

미래에셋증권 X NAVER Cloud 공동주관

제 8 회 2024 미래에셋증권
AI·Data
Festival

초보 투자자들을 위한 「맥주-기저귀」 관계의 테마 탐지와
재무제표 분석 및 증동 매매 방지 챗봇 서비스

팀 명 슈 퍼 통 개 미

팀 원 김 민 열
이 건 희
한 진 솔

소 속 동 국 대 학 교

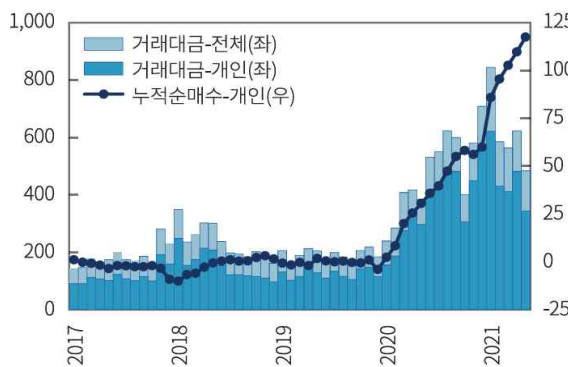
날 짜 2024-07-31

목 차

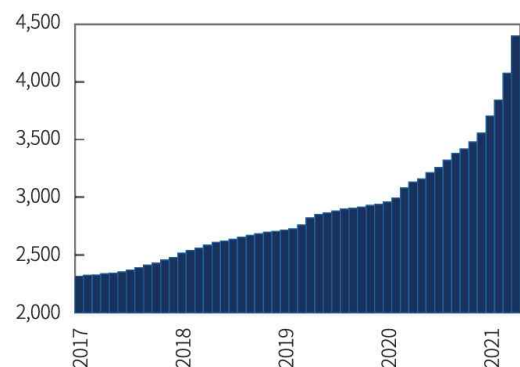
I . 연구 배경	01
II . 서비스 개요	02
2.1 실시간 테마 탐지 서비스 - ‘맥주와 기저귀’	02
2.2 재무제표 요약 및 챗봇 서비스 - ‘fin-bot’	03
2.3 감정적 투자자들을 위한 총동 매매 방지 서비스 - ‘COMMA’	04
III . 서비스 구현 과정	05
3.1 개발 파이프라인	05
3.2 실시간 테마 탐지 서비스 - ‘맥주와 기저귀’	06
3.2.1 이슈 데이터 수집	06
3.2.2 핫(HOT) 이슈 추출 및 본문 수집	07
3.2.3 기사 요약 및 테마 탐지	08
3.2.4 튜닝 데이터 생성 및 모델 튜닝	09
3.3 재무제표 분석 및 챗봇 서비스 - ‘fin-bot’	10
3.3.1 재무제표 데이터 수집	10
3.3.2 재무제표 분석	11
3.3.3 재무제표 세부 정보 대화형 챗봇	11
3.4 감정적 투자자들을 위한 총동 매매 방지 서비스 - ‘COMMA’	12
3.4.1 초보 투자자 클러스터링	12
3.4.2 챗봇 등장 조건	12
3.4.3 컴패니언쉽 요소의 활용	13
3.4.4 RAG 활용 개인화 투자 전략	13
IV . 결론 및 기대효과	14
V . 참고문헌	15

I. 연구 배경

[1]2020년 코로나19 팬데믹 직후 한국 주식시장에서 가장 주목되었던 현상 중 하나는 개인투자자들의 참여가 크게 확대되었다는 것이다. 개인투자자들은 2020년 3월부터 2021년 2월까지 1년간 87조원(KOSPI 69조원, KOSDAQ 18조원)을 순매수하였고 같은 기간 주식시장 활동 계좌수는 2,991만개에서 3,834만개로 843만개가 증가했다. 1년 남짓한 기간에 이 정도 규모의 신규 투자자금과 신규 투자자가 유입된 이른바 ‘동학개미운동’은 한국 주식시장에서 전례가 없었던 일이다. 거래도 크게 늘어, 이 기간 개인투자자의 월간 거래대금은 평균 404조원으로 2017~2019년 평균 130조원에 비해 약 3.1배 증가한 수준이다.



[그림 1] 개인 거래대금 및 누적순매수



[그림 2] 주식투자 활동계좌수

자료: DataGuide, 금융투자협회

개인투자자의 주식시장 참여 확대는 근본적으로 코로나19 충격 이후 경기회복에 대한 기대에서 출발했을 것으로 보인다. KOSPI지수는 코로나19의 전 세계적 확산과 함께 2020년 3월 중 1,400대까지 하락하였으나, 경기 하락 충격이 2020년 3분기부터 완화되기 시작하고 국내 상장기업 실적이 2020년 3분기에 전년 동기 수준을 회복하면서 빠르게 반등했다. 코로나19 충격에 따른 주식시장의 저평가와 이후의 신속한 반등은 장기적인 저금리 기조와 부동산 가격 상승으로 누적된 개인투자자의 잠재 투자수요를 자극했을 것으로 판단된다.

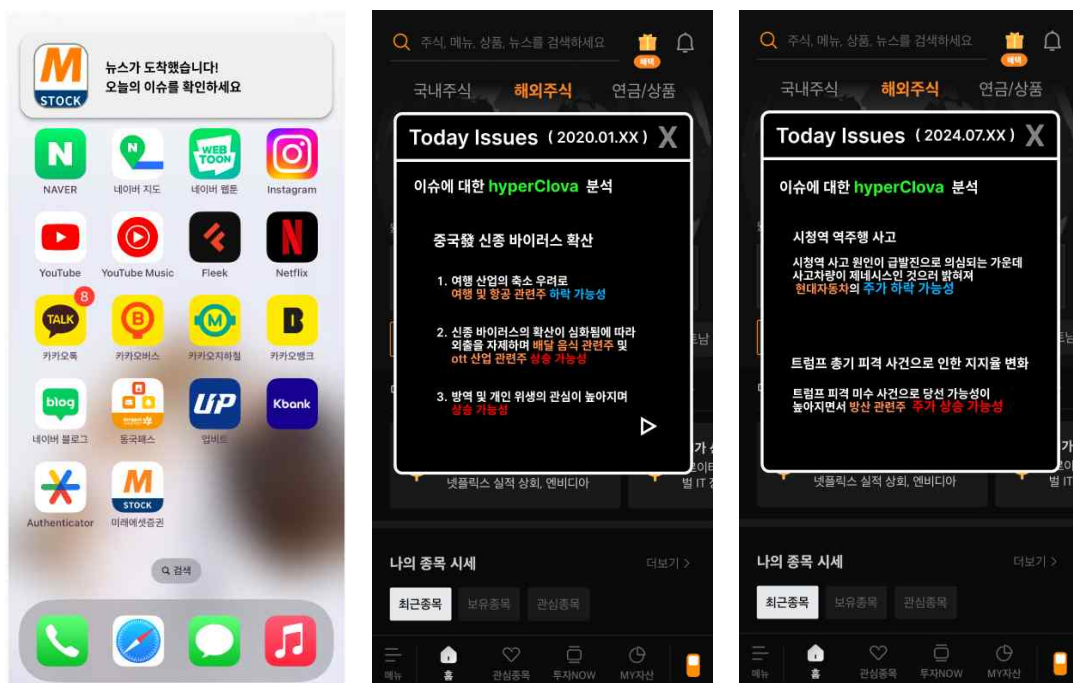
신규 투자자들의 대거 유입이라는 사회적 흐름에 편승하여 국내 주요 증권사들도 일명 ‘주린이’들을 주요 타겟으로 삼아 다양한 정책과 서비스를 출시하고 있다. [2]KB증권은 국내 금융권 최초로 챗GPT 기술을 기반으로 한 'KB증권 GPT'를 출시하였고 [3]토스증권에서 선보인 적립식 투자 서비스 ‘주식모으기’는 서비스 출시 2년만에 이용자 100만 명을 돌파했다. 그러나 이러한 서비스들은 대부분 정보 제공에 그치고 있어 매매 경험이 상대적으로 빈약한 초보 투자자들이 실제 투자 의사 결정 과정에서 체감할 수 있는 도움을 제공하는 데에는 한계가 있다. 따라서 본 프로젝트에서는 ‘주린이’들의 투자와 의사 결정에 실질적으로 도움을 주는 몇 가지 서비스들을 고안했다. 이를 통해 신규 투자자들이 보다 안정적이고 전략적인 투자를 할 수 있도록 지원할 예정이다.

