안녕하세요. 팀 네츄럴의 캡스톤디자인1 최종발표를 시작하겠습니다.

(다음장)

이번 발표는 프로젝트의 필요성과 목표, 팀원과 역할을 소개하는 Project Introduction, 프로젝트의 구성, 개발내용, 사용기술에 대해 기술하는 Project Content, 마지막으로 프로젝트의 검증과 결과분석, 미흡한점과 추후 보완할 내용들에 대해 얘기하는 Project Evaluation 세 파트로 이루어져 있습니다.

(다음장)

먼저 프로젝트 소개입니다.

(다음장)

2015년부터 암호화폐에 대한 관심이 많아졌고, 그 열풍은 지금도 이어지고 있습니다.

제테크가 선택이 아닌 필수가 된 시대에서, 암호화폐 역시 제테크의 일종으로 각광받고 있습니다. 그러나 현재 암호화폐에 투자하는 많은 사람들은 암호화폐에 대한 전문적인 지식이 전무하고, 단편적인 지식과 주변 조언만을 가지고 무작정 투자하고 있습니다.

종목에 대한 투자라는 점에서 주식과 비교하게 되는데, 주식 시장은 오랜기간에 걸쳐 쌓인 데이터와 진행된 많은 연구, 정비된 제도장치와 법안등이 충실하게 받쳐주고 있고, 다양한 글, 프로그램등을 통해 주식에 대해 지식이 부족한 사람도 쉽게 정보를 얻고 지식을 쌓아가기 좋습니다.

그러나 코인 시장은 형성된지 오랜 기간이 지나지 않아 쌓인 데이터도 부족하고, 제도나 법안등도 제대로 마련되지 않았습니다. 그렇기에, 코인에 대한 실제 투자자들의 감정분석을 통한 여론 파악이 큰 도움이 됩니다.

하지만, 특정 종목에 대한 현재 여론을 조사하기에 여러 불편한 점이 따름. 수많은 인터넷 데이터 중 특정 종목에 대한 데이터만 걸러내고, 그 중에서도 다양한 광고성 글들을 솎아내야지만 여론을 파악할 수 있습니다.

(다음장)

이러한 점에 착안해, 저희 프로젝트는 특정 코인에 대한 글들을 수집하고, 그중 광고성 데이터를 제거하여 실제 투자자들의 여론을 파악해 이를 다른 요소와 결합해 현재 코인시장에 대한 저희만의 지표를 산출하고, 이를 이용자들에게 제공하는 것을 목표로 삼았습니다.

(다음장)

저희 팀원은 이렇게 네명으로 구성되어있고, 각각 프론트엔드, 백엔드, 감정분석 모델 총괄, 광고분석 모델 총괄로 분야를 나누어 맡고 개발을 진행했습니다.

(다음장)

다음은 프로젝트 내용입니다.

(다음장)

저희 프로젝트의 플로우차트를 간단하게 그려봤습니다. 사용자가 웹에 접속하면 트위터 크롤링 후 암호화폐 관련 글들을 분류, 미리 학습시킨 모델을 바탕으로 자연어처리를 거쳐 광고분석, 감정분석을 수행한 뒤 결과 데이터를 서버에서 제공하여 웹이 사용자에게 전달하는 구성입니다.

(다음장)

프로젝트 구조 및 사용 기술은, 웹은 REACT와 CSS를 기반으로 구축하였고, 백엔드로는 python fastapi와 nodejs, aws server를 사용하였습니다. DB는 sqlite를 사용했습니다.

(다음장)

모델 학습 및 감정분석을 위한 데이터 크롤링은 트위터 공식 API와 python library인 tweepy를 사용하였습니다. 광고 학습 데이터셋을 생성할때도 사용했고, 서버에서 매 시간 일정량의 트윗을 수집하여 어떤 코인을 언급했는지 분류하고, 그 내용을 데이터베이스에 저장했습니다.

(다음장)

자연어 처리를 위한 라이브러리로는 텐서플로우, 케라스, 허깅페이스의 트랜스포머와 트레이너 등을 사용했고, pre-trained 모델로는 BERT 와 DISTILBERT 두가지 모델을 사용했습니다.

