

트럼프 말 한마디에 출렁이는 시장

: 연설문 감성과
주가 변동성의 비밀

: 떡잎방정대 (김진욱 김정인 신채원 이다현)



목차

Table of Contents

연구 주제 및 문제의식	03p
Topic & Motivation	
연구 질문	05p
Research Questions	
분석 전략 및 흐름	06p
Analytical Strategy & Flow	
분석 Framework	08p
Sequential Analytical Framework with Six Key Steps	
결과, 시사점 및 결론	36p
Results, Implications & Conclusion	

1. Topic



주제

트럼프 트윗 ➤ 시장을 움직인다는 인식

? 분석 결과,
트윗을 올렸다는 사실 자체 → 시장 반응 변동 X

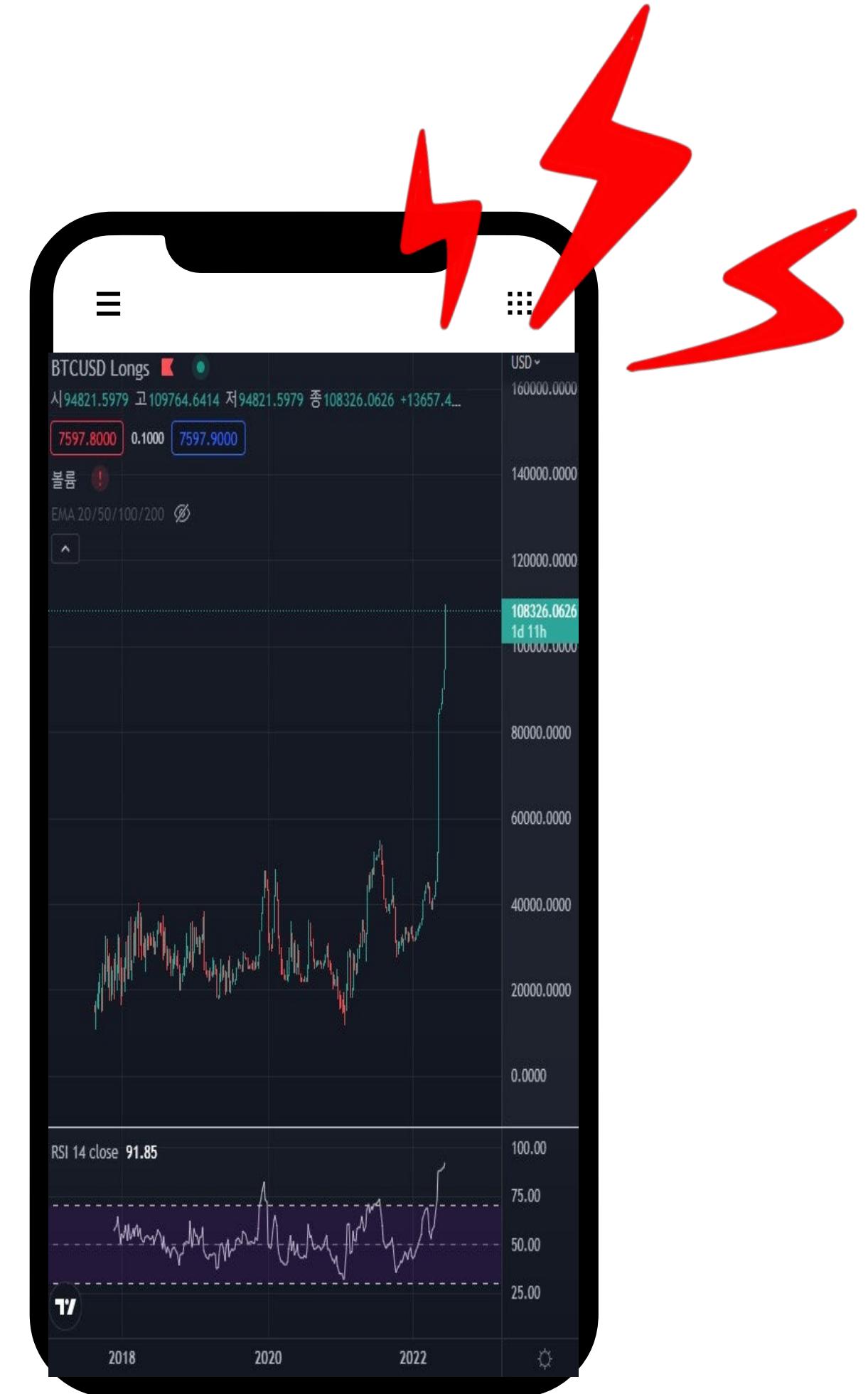
“그렇다면 어떤 트윗이, 언제 의미를 가지는가?”

단순한 트윗 여부를 넘어서
트윗의 주제, 핵심 단어, 감성(긍·부정)이
시장 반응을 설명할 가능성에 주목

본 연구는 트럼프 트윗의 내용과 맥락에 따라
시장 반응이 어떻게 달라지는지를 단계적으로 분석하고자 함

1. Topic

주제



2. Research Questions

연구 질문

트럼프 트윗 게시 여부 자체가 시장에 영향을 미치는가

→ 트윗 유무에 따른 단기 시장 반응 존재 여부 검증

어떤 주제의 트윗이 시장 반응을 유발하는가

→ 트윗 내용을 주제 단위로 분류하여 반응 차이 분석

특정 단어 또는 감성(긍·부정)은 시장 반응에 영향을 미치는가

→ 주제 내 핵심 단어와 감성 효과의 추가적 영향 확인

시기별로 트럼프 트윗의 역할은 달라지는가

→ 정책 이슈 강도에 따른 트윗 효과의 변화 분석



3. Overall Research Flow

연구 흐름

Step 1. 트윗 자체 효과 분석
트윗 게시 여부에 따른 시장 반응 확인

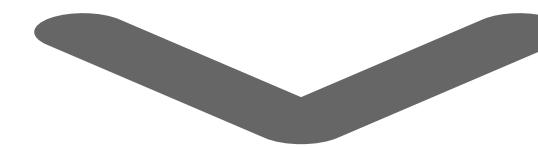


Step 2. 주제 분석 (Topic Modeling-LDA)
트윗을 주제 단위로 분류
시장 반응을 유발하는 주제 탐색

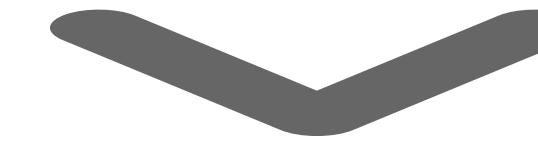


Step 3. 무역전쟁 주제 집중 분석
Trade War 주제의 시장 반응 검증

Step 4. 단어 수준 검증
Tariff 등 핵심 단어의 추가적 영향 확인



Step 5. 감성 분석
트윗의 긍·부정 감성에 따른 시장 반응 비교



Step 6. 시기별 비교 분석
정책 이슈 강도에 따른 트윗 효과 변화 확인



Step 7. 종합 해석
트럼프 트윗의 시장 반응 메커니즘 정리

4.Data

사용한 데이터 소스

1.Trump Twitter Archive

Data Source

- The Trump Twitter Archive
(<https://www.thetrumparchive.com>)

Data Characteristics

- 도널드 트럼프 원문 트윗 텍스트 데이터
- 트윗 단위의 시계열 이벤트 데이터
- UTC 타임스탬프 포함, 리트윗 제외

Data Collection

- 웹사이트에서 데이터 수집
- Selenium을 이용한 동적 페이지 로딩
- BeautifulSoup을 통한 파싱 및 JSON 추출
- 트윗 ID 기준 중복 제거

Variables

- 작성 시각 (datetime_utc), 트윗 ID (id), 리트윗 본문 (text), 리트윗 수 (retweets)

2.S&P 500 Minute Data

Data Source

- FX-1-Minute-Data (GitHub)
(<https://github.com/philipperemy/FX-1-Minute-Data>)

