일반연구 2021-05

서포트벡터머신(SVM) 기법을 활용한 해운시황 예측 연구

A Study on Forecasting Shipping Market Conditions with Support Vector Machine(SVM)

김병주 · 이호춘 · 최건우 · 황수진 · 류희영

KMI 한국해양수산개발원
KOREA MARITIME INSTITUTE

저자	**김병주, 이호춘, 최건우, 황수진, 류희영**
내부연구진	김병주 한국해양수산개발원 해운·물류연구본부 연구원 이호춘 한국해양수산개발원 해운·물류연구본부 부연구위원 최건우 한국해양수산개발원 해운·물류연구본부 전문연구원 황수진 한국해양수산개발원 해운·물류연구본부 전문연구원 류희영 한국해양수산개발원 해운·물류연구본부 연구원

연구기간	2021. 5. 17. ~ 2021. 10. 31.

보고서 집필내역

연구책임자	김병주 연구총괄 제1장, 제4장, 제5장
내부연구진	이호춘 제5장 최건우 제3장 황수진 제3장 류희영 제1장, 제2장, 제3장

발간사

해운산업은 국제적인 속성을 지닌 산업으로 국내외적 환경에 민감하게 반응한다. 특히 부정기선의 운임은 선박 수요, 공급량에 의한 완전 경쟁 원리에 따라 결정되기 때문에 운임을 예측하기 어려운 특징을 지니고 있다. 또한 경기순환 주기적인 고유의 특성이 있으며, 이로 인한 위기는 여러 시기에 걸쳐 해운선사에 중대한 영향을 미친다. 해운 위기에는 다양한 요인이 복합적으로 작용하지만 역사적으로 경기 변동에 따른 해운시황의 예측 실패와 위험 관리가 주요한 원인으로 분석된다.

아울러 우리나라 해운 위기의 원인으로 취약한 시장 관리 능력이 항상 문제시되었다. 2008년 리먼브러더스 사태로 인해 세계 경제에 대한 불확실성과 변동성이 심화되면서 해운시황의 불확실성이 높아져 전통적인 기존 방법으로 해운산업 시황을 예측하는 데 한계가 발생하였다.

최근에는 정보통신기술과 IT산업의 확산으로 전통적인 예측 방법을 보완할 수 있는 인공지능(AI: Artificial Intelligence), 빅데이터(big data), 사물인터넷(IoT: Internet of Things) 등 방대한 정보를 활용한 방법이 대두되고 있다.이에 이번 연구에서는 기존에 활용되던 정량적인 자료에서 벗어나 새로운 자료의 활용 가능성을 검토하는 데 목적이 있다. 이를 위해 해운산업과 관련된 비정형 자료를 수집하였으며 이를 분석할 수 있는 새로운 툴로 서포트벡터머신(SVM: Support Vector Machine) 및 감성분석(sentiment analysis)을 도입하였다.

본 연구는 그간 수치적으로만 접근하였던 예측 방법에서 해운시장을 분석할 수 있는 새로운 자료와 방법을 제공하였다는 점에서 선도적인 연구임과 동시에 해운기업에 실제로 활용될 수 있는 연구라 할 수 있다. 해운기업에 필요한 시황 분석에 대한 연구는 일부 대형 선사를 제외하고 현장 영업에서 담당하고 있어 직관적으로 이루어지는 것이 현실이다. 이번에 실시한 서포트벡터머신 및 감성분석을 활용한 연구는 정량적으로 전문가들이 시장을 보는 시각을 제시하여 중요한 참고 자료로 활용할 것으로 기대된다.

끝으로 본 연구의 책임을 맡은 해운·물류연구본부 김병주 연구원과 연구진의 노고에 깊은 감사를 표한다. 아울러 본 보고서의 감리를 맡아 적절한 지적과 조언으로 완성도를 높여 주신 우리 원의 김태일 해운·물류연구본부장과 보고서가 출판될 수 있도록 도움을 주신 모든 분께 이 자리를 빌려 감사를 드린다.

