기업 위기 뉴스 텍스트마이닝을 통한 주식수익률 극대화 모델 구축

- 효과적인 주식매매를 위한 위기뉴스 식별과 위기뉴스 군집분석 방법론 제작 -

팀명	앙상블
팀원	경희대학교 경영학과 박규리 이화여자대학교 통계학과

- 목차 -

1. 배경 및 목적
2. 알고리즘 순서도
3. 분석의 개요
4. 활용 데이터 수집
5. 분석
6. 결론 및 제언

1. 배경 및 목적

1) 분석의 배경

상장한 기업에 위기가 발생하게 되면 주식자체는 사람들의 심리적 요소가 반영되기 때문에 일시적 위기든 구조적 위기든 주가는 하락한다. 하지만 다른 점은 일시적위기는 회복이 가능하다는 점이고 구조적위기는 회복하지 못한다는 점이다.

	기업의 위기	
유형	**1. 일시적 위기**	**2. 구조적 위기**
개요	기업에게 치명적인 위기지만, 기업의 생존에는 큰 영향이 없어 다시 원래 상태로 돌아오는 위기	주가의 회복이 어려우며 기업의 생존에 치명적인 위기
예시	※ '삼성전자'의 갤럭시 노트7 폭발 사건 • 2016년 10월 삼성전자의 주가는 갤럭시 노트7 폭발 사건으로 인해 7% 가량 하락 • 하지만 위기에도 불구하고 빠른 시간 내에 최고가에 가까운 159만 원대로 회복	※ '남양유업'의 갑질 사건 • 2013년 대리점주를 향한 욕설과 물량 밀어내기 파동으로 국민적 공분을 사고 불매운동 수모까지 당한 남양유업의 주가는 1년 만에 반토막 • 2013년 5월 110만 원대 주가는 현재 60만 원대

2) 분석의 목적

이러한 생각을 바탕으로 증권뉴스나 네이버뉴스를 크롤링하여 텍스트마이닝 후, 기업의 위기상황이 나타난 경우를 골라 일시적위기인지, 구조적위기인지 판단하는 군집모델을 만드는 것을 목표로 한다. 위기의 유형을 분류하는 군집모델을 활용하여 주식모델을 만들 수 있을 것이다.

2. 알고리즘 순서도

3. 분석의 개요

주제	룰베이스를 활용한 기업의 위기 유형 군집분석
목적	기업뉴스 텍스트마이닝을 통해 기업의 위기상황을 파악 후 주식 수익률 극대화 모델 제작
분석 방법	파이썬 및 R 프로그램 활용
분석 과정	1) 기업뉴스 크롤링 ▶ 2013년 ~ 2018년 기간 동안의 네이버금융 장중특징주 기업뉴스 60,000여개 크롤링[1)] ▶ 기사제목 / 날짜 / 기사본문 / URL 2) 위기기사와 호재기사 분류 ▶ 기업뉴스의 제목을 기반으로 기업의 호황과 악재를 구분할 수 있는 단어 분류 ▶ 룰베이스를 통한 텍스트마이닝을 통해 뉴스 제목의 주요 단어를 추출한 후 분류한 단어를 활용하여 기업뉴스가 호재기사인지 위기기사인지 파악 3) 어떤 기업의 기사인지 분류 ▶ 기사본문 속 종목 코드 및 기사 제목을 분석하여 어떤 기업에 해당하는 기사인지 판단 4) 위기의 유형 군집분석 ▶ 위기로 판단된 기업뉴스들에 위기 유형을 라벨링 ▶ #실적부진 #미국이슈 #비리의혹 5) 기사에 해당하는 기업 재무제표 크롤링 후 병합[2)] ▶ 기업의 규모를 나타내는 매출액, 자산총액, 기업의 실속을 나타내는 영업이익률, 기업의 역사를 나타내는 설립년도, 주식의 PER를 직접적으로 나타내는 EPS, 기업의 안정성을 나타내는 부채비율만 뽑아 병합 5) 기업의 위기 발생 후 주가 최저점 도달 시기 계산 ▶ 중복된 기사를 제거한 후 기업의 위기를 나타내는 기사가 보도되었을 때, 주가가 하락하여 최저점에 도달하는 시기를 계산 6) 최저점 도달 후 원래의 상태로 회복하는 시기 계산 ▶ 기업의 주가가 최저점에 도달한 후 위기기사가 보도되기 전날의 주가 상태를 회복하는 데 걸리는 시간 계산

1) 출처 : https://finance.naver.com/news/market_special.nhn
2) 출처 : https://finance.naver.com/

4. 활용 데이터 수집

1) 기업뉴스 크롤링

▶ 사람들이 기업의 위기를 판단할 때, 대부분 인터넷 기사를 보고 생각하리라고 판단하였고, 그리하여 가장 영향력을 가진 네이버 금융 장중특징주 뉴스란을 크롤링하여 분석하였다. 다른 증권 뉴스도 있지만 장중특징란 뉴스가 특정한 사건을 표현한 기사가 주를 이루기 때문이다.

2) 기업 재무제표 크롤링

▶ 기업의 위기를 나타내는 기사가 보도되었을 때, 그 기업의 주가 변화를 예측하는 데에 매출액, 시가총액 등 기업의 규모가 중요하게 작용할 것이라고 판단하였다. 가령, 삼성전자와 같이 규모가 큰 대기업의 경우 위기가 발생하더라도 주가가 오히려 상승하거나 빠른 시기 내에 회복하는 사례가 많이 있다. 따라서 위기 기사가 나타내는 기업의 재무정보를 함께 활용하고자 네이버금융에서 크롤링을 실시하였다.

3) 기업명(종목명)과 종목코드 연결

▶ 제공받은 주식 데이터의 종목명과 거래소에 등록되어 있는 종목명이 다르다는 것을 발견하였다. 띄어쓰기의 차이 및 영어 표기 방식 등의 차이로 인해 종목명을 KEY로 잡기에는 무리가 있다고 판단하여 분석의 용이성을 위해 종목코드 열을 추가하였다. 종목명과 종목코드, 업종, CEO 등 간단한 기업정보가 포함되어 있는 데이터를 수집하였고 종목명이 다르게 표기되어 있는 경우, 직접 수작업으로 종목명을 같게 만들었다. 그리하여, R의 merge함수를 활용하여 주식 데이터에 종목 코드를 병합하였고 후에 기업을 판단할 때, 종목코드를 주요 KEY로 지정하였다.

