나의 부도일지

텍스트 마이닝을 활용한 기업 부도예측 모형 연구

초록

텍스트 마이닝을 활용한
기업부도예측 모형 연구

<요약>

본 연구는 기존의 연구를 기반으로 부도 예측 과정에서 뉴스텍스트와 같은 데이터를 데이터 적용 방법에 따라 부도예측력을 높일 수 있는지, 인공지능 기법을 통한 방법을 통해 예측성능이 향상되는지를 확인해보는 연구이며, 연도별 키워드를 다양한 키워드를 추출하기위해 코스닥기업 대상 2010년부터 2021년의 기간을 수집하였다. 분석결과로 텍스트데이터의 유의성을 확인할 수 있었으며, 데이터베이스 구축을 통해서 일반인이 사용할 수 있게 실증 모델을 구축하였다.

목차

1) 프로젝트 팀 소개
2) 프로젝트 개요
3) 프로젝트 수행절차 및 방법
4) 프로젝트 수행 결과
5) 자체평가

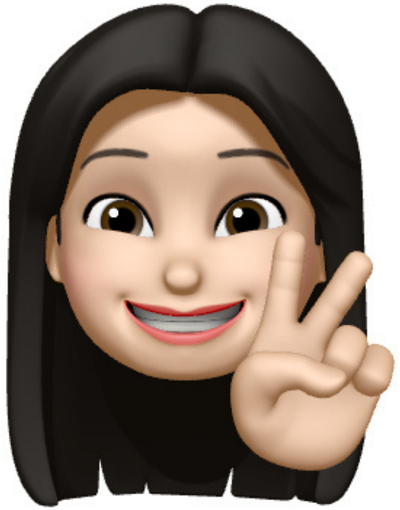
01

프로젝트 팀소개

01
팀원 소개

황윤재

팀리더 및 코딩



박희연

재무 데이터 수집
및 PPT제작

장동민

텍스트 데이터 수집
및 전처리

02

프로젝트 개요

01. 연구 배경
02. 주제 선정 이유

02 연구배경

서울파이낸스 | 2022.06.22.

[금융안정보고서] "기업대출 부실 가능성 높아져…은행, 충당금 …

잠재부실이 표면화될 가능성이 있다는 분석이다. 한국은행은 22일 '**금융안정보고서**'를 통해 코로나19 정책지원 종료시 기업대출의 신용손실 확대 가능성을 제기…

서울경제 | 4면 9단 | 2022.06.14. | 네이버뉴스

'코로나 유동성' 회수하자…좀비기업 부실폭탄 째깍

나타나면 **부도 기업**이 **증가**할 수 있다"고 밝혔다. 최근 주요 국제기구들은 전 세계적인 **부도 기업** 급증 위기를 강하게 경고하고 있다. 금융안정위원회(FSB), 세계은…

머니S | 2022.06.14. | 네이버뉴스

좀비기업, 잠재부실 터지나… 한은 "기업 채무조정제도 정비 시급"

벌어들인 돈으로 이자도 내지 못하는 좀비**기업**이 크게 늘어난데다 **부도기업**이 **증가**할 가능성이 있는만큼 **기업** 채무조정제도 정비가 시급하다는 지적이 나왔다. …

"금융지원조치 정상화"

"잠재된 부실리스크 드러나 부도기업 늘어날 가능성 증가"

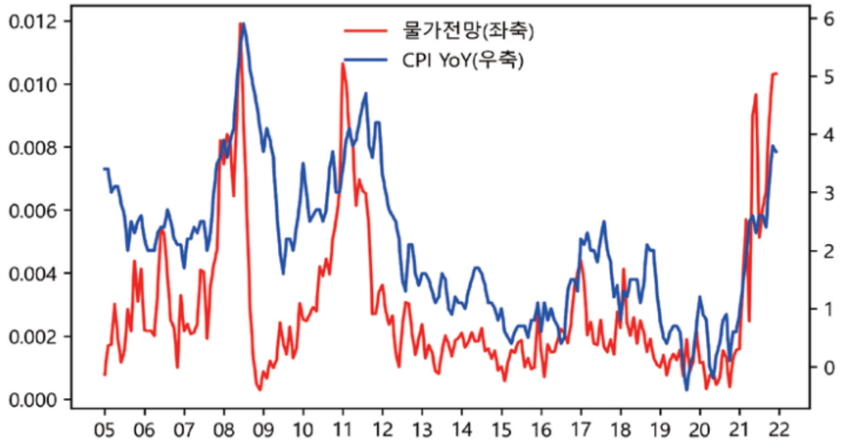
기업 부도 예측 필요

02 연구배경

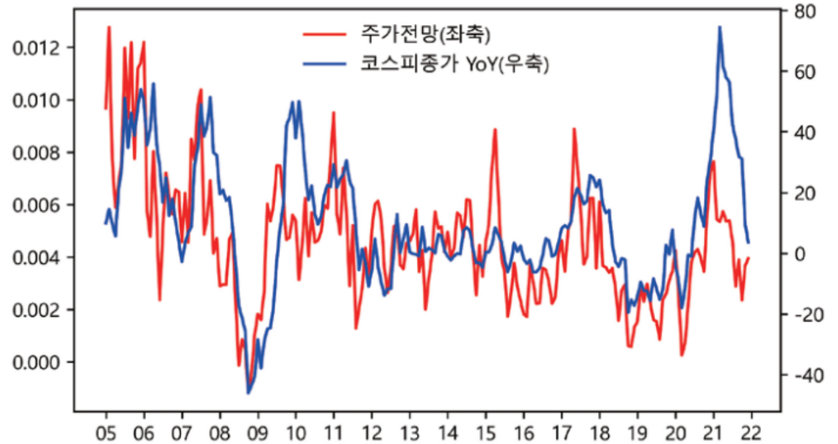
한은 "뉴스 기반 경제지표, 공식통계보다 최대 9개월 선행"

출처 : 매일 경제 2022.05.16

(물가전망 텍스트 지표와 소비자물가지수)



(주가전망 텍스트 지표와 코스피지수)



뉴스 텍스트는 다양한 전문가의 견해 · 전망 등
정성적 정보를 포함하고 실시간으로 입수 가능하므로
종합하고 정량화하여 경기예측에 활용할 필요가 있다.

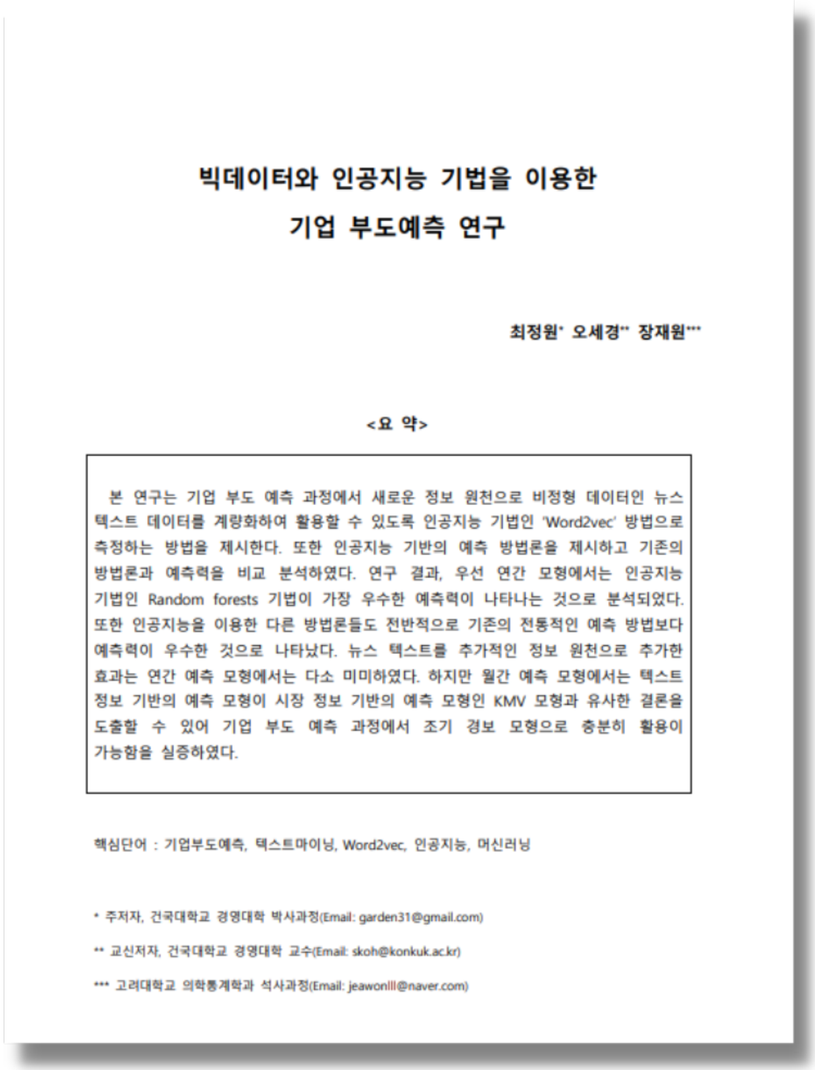
02
선행 논문

"빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구"
최정원,오세경,장재원

✓ 기업에 관한 뉴스는
해당 기업에 대한 가장 빠른 정보 중 하나

✓ 텍스트 정보 기반의 예측 모형이 시장 정보 기반의 예측 모형인 KMV 모형과 유사한 결론을 도출함

➡ 기업 부도 예측 과정에서 조기 경보 모형으로 충분히 활용이 가능함을 실증



빅데이터와 인공지능 기법을 이용한
기업 부도예측 연구

최정원* 오세경** 장재원***

<요 약>

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)

02

주제선정이유

"텍스트마이닝을 활용한 기업 부도 예측 모형 연구"

- ✓ 기업의 부실을 선제적으로 예측하기 위한 추가적인, 대체가능한 정보 원천으로 충분한 가치가 있다.
- ✓ 뉴스 텍스트 정보를 활용해 부도 예측 수준이 향상될 수 있는지 확인

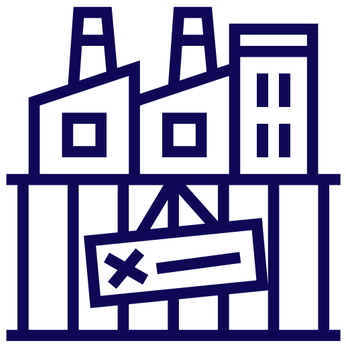
03

프로젝트 수행절차 및 방법

01. 부도의 정의
02. 데이터 수집
03. 데이터 전처리
04. 통계검정

03

부도의 정의



부도 발생 기업

상장폐지가 결정된 기업들 중
KIND(한국거래소-상장공시시스템)를 통해 상장폐지 사유를 확인

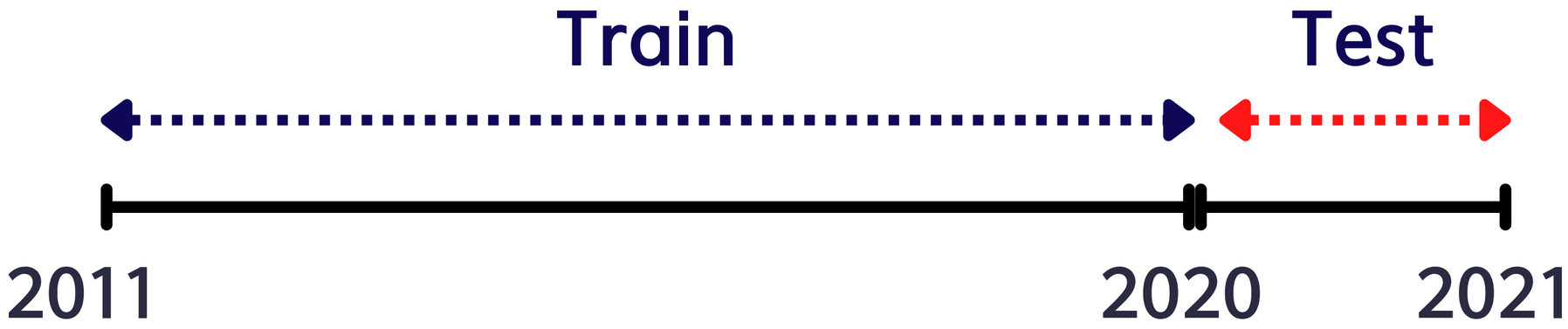
피흡수합병,자진상장폐지신청,유가증권상장 등 실적 부진과
관계없이 상장폐지된 기업은 **정상기업**으로 간주

