The Korean Journal of Applied Statistics
2023, Vol. 36, No. 5, 489–499

DOI: https://doi.org/10.5351/KJAS.2023.36.5.489
Print ISSN1225-066X / Online ISSN 2383-5818

Verification on stock return predictability of text in analyst reports

Young-Sun Lee[a], Akihiko Yamada[b], Cheol-Won Yang[c], Hohsuk Noh[1,a,d]

[a]Department of Statistics, Sookmyung Women's Univesity;
[b]Bigdata Convergence and Open Sharing System, Seoul National Univesity;
[c]School of Business Administration, Dankook Univerisity;
[d]Research Institute of Natural Sciences, Sookmyung Women's Univesity

Abstract

As sharing of analyst reports became widely available, reports generated by analysts have become a useful tool to reduce difference in financial information between market participants. The quantitative information of analyst reports has been used in many ways to predict stock returns. However, there are relatively few domestic studies on the prediction power of text information in analyst reports to predict stock returns. We test stock return predictability of text in analyst reports by creating variables representing the TONE from the text. To overcome the limitation of the linear-model-assumption-based approach, we use the random-forest-based F-test.

Keywords: stock return predictability, natural language processing, analyst reports, random forest F-test

1. 서론

일반적으로 애널리스트(analyst)는 특정 산업이나 기업에 대해 조사 · 분석을 수행하며, 보고서의 형태로 재무분석결과를 지속적으로 발표한다. 애널리스트 보고서에 공시되는 투자 관련 추천의견 및 목표주가, 이익예측치 등의 정량적인 정보가 투자에 유용한 정보를 제공하는지에 대해 최근 Barber 등 (2010), Bradley 등 (2014)에 의해 연구되었다. 하지만 애널리스트 보고서에는 이러한 정량적 정보 외에도 각종 예측치의 근거, 회계정보를 설명하는 보고서 본문과 본문의 내용을 축약한 보고서 제목이 존재한다. 투자자들은 이러한 텍스트들도 기업의 수익률을 예측하는데 중요한 정보로 인지한다. 하지만 국내에서 발간되는 애널리스트 보고서와 관련된 기존 연구들은 정량적인 정보의 투자유용성 검증에 집중되어 온 경우가 많았다. 최근들어서는 애널리스트 보고서내의 정량적인 정보뿐만 아니라 텍스트 정보도 기업수익률 예측에 유용한 정보가 있을 수 있다는 인식하에 여러 연구가 이루어지고 있다. Jang 등 (2016)은 애널리스트 보고서의 제목에서 bi-gram 기법을 활용하여 투자의견을 추출하고 이익예측 유용성을 살펴보았고, Yang (2021)은 보고서의 제목에서 사전기반 방법을 사용하여 어조를 추출하여 그것이 기업수익률 예측에 대한 정보력이 있는지 여부를 살펴보았다. 하지만 이러한 연구들은 텍스트에서 의견을 추출하는 방식이 문장의 문맥을 고려하지 하는 bag-of-words 방식이라는 점과 추출된 텍스트기반 변수의 유용성을 선형모형이라는 강력한 모형 가정하에서 검증한다는 한계점을 가지고 있다. 본 연구에서는 이러한 한계점을 해결하기 위해 보고서의 어조를 추출할 때 문장의 문맥적인 정보를 고려하는 자연어 처리 방식인 BERT 모형을 사용하는 것을 고려하였고, Coleman 등 (2022)의 연구결과를 활용하여 추출된 텍스트 기반 변수의 예측유용성 검증에 있어서 선형모형 가정을 완화하는 것을 고려하였다.

[1]Corresponding author: Department of Statistics, Sookmyung Women's Univesity, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Korea. E-mail: hsnoh@sookmyung.ac.kr

Published 31 October 2023 / journal homepage: http://kjas.or.kr

Table 1: Average return with respect to OPN in each recommendation-opinion class

OPN	Buy	Outperform	Hold	Underperform	Sell
5	-	1.69%	2.74%	2.58%	-
4	2.42%	1.21%	0.69%	−7.56%	-
3	2.67%	1.01%	−0.28%	5.11%	-
2	−0.94%	0.17%	−1.57%	−2.39%	-
1	−0.80%	−1.11%	−2.84%	−1.14%	−15.70%

본 논문의 구성은 다음과 같다. 2절에서는 애널리스트 보고서에 나타난 정보들의 특징, 애널리스트 보고서의 텍스트 분석에 관한 연구, 애널리스트 보고서 텍스트의 정보력 검증과 관련된 이슈에 대해 논의한다. 3절에서는 분석 자료에 대해 소개하고, 4절은 기존 정보력 검증 연구의 한계점을 보완할 수 있는 통계 방법론을 소개한다. 이를 토대로 애널리스트 보고서 텍스트의 정보력 검증 결과를 5절에서 보고한다. 6절에서는 연구 결과를 요약하였다.

2. 기존연구

2.1. 애널리스트 보고서에서 제공하는 정보들의 특징

애널리스트들의 정보 생성 활동 과정에서 관찰되는 대표적인 특징은 허딩(herding) 현상과 이해상충의 문제이다. 애널리스트와 관련된 허딩 현상은 애널리스트들이 자신의 보고서를 공시하기에 앞서 다른 애널리스트들의 이익예측치, 종목 매수/매도 추천 등 시장 전반의 컨센서스(consensus) 정보를 참고하여 보고하는 것을 의미한다. 정보를 생산하는 애널리스트와 평가 대상 기업 간의 이해관계가 대립하는 이해상충의 문제도 존재한다. 애널리스트는 평가 대상 기업과의 우호적인 사업 관계를 지속하기 위해 낙관적인 예측 결과를 발표하기 쉽다. 특히 이익예측치보다 투자 관련 추천의견이 더욱 낙관적으로 편향된다고 알려져 있다.