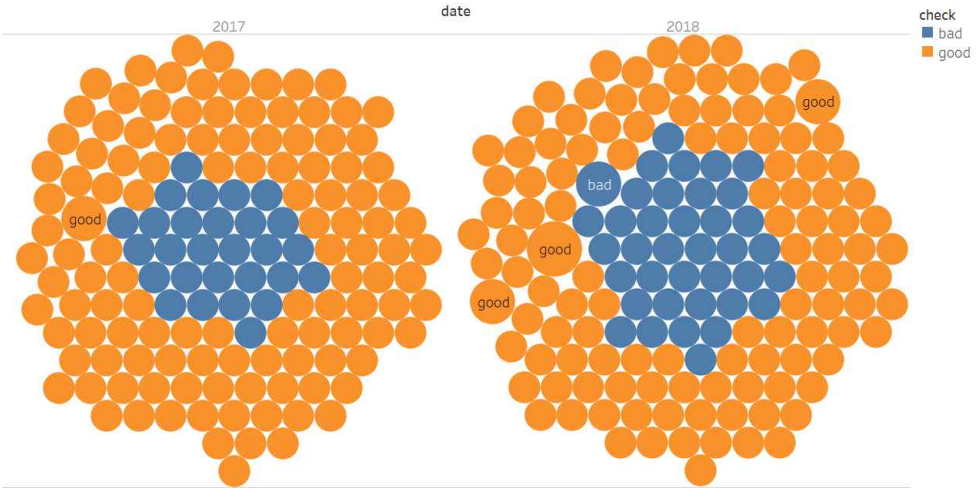
5. 분석

1) 위기기사와 호재기사 분류

<table>
<tr><td>1</td><td>▶ 기업뉴스의 제목을 기반으로 기업의 호황과 악재를 구분할 수 있는 단어 분류
네이버금융 장중특징주 기사 목록을 분석한 결과, 기업 주가의 변화에 대해 자주 사용되는 단어 표현들이 존재한다는 것을 알 수 있었다. 가령, 기업의 위기를 나타내는 단어에는 주식이 ‘곤두박질’, ‘하락세’, ‘내우외환’ 등이 있었고 호재를 나타내는 단어에는 ‘반등’, ‘신고가’ 달성, ‘성장세’ 등이 있었다. 이러한 분석 결과를 바탕으로 기업의 호황과 악재를 구분할 수 있는 단어사전을 구성하였다.
<table>
<tr><th>호재 단어</th><th>악재 단어</th></tr>
<tr><td>호재 / 반등 / 주목 / 상승 / 돌파 / 회복 / 성장 / 순매수 / 긍정 / 기대 / 호조 / 도약 / 급등 / 호실적 / 개선 / 강세 / 오름세 / 상승세 / 상한가 / 신고가 / 수혜 / 성장세 / 강보합 / 급증세 / ↑</td><td>위기 / 위축 / 휘청 / 하락 / 둔화 / 내리막길 / 내우외환 / 급락 / 부진 / 하향 / 불안 / 충격 / 우려 / 하락세 / 폭락 / 하회 / 신저가 / 감소 / 하한가 / 급락세 / 낙폭 / 미끄럼 / 곤두박질 / ↓</td></tr>
</table>
</td></tr>
<tr><td>2</td><td>▶ 텍스트마이닝을 통해 뉴스 제목의 주요 단어를 추출한 후 분류한 단어를 활용하여 기업뉴스의 성격 판단
크롤링한 기업뉴스 데이터의 기사제목을 텍스트 분석을 실시하였다. 분석방법으로는 R 프로그램의 텍스트 분석 패키지 ‘RHINO'를 활용하였고 기사 제목에서 주요 단어를 추출한 후, 사전에 구성한 호재와 악재 단어사전과 비교하여 해당 기사의 성격을 분류하였다. 명사 표현을 제외하고 주가의 상승과 하락을 표현하는 기사의 주요 표현인 ↑,↓ 화살표는 ’stringr' 패키지의 str_detect 함수를 활용하여 따로 분류하였다.</td></tr>
<tr><td>3</td><td>▶ 위기 기사 확인
Rule_based를 통해 텍스트 분석을 실시한 결과, 약 18,000개의 기업뉴스 데이터 중 2,700개의 데이터를 제외한 나머지 15,300개의 데이터를 기업의 위기와 호황 기사로 올바르게 분류하였다. 기사의 분포도는 호황 기사 12,000개, 위기 기사 3,300개로 기업의 호재를 나타내는 기사가 많았다.</td></tr>
</table>

	A	B	C	
1	lable	title1	date	news
2	4	금융그룹 통합감독 대상 대기업 금융사 하락	2018-07-02	(서울=연합뉴스) 권수현 기자
3	6	코스피·코스닥 급락에 증권주도 동반 하락세	2018-07-02	[머니투데이 이태성 기자] 국내
4	8	코스피·코스닥 급락...증권株도 하락세	2018-07-02	[이데일리 이명철 기자] 국내
5	9	정유화학株 약세...SK이노베이션 6%↓	2018-07-02	[아시아경제 임혜선 기자] 정유
6	10	조양호 회장 구속영장 청구에 한진그룹주 동반 신저가	2018-07-02	말없이 귀가하는 조양호 회장
7	11	POSCO, 증시 하락 영향에 2분기 이익 감소 전망에 4%↓	2018-07-02	[아시아경제 문채석 기자]POS
8	29	한국콜마, 1000억 유증 발표에 약세	2018-07-02	한국콜마가 대규모 유증에 따
9	30	제이콘텐트리, 1600억원대 유상증자 소식에 '하락'	2018-07-02	[이데일리 오희나 기자] 제이콘
10	32	바른손, 70억 규모 유상증자...약세	2018-07-02	바른손이 70억원 규모 유상증
11	34	제이콘텐트리, 1600억대 유상증자에 약세	2018-07-02	【서울=뉴시스】김형섭 기자 =
12	35	한국콜마, 1000억 규모 유상증자 결정에 약세	2018-07-02	【서울=뉴시스】이국현 기자 =
13	39	바른손, 대규모 증자 소식에 약세	2018-07-02	(서울=연합뉴스) 유현민 기자
14	43	바른손, 증자·CB발행 소식에 급락	2018-07-02	바른손이 유상증자 및 전환사
15	45	한국콜마, 1000억 규모 유상증자 결정...'하락'	2018-07-02	1000억원 규모의 유상증자를
16	47	제이콘텐트리, 1600억대 유증에 5%↓	2018-07-02	[아시아경제 문채석 기자]제이
17	49	제이콘텐트리, 대규모 유상증자 결정에 하락	2018-07-02	(서울=연합뉴스) 임은진 기자
18	50	제이콘텐트리, 대규모 증자에 '급락'	2018-07-02	제이콘텐트리가 대규모 증자를

▶ 위기 기사와 호재 기사 비율 (2017, 2018년도)



2) 어떤 기업의 기사인지 분류

1

▶ 기사본문 속 종목 코드를 추출하여 해당 기업명 병합

네이버금융 기업뉴스의 경우, 아래와 같이 기사의 본문 속 해당 기업의 종목코드가 포함되어 있다는 것을 발견하였다.