03
부도의 정의

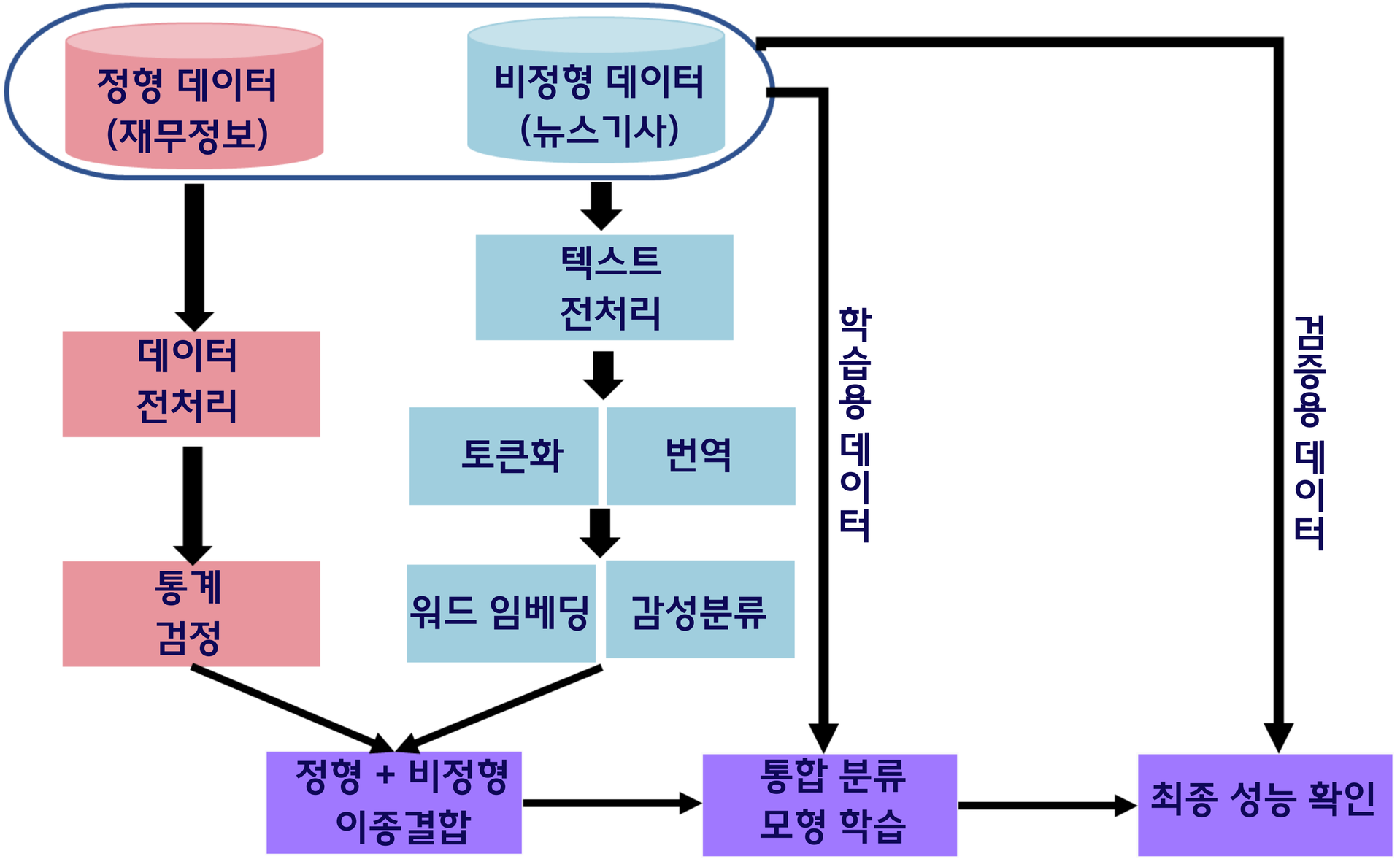
번호	회사명	폐지일자	폐지사유	비고
394	코 제이테크놀로지 폐 관	2019-12-12	기업의 계속성 및 경영의 투명성 등을 종합적으로 고려하여 상장폐지기준에 해당한다고 결정	
393	코 한화에이스스팩3호 폐 관	2019-12-06	상장예비심사 청구서 미제출로 관리종목 지정 후 1개월 이내 동 사유 미해소	
392	코 데코앤이 폐 관 환	2019-11-20	발행한 어음 또는 수표가 주거래은행에 의하여 최종부도로 결정되거나 거래은행에 의한 거래정지	
391	코 제이콘텐트리 X300	2019-10-18	유가증권시장 상장	유 콘텐트리중앙 X300
390	코 대신밸런스제4호스팩 폐 관	2019-10-10	상장예비심사 청구서 미제출로 관리종목 지정 후 1개월 이내 동 사유 미해소	
389	코 신한제3호스팩 폐 관	2019-10-01	상장예비심사 청구서 미제출로 관리종목 지정 후 1개월 이내 동 사유 미해소	
388	코 쿠첸 폐	2019-09-16	타법인의 완전자회사로 편입	
387	코 SK3호스팩 폐 관	2019-08-30	상장예비심사 청구서 미제출로 관리종목 지정 후 1개월 이내 동 사유 미해소	
386	코 현대정보기술 폐	2019-07-17	피흡수합병	

03
분석대상의 정의

기업 : 코스닥(KOSDAQ) 상장 제조업
기간 : 2011년 ~ 2020년



03
전체 프로세스


정형 데이터
(재무정보)
비정형 데이터
(뉴스기사)
데이터
전처리
텍스트
전처리
토큰화
번역
통계
검정
워드 임베딩
감성분류
학습용 데이터
검증용 데이터
정형 + 비정형
이종결합
통합 분류
모형 학습
최종 성능 확인

정형 데이터

- 재무 데이터 수집
- 재무 데이터 전처리

03
재무 데이터 수집

- 데이터 수집처: TS 2000
- 데이터 수집 기간: 2009~2018(부도 2년전 재무비율 수집)

회사명	회계연도	폐지일자
네이처글로벌	2009	2011-01-08

회계연도	상장 폐지일
2009	2011-01-08

> 재무비율 데이터는 Altman(1968)이 부도예측을위해 사용한 이후로 부도예측 연구에서 비교 표준으로 사용되고 있다. 후속 연구들을 통해 재무비율 변수가 부도예측에 효과가 있다는 것은 이미 입증이 되었다.
>
> "부도예측 모형에서 뉴스 분류를 통한 효과적인 감성분석에 관한 연구"-2019