(다음장)

광고 구분을 학습시킬 모델의 경우 DISTILBERT를 사용했습니다. DISTILBERT를 사용한 이유는, 광고학습의 경우 간단한 이진분류로 상대적으로 LAYER가 적어도 적절히 분류 학습이 가능하기 때문에 기존 BERT 대비 퍼포먼스를 중시한 DISTILBERT를 사용했습니다. 우측 표를 보면 알수 있듯이 DISTILBERT는 BERT대비 모델의 크기는 40프로 감소하였고, 연산속도는 60프로 증가했으나 성능은 기존 BERT의 97프로 수준을 유지하는 준수한 성능을 보였음을 알 수 있습니다.

(다음장)

광고 학습을 위한 데이터셋은 앞서 언급한 twitter API와 tweepy를 사용하여 각종 트윗을 수집하고, 그걸 매뉴얼하게 true / false로 이진분류하여 생성하였습니다. 학습을 위해 약 1500개의 학습 데이터셋을 제작했습니다.

(다음장)

데이터셋을 생성하며, 트위터에서 수집한 여러 트윗중 단순 광고 외에도 여러 불필요한 트윗이 많이 발견됬습니다. 특정 종목이 어느 가격인지 알려주는 트레이딩 봇이나 가상화폐관련 뉴스를 알리는 뉴스 봇, 특정 유저 거래내역을 알려주는 봇 등등 다양한 트윗 봇들과 이벤트 참여 트윗 등 많은 불필요 데이터가 발견되었고, 이에 광고 분석 모델을 단순 광고 뿐만 아니라 일반 사용자 트윗을 제외하고 전부 필터링 하도록 데이터셋 라벨링을 수행했습니다.

(다음장)

이런 데이터셋을 가지고 pre-trained된 DISTILBERT 모델을 바탕으로 텐서플로우, 케라스

등을 활용하여 dataset labeling, tokenizing, fine-tuning 과정을 거쳐 모델을 학습시켰습니다.

(다음장)

감정 분석을 위한 모델의 경우 좀더 높은정확도, 그리고 학습시킬 데이터셋 특징을 고려하여 BERT 모델로 결정했으며, 보이는 세가지 데이터셋을 사용하여 학습시켰습니다. (클릭두번)

위 두 데이터셋, twitter us airline sentiment와 tweet sentiment extRACTION은 감정분류가 완료된 트윗 기반 데이터셋이기 때문에 학습에 사용했고, 마지막 financial sentiment analysis는 가상화폐라는 종목이 경제용어가 자주 등장하기 때문에, 이를 학습시키고자 추가로 포함했습니다.

학습방법은 광고학습 모델과 비슷한 과정을 거쳐 학습시켰습니다.

(다음장)

서버는 FAST API, NodeJS, SQLite를 이용하여 구축했습니다.

백그라운드에서 매시간 1천개의 트윗을 수집하여, 학습시킨 광고분석 모델을 통해 불필요한 데이터를 필터링한 뒤, 해당 트윗이 어느 코인에 대한 트윗인지 분류합니다.

그 뒤, 감정분석 모델을 바탕으로 해당 코인에 대한 감정분석 결과를 db에 저장하도록 구현했습니다.

(다음장)

감정분석 결과를 바탕으로, 저희만의 FEAR and GREED 수치를 구하도록 했습니다.

우선 각 코인별 BITSENTIMENT SCORE의 경우 하루 전부터 지금까지 24시간동안 수집한 트윗의 감정분석 결과를 0~100으로 산출한 수치로 나타나도록 하였습니다.

그 뒤, 시장 전체의 마켓 fear and greed 수치를 MARKET MOMENTUM, SAFE HAVEN, VOLATILITY, SNS SCORE 4가지 지표를 바탕으로 결정했습니다.

MARKET MOMENTUM 같은 경우는 최근 60일간 코인 거래량의 평균과 현재 코인 거래량을 비교해 코인을 구매하는 사람이 얼마나 많아지는지를 나타내는 지표이고, SAFE HAVEN은 최근 60일간 전체 코인 시장에서 비트코인이 차지하는 비율의 평균에 비해 현재 비트코인의 전체 시장 대비 비율이 어느정도인지를 나타내는 지표이고, Volatility는 시세 변동이 얼마나 큰지를 나타내는 지표, SNS SCORE는 상위 10종 코인의 BITSENTIMENT 수치 평균을 나타내는 지표입니다.

이러한 네가지 지표를 0~100 사이의 scale로 매기고 각 25%비율로 반영하여 최종 fear and greed 수치를 결정합니다.