본 연구에서 사용될 한국경제 신문사의 애널리스트 보고서 자료에서도 애널리스트들의 추천의견이 낙관적인 쪽으로 치우쳐져 있음을 확인할 수 있었다. 추출한 총 9,691개의 보고서에서 추천의견이 존재하는 관측치는 9,425개인데 이 중에 종목 매도 또는 비중축소의 부정적 의견을 제시한 것은 0.14%(13건)정도 밖에 되지 않는다. 반면, 긍정적 의견인 종목 매수는 전체 추천의견의 93.05%(8,770건)를 차지하였다. Table 1은 각각의 추천의견 내에서 애널리스트 보고서 본문에 대한 어조 (OPN = 긍정문장의 비율-부정 문장의 비율)의 값에 따라 어떻게 수익률의 평균값이 달라질 수 있는지를 보여주고 있다. Table 1에서 OPN의 5에서 1까지의 값은 각각 OPN값이 상위 80%-상위 100%, 상위 60%-상위 80%, 상위 40%-상위 60%, 상위 20%-상위 40%, 상위 0%-상위 20%임을 나타낸다. Table 1에서 매수를 추천하는 종목임에도 불구하고 보고서 본문의 어조가 긍정적일수록 해당 종목의 수익률은 증가하는 경향을, 반대의 경우 수익률이 감소하는 추세를 보인다. 이는 추천의견이 낙관적인 쪽으로 편향되어, 상대적으로 매도 의견이 드물게 발견되는 정보의 희박성 문제를 보고서 텍스트의 어조를 분석함으로써 어느 정도 해결할 수 있는 가능성이 존재함을 의미한다.

2.2. 애널리스트 보고서의 텍스트 분석에 관한 연구

애널리스트 보고서의 텍스트에는 보고서 제목, 요약, 소주제별로 구분된 본문이 존재한다. 이와 같은 애널리스트 보고서 상의 텍스트 구성요소 중 어떤 요소를 선택하여 수익률 예측에 활용해야 하는가의 문제가 존재한다. 재무 분야에서 애널리스트 보고서의 텍스트를 분석하여 활용한 연구들은 크게 2가지 방식으로 구분된다.

첫 번째, 보고서 본문의 어조를 활용하는 방식이다. Huang 등 (2014)은 나이브 베이즈(Naive Bayes) 방식을 통해 애널리스트 보고서 본문의 문장들을 긍정과 부정으로 분류했다. Liang 등 (2022)는 나이브 베이즈

(Naive Bayes)대신 LSTM (long short-term memory)을 사용하여 보고서 본문의 문장들을 긍정, 부정으로 분류하였다. Huang 등 (2014), Liang 등 (2022)은 긍정/부정으로 분류된 보고서 본문의 문장들로부터 긍정문장과 부정문장의 비율을 계산한 후, 긍정문장의 비율과 부정문장의 비율의 차이를 보고서의 어조 변수로 사용하였다.

두 번째 방식은 애널리스트 보고서의 제목을 활용하는 방식이다. Jang 등 (2016)의 연구에서는 애널리스트 보고서에 공시된 투자의견은 관계사, 고객사, 주간사 등의 외적 영향을 많이 받는다는 점을 고려하여 외적 영향을 상대적으로 적게 받는 애널리스트 보고서의 제목에서 투자의견을 추출했다. Yang (2021)은 보고서의 제목에 보고서 전체적인 맥락이 반영되어 있다는 측면을 고려하여 제목에서 어조를 추출하였다.

2.3. 애널리스트 보고서 텍스트의 정보력 검증과 관련된 이슈

본 절에서는 애널리스트 보고서를 기업수익률 예측에 활용할 때 고려해야 하는 이슈들에 대해 살펴보고자 한다. 본 절부터 다양한 회귀모형이 소개되게 되는데 모두 오차항에 대한 가정은 평균이 0이고 분산이 유한이라는 가정($E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) < \infty$)으로 동일하다.

2.3.1. 보고서 어조의 수익률 예측 정보력을 선형모형 가정 하에 F-test로 검증하는 것이 타당한가?

Huang 등 (2014), Liang 등 (2022), Yang (2021)은 애널리스트 보고서 어조의 수익률 예측 정보력을 선형모형 가정하에서 검증하였다. Huang 등 (2014)과 Liang 등 (2022)의 연구는

$$\mathrm{CAR} = \alpha_0 + \beta \mathrm{OPN} + \gamma_1 \mathrm{REC\_REV} + \gamma_2 \mathrm{EF\_REV} + \gamma_3 \mathrm{TP\_REV} + \sum_j \delta_j \,\mathrm{controls}_j + \epsilon \tag{2.1}$$

로 선형모형을 가정하였다. 식 (2.1)에서 OPN은 애널리스트 보고서 본문의 전체 문장 중에서 긍정문장이 차지하는 비율에서 부정문장의 비율을 차감한 값으로 애널리스트 보고서 본문의 어조를 추출한 변수이다. REC_REV은 추천의견값이며, REC_REV가 1이면 매도, 2는 비중축소, 3은 중립, 4는 매수, 5는 적극매수인 경우를 나타낸다. EF_REV와 TP_REV는 각각 이익 예측치와 목표주가를 의미한다. 또한, 통제 변수 controls로 보고서 공시일 이전 열흘 간 누적된 기업수익률, 기업 규모, 주가순자산비율을 고려하였다. 반응 변수 CAR는 애널리스트 보고서 발표일 기준으로 당일과 전후 이틀, 총 5일의 누적 기업수익률이다. Huang 등 (2014)과 Liang 등 (2022)는 선형모형 (2.1)에서 $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$를 검정하여 보고서 본문에서 추출된 어조 변수 OPN이 기업수익률 예측 정보력를 가지고 있는지를 검정하였다.