2021년 12월

한국해양수산개발원

원장 김 종 덕

목차

요약 _i

Executive Summary _vii

01 서론 _1

제1절 연구 배경 및 목적 1
1. 연구 배경 1
2. 연구 목적 4

제2절 연구 내용 및 방법 5
1. 연구 범위 및 주요 내용 5
2. 연구 방법 6

02 선행연구 및 해운시장 비정형 자료 _9

제1절 선행연구 검토 9

제2절 해운시장 비정형 자료 개요 28
1. 기사 및 정기간행물 29
2. 해운시장 관련 연구기관 35
3. 해운중개업체 리포트 38

03 감성분석과 서포트벡터머신(SVM) _41

제1절 데이터마이닝과 텍스트마이닝 42
1. 데이터마이닝 42
2. 텍스트마이닝 45
3. 데이터마이닝과 텍스트마이닝의 차이 49

제2절 감성분석(Sentiment Analysis) 51
1. 감성분석 정의 51

2. 감성분석 방법 51

제3절 서포트벡터머신 56
1. 서포트벡터머신 정의 56
2. SVM 특징 및 장단점 59

04 SVM을 활용한 해운시장 분석 _61

제1절 자료 및 분석 방법 61
1. 자료 61
2. 분석 방법 65

제2절 분석 결과 70
1. 감성분석 71
2. SVM 분석 74

제3절 감성분석 및 SVM 예측 성공률 비교 81

05 결론 및 제언 _85

제1절 결론 85
1. 해운시황 예측에 비정형 데이터 분석 방법론 적용 85
2. SVM에 활용된 비정형 데이터는 '양'보다 '질'이 중요 86
3. 시황에 적합한 비정형 데이터 활용 필요 87
4. 단어의 중의성 해소 87

제2절 제언 88
1. 시장 심리를 측정한 해운시황 예측 방법 도입 및 서비스 제공 88
2. 다차원·다양한 연구 실험 필요 90
3. 체계적인 비정형 데이터 DB 구축 필요 91
4. 복합어 처리에 대한 해결 방안 필요 91

참고문헌 _93

표 목차

〈표 1-1〉 해운산업의 분산 ····· 2
〈표 1-2〉 기업 규모에 따른 선박 투자 의사결정에 미치는 요인 중요도 ····· 3
〈표 2-1〉 기업 부도 예측 방법론 요약 ····· 13
〈표 2-2〉 감성분석에 대한 주요 선행연구 ····· 16
〈표 2-3〉 감성분석을 활용한 주식시장 관련 선행연구 ····· 21
〈표 2-4〉 인공지능 기법을 적용한 해운시장 주요 선행연구 ····· 25
〈표 2-5〉 로이즈리스트 정보 제공 목록 ····· 30
〈표 2-6〉 PR News 정보 제공 서비스 목록 ····· 32
〈표 2-7〉 트레이드윈즈 정보 제공 목록 ····· 33
〈표 2-8〉 JOC 정보 제공 목록 ····· 35
〈표 3-1〉 전통적 통계분석과 데이터마이닝 ····· 43
〈표 3-2〉 텍스트마이닝의 단계 및 내용 ····· 48
〈표 3-3〉 데이터마이닝과 텍스트마이닝 비교 ····· 50
〈표 3-4〉 감성에 따른 단어(예) ····· 56
〈표 3-5〉 SVM의 장단점 ····· 60
〈표 4-1〉 실험 테스트 및 예측 샘플 수 ····· 64
〈표 4-2〉 감성분석 R 코드 ····· 66
〈표 4-3〉 SVM 분석 R 코드 ····· 69
〈표 4-4〉 Bing 감성사전(예) ····· 71
〈표 4-5〉 감성분석을 활용한 컨테이너 시황 예측 결과(로이즈리스트) ····· 72
〈표 4-6〉 감성분석을 활용한 건화물 시황 예측 결과(로이즈리스트) ····· 73
〈표 4-7〉 감성분석을 활용한 컨테이너 시황 예측 결과(PR News) ····· 74
〈표 4-8〉 로이즈리스트 문자 전처리 결과(컨테이너 기사) ····· 75
〈표 4-9〉 SVM을 통해 분류된 긍정어 및 부정어(컨테이너 기사) ····· 76
〈표 4-10〉 SVM 기법을 활용한 컨테이너 시황 예측 결과(로이즈리스트) ····· 76
〈표 4-11〉 로이즈리스트 문자 전처리 결과(건화물 기사) ····· 77
〈표 4-12〉 SVM을 통해 분류된 긍정어 및 부정어(건화물 기사) ····· 78
〈표 4-13〉 SVM 기법을 활용한 건화물 시황 예측 결과(로이즈리스트) ····· 78
〈표 4-14〉 PR News 문자 전처리 결과(컨테이너) ····· 79
〈표 4-15〉 SVM을 통해 분류된 긍정어 및 부정어(PR News) ····· 80

