처리 외에도 도출 빈도수가 높은 키워드 중에 특정한 의미의 동의어, 불용어 등은 분석자가 직접 지정하여 처리하여야 한다.

Rank	'부도' 기준		'부도' & '상장폐지' 기준	
	word	유사도	word	유사도
1	도산	0.74	퇴출	0.63
2	파산	0.63	관리종목	0.62
3	경영난	0.60	파산	0.62
4	외환	0.60	도산	0.61
5	자금난	0.60	분식회계	0.60
6	법정관리	0.57	법정관리	0.57
7	어음	0.57	원리금	0.56
8	연체	0.55	잠식	0.56
9	워크아웃	0.54	연체	0.55
10	대출금	0.53	자금난	0.55
11	원리금	0.53	손실	0.54
12	폐업	0.53	매매거래	0.53
13	부실화	0.53	워크아웃	0.53
14	부실	0.52	부실	0.53
15	채무	0.50	기업회생	0.52
16	손실	0.49	감사보고서	0.52
17	몰락	0.48	대출금	0.52
18	제때	0.48	회생	0.52
19	기업회생	0.48	부실기업	0.51
20	속출	0.47	정지	0.51

&lt;표 7&gt; 'Word2vec' 유사도 산출 결과

'부도'와 '상장폐지' 는 두 단어 간에도 유사도가 존재하기 때문에 일부 단어의 경우 중복하여 나타나는 것을 확인할 수 있다.

### 3.2.3.2. 부도 기사 비율 산출 결과

<표 7> 의 'Word2Vec' 기준으로 부도 연관 기사를 산출한 결과를 요약하면 <표 8> 과 같다.

구분		Total	2010	2011	2012	2013	2014	2015	2016
부도연관 단어(1)	부도연관 기사 수	380,673	16,586	48,636	59,214	65,863	60,729	59,473	70,172
	부도기사 비율(1) 평균	15.19%	15.05%	14.35%	15.15%	16.71%	15.08%	13.93%	15.87%

부도연관 단어(2)	부도연관 기사 수	389,952	14,496	46,398	59,157	69,142	64,457	61,718	74,584
	부도기사 비율(2) 평균	15.56%	13.15%	13.69%	15.14%	17.54%	16.00%	14.45%	16.87%
부도연관 단어(3)	부도연관 기사 수	221,523	11,616	29,269	31,948	37,553	35,198	35,176	40,763
	부도기사 비율(3) 평균	8.84%	10.54%	8.63%	8.18%	9.53%	8.74%	8.24%	9.22%

&lt;표 8&gt; 부도연관기사 및 부도기사비율 연간 추이

### 3.2.3.3. 부도 유사도 산출 결과

<표 7> 의 'Word2Vec' 기준으로 부도 연관 기사를 산출한 결과를 요약하면 <표 9> 과 같다.

구분	Total	2010	2011	2012	2013	2014	2015	2016
부도유사도(1) (‘부도’)	0.0206	0.0124	0.0216	0.0276	0.0279	0.0296	0.0247	0.0206
부도유사도(2) (‘부도’ & ‘상장폐지’)	0.0546	0.0309	0.0609	0.0730	0.0728	0.0749	0.0695	0.0546

&lt;표 9&gt; 부도연관기사 및 부도기사비율 연간 추이

### 3.2.4. 데이터 수집 결과 요약 및 데이터 세트(set) 적용 방안

정보 원천 별로 모형 예측의 영향을 평가하기 위하여 취합된 분석 DB 를 4 가지의 데이터 세트로 분류하여 각각의 모형에 적용하고자 한다. 분류된 데이터 세트의 구성은 <표 10> 와 같다. 데이터 세트는 기존 연구에서 활용도가 높았던 순서대로 재무 정보, 시장 정보, 거시경제 정보, 비정형 정보 순으로 점진적으로 반영하는 정보가 늘어나는 형태로 설계하였다.

방법론	Set 1	Set 2	Set 3	Set 4
적용 정보 (Source)	재무 정보	재무 정보 + 거시경제	재무 정보 + 거시경제 + (증권)시장 정보	재무 정보 + 거시경제 + (증권)시장 정보 + 미디어정보(Text)
데이터 수집가능기간	1998~2015년 (연간)	1998~2015년 (연간)	1998~2015년 (연간/월간)	2010~2015년 (연간/월간)
변수 정보	31개 변수 (21개 재무변수 +10개 기업특성)	42개 변수 (Set 1 + 거시 11개)	49개 변수 (Set 2 + 시장 7개)	53개 변수 (Set 3 + 미디어 4개)
이용가능 데이터 수	결측제외 총 33621 개 ->2291기업(부도 502개)	결측제외 총 30268 개 ->2291기업(부도 502)	결측제외 총 21402 개 ->2291기업(부도 502)	결측제외 총 9706 개 ->1586기업(부도 258)

&lt;표 10&gt; 모형 적용 데이터 세트 요약

재무 정보의 경우 부도 발생 전기 ( $t-1$  시점) 보다 이전부터 재무지표가 악화되어 부도에 영향을 줄 가능성이 있으므로 총 부도 발생 직전 3 기간( $t-1$ ,  $t-2$ ,  $t-3$ ) 기간의 재무 정보를 사용하여 예측모형을 산출하였다.

한편, 뉴스 텍스트 정보는 인터넷으로 뉴스 기사 수집이 가능한 시점인 2010 년 이후의 정보만 활용이 가능하다.<sup>14</sup> 따라서 분석 과정에서는 이러한 세트별 기간의 불일치를 고려하여 분석을 하여야 한다. 본 연구는 활용 가능한 데이터 수준에 따라 2 가지 분석 기준을 추가로 고려하였다. 따라서 <표 10>에서 구분한 정보 기준과 결합하면 총 7 개의 분석 Set 가 구성되었으며, 각 방법론에 모든 Set 를 반영하여 각각의 예측모형을 산출하고 상호 간의 비교 분석을 수행하였다.

