뉴스 기사 제목을 활용한 주가 변동여부 예측

이재용(응용통계학과), 이건이(응용통계학과), 서준영(AI학과), 김윤진(소프트웨어학부)

2021 CUAI 중앙대학교 인공지능 학회 하계 컨퍼런스
Proceeding of 2021 Chung-Ang University Artificial Intelligence's Summer Conference

요약

기사 제목이라는 텍스트(Text) 데이터를 통해서 주가를 예측하는 것이 본 연구의 목적이다. 기사 제목에는 기자들이 드러내고자 하는 핵심 내용이 들어갈 가능성이 크기 때문에 제목만으로도 해당 주식에 관한 정보를 얻을 수 있다. 특정 주식에 관한 기사 제목을 전처리 및 수치화 한 후 여러 텍스트 분류 모형으로 기계 학습하여 주가 변동여부를 예측하고 모형별 성능을 비교하였다.

연구방법 1: 연구대상 및 데이터 수집

1. 수집 기간: 2021년 1월 1일 ~ 2021년 7월 31일
2. 기사 수집 대상 기업
   한전기술, 셀트리온, 카카오게임즈, iMBC, 하이브, 삼성전자, 현대자동차, HMM, 대한항공, NAVER, 두산중공업, SK하이닉스
3. 주식별 기사 제목의 수
   한전기술: 84개, 셀트리온: 1,256개, 카카오게임즈: 232개, iMBC: 2개, 하이브: 316개, 삼성전자: 2,796개, 현대자동차: 2,692개, HMM: 999개, 대한항공: 1,610개, NAVER: 2,723개, 두산중공업: 325개, SK하이닉스: 1,272개
4. 주가 변동여부
   주가 변동여부를 예측하기 위해 이진형 변수를 생성
   기사가 쓰인 다음 날의 종가 > 그 전 장의 종가 == 1
   기사가 쓰인 다음 날의 종가 ≤ 그 전 장의 종가 == 0
5. 주가 변동여부의 비율
   0: 8,053, 1: 6,254

연구방법 2: 데이터 전처리

KoNLPy의 형태소 분석을 위한 여러 Class 중 Komoran Class 사용
-기사 제목을 형태소 분석 후 나온 토큰 중 일반명사, 고유명사, 동사, 형용사, 외국어, 관형사, 수사를 추출
-기사 제목에서 해당 주식의 이름은 미포함

	주식	기사 제목	언론사	주가변동여부
0	한전기술	우리기술 한전 전력연구원 과 풍력발전 제어시스템 개발 MOU	한국경제	0
1	한전기술	우리기술 한전 전력연구원과 풍력발전 제어시스템 개발 양해각서 체결	파이낸셜뉴스	0
2	한전기술	우리기술 한전 전력연구원과 풍력발전 제어시스템 개발 MOU	아시아경제	0
3	한전기술	코스피 종목 가 파란불 한전기술 신풍제약은 강세	이데일리	0
4	한전기술	한전 세계 최초로 해상풍력 일괄설치기술 개발 해상풍력 보급에 속도 낼 것	중앙일보	1
...	...	...	...	...
14302	SK하이닉스	SK하이닉스 만원 돌파 신고가 또 경신	연합뉴스	1
14303	SK하이닉스	삼성 SK하이닉스의 힘 정부 올 반도체 수출 억불↑	뉴스1	1
14304	SK하이닉스	삼성전자 만원 SK하이닉스 만원 반도체 투톱 목표가 줄상향	이데일리	1
14305	SK하이닉스	단독 SK하이닉스 기본급 보너스 연봉 잭팟도 터트릴까	디지털타임스	1
14306	SK하이닉스	신년사 박정호 SK하이닉스 부회장 경쟁자와도 협업해야	데일리안	1

본 연구에서 사용한 데이터

연구방법 3: 데이터 수치화 및 기계학습

1. 토큰화 한 기사 제목을 TF-IDF(Term frequency-inverse document frequency) 기반의 피처 벡터화
2. 로지스틱 회귀, 서포트 벡터 머신, Multinomial 나이브 베이즈를 이용한 기계학습

연구결과

모형별 분류성능 평가지표
-C와 a는 본 연구에서 쓰인 하이퍼 파라미터(Hyper Parameter)
-LR: Logistic Regression, SVM: Support vector machine, MNB: Multinomial Naive Bayes

모형		LR (C=1)	SVM (C=1)	MNB (a=0)
정확도		0.72	0.75	0.73
정밀도	상승: 1	0.71	0.75	0.70
	하락: 0	0.72	0.75	0.74
재현율	상승: 1	0.59	0.64	0.65
	하락: 0	0.81	0.84	0.79
F1-score	상승: 1	0.65	0.69	0.67
	하락: 0	0.76	0.79	0.76

혼동행렬 (Confusion Matrix)

모형	Actual Labels	Predicted Labels	
		상승: 1	하락: 0
LR	상승: 1	1,968	448
	하락: 0	768	1,109
SVM	상승: 1	2,020	396
	하락: 0	681	1,196
MNB	상승: 1	1,901	515
	하락: 0	663	1,214

결론

1. 정확도
   - 서포트 벡터 머신 75%, 나이브 베이즈 73%, 로지스틱 회귀 72%
   - 주가 변동여부 예측과 관련하여 뉴스 기사 제목이 데이터로 충분히 활용될 수 있음
2. 상승 기준 재현율
   - 세 모형 모두 주가 변동여부가 상승일 시 재현율은 70% 이하 실제로 주가가 상승하였던 것 중 70% 이하만 예측
   - 주가가 상승한 것을 많이 예측할수록 수익이 최대화가 되므로 재현율을 기준으로 분류의 성능을 더 높일 필요가 있음
3. 하락 기준 재현율
   - 주가 변동여부가 하락일 시 재현율이 모두 80% 근처, 실제로 주가가 상승하지 않았던 것 중 80%를 분류
   - 주가 변동여부를 하락을 기준으로 두면 본 연구의 모형이 투자의 손실을 최소화할 때 쓰일 수 있음
4. F1-score:
   - 세 모형 모두 70% 이하
   - 분류의 성능을 더 높여야 함