



MZ세대를 위한 도메인 별 투자의견 감성분석을 통한  
상위 긍/부정 문장 추출 서비스

# 세상을 읽다

팀 | MZ와 같이하자 가치투자



금융, 인공지능, 데이터 분석  
각 분야에 열정이 있는 세 명이 모여 기획했습니다.



황우진  
경희대학교 컴퓨터공학과

인공지능



금융

최지웅  
경희대학교 경제학과

데이터분석

이동환  
경희대학교  
소프트웨어융합학과

# INDEX

***01***

## 주제 소개

주제 선정 배경  
주제 설명  
내부데이터 분석  
시스템 구조

***02***

## 데이터 수집 및 가공

데이터 수집 기준  
유튜브 댓글  
유튜브 영상  
해외 뉴스  
증권사 리포트

***03***

## 모델링 및 분석 결과

모델 및 훈련 데이터 셋  
각 데이터 도메인 분석  
분석 결론

***04***

## m.club 적용

주요 기능  
기대 효과

## ✓ 왜 MZ에 집중해야 할까?



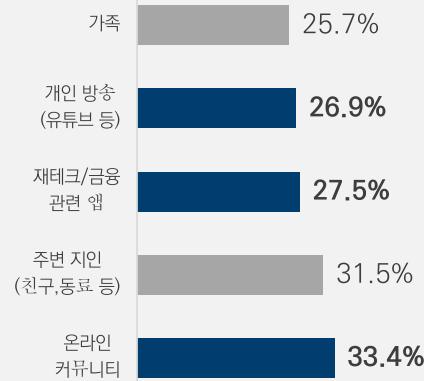
### 락인효과 *Lock-In Effect*

해가 지날수록  
MZ세대의 구매력은 상승할 것입니다.

보통 한번 정한 증권사 어플은  
잘 바꾸지 않음을 고려할 때  
지금이야 말로 MZ를 공략할 적기입니다.

## ✓ MZ는 어떻게 투자를 고려할까?

### [재테크 정보 획득 경로 TOP5]



[출처 : 대학내일, Base : 3년 내 재테크 경험자, n=721, 복수 응답]

2020년 말, MZ세대  
(19~34세)를 대상으로 한  
'대학 내일'의 조사에 따르면

**MZ개인투자자들은 한 곳에  
편향되지 않은 채 비슷한 비율로  
다양한 곳에서 의견을 들어보고  
투자합니다.**

**특히나 상위 5개 경로 중 3개가  
온라인 경로였습니다.**



즉, MZ세대는 직접 유튜브 영상을 보고, 댓글을 읽고, 뉴스, 증권사 리포트를 읽으며 투자 의견을 수집합니다.

**이러한 이들의 심리를 이해한 서비스가  
m.club에 포함되어 MZ세대를 락인할 시점입니다.**

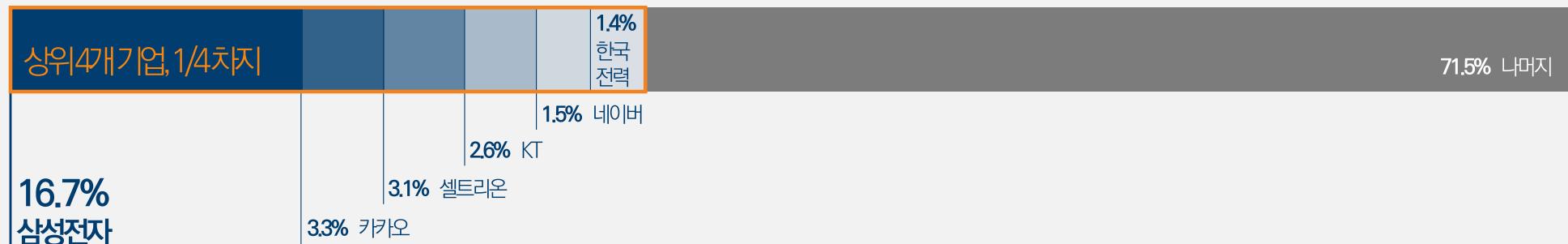
저희의 목적은 투자자들이 하나의 플랫폼에서 다양한 정보를 수집할 수 있도록 하는 것입니다.  
이러한 배경에서 저희는 서비스명을 ‘세상을 읽다’로 지었습니다.



- ✓ MZ세대는 단순히 한 곳의 목표 주가와 투자 의견을 듣고 투자로 진행하지 않습니다.
- ✓ 그들은 다양한 도메인에서 투자 의견을 수집하며 투자를 고민합니다.
- ✓ 또 하나의 도메인에서도 다양한 의견을 수집합니다.
- ✓ 즉, 유튜브 영상을 보더라도 하나가 아닌 여러 개를 봅니다.

프로토타입으로 분석할 종목은 내부데이터 분석을 통해 결정했습니다.

[BUY&HOLD 위주의 2030 주식보유 금액]



위 그래프는 20-30대, buy&hold비중이 50%를 넘기는 고객의 주식보유금액((20년 8월~21년 5월)에서 상위 6개 종목과 나머지 종목이 차지하는 비중을 시각화한 그림입니다.  
삼성전자 하나의 기업만 해도 16.66%를 차지했으며, 상위 4개 기업이  $\frac{1}{4}$ 을 차지함을 알 수 있습니다.