〈표 4-16〉 SVM 기법을 활용한 컨테이너 시황 예측 결과(PR News) ····· 80
〈표 4-17〉 컨테이너 시황 예측 결과(로이즈리스트) ····· 81
〈표 4-18〉 건화물 시황 예측 결과(로이즈리스트) ····· 82
〈표 4-19〉 컨테이너 시황 예측 결과(PR News) ····· 83
〈표 5-1〉 기업 규모에 따른 선박투자 의사결정에 미치는 요인 중요도 ····· 86
〈표 5-2〉 SVM 기법을 활용한 예측 결과(로이즈리스트 vs PR News) ····· 86
〈표 5-3〉 감성분석 vs SVM ····· 87
〈표 5-4〉 이론적 및 산업적 기여 내용 ····· 89

그림 목차

〈그림 2-1〉 연도별 감성분석 연구 동향 ····· 11
〈그림 2-2〉 감성사전 및 SVM 기반 분류 방법 ····· 27
〈그림 2-3〉 PR News 정보 제공 서비스 구성 ····· 31
〈그림 2-4〉 트레이드윈즈 제공 정보 구성 ····· 33
〈그림 2-5〉 JOC 제공 정보 구성 ····· 34
〈그림 2-6〉 드류리 발행 정기 보고서 ····· 36
〈그림 2-7〉 알파라이너 발행 월간 보고서 및 주간 보고서 ····· 37
〈그림 2-8〉 발틱거래소 정보 제공 ····· 38
〈그림 2-9〉 펀리스 정보 제공 주간 보고서 ····· 39
〈그림 3-1〉 데이터마이닝 과정 ····· 42
〈그림 3-2〉 텍스트마이닝 Areas ····· 47
〈그림 3-3〉 감성분석 시스템 순서도 ····· 53
〈그림 3-4〉 감성분석 방법론 구분 ····· 54
〈그림 3-5〉 서포트벡터머신 ····· 58
〈그림 4-1〉 로이즈리스트 컨테이너 기사 ····· 62
〈그림 4-2〉 감성분석 흐름도 ····· 65
〈그림 4-3〉 SVM 흐름도 ····· 68
〈그림 4-4〉 SCFI 및 BDI(주 단위) ····· 70
〈그림 4-5〉 SVM 결정경계 ····· 74
〈그림 5-1〉 연구 결과 활용 ····· 90

약어 목록

- AI: 인공지능(Artificial Intelligence)
- AIS: 선박 자동식별 장치(Automatic Identification System)
- ANN: 인공 신경망(Artificial Neural Network)
- ARIMA: 자동회귀 누적 이동 평균(Auto Regressive Integrated Moving Average)
- BCI: 발틱 케이프선 운임지수(Baltic Capesize Index)
- BCTI: 발틱 석유제품선 운임지수(Baltic Clean Tanker Index)
- BDI: 발틱 건화물선 운임지수(Baltic Dry Index)
- BDTI: 발틱 원유유조선 운임지수(Baltic Dirty Tanker Index)
- BERT: 구글의 인공지능 언어모델(Bidirectional Encoder Representations from Transformers)
- BPI: 발틱 파나막스선 운임지수(Baltic Panamax Index)
- BSI: 발틱 수프라막스선 운임지수(Baltic Supramax Index)
- CCFI: 중국 컨테이너 운임지수(China Containerized Freight Index)
- CRM: 고객관계 관리(Customer Relationship Management)
- DNN: 심층 신경망(Deep Neural Network)
- FBX: 세계 컨테이너 화물 운임지수(Freights Baltic Index)
- JOC: Journal of Commerce
- KNN: K-최근접 이웃 알고리즘(K-nearest neighbors)
- LSTM: 장단기 메모리(Long Short-Term Memory)
- MLP: 다층 신경망(Multi Layer Perception)
- NLP: 자연어 처리(Natural Language Processing)
- RF: 랜덤 포레스트(Random Forest)
- RMSE: 평균제곱근오차(Root Meen Square Error)
- RNN: 순환 신경망(Recurrent Neural Network)
- SCFI: 상하이 컨테이너 운임지수(Shanghai Containerized Freight Index)
- SVM: 서포트벡터머신(Support Vector Machine)
- XML: 확장성 생성 언어(eXtensible Markup Language)

