NN과 Text Mining을 이용한 종합주가지수 예측

'16 정보시스템학회 추계학술대회

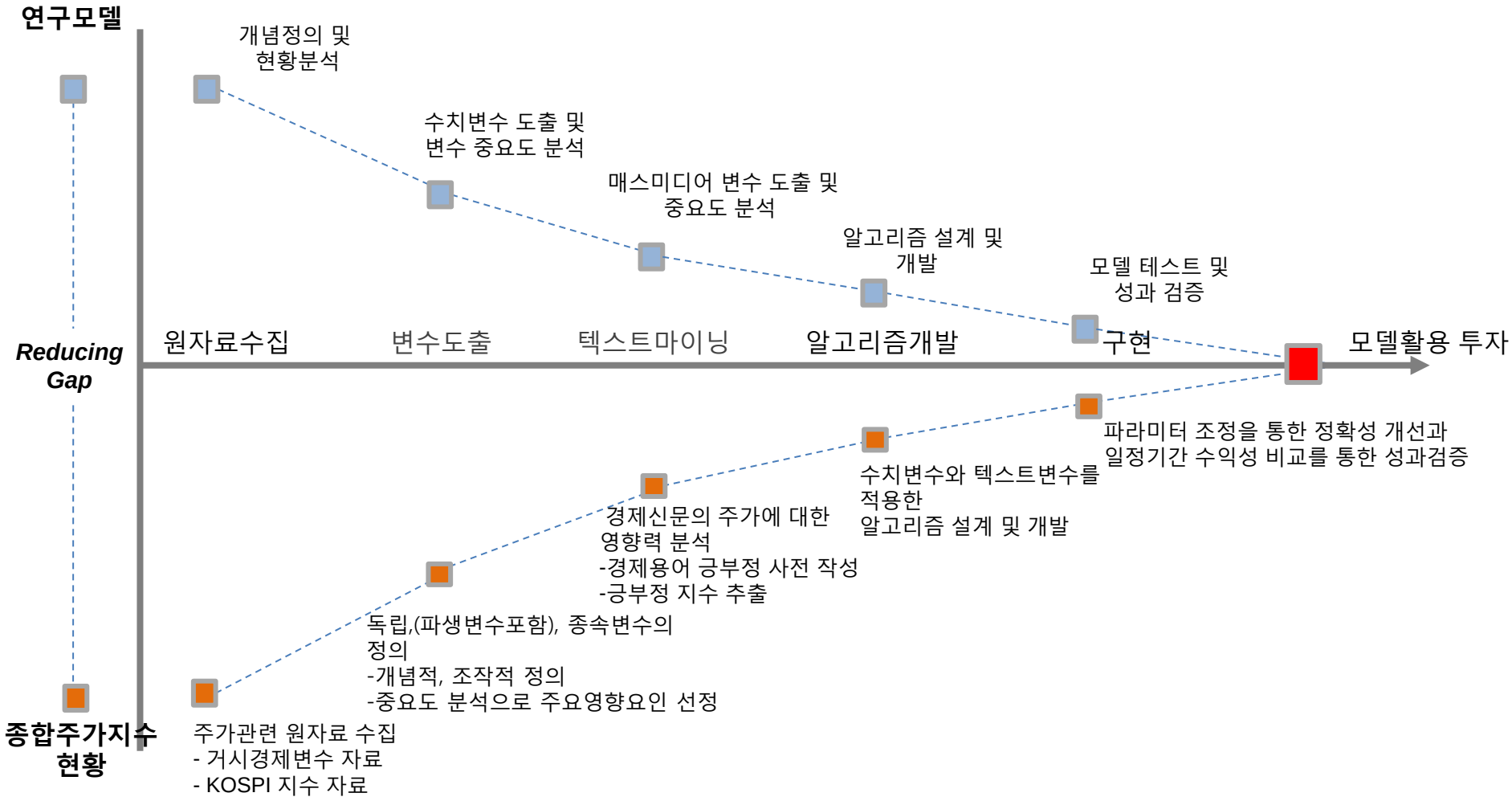
엄 혜미

문 윤지

목차

1. 개요: (1) 연구목적

본 연구는 종합주가지수에 영향을 미치는 핵심 설명변수들을 추출, 생성함으로써 종합주가지수의 상승, 하락을 보다 정확히 예측하는 모델을 만들어 ETF의 매수, 매도에 대한 의사결정에 활용 하고자 함.