- 1) **Set A:** 재무, 시장, 거시경제 정보(2001~2016 년). 총 2291 개 (부도 502 개) 기업 대상  
[SetA\_1] / [SetA\_2] / [SetA\_3]
- 2) **Set B:** 재무, 시장, 거시경제 정보(2010~2016 년). 총 1586 개 (부도 258 개) 기업 대상  
[SetB\_1] / [SetB\_2] / [SetB\_3] / [SetB\_4]

<sup>14</sup> 크롤링 Source 인 '네이버뉴스' 웹 페이지가 2010 년 이전 뉴스를 제공하지 않는다. 기타 Source 를 활용할 경우 분석 기간에 대한 확장이 가능하다.

### 3.3. 연간 예측 모형

재무정보를 포함하는 기업 부도 예측 모형은 연간 단위로 예측을 수행하여야 한다. 부도 여부(1: 부도, 0: 정상)를 목표(Target) 변수로 하여 각 방법론을 활용하여 예측 모형을 구성하였다.

#### 3.3.1 방법론별 최적 예측 모형 도출

상기 과정을 통하여 생성된 분석 DB 를 각 학습 세트(training set)를 기반으로 적용하여 <표 11>와 같은 방법론을 적용하여 모형을 적합(fitting)하고 최적 모형을 도출하였다.

방법론	세부 적용 방법론 및 가정	산출(fitting) 및 모형 평가 방법
<b>1. 로지스틱 (Logit)</b>	다중회귀분석 모형 (Stepwise) Engine: R (glm)	<ul style="list-style-type: none"> <li>• Cross-section 형태의 분석 방법이므로 시점 별(t-1,2,3) 변수를 모두 설명변수로 각각 적용</li> <li>• 변수가 많아 과다 적합 문제 발생 가능 → Stepwise 로 변수 선택 적용</li> <li>• F-value(P-value) 및 <math>R^2</math> 로 모형 평가</li> </ul>
<b>2. Cox-PH Hazard (Cox)</b>	Cox PH 모형(다중회귀, 층화, Stepwise) Engine: R (survival)	<ul style="list-style-type: none"> <li>• 추가, 거시경제, 비정형정보 등 Hazard 함수 설명 변수로 반영 가능</li> <li>• 산업별 생존함수를 추정하여 산업별 특성 반영</li> <li>• 변수 선택(Stepwise) 필요</li> <li>• F-value(P-value) 및 <math>R^2</math> 로 모형 평가</li> </ul>
<b>3. Decision Tree (Dtree)</b>	Max maxsurrogate(노드 수): 3 단계 Engine: R (Dtree)	<ul style="list-style-type: none"> <li>• 비교 모형으로 활용</li> <li>• Accuracy 로 사후적 모형평가</li> </ul>
<b>4. Random-Forest (RF)</b>	Sampling 을 통한 parameter 최적화 Engine: R (e1071)	<ul style="list-style-type: none"> <li>• 다양한 설정 값 시뮬레이션</li> <li>• Accuracy 로 사후적 모형평가</li> </ul>
<b>5. SVM</b>	Sampling 을 통한 parameter 최적화 Engine: R (e1071)	<ul style="list-style-type: none"> <li>• 다양한 설정 값 시뮬레이션</li> <li>• Accuracy 로 사후적 모형평가</li> </ul>
<b>6. 인공신경망 (DNN)</b>	Deep 구조: 512 EU * 8 Layer Activation Function: ReLU 초기값 설정: Xavier initializer <sup>15</sup> Engine: Python (TensorFlow)	<ul style="list-style-type: none"> <li>• Cost 함수(평균예측오차): <math>\frac{(\text{실제값}-\text{예측값})}{\text{평가 횟수}}</math></li> <li>→ 학습횟수 2 만 or Cost 기준 0.1 이하 까지</li> </ul>
<b>7. 인공신경망 (RNN)</b>	Deep 구조: 3 기간(LSTM Cell) 적용 Activation Function: ReLU 초기값 설정: Xavier initializer Engine: Python (TensorFlow)	<ul style="list-style-type: none"> <li>• Cost 함수(평균예측오차): <math>\frac{(\text{실제값}-\text{예측값})}{\text{평가 횟수}}</math></li> <li>→ 학습횟수 2 만 or Cost 기준 0.1 이하 까지</li> </ul>

<표 11> 각 모형의 세부 적용 방안 및 산출 모형 적합도 평가 방법

<sup>15</sup> 딥러닝 초기값에 대한 방법은 Glorot, X., Y. Bengio(2010)을 참고하였다.

### 3.3.2 예측 모형 성과 분석

#### 3.3.2.1. SET A 결과 (분석기간 2001 년~2016 년 적용)

각 기간별 데이터 세트와 예측 모형 추정 방법론을 적용한 예측 수준(정확도) 산출 결과는 <표 12> 과 같다. 가장 높은 정확도를 나타낸 방법론은 Random Forests 방법론이었다. 로지스틱 모형과 SVM 또한 0.9 에 상회하는 높은 정확도가 산출되었다. 그 외에 의사결정나무(Dtree)와 인공신경망(DNN, RNN) 등은 0.9 에 다소 못 미치는 정확도를 산출하였다. 기업의 재무정보, 거시 경제정보, 시장정보를 포괄하여 가장 정보가 많이 활용된 <SET3>의 정확도는 타 데이터 세트에 비하여 다소 높게 산출되긴 하였지만 유의미한 수준은 아니다.

