뉴스 기사 및 유튜브 텍스트의 감성 분석과 딥러닝을 이용한 주가 등락 예측 서비스 구현

6조 업빛투

발표자 : 김형림

〈p.1 STOCK CHANGE PREDICTION〉
안녕하십니까, 저는 6조 업빛투의 발표를 맡은 김형림이라고 합니다. 그럼 발표를 시작하겠습니다.

〈p.2 주의〉
저희 업빛투는 주식 관련 서비스를 구현했습니다. ***저희의 서비스를 통해 제공되는 정보는 투자 판단에 대한 참고 자료일 뿐임을 발표에 앞서 먼저 말씀드리겠습니다.*** 자 다시 한 번 말씀드리겠습니다~ (웃으면서) ***본 서비스에서 제공되는 투자 정보는 단지 투자 판단에 대한 참고임을 밝혀드립니다. 서비스에서 제공하는 주식의 가치, 가격 상승 및 가격 하락에 대해서는 업빛투가 보장하지 않습니다.***

〈p.3 목차〉
프로젝트 발표는 ***'개요 및 데이터 수집, 분석 과정, 웹 페이지 구현, 발전방향 및 느낀점'의 4단계***에 거쳐 진행하겠습니다.

〈p.4 프로젝트 개요〉
1단계, 프로젝트 개요입니다.

〈p.5 프로젝트 기획 배경〉
코로나19가 시작된 이후 코스피 지수는 1,400 선까지 하락하였다가, 낮은 기준금리, 소비 회복 등의 이유로 개인투자자가 급증하면서 3,000 선을 돌파하는 등 엄청난 주식 투자 열풍이 일어났습니다. 이러한 상황으로 ***'동학개미운동'***이라는 말까지 등장했습니다.

〈p.6 프로젝트 기획 배경〉
이렇게 유입된 새로운 주식 초보자들은 어떻게 정보를 얻을까요? 개인이 주식 관련 용어와 이슈들에 대해 직접 공부했던 지난 날과는 달리, 요즘에는 ***주식 관련 YouTube 콘텐츠***를 통해 공부하고 정보를 얻는 사람들이 많아졌다는 것을 여러 기사를 통해서 확인할 수 있습니다.

〈p.7 뉴스 기사 및 유튜브 텍스트의 감성 분석과 딥러닝을 이용한 주가 등락 예측 서비스 구현〉
이에 저희 업빛투는 주식가격 예측의 변수가 되는 다양한 방법들 중에서 '뉴스 기사와 유튜브 텍스트의 감성 분석'을 통해 주식가격의 등락을 예측해보기로 했고, 최종 주제를 ***[뉴스 기사 및***

유튜브 텍스트의 감성 분석과 딥러닝을 이용한 주가 등락 예측 서비스 구현]으로 정했습니다. ***서비스 이용 대상은 분석 대상 주식 종목에 관심이 있는 직장인***으로, ***출퇴근 시간을 활용하여 종목 관련 정보를 한눈에 확인할 수 있도록 서비스를 구현하는 것이 목표***입니다.

<p.8 주가 등락 예측 선행연구의 예측 정확도>
주가 등락 예측과 관련된 다양한 선행연구를 확인해본 결과, 일반적으로 45~55% 정도의 예측 정확도를 보였는데요, 저희 업빛투도 이 기준에 맞춰서 목표 등락 예측 정확도를 55%로 잡았습니다.

<p.9 구성원 및 역할>
이제부터 본격적으로 설명드릴 프로젝트를 열성적으로 진행해주신 업빛투 팀원들입니다. 정길종님을 팀장으로, 데이터 엔지니어링 파트의 인태우님, 데이터 사이언스 파트의 윤보람님, 채길호님 그리고 저까지 5명으로 구성되었습니다.

그리고 최근 미국 국채가 흔들리면서 우리나라 주식시장 또한 영향을 받아, 하락장이 지속되고 있어 걱정하시는 분들이 많을 것 같은데요. 저희 업빛투의 프로젝트가 추후에 더 발전되어 이런 불안함을 덜어드릴 수 있게 된다면 좋겠습니다.