Yang (2021)의 연구는

$$\begin{aligned}\mathrm{CAR} = {} & \alpha_0 + \alpha_1 \mathrm{POS\_TONE} + \alpha_2 \mathrm{NEG\_TONE} + \alpha_3 \mathrm{UP\_GR} + \alpha_4 \mathrm{DOWN\_GR} \\ & + \alpha_5\, \Delta\mathrm{TPRC} + \alpha_6\, \mathrm{Nanalyst} + \alpha_7\, \mathrm{size} + \alpha_8 (B/M) + \epsilon\end{aligned} \tag{2.2}$$

의 선형모형을 가정하였다. 식 (2.2)에서 POS_TONE과 NEG_TONE은 TONE의 값에 의해 정의되며, 애널리스트 보고서의 어조가 긍정적인지 부정적인지를 나타내는 지시 변수이다. TONE은 애널리스트 보고서 제목에 나타난 긍정어의 수에서 부정어의 수를 차감한 값인데 TONE이 0보다 크다면 POS_TONE이 1의 값, 반대로 TONE이 0보다 작다면 NEG_TONE이 1의 값을 가지도록 정의하였다. UP_GR와 DOWN_GR은 추천의견이 이전에 비해 상향되었는지, 하향되었는지를 나타내는 지시 변수로, 추천의견이 상향되면 UP_GR이 1, 하향되었으면 DOWN_GR이 1이다. ΔTPRC는 목표주가의 변화율, Nanalyst는 특정 연도에 해당 종목에 대한 보고서를 발표한 애널리스트 수, size는 기업 규모, B/M는 주가순자산비율을 나타낸다. Yang (2021)은 선형모형 (2.2)에서 $H_0 : \alpha_1 = \alpha_2 = 0$ vs. $H_1 : \alpha_1 \neq 0$ or $\alpha_2 \neq 0$를 검정하여 보고서 본문에서 추출된 어조 변수 POS_TONE와 NEG_TONE이 기업수익률 예측에 정보력이 있는지 여부를 판단했다.

McAlexander와 Mentch (2020)은 정량적 사회과학 연구에서 많은 경우에 반응 변수에 대한 공변량의 예측력 검증시 반응 변수와 공변량들 간의 관계가 선형 함수로 주어진다는 가정하에서 이루어짐을 지적하고 공변량의 예측력에 대해 편향된 결론을 피하기 위해 선형모형 가정을 완화해야 함을 주장하였다. Coleman 등 (2022)의 연구에서는 반응 변수와 공변량 사이의 관계가 복잡한 함수로 주어질 수 있음을 고려하여 랜덤 포레스트(random forest)를 통해 공변량의 marginal effect에 대해 검증하는 방법론을 제시하였다. 본 연구에서는 Coleman 등 (2022)의 연구 결과를 활용하여 선형성 가정을 완화한 가운데 보고서 어조의 수익률 예측 정보력을 검증해보고자 한다.

2.3.2. 보고서의 어조를 본문에서 추출할 것인가, 제목에서 추출할 것인가?

Huang 등 (2014), Liang 등 (2022)과 Yang (2021)은 기업의 수익률 예측을 위해 애널리스트 보고서에서 추출한 텍스트의 종류가 다르다. Huang 등 (2014)과 Liang 등 (2022)은 애널리스트 보고서의 본문에서, Yang (2021)은 제목에서 어조 변수를 추출하였다. 하지만 본문과 제목 중 어디에서 애널리스트 보고서의 어조를 추출하는 것이 기업수익률 예측에 좋은지 비교하는 연구는 이루어지지 않았다. 따라서 본 연구는 애널리스트 보고서의 텍스트 추출 위치에 따라 구분되는 2가지의 어조 변수에 대한 기업수익률에 대한 예측력을 비교해보려고 한다.

2.3.3. 보고서 텍스트에서 어조를 어떤 방식을 사용하여 추출할 것인가?

애널리스트 보고서 텍스트의 어조를 추출하는 것은 텍스트의 감성을 분석하는 것이며, 이를 감성분석이라 한다. 전통적인 텍스트 감성분석은 어휘 사전에 기반하여 긍정어와 부정어의 빈도수를 비교하여 텍스트의 감성을 얻어내는 것이다. Yang (2021)의 연구에서는 Kim과 Lee (2013)이 제안한 긍정/부정 카테고리로 구분된 어휘 사전을 이용했다. 사전에 기반한 감성분석은 단어의 사전적 의미만을 기초로 하기 때문에 단어 간 순서를 고려하지 않으며, 단어가 사용되는 상황 및 문맥에 대한 이해 없이 어휘의 통상적인 의미에만 의존한다. 하지만 텍스트의 감성을 분석할 때 단어의 사전적 의미뿐만 아니라 문맥을 파악하는 것도 매우 중요하기 때문에 사전기반 방법으로 감성분석을 하는 것에는 한계가 있다.