03
재무 데이터 수집

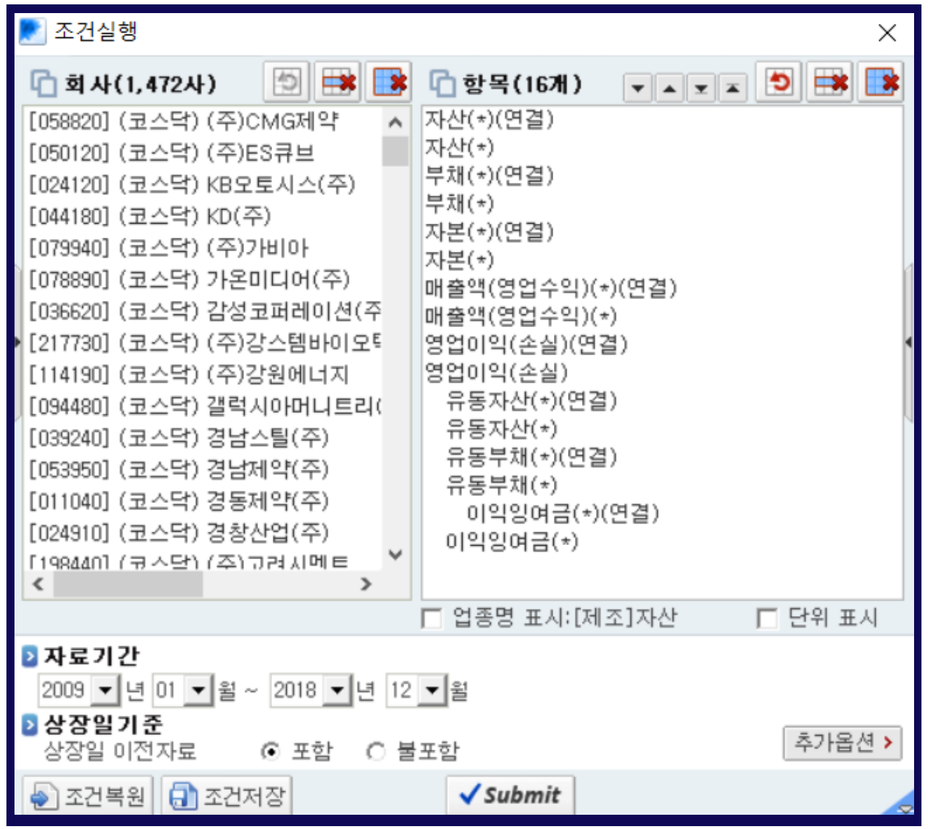
알트만 z-score

$$Altman\ Z-Score = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E$$

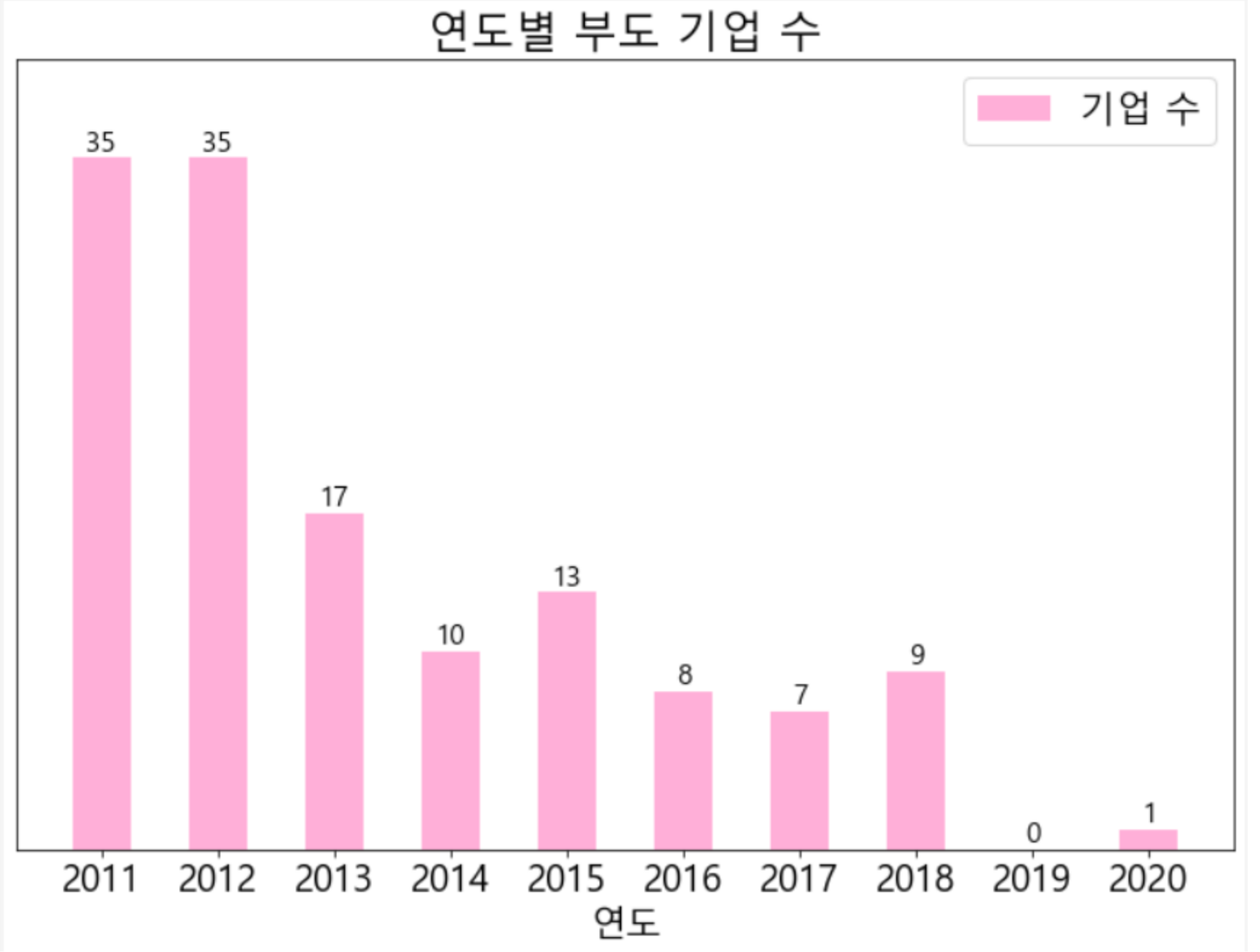
기업 연간보고서에 있는 재무제표 수치를 통해 계산할 수 있는 5개의 재무 비율

- 자산
- 자본
- 부채
- 매출액
- 영업이익
- 유동자산
- 유동부채
- 이익잉여금

} 8가지 수집



03
재무 데이터 수집


연도별 부도 기업 수
기업 수
35
35
17
10
13
8
7
9
0
1
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
연도

03

분석 대상 기업 표본

정상기업에 비해 부도 기업의 수는 상당히 적은 편, **불균형데이터**는 부도예측 모형이 정상 기업만 편향되게 예측하도록 하여 표본에서 부도기업을 예측하지 못하는 문제 발생

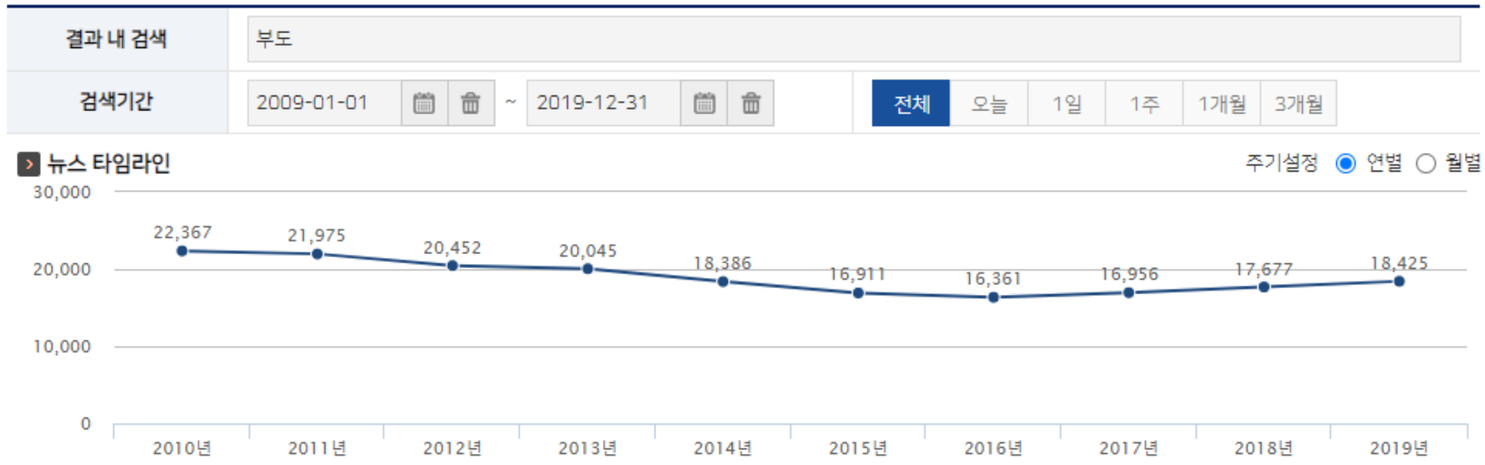
연도별로 부도기업 표본 수와 동일하게 정상 기업의 표본 수를 할당하고, 부도기업을 제외한 전체 기업에서 **시가 총액** 순으로 추출

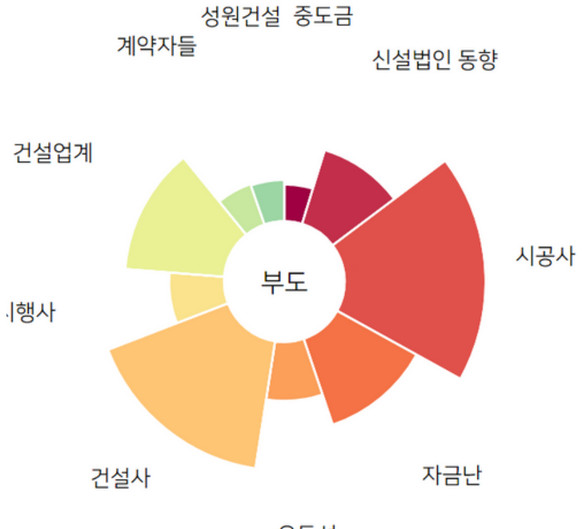
총 270개의 기업표본을 구성

기업 수											
연도	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
정상	35	35	17	10	13	8	7	9	0	1	135
부도	35	35	17	10	13	8	7	9	0	1	135
Total	70	70	34	20	26	16	14	18	0	2	270

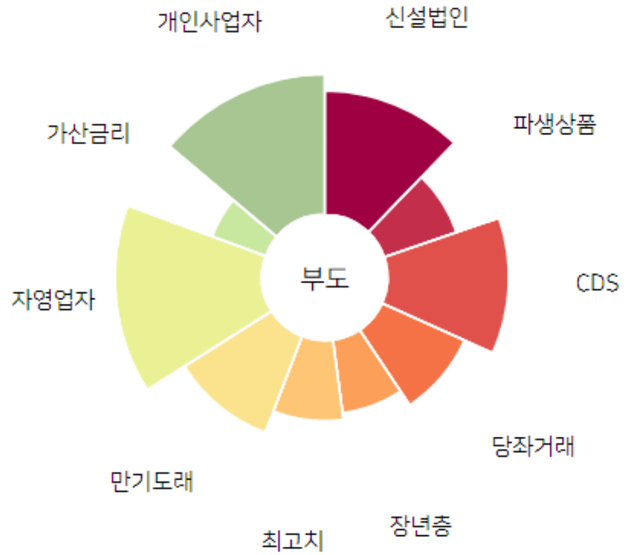
03
뉴스기사 연도별 경향

출처 :K2Base(KISTEP Knowledge Base)과학기술정책지원서비스

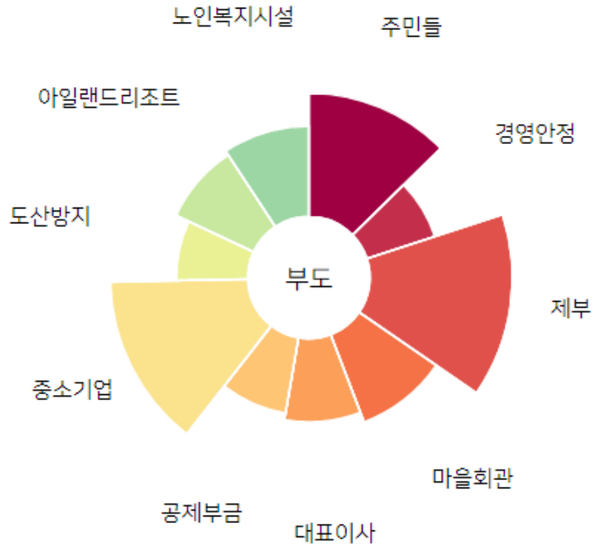




2010-01



2015-01



2019-01

비정형 데이터

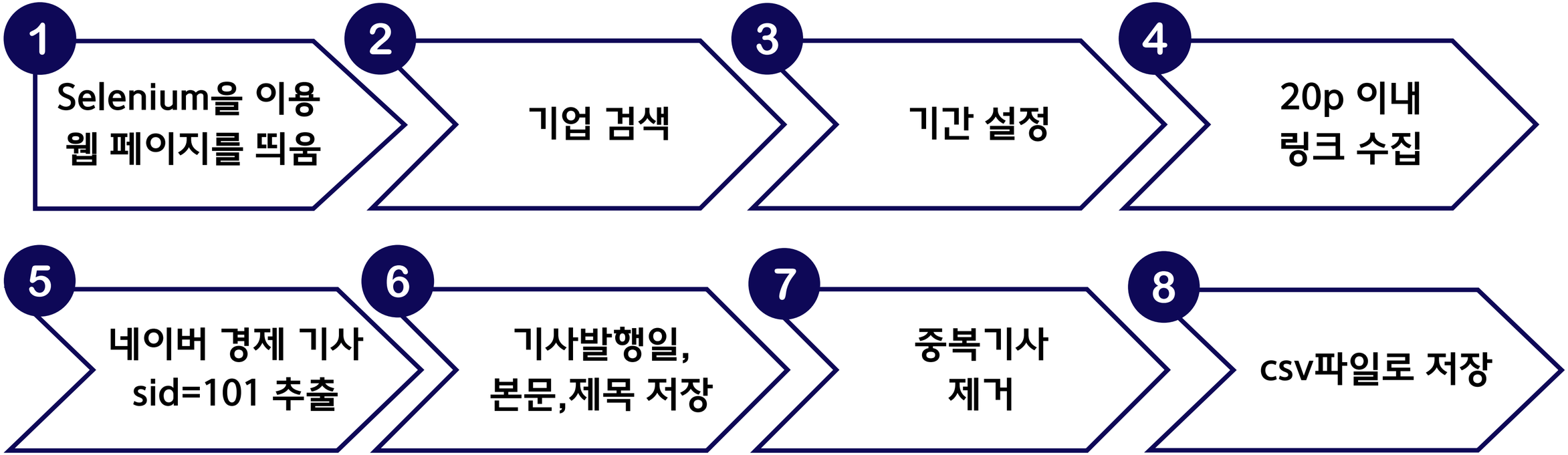
- 뉴스 텍스트 수집
- 뉴스 텍스트 전처리

03
뉴스기사 수집

- 데이터 수집처 : 네이버 뉴스
- 데이터 수집 기간 :
  상장폐지일 - 3개월 기준 1년간 뉴스데이터 수집

	언론사
경제	그린데일리,글로벌이코노미, 나이스 경제, 뉴스토마토, 대한 금융신문, 데일리 경제, 디지털 데일리, 매일경제,매일경제TV,머니투데이,블로터,서울경제,서울파이낸스, 시장경제신문,아시아경제, 아시아타임즈, 이뉴스 투데이, 이데일리, 이코노 뉴스, 이투데이,조선비즈, 조세일보,중소기업신문, 증권경제신문,초이스 경제,핀포인트뉴스,한국경제,헤럴드경제 등

