NH투자증권

SNS 텍스트 데이터를 활용한 이벤트 효과 투자 전략

AI/Big Data Strategy | 2022. 8. 10

실적 서프라이즈, ESG 이슈 등 이벤트는 주가의 단기 변동성을 확대시킵니다. 이러한 이벤트 효과를 정량적으로 측정할 수 있다면 주가 상승 가능성이 높은 종목을 찾을 수 있습니다. SNS 데이터와 자연어처리 기법을 활용하여 이벤트 효과를 측정하고, S&P 100 유니버스 내 종목 선정을 통한 알파 전략을 제시합니다.

김규진 (AI/빅데이터전략)
02)2229-6737, kyoujin.kim@nhqv.com

SNS 텍스트 데이터 활용해 기업 간 상관관계에서 이벤트 효과 분해

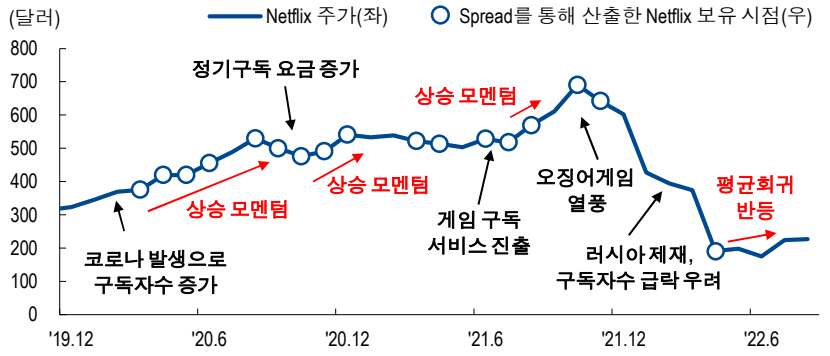
실적 서프라이즈, ESG 이슈, 시장 지배력 감소 등 이벤트는 단기적으로 시장과 무관한 주가 변동을 초래한다. 이로 인해 기업 간 상관관계는 이벤트 효과를 반영한다. 본 자료에서는 기업 간 상관관계의 변동에서 이벤트 효과를 분해하여 주가 상승 확률이 높은 종목을 선정하는데 활용했다. 이벤트 효과는 SNS 채널 중 하나인 트위터의 텍스트 데이터와 Word2Vec 자연어처리 기법을 활용해 분해하였다.

이벤트 효과 기준으로 선정한 종목은 시장을 아웃퍼폼

이벤트 효과는 펀더멘털 관련성이 높은 SNS 상관관계(SNS 텍스트 기준 기업 간 상관관계)에서 이벤트 발생에 따른 단기 주가 변동이 반영된 주가 상관관계(주가 기준 기업 간 상관관계)를 뺀 Spread 값으로 측정했다.

결과적으로 Spread가 큰 종목일수록 주가가 상승할 가능성이 높았다. 이벤트 효과가 높게 측정된 종목군으로 구성한 포트폴리오는 시장지수(S&P 100)를 연평균 3.3%p 아웃퍼폼하였다. 특히 위기 국면에서 시장지수보다 상대적으로 작은 낙폭을 기록하였다.

정량화된 이벤트 효과(Spread)는 종목 보유 시점 파악에 유효: NETFLIX 사례



주: 보유시점은 이벤트 효과값이 상위 90% 이상인 경우
자료: Refinitiv, NH 투자증권 리서치본부

본 자료는 서울대 데이터마이닝센터와의 산학협력을 통해 작성되었습니다.

I. 기업 간 상관관계와 이벤트 효과

1. 기업 간 상관관계는 기업 관련 정보를 반영

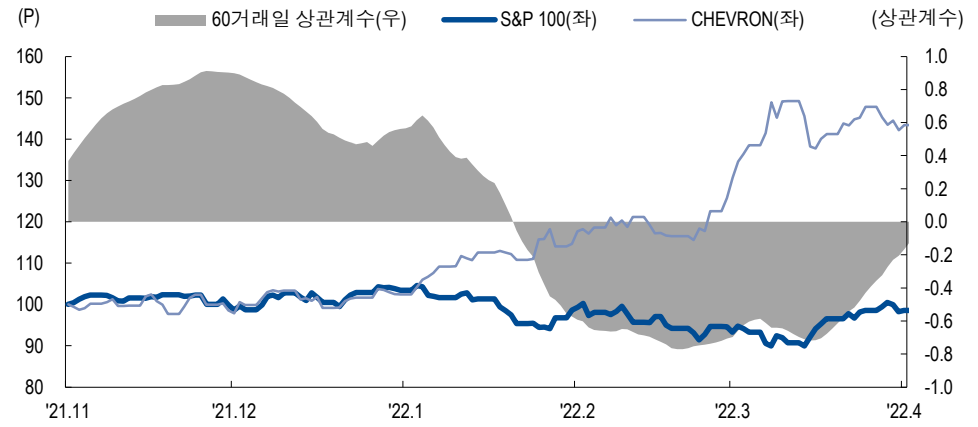
상관관계는 기업관련 이벤트 정보를 반영

2022년 3월 미국을 포함한 글로벌 지수는 동반 하락하였으나, 유가 및 원자재 가격은 러시아-우크라이나 전쟁의 영향으로 급등하였다. 이에 따라 에너지 섹터 기업의 주가도 상승세를 기록하였다.

에너지 기업 CHEVRON과 S&P 100의 상관관계를 살펴보면 2022년 1월부터 방향성이 엇갈리며 2월에는 상관계수가 음(-)으로 전환되었음을 확인할 수 있다(그림1 참조). 지수 하락 시기에 지수와의 상관계수가 하락하던 CHEVRON을 골라 투자하였다면 주가 하락을 방어할 수 있었을 것이다.

주가는 기업 관련 정보를 담는다. 심지어는 미래에 대한 기대와 현재의 루머까지도 반영한다. 기업 간 상관관계 역시 주가를 기준으로 산출하므로 기업 관련 정보를 지니고 있다고 볼 수 있다. 이에 기업 간 상관관계를 잘 활용한다면 정성적 판단 없이 정량적 기준만으로 종목 이벤트 발생 여부를 판단하고 주가가 상승할 가능성이 높은 CHEVRON과 같은 종목을 선정할 수 있을 것이다.

그림1. 기업 간 상관관계는 기업 관련 정보를 담고 있음



자료: Refinitiv, NH 투자증권 리서치본부

유의미한 정보로서의 효용을 지니는 상관관계는 이미 다양한 투자 분야에서 활용되고 있다. Harray Markowitz가 체계화한 MPT(Modern Portfolio Theory)의 평균-분산 전략에서 자산 간 상관관계는 중요한 지표이며, 헤지 펀드에서 핵심 전략으로 사용되는 롱숏 전략의 페어트레이딩에서도 자산 간 상관관계는 전략을 수립하기 위한 주요 지표로 활용된다.

기업 간 상관관계는 주가를 활용해 정량적인 방법으로 측정할 수 있지만, 이외에 GICS 섹터 분류, 거래소 산업 분류 등 사업의 내용을 통해 정성적으로 분류할 수도 있다(그림2 참조).

그림2. S&P Global에서 제작한 GICS 기업 섹터 분류 체계

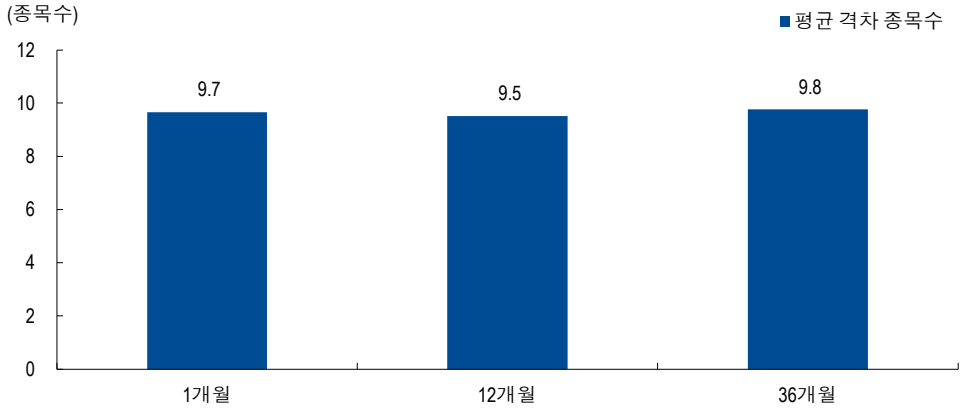
자료: S&P Global, NH 투자증권 리서치본부

주가는 장기적으로는 기업이 소속된 시장 및 산업의 상황에 따라 움직인다. 따라서 직관적으로는 주가 기준 기업 간 상관관계와 섹터 기준 기업 간 상관관계가 동일해야 한다. 그러나 실제 주가를 기준으로 분류된 섹터와 GICS 기준으로 분류된 섹터의 차이가 비교적 크다.

주가수익률 추세를 기준으로 군집화 알고리즘(K-Means*)을 사용하여 S&P 100의 기업을 24개 산업 그룹으로 구분해 실제 GICS 섹터 분류와 비교한 결과, 종목당 평균 10개 종목이 다르게 배정되었다(그림3, 4 참조).