> [이데일리 이후섭 기자] 코아스템(166480)이 상한가를 나타내고 있다. 근위축성측삭경화증(루게릭병) 줄기세포치료제가 미국 식품의약국(FDA)로부터 희귀의약품에 지정됐다는 소식에 영향을 받은 것으로 보인다.

어느 기업을 주제로 한 기사인지 분류하는 데에 가장 중요한 부분이라고 판단하였고 기사의 본문에 종목코드가 있을 경우, 일순위로 기업명을 부여하였다. 또한, 하나 이상의 종목 코드

가 등장하는 기사의 경우 관련주, 계열사 등 중요 내용의 기사라는 점을 감안하여 포함하는 종목 코드의 기업명을 모두 나열하였다.

2

▶ 텍스트마이닝을 통해 기사 제목에서 기업명 추출

기사 본문에 종목 코드가 따로 나와 있지 않은 경우, 기사 내용을 파악하여 어느 기업에 해당하는 기사인지 분류하여야 한다. 기사 제목을 살펴본 결과, 아래와 같이 처음 단어 또는 콤마(,) 뒤의 단어가 기사가 정보를 주고자 하는 기업명이라는 것을 발견하였다.

따라서 공백과 콤마(,)를 기준으로 콤마 이후의 기사 제목의 첫 단어를 추출하여 대응하는 기업명을 찾아 부여하였다. (콤마가 존재할 경우 콤마를 기준으로, 존재하지 않을 경우 공백을 기준으로 하였다.)

3

▶ 기업명을 제대로 분류하였는지 확인

위 두 가지의 단계를 실시하여 해당 기사가 어느 기업의 기사인지 분류하는 작업을 완료하였고, 주식 데이터에 없는 기업이나 재무제표가 없는 기업 등 몇십 개의 기업을 제외하고 거의 모든 기사를 분류할 수 있었다.

	A	B	C	D	
1	label_all	cb_all	title1	date	news
2	1	LG유플러스	통신주, 통신비 인하 공약에 동반 약세	2017-04-12	[이데일리 유재흐
3	1	SK텔레콤	통신주, 통신비 인하 공약에 동반 약세	2017-04-12	[이데일리 유재흐
4	1	케이티	통신주, 통신비 인하 공약에 동반 약세	2017-04-12	[이데일리 유재흐
5	10	NAVER	NAVER, 외인 '팔자'에 3%↓	2017-04-11	[아시아경제 박니
6	100	아리온	아리온, 92억 규모 지분 매각계약 해지에 `약세`	2017-03-28	[이데일리 이후섭
7	10002	에이씨티	에이씨티, 상장 첫날 10%대 '급락'	2013-11-08	에이씨티가 코스
8	10006	삼성전자	삼성전자 1.53%↓..시총 상위주 일제히 하락 출발	2013-11-08	코스피 시가총액
9	10007	에이씨티	에이씨티, 상장 첫날 급락...공모가는 웃돌아	2013-11-08	[머니투데이 정인
10	10008	씨티씨바이오	씨티씨바이오, 3Q 부진한 실적에↓	2013-11-08	[머니투데이 김성
11	1001	씨에스윈드	신재생에너지株, 급락...트럼프 당선 가능성 커져	2016-11-09	[이데일리 박형수
12	1001	태웅	신재생에너지株, 급락...트럼프 당선 가능성 커져	2016-11-09	[이데일리 박형수
13	1001	현진소재	신재생에너지株, 급락...트럼프 당선 가능성 커져	2016-11-09	[이데일리 박형수
14	1001	OCI	신재생에너지株, 급락...트럼프 당선 가능성 커져	2016-11-09	[이데일리 박형수
15	1001	한화케미칼	신재생에너지株, 급락...트럼프 당선 가능성 커져	2016-11-09	[이데일리 박형수
16	10010	CJ	CJ CGV , 영업이익 둔화 영향.. 3.86% 하락	2013-11-07	CJ CGV가 영업이
17	10011	게임빌	게임빌, 게임중독법 제정 전망에↓	2013-11-07	게임을 마약, 술
18	10013	위메이드	위메이드, 나흘째 약세...'게임중독법' 규제 이슈	2013-11-07	위메이드가 '게임

3) 위기의 유형 군집분석

본 분석의 최종 목표를 위기의 유형 군집분석으로 하였으며, 군집분석 결과를 활용할 수 있는 방안에 대해서도 생각해보았다.

1	▶ 기업의 일시적 위기와 구조적 위기에 대한 사전 지식 필요 기업의 위기의 유형에는 어떤 것이 있는지, 주가에 영향을 주는 요인에는 어떤 것이 있는지 등 주식 시장에 대한 기본적인 스터디를 실시하였다. 사전 지식과 분류한 위기 기사 리스트를 통해 아래와 같이 기업의 주가가 하락한 상황의 유형을 분류하였다. 유상증자 / 생산중단 / 실적부진 / 기타 / 상장버블효과 / 파산위기 / 호황무산 / 국제유가위기 / 경쟁구도악화 / 제품불량 / 비리의혹 / 업종불황 / 투자의견하향 / 고점논란 / 재무부담 / 경영권분쟁 / 자회사 이슈 / 제도적위기 / 미국이슈 / 사업중단 / 북한이슈 / 대규모매도 / 소송 / 정치테마주 / 상장폐지 / 국내증시위기 등 총 57가지의 위기유형 분류
2	▶ 위기로 판단된 기업뉴스들의 위기 유형을 라벨링

