뉴스 텍스트 감성 분석을 통한 주가 변동성 상관관계 연구

3-C 202344086 이민서

목차

목적

1번 프로젝트는 단순 검색량과 주가를 비교했다면
이번의 프로젝트는 뉴스 헤드라인의 긍정/부정적 어조(Sentiment)가 실제 주가 변동과 유의미한 상관관계가 있는지를 데이터 기반으로 검증하는 것을 목적으로 하였습니다

연구 범위

대상 종목: NVIDIA (Ticker: NVDA)

분석 기간: 2024년 1월 1일 ~ 2024년 12월 31일 (1년간)

데이터 출처: Google News (뉴스 데이터),
Yahoo Finance (주가 데이터)

데이터 수집 및 전처리 요약

Python의 feedparser와 newspaper3k 라이브러리를 활용하여 뉴스 데이터를 수집하였습니다.

수집 채널: Google News RSS

검색 키워드: "NVIDIA" + (stock, earnings, AI, revenue 등 주요 키워드 조합)

수집 결과: 2024년 1년 치 총 1,193건의 유효 뉴스 기사 제목 수집 완료

데이터 수집

데이터 수집 전략

이번 프로젝트는 2024년 1월부터 12월까지의 데이터를 필요로 합니다.

그러나 일반적인 뉴스 검색 API나 RSS 피드는 최신 기사 위주로 제공되거나, 한 번의 요청에 제공하는 데이터 개수(약 100건)에 제한이 있다는 한계가 있었습니다.

데이터 수집

이를 해결하기 위해 **시계열 분할 수집** 방식을 고안하여 적용하였습니다.

단순 검색 시 2024년 전체 데이터 중 최신(12월) 데이터만 편중되어 수집되었습니다.

그래서 Python의 반복문(Loop)을 활용하여 수집 구간을 1개월 단위(Month-by-Month)로 쪼개어 총 12회 분할 요청을 수행하여 문제를 해결하였습니다.

데이터 수집

검색 쿼리에 after:2024-01-01 before:2024-01-31과 같은 날짜 지정 연산자를 동적으로 삽입하였습니다.

calendar 라이브러리를 활용하여 각 월의 마지막 날짜(29일, 30일, 31일)를 자동으로 계산하여 쿼리의 정확도를 높였습니다.

이 결과로 특정 시기에 데이터가 쏠리는 현상 없이, 1월부터 12월까지 매월 고르게 분포된(월 100건의 뉴스) 총 1,193건의 균형 잡힌 뉴스 데이터셋을 확보하였습니다.

데이터 수집

서버 차단(HTTP 429 Error) 회피

짧은 시간에 다량의 요청을 보낼 경우 구글 서버로부터 'Too Many Requests' 에러가 반환되며 수집이 차단되는 문제가 있었습니다.

그래서 찾은 해결책은 time.sleep() 함수를 사용하여 요청 간에 0.5초~2.0초의 랜덤한 지연 시간을 부여하면 사람의 검색 행위처럼 보이도록 트래픽 속도를 조절이 가능하다는 것을 알게 되었습니다. 추가로 구글링 해본 결과 HTTP 요청 헤더에 User-Agent 정보를 브라우저 환경으로 위장하여 차단을 우회할 수 있다는 것을 알게 되었습니다.

데이터 전처리(날짜 데이터 표준화)

수집된 비정형 텍스트 데이터를 시계열 분석이 가능한 형태로 가공하였습니다.

날짜 데이터 표준화

RSS 피드에서 제공하는 날짜 형식(예: Fri, 14 Dec 2024 10:00:00 GMT)은 텍스트 형태이므로 연산이 불가능하였습니다. 그래서 Python의 datetime 객체로 변환한 뒤, 주가 데이터와의 결합을 위해 시간대 정보를 제거하고 YYYY-MM-DD 형식으로 통일하여 해결하였습니다.

데이터 전처리(노이즈 제거 및 필터링)

중복 제거

여러 언론사에서 동일한 AP/Reuters 통신사의 기사를 전재하는 경우, 제목이 100% 일치하는 데이터는 drop_duplicates를 통해 제거하여 분석 왜곡을 방지하였습니다.