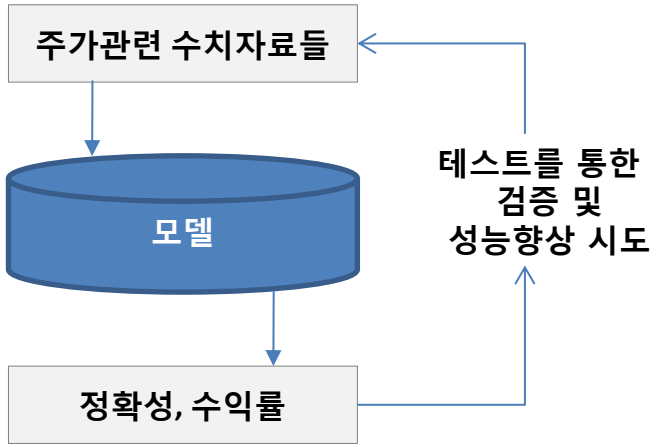


1. 개요: (2) 연구모델

기술적 분석에 텍스트마이닝 분석을 추가함으로써 기존 모델에서는 파악하기 어려웠던 사회적 심리적인 분위기를 반영하여 주가향방을 더욱 정교하게 예측할 수 있게 함

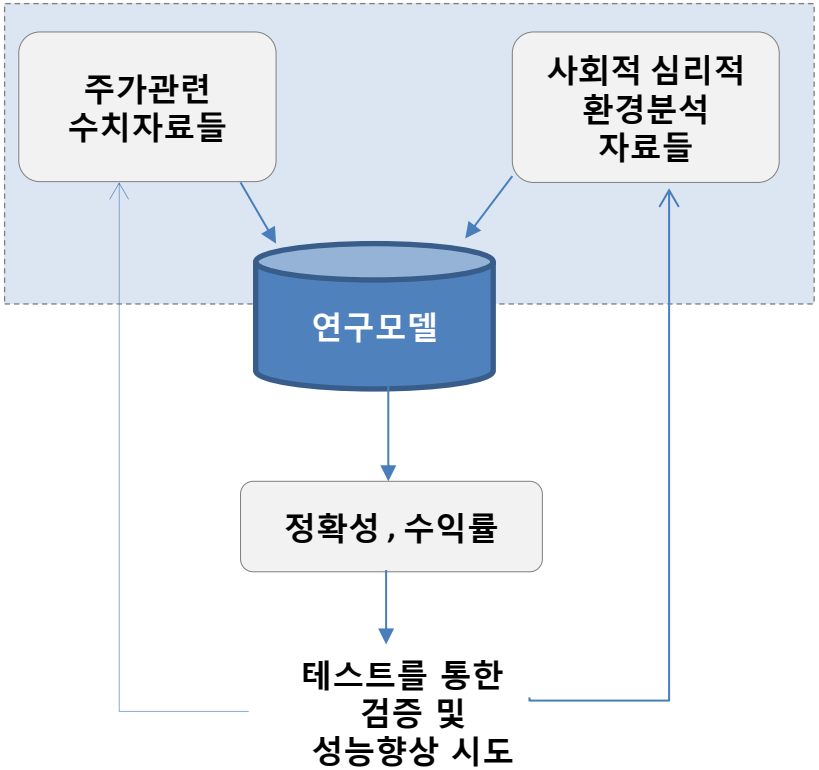
[기존모델]

대부분의 기존모델들이 다양한 알고리즘을 이용한 기술적 분석 시도



[제안모델]