Data Characteristics

- S&P 500 1분 단위(분봉) 가격 데이터
- 고해상도 금융 시계열 데이터
- 단기 시장 반응 분석에 적합

Rationale

- 일별 데이터는 다양한 정보가 누적되어 개별 트윗 효과 식별이 어려움
- → 분봉 데이터는 트윗 직후의 즉각적인 시장 반응 포착 가능

Usage

- 트윗 게시 시점과 시간 기준 정밀 매칭
- 트윗 직후 수익률 및 변동성 변화 분석에 활용

5. 분석 Framework

1 데이터 수집 및 기준선 설정

→ 트윗 분석의 출발점

- 트럼프 트윗 데이터 수집 및 정제

- 트럼프 공식 트위터 계정 트윗 수집
- 텍스트 전처리 및 불필요한 트윗 제거
- 분석 가능한 형태로 데이터 정리

- S&P500 시장 데이터 병합

- S&P500 지수 1분봉 시장 데이터 수집
- 트윗 게시 시점 기준으로 시장 데이터 정렬
- 트윗 시점 전후 수익률 계산

- 트윗 자체 효과에 대한 기준선 분석

- 트윗 존재 여부에 따른 시장 반응 비교
- 트윗 내용과 무관한 평균 효과 검증

2 트윗 내용 기반 주제 분석

→ 트윗 '내용' 차이가 시장 반응에 미치는 영향

- 트윗 텍스트 전처리

- 불필요한 기호, 링크, 특수문자 제거
- 불용어 제거 및 단어 형태 정규화
- 주제 분석을 위한 텍스트 정리

- 트윗 주제 분류

- LDA 기반 주제 모델링 적용
- 트윗을 여러 개의 의미 단위 주제로 분류
- 각 트윗의 주제별 비중 산출

- 주제별 시장 반응 비교

- 주제 비중을 설명변수로 활용
- 주제별 트윗 등장 시점 이후 시장 반응 비교
- 시장 반응이 두드러지는 주제 식별



5. 분석 Framework

3 무역전쟁 주제 심층 분석

→ 시장 반응이 나타난 특정 주제 집중

- 무역주제 관련 주제 식별

- 주제 분석 결과 중 무역전쟁 주제 선택
- 정책 불확실성 확대 가능성 확인

- 무역전쟁 주제의 시장 반응 분석

- 무역전쟁 트윗 등장 시점 기준 분석
- 시장 변동성 및 수익률 반응 확인

4 핵심 단어 수준 검증

→ 주제 분석 결과의 타당성 확인

- 핵심 단어 선정

- 무역전쟁 주제 내 주요 단어 식별
- 정책 수단을 의미하는 단어 설정

- 단어별 시장 반응 분석

- 관세 관련 단어 등장 시점 분석
- 일반 정치 표현과 반응 비교

- 주제 분석 결과 타당성 검증

- 단어 수준 결과와 주제 분석 결과 비교
- 주제 효과의 실제 원인 확인



5분석 Framework

5 트윗 감성 분석

→ 정책 이슈가 약한 시기의 보조 분석

- 트윗 감성 점수 산출

- 트윗별 긍,부정 감성 점수 계산
- 일별 평균 감성 지표 계산

- 감성과 시장 반응 관계 분석

- 감성 점수와 수익률 간 관계 분석
- 단기 시장 반응 확인

6 시기별 비교 및 종합 해석

→ 트윗 효과의 맥락 의존성 검증

- 시기별 데이터 분리

- 무역전쟁 이전 시기 분리
- 정치.정책 노이즈 차이 고려

- 2017년 집중 분석

- 2017년 데이터 단독 분석
- 트윗 감성 효과 상대적 변화 확인

- 종합 해석 및 시사점 도출

- 트윗 효과의 시기별 차이 정리
- 트레이딩 신호 vs 리스크 신호 구분
- 트윗의 시장 내 역할 재정의



**베이스라인 확인**

트윗 '내용'이 아닌 트윗 '발생 자체'가 시장을 움직이는지 먼저 파악

- 트윗 있다 vs 없다 평균 수익률 차이 검증
- 이후 LDA 토픽 분석의 기준점 마련

벤치마크 설정

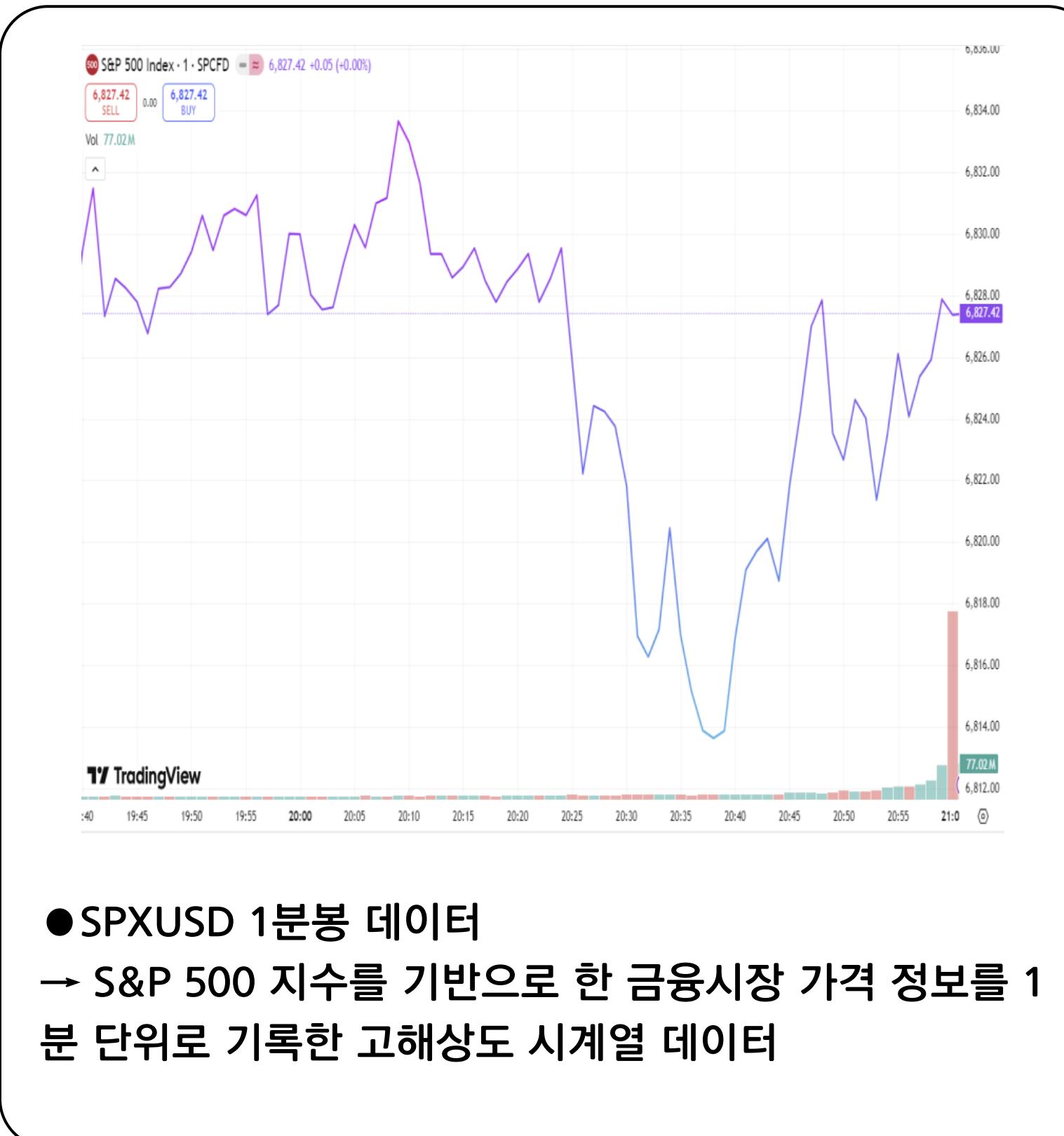
향후 토픽/단어 이벤트 분석을 위한 기준선 구축

- 효과 없음 : 특정 단어/토픽에서만 반응 가능성
- 효과 있음 : 어떤 내용이 더 강한 영향을 주는지 분석

데이터 수집 및 기준선 설정

사용 데이터

1. 가격데이터 : SPXUSD 1분봉



	datetime	open	high	low	close	volume
0	2016-11-07 00:02:00	2105.5	2105.75	2105.5	2105.75	0
1	2016-11-07 00:03:00	2105.75	2106	2105.75	2106	0
2	2016-11-07 00:04:00	2106.25	2106.25	2106	2106	0
3	2016-11-07 00:05:00	2105.75	2106	2105.75	2105.75	0
...						

- 데이터 출처
 - GitHub : FX-1-Minute-Data
 - 공개된 금융시장 1분 단위 시계열 데이터 저장소
- 데이터 성격
 - SPXUSD 1분봉 데이터
 - S&P 500지수를 기반으로 한 가격 정보를 분 단위로 기록한 고해상도 시계열 데이터
 - Open,High,Low,Close 가격 정보 포함
- 활용목적
 - 트럼프 트윗과 같은 단기 이벤트 발생 직후 시장의 즉각적인 반응을 포착하기 위해 사용
 - 일별 데이터로는 포착하기 어려운 단기 변동성 및 방향성 분석에 활용

1 데이터 수집 및 기준선 설정

| 사용 데이터

2. Trump Twitter Archive

The screenshot shows the Trump Twitter Archive V2 interface. At the top, there are links for 'Insights' and 'FAQ'. Below is a search bar with placeholder text 'Search for anything...'. Underneath the search bar are several filter buttons: 'Search tips', 'Truth Social filters', 'Retweet filters', 'Deleted filters', 'Date filters', 'Device filters', and 'Export'. A red box highlights the 'Search tips' button. Below these filters, it says '80,356 tweets found' and has a dropdown menu set to 'Latest'. At the bottom of the search area, it shows the date 'Nov 4th 2024 - 10:20:38 AM EST' and interaction counts: 2k likes, 4k retweets, and a 'Show' link.