	A	B	C	D	E	
1	label_all	category	cb_all	title1	date	news
2	1000	유상증자	한프	한프, 204억원 일반 유증 결정에 하락세	2018-06-11	한프가 204억원 규모의
3	10002	생산중단	파미셀	파미셀, 울산공장 생산 중단 소식에 하락	2017-12-14	[머니투데이 오정은 기자
4	1002	실적부진	펄어비스	펄어비스, 일매출 하락 우려에 4% 약세	2018-06-11	[머니투데이 하세린 기자
5	10023	기타	동부제철	동부제철, 2거래일 연속 하락세..8.11%↓	2017-12-13	동부제철이 2거래일 연
6	10037	인수중단	현대자동차	AJ렌터카, 매각 추진 중단 소식에 급락	2017-12-13	현대차그룹 매각설에 연
7	1004	유상증자	한프	한프, 대규모 증자 소식에 약세	2018-06-11	(서울=연합뉴스) 유현민
8	10046	수주중단	HSD엔진	두산엔진, '메가컨테이너선' 수주 불발 우려..약세	2017-12-13	[이데일리 함정선 기자]
9	10053	호황무산	금호타이어	금호타이어, 3거래일만에 다시 '하락'	2017-12-13	금호타이어의 주가가 3거
10	1006	실적부진	LG이노텍	LG이노텍, 아이폰 부품 주문 20% 감소 소식에 '약세'	2018-06-11	[이데일리 성선화 기자]LG
11	10064	실적부진	클리오	클리오, 2018년 실적 전망 공시에 '하락'	2017-12-12	[머니투데이 오정은 기자
12	10065	생산중단	인터플렉스	인터플렉스, "일부 생산라인 중단 여파 실적하향 불가피	2017-12-12	[아시아경제 임철영 기자
13	10073	생산중단	인터플렉스	인터플렉스, 일부 생산중단 여파에 주가도 하락	2017-12-12	인터플렉스가 일부 생산
14	10083	상장버블효과	BGF	BGF, 재상장 후 '반토막'...BGF리테일도 3일만에 하락	2017-12-12	기업분할 후 재상장한 지
15	10091	상장버블효과	BGF	분할상장 BGF, 3일 연속 하락세	2017-12-12	기업분할로 재상장한 BGF
16	10095	생산중단	인터플렉스	인터플렉스, 생산라인 중단으로 실적하향 분석에 '하락	2017-12-12	[이데일리 오희나 기자]
17	10097	생산중단	인터플렉스	인터플렉스, 점유율 하락 가능성...'약세'	2017-12-12	[머니투데이 오정은 기자
18	10106	상장버블효과	BGF	BGF, 인적분할 후 사흘 연속 급락...12%대 하락	2017-12-12	[머니투데이 오정은 기자

4) 기업의 위기 발생 후 주가 최저점 도달 시기 계산

1	▶ 중복된 기사 제거 위기기사 데이터 수집과 어느 기업의 기사인지 분류를 완료한 후 동일한 기업의 동일한 사건에 대한 기사가 있다는 것을 발견하였다. 특히 대기업의 비리의혹과 같은 사건의 경우 하루 안에도 다수의 기사가 보도되며, 주말을 제외하고 2-3일 내에도 동일한 기사가 보도된다. 주식 시장과 주식관련 뉴스는 주말과 공휴일을 제외한 평일에만 활동하므로 날짜 기준 3일 이내에 같은 기업의 기사가 보도되면 동일한 기사라고 판단하여 처음 보도된 위기 기사만 추출하였다.

9430	DMS	DMS, 147억 공급계약 해지에 8일만에 ↓	2014-01-08	[머니투데이 최경민 기자]147억원
9431	DMS	DMS, 약세...中 솔라셀장비 판매법인 청산	2014-01-08	[이하나 기자] DMS가 중국에서 태
5572	DMS	DMS, 적자 확대 소식에 '급락'	2015-03-09	[노정동 기자] DMS가 적자 확대 소
4909	DMS	[오늘 특징주]②NHN엔터, 약세...네오위즈 인수 우려?	2015-05-08	[이데일리 안혜신 기자] 다음은 8일
4906	DMS	[전일 특징주]②NHN엔터, 약세...네오위즈 인수 우려?	2015-05-09	[이데일리 안혜신 기자] 다음은 8일
785	DSC인베스트먼트	DSC인베스트먼트, 상장 첫날 하락세...공모가는 상회	2016-12-19	[머니투데이 김주현 기자] DSC인베
788	DSC인베스트먼트	DSC인베스트먼트, 상장 첫 날 하락...공모가는 웃돌아	2016-12-19	[박상재 기자] DSC인베스트먼트가
223	EG	EG, 탄핵 인용 결정과 함께 급락	2017-03-10	[아시아경제 박철응 기자]EG가 헌법
5564	EMW	EMW, 적자전환 소식에 ↓	2015-03-09	[아시아경제 최동현 기자] EMW가
5583	EMW	EWM, 급락...지난해 이익 적자로 전환	2015-03-09	[이데일리 경계영 기자] EMW(0791

↓

9430	DMS	DMS, 147억 공급계약 해지에 8일만에 ↓	2014-01-08	[머니투데이 최경민 기자]147억원
5572	DMS	DMS, 적자 확대 소식에 '급락'	2015-03-09	[노정동 기자] DMS가 적자 확대
4909	DMS	[오늘 특징주]②NHN엔터, 약세...네오위즈 인수 우려?	2015-05-08	[이데일리 안혜신 기자] 다음은 8
785	DSC인베스트먼트	DSC인베스트먼트, 상장 첫날 하락세...공모가는 상회	2016-12-19	[머니투데이 김주현 기자] DSC인
223	EG	EG, 탄핵 인용 결정과 함께 급락	2017-03-10	[아시아경제 박철응 기자]EG가 헌
5564	EMW	EMW, 적자전환 소식에 ↓	2015-03-09	[아시아경제 최동현 기자] EMW가

2

▶ 위기기사가 보도된 후, 주가의 최저점 찾기

기업의 위기를 나타내는 기사가 처음 보도되었을 때, 위기의 유형이 일시적인 위기인지 구조적 위기인지 판단한다. 주가가 하락한다는 가정 하에 최대의 이익을 위해 주가의 최저점으로 매수 지점을 잡아야한다.

기사가 보도된 이후 10일 동안의 주가 변화를 살펴보았을 때,

(1) 최저점이 첫날인 경우, 위기의 상황에도 주가는 상승한다고 판단하여 매수를 진행하지 않는다.

(2) 최저점이 1-3일인 경우, 충분히 주가가 하락하지 않아 이익이 나지 않는다고 판단하여 매수를 진행하지 않는다.

(3) 최저점이 4-8일인 경우, 기업의 위기 사건에 영향을 받아 주가가 하락하여 회복이 가능하다면 매수의 기회가 있기에 해당 기업의 주식을 매수한다.

(4) 최저점이 9일 이상인 경우, 계속해서 기업의 주가가 하락하여 회복이 불가능하다고 판단하여 매수를 진행하지 않는다.