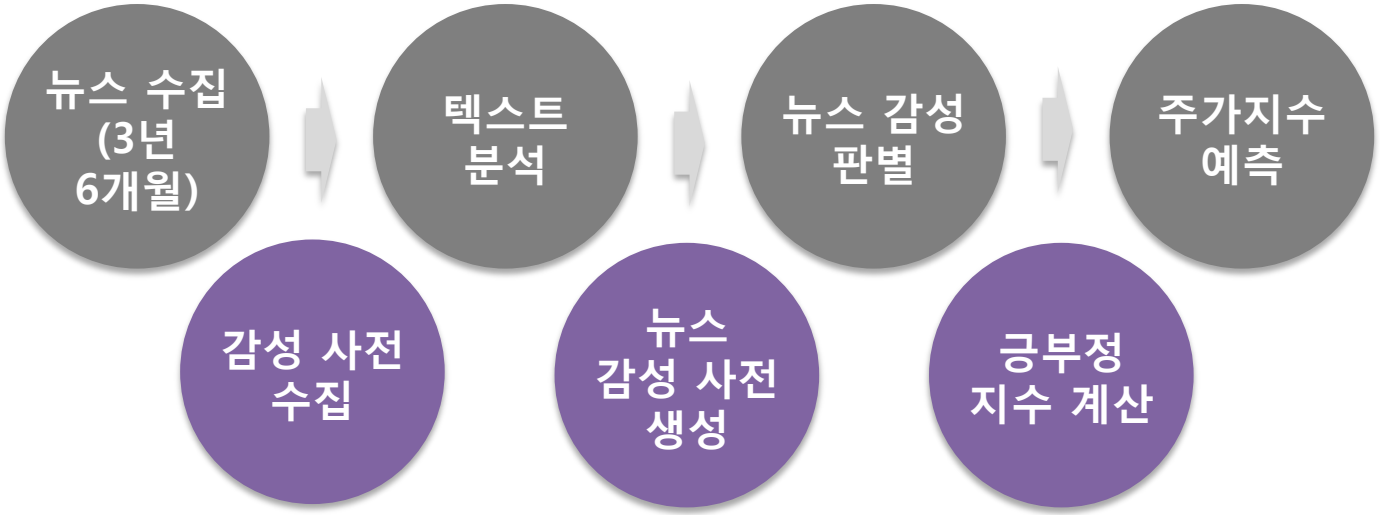
기술적 분석뿐 아니라 기존모델이 반영하지 못한 사회적 심리적 환경 분석 반영



2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

Contents

2011년부터 2014년 6월까지 3년 6개월간의 온라인 경제 뉴스를 수집하여 뉴스의 감성을 분석하고 종합주가 지수를 예측하기 위한 감성사전과 주가지수 예측모형을 수립함



✓감성 분석
온라인 뉴스와 소셜 미디어의 코멘트 등 사용자가 다양한 컨텐츠를 통해 표출한 의견을 추출, 분류, 이해, 자산화하는 과정

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

Contents

웹 사이트와 빅데이터 분석 전문가 과정 페이스 북, R 사용자 그룹 홈페이지 등으로부터 경제 뉴스와 한국어 긍/부정 사전을 수집하고 통합

뉴스 수집 (3년 6개월)

- **헤럴드 경제 뉴스**
- **기간 : 2011.1.1 ~ 2014.6.30**
- **666일 기준 총 2175건**

감성 사전 수집

- **일반 한국어 감성 사전 수집 및 통합**

1. **빅데이터 분석 전문가 과정, 김경태 강사 제공**
   - **긍정 단어 882개**
   - **부정 단어 1679개**
2. **웹 검색 수집**
   - **긍정 단어 883개**
   - **부정 단어 1235개**
3. **긍부정 사전 통합(중복 제거)**
   - **긍정 단어 842개**
   - **부정 단어 2388개**