방법론	SET A_1	SET A_2	SET A_3	평균
<b>logit</b>	0.9258 0.0146	0.9208 0.0153	0.9272 0.0142	<b>0.9246</b>
<b>Cox</b>	0.7798 0.0183	0.7033 0.0237	0.7115 0.0199	<b>0.7315</b>
<b>Dtree</b>	0.8998 0.0183	0.8984 0.0179	0.8956 0.0180	<b>0.8979</b>
<b>R.F</b>	0.9357 0.0133	0.9350 0.0127	0.9381 0.0125	<b>0.9363</b>
<b>SVM</b>	0.9217 0.0153	0.9082 0.0179	0.9212 0.0226	<b>0.9170</b>
<b>DNN</b>	0.8533 0.0200	0.8584 0.0184	0.9052 0.0148	<b>0.8723</b>
<b>RNN</b>	0.8867 0.0210	0.9065 0.0232	0.9046 0.0279	<b>0.8992</b>
<b>평균</b>	<b>0.8861</b>	<b>0.8758</b>	<b>0.8862</b>	

<표 12> 모형별 예측 정확도 산출 결과(SET A)<sup>16</sup>

#### 3.3.2.2. SET B 결과 (분석기간 2010 년~2016 년 적용)

<표 13>은 2010 년~2015 년 까지 데이터를 적용(SET B)하여 각 모형 예측 정확도를 산출한 결과이다.<sup>17</sup> 이 분석 결과에서도 역시 Random Forests 방법론이 가장 우수한 예측력을 보였고, SVM, 인공신경망(DNN) 순으로 예측력이 좋았다. 앞서 예측력 수준이 높았던 로지스틱 모형은

<sup>16</sup> 정확도는 총 100 회 Sample 세트 별 예측 정확도의 평균값이고, ()안은 표준편차이다. (<표 11> 동일)

<sup>17</sup> 인공신경망(RNN) 의 경우 분석 과정에 3 개년 연속된 데이터가 필요한 데, 이럴 경우 Set B 는 Data Sample 수의 손실이 너무 심해서 유효한 분석이 어렵다. 따라서 <Set B> 분석에서는 인공신경망-RNN 은 제외하고 분석하였다.

상대적으로 모형 예측력이 하락하였으나 인공지능 기법들의 예측력은 유지되거나 오히려 다소 상승하였다. 이는 데이터가 줄어드는 경우에도 인공지능 예측 방법론들이 상대적으로 모형 예측력이 강건하게 유지될 수 있음을 의미한다.

또한 기존의 <SET B\_3>에 뉴스 텍스트 정보까지 추가로 반영된 <SET B\_4>가 타 모형에 비하여 모형 예측력이 높게 나타났다. 이는 비정형 정보도 부도예측 성능 향상에 영향을 줄 수 있음을 실증하는 결과이다. 다만 역시 평균과 표준편차 수준으로 볼 때 정보를 미반영한 SET와 차이가 통계적으로 유의한 수준이라 보기는 어렵다.

방법론	SET B_1	SET B_2	SET B_3	SET B_4	평균
<b>logit</b>	0.8651 0.0427	0.8804 0.0410	0.8989 0.0383	0.9093 0.0338	<b>0.8884</b>
<b>Cox</b>	0.8280 0.0312	0.8235 0.0335	0.8473 0.0335	0.8745 0.0282	<b>0.8433</b>
<b>Dtree</b>	0.8910 0.0293	0.8895 0.0288	0.8868 0.0274	0.8862 0.0271	<b>0.8884</b>
<b>R.F</b>	0.9369 0.0224	0.9373 0.0226	0.9381 0.0225	0.9392 0.0222	<b>0.9379</b>
<b>SVM</b>	0.9217 0.0273	0.9148 0.0263	0.9271 0.0278	0.9178 0.0282	<b>0.9203</b>
<b>DNN</b>	0.9071 0.0285	0.9053 0.0282	0.9215 0.0286	0.9317 0.0299	<b>0.9164</b>
<b>평균</b>	<b>0.8916</b>	<b>0.8918</b>	<b>0.9033</b>	<b>0.9098</b>	

<표 13> 모형별 예측 정확도 산출 결과(SET B)

### 3.3.2.3. 연간 모형 예측 결과 종합 해석

분석결과 인공지능 중 Random Forests 방법론이 두 데이터 SET 모두 가장 높은 수준의 예측력을 보여주었다. 특히 데이터 수가 상대적으로 적은 <SET B>에서도 우수한 예측력을 유지함으로서 인공지능 기법이 강건하게 기업의 부도에 대한 예측을 잘 수행할 수 있음을 실증하는 결과이다.

한편, 현재 적용된 인공지능-DNN 체계의 은닉층 구조는 1 열 8 개층(layer) 중첩 구조이고, RNN 은 3 기간 10 개층(layer) 구조이다. 컴퓨터 하드웨어를 보강하고 추가적인 효율화 방안을 도입하여 이러한 구조를 개선하면 현재보다 더 높은 예측 정확도를 얻을 가능성이 있다.

인공지능(DNN)을 적용한 결과를 보면 Sample 데이터 수가 많은 <SET A>에 비하여 <SET B>의 예측 정확도가 오히려 높게 나오는 현상이 발생하였다. 이 역전 현상은 과잉 적합(over-

fitting)하여 오히려 예측력이 떨어지는 현상이 나타난 것으로 추정된다. 따라서 향후 변수간의 관계를 고려하여 일부 변수를 정리하거나 과잉적합을 해결할 수 있는 추가적인 방법론을 적용해준다면, 인공지능 기법의 예측 정확도 결과는 현재보다 높아질 수 있다.