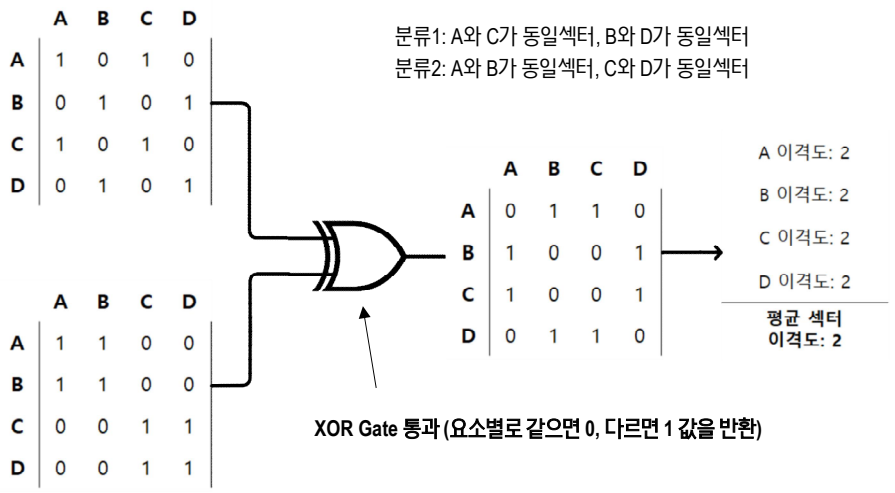
* K-Means 방법론에 대한 세부사항은
[김규진. (2021). AI와 대체데이터를 활용한 KOSPI 예측 모델. 2021, 26pg] 참조

그림3. 섹터 분류 격차는 평균 10개 종목



자료: NH투자증권 리서치본부

그림4. 섹터 분류 격차 계산 예시



자료: NH투자증권 리서치본부

주가 기준 기업 간 상관관계 관련 종목 정보 (소속시장+소속산업 +이벤트 효과)

이와 같은 현상이 발생하는 이유는 두 가지 기준의 기업 간 상관관계와 관련된 종목 정보에 차이가 있기 때문이다.

주가 기준 기업 간 상관관계 관련 종목 정보: 소속시장 + 소속산업 + 이벤트 + ε

* ε는 수급으로 인해 발생한 일시적인 상관관계 변동 요인. 만약 기업이 충분히 안정되어 있으며, 산업의 변동이 단기간에 이루어지기 어려운 대형종목이라면 ε는 장기적으로는 0에 수렴

섹터 기준 기업 간 상관관계 관련 종목 정보: 소속시장 + 소속산업(지연)

소속시장 + 소속산업 분류는 종목의 장기 상관관계 추세에 영향을 미치는 요인인 반면, 이벤트는 영향 범위에 따라 산업 이벤트, 종목 이벤트로 나눌 수 있으며 종목의 단기 상관관계 추세에 영향을 미치는 요인이다.

앞서 살펴본 CHEVRON 사례의 경우 기업 소속시장과 소속산업에는 변동이 없었으나, 유가 상승에 따른 실적 증가라는 이벤트가 발생하며 주가가 상승하며, 다른 기업과의 상관관계가 하락하였다. 즉, CHEVRON의 기업 간 상관관계의 변동을 통해 이벤트 효과를 파악할 수 있다.

그러나 모든 경우에 단순히 상관계수를 사용하는 것에는 문제가 있다. 우선 정성적인 분석 없이는 상관관계 변동이 소속시장+소속산업 변동이 아닌 이벤트 발생만으로 초래되었는지 파악하기 어려우며, 어느 정도 변동했을 때 투자하여야 할 지에 대한 기준점을 설정하는 것 역시 어렵다.

만약 '주가 기준 기업 간 상관관계 관련 종목 정보'에서 '섹터 기준 기업 간 상관관계 관련 종목 정보'를 제하여 이벤트 효과를 분해해낼 수 있다면 보다 용이한 투자가 가능할 것이다. 문제는 GICS 섹터 분류는 정량 데이터에 기반하지 않으므로 주가 기준 기업 간 상관관계로부터 산술적으로 이벤트 효과를 분해해낼 수 없다는 점에 있다. 또한 섹터 분류는 정기적으로 발표되므로 실제보다는 정보가 지연되었을 수 있다.

이에 실시간 '소속시장 + 소속산업' 정보를 갖는 데이터를 찾고 이를 정량화하는 방법을 논의한다.

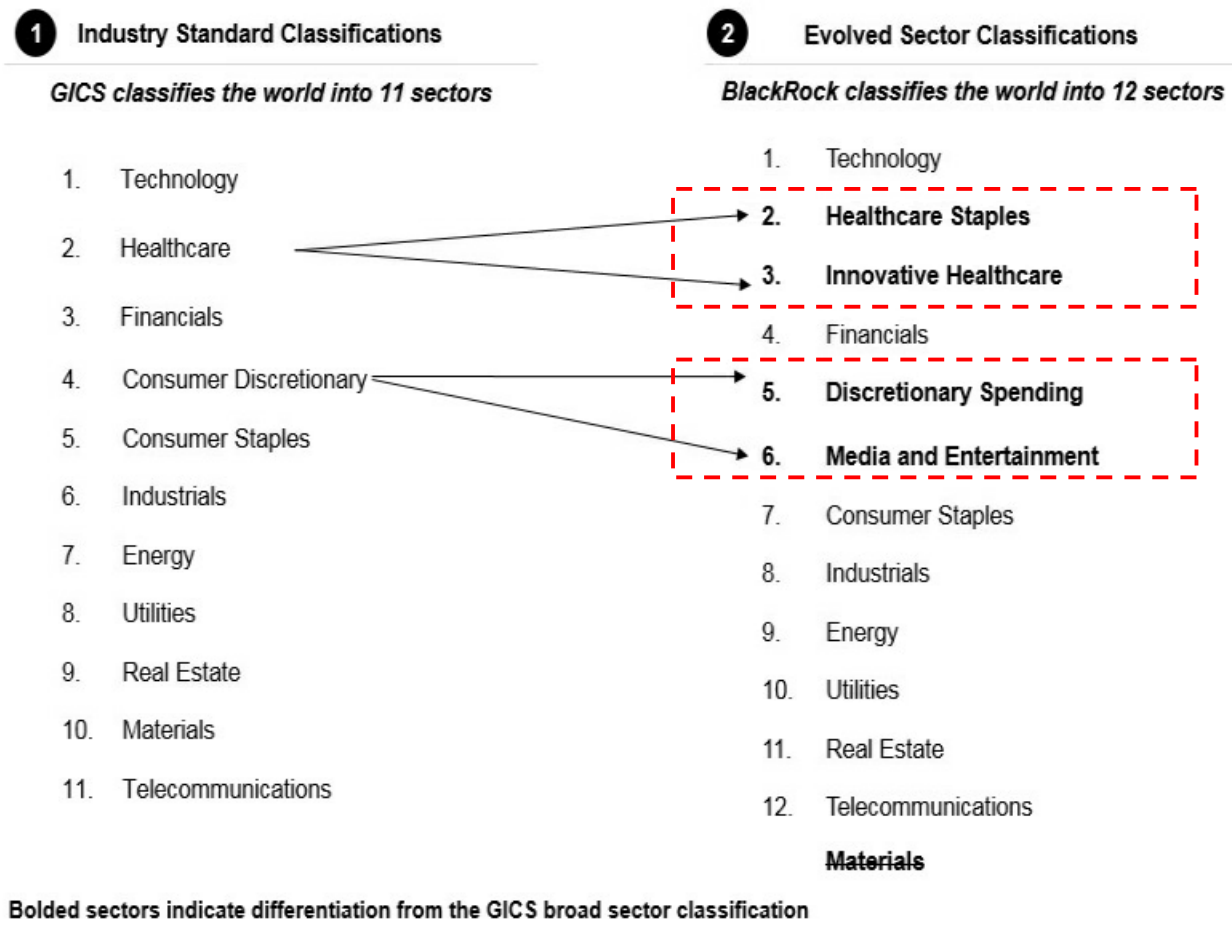
2. SNS 상관관계를 활용하여 이벤트 효과 분해

기업 간 상관관계는 머신러닝 자연어처리 기법을 활용해 측정하기도 한다. 기업 관련 텍스트를 분석하여 해당 기업 관련 텍스트가 다른 기업 관련 텍스트와 얼마나 유사성을 갖는지 측정하는 방식이다. 이 때 사용되는 텍스트에는 기업공시(사업보고서), 뉴스, SNS 등이 있다.

BlackRock이 출시한 ishares Evolved Sector ETF는 기업 10k 공시나 어닝 콜 자료에 쓰여진 텍스트 데이터를 NLP 분석해 섹터를 새롭게 분류한 상품이다. 기존 GICS와는 차별화된 섹터 분류 체계를 추가했다는 점 외에, 여러 섹터에 포함되는 기업에 대한 새로운 투자 방법론을 제시한다는 점에서도 의의가 있다(그림5 참조).