단어의 순서를 고려하여 문맥을 반영한 대표적인 자연어 처리 방법으로 BERT (bidirectional encoder representations from transformers) 모델이 있다 (Devlin 등, 2019). BERT 모델은 학습 과정에서 문장 내 어휘, 통사, 의미 정보를 포착하여 단어의 임베딩(embedding; 사람이 사용하는 자연어를 기계가 이해할 수 있는 형태의 벡터로 전환한 결과 혹은 그 과정)을 얻을 수 있다고 알려져 있다. KR-FinBert-SC (Kim과 Shin, 2022) 모형은 금융분야 코퍼스(Corpus; 말뭉치)에서 BERT 모델을 훈련시키고, 그것을 기반으로 금융 관련 텍스트의 감성을 분류하는 한국어 기반 자연어 처리 모형이다. KR-FinBert-SC를 이용해서 애널리스트 보고서 제목이나 본문 내 각 문장의 어조가 긍정/부정/중립일 확률을 예측할 수 있다. 본 연구는 KR-FinBert-SC 모형을 이용하여 제목 또는 본문 내 문장의 감성을 분류하고자 한다.

애널리스트 보고서 제목의 예시를 통해 사전기반 감성분류와 KR-FinBert-SC 기반 감성분류를 비교하면, 두 방식의 차이점을 뚜렷하게 확인할 수 있다. 예를 들어, '4월 손해율 하락으로 견조한 이익흐름 지속'이라는 애널리스트 보고서 제목의 단어 중 '손해', '하락'은 어휘사전의 부정적 단어 카테고리에 포함되기 때문에 부정어로 본다. '이익'은 긍정적 단어 카테고리에 포함되므로 긍정어로 판단한다. 부정적 단어의 개수는 2개, 긍정적 단어의 개수는 1개이기 때문에 사전기반 방법에 의하면 주어진 문장은 부정적 어조로 분류된다. 하지만 실제로 '견조한 이익의 흐름이 지속된다'는 것은 아주 긍정적인 표현이다. KR-FinBert-SC 모델은 '4월 손해율 하락으로 견조한 이익흐름 지속'이라는 문장에 대해 긍정적 어조일 확률을 0.911로 추정한다. 이로써 사전기반 감성분류와 달리 KR-FinBert-SC는 단어의 사전적 의미뿐만 아니라 단어가 사용된 문맥의 의미도 함께 고려하는 자연어 처리 모형임을 확인할 수 있다.

3. 분석자료

본 절에서는 분석에 사용할 자료를 소개한다. 한국경제 신문사에서 운영하는 한경 코리아마켓 웹페이지를 통해 총 9,691개의 국내 기업분석 애널리스트 보고서를 수집하였으며, 분석에 활용할 주요 변수가 모두 있는 보고서는 총 9,425개다. 자료 수집 기간은 2017년 1월부터 2018년 12월까지이다. 애널리스트 보고서에는 보고서 대상 기업, 작성 애널리스트 이름, 애널리스트의 소속 증권사명, 투자 관련 추천의견, 목표주가, 이익 예측치 등의 정량적 정보와 함께 보고서 제목, 보고서 내의 회계 정보 수치에 대한 근거를 설명하는 본문이 존재한다. 개별 애널리스트 보고서와 보고서가 발표된 시점에서의 분석 대상 기업의 초과 수익률을 매치하여 분석에 활용하였다.

모형에 사용하는 변수들을 반응 변수, 관심 변수, 통제 변수 3가지 범주로 구분하여 소개하고자 한다. 첫 번째 반응 변수는 애널리스트 보고서 공시 당일의 초과수익률에 애널리스트 보고서 공시일 기준 전후로 각각 이틀치의 초과수익률을 누적하여 총 5일 동안의 누적초과수익률을 사용하였으며, CAR$(-2, 2)$로 표현하였다. 애널리스트 보고서 발표일 전 이틀치의 수익률을 같이 고려한 이유는 보고서 발표 전에도 투자 관련 정보 유출이 발생한다는 선행연구를 참고하였다 (Cho 등, 2012). 애널리스트 보고서 발표 후 이틀치의 수익률은 투자 정보로 인한 효과가 영향을 미치는 기간의 수익률이라고 판단하여 반응 변수에 포함시켰다.

두 번째, 관심 변수는 본 연구에서 기업수익률에 대한 예측력 존재 여부를 검증하고자 하는 어조 변수이다. 본문과 제목 중 어떤 텍스트가 기업수익률 예측력에 더 도움이 되는 정보인지 비교하기 위해 애널리스트 보고서의 어조를 본문과 제목에서 각각 추출했다. 본문에서 추출한 어조 변수는 PCT_POS와 PCT_NEG로 구성된다. PCT_POS는 보고서 본문에서 긍정문장 수의 비율을, PCT_NEG는 보고서 본문에서 부정문장 수의 비율을 뜻한다. 제목에서 추출한 어조 변수는 어조가 긍정으로 판단되는 확률 PROB_POS와 어조가 부정으로 판단되는 확률 PROB_NEG로 정의하였다.

마지막으로 통제 변수는 기업수익률 예측에 관한 기존 문헌을 참고하여 설정했다. 모형의 통제 변수로 CAR$(-7, -3)$, size, B/M , UP_GR, DOWN_GR, ΔTPRC, Nanalyst를 포함시켰다. CAR$(-7, -3)$은 애널리스트 보고서 공시일로부터 7일 전에서 3일 전까지, 총 5일 동안의 누적초과수익률이다. CAR$(-7, -3)$을 통제 변수로 포함시킨 이유는 애널리스트가 최근 뉴스나 사건에 의존하는 것에 대한 효과를 통제하기 위해서다. 그 외의 통제 변수들은 2.3.1에서 설명되었던 변수들이다. 추천의견은 증권사마다 차이가 존재하지만, Kim과 Eum (2006)의 분류 방식을 차용하여 매도(1), 비중축소(2), 중립(3), 매수(4), 적극 매수(5)의 5단계로 구분했다. 그리고 이로부터 UP_GR, DOWN_GR를 계산하였다.