II. 서비스 개요

2.1 실시간 테마 탐지 서비스 - ‘맥주와 기저귀’

본 프로젝트에서는 초보 투자자들의 의사 결정에 직접적으로 도움을 줄 수 있는 서비스 중 첫 번째로 실시간 테마 탐지 서비스 ‘맥주와 기저귀’를 구현했다. 한때 미국의 일부 대형 마트에서는 기저귀 옆에 맥주를 진열해 두었다. 이는 고객들의 장바구니 안에 어떤 것들이 ‘같이’ 담기는지에 관한 궁금증으로부터 시작되었는데, 고객들의 구매 내역을 분석한 결과 전혀 어울리지 않는 두 상품인 맥주와 기저귀를 함께 구매하는 고객이 상당히 많다는 사실을 알아냈다. 이는 남편들이 퇴근길에 아내의 심부름으로 마트에 들러 기저귀를 사면서 집에 가서 마실 맥주도 함께 사기 때문일 것으로 추정했다.

주식시장에서 맥주와 기저귀 관계에 놓인 이슈와 테마를 파악할 수 있는 능력은 매우 중요하다. 코로나19 팬데믹이 처음 발병했을 당시를 예로 들어 보자. 정체불명의 전염병이 유행하고 사회적 불안감이 고조되면서 마스크나 손소독제와 같은 개인 위생 용품의 수요가 급증할 것이고, 이를 생산하는 회사의 매출 증대에 관한 기대감으로 관련주의 주가가 상승할 것은 누구나 예측할 수 있다. 그러나 코로나19의 장기화로 외출을 자제하는 사람들이 늘어나고 집에서 보내는 시간이 많아지며 넷플릭스 등의 OTT 서비스 가입자 수가 증가할 것이라는 판단 하에 관련 주식을 매수할 수 있는 사람은 많지 않다. 뉴스와 주식에 관한 안목이 부족한 초보 투자자들의 경우 더욱 그러하다.



[그림 3] 실시간 테마 탐지 서비스 - ‘맥주와 기저귀’

실시간 테마 탐지 서비스 ‘맥주와 기저귀’는 전일에 보도된 뉴스 기사들 가운데 조회수와 댓글, 그리고 해당 이슈로 작성된 기사의 수 등을 종합적으로 고려하여 하루 동안 가장 이슈가 되었던 뉴스를 파악한다. 그 후 테마 탐지에 유리하도록 튜닝된 CLOVA 모델을 사용해 해당 이슈로 직접적으로 또는 간접적으로 영향을 받을 만한 주식 테마나 종목을 선정하고 매일 개장 직전 알림과 팝업 창을 통해 제공한다. 이때 중요한 것은 해당 이슈가 영향을 줄 것으로 누구나 쉽게 예상할 수 있는 테마 뿐만 아니라 맥주와 기저귀 관계에 있는, 해당 이슈와 숨겨진 상관 관계를 갖는 테마를 빠르고 정확히 탐지해 제공하는 것이다. 이를 통해 초보 투자자들의 의사 결정에 도움을 주고 뉴스와 주식의 상호작용에 관한 이해도를 높이며, 상황에 따라 유리한 종목을 선정하는 안목을 높일 수 있다.

2.2 재무제표 분석 및 챗봇 서비스 - ‘fin-bot’

현대 기업 환경에서 재무제표는 기업의 성적표라고 할 수 있으며, 재무 상태와 성과를 평가하는데 필수적인 도구로 활용된다. 재무제표는 기업의 자산, 부채, 자본, 수익, 비용 등을 종합적으로 나타내는데 이는 투자자와 경영진, 채권자를 비롯하여 다양한 이해관계자가 기업의 재무 상태를 평가하고 전략적 결정을 내리는 데 중요한 자료로 활용된다. 예시로 최근 대규모 정산 지연사태로 연쇄 부도의 가능성까지 제기되고 있는 티몬의 재무제표로부터 지속적인 영업손실과 매출 감소, 높은 부채 수준, 마이너스 이익률, 낮은 투자 수익률 등을 확인할 수 있다. 이러한 지표들은 회사가 운영을 지속하는데 어려움을 겪고 있으며 파산 위험이 높다는 것을 시사하고 있었다.

이같은 사례는 투자에 관한 의사결정을 내릴 때 재무제표를 통해 해당 기업의 재무상태를 우선적으로 확인하고 위험성을 판단하는 것이 중요함을 의미한다. 그러나 재무제표는 복잡하고 전문적인 용어와 개념을 포함하고 있기 때문에 재무 관련 지식이 부족한 일반인이나 초보 투자자들이 이해하기에 어려움을 겪고 있다.

본 프로젝트에서는 초보 투자자들의 의사 결정에 직접적으로 도움을 줄 수 있는 두 번째 서비스로 재무제표 분석 및 챗봇 서비스 ‘fin-bot’을 제안한다. 기존 미래에셋증권 MTS의 종목별 재무제표 페이지에서 제공하고 있는 시각화와 함께 종목 투자 전 재무제표에서 반드시 확인해야 할 항목들을 5가지로 나누어 비전문가가 알기 쉽게 요약한다. ‘fin-bot’이 중요하게 판단하는 5가지 관점은 [4]관련 논문을 참고해 설정하였으며 그 내용은 다음과 같다.