요약

1. 연구 배경 및 목적

- **해운산업에서는 여전히 업계 전문가들의 경험이 쌓인 노하우를 통해 직관에 의해 많은 거래가 이루어지고 있음**
  - 직관적 분석이 과학적인 분석 방법보다 우수할 수 있겠지만, 시장을 바라보는 관점에 차이가 발생할 수 있으므로 시장을 객관화할 수 있는 정량적인 방법론이 필요함

〈요약 표 1〉 기업 규모에 따른 선박 투자 의사결정에 미치는 요인 중요도

저자	대기업	중소기업
과학적 분석	0.574	0.491
휴리스틱(직관)	0.345	0.293
시장 정서	0.081	0.216
합계	1.000	1.000

- **전통적인 계량분석 방법에서 벗어나 기존에 활용되지 않은 자료(보고서, 기사 등)를 이용해 해운시황 예측 방법을 실시함.**
  - '시장 심리'를 정량화한 예측 기법을 도입해 변동성이 높은 해운시황의 예측 정확도를 높이고 신뢰도를 향상시키고자 함

2. 연구 특징 및 방법

- 기존 해운시황을 예측하는 연구는 정량적 데이터를 활용한 전통적인 시계열 방법이나 인공지능기법(LSTM, DNN 등)이 대부분임
  - 이와 같은 방법론은 과거 및 현재의 정량적 데이터 특징이나 패턴을 분석하여 미래의 해운시황을 예측하는 방법임
- 본 연구에서는 비정형 데이터를 정량화(객관화)할 수 있는 SVM 분석 기법을 통해 시장 심리를 측정한 해운시황 예측 방법을 도입함
  - 본 연구에서는 비정형 데이터를 기반으로 감성분석과 SVM 기법을 활용해 해운시황 예측 및 성능 비교를 실시함

3. 연구 결과

1) 로이즈리스트를 활용한 예측 결과

- 로이즈리스트 컨테이너 시황[SCFI(종합)]의 감성분석 및 SVM 방법을 활용한 예측 실험 결과, 감성분석 예측 성공률이 51.7%, SVM 예측 성공률이 58.6%를 기록함

〈요약 표 2〉 컨테이너 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총예측 횟수	감성분석 예측 성공 횟수(성공률)	SVM 예측 성공 횟수(성공률)
SCFI(종합)	29(100.0)	15(51.7)	17(58.6)
SCFI(중국~미국 서안)	29(100.0)	17(58.6)	15(51.7)
SCFI(중국~유럽)	29(100.0)	17(58.6)	14(48.2)

- SCFI 세부 항로 지표에 대한 예측 성공률에서는 감성분석이 SVM 방법론 대비 모두 소폭 높은 예측 성공률을 기록함

▪ **로이즈리스트 건화물 시황(BDI)의 감성분석 및 SVM 방법을 활용한 예측 실험 결과, 감성분석 예측방법이 SVM 예측방법 대비 약 18%p 높은 것으로 나타남**

- 결론적으로 로이즈리스트 건화물 시황의 비정형 데이터 분석은 감성분석 방법이 SVM 방법론과 비교하여 더 적합한 모델이라 판단할 수 있음

〈요약 표 3〉 건화물 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총예측 횟수	감성분석 예측 성공 횟수(성공률)	SVM 예측 성공 횟수(성공률)
BDI	35(100.0)	21(60.0)	15(42.9)
BCI	35(100.0)	19(54.3)	15(42.9)
BPI	35(100.0)	23(65.7)	16(45.7)
BSI	35(100.0)	24(68.6)	15(42.9)