한편, 텍스트 데이터를 추가로 반영한 <SET B\_4>의 예측 정확도는 방법론에 따라 약간의 차이는 있지만 전반적으로 텍스트 데이터를 반영하지 않은 SET 에 비하여 정확도 수준의 유의한 차이가 나타나지 않았다. 또한 재무정보만 활용한 <SET A\_1>, <SET B\_1>의 예측력도 타 SET 에 비하여 큰 차이가 없었다. 이는 상장 기업의 경우 다양한 공시 요구 및 규제에 의하여 기업의 정보가 재무정보에 이미 충분히 반영되어 나타나는 결과라 판단된다.

### 3.4. 월간 예측 모형

미디어의 뉴스 기사는 시장 정보(주가)와 마찬가지로 실시간으로 공개되는 정보이다. 따라서 시장정보를 활용한 예측 모형인 KMV 모형과 유사한 형태의 부도예측 모형 구축이 가능하다. 본 연구는 기업의 부도 관련 뉴스가 실제 부도가 발생하는 시점 이전에 부도 가능성을 선제적으로 알려줄 수 있는지, 조기 경보 지표(early warning index)로서 활용 가치가 있는지 연구하였다.

#### 3.4.1. 예측 모형 설계

먼저 예측 모형 추정의 대상이 되는 Sample 데이터를 정의한다. 해당 기업의 기사 수가 너무 적은 경우 1,2 건의 부도 관련 기사로 인하여 과민한 예측 결과가 발생할 수 있다. 따라서 신뢰성 있는 모형 결과를 위하여 일정 건수 이상의 기사가 확보된 기업을 대상으로 예측 모형을 산출하였다.

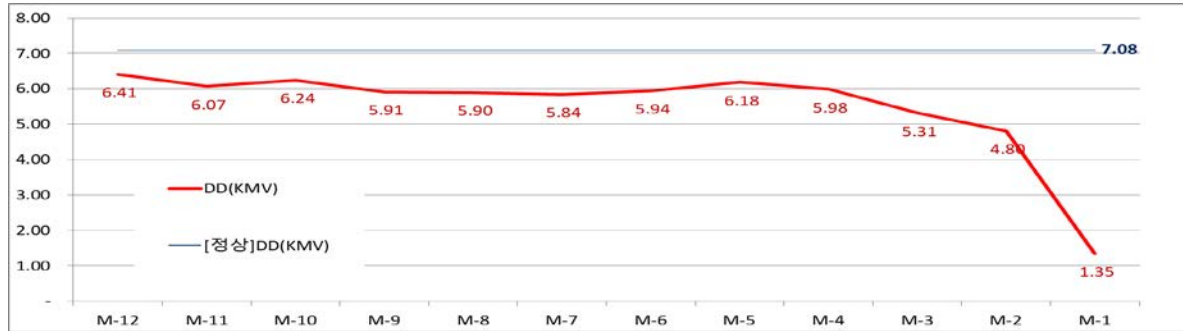
- 1) 대상기간: 2010 년~2016 년 (텍스트 DB 확보 가능 기간)
- 2) 기사 수 기준: 대상기간 동안 총 기사 수 합계 100 건 이상
- 3) 정보 확보 여부: 대상 기간 동안 주가, 시가총액, 재무정보 모두 확보 가능한 기업  
(신생기업, 중도 이탈 기업 제외. 부도 기업은 부도(상장폐지) 이전 시점 까지만 해당)

상기 기준을 적용하여 기존에 확보한 데이터를 대상으로 선별한 결과 부도 기업 52 개, 정상기업 855 개를 Sample 분석 대상으로 확보하였다. KMV 모형 및 텍스트 기반 모형의 부도예측 단위는 월간이며, 부도 기준 직전 12 개월의 추이를 분석하였다.



### 3.4.2. KMV 모형 산출 결과

부도기업의 부도발생 전 12 개월의 D.D. 의 평균 수준 추이는 <그림 6> 와 같다.

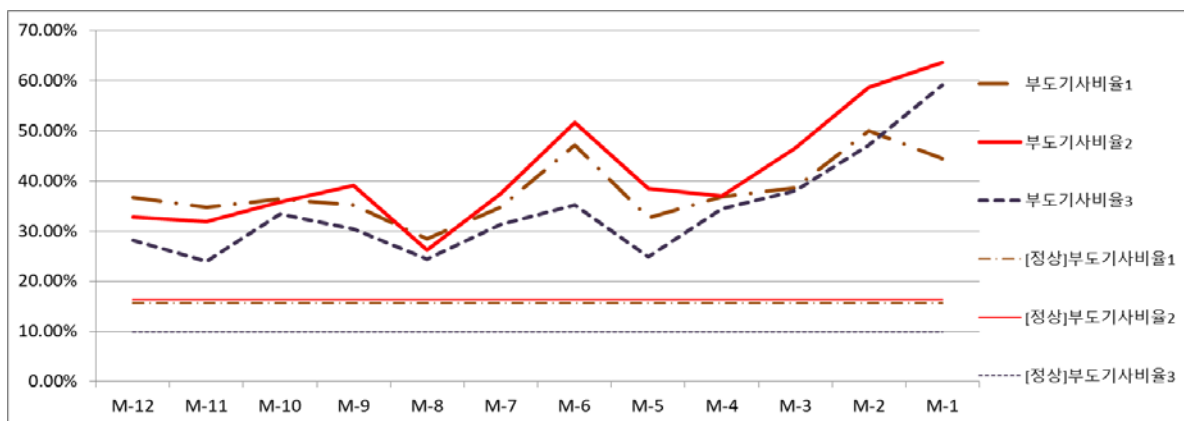


<그림 6> 부도 발생 12 개월 전 D.D. 평균 추이<sup>18</sup>

부도 기업의 경우 부도 발생 1 년전부터 점진적으로 평균 수준에 비하여 다소 낮은 수준으로 D.D. 가 하락하다가, 부도 발생 3 개월 전부터 급격하게 하락하는 것을 확인할 수 있다.