No text

Nov 4th 2024 - 10:20:27 AM EST 713 likes 3k retweets Show

Join me live in Raleigh, North Carolina at 10:45amEDT! Will be on @RSBN (<https://truthtwitter.com/@RSBN>) , @OAN (<https://truthtwitter.com/@OAN>) , @NewsMax (<https://truthtwitter.com/@NewsMax>) , @realAmericasVoice (<https://truthtwitter.com/@realAmericasVoice>) , and Donald J. Trump @Rumble (<https://truthtwitter.com/@Rumble>) , among others. See you soon—MAGA2024! <https://rumble.com/v5lrfjz-live-president-trump-in-raleigh-nc.html> (<https://links.truthtwitter.com/link/113425359035108189>)

Nov 4th 2024 - 9:56:53 AM EST 1k likes 5k retweets Show

<https://swampthevoteusa.com/> (<https://links.truthtwitter.com/link/112559845980056031>)

Nov 4th 2024 - 9:36:27 AM EST 2k likes 7k retweets Show

MAKE AMERICA GREAT AGAIN!

● Trump Twitter Archive는 트럼프의 모든 트윗 (2009~2021)을 수집, 보존한 비공식 데이터아카이브

<https://www.thetrumparchive.com/>

	datetime_utc	id	text	retweets	favorites
0	2016-11-08 06:42:00+00:00	7.95879E+17	Today we are going to win the great state of MICHIGAN and we are going to WIN back the White House! Thank you MI!... https://t.co/onRpEvzHrW	14011	45371
1	2016-11-08 11:43:00+00:00	7.95955E+17	TODAY WE MAKE AMERICA GREAT AGAIN!	281289	498035
2	2016-11-08 16:39:00+00:00	7.96029E+17	VOTE TODAY! Go to https://t.co/MXrAxYnTjY to find your polling location. We are going to Make America Great Again!... https://t.co/KPQ5EY9VwQ	24105	55870
3	2016-11-08 18:03:00+00:00	7.96051E+17	We need your vote. Go to the POLLS! Let's continue this MOVEMENT! Find your poll location: https://t.co/VMUdviltx1... https://t.co/zGOx74Ebhw	19145	51888

- 데이터 구성

- 원문 텍스트 데이터: 수집된 데이터는 트윗 단위의 원문 텍스트 데이터로 구성되며, 각 트윗에는 UTC 기준 작성 시각이 포함되어 있어 시계열 이벤트 데이터로 활용이 가능하다.
- 리트윗 제외: 본 연구에서는 시장 반응 분석의 정확성을 높이기 위해 리트윗을 제외하고 (retweet = false) 트럼프 본인이 직접 작성한 트윗만을 분석 대상으로 사용하였다.

- 데이터 수집 방법

- 웹 스크래핑: 데이터는 Trump Twitter Archive 웹사이트로부터 수집되었으며, 동적 페이지 구조를 고려하여 웹 스크래핑 방식으로 확보
- 중복 제거 및 전처리: 이후 트윗 고유 ID를 기준으로 중복을 제거하고, 텍스트 전처리를 거쳐 최종 분석용 데이터셋을 구성

- 분석 정보

- 트윗 정보: 각 트윗에 대해 작성 시각, 트윗 ID, 본문 텍스트, 리트윗 수, 좋아요 수 등의 정보가 포함되며, 이를 시장 데이터와 시간 기준으로 병합하여 트윗 발생 직후의 시장 반응을 분석하는 데 활용

| 사용 데이터

3. 핵심변수 정의

시장반응을
"무엇으로" 정의했는가?

"수익률"

트윗 이벤트를 "어떻게"
정의했는가?

"tweet_dummy"

- 시장 반응
 - 시장 반응은 SPXUSD 1분봉 가격을 이용한 단기 수익률로 측정
 - 트윗 발생 시점을 기준으로 이후 h 분 동안의 가격 변화를 계산하였으며, h 는 {1, 10, 20, 30, 45, 60, 120}분으로 설정
- 수익률 정의
 - 수익률은 각 시점의 시가(Open) 기준으로 다음과 같이 정의
 - $ret_{hm}(t) = \text{open}(t+h) / \text{open}(t) - 1$
- 시장 반응 분석
 - 이를 통해 트윗 직후부터 단기·중기까지의 시장 반응을 단계적으로 분석

- 트윗 이벤트 정의
 - 트윗 이벤트는 "분 단위 더미 변수(tweet_dummy)"로 정의
- 트윗 존재 여부
 - 해당 1분 구간에 트럼프의 트윗이 하나라도 존재하면 1, 존재하지 않으면 0의 값을 가짐
 - $tweet_dummy(t) = 1 \{ n_tweets(t) > 0 \}$
- 시장 반응 비교
 - 이를 통해 트윗이 있었던 분과 없었던 분의 시장 반응 차이를 비교

| 분석방법

01

회귀모형

각 시간 지평 $h \{1, 10, 20, 30, 45, 60, 120\}$ 에
대해 다음 모형을 추정

$$\text{Return}_{t,t+h} = \alpha_h + \beta_h \cdot \text{TweetDummy}_t + \varepsilon_{t,h}$$

02

변수정의

1. 종속변수: 수익률

$$\text{Return}_{t,t+h} = \frac{\text{Open}_{t+h} - \text{Open}_t}{\text{Open}_t} - 1$$

시점 t 에서 시작해 h 분 뒤까지의 forward 단순수익률

2. 독립변수: 트윗더미

$$\text{TweetDummy}_t = \begin{cases} 1, & \text{분 } t \text{에 트윗이 하나라도 존재할 때} \\ 0, & \text{그 외} \end{cases}$$

- α_h : 평균 수익률의 기준 수준 (절편)
- β_h : 트윗 발생 여부가 수익률에 미치는 평균 효과
- $\varepsilon_{t,h}$: 오차항

| 분석방법

03

추정 방법

• 로버스트 회귀 사용

- 극단적인 가격 변동: 1분봉 수익률 데이터는 뉴스, 트윗, 알고리즘 거래 등으로 인해 극단적인 가격 변동(outlier)이 자주 발생하는 특성을 가진다.
- OLS 추정의 민감성: 일반적인 OLS 추정은 이러한 극단값에 민감하여 결과가 왜곡될 수 있다.
- Huber 방식의 로버스트 회귀: 이에 본 분석에서는 Huber 방식의 로버스트 회귀(Robust Linear Model)를 사용하여 극단값의 영향을 완화한 상태에서 트윗의 평균적 효과를 추정하였다.