2) PR News를 활용한 예측 결과

▪ **PR News 컨테이너 시황[SCFI(종합)]의 감성분석 및 SVM 방법을 활용한 예측 실험 결과, 성공률은 감성분석이 37.5%, SVM이 74.3%를 기록함**

- SCFI 세부 항로 지표에 대한 예측 성공률 또한 SVM 방법론이 감성분석 대비 모두 높게 나타나, PR News에 더 적합한 모델이라 판단됨

〈요약 표 4〉 컨테이너 시황 예측 결과(PR News)

단위: 개, (%)

구분	총예측 횟수	감성분석 예측 성공 횟수(성공률)	SVM 예측 성공 횟수(성공률)
SCFI(종합)	35(100.0)	13(37.5)	26(74.3)
SCFI(중국~미국 서안)	35(100.0)	17(48.6)	23(65.7)
SCFI(중국~유럽)	35(100.0)	16(45.7)	21(60.0)

4. 결론 및 제언

▪ SVM에 활용된 비정형 데이터는 '양'보다 '질'이 중요함

- 분류된 단어 수는 로이즈리스트가 PR News 대비 약 4천 개 많은 것으로 나타났지만, 예측 성공률은 PR News를 활용한 수치가 74.3%로, 로이즈리스트 대비 약 두 배 높은 수준을 기록함
- 결론적으로 SVM에 활용된 비정형 데이터는 자료의 '양'과 예측 성공률이 비례하지 않는 것으로 나타남.

〈요약 표 5〉 SVM 기법을 활용한 예측 결과(로이즈리스트 vs PR News)

단위: 개, %

구분	분류된 단어 수 (중복 단어 제외)	SCFI 예측 성공률
로이즈리스트	9,569	37.5
PR News	5,565	74.3

▪ 시황에 적합한 비정형 데이터 활용 필요

- 감성분석과 SVM을 활용한 예측 실험 결과는 특정 분석 방법이 모든 시황 지표에서 우수한 결과를 기록하지는 않는다는 것을 보여 줌

- 결론적으로 비정형 데이터 활용한 해운시황 예측에서는 시황에 적합한 데이터를 탐색하는 작업 또한 매우 중요한 것으로 판단됨

〈요약 표 6〉 감성분석 방법 vs SVM 방법

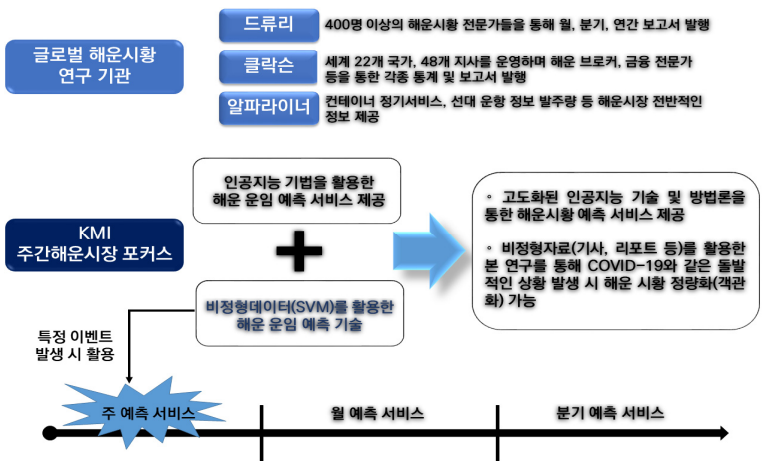
단위: %

예측 대상	감성분석 예측 성공률	SVM 예측 성공률
SCFI(로이즈리스트)	51.7	58.6
SCFI(PR News)	37.5	74.3
BDI(로이즈리스트)	60.0	42.9

■ 시장 심리를 측정한 해운시황 예측 방법 도입 및 서비스 제공

- 본 연구 결과에서 도출된 시장 심리를 정량화(객관화)하는 기법은 한국해양수산개발원에서 제공하는 『KMI 주간해운시장 포커스』에 해운시황 예측 서비스로 활용될 수 있음
- 특히 코로나19와 같이 돌발적인 이벤트가 발생했을 때 실시간으로 배포되는 비정형 데이터(기사, 리포트 등)의 정량화는 단기적인 예측 서비스의 질적 향상에 큰 도움이 될 것으로 기대됨

〈요약 그림 1〉 연구 결과 활용



EXECUTIVE SUMMARY

1. Purpose and Background

- A large number of transactions have been concluded in the shipping industry based on heuristic intuition from know-how accumulated through years of experience from industrial experts.
  - To a certain extent, intuitive analyses may exceed the capabilities of science-based analyses. As perspectives looking at the market may vary, a quantitative methodology is required to maintain objectivity when analyzing the market.