Evolved Sector ETF Series의 섹터 분류 예시는 다음과 같다. 우선 머신러닝 방법론에 의하여 Home Depot이 75%는 자유소비재 산업에, 25%는 산업재 산업에 속하는 것으로 분류한다. 이후 Home Depot 시총의 75%를 자유소비재 산업에 배정하며, 나머지 25%는 산업재 산업에 배정한다.

그림5. iShares Evolved Sector ETF Series의 분류 체계



자료: BlackRock iShares.com, NH투자증권 리서치본부

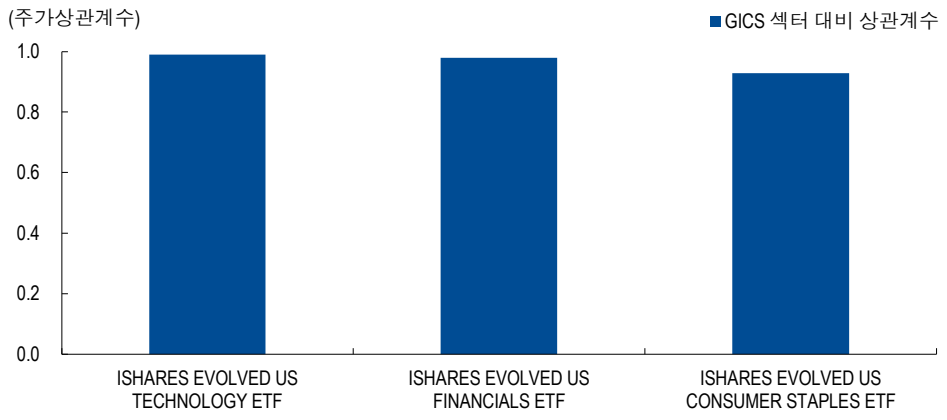
SNS 데이터 기준 기업 간 상관관계 종목 정보 (소속시장+소속산업)

Evolved Sector ETF의 섹터 분류는 텍스트에 기반하고 있음에도 4개 섹터를 제외하면 기존 GICS 섹터 ETF와 거의 동일한 주가 추세를 기록하며 높은 상관관계를 보인다(그림6 참조). 자연어처리 기법이 기업 간 관계 파악에 유효함을 알 수 있다.

다만 iShares Evolved ETF 섹터 분류에 사용된 기업공시 등의 데이터 역시 공시 주기가 분기라는 점에서 지연된 정보라는 단점을 지닌다.

따라서 **실시간으로 업데이트되는 SNS 데이터를 활용하여 지연되지 않은 '소속시장 + 소속산업' 정보를 추출**할 것이다. 이를 활용하면 주가 기준 종목 정보로부터 이벤트 효과를 분해하기 수월할 것이다.

그림6. Evolved 섹터 ETF와 GICS 섹터 ETF의 주가 상관관계 강해



자료: Refinitiv, NH투자증권 리서치본부

II. SNS 상관관계 추출 모델

1. SNS 텍스트 데이터 수집

SNS 채널 중 트위터에서 월간 기업별 1,000개의 텍스트를 추출하였다.

트위터 텍스트는 기업명을 기준으로 검색했으며, 유의미한 텍스트를 선별하기 위하여 '좋아요'가 1개 이상인 텍스트 만을 수집하였다. 월간으로 작성된 텍스트를 월말부터 월초까지 역순으로 추출하는 방식을 택해 작성된 텍스트의 개수가 많은 종목의 경우 텍스트 작성일이 월말에 치우친다는 한계점이 존재한다.

추출 기간은 2013년 1월부터 2022년 4월까지이며, 대상 종목군은 각 시기별 S&P 100 구성종목이다.

결과적으로 월 평균 5만 4천개, 연 평균 65만개의 텍스트 데이터를 추출했다.

2. SNS 텍스트 데이터를 활용한 기업 간 상관관계 측정 방법론

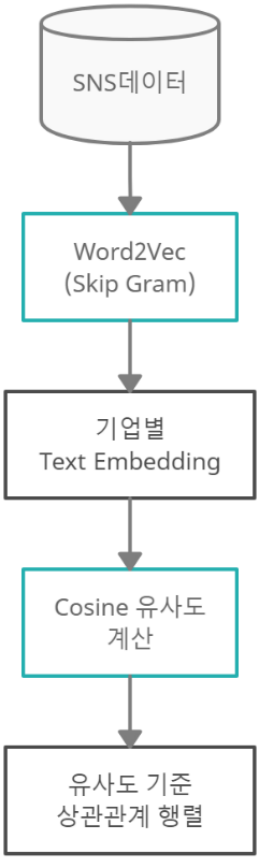
수집된 텍스트는 12개월 단위로 월별 Word2Vec 모델의 학습에 활용했다.

Word2Vec(Skip Gram) 모델은 특정 단어의 전후로 등장할 수 있는 단어를 예측하는 성능이 극대화되도록 단어의 임베딩을 정하는 모델이다. 참고로 임베딩이란 단어를 숫자 벡터로 정량화시킨 값을 의미한다.

이후, 학습된 모델을 통하여 정해진 기업별 임베딩을 출력한 후, 임베딩의 코사인 유사도를 측정하여 기업 간 상관관계를 측정한다.

* Word2Vec에 대한 추가설명은 Appendix A 참조
* 코사인 유사도는 벡터 간 유사도를 측정하기 위한 방법으로 평균값이 0인 경우, 코사인 유사도와 피어슨 상관계수는 동일

그림7. SNS 데이터를 활용한 기업 간 상관관계 측정 방법론



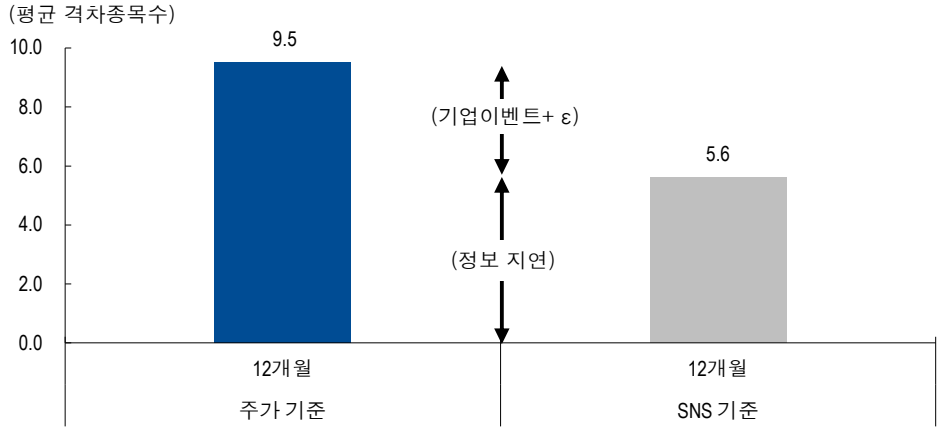
자료: NH투자증권 리서치본부

3. SNS 상관관계의 유효성 검증

앞서 섹터 분류 격차 측정 방법을 동일하게 적용하였을 때, SNS 기준 섹터 분류의 격차 값은 평균 5.6개 종목이었으며, 주가 기준 섹터 분류의 격차 값보다는 GICS 섹터와 적은 격차를 기록하였다(그림8 참조). 격차 값의 차이는 앞서 언급한 관련 종목 정보의 차이에 기인하는 것으로 보인다.

주가 상관관계 정보는 '소속시장 + 소속산업 + 이벤트 + ε'인 반면 SNS 상관관계 정보는 '소속시장 + 소속산업'이기 때문에 SNS 기준 섹터 분류가 주가 기준 섹터 분류보다 GICS 섹터 분류와 적은 격차를 기록하였을 것이다. 한편, SNS 기준 상관관계에서 발생한 일부 격차값은 GICS 섹터 분류가 지니는 지연된 '소속시장 + 소속산업' 정보에서 비롯된 것으로 해석할 수 있다.

그림8. SNS 텍스트 기준 섹터 분류는 주가 기준 섹터 분류보다는 GICS 섹터 분류와 더 유사하였음



자료: NH투자증권 리서치본부

SNS 기준 섹터 분류가 GICS 섹터 분류와 가장 유사했던 시기는 2018년 7월이다. 2018년 7월은 글로벌 주가지수가 횡보했던 시기로 수급 요인이 전반적으로 적어 주가 기준 섹터 분류도 어느 정도 GICS 섹터 분류와 유사했다.

SNS 기준 섹터 분류에서는 DISNEY[Media], NETFLIX[Retailing], TWENTY FIRST CENTURY FOX[Media]를 동일 섹터로 분류하고 있다는 점이 특징적이다. 2018년은 NETFLIX가 오리지널 콘텐츠에 투자하기 시작한 시기이며, 2019년은 DISNEY가 스트리밍 서비스 계획을 공개하는 한편, TWENTY FIRST CENTURY FOX 인수전에 뛰어들었던 시기이다. 이를 통해 **SNS 기준 섹터 분류가 해당 종목 소속산업의 변화를 반영**함을 알 수 있다.