수익성	영업 이익 및 당기 순이익의 변화를 바탕으로 회사의 자금 운용 능력을 판단.
안정성	자본 대비 부채 비율이 변화를 바탕으로 채무 상황을 판단.
성장성	매출액의 변화와 현금 흐름을 바탕으로 회사의 성장 가능성을 판단.
유동성	유동 자산과 유동 부채의 변화를 바탕으로 유동 비율을 판단.
생산성	매출 총 이익률과 발생하는 비용의 변화를 바탕으로 생산성을 판단.

[표 1] fin-bot의 5가지 관점

이러한 방식으로 ‘fin-bot’은 초보 투자자들의 재무제표에 관한 기본적 이해와 투자 판단을 돕는다. 또한 투자자들은 세부 재무지표나 용어들에 관해 추가적으로 궁금한 점을 ‘fin-bot’에게 질문함으로써 깊이 있는 정보를 얻을 수 있다.



[그림 4] 재무제표 분석 및 챗봇 서비스 - ‘fin-bot’

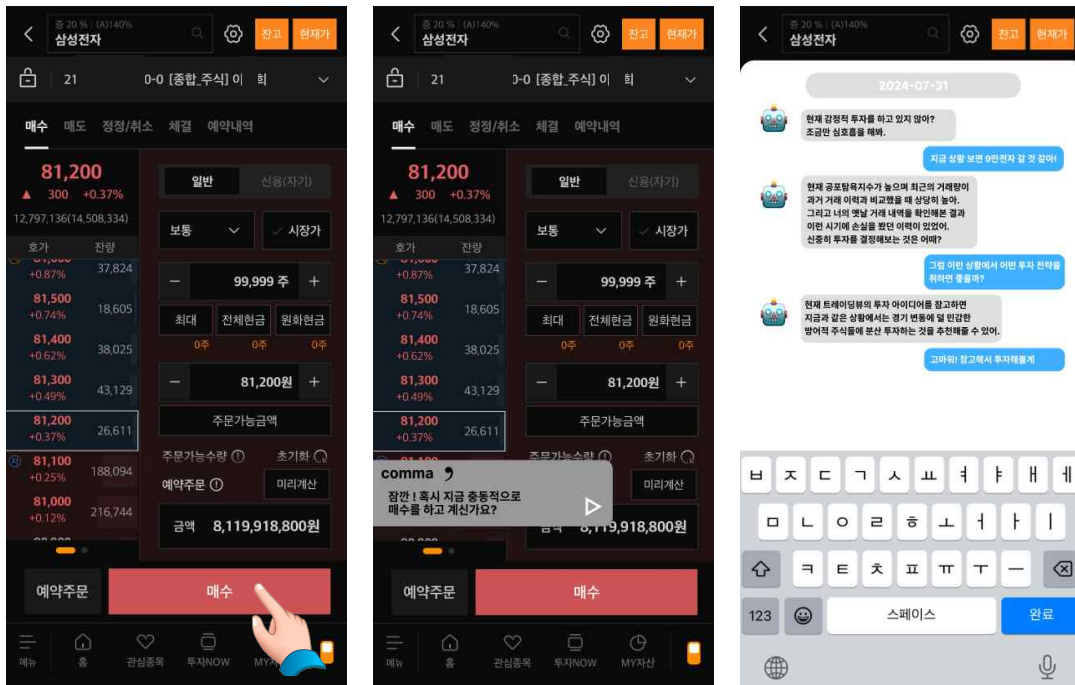
2.3 감정적 투자자들의 충동 매매 방지 서비스 - ‘COMMA’

감정은 투자 과정에서 중요한 역할을 한다. 강세장에서 기회를 놓치는 것에 대한 두려움이나 약세장의 패닉이든 투자자들은 시장 정서에 사로잡히기 쉽다. 감정적 투자는 성급한 결정으로 이어질 수 있으며, 이는 포트폴리오에 부정적인 영향을 미칠 수 있다. [5] 투자에 익숙하지 않은 초보 투자자일수록 자신의 감정에 따라 투자 결정을 내리는 경향이 있으며, 이는 잘못된 투자 선택으로 이어질 수 있다. 이들은 또한 과거 손실 등의 투자 결과에 대한 잔존감정으로 인해 공격적인 투자를 감행하는 등 비합리적인 의사결정을 반복하게 된다. 이는 매매 횟수는 증가하지만 수익률이 감소하는 결과를 초래할 수 있다.

초보 투자자들을 위한 충동 매매 방지 서비스 ‘COMMA’는 감정적 투자를 예방하고, 논리적인 투자 결정을 지원하기 위해 고안된 서비스이다. 현재 투자자가 감정적이라고 판단되면 주식 매수 및 매도 시 챗봇이 출현한다. 심호흡을 하면 마음이 진정되듯, 투자자들에게 한 번 더 생각할 수 있는 시간과 함께 도움이 되는 정보를 제공한다. 그러나 챗봇이 모든 투자자에게 매번 등장하면 사용자들이 불만을 제기하거나 불편을 야기할 수 있다. 그러므로 고객 데이터를 활용하여 군집화를 진행했고, 초보 투자자에 해당하는 군집을 대상으로 제공하기로 했다. 군집화를 통해 결정된 초보 투자자들 중에서도 서비스를 이용하길 원치 않는 사람이 있을 수 있으므로, 서비스 사용 여부를 물어본 후, 사용하는 초보 투자자들을 대상으로만 서비스를 제공할 예정이다. 그러므로 시장의 분위기를 알 수 있는 ‘공포-탐욕지수’

를 활용하거나, 투자자의 평균 매매 횟수와 비교함으로써 투자자에게 가장 도움을 줄 수 있는 최적의 출현 타이밍을 결정해보았다.

COMMA의 가장 큰 특징은 개인화되었다는 것이다. 고객의 과거 데이터를 학습함으로써 이를 기반으로 다양한 상황에 대해 과거의 실수를 반복하지 않는 방향으로 적절히 대응할 수 있게 한다. 특히 친근한 언어와 예사말을 사용함으로써 챗봇을 친밀하고 신뢰할 수 있는 대상으로 인식하게 하여 부정 감정을 낮추고, 자기통제감을 높일 수 있는 컴패니언십(companionship) 방식을 적용했다. 이로써 고객들이 냉철하고 이성적인 판단을 할 수 있도록 도우며 이를 통해 고객들은 안정적이고 성공적인 투자 경험을 쌓을 수 있게 된다.



[그림 5] 감성적 투자자들의 총동 매매 방지 서비스 - 'COMMA'

Ⅲ. 서비스 구현 과정

3.1 개발 파이프라인

[그림 5]는 본 프로젝트의 전체적인 흐름을 간략히 정리한 파이프라인이다. 본 장에 수록한 서비스 구현 과정을 파이프라인과 함께 읽어나가기를 추천한다.