03
뉴스기사 수집 프로세스



03
뉴스기사 수집 프로세스

03
전처리 전 뉴스 기사

기업	기사발행일	기사제목	기사본문
네이쳐글로벌	2010.10.01. 오전 9:23	[특징주] 네이쳐글로벌 상한가…"전 대표 횡령 확인 안돼"	비메모리 반도체 업체인 네이쳐글로벌이 전 대표 횡령 혐의에 대해 확인된 바 없다고 ...
네프로아이티	2011.07.14. 오전 4:08	[코스닥 메모] (14일) 일반공모청약=네프로아이티 등	◇변경상장(이익소각)=다음 (무상감자)=헤스본₩n₩n◇무상기준일=다원시스 파트론₩n...
코썬바이오	2019.12.23. 오후 5:17	코썬바이오 불성실 공시법인 지정예고	코썬바이오는 공시불이행으로 불성실 공시법인으로 지정예고 됐다고 23일 공시했다.₩n...
세븐코스프	2010.10.06. 오후 3:23	[코스닥 마감]500선 코앞..외인·기관 '쌍끌이'	[머니투데이 김성호 기자]코스닥시장이 500선에 조금 못 미친 채 마감했다. 외국인...

데이터 전처리

재무 비율

- 결측치 처리
- 변수 생성

03
재무 데이터 전처리

회사명	자산(연결)	자산	부채(연결)	부채	자본(연결)	자본	매출액(영업수익)(연결)	매출액(영업수익)	영업이익(손실)(연결)	영업이익(손실)	유동자산(연결)	유동자산	유동부채(연결)	유동부채	이익잉여금(연결)	이익잉여금
(주)CMG제약		34556966		9803795		24753170		10808242		-3785500		16725293		2971345		-18256069
(주)ES큐브		47490304		5480667		42009637		34640205		704325		26290142		4769552		-2249725
(주)가비아	43119087	31434599	17384906	16130784	25734181	15303815	32695039	20479549	5522735	3900714	21050302	11900139	9454572	8527994	10384246	10384246

- ✓ 연결 재무제표의 결측값은 개별 재무 제표로 대체, 나머지 결측값은 제거
- ✓ 알트만 5가지 재무비율 계산하여 새로운 변수 생성

05

재무 데이터 전처리

재무비율	공식
운전자본비율	$\frac{\text{운전자본}}{\text{총 자산}} = \frac{\text{유동자산-유동부채}}{\text{총 자산}}$
이익잉여금 총자산비율	$\frac{\text{이익잉여금}}{\text{총 자산}}$
총자산영업이익률	$\frac{\text{이자 및 세전이익}}{\text{총 자산}}$
시장가부채비율	$\frac{\text{자본의 시장가치}}{\text{총 부채의 장부가치}}$
매출액회전율	$\frac{\text{매출액}}{\text{총 자산}}$

05
재무 데이터 전처리 후

회사명	운전자본비율	이익잉여금총자산비율	총자산 이익률	시장가 부채비율	매출액 회전율
네이쳐글로벌	74.49204859	-49.80358923	-3.37118151	673.4326222	6.028268015
네프로아이티	-0.465153588	19.33814966	13.88472606	157.3443634	76.78071108
뉴젠아이씨티	27.59367683	-1075.333708	-28.34280817	136.3422624	40.74731592

부도
기사비율

- 부도 유사도
- 부도 기사비율

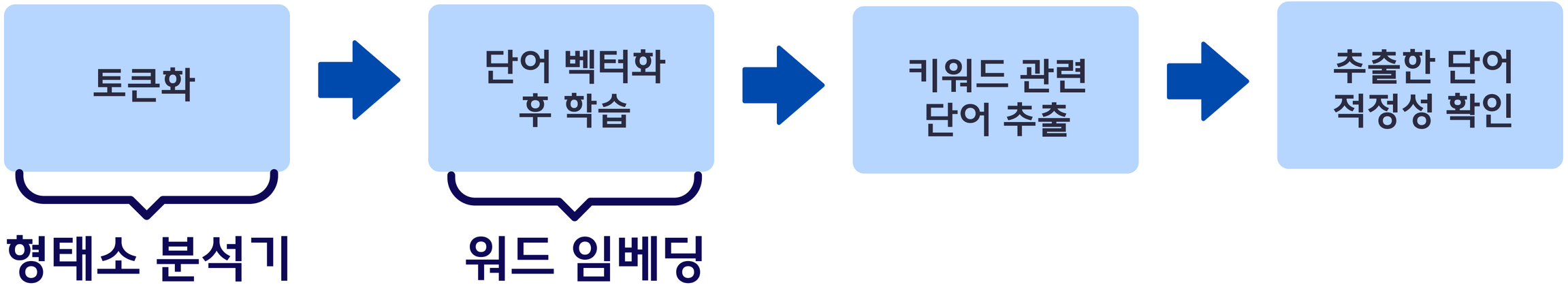
03
뉴스 데이터 전처리

'공시' 또는 '광고'와 같은 불필요한 기사 뉴스 제목에서 등장하는 단어들 중 '증시일정', '장마감후','주식왕 따라잡기', '주식컨설팅'와 같은 단어들을 포함하는 기사 제거

연도별 기사 수													
연도	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
전처리 전	8	2003	4833	3645	1857	1388	1138	839	898	892	10	45	17556
전처리 후	7	1658	4014	3095	1534	1223	1001	695	817	813	8	49	14879

03

뉴스 데이터 전처리

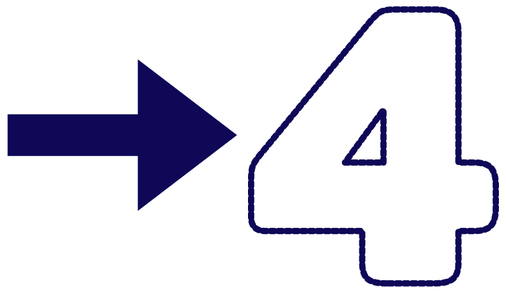


03
뉴스 데이터 전처리

Step1) 크롤링한 뉴스 기사 본문 토큰화
형태소 분석기 : Mecab,Okt

Step2) 학습 후에 단어를 벡터화 , 단어 유사도 추출
워드임베딩 기법 : word2vec , fasttext

2개의 형태소 분석기
X
2개의 워드임베딩 기법



가지

1) mecab-word2vec
2) mecab-fasttext
3) Okt-word2vec
4) Okt-fasttext

03
뉴스 데이터 전처리

1 Mecab+ Word2Vec

단어	유사도
격성	0.773562
상장	0.765573
퇴출	0.764932
실질	0.720605
심사	0.719068
속개	0.707785
이의	0.673985
심의	0.66827
거절	0.667534
확인서	0.655585
모면	0.645451
주권	0.622341
반기	0.621481
수순	0.604381
신청서	0.602632
거래소	0.599945
만료일	0.598995
제출	0.589584
여부	0.589018
회계감사인	0.584685

2 Mecab + FastText

단어	유사도
상장	0.793303
퇴출	0.791238
심사	0.754463
격성	0.745931
실질	0.727271
거절	0.691061
속개	0.690446
심의	0.683694
이의	0.679556
확인서	0.663792
신청서	0.630467
모면	0.62622
반기	0.625649
회계감사인	0.610378
여부	0.607254
잠식	0.6052
제출	0.602151
회계감사	0.600304
수순	0.596035
거래소	0.595941

3 Okt + Word2Vec

단어	유사도
상장폐지	0.770364
상폐	0.711783
적격성	0.685097
이의신청	0.680455
규정	0.673589
심사	0.673422
실질	0.646441
날로	0.637251
현행	0.63207
요건	0.627837
영업일	0.620572
폐가	0.615732
의거	0.61341
회피	0.607404
존립	0.594494
무더기	0.592312
모면	0.590389
행세	0.586581
가까스로	0.586334
이의	0.581774

4 Okt + FastText

단어	유사도
상장폐지	0.786027
상폐	0.760817
이의신청	0.693378
적격성	0.683094
심사	0.679682
규정	0.644429
심의	0.635623
실질	0.629092
거절	0.628239
무더기	0.610584
영업일	0.610526
날로	0.608194
타당	0.598166
행세	0.597805
비적정	0.591826
통보	0.587716
이의	0.58765
위원회	0.587023
주권	0.582306
현행	0.581516

03
뉴스 데이터 전처리

분류단어 : 폐지, 격성, 퇴출, 실질, 심사, 속개, 이의, 심의, 거절

재무비율	공식	비율
부도기업 부도기사 비율	$\frac{\text{부도기업 부도기사수}}{\text{부도기업 전체기사수}} = \frac{1770}{8494}$	20.9%
정상기업 부도기사비율	$\frac{\text{정상기업 부도기사수}}{\text{정상기업 전체기사수}} = \frac{460}{6882}$	6.7%
부도기사수 차이	1770 - 460 = 1310	14.2%

03

뉴스 데이터 전처리

$$\text{부도 기사비율} = \frac{\text{해당 기업의 부도 기사 수}}{\text{해당 기업의 전체 기사 수}} \times 100$$

회사명	회계년도	운전자본 비율	이익잉여금 총자산비율	총자산 이익률	시장가 부채비율	매출액 회전율	부도기사 비율	부실기업 여부
네이쳐글로벌	2009	74.49204859	-49.80358923	-3.37118151	673.4326222	6.028268015	5.41	1
네프로아이티	2009	-0.465153588	19.33814966	13.88472606	157.3443634	76.78071108	36.27	1
뉴젠아이씨티	2009	27.59367683	-1075.333708	-28.34280817	136.3422624	40.74731592	14.43	1

감성 분석

- 뉴스 기사 번역
- 감성 분석

03
뉴스 데이터 전처리

"딱딱한 경제 기사에서 '감성'찾아내는 뉴스심리 지수"