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

KoNLP와 tm 패키지를 활용하여 뉴스 데이터에 대해 한국어 명사를 추출하고 감성 사전에 포함된 단어들을 대상으로 출현 빈도수를 계산함

Contents

텍스트 분석

➢감성 분석에 사용될 텍스트 요소 도출

1. 한국어 명사 추출
2. 감성 사전을 기반으로 감성 단어 식별하고 출현 빈도수 계산

➢ 한국어 형태소 분석 : KoNLP 패키지 활용

➢ 텍스트 마이닝 : tm 패키지 활용

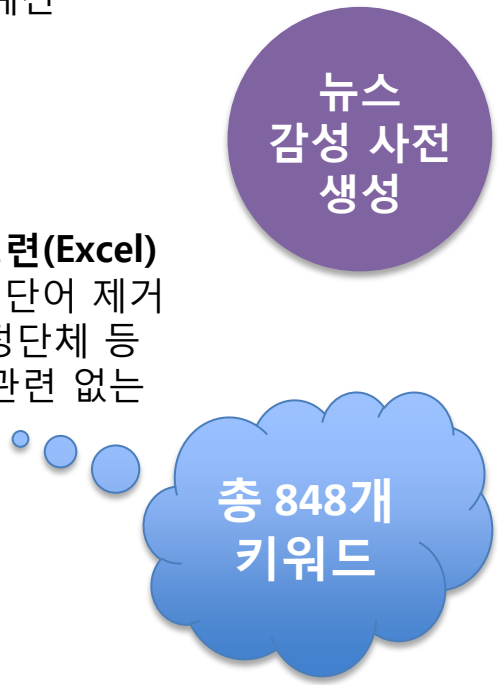
2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

Contents

일반적으로 사용되는 한국어 감성 사전은 뉴스가 종합주가 지수에 미치는 긍/부정 적인 영향을 판별하기에 부족할 수 있다고 판단하여 경제 뉴스에서 사용되는 긍/부정 사전을 생성 함

1. 뉴스 키워드 추출(KoNLP)
- 한국어 명사 추출
- 출현 빈도수 계산

2. 뉴스 키워드 정련(Excel)
- 빈도수 3이하 단어 제거
- 기자이름, 특정단체 등 감성 분석과 관련 없는 단어 제거



3. 뉴스 키워드 극성 정의
- 다음날 주가지수 등/하락 여부를 기반으로 해당 뉴스 키워드의 극성(긍정/부정) 정의
- 긍정 : +1, 부정 : -1

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

동일한 키워드에도 경우에 따라 긍부정 극성이 다르게 나타나므로 해당 단어가 종합주가 지수등락에 미친 영향을 종합하여 긍부정 강도를 계산

Contents

4. 뉴스 키워드 강도 정의

- 해당 단어가 뉴스에 나타난 다음날 종합지수가 상승하면 양수 1, 하락하면 -1로 설정
- 해당 단어의 모든 출현 경우에 대해 +1, -1의 영향도를 계산하고 이를 누계하여 전체 출현 빈도로 나눠 긍부정 강도를 계산함(-1 ~ +1 사이의 값으로 Normalize)

$$\text{키워드의 긍부정 강도} = \frac{\text{출현 시 상승 횟수 + 하락 횟수}}{\text{키워드 출현 빈도}}$$

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

Contents

뉴스의 긍/부정 여부를 판별하기 위해 일반적으로 사용하는 긍부정 단어의 출현 빈도법과 긍부정 강도 기반의 긍부정 지수법 두 가지를 사용

➢ **감성 사전을 이용한 뉴스 극성 분석**

뉴스 감성 판별

1. **긍부정 단어 출현 빈도 기반 방법**
   - 긍정 단어의 수가 많으면 긍정적 뉴스
   - 부정 단어의 수가 많으면 부정적 뉴스

   *** 뉴스 극성 = 긍정 단어 수 – 부정 단어 수**

2. **긍부정 강도와 출현 빈도 기반 방법 : 긍부정 지수**
   - 긍정 단어의 강도가 임계값 이상이면 긍정적 뉴스
   - 부정 단어의 강도가 임계값 미만이면 부정적 뉴스