결측치 처리

newspaper3k 라이브러리로 본문 추출이 실패한 경우, 사실상.. 가장 중요한 기사 제목만으로도 감성 분석이 가능하다고 판단하여 해당 데이터를 유지하였습니다.

데이터 전처리(감성 사전 구축 및 점수화)

주식 시장에 특화된 도메인 사전을 직접 정의하여 사용해보았습니다.

사전 정의

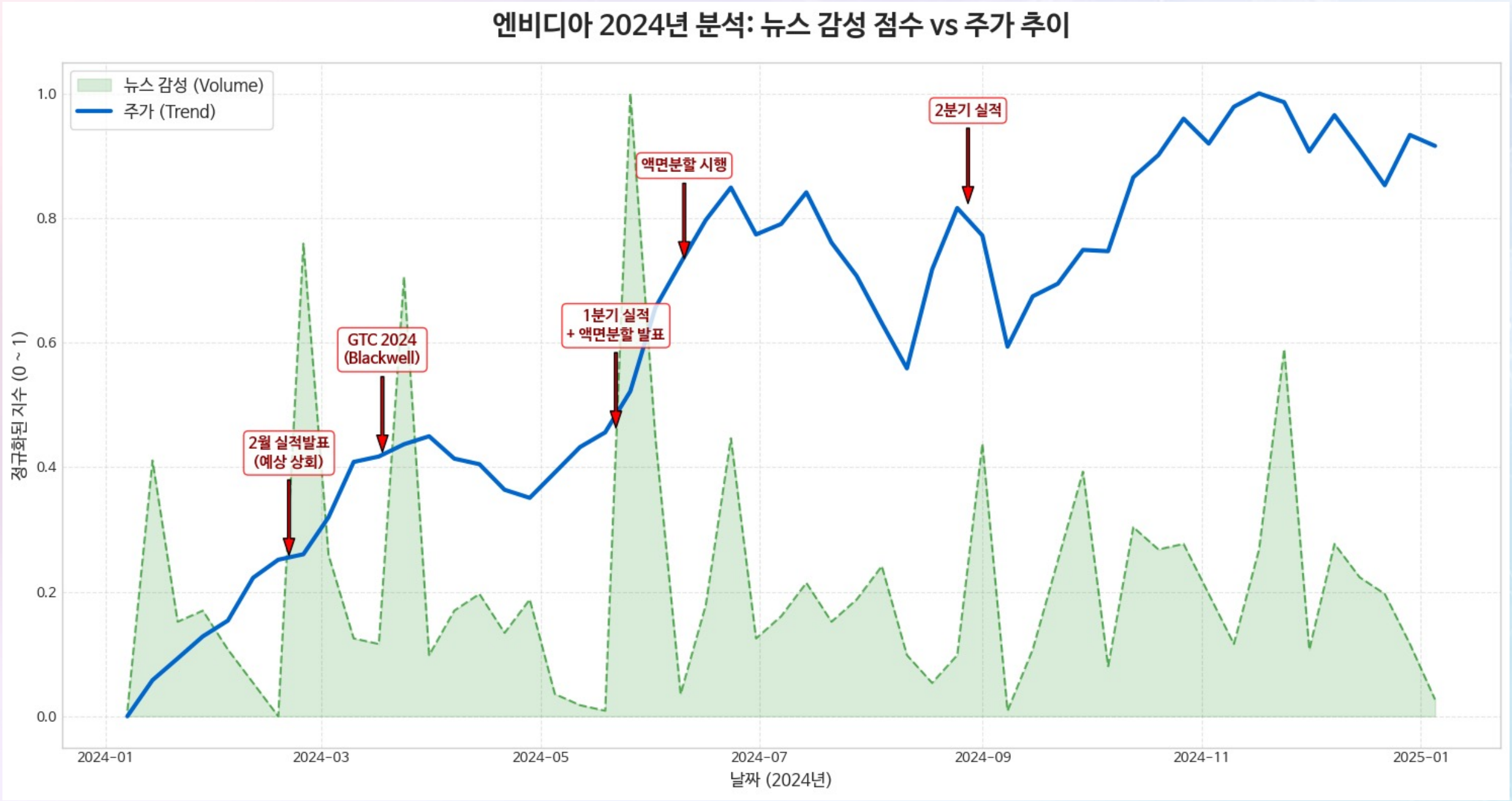
Positive(+): Surge(급등), Soar(치솟다), Beat(상회하다), AI, Growth 등 30개 핵심어

Negative(-): Plunge(급락), Miss(하회하다), Ban(규제), Delay(지연), Risk 등 28개 핵심어

데이터 전처리(감성 사전 구축 및 점수화)

점수 산출
기사 제목에 긍정 단어가 포함되면 +1점,
부정 단어가 포함되면 -1점을 부여하고,
이를 날짜별로 합산하여 일일 뉴스 감성 지수를 산출하였습니다.