시사 in 2021.04.29

한국은행은 이 지수를 만들면서 미국 샌프란시스코 중앙은행의 '데일리 뉴스 센티멘트 인덱스'를 참고했다. 데일리 뉴스 센티멘트 인덱스는 〈뉴욕타임스〉를 비롯한 미국 16개 언론의 경제 기사 단어를 분석해 지수를 산출한다. 국내의 뉴스심리지수는 미국 샌프란시스코 중앙은행의 방식과는 차이가 있다. 영어는 언어 사용자가 한국어 사용자보다 훨씬 많아서 데이터셋 형식으로 된 '감성 사전'이 있다. 샌프란시스코 중앙은행은 일반사전인 'WordNet'에 감성 정보를 추가한 'SentiWordNet'과 같이 공개된 감성 사전을 활용한다. 가령 'crisis(위기)'는 마이너스 몇 점 하는 식으로 정해진다. 그런데 한국어는 이런 '감성 사전'이 없다. 게다가 한 단어에 '감성 점수'를 매길 때 난점이 생긴다. 가령 예전에 '코로나'라고

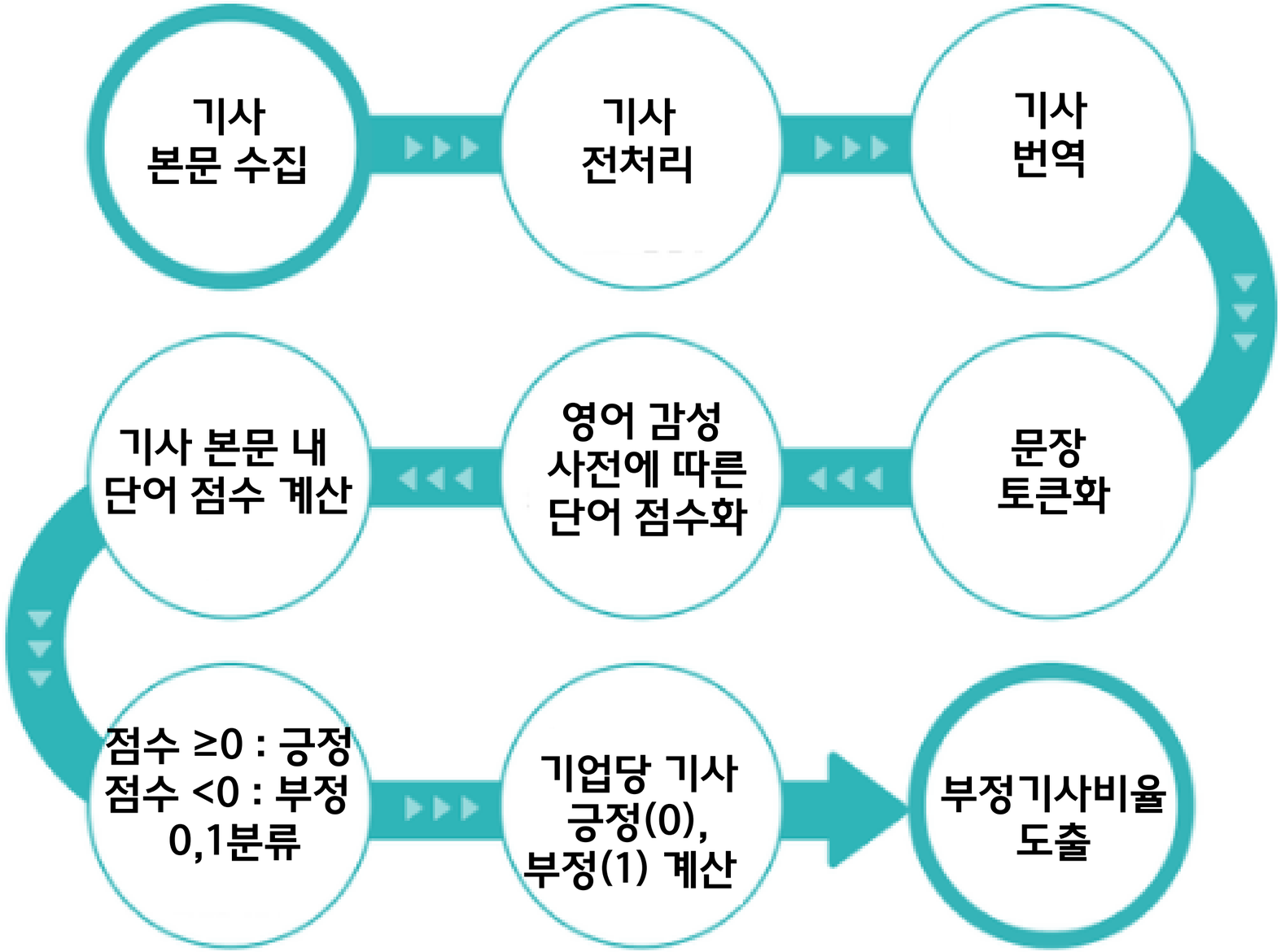
샌프란 시스코 중앙은행에서 경제 기사 단어를 분석하기 위해 사용하는 감성사전 SentiWordNet을 이용해 감성분석

03
뉴스 데이터 전처리

sentiwordnet txt파일

품사	단어 인덱스	긍정	부정	단어
a	1450969	0	0.625	lost
a	1451225	0	0.375	unsaved
a	1451402	0.500	0	ruined
n	8458912	0.125	0.000	progression
n	8459087	0.000	0.500	rash
n	8459252	0.000	0.000	sequence

03
뉴스 데이터 전처리


기사
본문 수집
기사
전처리
기사
번역
문장
토큰화
영어 감성
사전에 따른
단어 점수화
기사 본문 내
단어 점수 계산
점수 ≥0 : 긍정
점수 <0 : 부정
0,1분류
기업당 기사
긍정(0),
부정(1) 계산
부정기사비율
도출

03
뉴스 데이터 전처리

Step 1) 파파고 번역

감성분석을 위한 구글 번역 라이브러리
일정 용량 이상 넘어가면 API 차단

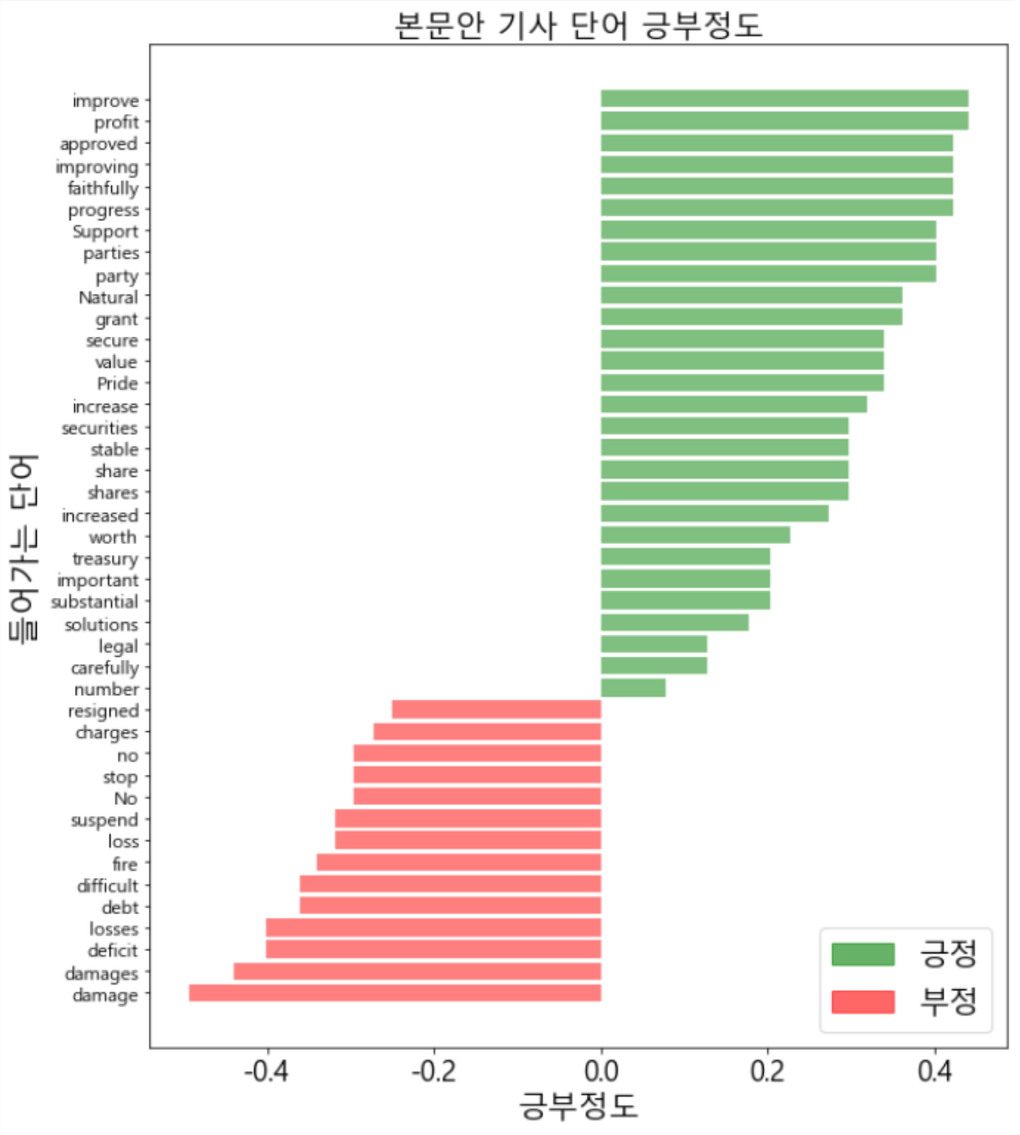
셀레니움을 이용한 파파고 번역

papago

03
뉴스 데이터 전처리

Step 2) 번역 후 토큰화한 긍부정도 점수

단어	Ratings	긍정/부정
deficit	-0.4019	Negative
losses	-0.4019	Negative
profit	0.4404	Positive
worth	0.2263	Positive
won	0.5719	Positive
increased	0.2732	Positive
Best	0.6369	Positive
increase	0.3182	Positive
stop	-0.2960	Negative
fire	-0.3400	Negative
damage	-0.4939	Negative
trust	0.5106	Positive
charges	-0.2732	Negative