〈Summary Table 1〉 Importance of factors on ship investment decisions by corporate size

Author	Large Corporation	SME
Scientific analysis	0.574	0.491
Heuristic analysis (intuition)	0.345	0.293
Market sentiment	0.081	0.216
Total	1.000	1.000

- Departing from traditional quantitative analyses, this study intends to present a method for forecasting market conditions of the shipping industry with data that has yet to be utilized thus far (reports, news articles, etc.).
  - With the adoption of a forecasting methodology of quantifying 'market sentiment', this study aims to improve the accuracy as well as the reliability of forecasts surrounding shipping market conditions which have been considered rather volatile.

2. Methodology and Features

- Most of the existing studies on forecasting shipping market conditions include traditional methods including time-series analyses or artificial intelligence-based methods (LSTM, DNN), utilizing quantitative data.
  - These methodologies were used to forecast the future prognosis of the shipping market by analyzing characteristics and patterns of past and present quantitative data.
- This study differentiates itself from existing studies as it adopts a forecasting method of shipping market conditions by measuring market sentiments via a SVM analysis method.
  - Utilizing sentiment analysis and SVM methods based on unstructured data, this study intends to predict market conditions of the shipping industry, while comparing performance.

3. Results

1) The result of forecasts utilizing Lloyd's List

- According to the result of testing forecasts utilizing a sentiment analysis based on Lloyd's List Container Market (SCFC (Composite)) against an SVM-based analysis, the success rate of the sentiment analysis stood at 51.7%, while SVM-based forecasts stood at 58.6%.
  - As for the success rate for forecasting specific indicators of SCFI by sea route, the sentiment analysis recorded a modestly higher success rate than the SVM-based methodology in all indicators.

〈Summary Table 2〉 The result of container market forecasts (Lloyd's List)

Unit: unit of cases, (%)

Category	Total number of predictions	The Number of successes by forecasts based on sentiment analysis (success rate)	The Number of successes by SVM forecast (success rate)
SCFI (Composite)	29(100.0)	15(51.7)	17(58.6)
SCFI (China~ west coast of the US	29(100.0)	17(58.6)	15(51.7)
SCFI (China ~ Europe)	29(100.0)	17(58.6)	14(48.2)

- The result of testing the forecasts based on Lloyd's List Dry Bulk Market (BDI) against the SVM-based method showed that the success rate of sentiment analysis was 18% points higher than the SVM method.

- In conclusion, the analysis of unstructured data from Lloyd's List Dry Bulk Market (BDI) indicates that the sentiment analysis method is more suitable than the SVM-based method.

〈Summary Table 3〉 The result of dry bulk market forecasts (Lloyd's List)

Unit: unit of cases, (%)

Category	Total number of predictions	The Number of successes by forecasts based on sentiment analysis (success rate)	The Number of successes by SVM forecast (success rate)
BDI	35(100.0)	21(60.0)	15(42.9)
BCI	35(100.0)	19(54.3)	15(42.9)
BPI	35(100.0)	23(65.7)	16(45.7)
BSI	35(100.0)	24(68.6)	15(42.9)

2) The result of PR News-based forecasts

- According to the result of testing forecasts utilizing a sentiment analysis based on PR News Container Market (SCFC (Composite)) against an SVM-based analysis, the success rate of sentiment analysis stood at 37.5%, while SVM-based forecasts stood at 74.3%.
- The success rate of SVM-based forecasts was found to be superior to that of forecasts based on sentiment analysis in all indicators of SCFI by sea route, meaning that SVM was found to be a more befitting model for PR News in forecasting market conditions.