### 3.4.3. 텍스트 정보기반 예측 모형 산출 결과

<그림 7>은 동일한 기간과 동일한 기업에 대하여 기사 텍스트 데이터베이스를 기반으로 산출한 부도 기사 비율 및 부도 유사도를 적용하여 도식한 결과이다.

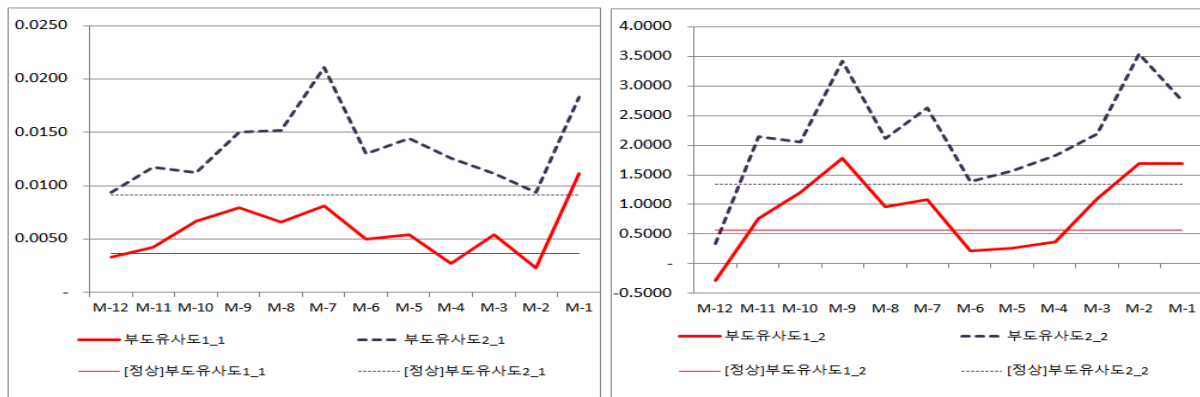


<그림 7> 부도 발생 12 개월 전 부도 기사 비율 추이

<sup>18</sup> 정상기업은 부도시점을 설정할 수 없기 때문에 2014~2016 년 3 개년도 기간의 월간 수치 평균값을 사용하였다. 이후 텍스트 지표도 동일한 기준을 적용하였다.

KMV 모형과 마찬가지로 부도기사 비율은 부도 발생 12 개월 이전부터 점진적으로 상승하여 지속적으로 정상기업에 비하여 높은 수준으로 산출되는 것을 확인할 수 있다. 부도기사 비율 중에는 '부도'와 '상장폐지'를 동시에 'Word2vec'을 활용하여 상위 20 개 단어가 포함된 기사를 부도기사로 간주한 [부도기사 비율 2] 가 정상 수준에 대비하여 가장 유의한 차이를 보이고 있다.

<그림 8>은 '부도' 단어와의 'Word2vec' 유사도 수준의 산출 결과이다.



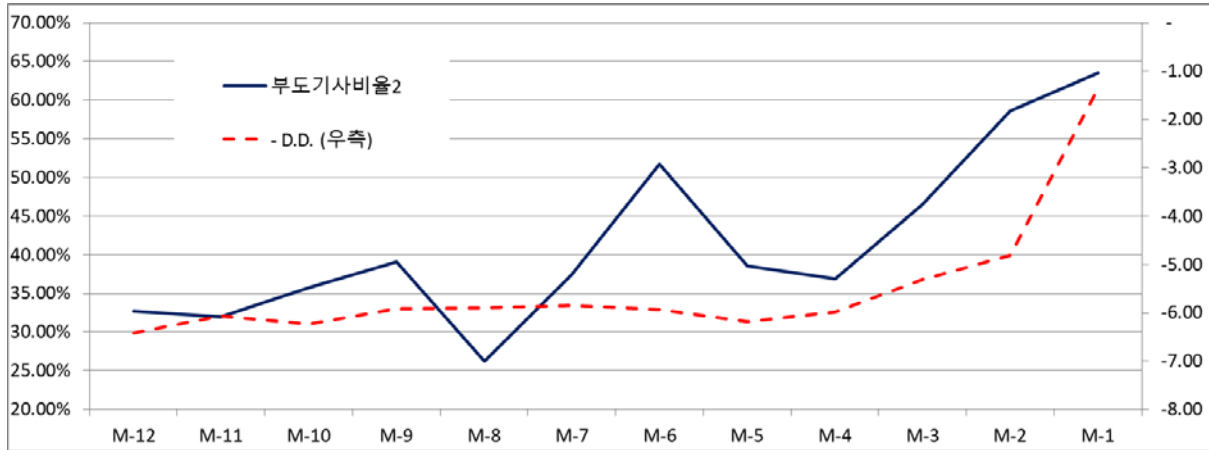
<그림 8> 부도 발생 12 개월 전 부도 유사도(평균, 기사단위 평균) 추이

부도 유사도 역시 부도 기사비율과 마찬가지로 부도발생 이전부터 정상기업과 차이가 나타난다. 다만, KMV 와 부도 기사비율과는 달리 점진적 상승 추세가 다소 약하고, 부도 시점에 가까워지면서 오히려 정상기업 보다 떨어지는 수준도 나타나는 것을 확인할 수 있다. 이는 기사수가 많아지면서 절대적인 단어수가 증가하여 부도 유사도가 높은 단어의 영향을 중화하는 현상이 발생한 것으로 파악되었다.