주가 기준 섹터 분류에서 특징적인 부분은 INTEL을 NVIDIA나 여타 IT 종목들과 다른 섹터로 분류하였다는 점이다. 이는 2018년 당시 INTEL의 시장 지배력 감소 및 반도체 생산 지연 등의 기업 이벤트로 인해 주가가 급락하며 발생한 현상이다. 그러나 SNS 텍스트 기준으로 INTEL과 NVIDIA는 동일 섹터로 분류된다. 즉, **주가 기준 섹터 분류가 SNS 기준 섹터 분류 대비 이벤트를 민감하게 반영**하고 있다.

한편 2020년 3월은 코로나19 확산에 따른 시장 공포 국면으로 대규모 하락 장세가 연출되었던 시기이다. SNS 기준 섹터 분류는 GICS 섹터 분류와 유사하게 유지되었으나, 주가 기준 섹터 분류는 시장 공포에 수급이 동기화되어 다수의 연관 없는 기업이 동일 섹터로 분류되었다. 즉, **주가 기준 섹터 분류에는 수급 등의 ε 요인이 추가적으로 반영**된다.

III. 이벤트 효과 투자 전략

1. 이벤트 효과 산출

'SNS 상관관계 – 주가 상관관계' 통해 이벤트 효과 분해

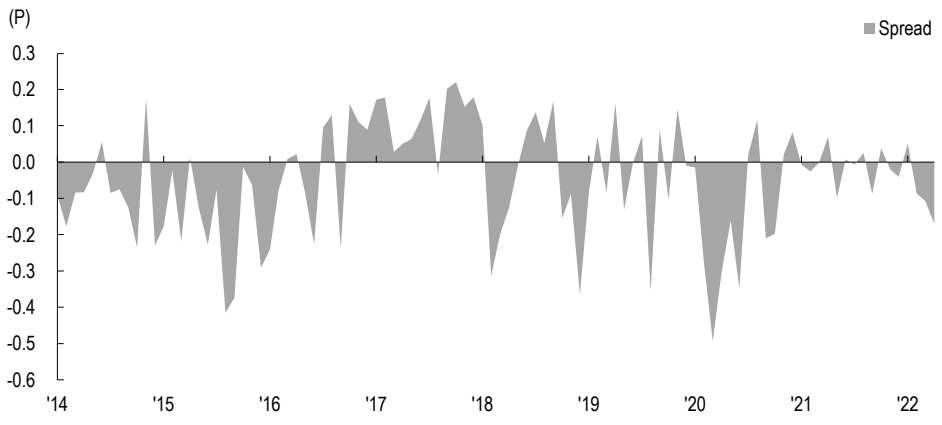
앞서 분석한 바에 따르면 SNS 기준 기업 간 상관관계(이하 SNS 상관관계)에 영향을 미치는 정보는 '소속시장 + 소속산업'이다. 주가 기준 기업 간 상관관계(이하 주가 상관관계)에 영향을 미치는 정보는 '소속시장 + 소속산업 + 이벤트 + ε'이다. 이 때 이벤트는 기업 간 상관관계에 일반적으로 음(-)의 영향을 미치므로 '소속시장 + 소속산업 – 이벤트 + ε'로 표현 가능하다.

따라서 이벤트 효과는 'SNS 상관관계 – 주가 상관관계' Spread와 같다.

SNS 상관관계 : 소속시장 + 소속산업
–) 주가 상관관계 : 소속시장 + 소속산업 – 이벤트 + ε
= (SNS 상관관계 – 주가 상관관계) Spread
≈ 이벤트(ε가 매우 작다면)

앞서 계산한 종목별 Spread를 기준으로 Spread가 클수록 이벤트 효과가 크게 발생했으며, 단기 주가 모멘텀이 존재할 것으로 가정한다.

그림9. S&P 100의 평균 종목별 Spread 추이



자료: Refinitiv, NH투자증권 리서치본부

2. 긍정적 이벤트와 부정적 이벤트 효과

상관관계 Spread는 주가 상승 이벤트 구분에 유효

앞서 이벤트 효과를 'SNS 상관관계 – 주가 상관관계' Spread로 정의하였다.

SNS 상관관계는 기업이 시장 내 기업들과 많은 관계를 맺고 있어 SNS상의 다양한 텍스트와 다양한 문맥에서 많이 이름이 언급될수록 상승하게 된다. **시장 전체에서 기업의 인지도나 점유율이 상승하는 경우에도 산업, 시장 내 기업의 비중이 상승하므로 SNS 상관관계 상승**으로 반영된다(시장/산업과의 상관관계 상승). 반대로 **기업의 인지도 및 점유율 하락은 SNS 상관관계 하락**으로 반영된다(시장/산업과의 상관관계 하락).

단기적인 이벤트는 12개월 텍스트를 기준으로 하는 SNS 상관관계에 반영되는 비중이 낮으며, 주가 상관관계에 더 먼저 반영된다. 단기 이벤트는 시장 대비 주가 변동성을 상승시키므로 긍정/부정 여부와는 관계없이 주가 상관관계를 하락시킨다.

따라서 장기 성장/침체 + 단기 긍정/부정의 이벤트 효과를 구분하면 다음과 같다.

(1) 장기 성장 + 단기 긍정 : 주가 상승 모멘텀 발생 (+)
- Spread 상승폭 확대 (SNS 상관관계 ↑, 주가 상관관계 ↓)

(2) 장기 성장 + 단기 부정 : 평균 회귀성 주가 반등 발생 (+)
- Spread 상승폭 확대 (SNS 상관관계 ↑, 주가 상관관계 ↓)

(3) 장기 침체 + 단기 긍정 : 평균 회귀성 주가 반락 발생 (-)
- Spread 상승폭 축소 (SNS 상관관계 ↓, 주가 상관관계 ↓)

(4) 장기 침체 + 단기 부정 : 주가 하락 모멘텀 발생 (-)
- Spread 상승폭 축소 (SNS 상관관계 ↓, 주가 상관관계 ↓)

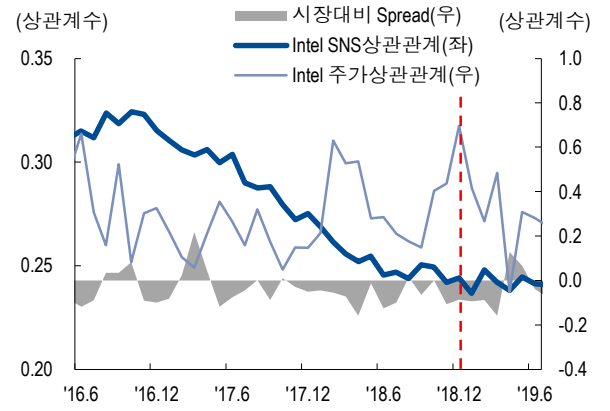
따라서 **Spread가 큰 종목에서 (1)주가 상승 모멘텀 또는 (2)평균 회귀성 주가 반등이 발생**하는 것으로 추정할 수 있다.

앞서 언급한 2018년 7월 INTEL의 반도체 시장 지배력 감소 이슈가 불거지기 이전인 2016년 12월부터 INTEL의 SNS 상관관계는 하락 추세를 보여왔다(그림10 참조). 따라서 2018년 7월 단기 이벤트 발생으로 INTEL의 주가 상관관계가 하락하였음에도 Spread 값은 크지 않았다.

반대의 사례도 있다. 2020년 3월 UTC와 RAYTHEON의 합병으로 2020년 3월 RAYTHEON은 세계 최대의 방산업체로 등극하였다. 시장 점유율 상승으로 SNS 상관관계가 큰 폭으로 상승하였으며, 합병이라는 단기 이벤트 발생으로 주가 상관관계도 일부 하락하여 Spread가 큰 폭으로 확대되었다(그림11 참조).

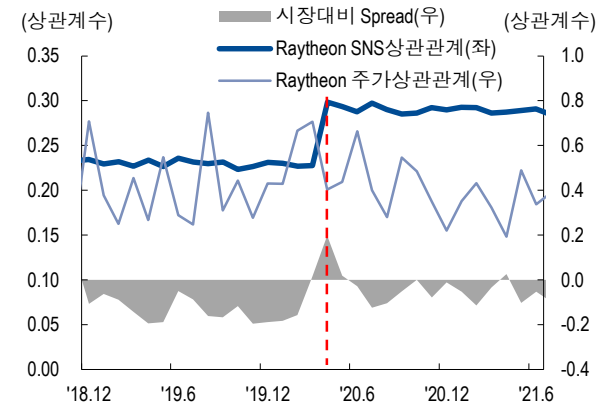
즉, 앞서 추정한 바와 같이 SNS 상관관계와 주가 상관관계 Spread가 이벤트 효과 측정에 유효함을 알 수 있다.

그림10. INTEL SNS 상관관계의 하락으로 주가 상관관계와는 무관하게 Spread 하락



자료: Refinitiv, NH투자증권 리서치본부

그림11. RAYTHEON의 SNS 상관관계 상승으로 Spread 확대



자료: Refinitiv, NH투자증권 리서치본부

3. 백테스팅 결과

이벤트 효과 투자 전략의 결과는 '10분위, Spread가중', '3분위, Spread가중'의 2가지 백테스팅 조건으로 확인하였다.