먼저 전체적인 프로젝트 진행 과정을 보여드리겠습니다. (((WBS))) 총 7주의 기간 동안, 주제를 선정하고, 크롤링으로 데이터를 수집하여 DB에 적재했습니다. EDA와 전처리 과정을 거친 후에 모델링을 하고, 최종 분석 결과를 시각화해서 이를 웹 페이지에 구현했습니다. 이 과정을 순서대로 하나씩 설명드리겠습니다.

<p.10 데이터>
데이터 수집 관련 설명입니다.

<p.11 시가총액 상위 종목>
저희 업빛투는 5개의 분석 대상 종목을 선정했습니다. 이때 ***시가총액 상위 10개의 종목 중, 동일 그룹사인 기업과 뉴스 기사 크롤링 시 좋지 못한 질의 데이터가 상당수 포함되는 네이버와 카카오를 제외***하여, 최종적인 분석 대상 종목은 '***삼성전자, 현대차, LG화학, SK하이닉스, 그리고 셀트리온***'으로 정했습니다.

<p.12 주식·재무 정보>
분석에 사용한 주식 거래 데이터는 FinanceDataReader 패키지로, 재무지표의 경우에는 DART API를 통해 수집했습니다.

<p.13 경제 일간지·주식 콘텐츠 유튜브 채널>
텍스트 데이터는 다음과 같이 뉴스 기사의 경우 제목, 업로드 날짜, 기사 내용, URL을, 유튜브

채널에서는 기사 내용 대신 유튜브에서 제공하는 ***자동 생성 STT(Speech-To-Text) 스크립트***를 크롤링합니다.

〈p.14 내역〉
이때 뉴스 기사는 일일 데이터 수가 평균 10개 정도이지만, 유튜브 스크립트의 경우에는 채널마다 업로드 주기가 불규칙해서 전처리 과정에서 어려움이 있었습니다. (((→ 어떻게 해결?)))

〈p.15 프로세스〉
전체적인 데이터 수집 프로세스입니다. 셀레니움과 파이튜브를 이용하여 크롤링한 텍스트 데이터와 FinanceDataReader의 주식 데이터를 DB에 저장하고, 이를 통해 분석한 결과를 다시 DB에 저장한 최종 결과 테이블을 웹 페이지와 연동하여 대시보드를 구현했습니다.

〈p.16 경제 일간지·주식 콘텐츠 유튜브 채널〉
뉴스 매체 및 유튜브 채널은 ***구독자수 기준*** 상위 매체 및 채널로, ***뉴스 매체는 '매일경제와 아시아경제', 유튜브 채널은 '삼프로TV, 한국경제TV, 슈카월드'*** 이렇게 총 5개를 선정했습니다. 정상적인 서비스 구현을 위해 뉴스 기사는 오전 9시 30분, 유튜브 채널은 오후 11시 30분에 매일 자동으로 크롤링해서 DB에 저장하도록 설계했습니다.

〈p.17 ERD〉
수집한 데이터의 ERD, 즉 개체-관계 다이어그램은 다음과 같습니다. (3초)

〈p.18 내역〉
매일경제와 아시아경제로부터 수집한 종목별 데이터 수는 매체별로 삼성전자 약 14,000건, SK하이닉스 약 4,000건, LG화학 약 3,500건 등입니다.

〈p.19 총 데이터 수〉
5개의 뉴스 매체 및 유튜브 채널로부터 수집한 전체 데이터 수는 다음과 같습니다.

〈p.20 데이터 전처리〉
크롤링한 데이터의 전처리는 크게 세 가지로 진행했습니다.
1. 2018~2022년까지의 휴장일 데이터를 구축했습니다. 그리고 ***휴장일 및 장 마감시간을 고려***하여 뉴스 기사나 유튜브 영상이 업로드된 시점이 휴장일이나 장 마감시간에 해당할 경우 기사와 영상에 대한 분석이 다음 개장일에 반영되도록 날짜를 조정했습니다.
2. 전일대비 주가 변동률에 따라 상승/보합/하락 여부를 판단하여 새로운 updown 컬럼을 만들었습니다.
3. ***한글 형태소 분석기인 KoNLPy의 Okt를 사용***해서 품사를 태깅하여 명사 및 형용사를 추출하여 토큰화하고, 불용어 처리 및 한 글자 단어를 제거했습니다.

