뉴스기사 감성분석을 활용한 주가 예측 모델링

-감성분석을 위한 논문 정리-

1. 감성 라벨링을 이용한 주식 트레이딩 시스템 개발 (박명석, 김재윤)

전처리

- 주말뿐만 아니라 공휴일에 나온 뉴스기사도 다음 개장일에 영향을 미치도록 설계
- 뉴스 라벨링: 뉴스심리지수 사용 (한국은행이 뉴스기사에 나타난 경제심리를 일 단위로 지수화한 것)

$$NSI = \frac{positive\ sentences - negative\ \ sentences}{positive\ sentences + negative\ \ sentences} \times 100 + 100$$

감성분석

- 사용한 감성 사전은 KOSELF, KNU
- 감성 라벨링 방법: 해당 날짜 뉴스에 감성사전 긍정 단어 개수만큼 +1, 부정단어 개수만큼 -1
- 사용한 모델은 KoBERT(max_seq_len, batch_size, warmup_steps, epoch, max_grad_norm, logging_steps, learning_rate), LSTM 모형(epoch, batch size)

2. 뉴스 텍스트 데이터를 활용한 사용자 정의 감성사전 구축 (장현지)

전처리

- 데이터: 한국경제신문-경제부문 약 5 년
- 주식시장 시간: 9 시~15 시 30 분 이기 때문에 15 시 30 분 이후의 뉴스는 같은 날짜의 주가에 영향을 줄 수 없음 -> 오후 3 시 30 분 이후의 기사들을 다음 날짜의 주식예측에 사용
- KoNLPy 사용 형태소 분석 진행 -> 명사만 추출 이후 다시 Mecab 을 사용해서 명사만 걸러냄, 불용어, 영어, 한자, 특수기호 모두 제거

감성분석

- TF-IDF 를 사용해 단어들의 중요도 구함 (min_df=200 옵션 통해 단어 차원 200 개로 조절)
- x: 일별 각 단어별 중요도 행렬 -> y: 일별 KOSPI 지수
  회귀분석 진행하여 얻은 회귀계수값으로 각 단어의 감성점수 도출

3. 뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형 (김유신, 김남규, 정승렬)

감성분석

- 1 개의 뉴스에 대해 긍정/부정 감성사전과 비교해서 극성 태깅 -> 긍정/부정 비율 계산

$$\text{NewsPNr} = \frac{\sum_{i=1}^{n} wordPN(i)}{n} \times 100\%$$

if(NewsPNr > 52), NewsPN = Positive;
else if(52 >= NewsPNr >= 48), NewsPN = Neu;
else if(NewsPNr < 48), NewsPN = Negative

- 하루동안 발생한 모든 뉴스의 긍정/부정비율 평균내서 최종 그날 하루의 긍정/부정비율 도출
- 다른 변수들(경제뉴스, 시황뉴스, 전망뉴스)을 같이 고려해서 회귀모델 모델링
- NewsPNr 변수가 KOSPI 지수 상승/하락에 유의미한 관계를 가지는지 T-검정 실시
- 상승정확도와 하락 정확도, 통합정확도 같이 제시

4. 온라인상의 뉴스 감성분석을 활용한 개별 주가 예측에 관한 연구 (정지선, 김동성, 김종우)

전처리

- 데이터: 네이버 증권정보서비스-종목뉴스, 약 3 년
- 기업 선정기준: KOSPI200 섹터지수 기준 40 개
- 불용어 (무단전재, 기자, 증권시황,...) 제거, 단음절 체언 및 용언 제거, 뉴스 작성자/메일주소/특수기호 제거
- 명사추출
- 뉴스라벨링: 초과수익률 사용

감성분석

- 각 어휘 감성점수 계산 방법: 긍정뉴스에서 출현한 i 의 빈도수/전체 i 의 빈도수 + 부정뉴스에서 출현한 i 의 빈도수/전체 i 의 빈도수 로 계산

$$TermScore(i_p) = \frac{Num(i) \in PosDocs}{Total\ Num(i)} \quad 식\ (1)$$

$$TermScore(i_n) = \frac{Num(i) \in NegDocs}{Total\ Num(i)} \quad 식\ (2) \qquad TermScore(i) = TermScore(i_p) + TermScore(i_n) \quad 식\ (3)$$

- 해당일 발생한 기업별 전체뉴스의 극성 점수 계산:

$$ComScore(j_t) = \frac{\sum_{i=1}^{n} Num(i_t) \times TermScore(i)}{\sum_{i=1}^{n} Num(i_t)}$$

Num= t 시점에 발생한 모든 뉴스에서의 어휘 i 출현빈도

TermScore(i)= 어휘 i 의 극성점수

5. 주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축방안 (김재봉, 김형중)

전처리

- 데이터: 증권 전용 투자자 게시판데이터 -> 전문성 높이고자, 3 년
- 명사+동사+형용사 추출

감성분석

- PMI 계산해서 PMI 값이 양수인 (두 단어가 같은 문서에 나타날 확률이 높아 비슷한 의미극성을 가진다는 뜻) 빈도가 높은 단어를 중심으로 말뭉치 구성
- 말뭉치 구성 후 classification -> 결과: 297 개의 핵심어와 1614 개의 말뭉치 구축, 긍정(주가상승) 731 개/ 부정(주가하락) 654 개/ 중립 229 개
- 말뭉치 기반 긍정지수 도출

$$긍정지수 = \frac{긍정\ 말뭉치의\ 수}{긍정\ 말뭉치의\ 수 + 부정\ 말뭉치의\ 수}$$