#### 3.4.4. KMV 와 텍스트 정보기반 예측 모형 비교

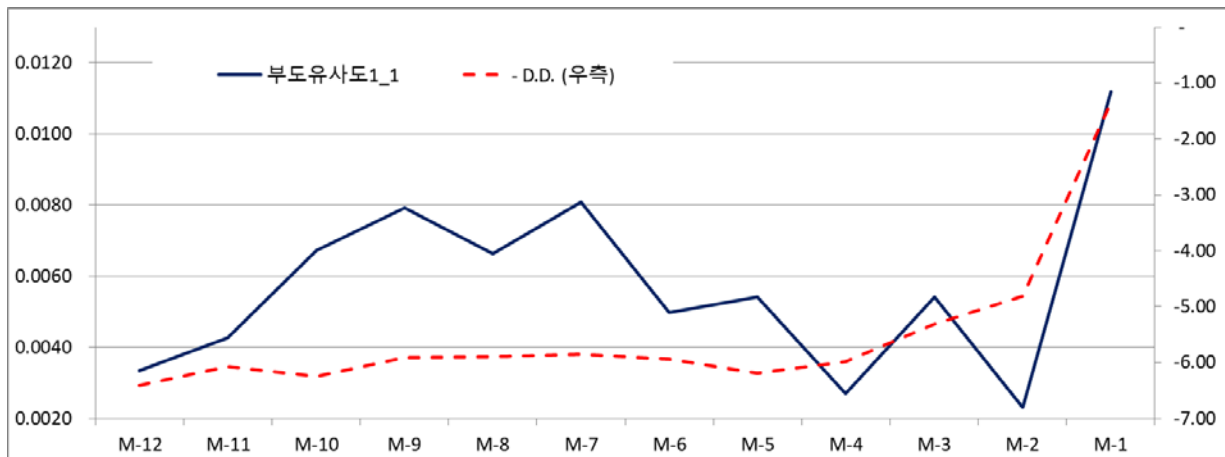
앞서 분석한 KMV 모형과 텍스트 정보기반의 예측 모형은 각각 부도 발생 이전 시점부터 부도 가능성이 상승함을 보여주는 것을 확인할 수 있었다. 이제 두 모형을 비교하기 위하여 함께 그래프로 도식화하여 비교하여 보았다.

부도 기사 비율은 KMV 모형의 결과인 D.D. 와 비슷한 형태로 부도 가능성에 대한 신호를 주고 있는 것을 볼 수 있다(그림 9). 특히 부도 발생 6 개월 이전 시점 부터는 지속적으로 KMV 모형보다 다소 높은 수준으로 부도 기사 비율이 나타난다.



<그림 9>. 부도 기사비율과 D.D.의 비교<sup>19</sup>

<그림 10>은 부도유사도 중 가장 예측수준이 높은 것으로 나타난 [부도기사비율 1\_1] 과 D.D. 수준 추이를 비교하여 보았다.



<그림 10>. 부도 유사도와 D.D.의 비교

기사의 부도 유사도 수준은 D.D. 에 비하여 부도 발생 7~10 개월 전에 매우 큰 차이를 보인다. 다만, 부도 발생 2~4 개월 기간은 D.D. 보다 낮은 수준으로 부도 가능성을 예측하고 있다. 이러한 현상은 부도 기업의 경우 실제 부도가 나타나기 오래 전부터 부도 관련 단어가 기사에서 많이 나타나는 현상을 실증한다.

<sup>19</sup> D.D. 는 부도로 부터의 거리로서 값이 작아질수록 부도 가능성이 증가하는 지표이고, 부도 기사비율은 값이 커질수록 부도 가능성이 높아지는 지표이다. 따라서 두 지표의 비교를 위하여 D.D. 는 음수로 표현하였다. <그림 x> 부도 유사도와 비교 그래프로 동일하다.

상기 분석 결과를 보면, 부도 기사비율과 부도 유사도를 활용할 경우 KMV 모형과 유사한 형태로 부도 예측이 가능함을 알 수 있다. 또한 부도 발생 시점을 기준으로 KMV 모형 보다 이전 기간에 부도 유사도가 상승하여 기업 부도에 대한 조기경보 지표로서 기사 정보를 이용한 텍스트 기반의 모형 결과가 활용될 수 있는 충분한 가능성을 보여주었다. 더욱이 텍스트 정보 기반의 부도예측은 주가 정보가 없는 비상장기업에도 활용이 가능하다는 점에서 KMV의 단점을 보완하는 방법론으로 더욱 의미가 있다.

텍스트 기반의 부도 예측 방법 또한 단점을 가지고 있다. 먼저 기업관련 뉴스의 편중 문제이다. 대부분의 기업 뉴스는 일부 매우 우량하고 유명한 대기업에 대한 기사가 많이 생성되고, 정작 부도가 많이 발생하는 규모가 작은 기업에 대한 뉴스는 상대적으로 매우 적다. 따라서 본 방법을 적용할 수 있는 분석 대상이 한계가 있었다. 향후 이를 보완하기 위해서는 텍스트 데이터 확보 정보 원천을 미디어 뉴스 뿐만 아니라 기업 공시자료, 증권/투자 관련 게시판, 해당기업 홈페이지 등으로 확대하여 보다 광범위한 텍스트 데이터의 확보가 필요할 것이다.

## 5. 결론 및 시사점

본 연구는 기업 부도 예측 과정에서 우선 비정형 데이터인 뉴스 텍스트 데이터를 계량화하여 새로운 정보 원천으로 활용할 수 있는 방법을 제시하였다. 또한 기존 정보 원천과 함께 텍스트 정보를 포함한 인공지능 기반의 예측 방법론을 제시하고 기존의 방법론과 예측력을 비교 분석하였다.

연구 결과, 우선 연간 모형에서는 인공지능 기법인 Random forests 기법이 가장 우수한 예측력이 나타나는 것으로 분석되었다. 또한 인공지능을 이용한 다른 방법론들도 전반적으로 기존의 전통적인 예측 방법보다 예측력이 우수한 것으로 나타났다. 뉴스 텍스트를 추가적인 정보 원천으로 추가한 효과는 연간 예측 모형에서는 다소 미미하였다. 하지만 월간 예측 모형에서는 텍스트 정보 기반의 예측 모형이 시장 정보 기반의 예측 모형인 KMV 모형과 유사한 결론을 도출할 수 있어 기업 부도 예측 과정에서 조기 경보 모형으로 충분히 활용이 가능함을 실증하였다.

