

딥러닝을 활용한 주식시장 예측의 최근 트렌드

Ask. 일단 논문이 다른 100여개의 논문들을 요약한 것이고, 2019년의 snapshot을 잘 나타내고는 있지만, 지금의 것이랑 거리가 많이 멀다는 점을 먼저 표기해 주세요. 관련 논문들을 분석하는 시각에 대해서는 좋은 인사이트들이 있는데,

Q. '예측'이긴 하지만 제안이 잘 동작하는지를 보려면 과거의 어느 시점 기준으로 그 때 사용가능한 정보들로 미래를 어떻게 예측했었고, 그 결과가 어땠는가 등을 "평가"하고 있을텐데 거기서 어떤 좋은 결과들이 있었는지가 관건들이겠습니다.

자료 조사하던 중에 지나간 거라 출처가 정확히 기억나지 않지만 과거의 방식, 일반적인 회귀분석이나 머신러닝 결과와 비교했을 때 딥러닝 예측이 특별히 아웃퍼폼하는 결과냐에 대해서는 그렇다 아니다가 혼재되어 있는 것으로 보였습니다.

Q. input 으로 쓰는 feature 들이 어떤 것을 쓰고 있는지에 대해 insight 들을 얻을 수 있을까요 ?

발표 내용에도 포함되어 있지만 주로 시장 가격, 또는 가격을 가지고 계산한 기술적 지표들(이동평균 등)만 활용한 연구들이 많았습니다. 생각보다 현업에서 많이 보는 경제지표라던지, 거래량, 심리지표 등등 활용한 연구는 많지 않았는데 추가를 한다면 더 괜찮은 결과가 나오지 않을까 생각해봤습니다.

Q. 나라마다 다른 시장을 다르게 접근하고 있을텐데, 어디까지 일반화가 가능할까요 ? 예를 들면 Yahoo Finance 는 한국 내용을 찾지 못하지 않을까요 ?

데이터 출처로 Yahoo Finance를 가장 많이 활용한 이유가 세계 각국의 거래소를 무료 벤더 중에서는 가장 많이 커버하고 있기 때문일 거라고 생각합니다.

단, 한국 주식에 집중해서 정보들을 받아야 하는 경우라면 키움, 대신, 한투 같은 증권사들의 무료 API를 활용하는 케이스도 생각해볼 수 있을 듯 합니다. Yahoo Finance에서 제공되는 한국 주식 정보 중 간혹 missing된 데이터나 수정종가 계산이 잘 반영되지 않은 경우가 있어 참고해서 사용해야 할 것 같습니다.

Q. 몇 유명한 업체의 미래를 예측하는 것과 덜 유명한 업체의 미래를 예측하는 것이 같은 선상에 있을 거 같지 않습니다. 예를 들면 여기저기 많이 언급되는 회사는 sentiment 등 사용할 수 있는 부분이 많지만, 반대의 경우는 다르고, 게다가 authority 있는 news 이나 just random gossip 이나 따라서도 다양한 해석들이 될 거 같은데, 어떻게 정리하면 좋을까요 ?

Text data를 활용한 모델 구축 시 Quality와 Quantity 간의 선택이 필요할 듯합니다. 예를 들면 정제되고 informative한 데이터가 많은 금융 뉴스지와 개인들의 unreliable 데이터가 무수히 혼재되어 있는 소셜 미디어를 함께 사용하면서 데이터 원천의 비중을 달리하는 방법을 생각해볼 수 있을 것 같습니다. 그 밖에 방법으로는 sentiment lexicon과 같은 전처리 기술로 팩트와 가십을 구분하는데 도움을 받거나, 금융과 관련된 용어로 학습된 분석 툴을 사용할 수 있겠습니다. 중소기업은 대기업만큼 메이저 언론이나 증권사 분석에서 많이 언급되고 사용할 데이터가 충분치 않을 수 있습니다. 이 경우 소셜 미디어나 지역 언론, 기업들의 어닝콜 내용 등 데이터 수집의 범위를 다양하게 넓히는 방식을 고려해볼 수 있을 것입니다.