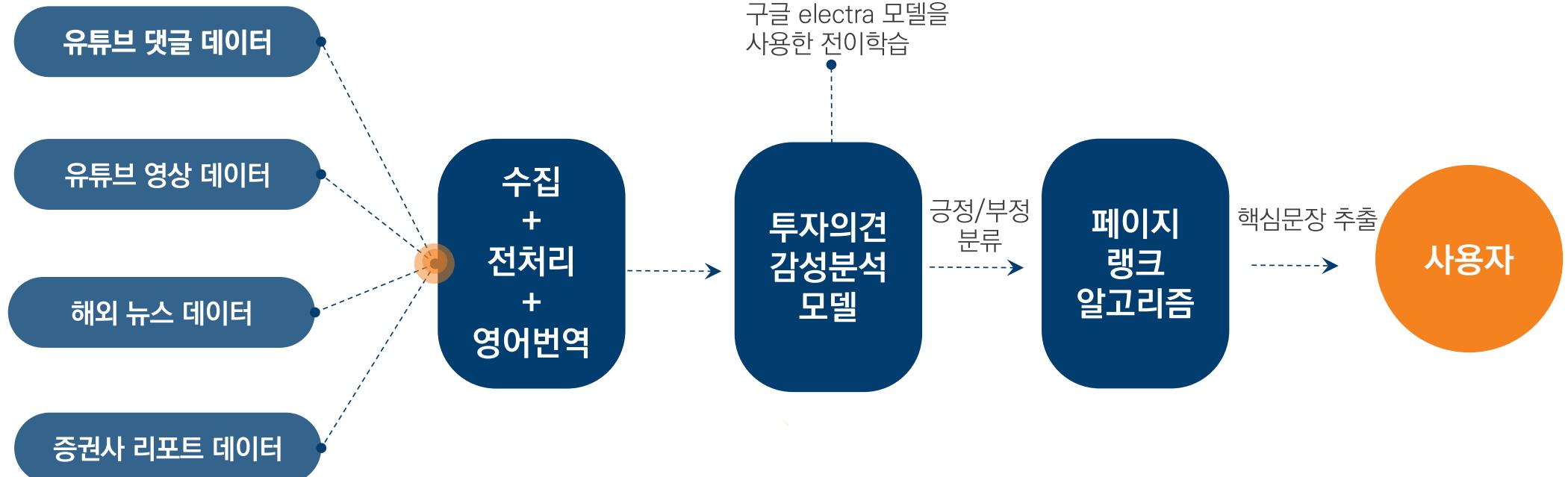
즉, MZ세대는 대장주에 편중해 자산을 보유함을 확인했습니다.

따라서,

‘세상을 읽다’는 대장주에 집중한 서비스를 기획하기로 했습니다.  
그 중에서도, 가장 비중이 높은 삼성전자(005930)를 분석했습니다.

‘세상을 읽다’는 총 4가지 도메인에서 수집한 투자의견의  
 ①긍/부정 비중과 ②각 근거의 순위를 도출해 보여주는 서비스입니다.

대중과 전문가 의견 모두 수집



## 데이터 수집 기준



## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외뉴스
- 증권사리포트



## 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

‘세상을 읽다’를 잘 구현하기 위한 관건인 ‘투명성’과 ‘다양성’을 만족하기 위해 다음과 같은 기준을 세우고 삼성전자 데이터를 수집하였습니다.

**01**

### 중,단기 데이터

중단기 데이터(2021년 8월~ 9월)만을  
수집하여 분석했습니다.

**02**

### 다양한 도메인

최대한 다양한 도메인  
(유튜브, 뉴스, 리포트)에서  
데이터를 수집했습니다.

또 각 도메인에서도  
최대한 다양하게 데이터를  
수집했습니다.

**03**

### 문장 데이터

공부정 처리를 위해  
자연어 문장 형식의 데이터만  
수집했습니다.