본 연구는 현재 도출된 결과로도 의미가 있지만 향후 연구의 확장이 필요하다. 현재 분석 대상인 상장기업의 경우 재무 정보가 기업 현황을 비교적 잘 반영하고 있고, 기업에 대하여 발생하는 정보 또한 주가에 즉각 반영되고 있는 편이기 때문에 텍스트 정보 및 인공지능 도입에 대한 예측 증가 수준이 미미할 수 있다고 판단된다. 따라서 재무정보의 신뢰도가 떨어지고 시장 정보의 확보가 어려운 중소기업이나 개인에 대하여 본 연구의 부도 예측 방법을 적용한다면 기존의 방법에 대하여 추가적인 예측 수준 증대를 얻을 수 있을 것이다. 다만, 이러한 연구 시 본 연구가 대상으로 한 뉴스 텍스트 정보 뿐만 아니라 웹 페이지, 공시자료 등 추가적인 정보 원천을 포괄하여 적용하여야 유의미한 결과를 얻을 수 있을 것이다.

빅데이터 및 딥-러닝 분야는 아직까지 국내 금융, 재무 분야에서 관련 연구가 부족한 상황이다. 하지만 본 연구에서 활용한 방법론은 타 연구에서도 충분히 응용하여 활용이 가능하다. 향후 관련 연구자들의 괄목할만한 연구 성과가 많이 도출되기를 기대한다.

### <참고문헌>

- 김민수 · 구평회(2013), "인터넷 검색추세를 활용한 빅데이터 기반의 주식투자전략에 대한 연구," 「한국경영과학학회지」, 제 38 권 제 4 호, pp. 53-63.
- 김성규(2010), "경기변동을 반영한 부도예측모형에 관한 실증연구 : 중소기업 회계정보 기반 동태적 모형을 중심으로", 「한양대학교 박사학위 청구논문」, pp. 1~142.
- 김성진 · 안현철(2016), "기업신용등급 예측을 위한 랜덤포레스트의 응용", 「산업혁신연구」, 제 32 권 제 1 호, pp. 187-211.
- 김유신 · 김남규 · 정승렬(2012), "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형," 「지능정보연구」, 제 18 권 제 2 호, pp. 143-156.
- 박재빈(2006), 「생존분석 이론과 실제」, 신광출판사.
- 안성원 · 조성배(2010), "뉴스 텍스트마이닝과 시계열 분석을 이용한 주가예측," 「한국컴퓨터종합학술대회 논문집」, 제 37 권 제 1 호, pp. 364-369.
- 오세경(2001), "다변량 판별분석모형과 주식옵션모형을 이용한 기업도산 예측", 「산은조사월보」, 제 549 호, pp. 1-29.
- 이광석(2014), "빅데이터 기반의 거래기업 모니터링", 「기술금융연구」, 제 4 권 제 1 호, pp. 91-131.
- 이인로 · 김동철(2015), "회계정보와 시장 정보를 이용한 부도예측모형의 평가 연구", 「재무연구」, 제 28 권, 제 4 호, pp. 626-666.
- 이재식 · 한재홍(1995), "인공신경망을 이용한 중소기업 도산 예측에 있어서의 비재무정보의 유용성 검증", 「한국전문가시스템학회지」, 제 1 권, 제 1 호, pp. 123-134.
- 조남옥 · 신경식(2016), "빅데이터 기반의 정성 정보를 활용한 부도예측 모형 구축", 「지능정보연구」, 제 22 권, 제 2 호, pp. 33-56.
- 최정원 · 오세경(2016), "생존분석과 KMV 모형을 이용한 기업부도예측", 「상경연구」, 제 41 권 제 1 호, pp. 91-136.
- 최정원 · 한호선 · 이미영 · 안준모(2015), "텍스트마이닝 방법론을 활용한 기업 부도예측 연구", 「생산성논집」, 제 29 권 제 1 호, pp. 201-228.
- Addal, S.(2016), "Financial forecasting using machine learning", *African Institute for Mathematical Science(AIMS)*, pp. 1-32.
- Altman, E.(1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy.", *The Journal of Finance*, Vol. 23, No. 4, pp. 589-609.
- Breiman, L.(2001), "Random forests", *Machine Learning*, Vol. 45, No. 1, pp. 5-32.
- Campbell, J. Y., J. Hilscher, J. Szilagyi(2008), "In search of distress risk", *The Journal of Finance*, Vol. 63, No. 6, pp. 2899-2939.
- Chen, H., P. De, Y. Hu, and B. Hwang(2014), "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media", *Review of Financial Studies*, Vol. 27, No. 5, pp. 1367-1403.