〈p.21 분석 모형〉
이렇게 전처리된 데이터를 가지고 저희 업빛투는 다양한 분석 모델을 통해 주가 등락을 예측했

습니다. 먼저 뉴스 기사의 감성점수를 계산하고, 이를 피처로 활용하여 LSTM, ARIMA, FBProphet, RL 분석 결과와 함께 ***최종 앙상블 모델***을 만들었습니다. 한 가지 모델만으로는 높은 예측 정확도를 기대하기는 힘들기 때문에, 저희 업빛투는 다양한 모델을 통한 분석을 시도해보는 것을 목표로 했습니다. 마지막으로는 웹 페이지에 종목별 등락 예측 확률을 제공하고, YouTube 텍스트로 이슈 키워드를 표현했습니다.

〈p.22 감성 분석〉
분석 과정 중에서 감성 분석 부분을 먼저 설명드리겠습니다.

〈p.23 KOSELF〉
저희는 단순히 텍스트의 긍정·부정 여부를 판단하는 것이 아니라, 텍스트가 주식가격 등락에 대해 긍정적인지, 혹은 부정적인지를 판단해야 하기 때문에, 일반적인 감성사전이 아닌 기업 재무분석을 위한 KOSELF라는 한국어 감성사전을 사용했습니다. KOSELF에는 48개의 긍정어와 47개의 부정어가 포함되어 있습니다.

〈p.24 감성사전 제작〉
저희 업빛투는 처음에는 자체적인 감성사전을 구축하려고 했습니다. 먼저 주가의 등락을 기준으로 기사의 긍정·부정 여부를 라벨링하고, 〈〈다음 페이지〉〉 다음과 같이 Facebook의 FastText라는 라이브러리를 활용하여 수집한 뉴스 기사로부터 KOSELF 단어들과의 코사인 유사도가 0.5 이상인 단어들을 추가하여 자체적인 사전을 구축해서 감성점수를 계산했습니다. 〈〈이전 페이지〉〉 하지만 감성점수를 토대로 예측한 주가 등락은 정확도가 50% 수준 혹은 미만이었습니다. 따라서 결국 KOSELF 감성사전으로 뉴스 기사마다의 감성점수를 계산하고, 계산 결과로 주가를 예측하는 것이 아니라 이를 최종 앙상블 모델의 피처로 포함했습니다.

〈p.25 기존 사전 이용〉
Pass

〈p.26 감성점수 계산 후 앙상블 피처 사용〉
KOSELF 감성사전을 채택한 이유는, 다음과 같이 저희 팀 5명이 50개의 뉴스 기사를 랜덤으로 추출하여 수작업으로 긍정·부정 라벨링을 진행하고, 계산된 감성점수로 긍정·부정 판단 정확도를 확인했을 때 해당 사전이 가장 높게 나왔기 때문입니다.

〈p.27 NSI〉
또 하나 최종 앙상블 모델에 피처로 포함한 것은 한국은행 경제통계국에서 개발한 ***'뉴스심리지수(NSI)'***입니다. NSI는 긍정문장과 부정문장의 수를 통해 계산되지만, 저희 업빛투는 이 부분에서 문장을 단어로 바꿔 계산했습니다.

〈p.28 유튜브 스크립트〉

유튜브 스크립트의 경우에는 불용어 처리 시 반복적으로 나오는 의미 없는 단어를 수작업으로 찾아내어 전처리를 진행했습니다.

〈p.29 유튜브 스크립트〉
유튜브 스크립트 또한 긍정·부정 라벨링을 진행했습니다. 기존의 VADER와 rhinoMorph 감성사전을 사용할 경우에는 한쪽으로 치우친 라벨링 결과를 보였습니다. 이에 두 번째 방법으로 위키독스에서 제공하는 네이버 영화 댓글 데이터를 사용하여 학습한 모델로 라벨링을 진행했습니다.