4. 분석방법

본 절은 먼저 관심 변수가 반응 변수에 대한 예측력이 있는지를 선형모형 가정을 완화하여 검증할 수 있는 Coleman 등 (2022)의 방법을 소개한다. 추가적으로 변수들간의 예측력을 비교하기 위해 사용되는 Davidson-Mackinnon test (Davidson과 MacKinnon, 1981)가 Coleman 등 (2022)의 방법을 이용함으로써 어떻게 확장될 수 있는지 제시한다.

4.1. Random forest (RF) 기반 F-test

먼저, 전체 데이터셋을 학습 데이터셋 $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$과 테스트 데이터셋 $T = \{(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_N, \tilde{y}_N)\}$로 나눈다. 그리고 학습 데이터셋을 관측치를 행으로 하고, 변수들을 열로 하는 (종속 변수를 제1열로 둠) 행렬의 형태로 정리한 것도 D_n으로 표시하기로 하자. 관심 변수들의 유의성 검증을 위해 학습 데이터 행렬 D_n에서 관심 변수 집합에 속하는 변수들의 열을 모아 놓은 행렬에 행을 섞어 주는 permutation을 적용한 다음 각각의 permutation된 열을 D_n의 원래 자리의 열과 교체해줌으로 permutation된 학습데이터 행렬 D_n^{π}을 생성한다.

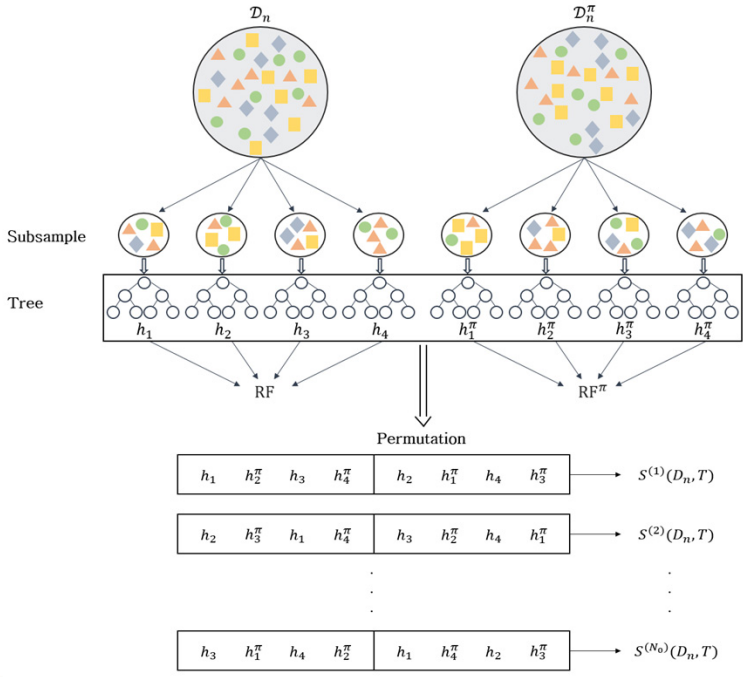

Figure 1: *Illustration of permutation idea in Coleman et al. (2022) (h_i is the i^{th} individual tree of the random forest and N_0 is the total number of the permutation).*

Permutation 여부에 따라 구분되는 2가지 학습데이터 행렬 D_n과 D_n^π을 랜덤 포레스트 모델로 각각 훈련시킨다. Permutation을 하지 않은 학습 데이터셋에 대해 훈련한 랜덤 포레스트 모델은 RF, permutation을 한 학습 데이터셋에 대해 훈련한 랜덤 포레스트 모델은 RF^π로 정의된다. 이때, 학습 데이터와 동일한 size를 갖는 부스트랩 자료에 tree를 적합하여 앙상블하는 것이 아니라 학습 size보다 적은 개수인 $k(\leq n)$의 자료로 subsampling한 자료에 tree를 적합하여 앙상블을 진행한다. 2종류의 학습기 RF과 RF^π를 테스트 데이터셋에 적용하여 구한 평균제곱오차(MSE)는 다음과 같이 정의된다.

$$\mathrm{MSE}_{RF}(T) = \frac{1}{N_t}\left\|\frac{1}{B}\sum_{b=1}^{B} T_b - y\right\|_2^2, \quad \mathrm{MSE}_{RF^\pi}(T) = \frac{1}{N_t}\left\|\frac{1}{B}\sum_{b=1}^{B} T_b^\pi - y\right\|_2^2 \quad (\|\cdot\|_2 : L_2 \text{ norm}). \tag{4.1}$$

식 (4.1)에서 B는 앙상블한 tree의 개수를 의미하고 y는 테스트 데이터셋의 반응 변수 값을 모은 벡터이다. T_b는 D_n의 b 번째 부스트랩 자료에 적합된 랜덤 포레스트를 테스트 데이터셋에 적용하여 얻은 y의 추정치 벡터이고, T_b^π는 D_n^π의 b 번째 부스트랩 자료에 적합된 랜덤 포레스트를 테스트 데이터셋에 적용하여 얻은 y의 추정치 벡터이다.