표1. 백테스팅 조건

백테스팅 조건	세부사항
종목군	S&P 100 구성종목
산출기간	2013년 12월 31일 ~ 2022년 5월 31일
수익률	배당, 기업분할 포함 총수익률
리밸런싱 주기	1개월
상관관계 산출 주가 시계열 기간	1개월
SNS상관관계 산출 텍스트 시계열 기간	12개월
수수료	0.015%
슬리피지	0.2%
분위별	- Spread 기준 10분위(내림차순) - Spread 기준 3분위(내림차순)
비중	Spread가중: 가장 Spread가 낮은 종목의 Spread를 0으로 조정한 뒤 Spread 가중 비중 배분

자료: NH투자증권 리서치본부

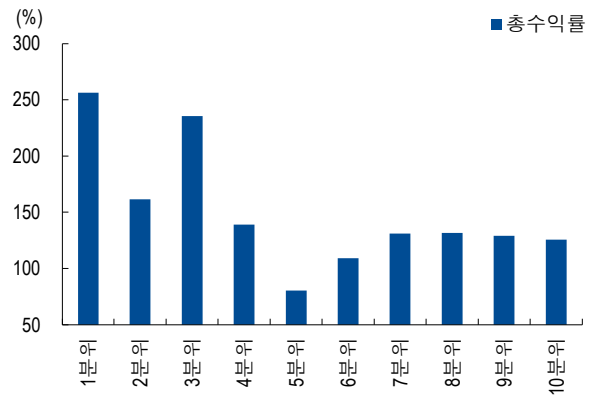
'10분위, Spread가중'으로 백테스팅을 진행한 결과, Spread(이벤트 효과)가 컸던 분위일수록 더 우수한 성과를 기록하였다. 변동성과 최대낙폭도 높은 분위일수록 낮은 수치를 기록하였다(표2, 그림12, 13, 14, 15, 다음 페이지 그림16 참조).

표2. Spread가중 10분위 성과표

분류	1분위	2분위	3분위	4분위	5분위	6분위	7분위	8분위	9분위	10분위
총수익률(%)	256.15	161.42	235.50	139.00	80.36	109.17	131.05	131.51	129.01	125.73
최대낙폭(%)	-24.45	-35.19	-21.96	-33.76	-46.99	-36.21	-39.52	-36.38	-33.17	-44.39
일간 샤프지수(연율화)	0.94	0.71	0.91	0.66	0.45	0.56	0.59	0.61	0.58	0.58
일간 소티노지수(연율화)	1.51	1.11	1.48	1.02	0.67	0.85	0.91	0.95	0.91	0.89
일간 변동성(연율화)	17.69	18.52	17.48	18.10	20.19	19.05	20.26	19.59	20.57	20.12

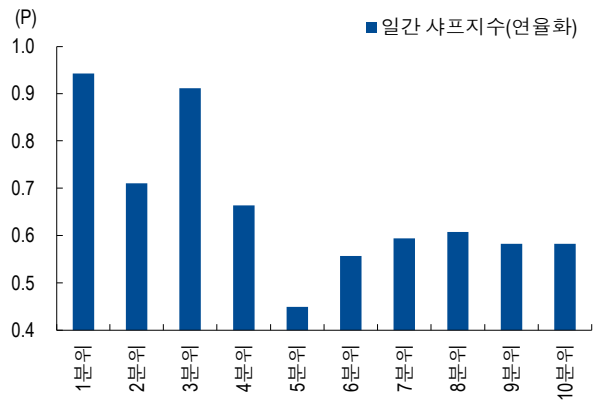
자료: S&P Clarifi, NH투자증권 리서치본부

그림12. Spread가중 분위별 총수익률



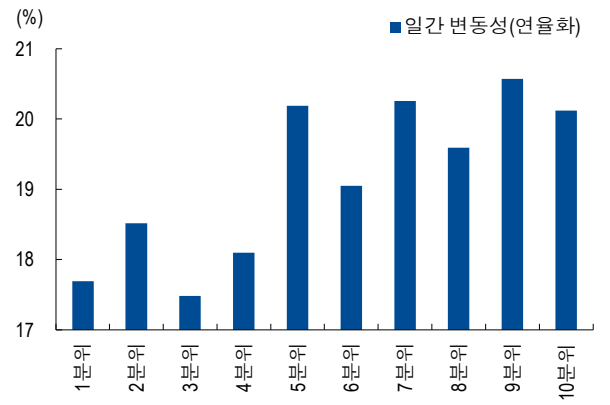
자료: S&P Clarifi, NH투자증권 리서치본부

그림13. Spread가중 분위별 샤프지수



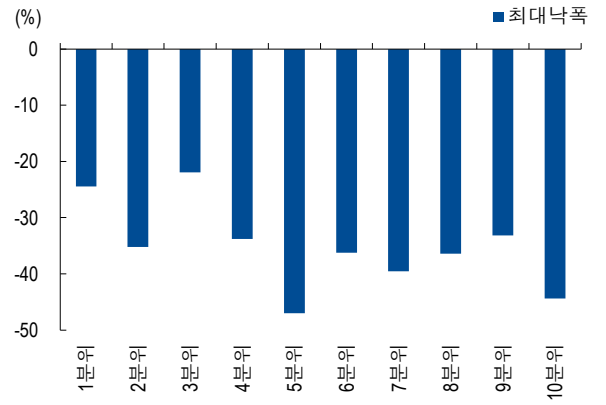
자료: S&P Clarifi, NH투자증권 리서치본부

그림14. Spread가중 분위별 변동성



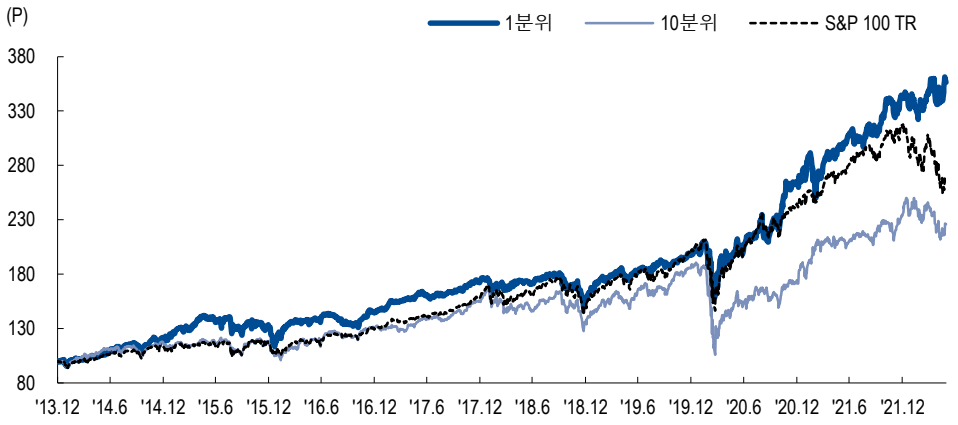
자료: S&P Clarifi, NH투자증권 리서치본부

그림15. Spread가중 분위별 최대낙폭



자료: S&P Clarifi, NH투자증권 리서치본부

그림16. (10분위, Spread가중) 백테스팅 결과 비교



자료: S&P Clarifi, NH투자증권 리서치본부

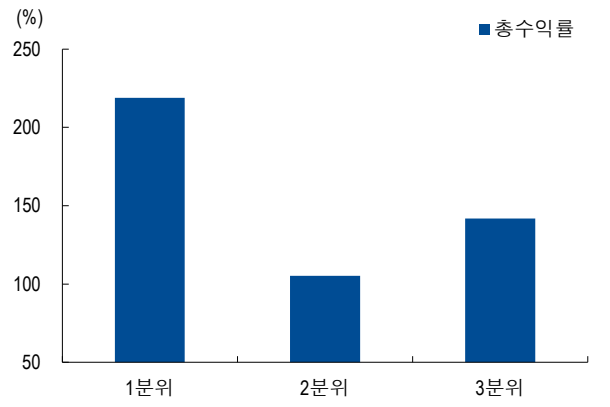
분위의 개수를 줄여 (3분위, Spread가중)으로 백테스팅을 진행한 결과 역시 마찬가지로 높은 분위가 더 우수한 성과를 기록하였다. 변동성과 최대낙폭도 높은 분위일수록 낮은 선형적 관계를 보이고 있다(표3, 그림17, 18, 19, 20, 다음 페이지 그림21 참조).