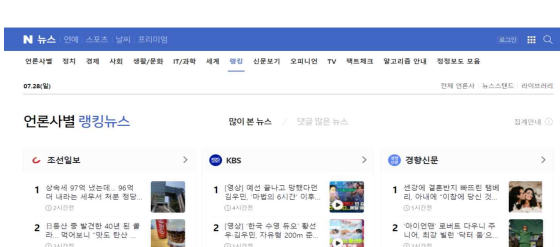
[그림 6] 개발 파이프라인

3.2 실시간 테마 탐지 서비스 - '맥주와 기저귀'

3.2.1 이슈 데이터 수집

먼저, 이슈에 관해 직·간접적으로 연관된 테마를 적절히 추출하는 CLOVA 모델을 튜닝할 데이터가 필요하다. AI 모델의 고질적 문제인 할루시네이션(hallucination)을 최소화하고 답변 형식을 고정하기 위해 모델 튜닝이 필요하다. 그러나 특정 이슈에 관해 영향을 받았던 테마 데이터를 직접 수집하기에 무리가 있었다. 본 프로젝트는 서비스 제안에 목적을 두므로 GPT 모델을 사용해 가상의 튜닝 데이터를 생성하기로 한다. 향후 실제 서비스 상용화 과정에서 모델의 튜닝 데이터는 미래에셋증권에서 생성해야 할 것으로 보인다.

튜닝 데이터를 만들기 위한 첫 단계는 특정 테마나 종목에 충분한 영향을 미칠 만한 핫이슈를 선정하는 것이다. 먼저, 2023년 7월 1일부터 2024년 7월 1일까지 네이버 뉴스의 언론사별 랭킹 뉴스에서 '많이 본 뉴스'와 '댓글 많은 뉴스'의 헤드라인을 일별로 크롤링했다. 이는 하루마다 쏟아지는 수천 개의 기사들 중 가장 이슈가 되었던 뉴스를 선정할 후보 데이터로, 수집된 데이터의 크기는 총 294,322개다.



[그림 7] 네이버 뉴스 언론사별 랭킹 홈

	Date	Title
0	20230701	'제로 콜라' 자주 마셨는데 어찌나...인공감미료 '아스파탐' 발암물질 분류 예정
1	20230701	태슬라 2분기 판매량 공개...주가 상승 모멘텀 강화될까
2	20230701	후쿠시마 오염수 방류 불안감에...국내산 '육상 양식 연어'에 불리는 눈
...
294319	20240701	[단독]'손흥정 고소' 학부모 '혹독 훈련' 동의 안 구해...손흥준도 육설
294320	20240701	[속보]서울 시청역 앞 차량 돌진 "대표저렴 큰 소리"...경찰 "70대 운전자, 급발...
294321	20240701	[단독] 머리 때리고 넘어뜨리고...어린이집 3살 반에서 벌어진 일
...

294322 rows × 2 columns

[그림 8] 23.07~24.07 '많이 본 뉴스' 및 '댓글 많은 뉴스' 헤드라인

3.2.2 핫(Hot) 이슈 추출 및 본문 수집

다음으로 크롤링한 뉴스 헤드라인 중 하루 동안 가장 많이 다루어진 이슈의 키워드를 3개씩 추출했다. 특정 테마나 종목에 영향을 미칠 만한 핫 이슈는 단순히 기사의 조회수나 댓글수를 기준으로 판단하기에는 부족하기 때문이다. 따라서 하루 동안 동일한 이슈가 반복적으로 작성된 기사의 수를 고려해 일별 핫 이슈 TOP3의 키워드를 추출했다. 이 때, 모델은 GPT-4o 모델을 사용했다. CLOVA-X가 아닌 GPT 모델을 사용한 이유는 프롬프트에 입력할 헤드라인의 수가 일별로 대략 800개 정도인데, CLOVA-X 모델에서는 토큰 수 제한으로 한 번에 입력하지 못하기 때문이다. 프롬프트 엔지니어링에서 고려한 부분은 다음과 같다. 먼저, CLOVA-X와 달리 GPT 모델의 경우 한국어보다 영어를 입력으로 주었을 때 답변의 성능이 더 높으므로 기본 설명은 영어로 제공한다. 모델의 역할을 정확히 기술하고 구체적인 답변 예시를 세 가지 제시함으로써 모델의 일관적인 답변을 유도한다. GPT 모델에 입력한 프롬프트는 다음과 같다.

```
## Role
I'm going to give you a txt file, which is a concatenated string of the titles of the news stories from the day, separated by tabs. Read through the titles of the news stories and pick out the three most common issues, regardless of field. Your job is to pick three issues. This is a Korean article, so please answer in Korean.
Tell me three specific incidents. And we don't need a description of the incident, just a keyword of incidents.

## Example 1
1. 도널드 트럼프 총격 미수 사건
2. 백종원과 연돈볼카츠 논란
3. 로또 당첨자 수 증가
...
```

다음으로 각 이슈가 발생한 날짜를 기준으로 해당 이슈의 키워드를 검색해 상세한 뉴스 본문을 크롤링했다. 이로써 처음 수집되었던 294,322개의 데이터 중 1,101개의 일별 핫 이슈와 기사 본문을 데이터가 생성되었다. 다음은 2023년 7월 1일부터 2024년 7월 1일까지 일별 핫 이슈 세 가지와 그 본문을 수집한 데이터프레임이다.

	Date	Title	Contents
0	20230701	정유정 살인 사건	신상공개 이후 '놀잇감' 된 정유정 얼굴생성형 AI로 왜곡하고 "예쁘다""처벌은 하...
1	20230701	일본 오염수 방류 논란	김영주 국회의원 photo 뉴시스더불어민주당이 후쿠시마 오염수 방류에 반대를 촉구...
2	20230701	80대 공연 원로 구속	법원, 구속영장 발부..."증거인멸 우려"가해자, 혐의 부인안산단원경찰서. /사진=연합..
...
1098	20240701	손웅정 감독 고소 사건	손웅정 SON축구아카데미 감독 / 사진=연합뉴스아동학대 논란이 불거진 손웅정 'SO...
1099	20240701	서울 시청역 차량 돌진 사고	1일 밤 대형 교통사고가 발생한 서울 시청역 인근 교차로에서 경찰이 완전히 파괴된 ...
1100	20240701	윤석열 대통령 탄핵 청원	이태원 참사 음모론 논란 탄핵 여론 불지퍼정청래 등 강경파 탄핵 시사 발언 잇따라민...

1101 rows × 3 columns

[그림 9] 23.07~24.07 일별 핫 이슈 TOP3

3.2.3 기사 요약 및 테마 탐지

네이버에서 제공하는 CLOVA Summary API를 활용해 앞 절에서 수집한 핫 이슈의 기사 본문을 세 문장으로 요약했다. 그리고 요약된 기사를 GPT 모델에 입력해 이슈와 연관된 테마를 탐지함으로써 CLOVA 튜닝 데이터를 생성하는데, 이 때 기사 원문을 직접 입력하지 않고 요약한 데는 두 가지 이유가 있다. 첫 번째로 프롬프트에 입력되는 토큰 수를 줄여 발생하는 비용을 절감하기 위함이다. 두 번째로는 기사의 핵심 내용만 입력해 모델이 더 정확히 테마를 탐지할 수 있게 하기 위함이다.