관심 변수 집합의 유의성 여부를 판단하기 위한 귀무가설과 대립가설은 아래와 같다.

$$H_0 : E\,\mathrm{MSE}_{RF}(T) = E\,\mathrm{MSE}_{RF^\pi}(T) \quad \text{vs.} \quad H_1 : E\,\mathrm{MSE}_{RF}(T) < E\,\mathrm{MSE}_{RF^\pi}(T). \tag{4.2}$$

(4.2)의 가설을 검정하기 위한 검정 통계량으로는 $S(T) = \mathrm{MSE}_{RF^\pi}(T) - \mathrm{MSE}_{RF}(T)$을 사용하고 검정통계량 값이 클수록 관심 변수 집합이 반응 변수 예측에 유의성을 갖는다고 할 수 있다. Coleman 등 (2022)에서는 검정통계량의 귀무가설 하에서의 분포를 유도하기 위해 RF, RF^π를 구성하는 $2B$개의 tree를 permutation하는 idea를 사용하였다. 이를 도식화하면 Figure 1과 같다.

Table 2: Decision process in the encompassing test

Model 1 vs. Model 3	Model 2 vs. Model 3	Result
Can be reduced to Model 1	Cannot be reduced to Model 2	Choose Model 1
Cannot be reduced to Model 1	Can be reduced to Model 2	Choose Model 2
Can be reduced to Model 1	Can be reduced to Model 2	Cannot determine superiority between models

애널리스트 보고서의 본문 기반 어조 변수들과 제목 기반 어조 변수들이 기업수익률에 예측력이 있는지를 검증하고 두 가지 종류의 변수들의 예측력을 비교하기 위해 모델을 아래와 같이 설정하였다 (controls: 통제변수 전체).

$$\text{Model 1} : \text{CAR}(-2,2) = f_1(\text{controls}, \text{PCT\_POS}, \text{PCT\_NEG}) + \epsilon, \tag{4.3}$$

$$\text{Model 2} : \text{CAR}(-2,2) = f_2(\text{controls}, \text{PROB\_POS}, \text{PROB\_NEG}) + \epsilon. \tag{4.4}$$

각각의 모델에 Coleman 등 (2022)이 제안한 랜덤 포레스트 F-test를 적용시켜 보고서의 어조 변수들의 유의성을 검정하고자 한다.

4.2. Davidson-Mackinnon test의 확장

본 연구에서는 보고서 본문에서 추출된 변수와 보고서 제목에서 추출된 변수 중 어떤 것이 기업의 수익률 예측에 더 도움이 되는지를 검증하고자 한다. 이는 4.1에서 제시한 Model 1과 Model 2중 어느 모델의 기업수익률 예측력이 더 좋은지를 알아내는 non-nested 모델 간 비교문제이다. 이러한 non-nested 모델 간 비교를 위해 사용하는 전통적인 방법은 Davidson-Mackinnon test이다. 하지만, 전통적인 Davidson-Mackinnon test는 선형모형 기반의 F-test를 기반으로 한다. 이에 본 연구에서 선형모형 가정으로 인한 편향된 결론을 피하기 위해 해당 F-test를 모두 Coleman 등 (2022)에서 제안된 랜덤 포레스트 기반 F-test로 대체하여 Model 1과 Model 2의 예측력을 비교하였다. 구체적으로 본 연구에서는 Davidson-Mackinnon test의 두 가지 버전 중 하나인 Encompassing test를 사용하였다. Encompassing test는 Model 1과 Model 2의 변수들이 모두 포함된 encompassing 모델인 Model 3

$$\text{Model 3} : \text{CAR}(-2,2) = f_3(\text{controls}, \text{PCT\_POS}, \text{PCT\_NEG}, \text{PROB\_POS}, \text{PROB\_NEG}) + \epsilon \tag{4.5}$$

을 생성하고 Table 2과 같이 모형간 비교를 변수들의 유의성 검정을 통해 실시하여 non-nested 관계에 있는 두 모형의 예측력을 비교한다.

5. 분석결과

5.1. RF F-test 기반 검정 결과

RF F-test를 위한 데이터셋의 분할은 데이터의 관측순서를 고려하여 시간순서상 앞의 70퍼센트를 학습데이터, 뒤의 30퍼센트를 테스트 데이터로 실시하였고, 가설검정을 위한 유의수준은 편의상 0.05로 하였다. subsample size인 k는 $\sqrt{n}$ (n: 학습데이터의 크기)로 하였고 RF 학습을 위한 bootstrap iteration 수인 B는 1,000으로 하였다. RF F-test의 구현을 위해 https://github.com/tim-coleman/RFtest/에 제시되어 코드를 수정하여 사용하였다.

5.1.1. 본문 기반 어조 변수와 제목 기반 어조 변수의 유의성 검정

Model 1에서 PCT_POS, PCT_NEG에 대해 RF F-test로 유의성을 검정하면 검정 통계량 $S(T)$의 값은 0.8054이고 P-value는 0.0020으로 귀무가설을 기각한다. 즉, 본문 기반 어조 변수는 기업수익률에 대한 예측력이 있다고 판단된다. Model 2에서 PROB_POS, PROB_NEG에 대해 RF F-test로 유의성을 검정하면 검정 통계량 $S(T)$의 값은 0.3704이고 P-value는 0.0020이다. 따라서, 귀무가설을 기각하게 되어 제목기반 어조 변수 역시 기업수익률 예측력이 있다고 판단할 수 있다.