   *** 뉴스 긍부정 강도 = 긍정 단어의 강도 합 – 부정 단어의 강도 합**

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

Contents

예측의 정확도를 비교하기 위해 3가지 감성 사전을 이용하며, 일별 예측을 위해 개별 뉴스의 긍부정 지수를 종합하여 일별 긍부정 지수 생성

➢ **긍부정 강도를 기반으로 일별 뉴스의 긍부정 지수 계산**

① **개별 뉴스의 긍부정 지수 계산**

② **일별 뉴스 긍부정 지수 합산**

➢ **뉴스 감성 판별에 이용된 사전**

① 한국어 감성 사전 : 웹 사이트에 공개된 한국어 긍/부정 사전 통합**(3,171개 단어)**

② 뉴스 감성 사전 : 뉴스 데이터에서 추출한 단어를 기반으로 해당 단어의 극성과 강도를 정의한 것 **(848개 단어)**

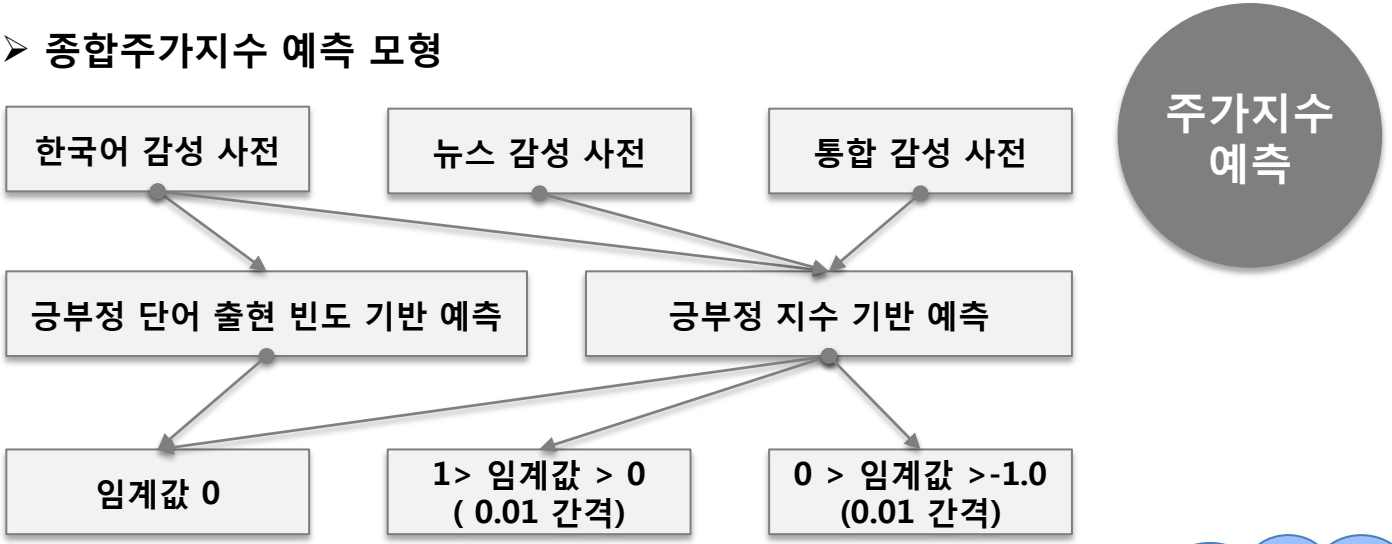
③ 통합 감성 사전 : 뉴스에서 추출한 단어와 한국어 감성 사전에 포함된 단어를 통합하고 해당 단어들의 극성과 강도를 정의한 것**(3,844개 단어)**

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

감성 사전과 극성분석방법과 극성 판별을 위한 임계값을 다양하게 변화시키면서 주가지수를 예측함

Contents

➢ **종합주가지수 예측 모형**



① 한국어 감성 사전 + 긍부정 출현 빈도 + 임계값 0
② 한국어 감성 사전 + 긍부정 지수 + 임계값 0
③ 뉴스 감성 사전 + 긍부정 지수 + 임계값 0
④ 통합 감성 사전 + 긍부정 지수 + 임계값 0
⑤ 한국어 감성 사전 + 긍부정 지수 + 임계값 -0.1
⑥ 뉴스 감성 사전 + 긍부정 지수 + 임계값 -0.1
⑦ ....