테마를 탐지하는 모델은 GPT-4o-mini 모델을 사용했다. CLOVA 모델의 튜닝 데이터를 생성하는 데 CLOVA 모델을 사용하는 것은 오류가 있다고 판단했기 때문이다. 3.2.2절에서 설명한 바와 같이 입력 프롬프트는 영어로 제공한다. 모델이 테마를 탐지하는 <기준>을 명확히 제시하고 <형식>을 지정해 일관된 답변을 유도했다. 모델의 역할은 요약된 기사를 입력으로 받아 해당 이슈와 직·간접적으로 관련된 종목 테마를 출력하는 것이다. 이 때, 출력하는 테마는 네이버페이 증권에서 크롤링한 테마 분류 데이터를 활용했다. 아래는 테마 탐지를 위해 GPT모델에 입력한 프롬프트이다. GPT 모델이 생성한 테마 탐지 결과를 열로 추가한 데이터프레임은 [그림 10]과 같다.

	Date	Title	Contents	Recommend
0	20230701	정유정 살인 사건	신상공개 이후 '놀잇감' 된 정유정 얼굴생성형 AI로 왜곡하고 "예쁘다""처벌은 하...	sector : 지능형로봇/인공지능(AI)Wnimpact : 긍정Wnreason : ...
1	20230701	일본 오염수 방류 논란	김영주 국회부의장 photo 누시스더불어민주당이 후쿠시마 오염수 방류에 반대를 촉구...	sector : '수산'Wnimpact : 부정Wnreason : 후쿠시마 오염수 ...
2	20230701	80대 공연 원로 구숙	법원, 구속영장 발부..."증거인멸 우려"가해자, 혐의 부인안산단원경찰서, /사진=연합...	sector : [엔터테인먼트]Wnimpact : [부정]Wnreason : 이변 ...
...
1098	20240701	손용정 감독 고소 사건	손용정 SON죽구아카데미 감독 / 사진=연합뉴스아동학대 논란이 불거진 손용정 'SO...	sector : [교육/온라인 교육] Wnimpact : [부정] Wnreason : ...
1099	20240701	서울 시청역 차량 돌진 사고	1일 밤 대형 교통사고가 발생한 서울 시청역 인근 교차로에서 경찰이 완전히 파괴된 ...	sector : [자동차 대표주]Wnimpact : [부정]Wnreason : 이변...
1100	20240701	윤석열 대통령 탄핵 청원	이태원 참사 음모론 논란 탄핵 여론 불지퍼정청래 등 강경파 탄핵 시사 발언 잇따라민...	sector : [정치 관련 주식]Wnimpact : [부정]Wnreason : 이...

1101 rows × 4 columns

[그림 10] 일별 핫 이슈에 대한 테마 탐지 결과

```
# system prompt
You're an ai assistant who reads a news article and detects the stock sector it's going to influence

# human prompt
{content} is news that has a lot of views and comments and deals with issues of great interest.
Based on the following <Criteria>, please present the stock theme that this news will affect and why
according to <Form>.
As for the theme, just present the theme in the theme list <Theme list>.
Please answer in Korean.
<Criteria>
First, read the news article and select the stock sector that you can influence and let me know the
impact and why.
Please recommend not only the sectors that the article directly affects, but also those that can
indirectly affect.
However, in this case, you need to recommend a sector with a clear causal relationship.
```

For example, the more time people spend at home during the COVID-19 pandemic has led to a significant increase in revenue for the alt platform, and shares in related sectors have risen.
Please present only the sectors that you expect to have a definite impact on.
Please explain in detail why it affects the sector.
If there are multiple sectors that can have an impact, please present them all.

<Form>

sector : [주식 섹터]

impact : [긍정] or [부정]

reason : [이유]

<Theme list>

(생략)

3.2.4 튜닝 데이터 생성 및 모델 튜닝

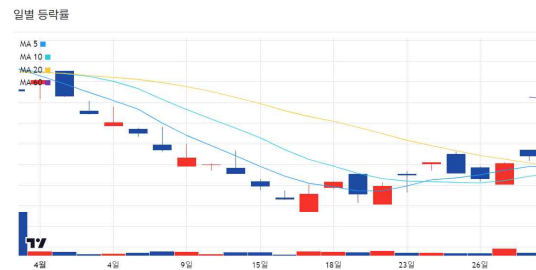
앞서 언급했듯 튜닝 데이터는 모델의 할루시네이션을 줄이기 위해 정확한 내용이 들어 가야 하며 실제 서비스 상용화에는 양질의 이슈별 테마 데이터를 수집하는 과정이 필수적이다. 그러나 본 프로젝트에서는 실제 데이터 수집의 어려움으로 가상의 데이터를 생성했기 때문에 모델이 탐지한 테마가 실제 이슈에 그러한 영향을 주었는지 판단하는 과정이 필요하다. 따라서 각 이슈별로 추출된 테마가 실제 해당 이슈의 발생 시점에 움직였던 방향을 직접 확인함으로써 튜닝 데이터의 품질을 높였다.

예시로, 2024년 4월 3일 추출된 HOT 이슈 중 하나는 농축산물이 역대 최고 상승률을 기록했다는 내용이었다. 위 이슈에 관해 탐지된 첫 번째 테마는 ‘농업’ 테마로, 농축산물의 가격이 급등하면서 시장의 수요가 줄어들어 농업 테마주의 하락을 예측했다. [그림 11]은 2024년 4월 농업 테마주의 등락률을 시각화한 것으로, 해당 이슈가 발생한 이후 테마주 내 종목들의 주가가 급락한 사실을 확인할 수 있다. 이 경우 모델의 예측이 참이므로 튜닝 데이터에 포함한다.

같은 이슈로 탐지된 또 다른 테마는 ‘비료’ 테마로, 농축산물 공급 문제를 완화하기 위해 비료 사용이 늘어남에 따라 비료 테마주의 상승을 예측했다. 그러나 [그림 12]를 통해 실제 2024년 4월 비료 테마주의 등락률을 확인해 본 결과 테마주 내 종목들의 주가가 하락했음을 알 수 있다. 이는 모델의 예측이 실제 결과와 다른 예로, 과도한 추론으로 간주해 튜닝 데이터에서 제외한다.



[그림 11] 24.04 농업 테마주 등락률



[그림 12] 24.04 비료 테마주 등락률

위 과정을 통해 핫 이슈별 실제 영향을 받은 테마 데이터를 최종 생성하고, 이를 활용해 CLOVA-X 모델을 튜닝했다. 튜닝된 CLOVA-X 모델을 활용해 실제 이슈와 연관된 테마를 탐지한 결과는 다음과 같다. 입력으로 사용한 이슈는 최근 발생한 미국 대선 후보 트럼

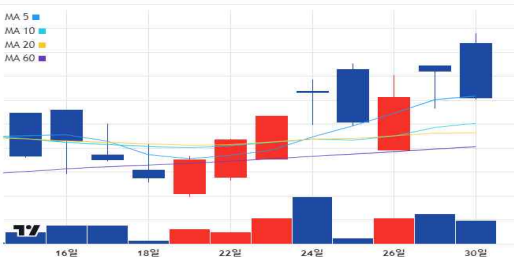
프 피격 사건으로, 해당 이슈에 관련된 두 가지 테마를 탐지한 것을 확인할 수 있다. [그림 14]와 [그림 15]를 통해 실제 피격 사건 이후 방산 테마와 제약바이오 테마의 주가 흐름이 모델이 예측한 결과와 일치한 것을 확인할 수 있다.