표3. Spread가중 3분위 성과표

분류	1분위	2분위	3분위
총수익률(%)	218.71	105.15	141.79
최대낙폭(%)	-27.33	-38.71	-37.43
일간 샤프지수(연율화)	0.92	0.57	0.64
일간 소티노지수(연율화)	1.44	0.85	0.98
일간 변동성(연율화)	16.50	17.99	19.45

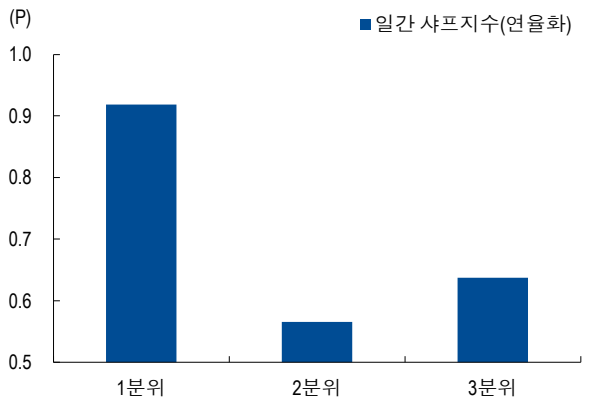
자료: S&P Clarifi, NH투자증권 리서치본부

그림17. Spread가중 분위별 총수익률



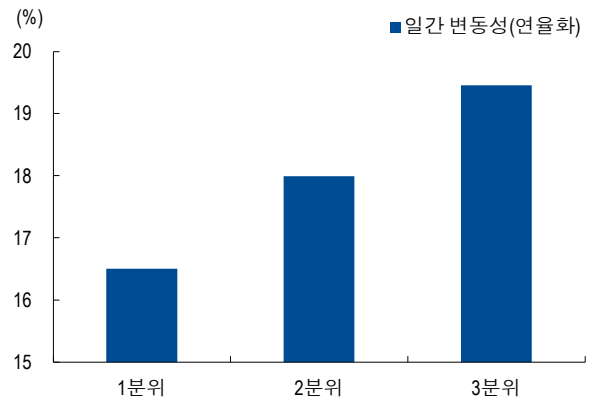
자료: S&P Clarifi, NH투자증권 리서치본부

그림18. Spread가중 분위별 샤프지수



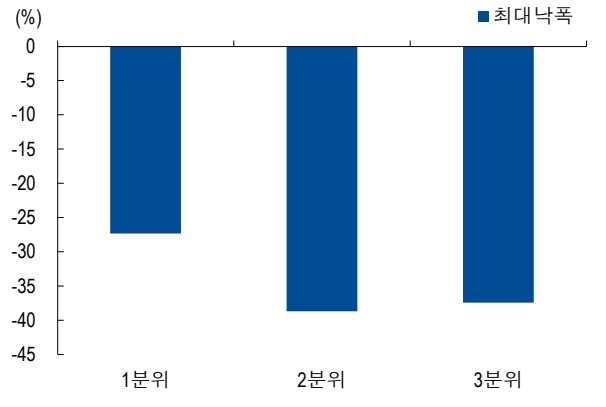
자료: S&P Clarifi, NH투자증권 리서치본부

그림19. Spread가중 분위별 변동성



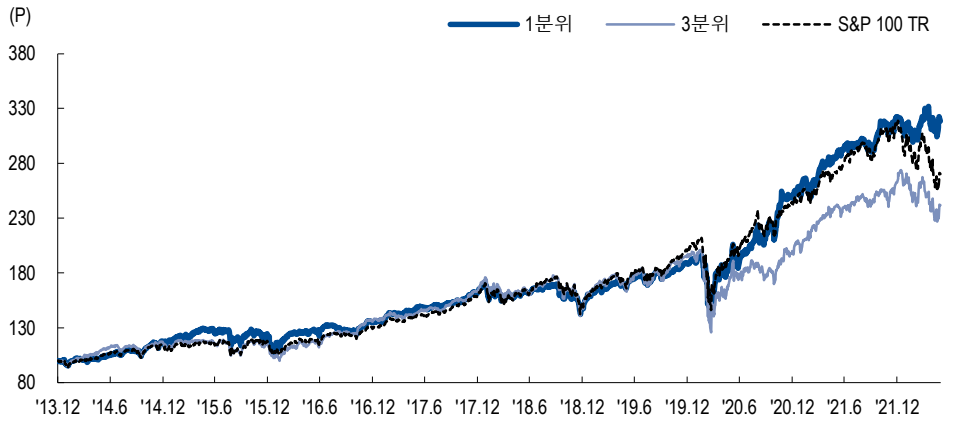
자료: S&P Clarifi, NH투자증권 리서치본부

그림20. Spread가중 분위별 최대낙폭



자료: S&P Clarifi, NH투자증권 리서치본부

그림21. (3분위, Spread가중) 백테스팅 결과 비교



자료: S&P Clarifi, NH투자증권 리서치본부

다른 주식 포트폴리오 최적화 방법론과 대비하여도 Spread를 활용한 종목 선택 효과가 우수하였다. 변동성은 리스크 패리티 수준으로 낮은 반면, 최대낙폭과 성과는 방법론 중 가장 양호하였다(표4 참조).

특히 위기 국면에서는 시가총액 가중 방법론 대비 적은 낙폭을 기록하며 하락장에서 방어 효과를 보였다(표5, 그림22 참조).

표4. 주식 포트폴리오 최적화 방법론 대비 이벤트 효과 투자 전략의 성과

분류	1분위 (총 3분위)	시가총액 가중	동일가중	계층적 리스크 패리티	리스크 패리티	평균분산
총수익률(%)	218.71	170.25	148.68	146.37	150.38	117.47
최대낙폭(%)	-27.33	-30.90	-34.62	-32.64	-33.52	-31.38
일간 샤프지수(연율화)	0.92	0.76	0.71	0.77	0.76	0.54
일간 소티노지수(연율화)	1.44	1.16	1.08	1.16	1.14	0.83
일간 변동성(연율화)	16.50	17.66	17.34	15.48	16.17	21.39

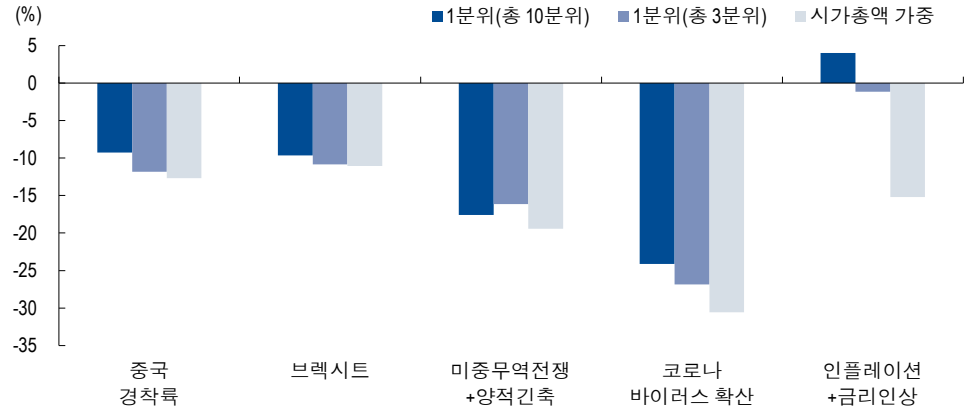
자료: Refinitiv, NH투자증권 리서치본부

표5. 위기 국면별 하락폭 비교

위기 국면	1분위(총 10분위)(좌)	1분위(총 3분위)	시가총액 가중
중국 경착륙 (2015년 7월~8월)	-9.3	-11.8	-12.7
브렉시트 (2015년 12월~2016년 2월)	-9.7	-10.8	-11.1
미중무역전쟁+양적긴축 (2018년 10월~12월)	-17.6	-16.1	-19.4
코로나 바이러스 확산(2020년 2월~3월)	-24.1	-26.8	-30.6
인플레이션+금리인상 (2022년 1월~5월)	4.0	-1.2	-15.2

자료: NH투자증권 리서치본부

그림22. 이벤트 효과 투자 전략의 성과가 시장 지수 대비 위기 국면에 안정적

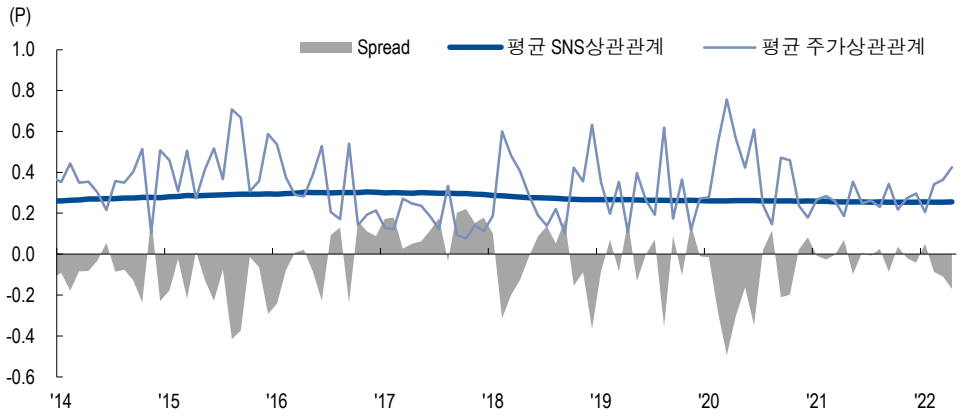


자료: S&P Clarifi, NH투자증권 리서치본부

4. Spread(이벤트 효과)를 활용한 종목 선택 효과

시계열로 보면, SNS 상관관계의 평균은 변동이 적은 반면, 주가 상관관계 평균은 변동이 크다(그림23 참조). 따라서 'SNS 상관관계 – 주가 상관관계'인 Spread는 주가 상관관계의 변동에 크게 영향을 받는다. 따라서 단순히 '– 주가 상관관계'만을 활용해도 어느 정도의 성과를 달성할 수 있다(표6 참조).