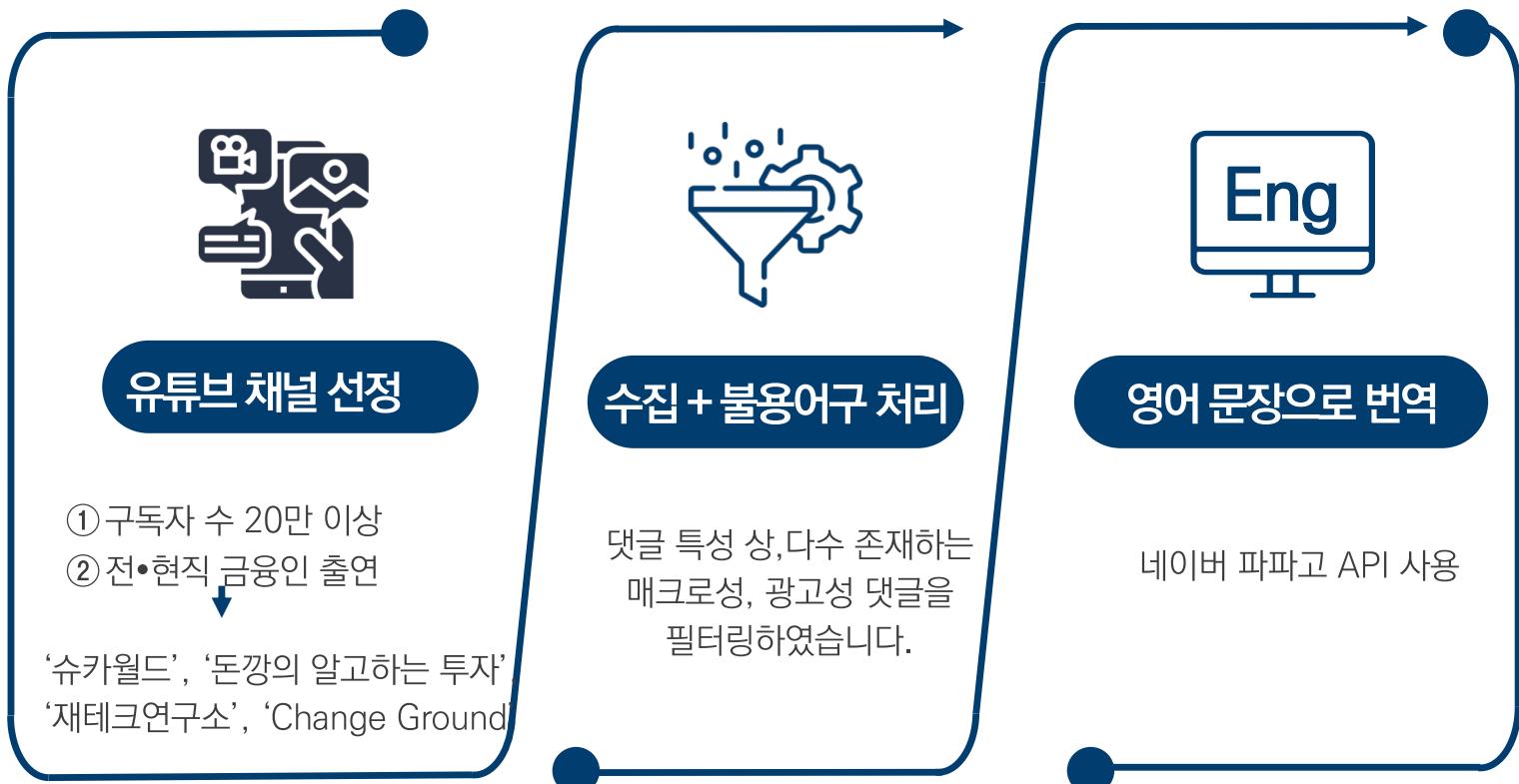
### 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

### 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

대중의 생각을 알기 위해 국내 유튜브 영상 내 댓글을 수집했습니다.





## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외뉴스
- 증권사리포트



## 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

국내 유튜브 영상의 전문가 의견을 수집했습니다.



## 유튜브 채널 선정

- ① 구독자 수 20만 이상
- ② 전·현직 금융인 출연

'김작가 TV', 'Super K-슈퍼개미김정환', 'Change Ground'



## STT 변환

네이버 STT  
(Speech To Text)  
API를 사용하여 유튜브 영상  
에서 문장을 추출하였습니다.



## 영어 문장으로 번역

네이버 파파고 API 사용



## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외뉴스
- 증권사리포트



## 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

국내 주식에 큰 영향을 끼치는 해외의 시각도 수집했습니다.



## 구글 뉴스 수집

구글 검색에서 증권 관련 뉴스\*를 키워드 기반으로 직접 크롤링하여 수집했습니다.



## 문장 분리

영어 문장 세그멘테이션을 사용해서 분석 성능을 높였습니다.

\*구글(미국)뉴스: investing.com, seekingalpha.com, finance.yahoo.com에 올라온 뉴스 기사



## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트



## 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

국내 최고 전문가인 증권사 애널리스트의 의견도 수집했습니다.



## 증권사 리포트 수집

한경컨센서스에서 21년 8, 9월  
삼성전자를 다룬 키움증권,  
유진투자증권, DGB 금융그룹  
리포트를 수집했습니다.



## 문장분리

한글 문장 세그멘테이션을  
사용해 분석 성능을  
높였습니다.



## 영어 문장으로 번역

네이버 파파고 API 사용

 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

 모델링 및 분석 결과

- 모델 및 훈련 데이터셋
- 각 데이터 도메인 분석
- 최종 결론

훈련 데이터셋은 Kaggle의 금융 뉴스 감성 분석 데이터셋을 사용했습니다.

## 훈련 데이터셋 **kaggle**

- Kaggle의 Sentiment Analysis for Financial News
- 투자의견 감성분석 데이터셋
- 금융 뉴스를 부정, 중립, 긍정 3가지로 분류한 총 4846개의 데이터
- 학습을 위해 검증 데이터셋과 테스트 데이터셋으로 각각 500개씩 분리해 학습
- 기본적, 기술적 분석을 비롯한 주가에 영향을 끼치는 다양한 변수를 다루고 있는 데이터셋

Positive	Earnings per share (EPS) for the first quarter 2007 amounted to EURO .07, up from EURO .04
Negative	Net sales have been eaten by the weak US dollar.
Neutral	This would be a huge process.

 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

  
electra 모델 Google

- SST(Stanford Sentiment Treebank) Benchmark에서 96.9%의 높은 성능을 보이고, 이외에 GLUE, Quora 등 여러 NLP 성능지표에서도 높은 성능을 보이는 모델입니다.
- ELECTRA 모델은 **GAN**의 원리를 사용한 데이터 증강을 통해 **적은 데이터나 파라미터를 가지고도 학습이 효과적으로 수행**되도록 합니다. **데이터가 적어서 학습이 잘 안되는 것을 방지하기 위해 구글에서 만든 ELECTRA 모델을 사용하였습니다.**

‘세상을 읽다’는 투자의견 감성분석 모델로  
GAN Based이면서도 무겁지 않고, 성능이 높은  
**electra-small** 모델의 가중치를 전이학습하여 사용했습니다.