뉴스 기사 입력 : 11월 미국 대선을 3개월여 앞둔 13일 오후(현지시간) 공화당 대선 후보인 도널드 트럼프 전 대통령의 야외 유세 도중에 그를 겨냥한 총격 사건이 발생했다. 트럼프 전 대통령은 귀에 피를 흘리며 경호원에 둘러싸인 가운데 긴급히 대피했으며 다행히 총알이 스쳐 지나가 무사한 것으로 확인됐다. 총격범은 현장에서 사망했으며 유세를 지켜보던 지지자 1명도 희생된 것으로 알려졌다. (중략) 트럼프 전 대통령은 곧바로 차를 타고 유세장을 빠져나갔다.

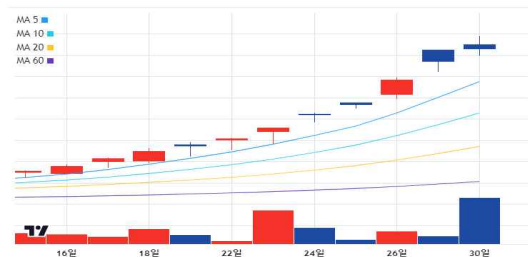
sector : 방위산업
 impact : 긍정
 reason : 트럼프는 과거 펼친 정책 중 미국의 군사력을 강화하는 데에 많은 예산을 사용했고, 국방 예산을 대폭 증가시킨 적이 있기 때문에 당선 가능성이 오름에 따라 방위산업이 오를 것으로 예측됩니다.

sector : 제약업체
 impact : 긍정
 reason : 이전 FDA 신약 승인 절차 간소화를 진행해 신약 개발과 출시가 빠르게 이뤄져 바이오 제약 회사들에게 긍정적인 영향을 미쳤습니다. 이를 바탕으로 바이오/제약 관련주의 주가가 상승할 것으로 예상됩니다.

[그림 13] 튜닝된 모델을 활용한 결과 예시



[그림 14] 이슈 발생 후 방산 테마주 등락률



[그림 15] 이슈 발생 후 제약-바이오 테마주 등락률

3.3 재무제표 분석 및 챗봇 서비스 - ‘fin-bot’

3.3.1 재무제표 데이터 수집

먼저, 기업의 재무제표 데이터를 수집하는 과정이다. 각 종목의 재무제표 정보는 CompanyGuide에서 크롤링을 통해 수집했다. 각 종목별로 2021년부터 2024년 1분기까지 기록되어 있는 포괄손익 계산서, 재무상태표, 현금흐름표를 크롤링하고 SQLite를 활용해 데이터베이스를 구축했다. 이후 요청에 따라 데이터베이스에 저장된 재무제표 데이터를 json 형식으로 출력 받아 CLOVA-X에 바로 데이터를 적용할 수 있도록 구현했다. 다음은 각 종목별 재무제표를 저장한 데이터베이스의 일부 예시이다.

	code	ITEM	202112	202212	202312	202403
			Filter...	Filter...	Filter...	Filter...
1	A003490	매출액	90168	140961	161118	42914
2	A003490	매출원가	69120	102454	128503	33604
3	A003490	매출총이익	21048	38507	32615	9311
4	A003490	판매비와관리비계산	6868	10201	14714	3935
5	A003490	영업이익	14180	28306	17901	5375
6	A003490	영업이익(발표기준)	14180	28306	17901	5375

[그림 16] 각 종목별 재무제표 데이터베이스

3.3.2 재무제표 분석

[6]CLOVA-X 모델이 입력받은 재무제표 데이터를 초보 투자자가 알기 쉽게 분석하도록 다음과 같은 프롬프트를 설정했다.

```
{ "role": "system", "content": "당신은 재무 용어에 익숙하지 않은 사람들이 재무제표를 쉽게 이해할 수 있도록 돕는 재무 분석 봇입니다. 사용자가 제공하는 재무제표 데이터를 바탕으로 수익성, 안정성, 성장성, 유동성, 생산성을 분석하고, 이를 바탕으로 투자자들에게 균형 잡힌 평가를 제공하세요." },
{ "role": "system", "content": "다음 지침을 따르세요:" },
{ "role": "system", "content": "1. 간단하고 직관적인 언어를 사용하세요. 기술 용어는 피하고, 필요할 경우 해당 용어를 항상 설명해 주세요." },
{ "role": "system", "content": "2. 수익성, 안정성, 성장성, 유동성, 생산성 측면에서 나눠서 단계적으로 분석하세요." },
{ "role": "system", "content": "3. 데이터의 주요 변화와 트렌드를 강조하세요. 예를 들어 매출이나 이익의 증가 또는 감소를 설명하세요." },
{ "role": "system", "content": "4. 중립적이고 객관적인 어조를 유지하세요. 편견 없는 분석을 제공하세요." },
{ "role": "system", "content": "5. 기업에 대한 긍정적 평가와 부정적 평가를 모두 포함하세요. 예를 들어, 긍정적인 측면에서는 강점을 강조하고, 부정적인 측면에서는 잠재적인 위험 요소나 미래의 불확실성을 설명하세요." }
```

먼저, 프롬프트의 구조와 명확성을 확보하기 위해 타겟층을 명시하여 모델이 누구를 대상으로 답변을 제공해야 하는지 이해할 수 있도록 한다. 또한, 재무제표를 통해 파악할 수 있는 중요한 정보를 놓치지 않고 구체적인 답변을 제공하기 주요 지표를 제시한다. 이 다섯 가지 지표는 전통적인 재무제표 분석에서 중요하게 여겨지며 기업의 경영성과 재무 상태를 평가하고 미래 상황을 예측하는 데 중요한 역할을 한다. [7]마지막으로, 분석한 내용을 기반으로 기업의 전망에 대한 종합 평가를 제공하도록 한다. 그러나 재무제표 분석 결과만으로 기업의 경영성과 미래를 완벽하게 평가하기에는 한계가 있기 때문에, 객관적 지표를 바탕으로 개별 항목을 판단하되, 기업에 대한 종합적 평가는 한쪽으로 치우치지 않게 균형적으로 평가하도록 했다. 이러한 원칙들을 바탕으로 프롬프트를 설계하고, 다양한 테스트를 통해 프롬프트의 효과를 검증하며 지속적으로 개선해 나가, 사용자에게 더 정확하고 유용한 정보를 제공하는 재무제표 분석 서비스를 구현했다.

