뉴스 기사 제목을 활용한 주가 변동여부 예측

CUAI 4기 금융 A팀
이재용(응용통계), 이건이(응용통계), 서준영(AI), 김윤진(소프트웨어)

[요약] 본 연구의 목적은 뉴스 기사 제목을 통해 주가변동여부를 예측하는 데 있다. 본 연구에서는 총 12개의 주식에 관한 2021년 1월부터 2021년 7월까지의 기사 제목을 활용한다. 주가 변동여부는 기사가 쓰인 다음 날의 종가와 이전 장의 종가를 비교해서 구한다. KoNLPy로 전처리 후 TF-IDF 기반으로 수치화된 기사 제목으로 주가 변동여부를 분류한 결과 서포트 벡터 머신이 75%의 정확도를 보였다.

1. 서 론

주식 시장은 전반적인 시황과 거래자들의 기업에 대한 기대에 따라 움직인다. 이러한 기대감은 다양한 원천의 데이터로부터 추정해볼 수 있으며 이를 통해 주가의 움직임도 예측해볼 수 있다. 본 연구에서는 기사 제목이라는 텍스트(Text) 데이터를 통해서 주가를 예측하고자 한다.

텍스트라는 비정형 데이터로부터 통계적인 특성과 패턴이나 추세 등의 정보를 끌어내는 과정을 텍스트 마이닝(text mining)이라 하며, 이에 관한 연구가 최근에 빠르게 진행되고 있다[1]. 텍스트 마이닝은 여러 분야에서 연구되고 있는데, 그중 주제에 따라 문서를 나누는 텍스트 분류(text classification), 문서 작성자의 심리, 감정 상태 등을 파악하는 감성분석(sentiment analysis) 등에서 집중적으로 활용되고 있다(백두현 외, 2020).

최근 주가 변동을 예측하는 여러 연구 중, 기사의 특정 단어를 활용하여 주가의 상승, 하락을 분류하는 연구가 진행되어 최대 78%의 예측률을 보이기도 하였다(이민식, 이홍주, 2017). 기사 제목에는 기자들이 드러내고자 하는 핵심 내용이 들어갈 가능성이 크기 때문에 제목만으로도 해당 주식에 관한 정보를 얻을 수 있다. 이에 본 연구에서는 특정 주식에 관한 기사 제목을 전처리 및 수치화한 후 여러 텍스트 분류 모형으로 기계학습하여 주가 변동여부를 예측하고 모형별 성능을 비교하였다.

2. 본 론

1) 연구대상 및 데이터 수집

본 연구의 연구 대상은 2021년 1월 1일부터 2021년 7월 31일까지의 한전기술, 셀트리온, 카카오게임즈, iMBC, 하이브, 삼성전자, 현대자동차, HMM, 대한항공, NAVER, 두산중공업, SK하이닉스에 관한 기사 제목이다. 산업 특성상 해당 기간 중 주가가 잘 오르는 산업이 있고 그렇지 않은 산업도 있기에 최대한 다양한 산업의 기업들을 연구 대상으로 삼고자 하였다. 본 연구에서는 웹스크레이핑(Web Scraping)으로 네이버 금융 페이지의 뉴스 검색기를 통해 총 14,307개의 기사 제목을 추출하였다[2]. 주식별 기사 제목의 수는 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: 1,272개이다.

본 연구는 해당 주식에 관한 기사가 쓰인 다음 날의 주가가 올랐는지 예측하기 위해 주가 변동여부라는 데이터를 수집하였다. 네이버 금융에서 제공하는 해당 주식의 일별 시세 페이지를 통해 종가를 추출하였고, 기사가 쓰인 다음 날의 종가가 그 전 장의 종가보다 높으면 1, 같거나 낮으면 0인 이진형 변수를 생성하였다. 이때, 위에서 구한 기사 제목은 기사가 쓰인 다음 날이 휴장인 날이 없도록 구한 데이터이다. 주가 변동여부의 비율은 0: 8,053, 1: 6,254로 다소 불균형적인 데이터이다.

2) 데이터 전처리

텍스트 데이터는 비정형 데이터로 반드시 전처리 해주어야 한다. 이때, 문장으로 이루어진 텍스트 데이터는 토큰화를 통해서 토큰이란 단위로 나누어줄 수 있다. 이때, 한국어는 영어와 달리 띄어쓰기가 아닌 형태소 단위로 토큰화를 해주어야 한다. KoNLPy는 한글 자연어 처리를 위해 쓰이는 패키지이다(Park and Cho, 2014). KoNLPy의 형태소 분

[1] 네이버 지식백과, 국립중앙과학관-빅데이터

[2] https://finance.naver.com/news/ 에서 발췌하였다.