주간 집계
일별 데이터의 불규칙성(Noise)을 줄이기 위해 데이터를
주간 단위로 재집계하여 주가 흐름과의 장기적 추세를 비교하였습니다.


엔비디아 2024년 분석: 뉴스 감성 점수 vs 주가 추이
뉴스 감성 (Volume)
주가 (Trend)
2월 실적발표
(예상 상회)
GTC 2024
(Blackwell)
1분기 실적
+ 액면분할 발표
액면분할 시행
2분기 실적
정규화된 지수 (0 ~ 1)
1.0
0.8
0.6
0.4
0.2
0.0
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
날짜 (2024년)

분석 결과(상관관계 분석)

데이터 전처리 및 주간 리샘플링을 거쳐 뉴스 감성 지수와
주가 간의 피어슨 상관계수를 산출한 결과는 아래와 같습니다.

주간 상관계수: 0.066

기존의 단순 일별 분석과 달리, 주간 단위로 노이즈를 제거한 결과
0.066의 양의 상관계수가 도출되었습니다.

분석 결과(상관관계 분석)

상관계수가 낮게 나온 이유는 두 가지로 해석됩니다.

첫째는 **매크로 변수의 지배력**입니다. 2024년은 개별 기업의 뉴스보다 '미국 금리 향방'이나 'AI 산업 전체의 수급 쏠림'과 같은 거시적 요인이 엔비디아 주가에 더 강력한 영향력을 준 것 같습니다.

둘째는 **재료의 소멸**입니다. 뉴스가 나왔을 땐 이미 주가가 다 오른 뒤인 경우가 많았습니다.
즉, 뉴스는 주가를 움직이는 '원인'이기도 하지만,
이미 반영된 주가를 설명하는 결과적 성격도 강하기 때문에
수치적 상관관계가 낮게 측정된 것입니다.

분석 결과(시각화 및 구간별 상세 분석)

전체 기간의 선형적 상관관계는 낮았으나(0.066), 시각화를 통해 대형 이벤트 발생 구간에서는 강한 동조화가 나타남을 확인하였습니다.

상승 동조화 구간 (2월~3월): 엔비디아의 GTC 2024 컨퍼런스(Blackwell 발표)를 전후로 감성 점수가 급등하였고, 주가 역시 $800선을 돌파하며 강한 양의 상관관계를 보였습니다.

최대 호재 반영 (5월~6월): '1분기 어닝 서프라이즈'와 '10:1 액면분할' 발표가 겹친 5월 말, 뉴스 감성 지수가 연중 최고치를 기록하였으며 주가 또한 신고가를 경신하며 완벽하게 동행하였습니다.

괴리 발생 구간: 8월 초 '블랙웰 칩 설계 결함 루머' 등으로 감성 점수가 급락했으나, 저가 매수세 유입으로 주가는 반등하는 등 뉴스와 주가가 엇갈리는 구간도 관찰되었습니다.
이는 낮은 상관계수(0.066)의 주요 원인이 되었습니다.

최종 결론

1. 본 연구 결과, 전체 기간의 뉴스 감성과 주가는 **0.066의 낮은 상관계수**를 보여 키워드 뉴스만으로 주가를 예측하는 데는 한계가 있음을 확인하였습니다.
2. 그러나 단순 통계 수치와 달리, 실적 발표나 신제품 공개 등 **‘핵심 이벤트(Mega Event)’ 구간에서는 뉴스와 주가가 강력하게 동조화**되는 현상이 시각적으로 입증되었습니다.
3. 결론적으로 뉴스 데이터는 상시 예측 도구보다는, **특정 이벤트 전후의 시장 기대 심리와 변동성을 포착하는 핵심 보조 지표**로 활용할 때 가장 유효함을 확인하였습니다.