3.3.3 재무제표 세부 정보 대화형 챗봇

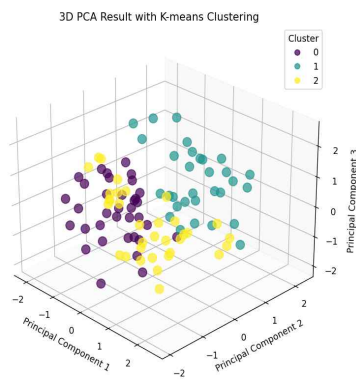
위의 재무제표를 분석한 내용에 관해 고객이 이해를 하지 못하거나 모르는 용어가 있을 수 있는 상황을 대비하여 관련 내용을 전문적으로 설명하는 챗봇을 구현했다. 우선 챗봇의 역할과 목표를 명확히 정의함으로써 재무 데이터 분석에 대한 전문적인 도움을 줄 수 있게 했다. 또한, 사용자의 질문에 대해 명확하고 자세한 답변을 제공해주기 위해 정확한 지침을 줌으로써 도움이 되도록 답변했다. 이를 통해 챗봇은 사용자의 특정 질문이나 이해 부족을 즉시 해결하여 재무제표 이해도를 높임으로써 더 나은 거래를 할 수 있도록 한다. 다음은 챗봇 구현에 사용한 프롬프트이다.

```
{ "role": "system", "content": "당신은 재무제표 용어와 재무 데이터 분석에 대해 설명해주는 전문적인 도움을 주는 챗봇입니다. 사용자가 제공한 재무 데이터 분석 내용을 바탕으로 모르는 용어 또는 이해되지 않는 부분에 대해 질문하면, 명확하고 자세한 답변을 제공합니다." }
```

3.4 감정적 투자자들의 충동 매매 방지 서비스 - 'COMMA'

3.4.1 초보 투자자 클러스터링

우선, 초보 투자자의 기준을 정해야 한다. 실제 고객 데이터가 없으므로, 랜덤으로 변수를 생성하여 진행했다. 고객 데이터는 고객 ID, 일일 매매 횟수, 총 수익률, 총 투자 금액, 투자 경력으로 구성했다. 군집 시각화를 위해 PCA를 진행해 주성분의 수를 줄여주었다. 누적 분산 비율이 75% 이상일 때를 기준으로 주성분의 개수를 정해주었고, 예시 데이터의 경우 주성분의 개수가 3개로 정해졌다. 그 이후 k-means 클러스터링을 통해 군집화를 진행했다. 옆은 주성분에 대한 변수 기여 정도를 수치화한 값이다.



[그림 17] 클러스터링 결과

	PC1	PC2	PC3
일일 매매 횟수	0.598	-0.065	0.751
총 수익률	-0.734	0.175	0.222
총 투자금액	0.498	0.575	-0.476
투자 경력	-0.142	0.836	0.34

[표 2] 변수별 주성분 분석

위의 시각화와 주성분 값을 바탕으로 군집을 분석해보았을 때, cluster 2에 해당하는 고객들은 수익률이 낮고, 투자 경력이 적은 것을 확인할 수 있다. 이러한 방식으로 초보자 군집을 설정할 수 있다. 실제 고객 데이터의 많은 변수를 활용하면 위보다 더 정확하고 세분화하여 군집을 설정할 수 있을 것이다. 초보자 군집을 설정한 이후, 이 군집에 해당되는 투자자는 보유 종목에 대한 주가 변동이나 시장 상황에 따라 두려움, 욕심 등과 같은 감정으로 인해 옳지 못한 투자 의사결정을 내릴 수 있으므로, 고객이 투자 결정을 내리기 전에 실시간으로 경고 알림을 제공할 수 있도록 했다.

3.4.2 챗봇 등장 조건

그렇다면, 초보자 군집에 속하는 투자자들에게 어떠한 경우에 챗봇이 나타나야 할까? 첫 번째로 고려한 변수는 공포-탐욕 지수이다. [8]공포 탐욕 지수란 금융 시장에서 투자자들의 감정을 나타내는 지표 중 하나이다. 이 지수는 투자자들이 시장에서 얼마나 두려워하거나 탐욕스러워하는지를 수치로 나타낸다. 지표는 주식 가격 모멘텀, 시장 거래량, 안전 자산에 대한 수요 등을 포함한 7가지 지표를 활용해 측정된다. 만약 이 지표가 0~30 사이의 극단적 공포 상태이면 시장이 과매도 상태임을 시사하며 오히려 좋은 매수 기회가 될 수 있다. 반대로 지표가 70~100 사이의 탐욕 상태이면 시장이 과열된 상태일 수 있으며, 조정이나 하락이 일어날 가능성이 있기 때문에 매도 전략을 고려하거나 보유 자산을 재조정하는

것이 좋다. 공포 탐욕 지수는 현 시장의 상황을 직관적으로 바라볼 수 있는 지표로 판단되기 때문에 이를 변수로 활용했다.

두 번째는 초보자 군집 투자자의 평균 매매 회수보다 상회하는 매매 회수를 기록했을 때이다. 투자자의 수익률이 낮으면서 매매 횟수가 증가하는 경우는 투자자가 심리적으로 불안정한 상태에 있음을 의미한다. [9]이 논문을 보면, 투자자의 매매 회수가 심리에 어떻게 영향을 미치는지 확인할 수 있다. 과잉 자신감을 가진 투자자들은 자신의 정보와 분석 능력을 과대평가하여 잦은 거래를 하게 되며, 이는 결과적으로 낮은 수익률로 이어진다. 이들은 손실을 피하기 위해 더 많은 거래를 시도하며, 이는 결국 거래 비용 증가와 함께 투자 성과에 부정적인 영향을 미친다. 또한 손실 회피 성향이 강한 투자자들이 손실이 난 주식을 매도하지 않고 보유하는 경향이 있음을 발견했다. 이는 손실을 확정 짓지 않으려는 심리적 요인 때문인데, 이러한 행동은 투자자가 손실을 만회하기 위해 더 많은 거래를 유발할 수 있다. 이는 투자자가 감정적으로 불안정할 때 거래 빈도가 증가하는 현상과 일치한다.

위의 내용을 근거로 공포-탐욕 지수와 거래 빈도 변수를 활용하여 COMMA 등장 조건을 설정하였고, 투자자들이 더 나은 선택을 할 수 있도록 설계했다.

3.4.3 컴패니언쉽 요소의 활용

[10]컴패니언쉽은 '우정, 동료애'라는 뜻으로 컴패니언쉽이 적용된 챗봇이 투자자의 감정적인 투자를 예방하는 데에 효과적이라는 사실을 기반으로 챗봇을 만들었다. 고객과 챗봇 간의 친근감을 형성하기 위해 예사말을 활용한 데이터셋이 필요했다. 그러나 예사말을 활용한 금융 관련 텍스트 데이터셋을 찾을 수 없었기 때문에 금융 관련 지식을 묻고 답하는 데이터셋을 직접 생성했다. CLOVA-X의 튜닝을 위해 최소 400개의 데이터셋이 권장되므로 이의 10프로인 40개를 만들었다.