석을 위한 여러 Class 중 본 연구에선 Komoran Class를 사용하였다. 기사 제목을 형태소 분석 후 나온 토큰 중 일반명사, 고유명사, 동사, 형용사, 외국어, 관형사, 수사를 추출하였다. 이는 기사 제목 특성상 문장의 길이가 짧고 단어가 많지 않기 때문에 최대한 많은 정보를 추출하기 위해 관형사와 수사 같은 품사도 포함하였다. 또한 기사 제목에서 해당 주식의 이름은 포함하지 않았는데, 이는 주가가 상승 중인 주식이 있고 하락 중인 주식도 있으므로 이러한 추세를 분석에 반영하지 않기 위함이다.

3) 데이터 분리

주식의 기사 제목과 주가 변동여부를 70:30으로 분리하여 훈련용(n=10,014)과 검증용(n=4,293) 데이터 세트(Data set)를 만들어 분석하였다. 분리된 데이터 세트에서도 주가 변동여부의 비율을 유지해주는 표본추출(Sampling) 방법으로 층화추출법(Stratified sampling)을 사용하였다.

4) 데이터 수치화

토큰화된 기사 제목을 TF-IDF(Term frequency-inverse document frequency) 기반의 피처 벡터화(Feature vectorization)를 하였다[3]. TF-IDF는 비슷한 기법인 Count Vectorizer의 단점인 언어 특성상 자주 나타나는 단어의 중요도를 높이는 문제를 보완해준다(권철민, 2020). 또한 데이터가 금융과 관련된 기사 제목들이다 보니 주제 특성상 자주 쓰이지만 별로 의미가 없는 단어들이 있을 수 있으므로 TF-IDF가 적합하다고 판단되었다. 본 연구에선 Scikit-Learn에서 제공하는 TfidfVectorizer 함수를 사용하였다.

5) 기계학습

본 연구에서 사용된 기계학습(Machine learning) 모형은 텍스트 분류에서 자주 사용되는 로지스틱 회귀(Logistic Regression, LR), 서포트 벡터 머신(Support Vector Machine, SVM), Multinomial 나이브 베이즈(Multinomial Naive Bayes, MNB)로 총 세 가지이다. 이 모형들의 특징은 다음과 같다.

로지스틱 회귀(Logistic Regression)는 선형회귀와 비슷하나, 종속변수의 값이 [0, 1] 범위 내에서만 존재한다. 종속변수의 값이 특정 임계치 이상이면 해당 범주에 속하는 것으로, 미만이면 해당 범주에 속하지 않는 것으로 판단하기 때문에 분류 문제에서 자주 사용하는 모형이다. 결정경계로부터 멀리 떨어진 표본이 모수에 과도하게 영향을 주는 것을 방지한다는 장점이 있다(김수현 외, 2020).

서포트 벡터 머신(Support Vector Machine)은 두 범주 사이의 거리, 마진(margin)을 최대로 하는 선형경계를 찾아 새로운 데이터가 주어졌을 때, 경계선을 기준으로 특정 범주로 분류하는 비확률적 분류 모형이다(황화목, 2017). 노이즈(Noise) 데이터에 영향을 크게 받지 않고 과적합이 잘 일어나지 않는다는 장점이 있다(황화목, 2017).

나이브 베이즈(Naive Bayes)는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 조건부 확률 모형이며 조건부 사후 확률을 극대화하는 값으로 범주를 추정한다. 나이브 베이즈 가정하에 고차원의 확률추정 문제를 반복적인 일차원의 확률추정 문제로 단순화 시킬 수 있다는 장점이 있으며, 스팸메일 분류 등에서 자주 사용되는 모형이다(강민선, 2017).

6) 기계학습 성능평가

각 기계학습 모형의 분류 성능평가 시 사용되는 지표는 총 네 가지로 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score이다.

<표 1> 혼동행렬(Confusion Matrix)

		Predicted Condition	
		Positive	Negative
True Condition	Positive	TP(True Positive)	FN(False Negative)
	Negative	FP(False Positive)	TN(True Negative)

표 1을 보았을 때, TP는 해당 범주에 알맞게 분류된 경우, FN은 해당 범주에 속하는 것이 다른 범주로 잘못 분류된 경우, FP는 다른 범주의 것이 해당 범주로 잘못 분류된 경우, TN은 다른 범주의 것이 다른 범주로 알맞게 분류된 경우를 말한다.