3

▶ 최저점을 제대로 계산했는지 확인

title1	date	news	low
우리은행, 외국인 매도세 전환에 '하락'	2017-07-27	우리은행이 외국인의 매도세 전환에 하락하고 있다.27일 오...	infinity
은행주, 연체금리 규제 리스크 부각에 3%대 급락	2017-09-26	[이데일리 최정희 기자] 은행주가 금융당국의 연체금리 규제 ...	infinity
美 국채금리 하락에 은행주 약세	2018-04-03	[이데일리 최정희 기자] 무역분쟁 우려 등에 안전자산 선호심...	2018-04-04
KR모터스, 300억 규모 BW 발행 소식에↓	2017-06-26	[이데일리 이명철 기자]?#KR모터스가 300억원 자금 조달을 ...	2017-06-28
하이트진로, 노조 파업 따른 생산중단 소식...9%↓	2017-10-16	[아시아경제 임철영 기자]하이트진로가 노조 파업에 따른 생...	2017-10-18
유한양행, 예상 못미치는 실적에 급락	2017-07-28	[아시아경제 권성회 기자] 유한양행이 시장 예상치에 미달하...	2017-08-04
유한양행, 3Q 실적 부진 전망에 하락세	2017-10-10	[이데일리 윤필호 기자] 유한양행(000100)이 올해 3분기 실적...	2017-10-13
유한양행, 지난해 수익성 부진...'하락'	2018-02-21	[머니투데이 오정은 기자] 유한양행이 지난해 수익성 악화로 ...	infinity
CJ대한통운, 합병 우려감에 '약세'	2017-12-20	[머니투데이 조한송 기자] CJ대한통운이 CJ건설과의 합병으...	2017-12-22
CJ대한통운, 수익성 우려에 신저가	2018-01-23	CJ대한통운이 영업이익률 하락 우려에 신저가를 새로 썼다.2...	increase
CJ대한통운, 아시아나항공 블록딜에 하락세	2018-03-16	[아시아경제 조호윤 기자]CJ대한통운 주가가 아시아나항공의...	2018-03-23
"CJ대한통운, 최저임금 인상 타격 크다"...목표가↓	2018-03-27	자료 제공 = 유진투자증권유진투자증권은 27일 CJ대한통운...	2018-03-28

5) 최저점 도달 후 원래의 상태로 회복하는 시기 계산

1	▶ 주가 회복의 기준 설정 기업의 위기 기사가 보도되면 주가가 하락한다는 가정 하에, 매수와 매도의 기회를 잡아 이익을 내기 위해 주가가 회복했다고 판단하는 기준을 정해야한다. 본 분석에서는 처음 위기기사가 보도된 날의 전날 종가를 기준으로 정하였다.
2	▶ 주가 회복지점 찾기 기사가 보도된 이후 40일(최저 도달 기간 10일 + 회복 기간 30일) 동안의 주가 변화를 살펴보았을 때, (1) 회복 지점이 1-8일인 경우, 최저점에 도달했다가 회복하는 기간이 충분히 길지 않다고 판단하여 이익이 나지 않아 매수 및 매도를 진행하지 않는다. (2) 회복 지점이 9-30일인 경우, 최저점에 도달한 후 기업의 규모 및 노력에 의해 주가가 회복했다고 판단하여 매도를 진행한다. (3) 회복 지점이 30일 이상인 경우, 기업이 마주한 위기를 극복하지 못한다고 판단하여 매도 및 매수를 진행하지 않는다.
3	▶ 위기 기사가 터진 처음 날부터, 10일 안에 주가가 최저점에 도달하기까지 걸린 날짜를 계산 후, 그 때부터 원래 가격까지 회복한 날짜 또한 계산하였다.

EPS	buildingyear	Psell	amount	low	recovery	recovery_day	low_day
1866	1899	350.000000	1926614	infinity	infinity	infinity	infinity
1866	1899	100.000000	2004370	infinity	infinity	infinity	infinity
2237	1899	100.000000	1880966	2018-04-04	2018-04-10	5	1
-89	1917	3.009231	1383529	2017-06-28	2017-07-06	8	2
538	1954	-50.000000	879409	2017-10-18	2017-10-27	9	2
12956	1936	16000.000000	132263	2017-08-04	infinity	infinity	5
12956	1936	1500.000000	50155	2017-10-13	2017-10-18	6	3
8761	1936	-1000.000000	42923	infinity	infinity	infinity	infinity
2446	1930	7000.000000	91757	2017-12-22	2018-04-19	80	2
1380	1930	1000.000000	72851	increase	increase	increase	increase

6) 주가 최저점, 회복시기 찾는 모델 적합

1	▶ 주가 최저점, 회복시기 찾는 모델 제작 기업의 위기 기사가 보도되면 주가가 하락하고, 하락했을 때 회복한다는 것을 미리 예상한다면, 주가가 최저점이 도달하는 시기에 주식을 매수하고, 회복하면 매도한다면, 수익을 얻을 수 잇을 것이므로, 주가의 최저점, 회복시기를 찾는 모델을 찾는다

2	▶ 주가 최저점 찾는 모델1 독립변수 : 그 기업의 자산총액, 매출액, 거래량, 매도의 힘(시가 - 종가), 거래량, EPS, 부채비율, 영업이익률, 매도의 힘의 비율 (매도의 힘/시가) 종속변수 : 그 기업의 위기 기사가 등장했을 때 10일 안에 며칠만에 최저점에 도달하는가 종속변수 명목화된 구간 1) 1일-3일 4) 4일-8일 9) 9일(무한대) ▶ 주가 회복시기 찾는 모델2 독립변수 : 그 기업의 자산총액, 매출액, 거래량, 매도의 힘(시가 - 종가), 거래량, EPS, 부채비율, 영업이익률, 매도의 힘의 비율 (매도의 힘/시가) 종속변수 : 그 기업이 최저점에 도달했을 때 10일 안에 며칠만에 그 기업의 위기기사 하루 전 본래 주식 가격으로 도달하는가 종속변수 명목화된 구간 1) 1일-8일 2) 9일-30일 3) 31일 이상(무한대)
3	▶ 위기 기사가 터진 처음 날부터, 10일 안에 주가가 최저점에 도달하기까지 걸린 날짜를 계산 후, 그 때부터 원래 가격까지 회복한 날짜 또한 계산하였다. ▶ R 프로그램의 'caret' 패키지의 train 함수를 활용하여 randomForest, nnet.ddn, logistic regression, XGboost, GBM 등 다양한 모델을 적합하였다. 가장 정확도가 높은 랜덤포레스트 모델을 최종 모델로 선정하였다.