뉴스 데이터 전처리 - 단어의 긍부정도 시각화

Step 3) 기사 본문 단어들 긍부정도 합계

Rating ≥ 0 긍정 단어

Rating < 0 부정 단어

→ 기사 본문 내 모든 단어 점수화

Rating 합 ≥ 0 → 긍정 기사

Rating 합 < 0 → 부정 기사

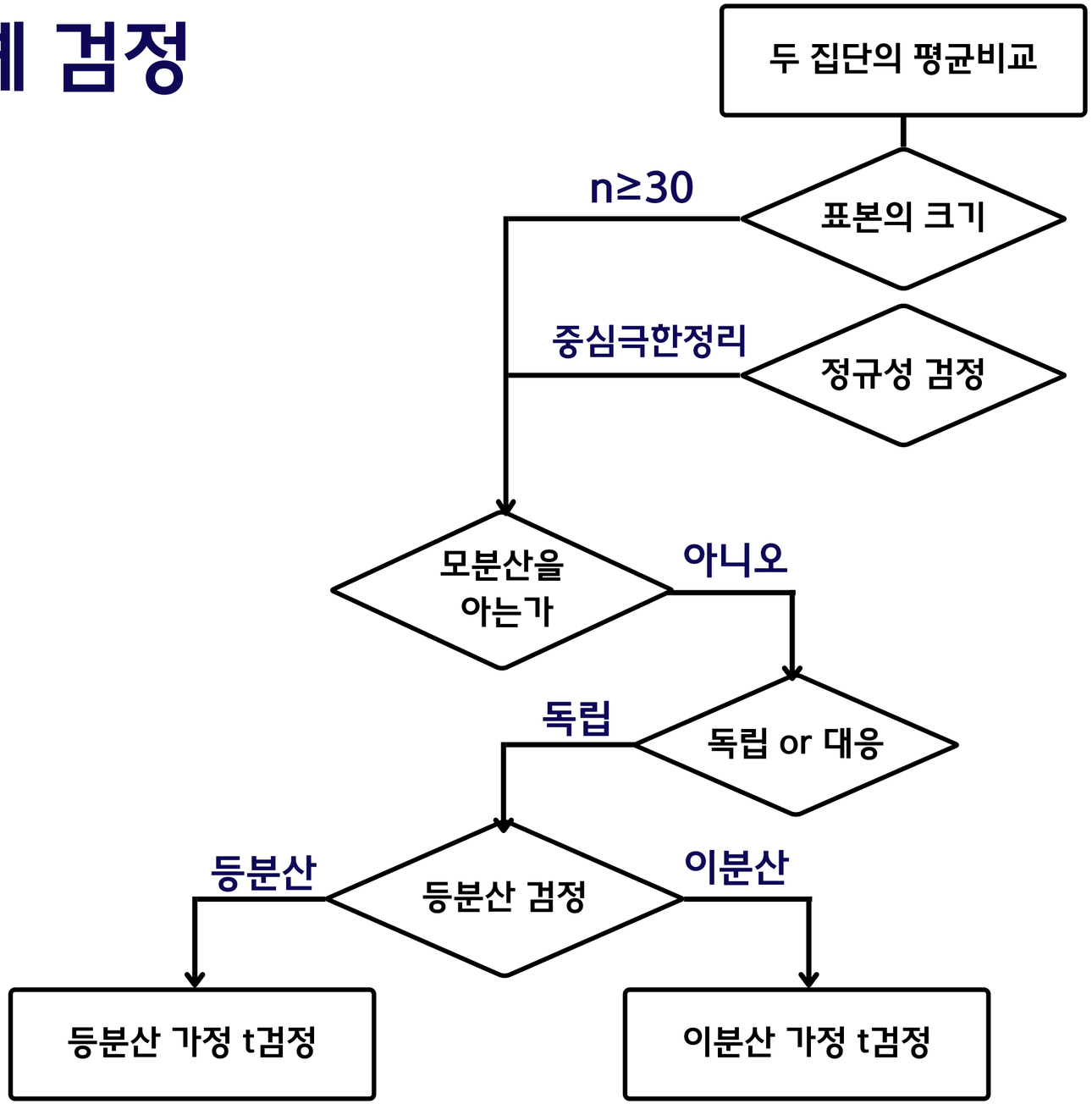
03
부정 기사 비율

$$\text{부정 기사비율} = \frac{\text{해당 기업의 부정 기사 수}}{\text{해당 기업의 전체 기사 수}} \times 100$$

회사명	운전자본 비율	이익잉여금총 자산비율	총자산 이익률	시장가 부채비율	매출액 회전율	부정 기사 비율	부실기업 여부
네이쳐글로벌	74.49204859	-49.80358923	-3.37118151	673.4326222	6.028268015	29.33	1
네프로아이티	-0.465153588	19.33814966	13.88472606	157.3443634	76.78071108	29.8	1
뉴젠아이씨티	27.59367683	-1075.333708	-28.34280817	136.3422624	40.74731592	24.24	1

통계 검정

03
통계 검정



☑ 독립 2표본 t검정

① 분산의 동일성 검사
:Levene's Test

② 등분산일경우
Student's T-test

이분산일 경우
Welch's T-test

03

통계 검정

운전자본 비율

유의수준0.05

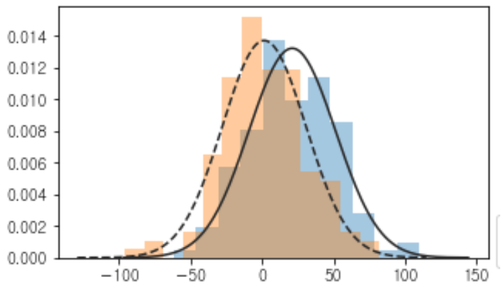
귀무가설 : 정상기업과 부도기업의 운전자본 비율 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 운전자본 비율 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.0677 > 0.05

귀무가설 채택 ⟶ 등분산이다



☑ T-test

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 운전자본 비율의 평균차이가 있다.

03
통계 검정

이익잉여금 총자산 비율

유의수준0.05

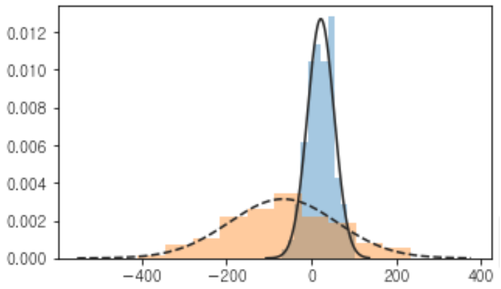
귀무가설 : 정상기업과 부도기업의 이익잉여금 총자산 비율 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 이익잉여금 총자산 비율 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 이분산이다



☑ Welch's T-test

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 이익잉여금 총자산 비율의 평균차이가 있다.

03
통계 검정

총자산 이익률

유의수준0.05

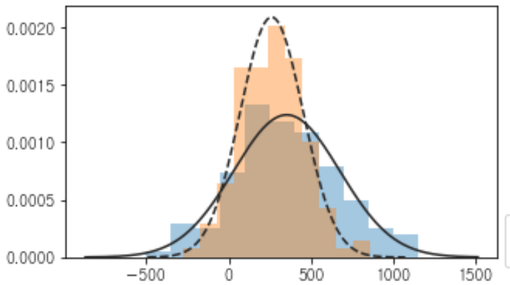
귀무가설 : 정상기업과 부도기업의 총자산 이익률 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 총자산 이익률 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.0841 > 0.05

귀무가설 채택 ⟶ 등분산이다



☑ T-test

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 총자산 이익률의 평균차이가 있다.

03
통계 검정

매출액 회전율

유의수준0.05

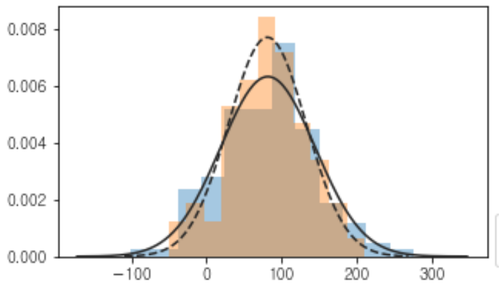
귀무가설 : 정상기업과 부도기업의 매출액 회전율 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 매출액 회전율 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.3552 > 0.05

귀무가설 채택 ⟶ 등분산이다



☑ T-test

p-value : 0.0087 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 매출액 회전율의 평균차이가 있다.

03

통계 검정

시장가 부채비율

유의수준0.05

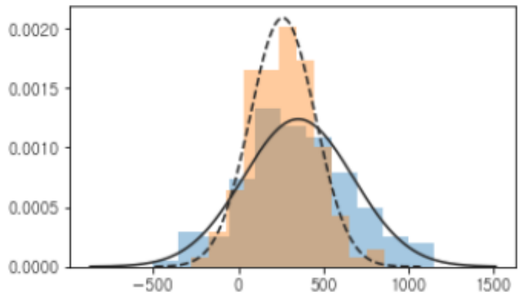
귀무가설 : 정상기업과 부도기업의 시장가부채비율 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 시장가부채비율 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.0107 < 0.05

귀무가설 기각 ⟶ 이분산이다



☑ Welch's T-test

p-value : 0.0005 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 시장가부채비율의 평균차이가 있다.

03
통계 검정

부정기사 비율

유의수준0.05

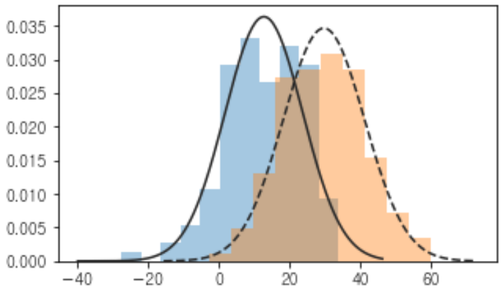
귀무가설 : 정상기업과 부도기업의 부정기사 비율 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 부정기사 비율 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.4471 > 0.05

귀무가설 채택 ⟶ 등분산이다



☑ T-test

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 부정기사 비율의 평균차이가 있다.

03

통계 검정

부도기사 비율

유의수준0.05

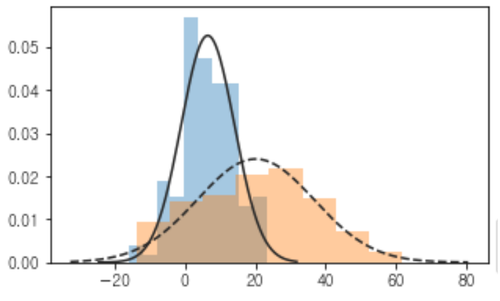
귀무가설 : 정상기업과 부도기업의 부도기사비율 사이의 평균의 차이가 없다.
대립가설 : 정상기업과 부도기업의 부도기사비율 사이의 평균의 차이가 있다.

☑ 등분산 검정 Levene' Test

귀무가설 : 등분산이다.
대립가설 : 이분산이다.