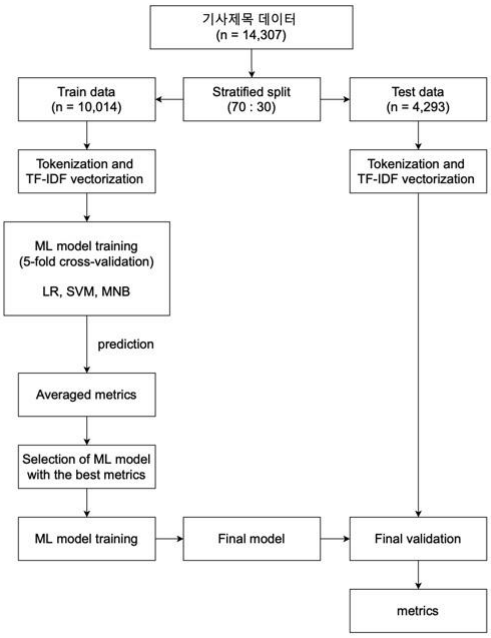
이를 토대로 각 지표는 다음과 같이 계산한다.

<표 2> 분류성능평가지표 계산방식 및 비고

지표	계산방식	비고
정확도	(TP+ TN)/ (TP+ FN+ FP+ TN)	정확한 분류/ 전체 자료
정밀도	TP/(TP+ FP)	정확한 양성분류가 중요할 시 유용
재현율	TP/(TP+ FN)	
F1-score	2*(정밀도*재현율)/ (정밀도+ 재현율)	데이터 레이블이 불균형 구조일 시 유용

[3] 피처 벡터화란 텍스트를 수치화하는 작업을 뜻한다.

7) 결과



<그림 1> 기계학습 기반 기사 제목을 통한 주가 변동여부 예측의 Flow Diagram. ML: Machine learning, LR: Logistic regression, SVM: Support vector machine, MNB: Multinomial naive bayes, TF-IDF: Term frequency-inverse document frequency

전처리를 거친 데이터 세트를 세 모형(로지스틱 회귀, 서포트 벡터 머신, 나이브 베이즈)으로 기계학습시켰다. 모형마다 훈련용 데이터에서 5-겹 교차검증(5-fold Cross validation)을 통해 최적의 하이퍼 파라미터를 구한 후 검증용 데이터로 최종적으로 성능을 평가하였다. 이의 결과는 다음의 표와 같다.

<표 3> 모형 별 분류성능평가지표; C와 a는 본 연구에서 쓰인 하이퍼 파라미터(Hyper Parameter)이다.

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

LR: Logistic Regression, SVM: Support Vector Machine, MNB: Multinomial Naive Bayes

<표 4> 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

LR: Logistic Regression, SVM: Support vector machine, MNB: Multinomial Naive Bayes

표 3을 보면 세 모형 모두 정확도가 70% 이상인 것을 볼 수 있다. 주가 변동여부가 상승일 때의 정밀도도 세 모형 모두 70% 이상이다. 하지만 주가 변동여부가 상승일 때의 재현율은 세 모형 모두 70% 이하이다. 반대로, 주가 변동여부가 하락인 경우의 재현율은 세 모형 모두 80% 근처로 다른 지표에 비해 상당히 높다. 세 모형 모두 F1-score는 70% 이하이다. 마지막으로, 서포트 벡터 머신이 다른 모형보다 전반적으로 성능이 더 좋았고 정확도가 75%로 가장 높았다.

3. 결 론

본 연구는 12개의 주식에 관한 기사 제목을 수집하고 분석하여 주가 등락을 예측하였다. KoNLPy의 Komoran Class를 활용하여 텍스트 전처리를 하였으며 토큰화된 기사 제목을 TF-IDF 기반의 피처 벡터화 후 로지스틱 회귀, 서포트 벡터 머신, 나이브 베이즈 모형으로 기계학습을 시켰다. 이들의 분류 결과는 다음과 같다.