(다음장)

이러한 결과를 사용자에게 표현해줄 웹의 경우, REACT와 CSS를 사용하여 구현했습니다.

사용자들에게 상위 10종 코인의 각종 정보, 각 코인의 시가, 시가총액, 24시간동안 변화량, BitSentiment 수치 등의 정보와 이전에 계산했던 시장 전체의 FEAR and GREED 수치를 보여줄 수 있도록 구현했습니다.

(다음장)

마지막으로 프로젝트 평가입니다.

(다음장)

학습시킨 모델 평가의 경우 KERAS 라이브러리의 EVALUATE 기능을 사용하여 모델의 정확성을 평가했습니다.

(다음장)

광고 학습 데이터셋중 20%를 모델 테스트용 데이터셋으로 따로 분리했고, 나머지 80%의 데이터셋을 바탕으로 학습을 진행한 뒤 테스트용 데이터셋을 바탕으로 EVALUATION을 실행한 결과입니다. 보이는 바와 같이 정확도 0.83, 즉 84% 정도의 정확도를 보였습니다.

(다음장)

동일한 방법으로 감정분석 모델 역시 학습 데이터셋중 일부를 테스트용 데이터셋으로 분리해 학습 뒤 EVALUATION을 진행했습니다. 감정분석 모델의 경우 77%라는 비교적 저조한 정확도를 보였습니다.

(다음장)

저희가 산출한 Fear and Greed 수치가 실제 코인 시장에 얼마나 잘 반영되는지를 확인하기 위해, 최근 7일간의 Fear and greed 수치와 전체 코인시장의 총 시가총액을 비교했습니다.

(다음장)

주황색 선이 코인시장의 시총, 파란 선이 저희가 측정한 fear and greed 수치입니다.

그래프 결과를 해석하기 전 알아둘 사항으로,

최근 암호화폐의 급격한 폭락으로, 6월 6일 시가총액이 1.27조 달러였던데 반해 최저점인 6월 13일에는 939억달러로 9일만에 26%의 급락을 보였습니다.

이러한 대형 악재 때문에 측정한 fear and greed 수치는 전부 50을 넘지 못하는, fear 에 머무르는 모습을 보여줬습니다. 그러나, 그래프를 보면 확인할 수 있듯 6월 13일 전후로 하락세이던 코인 시총이 상승세에 접어들며 조금씩 회복하는 모습을 보이자 동시에 fear and greed 수치가 따라가듯이 상승하는 경향을 보이는것을 확인할 수 있습니다. 이러한 점을 보아, 저희의 알고리즘을 바탕으로 측정한 fear and greed 수치가 실제 암호화폐 시장과 유의미한 연관관계를 보이고, 이를 통해 실제 암호화폐 시장의 여론과 동향을 파악할 수 있을 것이라 생각됩니다.

(다음장)

프로젝트를 진행하며 아쉬웠던 점으로는, 첫번째로 서버 선택 실수입니다. 프로젝트를 처음 시작할때는 미리 학습시킨 모델을 불러오는 과정에서 많은 부하가 걸리지 않을거라 판단하고 AWS free server를 사용했습니다. 그런데 실제로 텐서플로우를 통해 학습시킨 모델을 불러오는 과정에서 서버가 계속 멈추는 문제가 발생해, 어쩔수없이 팀원의 데스크탑을 통해 서버를 돌리는 방식으로 조치했습니다.

두번째로 아쉬웠던 점은 학습 모델의 정확도로, 약 85% 정도의 정확도를 목표로 했으나 광고분석과 감정분석 각각 83%, 77%로 다소 아쉬운 정확도를 보였다는 점, 그리고 마지막으로 아쉬웠던 점은 분석 내용이 트위터 데이터에 한정되어 있다는 점입니다.

후에 프로젝트를 개선할 기회가 된다면, 더 좋은 서버를 사용해 상시 서비스가 가능하도록 개선하고, 현재 분석 결과는 하루를 기준으로 하는데 이를 시간별로, 혹은 주나 달별 분석결과를 추가적으로 제공해 사용자가 더 다양한 데이터를 얻을 수 있도록 추가 기능을 구현하는 것.

그리고 더 다양한 데이터셋으로 모델을 추가적으로 학습시켜 모델 성능을 개선시키는 작업을 통해 정확도를 올리고자 합니다.

발표는 이상입니다. 감사합니다.