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 이분산이다



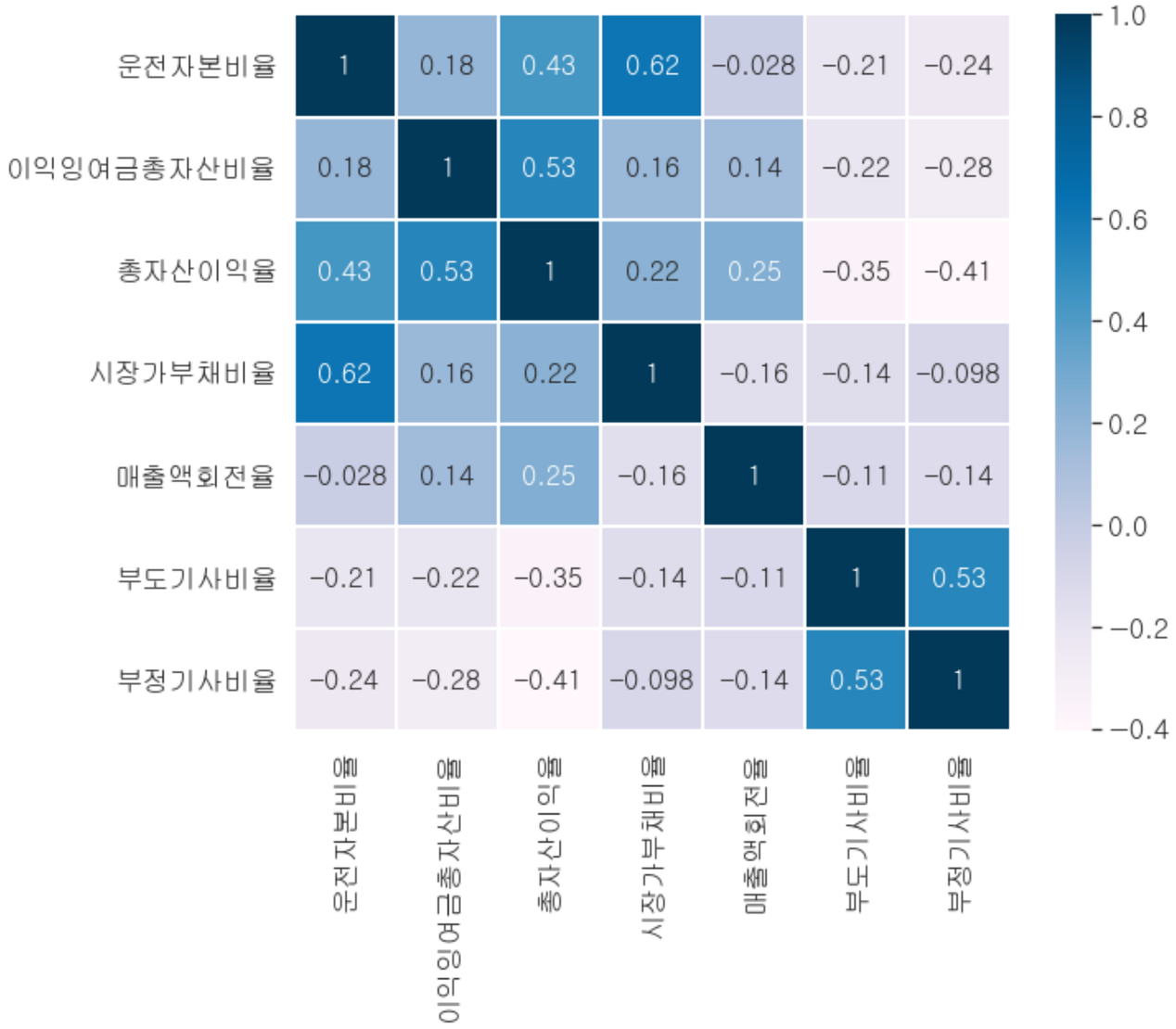
☑ Welch's T-test

p-value : 0.0000 < 0.05

귀무가설 기각 ⟶ 차이가 있다.

결론 : 정상기업과 부도기업의 부도기사비율의 평균차이가 있다.

03
변수 상관관계 -히트맵

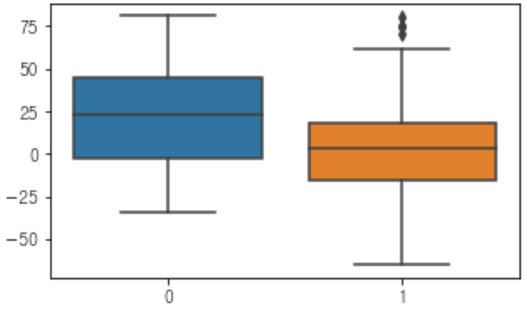


변수간의 상관성이 높지 않음

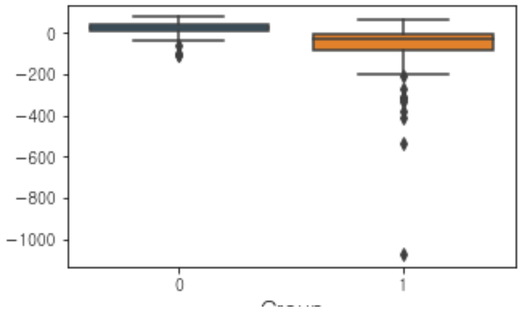
다중공선성 문제가 없음

03
데이터 분포

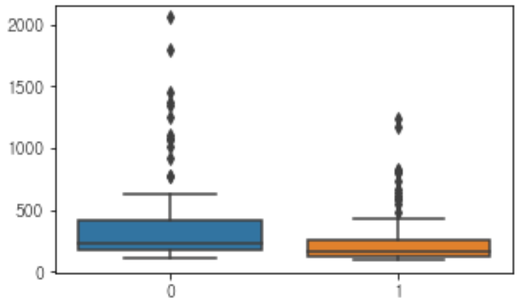
운전자본비율



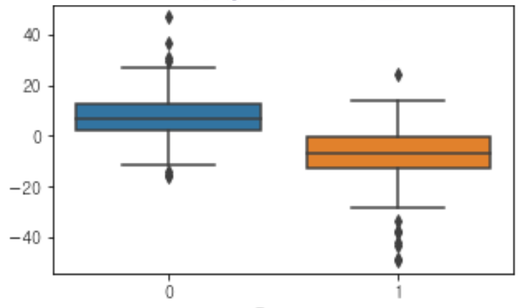
이익잉여금총자산비율



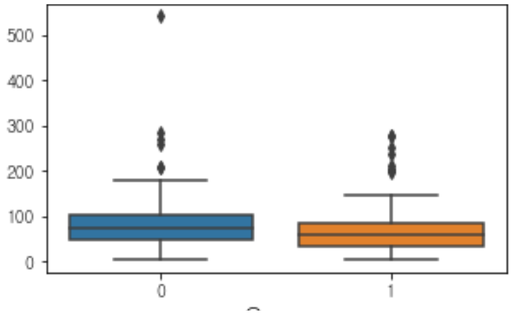
총자산이익율



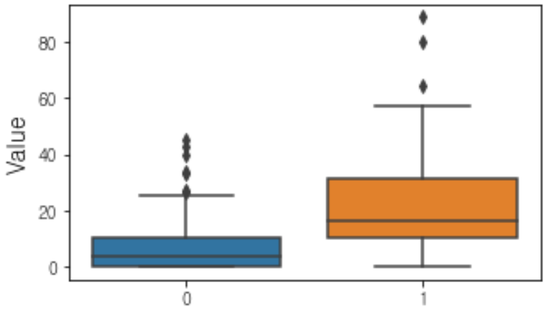
시장가부채율



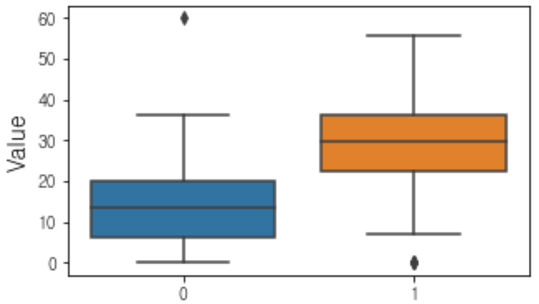
매출액회전율



부도기사 비율



부정기사 비율



값이 치우쳐져 있어 이상치 처리 → Robust Scaling

03
최종 데이터 세트

재무비율

회사명	운전자본비율	이익잉여금 총자산비율	총자산이익율	시장가 부채비율	매출액회전율
네이쳐 글로벌	74.492049	-49.803589	-3.371182	673.432622	6.028268
네프로 아이티	-0.465154	19.33815	13.884726	157.344363	76.780711
트루아워	52.368738	-19.102365	-25.573719	412.297157	33.436299
지니뮤직	45.063046	-14.623948	4.450491	262.390477	101.096537
앤디포스	81.281282	35.275974	18.934118	1071.39392	79.493948
에이치 엘비	31.643601	13.030604	0.544161	321.677578	11.901687

부도기사비율

부도기사비율
10.810811
38.235294
46.341463
4.761905
0
5.714286

부정기사비율

부정기사비율
29.333333
29.807692
27.380952
4.761905
8.77193
22.857143

DATA SET (1) : 재무 비율
DATA SET (2) : 재무 비율 + 부도 기사 비율
DATA SET (3) : 재무 비율 + 부정 기사 비율

04

프로젝트 수행 결과

01. 모델링 및 성능평가
02. 결론
03. 모델 활용

모델링 성능 및 평가

04
모델링 및 성능평가

성능 평가 지표: ROC_AUC_SCORE

	Logistic	RandomForest	SVM	XGBoost	CatBoost
DATA SET (1)	90.3	90.3	90.3	89.1	83.4
DATA SET (2)	93.7	92.5	93.8	92.6	81.5
DATA SET (3)	94.3	94.5	94.4	92.6	81.6

최종 데이터셋 : DATA SET(2)

최종 모델 : SVM

04
모델링 및 성능평가

단일 텍스트 데이터 컬럼 모델

단일 텍스트 데이터셋들에도 SVM이 가장 높은 성능을 띔

	Logistic	RandomForest	SVM	XGBoost	CatBoost
부도기사비율	82.9	82.0	83.0	79.5	66.3

	Logistic	RandomForest	SVM	XGBoost	CatBoost
부정기사비율	89.1	88.6	89.7	88.1	66.5

04
모델 평가 강건성 검증

부도 기업 수 : 20

정상 기업 수 : 20

전체 기업 수
40

맞힌 부도 기업 수 : 18

맞힌 정상 기업 수 : 16

맞은 기업 수
34

ROC_AUC_SCORE : 0.9025

Accuracy : 0.85

04
결론

- 텍스트 데이터만으로도 충분히 예측 가능
- 재무비율과 같이 사용함으로써 예측력 향상

모델 활용

텍스트 분류

- C-LSTM을 통한 텍스트 분류기

04
텍스트 분류

1. 재무데이터는 5가지 계산 후 DB에 저장
2. DB에 저장된 재무데이터와 기업명 가져옴
3. 기업명을 가지고 크롤링
4. 수집된 뉴스 전처리
5. 부도기사 분류 모델 (C-LSTM) 뉴스넣고 부도기사 여부 레이블
6. 부도기사여부로 기업의 부도기사비율 산출
7. 재무데이터에 부도기사비율 붙여서 학습한 모델에 넣고 예측
8. 결과 도출

04

텍스트 분류

☑ 문장에서는 단어의 뜻 뿐 아니라 단어 순서 중요

→ CNN은 문장의 지역 정보를 보존하여
단어의 등장순서를 학습에 반영

☑ 기사본문에서는 더 많은 문맥 필요

→ 장기의존성 문제를 보완해주는 LSTM 을 활용

“C-LSTM을 활용하여 기사 본문을 학습한다면
빠르고 정확하게 부도기업여부 예측가능”

04
텍스트 분류 - LSTM

단어 벡터의 최대길이100

기사 본문 토큰화 후 워드 임베딩

Drop out(0.5)

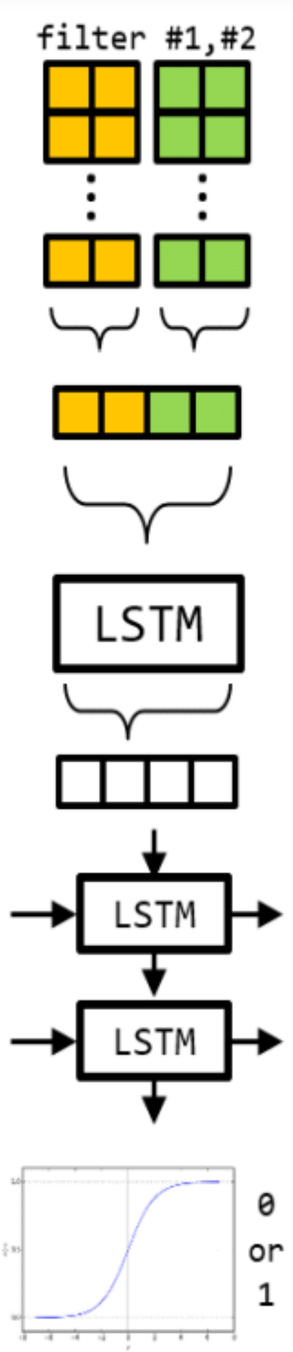
Conv1D layer

MaxPooling (pool size 4)