7) 기사 카테고리 라벨링을 자동으로 하는 모델 제작

1	▶ 기사 카테고리 라벨링을 자동으로 하는 모델 제작 기사 카테고리를 직접 분류하는 것은 여러모로 시간과 비용이 많이 소모되므로, 기사 호황과 불황을 나누는 모델과 비슷하게 기사의 제목 패턴을 찾아 기사 카테고리 라벨링을 자동으로 하는 모델을 제작한다
2	▶ 비리의혹 기사 카테고리 찾는 모델 기사 제목에 다음과 같은 단어가 나타나면 '비리의혹' 기사라고 판단한다 '압수','감사','위반','혐의','검찰', '구속','분식','감리','의혹','수색','뇌물','횡령','조사','비리','갑질', '수사','검사' ▶ 실적부진 기사 카테고리 찾는 모델 기사 제목에 다음과 같은 단어가 나타나면 '실적부진' 기사라고 판단한다 실적','이익','매출액','매출','부진','매출','적자','손실' ▶ 미국이슈 기사 카테고리 찾는 모델 기사 제목에 다음과 같은 단어가 나타나면 '미국이슈' 기사라고 판단한다 '트럼프','미국','美' ▶ 북한이슈 기사 카테고리 찾는 모델 기사 제목에 다음과 같은 단어가 나타나면 '북한이슈' 기사라고 판단한다 '경협','북한','北','대북','탄도','미사일','안보','북핵'

3	▶ 모델이 잘 분류하는지 확인
	`> table(real=pre_data_str$category,predict=pre_data_str$lable_pred)[c(4,5,6,8),]`
	`predict`
	`real 미국이슈 북한이슈 비리의혹 실적부진`
	`미국이슈 69 23 0 1`
	`북한이슈 0 85 0 0`
	`비리의혹 0 0 69 0`
	`실적부진 5 0 0 295`

6. 결론 및 제언

1) 분석 결과 - 주식 모델 적합 결과

<table>
<tr><td>1</td><td>

▶ 주식 모델 적합 결과
네이버 금융에서 18000여개를 가지고 크롤링 후, 상기된 바와 같이 주식 정보와 재무제표, 최저점과 최고점을 계산하여 병합하였다. 18000여개에서 불황인 기사 중, 재무제표와 주식 정보가 있는 정보만을 추출하면 1400여개의 기사들이 나온다. 이 데이터를 가지고 training 모델을 만들고, 추가로 60000여개의 기사를 추출하여 동일한 방식을 취한 후 4400여개의 데이터를 가지고 test하였다.

training방식과 test방식을 더욱 자세히 기술하면 다음과 같다.
1. 위기기사가 터진 후, increase 즉, 가격이 올라가는 주식은 고려하지 않는다. 왜냐하면 투자자들은 위기기사 후 주식가격이 올라가는 주식은 구매하지 않을 것이기 때문이다
2. 주식모델을 가지고 예측했을 때 최저점이 infinite(31일 이상)인 주식은 고려하지 않는다. 손실을 당할 가능성이 높기 때문이다. 최저점이 1-2일 인 주식 또한 고려하지 않는다. 최저점이 빠르게 도달하므로 이익을 얻을 가능성이 적기 때문이다.
3. training에서 회복점을 찾을 시, class 불균형이 심하여 downsampling을 실시하였다.

[주가 최저점 찾는 모델 결과]
<pre>
> confusionMatrix(pred.rf4, analysis_data_add_pre4$low_day2)
Confusion Matrix and Statistics

          Reference
Prediction 1-2 3-8 infinity
  1-2   39 116   38
  3-8   713 1877   656
  infinity   39   74   30

Overall Statistics

               Accuracy : 0.5433
                 95% CI : (0.5268, 0.5597)
    No Information Rate : 0.5771
    P-Value [Acc > NIR] : 1
</pre>
결과적으로 3개의 클래스에서 정확도가 54%가 나왔으므로, 높은 정확도를 보이는 모델은 아니다.

</td></tr>
</table>

```
Statistics by Class:

                     Class: 1-2 Class: 3-8 Class: infinity
Sensitivity             0.04930    0.90808        0.041436
Specificity             0.94482    0.09637        0.960462
Pos Pred Value          0.20207    0.57825        0.209790
Neg Pred Value          0.77811    0.43452        0.798197
Prevalence              0.22083    0.57705        0.202122
Detection Rate          0.01089    0.52401        0.008375
Detection Prevalence    0.05388    0.90620        0.039922
Balanced Accuracy       0.49706    0.50222        0.500949
>
```

다만, 3-8클래스에서 Sensitivity가 0.90808로 가장 높은 것은 눈여겨볼만 하다.

```
> confusionMatrix(pred.rf4, analysis_data_add_pre4$low_day2)
Confusion Matrix and Statistics

          Reference
Prediction  1-2  3-8 infinity
  1-2        39  116       38
  3-8       713 1877      656
  infinity   39   74       30

Overall Statistics

               Accuracy : 0.5433
                 95% CI : (0.5268, 0.5597)
    No Information Rate : 0.5771
    P-Value [Acc > NIR] : 1
```

$$Sensitivity = \frac{TP}{TP+FN} \quad (TPR:\ True\ Positive\ Rate)$$

Sensitivity는 실제로 3-8에 최저점인 것을 얼마나 3-8로 분류했는가를 나타낸다. 투자자들 입장에서는 3-8에 최저점이 도달한다면 손해가 날 가능성이 가장 적은 상태이므로 3-8을 선택하여 투자할 것이다. 3-8에서의 Sensitivity가 가장 높으므로, 투자자들이 참고할만한 모델이라고 할 수 있다.

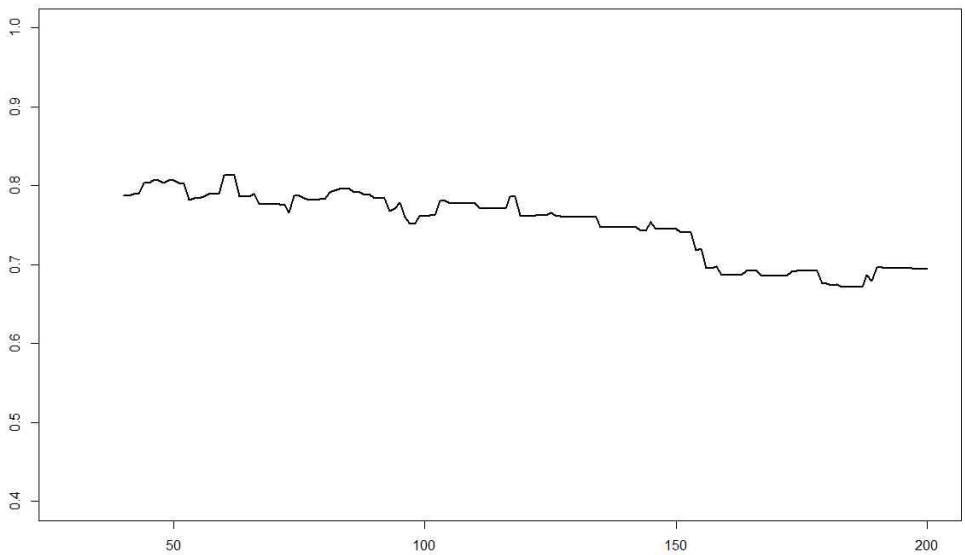
[주가 회복시기 찾는 모델 결과]

```
> confusionMatrix(pred.rf5, as.factor(data1_add$recovery_day3))
Confusion Matrix and Statistics

          Reference
Prediction 1-8 8-30 infinity
  1-8      338  244      600
  8-30     228  257      546
  infinity 219  316      834

Overall Statistics

               Accuracy : 0.3989
                 95% CI : (0.3829, 0.4152)
    No Information Rate : 0.5528
    P-Value [Acc > NIR] : 1
```