테스트 셋 500개로 **f1 score** 성능을 측정해본 결과 **0.95 이상의** 높은 점수를 얻었습니다.



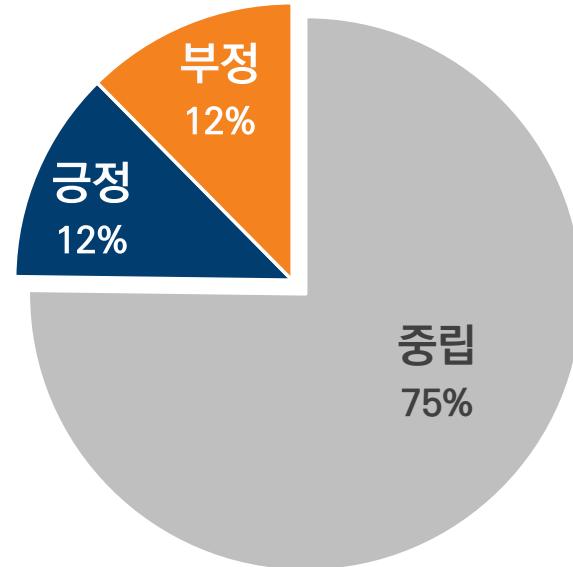
### 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

### 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

## 유튜브 댓글 분석



### 유튜브 댓글 긍정 상위 3문장

- **1순위:** 솔직히 누가봐도 오르지.. ddr5 라는 호재도있고 강 잠깐 조정정도로 봐야지 무슨 삼성전자가 망한다 이러는건 좀
- **2순위:** 슈카가 삼성전자 언급했다. 내일부터 반등있을듯! 슈카는 고점 아니면 저점 판독기임
- **3순위:** 반도체 업황에 대한 단기 우려는 이해하지만 지금 삼성전자는 너무 과매도임. 내년엔 주요 업체 서버 증설이 확실시되는 상황이라 길게 보면 좋음(DDR5 출시 + 인텔 새 서버CPU 출시 + 서버수요 증가 + 서버 교체주기 도래 등)

### 유튜브 댓글 부정 상위 3문장

- **1순위:** 삼성전자를 최소한 10년 들고 있지 않을거면 사지마라.. 매달 조금씩 사면 돼..
- **2순위:** 삼성전자도 세월가니 이제 서서히 막을 내리는구나 ?역시 영원한건 없구나 ? 이제 카카오시대가 오는구나~
- **3순위:** 형~ 삼성전자도 맛이 갔어



## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

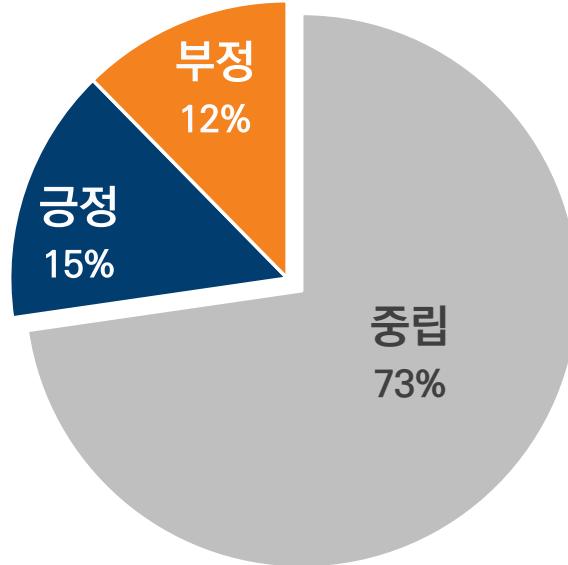


## 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론



## 유튜브 영상 분석



## 유튜브 영상 긍정 상위 3문장

- **1순위:** 엄청난 변동성과 함께 열심히 훈들고 있습니다. 제가 오늘 재미를 본다면, 몇몇 주식이 상승해서 좋다고 말할 것이고, 그러면 투자자들은 빚을 내서 그것을 살 수 있을 것입니다.
- **2순위:** 65000원 이하로 살 수 있는 기회를 주신다면, 그만 살 수 있을까요?
- **3순위:** 제 성격은 매우 낙천적이고 대통령 발언을 어디서 다루든 간에 그것은 다시 올라갑니다.

## 유튜브 영상 부정 상위 3문장

- **1순위:** 이러한 상황이 어떻게 발생하게 되었는지, 저는 최근 삼성전자의 하향 보고서를 작성했고 이 가격이 어떻게 떨어질지 경고했습니다
- **2순위:** 국내 6개 기관 중 한 곳은 대만의 시장조사기관인 트렌드플러스(Trend Plus)에서 D램 가격이 떨어질 수 있다는 이유로 목표주가를 내렸다.
- **3순위:** 메모리 업체들의 D램과 낸드 재고량은 상당히 낮은 수준에 머물 것



## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

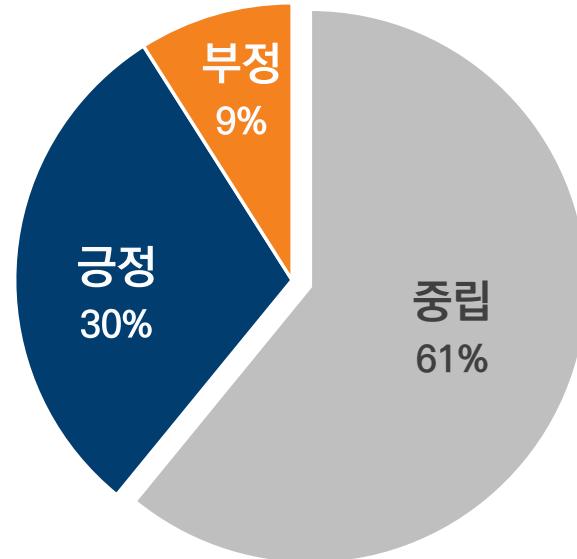


## 모델링 및 분석 결과

- 모델 및 훈련데이터셋
- 각데이터도메인분석
- 최종결론



## 해외 뉴스 분석



## 해외 뉴스 긍정 상위 3문장

- **1순위:** 한 고위 임원은 5G 롤아웃이 확대되고 업계 선두주자인 중국 화웨이 테크놀로지스가 국내 시장에 집중함에 따라, 삼성전자는 네트워크 장비 사업의 성장 모멘텀을 유지하기 위해 유럽에 기대를 걸고 있다고 말했다.
- **2순위:** 서버에 대한 DRAM 수요는 2018년과 2021년 사이에 17.5%의 연평균 성장률을 보일 것입니다.
- **3순위:** 브라질과 다른 신흥 경제국에서는 한때 식기세척기와 로봇청소기에 의존했던 가정들이 이제는 투자를 하고 있는 반면 삼성은 올 1월부터 7월까지 공기청정기 해외 판매가 작년 같은 기간에 비해 5배 이상 급증했다고 말한다.

## 해외 뉴스 부정 상위 3문장

- **1순위:** 지난 1월부터 3월까지보다 적은 스마트폰을 출하했음에도 불구하고 한국의 거대 기술기업인 하이닉스의 강력한 실적은 비축량을 고갈시키고 생산능력을 채운 칩에 대한 성종권 수요를 강조하고 있다.
- **2순위:** 그의 가석방에 대한 지지는 한국의 거대 기술기업에서 핵심 전략적 결정이 내려지지 않고 있다는 불안감 속에서 커져왔다.
- **3순위:** 여론조사가 이 대통령의 가석방에 대한 국민의 높은 지지를 보여준 반면, 많은 시민 단체들은 한국의 정계와 재계 엘리트들 사이의 부패에 대한 분노의 물결로 집권한 문재인 대통령의 위선을 비난하며 반대해 왔다.



## 데이터 수집 및 가공

- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

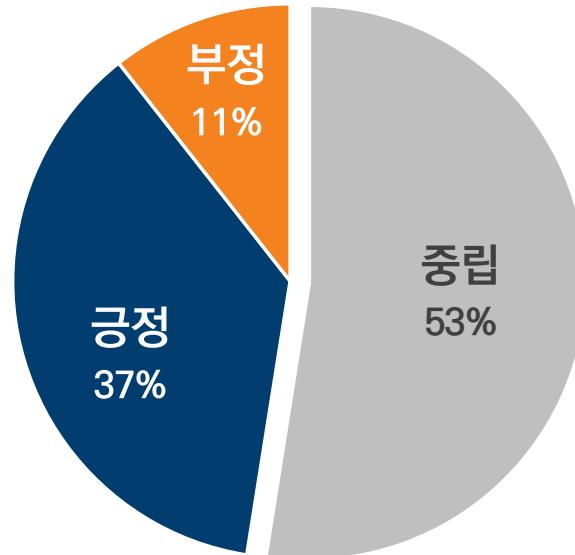


## 모델링 및 분석 결과

- 모델 및 훈련데이터셋
- 각데이터도메인분석
- 최종 결론



## 증권사 리포트 분석



## 증권사 리포트 긍정 상위 3문장

- **1순위:** 시스템반도체사업부도 오스틴 조기 정상화(단, 매출 감소에도 SCM 역량 강화, 마케팅 효율성 향상, 웨어러블 제품 기여 등으로 예상보다 높은 수익성 유지)로 실적이 개선됐다.
- **2순위:** 비트 출하량은 D램 15나노미터와 낸드 6세대 128단 변환의 가속화로 인해 증가할 것으로 예상된다.
- **3순위:** 파운드리사는 하반기 평택 S5의 공급능력을 확대하여 성장을 가속화할 계획이다. 신제품 출시를 통한 중소형 패널 성능 향상 기대

## 증권사 리포트 부정 상위 3문장

- **1순위:** 코스피지수(오른쪽,p)와 비교하면 중소형 비수기 영향으로 출하량은 줄었지만 전반적인 매매가격 상승과 일회성 이익으로 전분기보다는 호전됐다.
- **2순위:** 대형 LCD는 계절적 비수기, 부품 공급 부족, 코로나로 인한 인도와 베트남에서의 생산 차질 등의 영향을 받았는데, 이는 라디오 적자로 인한 스마트폰 수요 감소로 인한 것이다.
- **3순위:** 메모리 업체들의 D램과 낸드재고량은 상당히 낮은 수준에 머물 것