• 다중 시간 구간 반복 추정

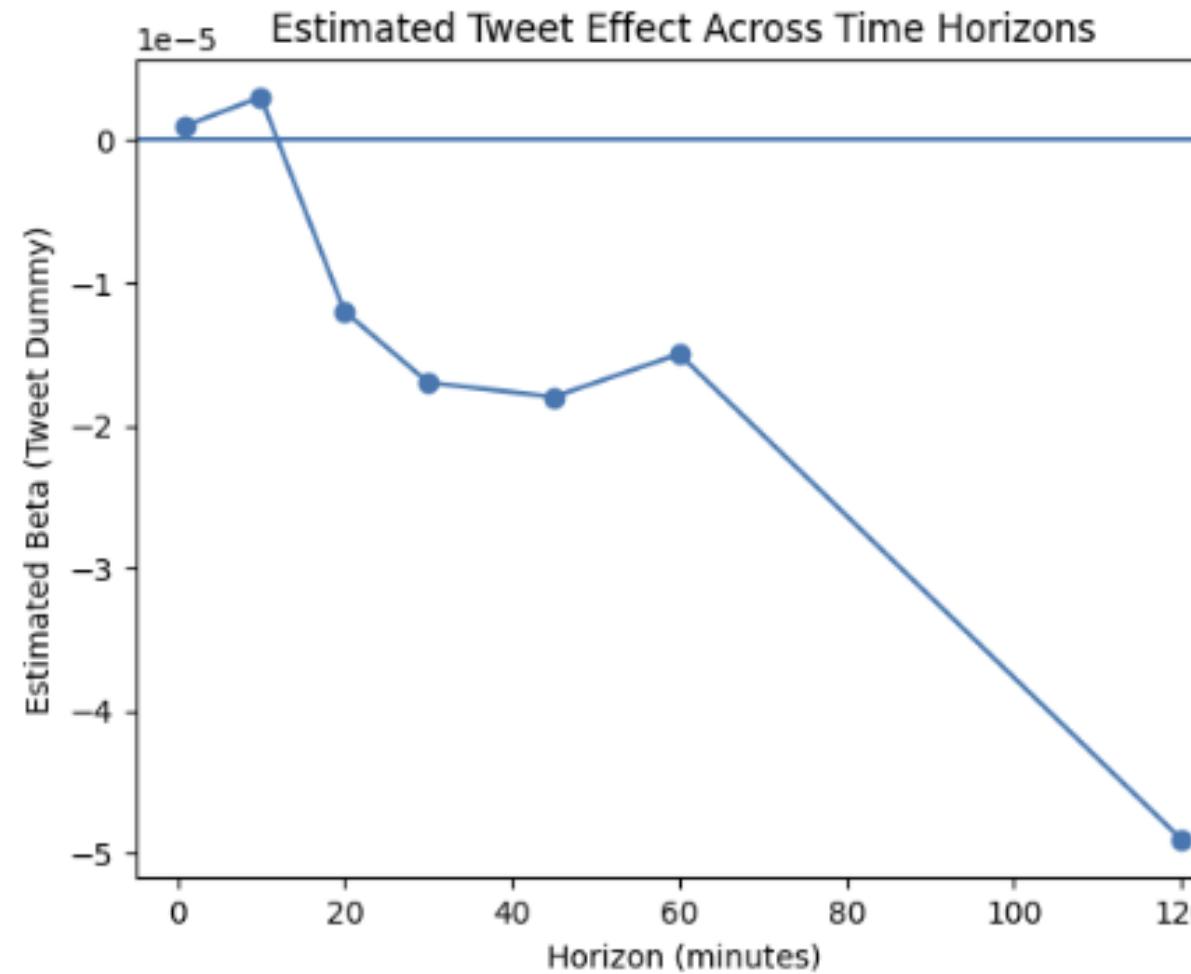
- 트윗 효과의 지속성: 트윗 효과의 지속성을 확인하기 위해 각 h 값에 대해 동일한 회귀모형을 반복 추정하였다.
- 효과 비교: 이를 통해 트윗의 효과가 단기적인지, 혹은 시간이 지나며 나타나는지를 비교한다.

Estimate (α_h, β_h) via Robust Linear Model (Huber M-estimator)

- $\beta > 0$
→ 트윗이 있었던 분 이후에 평균 수익률이 더 큼
- $\beta < 0$
→ 트윗이 있었던 분 이후에 평균 수익률이 더 작음
- $\beta \approx 0$ 또는 유의하지 않음
→ 트윗 존재 여부만으로는 시장 반응 차이가 뚜렷하지 않음

데이터 수집 및 기준선 설정

| 결과

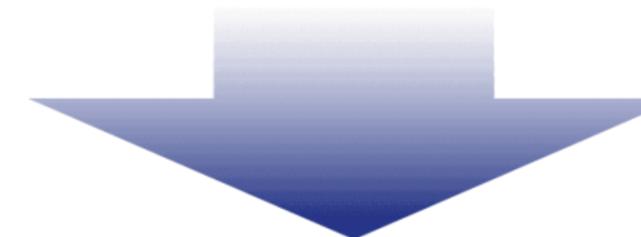


```
==== Tweet dummy 회귀 결과 요약 ====
   horizon_min      alpha    beta_Tweet    p_Tweet      obs
0           1  8.098145e-07  0.000001  0.555495  790516
1          10  1.170629e-05  0.000003  0.685314  790507
2          20  2.400724e-05 -0.000012  0.233538  790497
3          30  3.693491e-05 -0.000017  0.174949  790487
4          45  5.332024e-05 -0.000018  0.269778  790472
5          60  6.889934e-05 -0.000015  0.425812  790457
6         120  1.211033e-04 -0.000049  0.064668  790397
```

- x축
 - 시간 구간: (horizon, 분) → 1, 10, 20, 30, 45, 60, 120
- y축
 - β : (beta_Tweet) → 트윗 더미의 추정 계수
- 그래프 의미
 - 각 점: 해당 시간 구간에서의 트윗 효과 추정값
 - 가로선: $y=0$: “효과 없음” 기준선

“x축은 트윗 이후 시간 구간이고, y축은 트윗 더미의 추정 계수. 0선 위아래를 기준으로 보면, 대부분의 구간에서 효과가 크지 않고 통계적으로도 뚜렷하지 않습니다. 즉, 트윗이 있었다는 사실 자체만으로는 시장 수익률을 일관되게 설명하기 어렵다는 점을 보여줌”

“트윗 전체 효과는 약하다”



그렇다면 '어떤 트윗이' 시장을 움직이는가?

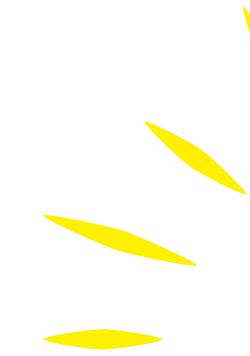
트윗 내용 기반 주제 분석

왜 LDA인가?



단문 특성

- 트윗이 너무 짧아 단어 하나만으로 주제 파악이 어려움



트윗을 주제(topic) 단위로 묶어주는 방법이 필요!!

대량 데이터

- 수천 개의 트윗을 효율적으로 분류 필요



→ LDA를 사용해줌으로써

표현 다양성

- 같은 이슈도 표현이 계속 바뀌어 패턴 인식 필요

-LDA 사용 목적

- 트윗을 주제별로 자동 분류를 위해서
- 무역전쟁, 가짜뉴스, 이민, 선거 등 큰 맥락 단위로 정리하기 위해서

트윗 내용 기반 주제 분석

| 전처리 과정

1. 데이터 로드 및 시간 필터

- 트윗 데이터를 불러온 뒤 `datetime_utc`를 `datetime`으로 변환했다.
- 이후 시간 처리의 일관성을 위해 `tz-naive`로 통일했다.
- 분석 대상 기간(2016-11-08 ~ 2019-09-22)으로 필터했다.
- 분 단위 분석을 위해 트윗 시각을 minute 단위로 내렸다.

2. 텍스트 기본 정리

- 모든 텍스트를 소문자로 변환했다.
- URL, 멘션(@)을 제거했다.
- 해시태그는 기호만 제거하고 단어는 유지했다.
- 이 단계의 목적은 텍스트 분석에 의미 없는 노이즈를 제거하는 것이다.

3. 토큰화

- 단어 단위로 토큰화했다.
- 기호를 제거했고, 길이 3 미만의 단어는 제외했다.
- 의미 없는 짧은 토큰을 줄여 LDA 품질을 높이기 위한 처리다.

6. 빈 문서 제거 및 동기화

- 전처리 후 토큰이 하나도 남지 않은 트윗은 제거했다.
- 이 과정에서 트윗 데이터와 LDA 입력 데이터의 길이를 정확히 맞췄다.

5. 불용어 제거 및 stemming

- 불용어(stopwords)를 제거했다.
- stemming을 적용해 단어 형태를 통일했다.(예: `countries` → `countri`, `people` → `peopl`)
- 표현 차이를 줄여 같은 의미의 단어들이 한 그룹으로 묶이게 했다.

4. Bigram 생성

- 자주 함께 등장하는 단어를 하나의 토큰으로 묶었다.(예: `fake_news`, `thank_you`, `our_countri`)
- 이를 통해 트럼프 트윗에 많은 관용구 표현을 살렸다.

LDA 모델학습

01

사전 및 코퍼스 생성

전처리된 토큰으로 단어 사전과 문서, 단어 행렬 구축

02

모델 선정

- 토픽수 20개
- passes10
- alpha-eta 자동 추정
- random_state: 고정

03

토픽 수 선정

```

19      countri    389
[30]: from gensim.models.coherencemodel import CoherenceModel
# c_v coherence (가장 많이 쓰이느 지표)
coherence_model = CoherenceModel(
    model=lda_model,
    texts=token_clean,
    dictionary=dictionary,
    coherence='c_v'
)

coherence_score = coherence_model.get_coherence()
print(f"Coherence (c_v): {coherence_score:.4f}")

```

Coherence (c_v): 0.3720

- 주제 선정 이유
 - 토픽 수(K) 선정 지표인 coherence 기준으로는 K=15가 약간 더 높았지만,
 - 토픽 해석 가능성과 주제 분리도를 종합적으로 고려해 K=20을 최종 선택했습니다.