- Flood, Mark D., H. Jagadish, L. Raschid (2016), Big data challenges and opportunities in financial stability monitoring
- Glorot, X., Y. Bengio(2010), "Understanding the difficulty of training deep feedforward neural networks", *Proceedings of the 13<sup>th</sup> International Conference on Artificial Intelligence and Statistics*, PMLR 9, pp. 249-256.
- Gu, X., H. Zhang, D. Zhang and S. Kim(2016), "Deep API Learning", *In Proceedings of the 24th ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE 2016)*, pp. 1-12.
- Kim, H., S. So(2010), "Support vector machines for default prediction of SMEs based on technology credit", *European Journal of Operational Reserch*, Vol. 201, pp. 838-846.
- Kim, K. (2003), "Financial time series forecasting using support vector machines", *Neurocomputing*, Vol. 55, pp. 307-319.
- Lu, Y.C., C.H. Shen and Y.C. Wei(2013), "Revisiting early warning signals of corporate credit default using linguistic analysis", *Pacific-Basin Finance Journal*, Vol. 24, pp. 1-21.
- Marko, K., R. T. Krishnamachari(2017), "Big data and AI Strategies", *Global Quantitative & Derivatives Strategy*, JP Morgan, pp. 1-280.
- Martinez, J. and R. Garcia, F. Sanchez(2012), "Semantic-Based Sentiment analysis in financial news", *Finance and Economics on the Semantic Web 9<sup>th</sup> conference*, pp. 38-51.
- McQuown, J. A.(1993), "A Comment on Market vs. Accounting-Based Measures of Default Risk", *KMV Corporation working paper*.
- Merton, R. (1973), "On the Pricing of Corporate debt: The Risk Structure of Interest Rates", *Journal of Finance*, Vol. 29, No. 2, pp.449-470.
- Nam, C., T. Kim, N. Park, and H. Lee(2008), "Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies", *Journal of Forecasting*, Vol. 27, no. 6, pp. 493-506.
- Ohlson, J. A. (1980), "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of accounting research*, Vol. 18, No. 1, pp. 109-131.
- Shumway, T.(2001), "Forecasting bankruptcy more accurately: A simple hazard model", *The Journal of Business*, Vol. 74, no. 1, pp. 101-124.
- Tinoco M. H., N. Wilson(2013), "financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables", *International Review of Financial Analysis*, Vol. 30, pp. 394–419.
- Vahala, J.(2016), "Prediction of financial markets using Deep learning", *Bachelor's Thesis, Masaryk University*, pp. 1-50.
- Yeh, S., C. Wang, M. Tsai(2015), "Corporate default prediction via deep learning", *Wireless and Optical Communication Conference (WOCC) 24<sup>th</sup>*, pp. 1-8

## &lt;Appendix 1 – 변수 정의&gt;

Code	분류	Index	산식
F01	건전성	부채비율	총부채/총자산
F02		시장부채비율	총부채/시장총자산*
F03		금융부채비율	금융부채/총자산
F04		금융부채비율 2	금융부채/총부채
F05		금융부채변동율	당기 - 전기 / 전기 금융부채
F06		이자보상배율	영업이익/이자비용
F07		유동비율	유동자산/유동부채
F08		고정자산비율	고정(비유동)자산/총자산
F11	수익성	총자산영업이익율	영업이익/총자산
F12		총자산순이익율	당기순이익/총자산
F13		시장자산영업이익율	당기순이익/시장총자산*
F14		자기자본순이익율	당기순이익/총자본
F15		총자산이익잉여금비율	이익잉여금/총자산
F21	성장성	총자산증가율	당기 - 전기 / 전기 총자산
F22		매출액증가율	당기 - 전기 / 전기 매출액
F23		당기순이익증가율	당기 - 전기 / 전기 당기순이익
F31	유동성	현금자산비율	현금 및 현금성자산/총자산
F32		시장현금자산비율	현금 및 현금성자산/시장총자산*
F33		자산대비영업현금흐름	영업현금흐름/총자산
F34		자산대비총현금흐름	총현금흐름/총자산
F41	활동성	자산회전율	매출액/총자산
F42		매출채권회전율	매출액/매출채권
F51	규모	총매출액	ln(총매출액)
F52		총자산	ln(총자산)
M01	시장 정보	주가수익률	당기 - 전기 / 전기 주가**
M02		주가초과수익률	주가수익률 - 시장수익률
M03		주가변동성	주가변동성(20 일)
M04		주가수준	ln(주가)
E01	거시경제 (연 기준 적용)	Kbond	국고채(3 년)
E02		CD	CD 유통수익률(91 일)
E03		GDP	국내총생산(실질성장률)
E04		USD	원/미국달러(매매기준율)
E05		KOSPI	KOSPI_증가
E07		dPPI	PPI 증감
E08		dCPI	CPI 증감
E09		House	주택매매가격지수(증감율)
E10		Oil	국제유가(Dubai)
E11		Unemp	실업률
C01	기업특성	직원수 증감	당기 - 전기 / 전기 직원수
C02		직원평균임금 증감	당기 - 전기 / 전기 인당평균임금
C03		최대주주지분율	최대주주 지분율
C04		최대주주지분율 증감	당기 - 전기 지분율
C05		배당수익률	배당수익률
articleNum	비정형정보	연간 기사수	해당 기업 관련 총 기사 수



NumNeg_1	(뉴스 텍스트)	부도기사비율_1	연간 부도(w2v-부도) 기사수 / 연간 기사수
NumNeg_2		부도기사비율_2	연간 부도(w2v-부도&상폐) 기사수 / 연간 기사수
NumNeg_3		부도기사비율_3	연간 부도(선행연구 단어) 기사수 / 연간 기사수
w2v1_1		부도 유사도_1	연관도평균(w2v-부도)
w2v1_2		부도 유사도_2	연관도합계 (w2v-부도) / 기사 수
w2v2_1		<부도+상장폐지>유사도_1	연관도평균(w2v-부도&상장폐지)
w2v2_2		<부도+상장폐지>유사도_2	연관도합계 (w2v-부도&상장폐지) / 기사 수
market	기업특성 (통제변수적용)	소속시장	KOSPI / KOSDAQ
industry		산업	표준산업분류기준 분석용 산업그룹 재분류
group		재별그룹여부	30 대 재별 그룹 소속 기업
KP200		공공기관여부	공기업 & 공기업이 대주주 기업
Gov		대기업여부	KOSPI200 기업

\* 시장총자산 = 주식의 시장가치 + 부채의 장부가치

\*\* 주가는 지분 변동 등을 고려한 수정주가 사용