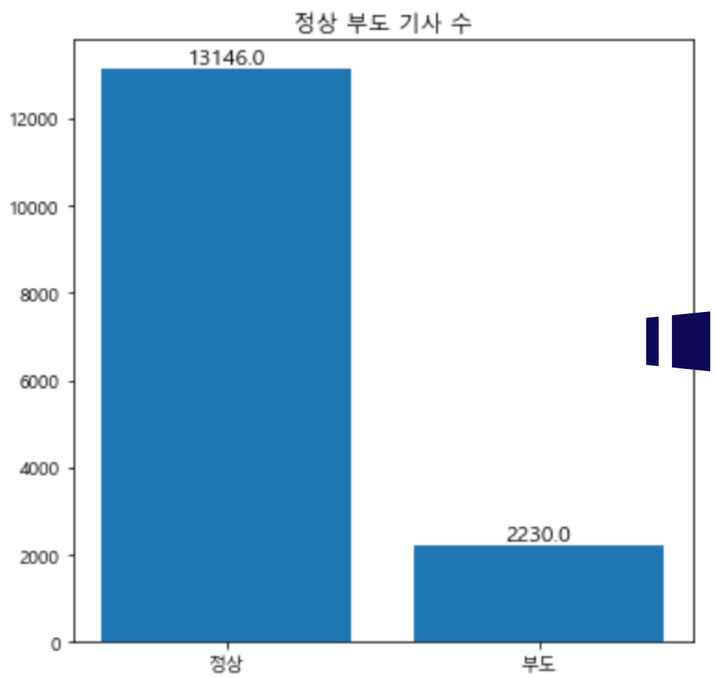
CNN의 장기의존성 보완

LSTM layer

Sigmoid



04
텍스트 분류



불균형데이터 샘플링 적용

기법	종류
UNDER SAMPLING	EditedNearestNeigghbours, RepeatedEditedNearestNeigghbours, AllKNN,NearMiss,OneSideSelection, NeighbourhoodCleaningRule,TomekLinks
OVER SAMPLING	SMOTE,BorderlineSMOTE,ADASYN
COMBINE SAMPLING	SMOTEENN,SMOTETOMEK, 파이프라인 이용 복합샘플링

제일 loss가 적었던 TomekLinks로 샘플링한 데이터로 학습한 모델을 최적 모델로 선정

f1-score : 0.896

05 자체평가기준

01. 한계점
02. 느낀점

05

한계점

1 소규모 기업 대상 적용 어려움

2 불균형 데이터 세트에 부적합 가능성

3 텍스트 수집 제한

05
느낀점

시간의 한계

많은 재무비율을 수집하여 피쳐셀렉션 과정을 통해 최적의 재무비율을 선정하려 했으나 5가지로 제한함

텍스트 데이터 수집이 기업과 페이지수가 제한하여
크롤링 과정에서 시행착오가 많았습니다.

감성사전을 직접 작성하지 못해 한국어 감성사전을 사용하지 못함

참고논문

06
참고논문

최정원, 오세경, 장재원. (2017). 빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구. 2017년 한국재무학회 추계학술대회

Flavio Barboza, Herbert Kimura, Edward Altman, Machine learning models and bankruptcy prediction, Expert Systems with Applications, Volume 83, 2017, Pages 405-417, ISSN 0957-4174

최정원, 한호선, 이미영, 안준모. 2015. 텍스트마이닝 방법론을 활용한 기업 부도 예측 연구. 생산성논집(구 생산성연구)

남기정, 이동명 and 진로. (2019). 비재무정보를 이용한 창업기업의 부실요인에 관한 실증연구. 벤처창업연구

하만석, 안현철.(2019).정형 데이터와 비정형 데이터를 동시에 고려하는 기계학습 기반의 직업훈련 중도탈락 예측 모형.한국콘텐츠학회논문지

김승수, 김종우.(2018).비정형 정보와 CNN 기법을 활용한 이진 분류 모델의 고객 행태 예측. 지능정보연구

김정미, 이주홍.(2017).Word2vec을 활용한 RNN기반의 문서 분류에 관한 연구.한국지능시스템학회 논문지

06
참고논문

김도우, 구명완.(2016).Doc2Vec을 활용한 CNN기반 한국어 신문기사 분류에 관한 연구.한국어정보학회 학술대회

Hwiyeol Jo, Jin-Hwa Kim, Kyung-Min Kim, Jeong-Ho Chang, Jae-Hong Eom, Byoung-Tak Zhang.(2017).Large-Scale Text Classification with Deep Neural Networks.정보과학회 컴퓨팅의 실제 논문지

주명길 and 윤성욱. (2019). 워드 임베딩과 CNN을 사용하여 영화 리뷰에 대한 감성 분석. (사)디지털산업정보학회 논문지

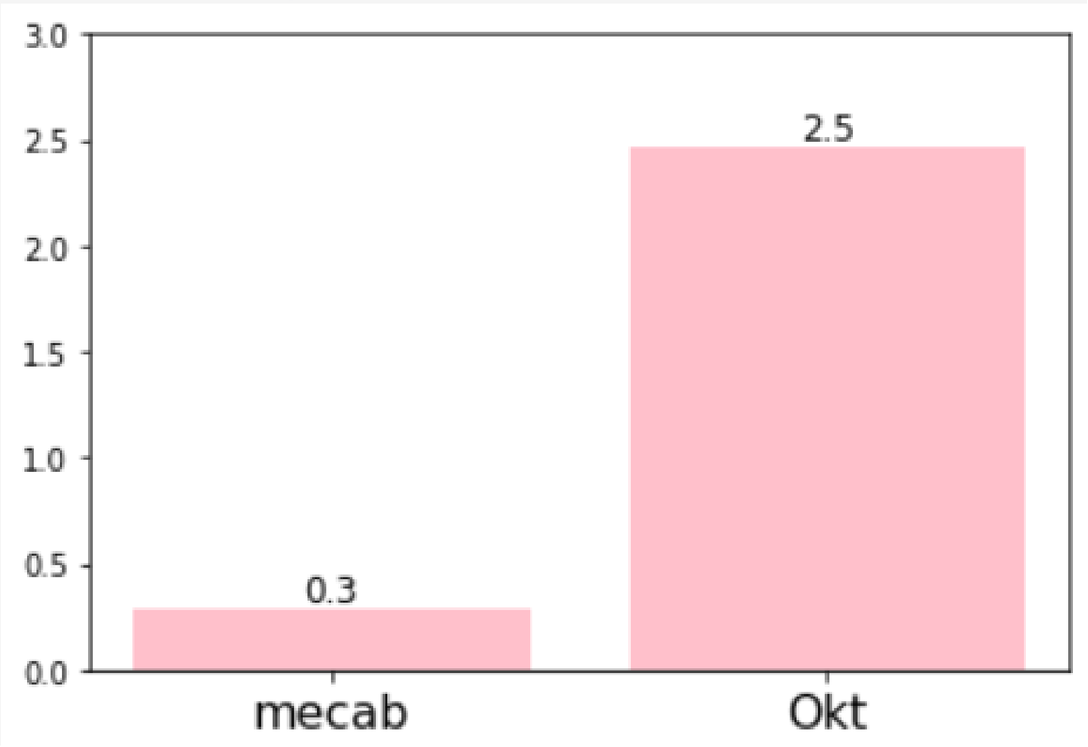
김찬송 and 신민수. (2019). 부도예측 모형에서 뉴스 분류를 통한 효과적인 감성분석에 관한 연구. 한국IT서비스학회지

03

뉴스 데이터 전처리

Mecab VS Okt

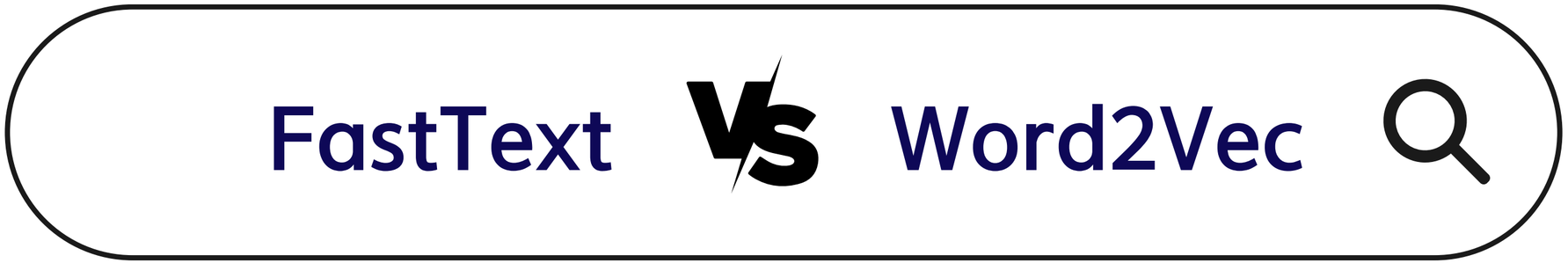
10만 문자의 문서를 대상으로 한 집단을 형태소로 분석했을 때 실행하는데 소요되는 시간(초)



형태소를 분석하는 방법 (거의 비슷하지만 약간의 차이)

Mecab	Okt
아버지 / NNG	아버지 / Noun
가 / JKS	가방 / Noun
방 / NNG	에 / Josa
에 / JKB	들어가신 / Verb
들어가 / VV	다 / Eomi
신다 / EP+EC	

03
뉴스 데이터 전처리



ex) 패스트텍스트에서 birthplace(출생지)란 단어를 학습하지 않은 상태일때

하지만 다른 단어의 n-gram으로서 birth와 place를 학습한 적이 있다면 birthplace의 임베딩 벡터(Embedding Vector)를 만들어낼 수 있다. 이는 모르는 단어에 대처할 수 없었던 Word2Vec와는 다른 점이다.

Word2Vec은 학습하지 않은 단어에 대해서 유사한 단어를 찾아내지 못 했지만, 패스트텍스트는 유사한 단어를 계산해서 출력하고 있음을 볼 수 있다.

04
모델 평가

2021년도 부실기업과 정상기업 부도예측

회사명	회계년도	운전자산총자본비율	이익잉여금총자산비율	총자산이익율	시장가부채비율	매출액회전율	부도기사비율	부도기업여부	예측
금빛	19-Dec	-29.204	-47.253	-16.201	143.204	57.027	0.000	1	1
맥스로텍	19-Dec	-7.785	-20.590	-11.272	159.409	49.384	20.000	1	1
미래SCI	19-Dec	0.371	-180.296	-9.221	145.128	27.199	33.333	1	1
아이엠텍	19-Dec	-26.400	-81.316	-5.069	146.827	25.509	37.500	1	1
에스제이케이	19-Dec	-52.182	-122.705	-8.280	131.571	7.453	20.000	1	1