투자자들은 이 모델을 보고 가장 참고할만한 부분은 infinity를 피하는 것일 것이다. 만약에 주가가 30일이 경과했는데도 불구하고 회복하지 못한다면, 극심한 손해를 입을 것이기 때문이다. 이 모델에서 실제 infinity를 infinity라고 예측하는 양이 834개로 가장 많다. 그러므로 투자자들은 834개의 infinity를 피할 수 있을 것이다. 하지만 그럼에도 불구하고 투자자들이 가장 관심있어할 만한 8-30 class 예측값에 546개의 infinity가 있다는 것은 모델이 아직 참고할만큼 완벽하지 않다는 뜻이다. 그러므로 투자자들은 모델은 위험을 줄일 수 있는 하나의 참고사항으로 보고, 위기기사가 났을 때 그 기사의 토픽이나 여러 가지 상황을 고려하여 투자를 진행해야 할 것이다.

2

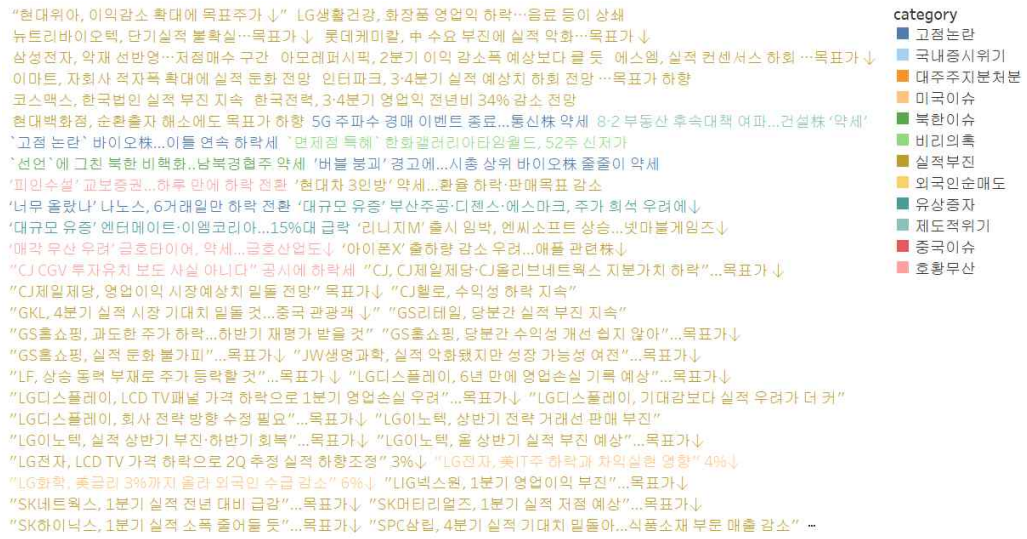
▶ 토픽모델 - 나이브 베이즈 분석 결과

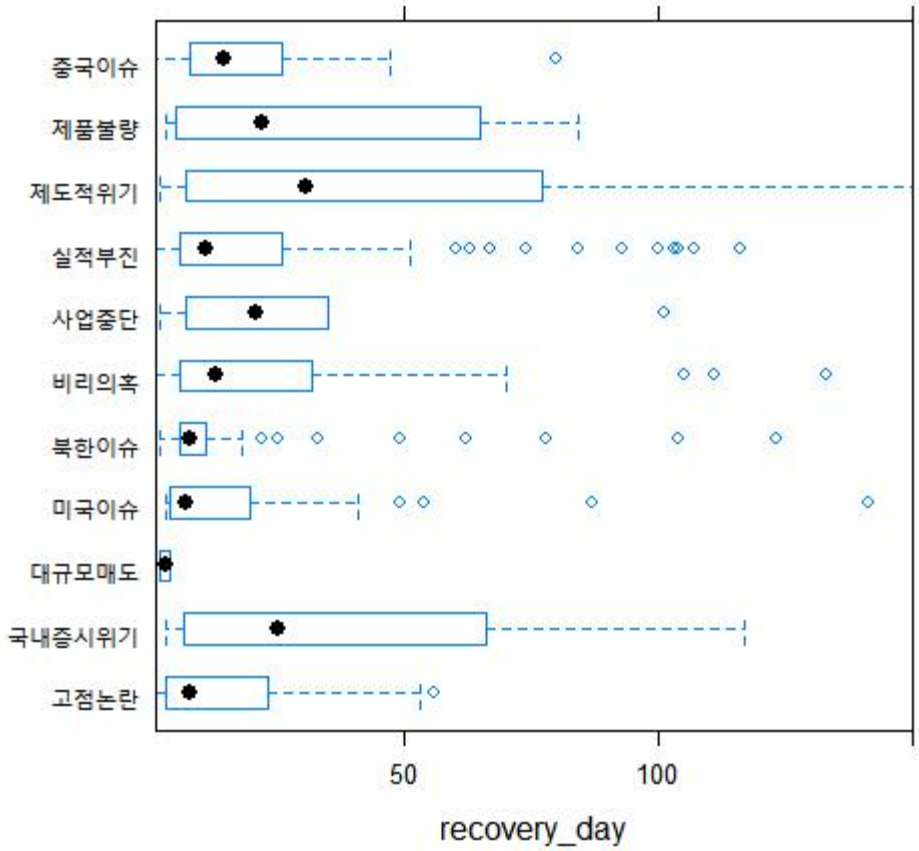


'실적부진','미국이슈','북한이슈','비리의혹' 카테고리만 뽑아서 나이브베이즈 한 결과이다. x축은 그 4개의 토픽 안에서 얼마만큼 단어를 고려하였는가를 본 것이고 y축은