5.1.2. 어조 변수 본문과 제목의 기업수익률 예측력 비교

Model 1과 Model 3을 비교하는 RF F-test의 검정 통계량 $S(T)$값은 0.1155, P-value는 0.1238가 나온다. 따라서 귀무가설을 채택하게 되고 Model 3에서 제목 기반 어조변수를 생략하여도 기업수익률 예측력에 대한 유의한 감소가 일어나지 않는다고 볼 수 있다. Model 2와 Model 3를 비교하는 RF F-test의 결과는 검정 통계량 $S(T)$의 값은 0.6500, P-value는 0.0020로 귀무가설을 기각하게 된다. Model 3에서 본문 기반 어조 변수를 생략하면 기업수익률 예측력에 대한 유의한 감소가 나타난다는 결론이다. 따라서, 두 RF F-test 결과를 종합하면 기업수익률 예측력 측면에서 어조 변수 본문을 포함한 모델인 Model 1이 어조 변수 제목을 포함한 모델인 Model 2보다 우수하다는 결론을 낼 수 있다. 이것은 애널리스트 보고서의 텍스트 정보력에 대한 해외 선행연구인 Huang 등 (2014), Liang 등 (2022)에서 본문 기반 어조 변수를 사용하고 있는 것에 대한 근거로써 이해할 수 있다.

5.2. 선형모형 하에서의 검정 결과

5.1절에서 이루어진 분석을 선형모형 기반의 F-test로 진행했을 때 얻은 결과들과 비교하여 살펴보고자 한다. 선형모형 하에서 본문 기반 어조 변수의 기업수익률 예측에 대한 유의성 검정 결과, 귀무가설을 기각하여 RF F-test 결과와 마찬가지로 본문 기반 어조 변수가 예측력이 있는 것으로 나타난다. 제목 기반 어조 변수의 유의성 검정 결과도 RF F-test 결과와 동일하게 유의한 것으로 나타난다.

본문 기반 어조 변수와 제목 기반 어조 변수의 예측력 비교를 위해 먼저 선형모형 가정 하에서 Model 3과 Model 1를 F-test로 비교했을 때, Model 3에서 어조 변수 제목을 생략할 수 없다는 결과를 얻었다. 또한 Model 3과 Model 2를 선형모형 가정 기반의 F-test로 비교하면, Model 3에서 어조 변수 본문을 생략할 수 없는 것으로 나온다. 즉, 선형모형 가정 하에서는 Model 1과 Model 2가 예측력 측면에서 한 모델이 다른 모델보다 낫다는 판단을 내릴 수 없다. 따라서, 선형모형 가정 하에서는 두 종류 어조 변수를 모두 포함한 모델을 고려하는 것이 적절하다고 볼 수 있다. 이 결과를 5.1의 결과와 비교하여 보면 선형모형이라는 제약하에서는 예측력 확보를 위해 비선형모형에서 필요한 변수 개수보다 더 많은 개수의 변수를 필요로 하는 측면이 있음을 보여준다.

5.3. 사전 기반 어조 추출 방식과의 비교

5.1과 5.2에서는 제목 기반 어조 변수에 대해 BERT 모델로 어조 변수를 생성하였다. 본 절에서는 사전기반 방법으로 애널리스트 보고서 제목의 어조 변수를 생성해 보고, BERT 모델을 이용했을 때의 결과와 비교해 보고자 한다. 사전 기반 방법에 따른 애널리스트 보고서 제목의 어조 변수는 CNT_POS와 CNT_NEG로 구성된다. CNT_POS는 보고서 제목을 구성하는 긍정단어의 수가 부정단어의 수보다 많으면 1, 그렇지 않으면 0의 값을 가진다. 이와 반대로 CNT_NEG는 부정단어의 수가 긍정단어의 수보다 많으면 1, 그렇지 않으면 0을 가진다. Model 2에서 BERT 모형으로 생성된 제목 기반 어조 변수를 사전 기반으로 계산된 어조 변수 CNT_POS와 CNT_NEG로 바꾼 모형을 Model 4라고 하고 Model 4에서 RF F-test와 선형모형 기반 F-test를 통해 제목 기반 어조 변수의 유의성 검정을 실시하였다.

RF F-test의 검정 통계량 $S(T)$의 값은 0.1190이고 P-value는 0.1756으로 귀무가설을 기각하지 못했다. 따라서, 5.1절에서 보고서 제목의 어조를 BERT로 추출한 어조 변수가 유의성을 갖는 결과와는 다르게 애널리스트 보고서 제목의 어조를 사전 기반으로 추출한 어조 변수는 기업수익률 예측에 정보력이 없다는 결론을 얻게 된다. 추가적으로 보고서 제목에서 사전 기반으로 추출한 어조 변수의 유의성을 선형모형 가정의 F-test로 진행하면, 귀무가설을 기각하게 되어 보고서 제목에서 사전 기반으로 어조를 추출한 어조 변수가 예측 정보력이 있는 것으로 확인된다. 하지만 분석자료에 대해 7 : 3의 비율로 학습 데이터셋과 테스트 데이터셋을 분할하여 랜덤 포레스트 모형을 적합한 후, 테스트 데이터셋에서 MSE를 비교하면 기존의 변수들에 사전기반 어조 변수를 추가로 포함시킨 모형이 그렇지 않은 모형보다 오히려 MSE가 더 낮게 나온다. 이는 사전 기반 어조가 예측 정보력이 있다는 결론이 적합하는 모형을 단순한 구조를 가지는 선형모형으로 제한함으로써 나타나는 제한적 결론임을 파악할 수 있다. 따라서 사전 기반 어조 변수의 유의성은 기업수익률 예측에 다양한 머신러닝 모델을 활발하게 사용할 수 있는 현재의 상황에서 의미있는 결론으로 이해되기 어렵다.