604개 예측모형

2.1. 뉴스 감성 분석을 이용한 종합주가지수 예측

Contents

2011년부터 2013년까지 3년치 뉴스 데이터는 학습용 데이터로 사용하고 최근 6개월 데이터는 검증용 데이터로 사용

	학습 기간(11.01~13.12) 예측정확도	정확도
1	한극어 감성 사전 + 긍부정 출현 빈도	52.36%
2	한국어 감성 사전 + 긍부정 지수	73.89%
3	뉴스 감성 사전 + 긍부정 지수	70.12%
4	통합 감성 사전 + 긍부정 지수	75.10%

일반 감성 사전 경우, 긍부정 단어의 개수보다 지수 사용시의 정확도가 20%나 우수하므로 지수 활용으로 1차 결정

	검증 기간(14.01~06) 예측정확도	정확도
2	한국어 감성 사전 + 긍부정 지수	53.72%
3	뉴스 감성 사전 + 긍부정 지수	54.55%
4	통합 감성 사전 + 긍부정 지수	56.20%

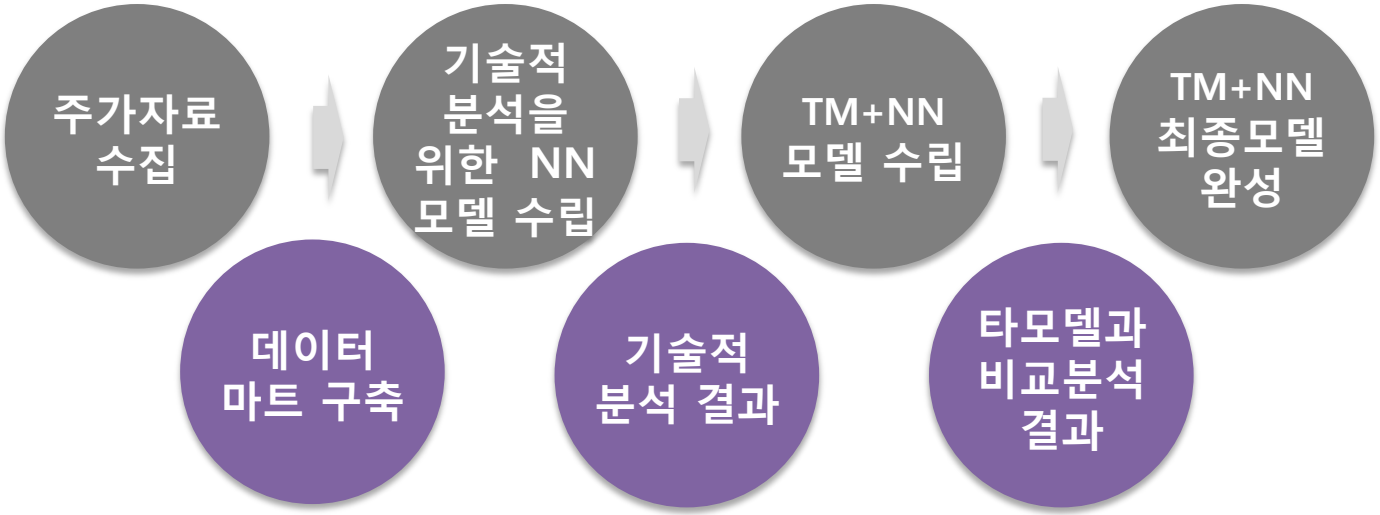
검증용data에 대한 예측 정확도가 가장 높은 통합 감성사전 활용으로 2차 결정했으나, 뉴스 감성 사전의 우수성 확인

뉴스 감성사전은 일반 감성사전에 비해 그 단어수가 4.5배 적음에도 불구하고 예측 정확도가 우수함

2.2. NN을 이용한 종합주가지수 예측

Contents

2011년부터 2014년 6월까지 3년 6개월간의 주가관련 수치자료를 기술적으로 분석하는 모형을 수립하고, 여기에 앞서 완성한 뉴스감성 사전의 긍부정지수를 통합한 종합주가지수 예측 최종모형을 수립함