트윗 내용 기반 주제 분석

| 토픽 워드클라우드 시각화



● 학습된 LDA

- **상위 단어:** 학습된 LDA에서 각 토픽의 상위 단어 10개를 뽑음
- **형태:** lda_model.show_topic(topic_id, topn=10) → (단어, 확률) 형태로 반환

● 워드클라우드

- **확률:** 이 확률을 워드클라우드에 넣어서 확률이 큰 단어일수록 글씨가 크게 보이게 만들었

● 결과

- **시각화:** 그래서 지금 그림은 “각 토픽에서 대표적으로 많이 나오는 단어(확률 높은 단어)가 무엇인지”를 시각화한 결과

2 트윗 내용 기반 주제 분석

I 무역전쟁 토픽 선정 이유

20개 토픽 중 Trade War 토픽이 가장 정책적 성격이 강했습니다.

- 명확한 경제 정책 이슈
- 시장 반응 측정 가능성
- 이벤트 분석에 적합한 구조

이 토픽이 강하게 등장했을 때 주식시장이 어떻게 반응하는지 확인하기로 결정

$$Return_{t,t+h} = \alpha + \beta \cdot TradeWar_t + \varepsilon_t$$

- $Return_{t,t+h}$:
이벤트 시점 이후 h 분 동안의 S&P 500 수익률
- $TradeWar_t$:
이벤트 분에서의 Trade War 토픽 강도
- β :
Trade War 토픽이 강할수록 수익률이 어떻게 변하는지를 나타내는 계수



YONHAP NEWS

2 트윗 내용 기반 주제 분석

| 이벤트 분석방법

Step 1

토픽 강도 계산

각 트윗의 Trade War 토픽 확률 측정

Step 2

분 단위 집계

같은 분의 최대 토픽 확률로 강도 정의

Step 3

이벤트 정의

중앙값 이상인 분만 선택 (노이즈 제거)

Step 4

정규장 필터

월-금 9:30-16:00 ET 시간대만 사용

Step 5

중복 제거

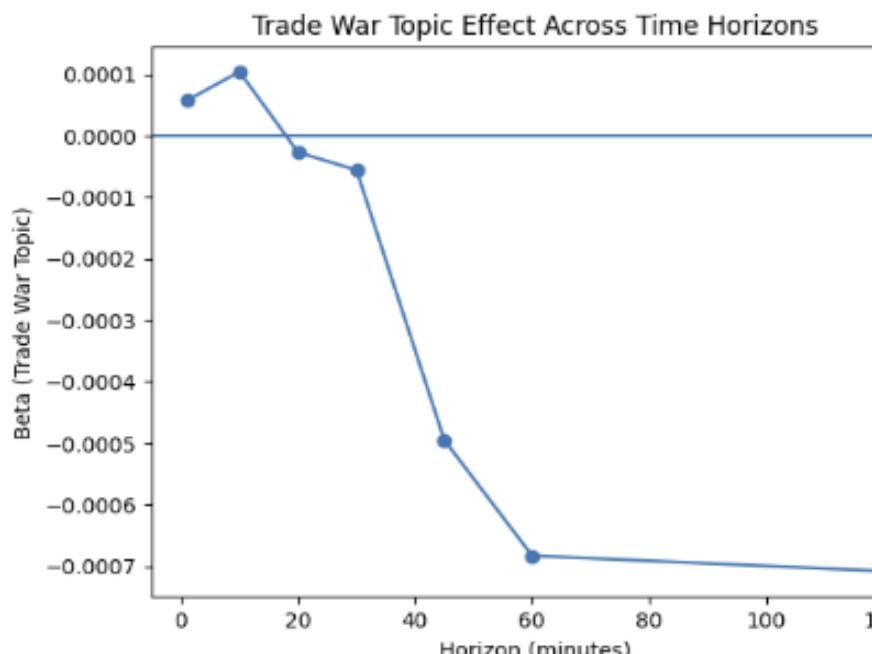
독립적인 이벤트 샘플만 유지

2 트윗 내용 기반 주제 분석

회귀분석 결과

- Trade War 토픽

시간 구간 (분)	β (Trade War)	p-value
1	0.000057	0.296
10	0.000104	0.533
20	-0.000027	0.909
30	-0.000056	0.853
45	-0.000495	0.180
60	-0.000683	0.108
120	-0.000708	0.270

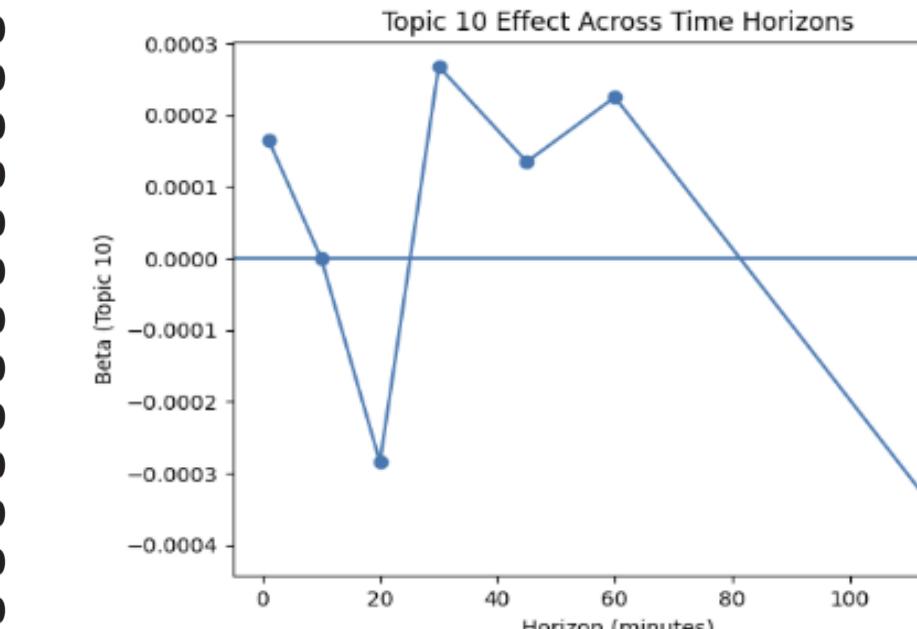


- 무역전쟁 토픽 결과 해석

- 무역전쟁 토픽은 단기(1~10분) 구간에서는 시장 수익률에 뚜렷한 반응을 보이지 않음
- 그러나 30분 이후부터 수익률 계수가 음(-)의 방향으로 전환되며, 시간에 지날수록 부정적 반응이 누적되는 경향이 관찰됨
- 이는 무역·관세와 같은 정책적 발언이 즉각적인 가격 변화보다는 점진적인 기대 조정 (expectation adjustment)을 통해 시장에 영향을 미칠 가능성 을 시사함

- Mexico 토픽

시간 구간 (분)	β (topic10)	p-value
1	0.000166	0.057
10	-0.000399	0.999
20	-0.000284	0.427
30	0.000268	0.569
45	0.000135	0.814
60	0.000225	0.740
120	-0.000408	0.683



- Mexico Trade 토픽 결과 해석

- Mexico Trade 토픽은 시간 구간에 따라 수익률 계수의 부호와 크기가 불규칙하게 변화
- 일부 구간에서 양(+) 또는 음(-)의 값이 나타나지만, 일관된 방향성이나 시간적 패턴은 관찰되지 않음
- 이는 해당 토픽이 시장에 체계적인 정보로 인식되기보다는 노이즈에 가까운 일반적 발언으로 해석 될 가능성을 시사함

2 트윗 내용 기반 주제 분석

| 결과

결과 해석

- 무역전쟁 토픽
 - 무역전쟁 토픽은 다른 토픽과 비교했을 때 상대적으로 더 뚜렷한 계수 방향성과 낮은 p-value를 보임
 - 이는 무역전쟁 키워드를 적절히 포착하였고 LDA 토픽 분리가 합리적으로 이루어졌음을 시사함
- 정책·무역 관련 주제
 - 트윗 전체 또는 일반 토픽 대비, 정책·무역 관련 주제가 시장에 더 민감하게 반응함을 확인

해석 시 고려사항

- 트윗 특성
 - 트윗은 매우 짧고, 하루에도 다수 게시됨
- 분석 방법
 - 1분 단위 분석에서는 노이즈가 크게 작용할 수 있음
- 단기 반응
 - 단기 구간에서는 반응이 희석되어 p-value가 높게 나타날 가능성이 존재
- 시간에 따른 변화
 - 그럼에도 불구하고 시간에 지날수록 계수가 음(-)으로 이동하고 p-value가 감소하는 패턴이 관찰됨



토픽 분석을 통해 시장반응이 주제에 따라 달라진다는 점을 확인
다만, 토픽 단위 분석은 여러 단어의 평균 효과로 인해 한계 존재

→ 관세·협상·중국 등 핵심 단어 수준 회귀 분석으로 확장

4 핵심 단어 수준 검증

| 단어 회귀 분석 구조

01

분석 단위

정규장(RTH) 전체 1분 봉 데이터를 기준으로 분석

02

설명변수 설정

- Tariff_t : 해당 분에 tariff가 한 번이라도 등장하면 1
- china_t : china가 등장하면 1
- deal_t : deal(파생형 포함)이 등장하면 1