〈p.30 데이터 분석〉
2단계, 이제부터는 텍스트 분석 외에 사용한 모델들에 대해서 말씀드리겠습니다.

〈p.31 ARIMA〉
첫 번째 모델은 ARIMA입니다. 시계열 데이터의 정상성을 위해 로그 변환과 차분을 하여 ADF 검정으로 정상성을 확인하였고, ***AR(자기회귀)과, MA(이동평균)의 차수를 조정***하여 ARIMA 모델의 유의성을 맞췄습니다. 종목마다 상이하지만, 삼성전자의 경우 AR의 차수는 2, 차분은 1, MA의 차수는 2일 때 유의한 결과를 얻을 수 있었습니다.

〈p.32 ARIMA〉
ARIMA로 마지막 10일을 예측한 결과인데요, 오른쪽의 그래프에서 초록색이 실제값, 노란색이 예측값으로 RMSE 692원의 좋은 예측 정확도를 보이는 것을 확인할 수 있습니다.

〈p.33 FBProphet〉
두 번째 모델은 Facebook에서 개발한 Prophet으로, 선형회귀모델인 ARIMA의 단점을 극복하기 위한 목적으로 분석해보았습니다. 주가 데이터의 특성상 진폭이 점점 증가하는 그래프가 나타나므로, 파라미터를 multiplicative(데이터의 진폭이 점점 증가하거나 감소)로 지정하였고, change point_prior_scale은 0.5, 0.6, 0.7 중에서 오차가 작은 값으로 지정하여 분석을 진행했습니다. 여기서 change point는 데이터의 트렌드가 변하는 지점을 의미(ex. 상승→하락 지점)하고, changepoint_prior_scale은 change point의 유연성을 결정하는 파라미터로, 값이 커질수록 유연성이 높아집니다.

〈p.34 FBProphet〉
Train-Test Set으로 나누어 Prophet으로 마지막 10일의 종가를 예측했고,

〈p.35 FBProphet〉
예측 결과는 다음의 그래프에서 실제값인 초록색 선과 예측값인 노란색 선을 보면, 다른 모델들에 비해 Prophet의 경우 예측 오차가 큰 편임을 확인할 수 있습니다.

〈p.36 LSTM〉
세 번째는 선행연구에서 주가 등락을 예측할 때 가장 많이 사용하는 모델인 LSTM입니다. 화면

양쪽의 그래프에서 볼 수 있듯이, 날짜만을 피처로 하는 것보다 재무지표를 포함한 데이터로 종가를 예측하는 모델이 확연히 더 높은 예측 정확도를 보여, DART로부터 가져온 재무지표를 모델에 반영했습니다. 앞선 ARIMA와 FBProphet과 같이 10일을 기준으로, 즉 window size를 10으로 shift를 진행했고, 재무지표까지 포함하여 12개 데이터의 총 120개 피처로 모델을 돌렸습니다.

<p.37 LSTM>
LSTM의 결과값을 산점도로 그려서, 종목별 산점도가 선형에 가까운 형태를 보여 높은 예측 정확도를 보이는 것을 알 수 있습니다.

<p.38 RL>
마지막 네 번째는 강화 학습입니다. 강화 학습은 매수, 매도, 유지의 3가지로 수익점을 계산하여 reward를 주는 방법으로 진행했으며, layer는 Linear와 ReLU를 반복한 층으로 쌓았습니다.

<p.39 RL>
강화 학습도 재무지표를 포함하지 않은 데이터와 포함한 데이터로 진행한 결과, 셀트리온을 제외한 나머지 4가지 종목은 재무지표를 포함한 결과가 더 높게 나왔습니다.

<p.40 앙상블 모델>
최종적으로 앞서 말씀드린 총 여섯 가지의 분석 결과, ***[1. 감성점수, 2. 뉴스심리지수(NSI), 3. ARIMA, 4. FBProphet, 5. LSTM, 6. RL]***의 결과값을 피처로 하여 최종 앙상블 모델을 구축했습니다. 각각의 분석 결과는 매일 뉴스 크롤링이 진행된 이후에 자동으로 업데이트 되도록 코드를 작성했습니다. 아래 그림은 앙상블 모델의 피처 예시입니다. (3초)

<p.41 앙상블 모델>
앙상블 모델이 최종 주가 등락 예측 결과를 도출하는 방법은 두 가지입니다. ***[첫째, 하드보팅, 둘째, 각 모델의 결과값으로 로지스틱 분류]!*** 이 중에서 정확도가 높은 방법을 택해서 최종적으로 웹 페이지에 표출했습니다.