2.2. NN을 이용한 종합주가지수 예측

Contents

기존 연구들을 기반으로 HTS과 통계청 등에서 제공하는 주가관련 수치자료들과 거시경제변수 자료들을 수집하고 필요한 파생변수도 생성하여 1차 데이터 마트를 구축함

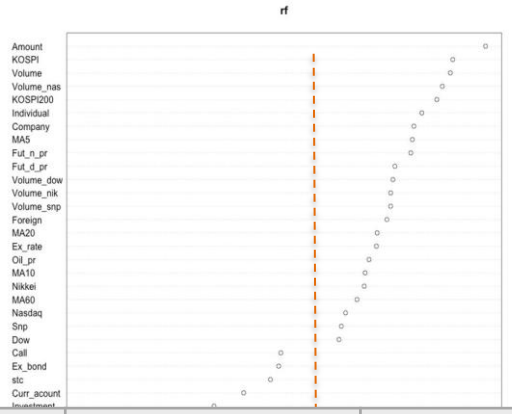
주가자료 수집

- **HTS, 통계청 공시자료 이용**
- **기간 : 2011.1.1 ~ 2014.6.30**
- **기존 주가예측 연구들에서 주요 입력변수로 선정되었던 변수들을 조사하고, 기술적 분석의 기초자료가 되는 주가변수와 거시경제 변수 35개를 선정하여 이를 수집**

2.2. NN을 이용한 종합주가지수 예측

Contents

초기 선정변수 총 35개 중 RF를 이용하여 변수의 중요도 분석 실시 후 기준 10을 상회하는 23개의 변수만을 선택, 최종 데이터마트를 구축함



데이터마트

주가 자료				거시경제 자료		파생변수	
KOSPI 지수	√	나스닥지수	√	환율	√	Stochastic	
거래총량	√	나스닥거래량	√	콜금리		5일 이동평균	√
거래총액	√	다우지수	√	CD금리		10일 이동평균	√
KOSPI200지수	√	다우거래량	√	국고채1년금리		20일 이동평균	√
외국인 매수액	√	니케이지수	√	유가(두바이유)	√	60일 이동평균	√
외국인 거래량		니케이거래량	√	경상수지			
기관 매수액	√	S&P지수	√	설비투자			
기관 거래량		S&P 거래량	√	경제심리지수			
개인 매수액	√	주간선물가	√	평잔			
개인거래량		야간선물가	√	소비자물가지수			

2.2. NN을 이용한 종합주가지수 예측

Contents

최종 데이터마트에 저장된 독립변수 23개를 NN 기법을 이용하여 분석함

2011년부터 2013년까지 3년치 데이터는 학습용 데이터로, 최근 6개월 데이터는 검증용 데이터로 사용하여 23개의 독립변수를 NN 기법으로 분석한 결과 70%에 가까운 예측 정확도를 나타냄

- **HTS, 통계청 공시자료**
- **학습용 데이터 기간 : 2011.1.1 ~ 2013.12.30**
- **검증용 데이터 기간 : 2014.1.1 ~ 2014.6.30**

기술적 분석 결과

23개 독립변수 학습용 data		실제	
		-1	1
예측	-1	253	110
	1	120	260

- **정확도 69.04%**

23개 독립변수 검증용 data		실제	
		-1	1
예측	-1	35	19
	1	21	46

- **정확도 66.94%**

2.2. NN을 이용한 종합주가지수 예측

Contents

기술적 변수와 통합 감성사전의 긍부정지수를 입력변수로 통합한 최종 종합주가지수 예측 모형을 완성함

➢ 23개 기술적 변수 + 텍스트마이닝 bit 변수(통합사전지수를 임계치 -0.7로 나눈)
= 총 24개 독립변수를 입력변수로 설정

TM+NN 모델 수립

24개 독립변수 학습용 data		실제	
		-1	1
예측	-1	253	84
	1	120	286

➢ 정확도 72.54%

24개 독립변수 검증용 data		실제	
		-1	1
예측	-1	34	20
	1	22	45

➢ 정확도 65.29%

기술자료만을 입력한 것 보다 텍스트 마이닝을 결합한 NN모델의 수익률이 더 우수함

	모형구분	학습용 정확도	검증용 정확도	수익률
1	23개 변수 NN	69.04%	66.94%	14.71%
2	TM 포함 24개 변수 NN	72.54%	65.29%	17.37%