03

종속변수 정의

1,10,20,30,45,60,120분 이후 수익률

04

회귀 방식

로버스트 회귀(Huber RLM)

정규장 전체 분 중에서 이 단어가 나온 분과 안 나온 분의 평균
수익률이 다른가를 직접 비교하는 구조

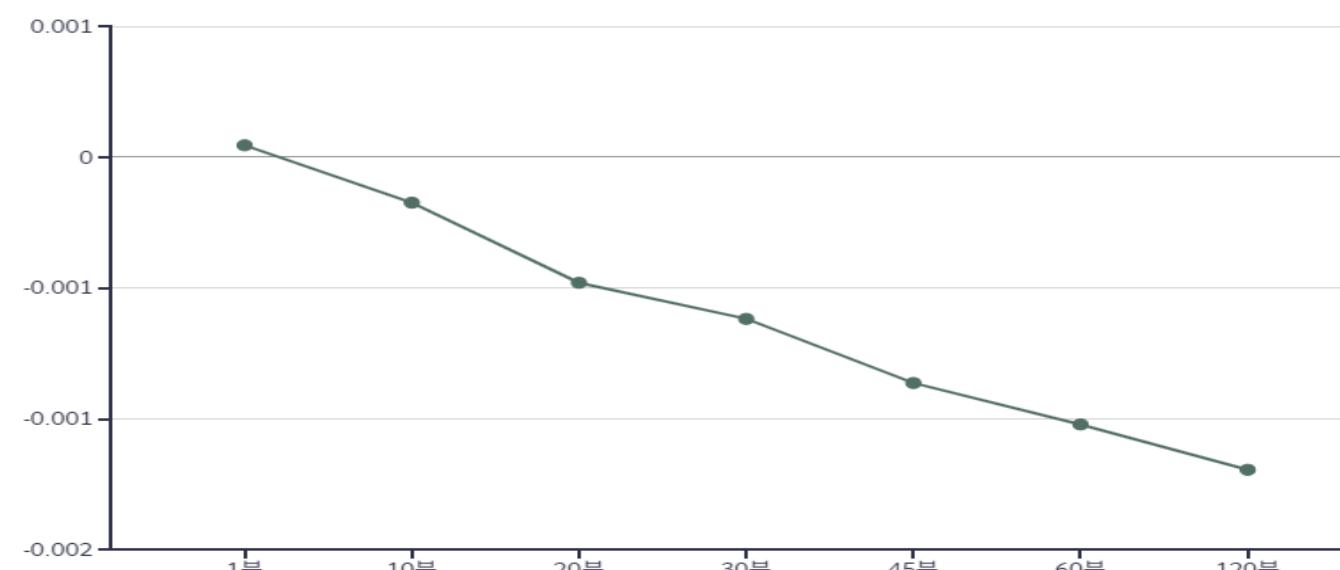
| 결과 요약: tariff vs china vs deal

- tariff

tariff 등장 분 수 (UTC): 150
RTH 분 수 : 161548

```
==== Tariff 단어 회귀 결과 (현재 메커니즘) ====
      horizon_min  beta_tariff  p_tariff    obs
0            1     0.000047  0.247333  161548
1           10    -0.000172  0.178092  161548
2           20    -0.000481  0.008004  161548
3           30    -0.000621  0.005115  161548
4           45    -0.000865  0.001366  161548
5           60    -0.001022  0.001002  161548
6          120   -0.001196  0.005431  161548
```

시간대별 Tariff 효과 분석



- 결과 패턴

- 1분: 유의하지 않음
- 20분 이후부터
- 계수: 일관된 음(-)
- p-value: 0.01 이하 수준으로 급격히 낮아짐
- 60분, 120분 구간에서도 음(-) 효과 유지

- 해석

- 관세는 단기 노이즈가 아니라
- 시간이 지나면서 시장 전반에 부정적 영향을 주는 정책 충격
- 분 단위 데이터에서 이 정도 패턴은 매우 강한 결과

시간이 지날수록 계수가 더 음(-)의 값을 보이며, 이는 정책 발언 -> 해석 -> 포지션 조정의 자연 반응 구조와 일치

4 핵심 단어 수준 검증

| 결과 요약: tariff vs china vs deal

- china, deal

china 등장 문 수 (UTC): 256
 deal 등장 문 수 (UTC): 320
 RTH 문 수 : 161548

==== china / deal 단어 회귀 결과 (현재 메커니즘) ===					
	word	horizon_min	beta	p_value	obs
0	china	1	1.775639e-05	0.523855	161548
1	china	10	-6.049340e-07	0.994459	161548
2	china	20	-2.318673e-04	0.061112	161548
3	china	30	-1.213681e-04	0.422203	161548
4	china	45	-4.233752e-04	0.021509	161548
5	china	60	-2.993624e-04	0.157770	161548
6	china	120	-2.659189e-04	0.364711	161548
7	deal	1	1.733443e-05	0.543432	161548
8	deal	10	-1.336060e-04	0.134244	161548
9	deal	20	-2.349585e-04	0.063884	161548
10	deal	30	-4.032465e-05	0.794559	161548
11	deal	45	-2.344815e-04	0.213760	161548
12	deal	60	-1.953690e-04	0.367982	161548
13	deal	120	-4.005579e-04	0.182465	161548

(1) China 결과

대부분 구간에서 유의하지 않으며, 일부 구간(45분)에서만 약한 신호를 보이고, 계수 방향도 일관적이지 않음.

- 너무 포괄적인 단어
- 외교, 정상회담, 칭찬/비난 등 맥락이 섞임
- 시장이 "중국 언급 자체"에는 반응하지 않음

(2) Deal 결과

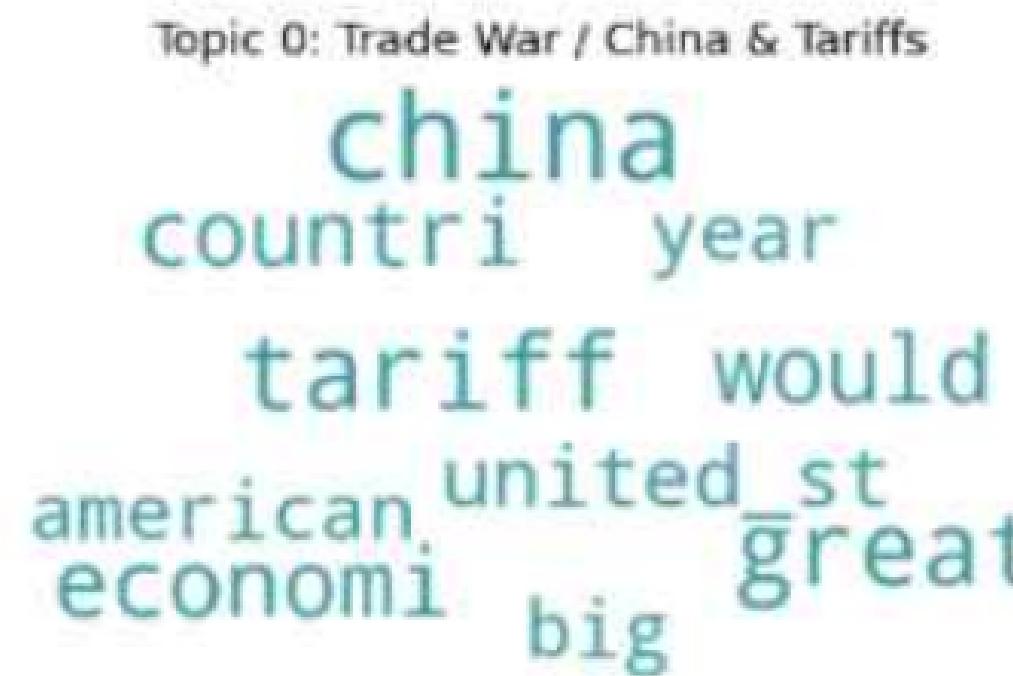
전 구간에서 유의성이 거의 없으며, 계수 크기도 작고 방향성도 불안정함.