### 데이터 수집 및 가공

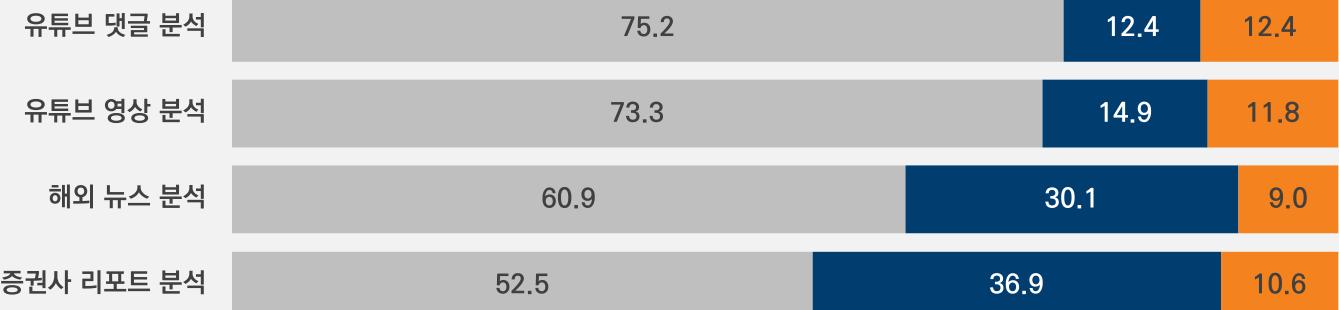
- 데이터수집기준
- 유튜브댓글
- 유튜브영상
- 해외 뉴스
- 증권사리포트

### 모델링 및 분석 결과

- 모델및훈련데이터셋
- 각데이터도메인분석
- 최종결론

#### [ 도메인 별 분석 ]

■ 중립 ■ 긍정 ■ 부정



삼성전자의 주가가 해당기간(21.08.14 ~ 21.09.17)에 큰 변동성을 가졌음에도 불구하고  
투자의견의 긍정 비율이 부정 비율보다 오히려 높음을 알 수 있습니다.  
특히, 증권사 리포트와 구글 뉴스의 경우 더 낙관적인 전망을 예상했습니다.

## ‘세상을 읽다’는 유튜브 영상, 유튜브 댓글, 해외 뉴스, 증권사 리포트

총 4가지 도메인에서 수집한 투자의견의 ①긍/부정 비중과 ②상위 긍/부정 근거를 제시해줍니다.

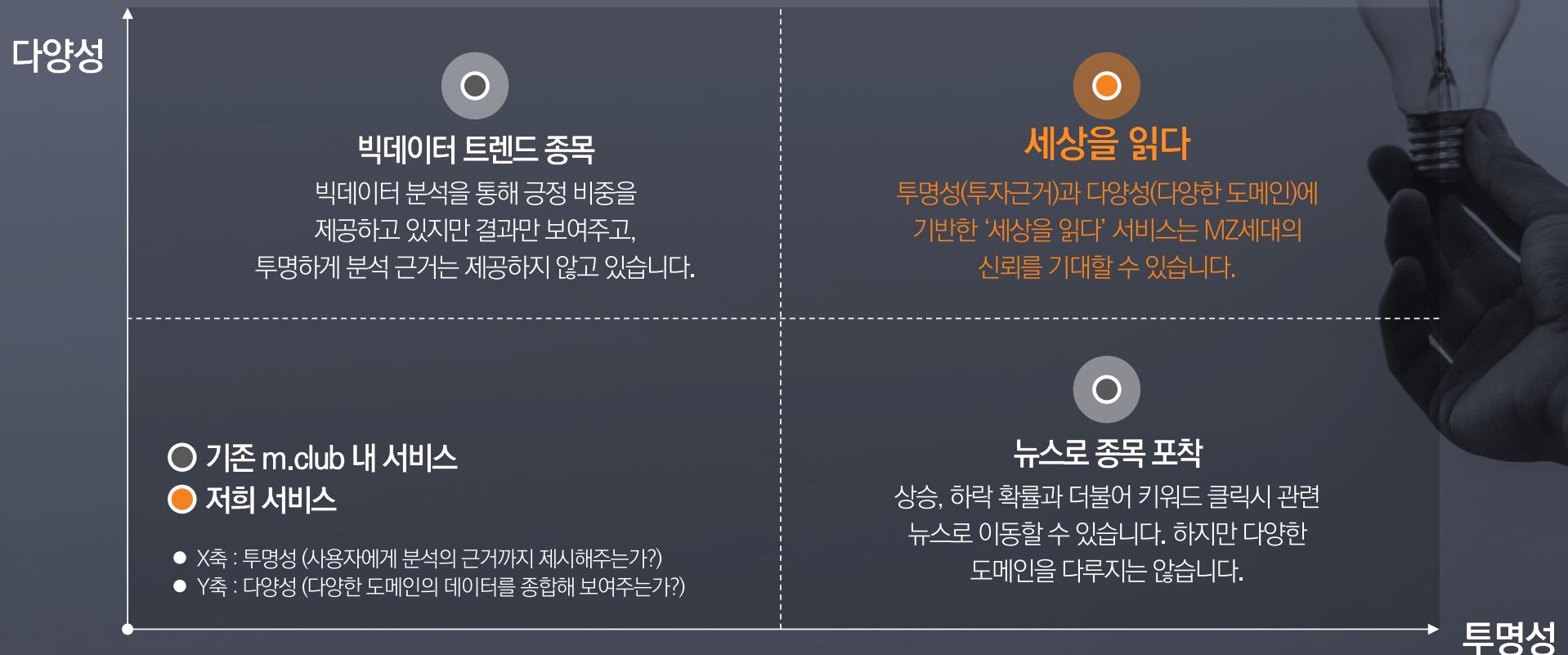


- ✓ ‘삼성전자’에 관심이 생긴 투자자는 유튜브와 네이버에 ‘삼성전자’를 검색하기 전, 저희 서비스를 먼저 방문해 종목에 대한 사회 분위기 및 경향을 쉽게 알게 될 것입니다.