2.2. NN을 이용한 종합주가지수 예측

Contents

본 프로젝트 최종모델 이용시 분석결과와 타모델 이용시 분석결과를 비교한 결과, 본 프로젝트 최종모델의 종합주가지수 예측 정확도와 수익률이 더 우수하다는 것이 검증됨

① RF (nod size=2, 10만번 반복)
② SVM (가우시안 RBF 커널함수 사용,
C 파라미터 = 100 , σ^2= 0.0034)
③ NN

다른 모델과 비교 시에도 최종 모델의 예측력이 더욱 우수함

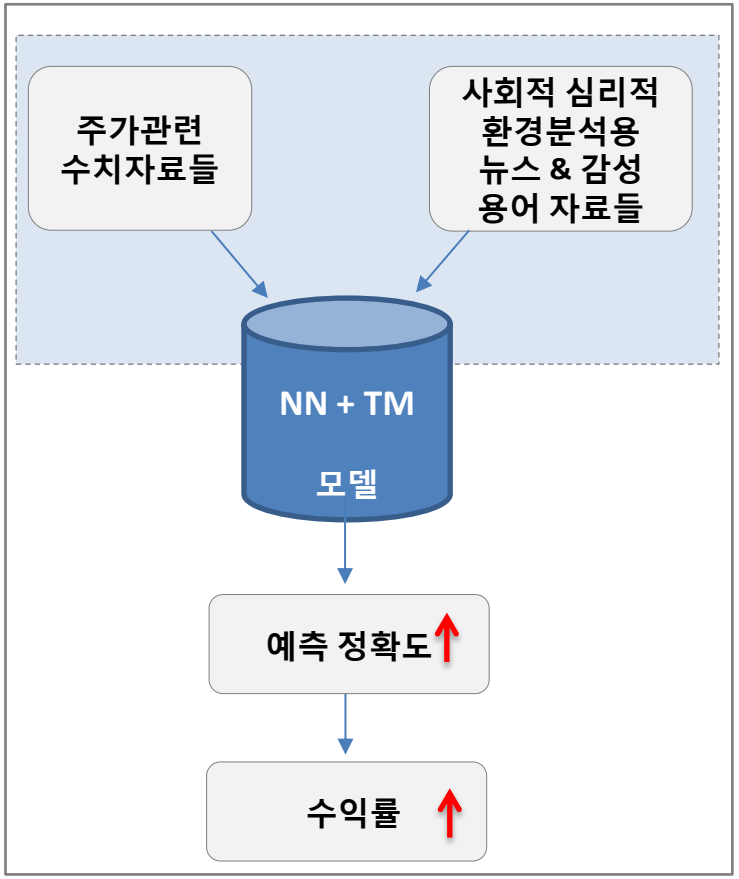
타모델과 비교분석 결과

	모형구분	학습용 정확도	검증용 정확도	수익률
1	TM 포함 RF	60.57%	55.37%	2.82%
2	TM 포함 SVM	56.39%	58.68%	5.18%
3	TM 포함 NN	72.54%	65.29%	17.37%

2.2. NN을 이용한 종합주가지수 예측

Contents

기술적 변수와 통합 감성사전의 긍부정지수를 입력변수로 통합한 최종 종합주가지수 예측 모형을 완성함



TM+SVM
최종모델
완성

✓ 기술적 분석과 텍스트마이닝 분석을 통합한 본 프로젝트의 최종 모델의 높은 예측정확도와 수익률로 검증되었듯이 뉴스용어사전을 이용을 통해 투자자들의 투자심리 변화를 보다 빨리 읽어냄으로써 주가의 방향성을 더욱 정교하게 예측한다고 할 수 있음