- 협상, 발언, 서사적 표현
- 실제 비용이나 정책 충격을 직접 의미하지 않음
- 시장 반응이 약한 것이 자연스러움

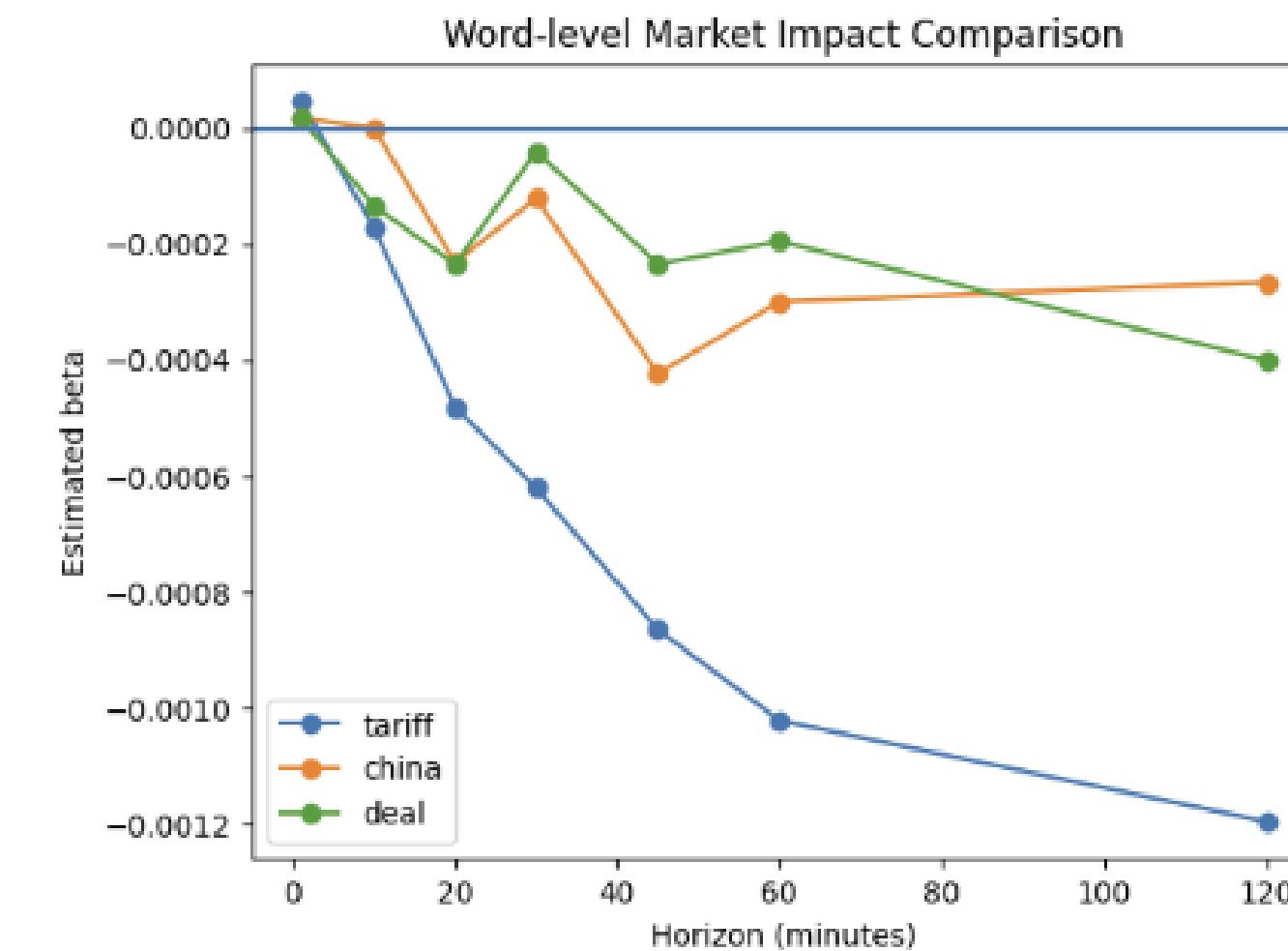
핵심 단어 수준 검증

| LDA 결과와 단어 회귀 결과의 연결성

- LDA 결과



- 단어 회귀 결과



LDA가 “아무 무역 단어”를 뚫은 게 아니라 실제로 시장에 영향을 주는 정책 키워드를 포함한 주제를 잘 분리해냈다는 뜻.

트윗 감성 분석



감성분석의 목적

감성이 시장 반응을 설명하는가

→ 앞선 분석에서는 트윗의 게시 여부, 주제(Topic), 핵심 단어(Word)를 중심으로 시장 반응을 확인

본 파트에서는 트윗의 톤(tone), 즉 긍·부정 감성(Sentiment)이 단기 시장 반응과 어떤 관계를 가지는지 검증



- 긍정적인 트윗 이후 시장 수익률은 더 높아지는가
- 부정적인 트윗 이후 시장 수익률은 더 낮아지는가
- 감성 효과는 모든 시기에서 동일한가,
- 아니면 특정 시기에서 더 뚜렷하게 나타나는가



[분석 구성]

Step 1. 트윗 감성 점수 산출

- 트윗 원문 텍스트를 기반으로 트윗별 감성 점수 계산
각 트윗을 정량 지표(숫자)로 변환

Step 2. 일별 감성 지표 생성

- 하루 동안 게시된 트윗의 평균 감성 점수 계산
시장 데이터와 시간 축 정합

Step 3. 감성과 시장 반응 관계 분석

- 감성 점수와 트윗 이후 단기 수익률 및 변동성 간 관계 확인
긍·부정 감성 구간별 시장 반응 비교

Step 4. 시기별 비교 분석

- 정책 이슈 강도가 다른 시기로 분리 분석
감성 효과가 언제 더 강하게 나타나는지 확인

- 단순한 트윗 게시 여부가 아닌
트윗의 톤(감성)이 시장 반응과 연결되는지 검증
- 감성은 주제·정책 이슈보다 미세한 신호이므로
단기 구간 중심의 반응에 초점

트윗 감성 분석

감성 점수 계산 방식 (Sentiment Scoring)

:감성 점수 산출 방법

▶ VADER 감성 분석 모델의 "compound score" 사용

감성 점수 범위

-1 : 매우 부정적

+1 : 매우 긍정적

트윗 하나당 하나의 감성 점수 부여

:해당 점수는 이후 회귀 분석에서

Sentiment 변수로 직접 사용

이 단계에서는
트윗의 톤(긍·부정)을
단일 숫자 지표로 변환

→텍스트 데이터를
시장 데이터와 결합 가능한 형태로 정량화

즉, VADER compound score →
트윗 감성의 정량적 지표화



왜 VADER인가?

- 소셜미디어 텍스트에 특화된 감성 분석 모델
짧은 문장과 감정 표현에 강점
- 트윗과 같은 짧은 텍스트 분석에 적합

감성 분석 결과 (Sentiment Effect)

회귀 분석 결과 요약

트윗 감성 점수와 이후 단기 수익률 간 유의한 관계 관찰

긍정 감성 트윗 이후 평균 수익률은 양(+)

부정 감성 트윗 이후 평균 수익률은 음(-)

[결과 해석]

- 감성 점수가 높을수록
이후 수익률이 상대적으로 증가하는 경향
- 감성 점수가 낮을수록
이후 수익률이 상대적으로 감소하는 경향

<효과의 크기>

감성 효과는 방향성은 존재하나 크기는 크지 않음
→단기 시장 반응을 설명하는 보조적 요인에 해당

트윗 감성은 시장 반응의 방향을 설명하지만,
영향력은 "제한적"



트윗 감성 분석

감성과 수익률의 관계 분석 (Regression Results)

→ "분석 설정 요약"

- 트윗 시점을 UTC 기준으로 정리 후 뉴욕시장 시간으로 변환

- 해당 시점 직전 1분봉 가격을 기준으로

이후 h분 수익률 계산

- 각 horizon별로 단순 선형회귀 수행

$$\text{ret}_{t,t+h} = \alpha + \beta \cdot \text{Sentiment}_t + \varepsilon_t$$

[변수 정의]

- 종속변수

$\text{ret}(t, t+h)$

→ 트윗 시점 :t 기준, h분 뒤 S&P 500 수익률

- 독립변수

Sentiment_t

→ 해당 트윗의 VADER compound 감성 점수

(-1: 매우 부정 ~ +1: 매우 긍정)

추정 방법

:OLS 회귀 사용

→ 분 단위 수익률의 이분산성 및 극단값 영향을 고려하여 HC3 로버스트 표준오차 적용



회귀 결과 요약

- 대부분의 horizon에서
감성 계수(β)의 부호는 양(+)
- 일부 단기 horizon에서만 통계적 유의성 관찰

--- horizons별 회귀 요약 ($\text{ret}_h \sim \text{sentiment}$) ---						
horizon_min	n_obs	beta_sentiment	t_sentiment	p_sentiment	R2	
0	1	0.000243	1.891397	0.058571	0.000499	
1	10	0.000185	1.410732	0.158324	0.000284	
2	20	0.000176	1.338328	0.180798	0.000255	
3	30	0.000260	1.954486	0.050644	0.000552	
4	45	0.000241	1.820567	0.068673	0.000476	
5	60	0.000216	1.594436	0.110838	0.000374	
6	120	0.000188	1.334551	0.182023	0.000260	

Horizon	p-value	해석
1분	≈ 0.059	10% 유의수준에서 유의
30분	≈ 0.051	10% 유의수준에서 유의
45분	≈ 0.069	경계적 유의성
그 외	> 0.1	통계적 유의성 없음