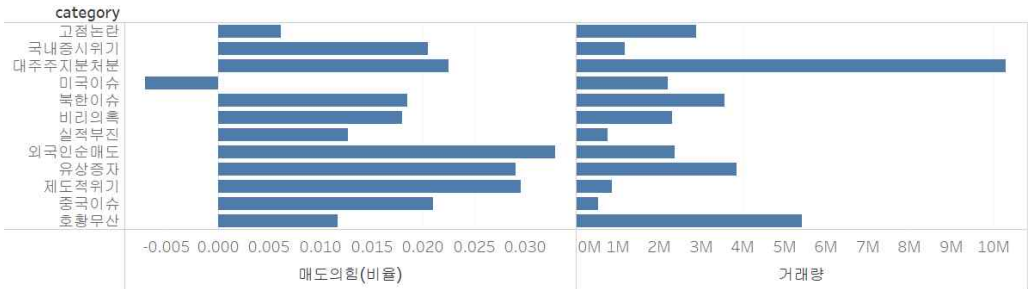
	나이브베이즈의 정확도다. 최고 정확도는 80%이었으며, 150개 단어를 넘어가는 순간 급격하게 정확도가 급감하는 것을 볼 수 있다. 4개의 클래스인 것을 감안했을 때 80%는 높은 정확도라고 볼 수 있지만, 실제 분류된 카테고리는 너무 많아 모든 카테고리를 넣으면 높은 정확도를 가지기 힘들다는 단점을 가지고 있다. 또한 처음 기사를 접했을 때는 4개의 카테고리 안에 이 기사가 포함되는 지 유무를 알 수 없어 활용도가 떨어지는 모델이라고 말할 수 있다.
3	▶ 토픽모델 - 룰베이스 분석 결과 `> table(real=pre_data_str$category,predict=pre_data_str$lable_pred)[c(4,5,6,8),]` predict real 미국이슈 북한이슈 비리의혹 실적부진 미국이슈 69 23 0 1 북한이슈 0 85 0 0 비리의혹 0 0 69 0 실적부진 5 0 0 295 나이브베이즈가 기사를 분류하는데에 있어 활용도가 떨어지는 부분이 있어, 기사의 제목 패턴을 파악해 룰을 만들어 기사 카테고리를 분류하였다. 위의 결과를 보면 매우 정확하게 기사를 분류할 수 있음을 볼 수 있었다. 기사의 패턴을 파악할 수 없거나, 모수의 부족 등의 이유로 4개의 카테고리만 분류하는 패턴만 파악하여 분류하는 룰을 만들었으나, 나이브베이즈같은 머신러닝 기법이 아닌, 단순한 룰베이스만으로도 기사의 카테고리를 정확하게 분류할 수 있음을 시사하는 결과였다.

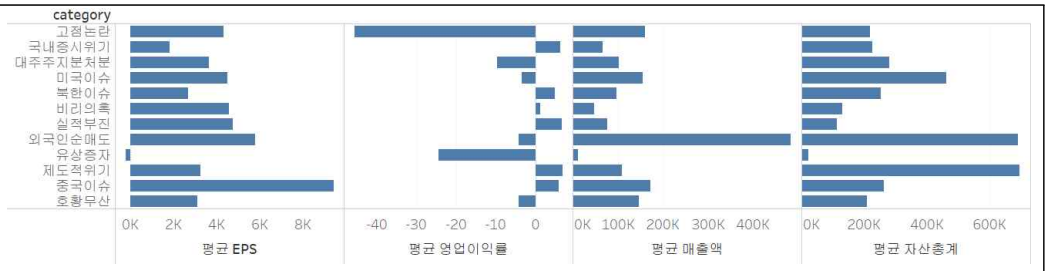
2) 분석결과 - 뉴스 카테고리 분류 시각화

1	▶ 뉴스 카테고리 분류 시각화  위의 그림은 40개 이상의 카테고리가 있는 기사의 제목을 나열한 것이다. 어떠한 기사의 종류가 많은지 한눈에 볼 수 있다. '실적부진'이 눈에 띄게 많은 것을 알 수 있다.



이 그림은 뉴스 카테고리 별로 회복점 시기를 boxplot으로 시각화 한 것이다. 연구 초반에 뉴스의 위기 카테고리 형식별로 어떠한 패턴이 있을 거라고 생각했고, 그 뉴스의 기업의 재무제표와 상호작용하여, 주식에 많은 영향을 미칠 것이라고 판단하였으나, 모델상에서는 눈에 띌만한 패턴과 영향을 발견하지 못하였다. 아마도 뉴스라는 정보는 정량화되지 않는 정성적인 부분이 많이 반영이 되어 있고, 또한 가지고 있는 데이터에 라벨링 할 수 있는 시간과 비용이 한정적이어서, 카테고리별 데이터 확보가 미진했던 것이 그 이유라고 볼 수 있을 것이다.





하지만 위기기사의 분류체계를 만든다는 것이 의미가 없는 것은 아니다. 위의 그래프를 살펴보면 '대주주지분처분', '외국인 순매도', '유상증자'의 이슈가 매도의 힘과 거래량이 높은 것을 발견할 수 있으며, '외국인 순매도'는 매출액과 자산총계가 높고, '유상증자'이슈의 경우 EPS와 매출액 자산총계가 모두 낮은 것을 볼 수 있다. 이것으로 비추어봤을 때 재무정보와 위기뉴스 카테고리는 서로상호작용을 하고 있다고 볼 수 있으며, 모델을 더욱 정교화 시킨다면, 카테고리와 재무정보, 그리고 다른 정보들과의 상호작용으로 주식 수익을 극대화시키는 모델을 더욱 정교화시킬 수 있을 것이다.

2) 제언 및 한계점

지금까지 뉴스시황을 살펴보고 주식매매를 시도하는 것은 줄곧 호황기사를 보고 이루어졌다. 하지만 진정한 기회는 위기기사에 있다. 주식은 사람들의 심리에 의해 움직이기에, 위기기사에서 해당하는 기업은 그 때 일어난 위기 사건에 비해, 그리고 기업 가치에 비해 주식이 더 가파르게 하락할 수 있기 때문이다. 그러므로 호황기사 뿐만 아니라 위기기사, 그리고 더 나아가 위기기사의 카테고리를 분류하여 다른 정보와의 상호작용을 파악한다면, 더욱 효과적으로 주식 수익률을 극대화 시킬 수 있을 것이다

이 연구의 한계점으로는 첫 번째, 이 주식 데이터에 위기를 극복하지 못하고 망해버린 기업의 데이터가 없어 데이터에 왜곡이 있다는 점, 두 번째, 동적인 뉴스기사 정보에 비해 정적으로 주어지는 재무정보만을 변수로 이용하였다는 점, 세 번째는 위기기사와 그 위기기사의 카테고리를 완벽하게 분리하는 모델을 만들지 못하였다는 점, 그리고 마지막으로 모델이 위기기사 이후 최고점과 최저점의 infinity지점을 불완전하게 예측한다는 점을 들 수 있다.

데이터 출처

- 네이버 금융 -
- 유가증권 표준코드 -
- http://media.kisline.com/ -