〈Summary Table 4〉 The result of container market forecasts (PR News)

Unit: unit of cases, (%)

Category	Total number of predictions	The Number of successes by forecasts based on sentiment analysis (success rate)	The Number of successes by SVM forecast (success rate)
SCFI (Composite)	35(100.0)	13(37.5)	26(74.3)
SCFI (China~ west coast of the US	35(100.0)	17(48.6)	23(65.7)
SCFI (China ~ Europe)	35(100.0)	16(45.7)	21(60.0)

4. Conclusion and policy suggestions

- ‘Quality’ is more important than ‘quantity’ for unstructured data utilized in SVM

- With regards to the number of classified words, Lloyd's List has approximately 4,000 more than PR News. However, the success rate of PR News-based forecasts was recorded to be 74.3%, twice that of Lloyd's List-based predictions.

- In conclusion, what is important for unstructured data utilized in SVMs is ‘quality’, rather than ‘quantity’ to apply the data more befitting to market condition.

〈Summary Table 5〉 The result of SVM-based forecasts (Lloyd's List vs PR News)

Unit: unit of words, %

Category	The number of classified words (excluding duplication)	Success rate of SCFI forecasts
Lloyd's List	9,569	37.5
PR News	5,565	74.3

▪ Necessity for utilizing unstructured data suitable for market conditions

- The result of testing the forecasts based on sentiment analysis and SMVs show that there is no specific methodology that dominantly excels in all indicators of market conditions.
- Conclusively, it is extremely important to search and find the data befitting to market condition surrounding the shipping industry in making unstructured data-based predictions.

〈Summary Table 6〉 Sentiment analysis vs SVM

Unit: %

Forecasting target	Success rate of sentiment analysis	Success rate of SVM forecasts
SCFI (Lloyd's List)	51.7	58.6
SCFI (PR News)	37.5	74.3
BDI (Lloyd's List)	60.0	42.9

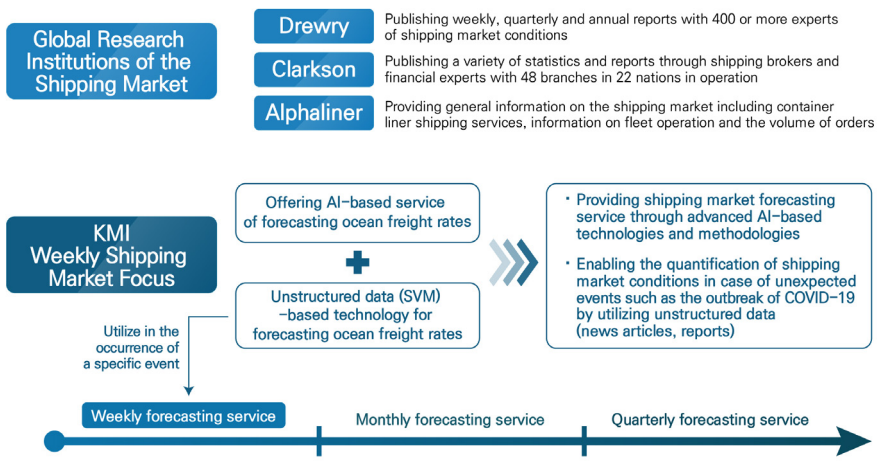
▪ Introduction of a methodology and service for forecasting shipping market conditions by measuring market sentiment

- Methods that quantify (maintain objectivity) market sentiment

extracted from this study can be utilized in 'Shipping Market Forecasts' of 'Weekly Shipping Market Focus' published by KMI.

- When an unexpected event occurs such as the breakout of COVID-19, in particular, quantification of unstructured data (news articles, reports) distributed in real-time will significantly improve the quality of short-term forecasting service.

〈Summary Picture 1〉 Utilization of study results



01
서론

제1절 연구 배경 및 목적

1. 연구 배경

해운산업은 국제적인 속성을 지닌 산업으로 국내외적 환경에 민감하게 반응한다. 특히 부정기선의 운임은 선박 수요, 공급량에 의한 완전 경쟁 원리에 따라 결정되기 때문에 운임을 예측하기 어려운 특징을 지니고 있다. 또한 해운산업은 경기순환 주기적인 고유의 특성이 있으며, 이로 인한 위기는 여러 시기에 걸쳐 해운선사에 중대한 영향을 미치고 있다. 해운 위기에는 다양한 요인이 복합적으로 작용하지만 역사적으로 경기 변동에 따른 해운시황의 예측 실패와 위험 관리가 주요한 원인으로 분석된다.