→ 10% 유의수준에서 통계적으로 유의

(2) 추가 분석

-긍정 vs 부정 감성 t-test 결과 해석

긍정 트윗 개수: 5128 부정 트윗 개수: 2758

```
--- horizons t-test (positive vs negative sentiment) ---
   horizon_min pos_n neg_n    t_stat   p_value   pos_mean   neg_mean
0           1     4491    2369  2.246788  0.024707  0.000133 -0.000258
1          18     4374    2333  1.861475  0.062742  0.000118 -0.000211
2          28     4388    2337  1.746661  0.088764  0.000103 -0.000288
3          38     4336    2295  2.325568  0.020887  0.000055 -0.000363
4          45     4364    2281  2.266734  0.023455  0.000086 -0.000322
5          68     4296    2313  1.813646  0.069808 -0.000035 -0.000365
6         120     4224    2284  1.859219  0.063062 -0.000039 -0.000392
```

Horizon	p-value	해석
1분	0.0247	5% 유의
10분	0.0627	10% 유의
20분	0.0308	5% 유의
30분	0.0209	5% 유의
45분	0.0235	5% 유의
60분	0.0598	10% 유의
120분	0.0631	10% 유의

① 통계적 유의성

- 긍정 감성 트윗과 부정 감성 트윗 간 평균 수익률 차이에 대한 t-test 수행
- 대부분의 horizon에서 5% 또는 10% 유의수준에서 통계적으로 유의한 차이 관찰
- 여러 시간 구간에서 반복적으로 유의한 결과가 나타나 감성 효과는 우연에 의한 단발적 현상이라기보다 일정한 패턴을 가지는 결과로 해석 가능

② 방향성 해석

- 모든 horizon에서 긍정 감성 트윗 이후 평균 수익률은 양(+), 부정 감성 트윗 이후 평균 수익률은 음(-)의 값을 보임
- 이는 트윗의 감성 톤과 이후 시장 수익률 방향이 일치함을 의미
- 시장은 트윗의 감정적 뉘앙스를 완전히 무시하지 않으며, 긍정적 메시지에는 상대적으로 우호적으로, 부정적 메시지에는 부정적으로 반응할 가능성을 시사



트윗 감성은 시장 반응의 방향성과 일관된 연관성을 보임

6 시기별 비교 및 종합 해석

<Sentiment Effect in 2017 분석 배경>

: 2017년은

무역전쟁 본격화(2018년 이후) 및 대형 정책 충격 이전 시기로
외생적 정책 노이즈가 상대적으로 낮은 구간

→ 정책 이슈에 가려지지 않은 트럼프 트윗 자체의 감성 효과를
보다 선명하게 관찰하기 위해 "2017년만" 분리 분석

1) 회귀 분석 결과 요약

== 2017 horizons별 회귀 요약 (ret_h ~ sentiment) ==						
horizon_min	n_obs	beta_sentiment	t_sentiment	p_sentiment	R2	
0	1486	0.001540	2.004238	0.045045	0.002739	
1	1458	0.001410	1.833874	0.066673	0.002317	
2	1405	0.001084	1.402843	0.160664	0.001401	
3	1406	0.001231	1.620329	0.105162	0.001863	
4	1373	0.001441	1.864339	0.062274	0.002544	
5	1458	0.001704	2.332269	0.019687	0.003612	
6	1185	0.000776	0.955712	0.339218	0.000756	

- 모든 horizon에서 감성 계수(β)는 양(+)의 값을 유지
- 특히 단기 구간에서 감성 효과가 상대적으로 뚜렷하게 관찰됨

[통계적으로 유의한 구간]

- 1분 ($p \approx 0.045$, 5% 유의)
- 10분 ($p \approx 0.067$, 10% 유의)
- 45분 ($p \approx 0.062$, 10% 유의)
- 60분 ($p \approx 0.020$, 5% 유의)



긍정적인 감성 트윗 =
이후 단기 수익률이
높아지는 방향성 일관성
확인

2) 긍정 vs 부정 감성 t-test 결과

[2017] 긍정 트윗 개수: 1336 부정 트윗 개수: 668

== 2017 horizons별 t-test (positive vs negative sentiment) ==						
horizon_min	pos_n	neg_n	t_stat	p_value	pos_mean	neg_mean
0	1	855	1.770047	0.077072	0.028892	0.027170
1	10	811	1.786405	0.074363	0.028637	0.026911
2	20	799	1.319130	0.187482	0.028925	0.027625
3	30	790	1.389762	0.164961	0.029236	0.027901
4	45	787	2.085828	0.037306	0.029026	0.027007
5	60	812	1.710768	0.087453	0.028549	0.026969
6	120	681	0.909738	0.363269	0.029694	0.028746

- 2017년은 전반적으로 상승장이 지속된 시기로,
부정 감성 트윗 이후에도 평균 수익률이 완전히 음(-)으로 전환되지는 않음
- 그럼에도 불구하고,
모든 horizon에서 긍정 감성 트윗 이후 평균 수익률이
부정 감성 트윗 이후보다 더 높게 나타남
- 일부 단기 horizon에서는
이 평균 차이가 통계적으로 유의

→ 정책 노이즈가 낮은 2017년에는
트윗 감성 효과가 전체 기간보다 상대적으로 더 선명

따라서, 감성은 시장을 지배하는 요인은 아니지만,
정책 이슈가 약한 환경에서는 의미 있는 설명력을 가질 수 있음

▶ 정책 이슈가 약한 시기에는
트윗 감성 효과가 상대적으로 더 뚜렷하게 관찰됨



결론

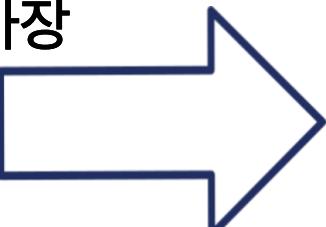
Result



- 전체 기간을 보면 S&P 500 Buy & Hold가 압도적으로 우수
- 그래프 모두에서 장기적으로는 S&P 500을 그냥 들고 가는 전략이 가장 성과가 좋음
- Trade War 토픽이 등장할 때마다

1시간짜리 솟 포지션을 취하는 전략은

누적 수익률 기준으로 Buy & Hold를 이기지 못함



주황색 선은 관세 전쟁 토픽이 관측된 시점 직후
1시간 동안 인버스 포지션을 취했을 때의 누적
수익률

LS 증권

가치투자자의 일상



아 원래 투자하다 보면 피도 나고 그래

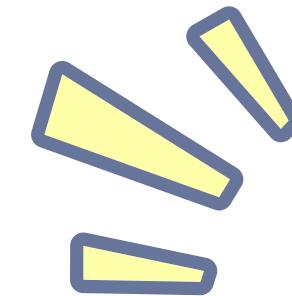
“관세 전쟁 트윗이 나올 때마다 솟을 치는 전략”은
투자 전략으로서 우월하다고 말하기 어렵다!

결론

Result

Trade War가 '트렌드'일 때의 누적 수익률 (2018-05 이후)

관세전쟁이 트랜드일때의 누적수익률(2017~)



트럼프 트윗의 토픽이 일회성 뉴스가 아니라 하나의 서사(트렌드)로
이어질 때 의미있는 리스크 관리 지표로 사용.

→ 2018-05 이후를 기준으로 리베이스한 누적 수익률 비교

[관세 전쟁 이슈가 집중적으로 등장한 구간을 별도로 관찰]

- 2018년 하반기~2019년 초반에는 시장 변동성이 확대되며 Trade War 이벤트 기반 솟 누적 수익률이 상대적으로 개선되는 '역전 구간'이 관찰됨

→ 해당 구간에서 Buy & Hold는 큰 낙폭을 겪는 반면 Trade War 신호 기반 포지션은 상대적으로 방어적인 움직임을 보임



분석에서의 어려움 및 한계점



Table Page

항목	발생한 어려움	해결방안
수집중 어려움	Selenium 기반 스크롤 크롤링 중 대량 데이터 수집 과정에서 잦은 비정상 중단 발생	<ul style="list-style-type: none"> 중단 지점부터 재크롤링을 반복 수행 데이터를 분할 수집 후 병합하는 방식으로 누락 최소화 고유 트윗 ID 기준 중복 제거 및 정합성 검증 수행
토픽 모델링 관련 한계, 어려움	전처리 방식과 토픽 수(K)에 따라 결과가 크게 달라지는 민감성 문제	<ul style="list-style-type: none"> K 값을 5-30 범위로 변화시키며 coherence 지표 반복 계산 정량 지표와 대표 단어 육안 검토를 병행하여 해석 가능성이 가장 높은 결과 선택

감사합니다

텍스트마이닝 Term Project 최종 발표

지금까지 떡잎방정대였습니다.