그림23. 이벤트 효과에 주가 상관관계가 미치는 영향이 큼



자료: Refinitiv, NH투자증권 리서치본부

표6. '– 주가 상관관계' 가중 10분위 성과표

분류	1분위	2분위	3분위	4분위	5분위	6분위	7분위	8분위	9분위	10분위
총수익률(%)	234.34	220.04	124.40	117.04	149.65	142.47	144.43	78.87	129.82	188.38
최대낙폭(%)	-25.91	-29.74	-41.70	-23.98	-37.11	-39.38	-36.53	-39.30	-36.15	-38.57
일간 샤프지수(연율화)	0.90	0.83	0.62	0.62	0.68	0.64	0.63	0.44	0.59	0.71
일간 소티노지수(연율화)	1.44	1.34	0.94	0.97	1.04	0.98	0.98	0.68	0.92	1.11
일간 변동성(연율화)	17.75	18.78	18.10	17.41	18.57	19.34	19.90	20.14	20.37	20.94

자료: S&P Clarifi, NH투자증권 리서치본부

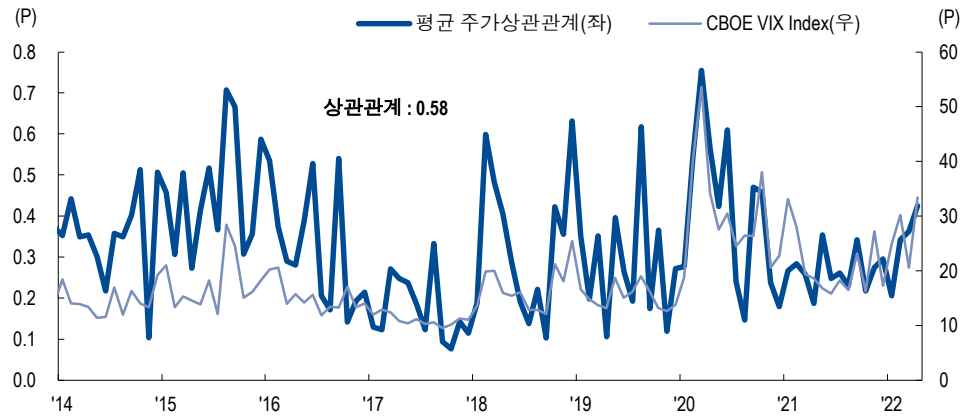
SNS 상관관계 Spread 활용 시 종목 선택 효과 증폭

음(-)의 주가 상관관계에 알파가 발생하는 것은 시장내 주가가 동반 하락하며 기업간 상관관계가 위기 국면에서 동기화되는 현상에서 기인한다. 실제로 VIX가 상승하는 시기에는 주가 상관관계가 동반 상승하여 주가 상관관계와 VIX 간에는 양(+)의 상관관계가 존재한다(그림24 참조).

즉 위기국면에서 시장과 음(-)의 주가 상관관계를 보이는 종목을 선택하는 것은 이벤트 효과로 인하여 하락장에서 주가가 상승하는 종목을 선택하는 것과 같다. 이러한 특성은 위기 국면에서 작은 변동성과 낙폭을 제공한다.

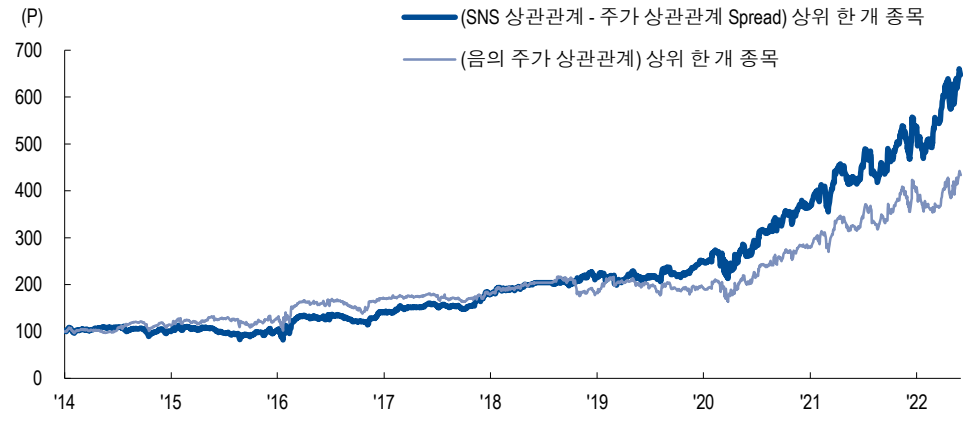
이 때, 단순히 주가 상관관계만 사용하면 정상국면에서도 시장과의 상관관계가 낮았던 종목도 선택되어 이벤트 효과 종목 변별력이 낮다. **'SNS 상관관계 – 주가 상관관계' Spread 활용 시 원래는 시장과의 상관관계가 높았으나, 이벤트 발생으로 인하여 시장과의 상관관계가 낮아진 종목을 선택**할 수 있다. 백테스팅에서도 Spread가 가장 높은 1개 종목을 선택하는 경우의 성과가 음(-)의 주가 상관관계가 가장 높은 1개 종목을 선택하는 경우보다 더 우수했다(그림25 참조).

그림24. 주가 상관관계와 VIX 지수의 상관관계가 높음



자료: Refinitiv, NH투자증권 리서치본부

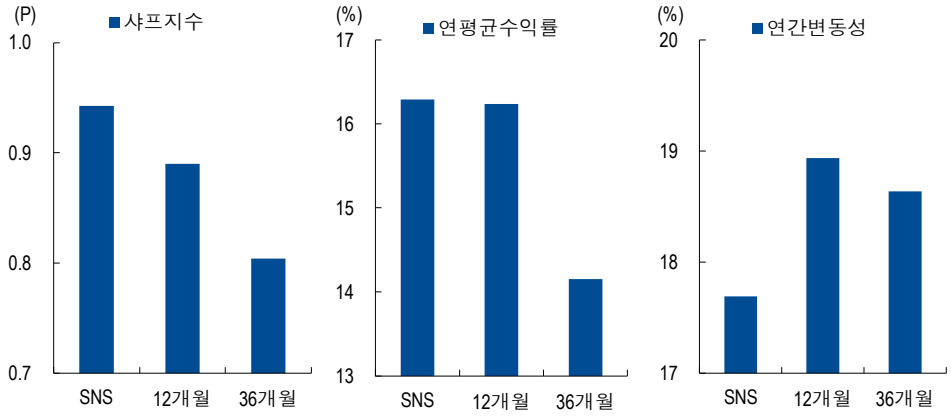
그림25. SNS 상관관계 Spread 활용 시 종목 선택 효과 증폭



자료: S&P Clarifi, NH투자증권 리서치본부

기업의 장기 주가 상관관계로 SNS 상관관계를 대체하여도 양호한 성과를 얻을 수 있다. 그러나 SNS 상관관계 Spread 활용 시와 비교하면 리스크 대비 성과 측면에서 저조한 모습을 보인다(그림26 참조).

그림26. SNS 상관관계와 장기 주가 상관관계 Spread 활용 시, 1분위 포트폴리오 성과 비교



자료: S&P Clarifi, NH투자증권 리서치본부

5. 긍정적 이벤트와 부정적 이벤트 효과별 성과 기여

이전달 수익률이 (+)이면 긍정이벤트, (-)이면 부정이벤트로 분류할 경우, 이벤트별 Spread 전략 수익률은 아래와 같다(표7 참조).

긍정이벤트와 부정이벤트 모두 1분위의 성과는 벤치마크를 상회하였다. 특히, 부정이벤트가 긍정이벤트 대비 더욱 강한 모멘텀을 제공하였다(그림29 참조). 그러나 부정이벤트는 1분위를 제외하면 분위별 최대낙폭과 성과 격차가 커 투자의 안정성이 떨어졌다.

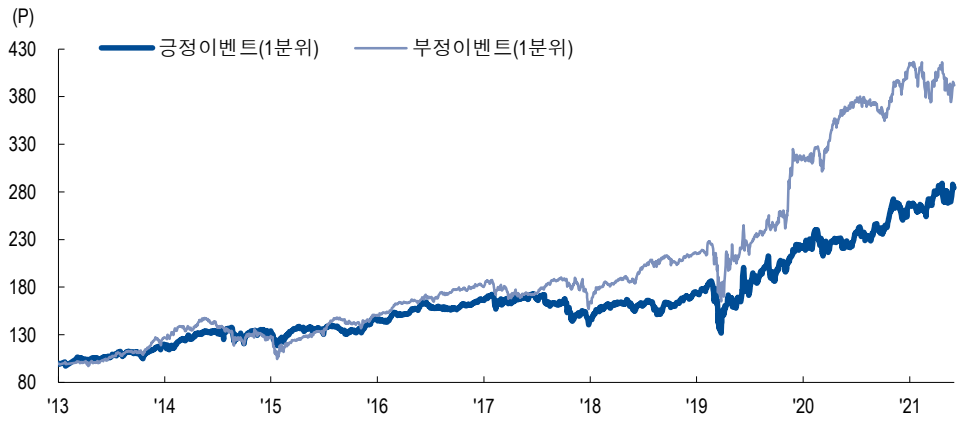
따라서 긍정이벤트 관련 주가 모멘텀 전략의 안정성과 부정이벤트 관련 주가 반등 전략의 성과를 혼합하는 방식으로 리스크 대비 수익을 극대화할 수 있을 것으로 판단한다.