즉, m.club이 주식계의  
‘실시간 검색어’, ‘플랫폼’이 될 수 있습니다.

- ✓ ‘매수’, ‘매도’ 아이콘을 추가해 플랫폼에서 실제 투자까지 가능하게 합니다.
- ✓ 단지 사라고 추천하는 것만이 아닌 투명하게 투자 근거를 제공합니다.

m.club은 이미 100 만명이 가입한 인기 있는 서비스입니다.  
 ‘세상을 읽다’가 추가되면 더 균형 잡힌 서비스가 될 것이라 생각합니다.



A dark, atmospheric photograph of a modern skyscraper at night. The building's glass facade reflects the surrounding city lights, and its name, "MIRAE ASSET", is prominently displayed in large, illuminated letters on the side. The sky above is filled with heavy, dark clouds.

저희의 아이디어가  
미래에셋증권의 빛이 되길 기원합니다.

감사합니다

팀 | MZ와 같이하자 가치투자

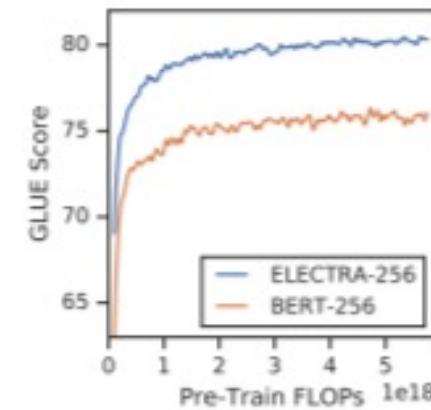
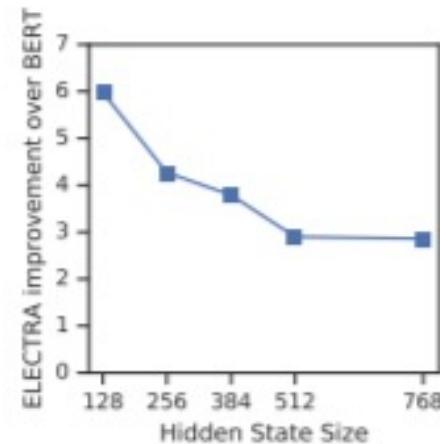
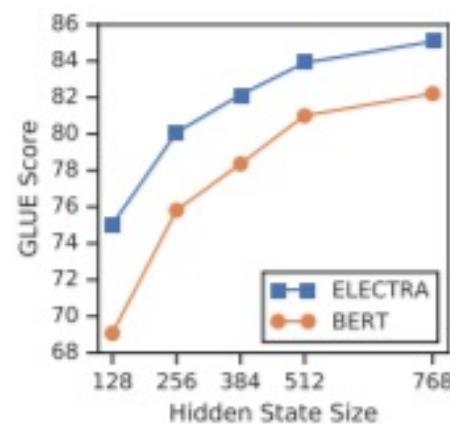
## 참조한 문서는 다음과 같습니다.

- ✓ [미래에셋대우, AI 금융서비스 '엠클립' 20만명 돌파] <https://m.etnews.com/20200122000327>
- ✓ [대학내일 20대 연구소] MZ세대 재테크 정보, 000에서 얻는다 <https://www.20slab.org/Archives/37599>
- ✓ Google Electra Model <https://arxiv.org/pdf/2003.10555.pdf>
- ✓ Sentiment Analysis on SST-2 Binary Classification Benchmark <https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary>
- ✓ GLUE Benchmark <https://gluebenchmark.com/leaderboard>
- ✓ Quora Benchmark <https://paperswithcode.com/sota/question-answering-on-quora-question-pairs>
- ✓ Naver Papago API <https://www.ncloud.com/product/aiService/papagoTranslation>
- ✓ Kaggle Dataset <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news>
- ✓ Sumy <https://github.com/miso-belica/sumy>
- ✓ Pygooglenews <https://github.com/kotartemiy/pygooglenews>
- ✓ Korean sentence splitter <https://github.com/likejazz/korean-sentence-splitter>
- ✓ spaCy <https://github.com/explosion/spaCy>
- ✓ Youtube API <https://github.com/youtube/api-samples>
- ✓ Tika <https://github.com/apache/tika>