<p.42 앙상블 모델>
하드보팅 결과 41% 예측 정확도를, (3초)

<p.43 앙상블 모델>
로지스틱 분류 결과 63%의 예측 정확도를 보여서, (3초)

<p.44 앙상블 모델>
이렇게 앙상블 모델은 로지스틱 분류를 통해 텍스트의 주가 등락에 대한 긍정 확률을 도출하도록 만들었습니다.

<p.45 웹 페이지 구현>
3단계, 분석 결과를 통한 웹 페이지 구현입니다.

<p.46 웹 페이지 구조도>
웹 페이지는 프레임워크로 Django를 사용했고, DB는 MySQL을, 디자인 및 시각화 부분은 아래 그림과 같이 4가지 툴, [부트스트랩, Chart.js, AnyChart, Tableau]를 사용해서 구현했습니다.

<p.47 Use Case Diagram>
저희 업빛투가 제공하는 주가 등락 예측 서비스의 Use Case Diagram입니다. 아래와 같이 사용자는 웹 페이지에 접속하여 코스피, 혹은 종목별 주가 등락 예측 결과와, 언론사별 긍정·부정 판단 결과, 그리고 종목별 뉴스와 YouTube 텍스트의 이슈 키워드를 확인할 수 있습니다. 자세한 내용은 직접 시연하면서 설명드리겠습니다.

<<<웹 페이지 시연>>>
(((화면설계서 참고)))
저희 업빛투의 서비스를 구현한 웹 페이지입니다.

<p.49 메인 페이지>
메인 페이지에는 (((상단 뉴스 기사))) 상단에 삼성전자, SK하이닉스, 현대차, LG화학, 셀트리온 순으로 긍정/부정점수가 가장 큰 기사를 표출하고, 링크를 통해 이동할 수 있도록 만들었습니다..

그리고 이 아래에 각각 (((코스피 지수 그래프))) 연도별로 월별 평균 코스피 지수 그래프와 (((긍정/부정 판단))) 언론사별 긍정/부정 판단 결과, (((이슈 키워드))) 그리고 제일 아래에는 오늘의 이슈 30선 워드 클라우드를 나타냈습니다.

<p.50 메인 페이지>
이렇게 연도를 선택하면 해당 연도의 코스피 지수를 볼 수 있습니다.

<p.51 종목별 페이지>
그리고 탭을 선택해서 종목별로 관련된 정보와 분석 결과를 확인할 수 있습니다. (((SK하이닉스))) 메인 페이지와는 달리 상단에 해당 종목의 전일 기준 종가와 거래량, 변동률을 표시했고, (((이모티콘))) 저희 업빛투의 앙상블 모델로 예측한 오늘의 주가 등락 예측 결과를 바로 옆에 표시했습니다. ***직관적으로 확인할 수 있도록 눈에 잘 띄는 색과 이모티콘을 활용***했습니다. (((화면 중간))) 그리고 메인 페이지와 마찬가지로, 종목별로 연도별 주가 데이터와 해당 종목에 대한 언론사별 긍정/부정을 판단한 결과를 나타냈습니다.