표7. 이벤트별 Spread전략 성과표

분류	긍정이벤트					부정이벤트					S&P 100TR
	1 분위	2 분위	3 분위	4 분위	5 분위	1 분위	2 분위	3 분위	4 분위	5 분위	
연평균수익률(%)	13.2	12.6	9.7	10.8	11.1	17.6	9.9	10.9	6.9	14.4	12.5
최대낙폭(%)	-29.4	-28.1	-39.5	-33.8	-33.2	-29.2	-31.5	-40.8	-40.8	-39.5	-30.9
일간 샤프지수(연율화)	0.75	0.74	0.58	0.62	0.62	0.95	0.59	0.61	0.42	0.76	0.76
일간 소티노지수(연율화)	1.17	1.15	0.90	0.96	0.96	1.56	0.95	0.93	0.65	1.19	1.16
일간 변동성(%, 연율화)	18.8	18.2	19.1	19.8	20.5	19.1	19.0	20.5	20.8	20.6	17.7
상승확률(%, 연간)	88.9	66.7	66.7	66.7	66.7	77.8	66.7	66.7	66.7	77.8	77.8

자료: S&P Clarifi, NH투자증권 리서치본부

그림29. 성과 측면에서는 부정이벤트의 반등 효과가 더욱 강한 모멘텀 제공



자료: S&P Clarifi, NH투자증권 리서치본부

6. 8월 이벤트 효과 상위 10개 종목

표8. 2022년 7월 29일 기준 이벤트 효과 상위 10개 종목

기업명	티커	SNS 상관관계	주가 상관관계	이벤트 효과(Spread)	포트폴리오 비중(%)
Kraft Heinz	KHC	0.223	-0.160	0.382	13.13
IBM	IBM	0.294	-0.061	0.354	12.61
Comcast	CMCSA	0.295	0.084	0.211	9.98
Merck	MRK	0.285	0.083	0.201	9.80
Procter & Gamble	PG	0.239	0.050	0.189	9.57
Charter Communications	CHTR	0.218	0.040	0.177	9.36
Duke Energy	DUK	0.294	0.130	0.165	9.12
Qualcomm	QCOM	0.276	0.126	0.150	8.86
3M	MMM	0.221	0.074	0.148	8.81
Johnson & Johnson	JNJ	0.280	0.135	0.145	8.76

주: 상기 종목은 특정 지표 관점에서 추출한 종목으로써 추천종목이 아님
자료: NH투자증권 리서치본부

IV. Appendix

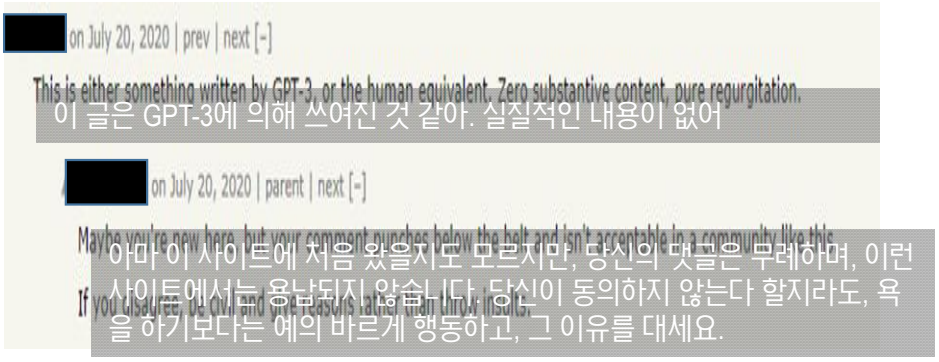
Appendix A. Word2Vec

인간이 구사하는 언어(Natural Language)를 컴퓨터가 분석 및 처리하는 것을 자연어처리(NLP: Natural Language Processing)이라 하며, 자연어처리를 수행하는 모델을 언어 모델이라 통칭한다.

현재의 언어 모델은 Transformer 구조를 바탕으로 하는 BERT(Bidirectional Encoder Representations from Transformers) 언어 모델이 등장한 이후 인간이 작성한 텍스트와 구분하기 어려울 정도로 자연스러운 문맥의 텍스트를 작성하는 등 엄청난 발전을 보이고 있다. MIT의 Liam Porr가 블로그에 게시한 GPT-3 언어 모델의 작성글 'Feeling unproductive? Maybe you should stop overthinking.'이 다수의 네티즌을 속인 것은 언어 모델의 성능을 보여주는 아주 흥미로운 일화 중 하나이다(그림A1 참조).

본문에서는 단어별 고유(Unique) 벡터를 산출하고, 벡터 간 유사도 측정이 용이한 Word2Vec 방법론을 사용하였다.

그림A1. GPT-3 언어 모델이 작성한 글에 속아버린 네티즌



자료: Hacker News, NH투자증권 리서치본부

언어 모델은 문맥에 부합하는 단어를 통계적으로 예측할 수 있다는 주장에 기반한다. 예를 들어 '나는 __에 간다'라는 문장에서 __에 들어갈 단어를 예측한다고 하자. __는 장소에 해당하므로 놀이터, 학교, 집 등의 단어가 들어갈 수 있다. 만약 문장이 '나는 공부하러 __에 간다'라면, 앞 단어가 '공부하러'이므로 학교, 학원 등의 단어가 들어갈 확률이 높아진다. 단어가 더 추가되어 문장이 '나는 방학에는 공부하러 __에 간다'가 된다면 __에는 '학원'이 들어갈 가능성이 매우 높아진다. Word2Vec이라는 언어 모델은 입력된 문장을 참조하여 이와 같은 확률을 극대화시키는 방식(Continuous Bag of Words)으로 학습을 진행한다.

이 과정을 반대로 보면, '학원'이라는 단어가 있을 때 앞에 들어갈 단어는 '공부하러', 뒤에 들어갈 단어는 '간다'일 가능성이 높다. 이처럼 앞, 뒤에 나올 단어의 확률을 입력된 단어에 맞추어 극대화시키는 학습 방법(Skip-Gram)도 있다.

이 때 단어는 임베딩(Embedding)이라는 형식으로 표현되고 학습된다. 임베딩이란 숫자로 구성된 1차원 행렬(벡터)로, 단어별 고유 번호를 부여하는 방식 대비 유사도 확률을 계산하는 데 있어 유리하다.

그림A2. 임베딩 값을 기준으로 하면 세밀한 유사도 계산에 유리

단어	고유번호
고양이	1
개	2
호랑이	3
쥐	4

단어	임베딩			
고양이	1	1	1	0
강아지	1	0	1	1
호랑이	0	1	1	0
쥐	1	0	0	0

임베딩 값 기준 코사인 유사도

고양이 – 강아지 유사도 : 66%
고양이 – 호랑이 유사도 : 82%
고양이 – 쥐 유사도 : 58%

자료: NH투자증권 리서치본부

그림A3. 코사인 유사도 계산식

$$Cosine\ Similarity = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{N} A_i * B_i}{\sqrt{\sum_{i=1}^{N} {A_i}^2} * \sqrt{\sum_{i=1}^{N} {B_i}^2}}$$

자료: NH투자증권 리서치본부

표A1. 본문 Word2Vec 모델 변수

백테스팅 조건	세부사항
학습방식	Skip-Gram
참조 단어 최대거리	8
임베딩 차원	300
학습 횟수	50
네거티브 샘플링	20
초기 학습률	0.03
최소 학습률	7E-4

주: 모델 변수 설정은 [Adewumi, T. P., Liwicki, F., & Liwicki, M. (2020). Word2vec: Optimal hyper-parameters and their impact on nlp downstream tasks]를 참조하였으며, 서울대 데이터마이닝센터와 설정

자료: NH투자증권 리서치본부

NH투자증권 AI/빅데이터 전략 발간 목록

작성일	종류	작성자	제목
2021-10-19	AI/빅데이터 전략	김규진	AI 와 대체데이터를 활용한 KOSPI 예측 모델
2022-03-15	AI/빅데이터 전략	김규진	AI 를 활용한 공모주 투자전략

Compliance Notice

- 당사는 자료작성일 현재 동자료에 언급된 종목의 발행주식 등을 1% 이상 보유하고 있지 않습니다.
- 당사는 동 자료를 기관투자자 또는 제 3 자에게 사전 제공한 사실이 없습니다
- 동 자료의 금융투자분석사와 배우자는 자료작성일 현재 동 자료상에 언급된 기업들의 금융투자 상품 및 권리를 보유하고 있지 않습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.

고지 사항

본 조사분석자료에 수록된 내용은 당사 리서치본부의 금융투자분석사가 신뢰할 만한 자료 및 정보를 바탕으로 최선을 다해 분석한 결과이나 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 투자자의 투자판단을 위해 작성된 것이며 어떠한 경우에도 주식 등 금융투자상품 투자의 결과에 대한 법적 책임소재를 판단하기 위한 증빙자료로 사용될 수 없습니다. 본 조사분석자료는 당사의 저작물로서 모든 지적 재산권은 당사에 귀속되며 당사의 동의 없이 복제, 배포, 전송, 변형, 대여할 수 없습니다.