- ✓ 유튜브 슈카월드 댓글 데이터 <https://www.youtube.com/c/%EC%8A%88%EC%B9%B4%EC%9B%94%EB%93%9C>
- ✓ 유튜브 돈깡의 알고하는 투자 댓글 데이터  
<https://www.youtube.com/c/%EB%8F%88%EC%9D%B4%EA%B9%A1%ED%8C%A8%EB%8B%B9>
- ✓ 유튜브 재테크 연구소 댓글 데이터 [https://www.youtube.com/channel/UCdSIR0\\_X9oGNP2hcDLW0okA](https://www.youtube.com/channel/UCdSIR0_X9oGNP2hcDLW0okA)
- ✓ 유튜브 체인지그라운드 댓글 및 영상 데이터  
<https://www.youtube.com/c/%EC%B2%B4%EC%9D%B8%EC%A7%80%EA%B7%B8%EB%9D%BC%EC%9A%B4%EB%93%9C>
- ✓ 유튜브 김작가TV 영상 데이터 <https://www.youtube.com/c/%EA%B9%80%EC%9E%91%EA%B0%80TV>
- ✓ 유튜브 Super K-슈퍼개미김정환 영상 데이터  
<https://www.youtube.com/c/SuperK%EC%8A%88%ED%8D%BC%EA%B0%9C%EB%AF%B8%EA%B9%80%EC%A0%95%ED%99%98>
- ✓ 유튜브 Scott's Sotck Due Diligence 영상 데이터 <https://www.youtube.com/user/Rubin16241>
- ✓ 유튜브 Cameroon Stewart, CFA 영상 데이터 <https://www.youtube.com/c/CameronStewartCFA>
- ✓ 키움증권 리포트(21.08.03) [http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report\\_idx=595494](http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report_idx=595494)
- ✓ 키움증권 리포트(21.09.14) [http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report\\_idx=597847](http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report_idx=597847)
- ✓ 유진투자증권 리포트 [http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report\\_idx=595239](http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report_idx=595239)
- ✓ 하이투자증권 리포트 [http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report\\_idx=598140](http://consensus.hankyung.com/apos.analysis/analysis.downpdf?report_idx=598140)
- ✓ 구글 미국 뉴스 [https://github.com/sw-membership/ji\\_wung/blob/master/Google\\_News\\_link\\_list.ipynb](https://github.com/sw-membership/ji_wung/blob/master/Google_News_link_list.ipynb)

## ELECTRA vs BERT

- Electra 모델은 일반적으로 사용되는 NLP 모델인 BERT과 비교하였을 때 GLUE 평가 지표에서 큰 성능 차이를 보입니다.
- 또한 Train 과정에서도 BERT에 비해 빠른 학습 능력을 가지고 있는 우수한 모델입니다.
- Electra 모델은 2020년에 공개 되었지만, 현재까지 나온 모든 모델과 비교하여도 높은 성능을 보이고 있으며, GLUE에서 제공하는 Benchmark에서 89.4의 Score를 기록하여 현재 11위에 랭크해 있습니다.



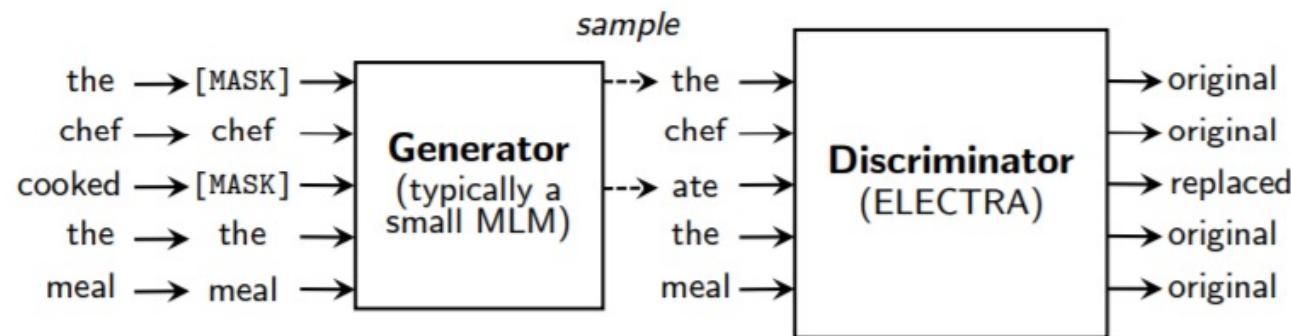
## ELECTRA Quora Benchmark

- Electra 모델은 다른 NLP Task에서도 높은 성능을 보이는 Generality가 우수한 모델입니다.
- 질문에 대한 답을 하는 능력을 평가하는 Quora Benchmark에서도 상위 지표를 나타내고 있습니다

Rank	Model	Accuracy↑	Paper
1	XLNet (single model)	92.3%	XLNet: Generalized Autoregressive Pretraining for Language Understanding
2	DeBERTa (large)	92.3%	DeBERTa: Decoding-enhanced BERT with Disentangled Attention
3	ALBERT	90.5%	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
4	T5-11B	90.4%	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
5	MLM+ subs+ del-span	90.3%	CLEAR: Contrastive Learning for Sentence Representation
6	RoBERTa	90.2%	RoBERTa: A Robustly Optimized BERT Pretraining Approach
7	ERNIE 2.0 Large	90.1%	ERNIE 2.0: A Continual Pre-training Framework for Language Understanding
8	ELECTRA	90.1%	ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

## ELECTRA GAN 구조

- Electra 모델은 GAN Based 모델이므로 Generator에서 생성한 새로운 형태의 데이터를 Discriminator를 거치면서 성능을 향상시킬 수 있습니다.
- Electra의 Generator에서는 특정 단어에 대해 Mask를 통해 다른 단어로 변경하는 방식을 사용합니다.
- 이와 같은 방법으로 Generator와 Discriminator가 zero-sum game을 하며 적은 데이터를 활용하여도 높은 성능을 보일 수 있습니다.



$$\begin{aligned} \mathcal{L}_{MLM}(\mathbf{x}, \theta_G) &= \mathbb{E} \left( \sum_{i \in \mathbf{m}} -\log p_G(x_i | \mathbf{x}^{masked}) \right) \\ &= \mathbb{E} \left( \sum_{t=1}^n -1(x_t^{corrupt} = x_t) \log D(\mathbf{x}^{corrupt}, t) - 1(x_t^{corrupt} \neq x_t) \log \right. \\ &\quad \left. (1 - D(\mathbf{x}^{corrupt}, t)) \right) \end{aligned}$$