대부분의 분류 성능 평가 지표에서 서포트 벡터 머신의 성능이 다른 모형들보다 높게 측정되었다. 특히, 정확도가 서포트 벡터 머신이 75%로 가장 높았고 나이브 베이즈는 73%, 로지스틱 회귀 모델이 72%로 이들 모두 70% 이상의 정확도를 보였다. 이는 주가 변동 예측과 관련해서 뉴스 기사 제목이 데이터로 충분히 활용될 수 있음을 시사한다. 세 모형의 정밀도도 모두 70% 이상으로, 기사 제목을 통해 주가가 상승하였다고 예측한 것 중 70% 이상이 실제로 상승하였다. 하지만 세 모형 모두 주가 변동여부가 상승일 시 재현율은 70% 아래로 실제로 주가가 상승하였던 것 중 70% 이하만 예측할 수 있었다. 실제로 주가가 상승한 것을 많이 예측할수록 수익이 최대화가 되므로 재현율을 기준으로 분류의 성능을 더 높일 필요가 있어 보인다. 반대로 주가 변동여부가 하락일 시 재현율이 모두 80% 근처로,

실제로 주가가 상승하지 않았던 것 중 80%를 분류할 수 있었다. 즉, 주가 변동여부를 하락을 기준으로 두면 본 연구의 모형이 투자의 손실을 최소화할 때 쓰일 수 있음을 시사하고 있다. 마지막으로, 불균형적인 데이터의 분류에 있어 더 정확한 성능 평가 지표인 F1-score가 세 모형 모두 70% 이하인 것을 고려하면, 이 또한 분류의 성능을 더 높여야 함을 시사하고 있다.

본 연구의 한계점은 다음과 같다. 본 연구에서 측정된 분류 성능 지표가 좋은 편은 아니다. 따라서, 더욱 정교한 전처리 및 하이퍼 파라미터 튜닝을 하거나 기계학습보다 성능이 더 뛰어나기로 알려진 딥러닝으로 새로 연구할 필요가 있다(Wibawa, 2018). 이지민 외(2020)에 따르면 뉴스 분석 시 한국어 뉴스 데이터를 영문으로 번역시켜서 전처리하였을 때 성능이 더 좋다고 한다. 특히 한글의 형태소 분석을 통해 얻은 토큰보다 영어 단어의 토큰이 수치화 때 희소행렬을 더 줄일 수 있을 것으로 판단된다. 따라서 추후 연구에선 이처럼 뉴스 기사 제목을 영문으로 번역해서 전처리한 후 분류하는 방법을 고려해볼 수 있다. 마지막으로, 본 연구에서 사용된 주가 변동여부 데이터는 단순히 전 장 기준 종가의 등락 여부로 이의 정도는 고려하지 않았다. 따라서 추후 연구에서는 적당한 주가의 상승 기준을 잡고 주가 변동여부 데이터를 구하는 것을 고려해보아야 한다.

참고 문헌

1) 텍스트 마이닝. Available: https://terms.naver.com/entry.naver?docId=3386330&cid=58370&categoryId=58370
2) 백두현, 황민규, 이민지, 우성일, 한상우, 이연정 and 황재욱, “텍스트 분류 기반 기계학습의 정신과 진단 예측 적용”, 생물정신의학, 27(1), 18-26, 2020.
3) 이민식, 이홍주, ”카테고리 중립 단어 활용을 통한 주가 예측 방안 텍스트 마이닝 활용”, 지능정보연구, 2017.6.
4) Park, Eunjeong L., and Sungzoon Cho. "KoNLPy: Korean natural language processing in Python." Annual Conference on Human and Language Technology. Human and Language Technology, 2014.
5) 권철민, 『파이썬 머신러닝 완벽가이드』, 위키북스, p.476-479. 2020.
6) 김수현, 이영준, 신진영, 박기영, “거시경제 분석을 위한 텍스트 마이닝”, 韓國經濟의 分析, Vol.26 No.1, 2020.
7) 황화목, “비정형 텍스트 자료 분석을 통한 다중 클래스 분류 기법의 성능 비교”, 인하대학교 대학원, 2017.2.
8) 강민선, “토픽모형을 이용한 텍스트 자료의 분류 성능비교”, 동국대학교, 2017.2.
9) M. S. Wibawa, "A Comparison Study Between Deep Learning and Conventional Machine Learning on White Blood Cells Classification," International Conference on Orange Technologies (ICOT), pp. 1-6, 2018.
10) 이지민, 정다운, 구영현, 유성준. "한국어 뉴스 분석 성능 향상을 위한 번역 전처리 기법." 한국방송미디어공학회 학술발표대회 논문집, 2020: 497-501, 2020.