이후 CLOVA-X의 확장 익스플로러를 활용하여 데이터를 400개로 증강하였고 이 데이터를 활용하여 튜닝함으로써 친근한 말투를 사용하는 금융 분야 모델을 구축했다.

	Text	Completion
0	감정적으로 대응하지 않기 위해 어떤 조언을 해줄 수 있어?	감정적인 대응을 피하기 위해서는 장기적인 투자 목표를 재확인하고, 분산 투자 원칙을 유지하며, 시장 변동에 휘둘리지 않는 것이 중요해.
1	변동성을 관리하기 위한 조언을 해줄 수 있어?	변동성을 관리하기 위해 자산 배분을 다양화하고, 변동성 낮은 자산에 비중을 두며, 시장 타이밍을 피하는 것이 중요해!
2	리스크 관리를 위해 어떤 방법을 추천해?	리스크 관리를 위해 자산 배분을 다양화하고, 헷징 전략을 사용하며, 정기적으로 포트폴리오를 재조정하는 것이 중요해!
3	장기적인 투자 전략은 어떻게 유지할 수 있어?	장기적인 투자 전략을 유지하기 위해 정기적으로 포트폴리오를 점검하고, 분산 투자를 유지하며, 시장 변동에 휘둘리지 않는 것이 중요해!
4	인플레이션에 대처하기 위한 투자 전략은 무엇이 있을까?	인플레이션에 대처하기 위해 실물 자산, 인플레이션 연동 채권, 그리고 고배당주에 투자하는 것이 좋아!

[그림 18] 컴패니언쉽 요소를 활용한 학습 데이터

3.4.4 RAG 활용 개인화 투자 전략

다음으로, 고객의 과거 투자 데이터를 기반으로 대화를 이끌어 나가기 위해 RAG (Retrieval-Augmented Generation) 모델을 활용했다. RAG 모델은 고객의 투자 이력과 같

은 대규모 데이터베이스에서 관련 정보를 검색하여 최신의 정확한 정보를 제공하고, 이를 바탕으로 챗봇이 맞춤형 응답을 생성할 수 있도록 도와준다. 이러한 접근 방식을 통해 챗봇은 고객의 특정 투자 상황에 대해 보다 정밀하고 개인화된 조언을 제공하여 고객 만족도를 높이고, 최신 금융 데이터를 기반으로 신뢰성을 향상시킬 수 있다.

개인화된 상담을 제공하는 챗봇을 설계하기 위해 고객의 투자 이력, 최신 금융 데이터, 투자 가이드라인을 사용한다. 먼저, 고객의 투자 기록이 수집된 링크와, Yahoo Finance 등에서 최신 금융 정보를, 트레이딩뷰의 커뮤니티 등에서 투자 원칙 및 전략 자료를 확보한다. 이 데이터를 적절한 크기로 나누고 임베딩하여 Vector DB에 저장한 후, 고객이 챗봇에게 질문하면 이를 임베딩하여 Vector DB에서 유사한 데이터를 검색한다. 검색된 데이터를 기반으로 금융 관련 지식을 튜닝 시킨 3.3.3절의 모델을 통해 맞춤형 답변을 생성함으로써 고객의 구체적인 투자 상황에 대한 개인화된 조언을 제공할 수 있다.

IV. 결론 및 기대효과

본 프로젝트에서는 초보 투자자들을 위한 증권 MTS의 핵심 서비스로 당일 뉴스 기사를 활용한 테마 추천 서비스 ‘맥주와 기저귀’, 재무제표 분석 서비스 ‘fin-bot’, 감정적 투자를 막아주는 챗봇 서비스 ‘COMMA’를 제안한다. 각 서비스는 초보 투자자들에게 필요한 정보를 제공하고 감정적 결정을 줄이며, 투자 의사 결정에 결정적인 도움을 주는 것을 목표로 한다. 이를 통해 초보 투자자들에게 빈번하게 일어나는 정보의 부족이나 감정적 결정으로 인해 발생할 수 있는 투자 실패를 줄이고, 데이터 기반의 논리적 결정을 통해 투자 성과를 높일 수 있도록 한다. 이는 투자자들의 만족도를 높이고, 장기적으로 미래에셋증권 MTS를 사용하는 고객 충성도를 증가시킬 수 있다.

또한 금융 시장과 재무제표에 대한 초보 투자자의 이해를 돕고 자신감 있는 투자 결정을 내릴 수 있도록 한다. 이러한 교육적인 측면은 사용자의 지속적인 참여와 MTS 사용률을 높일 수 있다.

이러한 긍정적인 연쇄반응은 최종적으로 미래에셋증권의 브랜드 이미지 강화 효과를 불러일으킨다. 인공지능을 활용해 사용자 중심의 혁신적인 서비스를 제공함으로써, 기업의 브랜드 이미지를 강화하고, 경쟁사 대비 차별화된 서비스를 통해 시장에서의 입지를 확립할 수 있다. 이는 장기적으로 기업의 성장과 시장 점유율 확대에 기여할 것으로 보인다.

이러한 기대효과를 통해 현 미래에셋증권 MTS가 보고서에서 제안한 서비스를 고객들에게 제공한다면 신규 투자자들의 유입을 늘릴 수 있으며, 동시에 기업의 수익성과 브랜드 가치를 높이는 데 중요한 역할을 할 것이다.

V. 참고문헌

- [1] 김민기, 김준석. 자본시장연구원 이슈보고서 21-11 - 코로나19 국면의 개인투자자: 투자행태와 투자성과
(https://www.kcmi.re.kr/report/report_view?report_no=1243&s_report_subject=&s_report_type=&thispage=2)
- [2] 임성원 (2024.02.20). KB증권, GPT스토어에 AI챗봇 'KB증권 GPT' 출시. 뉴스저널리즘
(<https://www.ngetnews.com/news/articleView.html?idxno=502013>)
- [3] 서한나 (2024.03.13.). '주린이' 품는 토스증권, MTS 잠재력 폭발할까. 뉴스저널리즘
(<https://www.ngetnews.com/news/articleView.html?idxno=503179>)
- [4] 윤주홍. 재무제표분석을 통한 경영성과분석 : 사례연구 : J 은행의 재무제표를 중심으로 1p
- [5] 이승현. 2030세대 개인투자자의 감정적인 투자를 예방하기 위한 컴패니언쉽과 리추얼 요소 기반의 금융 챗봇 인터랙션 디자인 연구 6-7p
- [6] Sondos Mahmoud Bsharat, Aidar Myrzakhan, Zhiqiang Shen. Principled Instructions Are All You Need for Questioning
- [7] 윤주홍. 재무제표분석을 통한 경영성과분석 : 사례연구 : J 은행의 재무제표를 중심으로 49p
- [8] Jyoti Singhal. Role of Greed and Fear Index in Investment Decision Making 11~12p
- [9] Terrance Odean. Are Investors Reluctant to Realize Their Losses?
- [10] 이승현. 2030세대 개인투자자의 감정적인 투자를 예방하기 위한 컴패니언쉽과 리추얼 요소 기반의 금융 챗봇 인터랙션 디자인 연구 11p