<p.52 종목별 페이지>
Pass

<p.53 종목별 페이지>
아래로 스크롤을 내리면 (((이슈 기사))) 이렇게 언론사별, 즉 매일경제와 아시아경제의 당일 기사 하나씩을 표출했습니다. 말풍선의 색깔로 파랑은 긍정, 회색은 보합, 빨강은 부정으로 표현해 직관적으로 받아들일 수 있도록 했습니다. 링크를 누르면 해당 기사로 이동합니다. (((서비스 이용자 선택))) 또 이렇게 ***서비스 이용자가 각각의 기사를 읽고 해당 기사가 긍정/중립/부정 중에 어떤 걸 의미하는지 직접 판단할 수 있도록 버튼을 만들었습니다. 기사를 읽고 (((버튼 선택))) 이렇게 버튼을 선택하면, 이 판단 결과를 저희 업빛투의 DB에 저장되도록 해서 추후 모델 개선에 사용할 수 있도록 발전시킬 계획***입니다. (((나머지))) 이렇게 각각의 모델의 분석 결과와 현재 가격과 예측 결과의 차이를 보여주고, 유튜브 스크립트로 뽑아낸 이슈 키워드를 하단에 배치했습니다.

<p.54 태블로 페이지>
(((주가 차트))) 또 태블로로 구현한 분석 결과 시각화 자료를 따로 탭을 만들어서 나타내고, 여기에서는 종목별 시가총액과 (((마우스 오버레이))) 캔들차트를 확인할 수 있도록 구현했습니다. (((분석 차트))) 분석 차트에서는 종목별로 저희 업빛투가 분석에 사용한 ARIMA, FBProphet, LSTM, RL의 일자별 분석 결과를 확인할 수 있습니다. (((네이버 금융))) 그리고 마지막으로 필요 시 네이버 금융으로 이동해서 정보를 얻을 수 있도록 링크를 걸어 버튼을 만들었습니다.

+) 그리고 가장 중요한 부분은, (((모바일 시연))) 직장인들이 출퇴근 시간을 활용해 정보를 얻을 수 있도록 하는 것이 저희 업빛투 서비스의 목표이기 때문에, 모바일로도 접속했을 시에도 한눈에 정보를 확인할 수 있도록 설계했습니다. (((인터넷 접속))) 이렇게 모바일로도 접속하여 정보를 확인할 수 있구요, 좀 더 부드러운 시연을 위해 사전에 녹화해놓은 영상을 잠깐 보여드리겠습니다. (((카톡)))

<p.55 발전방향 및 느낀점>
마지막 4단계, 발전방향 및 느낀점입니다.

<p.56 한계점 & 발전방향>
먼저 가장 큰 한계점은 아직은 유튜브의 한글 STT(Speech-To-Text) 스크립트가 부정확해서 뉴스 기사와 똑같은 방법으로 감성사전을 통해 감성점수를 매길 수 없다는 점입니다. 긍정·부정 라벨링도 저희 업빛투 팀원 5명이 50개라는 적은 수의 뉴스 기사만을 가지고 진행한 것도 분석의 논리를 뒷받침하기에는 부족합니다. 또 아직은 5개의 매체 및 채널, 5개의 종목만을 대상으로 한다는 것입니다. 게다가 범용적인 모델이 아니라 종목별로 모델을 따로 구축해야 한다는 문제도 있습니다.

이를 바탕으로 다음과 같은 발전방향을 제시합니다.
1. 유튜브 STT 기술의 발전 시 동반 발전 가능성
2. 서비스 이용자의 의견을 반영한 객관적인 기사 라벨링으로 추후 모델 개선

3. 분석 대상 매체와 채널, 종목의 확장입니다.

〈p.57 느낀점〉
이번 프로젝트를 통해 텍스트 분석과 다양한 모델로 분석해보고, 협업하면서 분석부터 웹 페이지 구현까지 많은 경험을 할 수 있었습니다. 저희 업빛투는 이번 프로젝트 경험을 토대로 이후 모델을 더 발전시키거나 다른 분야에서도 좋은 결과를 낼 수 있도록 노력할 것입니다.

〈p.58 참고 자료〉
Pass

〈p.59 예측 결과와 실제 종가 비교〉
10/06에 예측한 10/07 종가와 실제 10/07일 종가입니다. 5개 종목 중에 3개 종목의 주가 등락 여부를 맞혔습니다!

〈p.60 Thank You〉
이상으로 6조 업빛투의 ***[뉴스 기사 및 유튜브 텍스트의 감성 분석과 딥러닝을 이용한 주가 등락 예측 서비스 구현]***에 대한 발표를 마치겠습니다. 경청해주셔서 감사합니다.