이상의 분석을 정리하면, 보고서 제목의 어조는 추출하는 방식의 문제로 예측력을 상실하는 문제가 발생할 수 있다. 또한 선형모형 가정하에서 변수의 예측력을 판단할 경우, 예측유용성이 떨어지는 변수를 유용성 있는 변수로 인식할 우려가 존재한다.

6. 결론

본 연구의 주요 결과는 다음과 같다. 첫 번째, 애널리스트 보고서 상의 본문 기반 어조와 제목 기반 어조를 비선형모형 가정에서 비교했을 때, 선형모형 가정에서 비교했을 때와 다르게 본문 기반 어조 변수가 좀 더 기업수익률 예측 능력이 있다고 판단되었다. 두 번째, 선형모형 기반으로 기업수익률 예측 정보력을 검증하는 경우 수익률 예측을위해 다양한 머신러닝 모델을 사용할 수 있는 현재의 상황에서는 예측정보력이 없는 변수를 예측정보력이 있는 변수로 불필요하게 식별하는 잘못된 판단의 가능성을 확인했다. 따라서, 의미있는 변수식별을 위해서는 선형모형 기반 F-test 뿐만 아니라 비모수모형 기반 F-test를 함께 사용하는 것을 검토해야 한다고 판단된다. 마지막으로 보고서 어조의 기업수익률 예측정보력이 있는가에 대한 여부는 보고서의 어조를 계산하는 방법의 영향을 받는 것을 확인했다. 어휘사전 기반으로 어조 변수를 생성한 경우, 기업수익률 예측정보력이 없는 것으로 나왔기 때문이다. 따라서 본 연구에서 사용한 KR-FinBert-SC와 같이 금융 텍스트의 어조 추출 시 문맥을 고려할 수 있는 자연어 처리 모형을 사용하는 것을 추천한다.

References

Barber BM, Lehavy R, and Trueman B (2010). Ratings changes, ratings levels, and the predictive value of analysts' recommendations, *Financial Management*, **39**, 533–553.

Bradley D, Clarke J, Lee S, and Ornthanalai C (2014). Are analysts' recommendations informative? intraday evidence on the impact of time stamp delays, *Journal of Finance*, **69**, 645–673.

Cho SS, Byun JH, and Park SH (2012). Short-Selling behavior of investor groups before analyst downgrades, *The Korean Journal of Financial Management*, **29**, 191–231.

Coleman T, Peng W, and Mentch L (2022). Scalable and efficient hypothesis testing with random forests, *Journal of Machine Learning Research*, **23**, 1–35.

Davidson R and MacKinnon JG (1981). Several tests for model specification in the presence of alternative hypotheses, *Econometrica*, **49**, 781–793.

Devlin J, Chang MW, Lee K, and Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 4171–4186, Available from: arXiv preprint arXiv:1810.04805

Huang AH, Zang AY, and Zheng R (2014). Evidence on the information content of text in analyst reports, *The Accounting Review*, **89**, 2151–2180.

Jang JK, Lee KH, and Lee ZK (2016). How the title of investment strategy report affects stock price forecast: Using text mining method, *The Korean Journal of Bigdata*, **1**, 21–34.

Kim DS and Eum SS (2006). The impact of analysts' revisions in their stock recommendation and target prices on stock prices, *Asia-Pacific Journal of Financial Studies*, **35**, 75–108.

Kim E and Shin H (2022). KR-FinBert: Fine-tuning KR-FinBert for sentiment analysis, Available from: https://huggingface.co/snunlp/KR-FinBert-SC

Kim TH and Lee SY (2013). Do the firm's exposures to SNS affect their stock prices in Korea?, *The Korea Society of Management Information Systems*, 491–499.

Liang D, Pan Y, Du Q, and Zhu L (2022). The information content of analysts' textual reports and stock returns: Evidence from China, *Finance Research Letters*, **46**, 102817.

McAlexander RJ and Mentch L (2020). Predictive inference with random forests: A new perspective on classical analyses, *Research & Politics*, **7**.

Yang CW (2021). Information content of analyst report title: Focusing on the TONE of text, *Korean Journal of Financial Management*, **38**, 1–38.

Received April 7, 2023; Revised May 1, 2023; Accepted May 13, 2023

애널리스트 보고서 텍스트의 주가예측력에 대한 검증

이영선[a], 야마다 아키히코[b], 양철원[c], 노호석[1,a,d]

[a]숙명여자대학교 통계학과; [b]서울대학교 빅데이터 혁신융합대학; [c]단국대학교 경영학부;
[d]숙명여자대학교 자연과학연구소

요 약

온라인 플랫폼을 통한 애널리스트 보고서의 공유가 가능해짐에 따라 애널리스트들이 생성한 보고서는 시장 참여자들 간 금융 정보 격차를 줄일 수 있는 유용한 도구가 되었으며, 애널리스트 보고서의 정량적 정보가 주식수익률 예측에 다수 활용되었다. 하지만 상대적으로 애널리스트 보고서 내 텍스트 정보의 주식수익률 예측 정보력에 대한 국내 자료 기반 연구는 상대적으로 많이 부족하다. 본 연구는 애널리스트 보고서에서 추출 가능한 텍스트로부터 어조 변수를 생성하여 주식수익률 예측에 정보력이 있는지를 검증하되, 기존 연구들의 선형모형 가정 기반 검정의 한계를 해결하고자 랜덤 포레스트 기반의 F-test를 사용하여 기업수익률 예측력을 검증하였다.

주요용어: 주식수익률 예측가능성, 자연어 처리, 애널리스트 보고서, 랜덤 포레스트 F-test

[1]교신저자: (04310) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과. Email: hsnoh@sookmyung.ac.kr