아울러 해운산업은 선박이 수송하는 화물 특징, 입·출항지 상황 등 다양한 요인으로부터 영향을 받는 파생수요 산업으로 변동성이 큰 구조적 특징을 지니고 있다. 특히 해운산업은 거래 비용의 변동성이 매우 크며, 변동성이 크다는 의미는 평균에서 많이 떨어져 있다는 뜻으로 분산이 큰 것으로도 표현할 수 있다.[1)]

1) 해양수산부 블로그(검색일: 2021. 8. 21)

〈표 1-1〉 해운산업의 분산

INDEX(지표)	분산(%)
ClarkSea(클락슨평균운임지수)	**34.68**
Baltic Dry Index(벌크운임지수)	**53.09**
Bulker Secondhand Price(벌크중고선가지수)	31.98
Bulker Newbuilding Price(벌크신조선가지수)	7.54
Stock Marker Index(주식)	**15.73**
Government Bond Index(국채지수)	6.59

자료: Stefan Albertijn et al.(2011), 해양수산부 블로그 재인용(검색일: 2021. 8. 21)

Stefan Albertijn et al.(2011)은 해운시장의 분산이 34.68%(클락슨평균운임지수 기준) 및 53.09%(벌크운임지수 기준)로 주식의 분산 15.73% 대비 매우 큰 편이며, 이로 인해 해운시장의 운임 변동 폭도 심하다고 설명했다.[2] 이처럼 해운시장은 변동성이 크며, 이를 예측하기는 매우 힘들다.

따라서 해운산업을 안정적으로 지속하기 위한 해운시황 예측은 중요한 과제로 인식되고 있으며 많은 연구들이 발표되고 있다. 하지만 시대가 점차 변함에 따라 해운산업의 변동성, 불확실성, 복잡성이 더욱 확대되어 시계열 등 전통적인 해운시황 분석 방법은 한계가 발생할 가능성이 존재하며, 이에 기존의 전통적·일반적 예측 방법을 보완할 수 있는 인공지능, 빅데이터 등을 활용한 분석 방법이 대안으로 대두되고 있다.

〈표 1-2〉와 같이 대기업과 중소기업 모두 선박 투자 시 과학적 분석이 의사결정에 미치는 중요도가 가장 큰 것으로 나타났다. 하지만 직관적인 판단과 시장 정서가 의사결정에 미치는 영향이 여전히 큰 상황으로, 의사결정으로 인한 시장 위험을 최소화하기 위해서는 해운시황을 예측하는 방법론의 고도화로 과학적 분석의 질을 향상시킬 필요가 있다.

2) Stefan Albertijn et al.(2011), pp. 70- 82, 해양수산부 블로그 재인용(검색일: 2021. 8. 21)

〈표 1-2〉 기업 규모에 따른 선박 투자 의사결정에 미치는 요인 중요도

저자	대기업	중소기업
과학적 분석	0.574	0.491
휴리스틱(직관)	0.345	0.293
시장 정서	0.081	0.216
합계	1.000	1.000

자료: 김경수·박은식·우수한(2014), p. 183

4차 산업혁명 시대가 도래하면서 대부분의 산업 분야는 빅데이터, 인공지능을 활용할 수 있게 되었으며 기존에는 활용이 어려웠던 데이터와 방법론을 시장 분석에 활용할 수 있게 되었다.

기존의 계량경제학에 기반한 방법론들이 변수들 간의 인과관계 증명을 목적으로 정형적 모형 제시를 추구했다면 최근에는 빅데이터를 활용한 인공지능, 기계학습, 딥러닝 등과 같은 새로운 기법은 변수들 간의 다차원 상관관계 파악을 통해 예측(forecasting) 정확도를 향상시키고 있다. 시장 예측과 관련하여 많은 선행 연구가 수행된 금융 분야는 특히 이러한 새로운 기법들을 신속하게 받아들여 접목하고 있다. Shen et al.(2012)[3]은 주식시장 예측에 서포트벡터머신(SVM: Support Vector Machine) 기반의 기계학습 알고리즘을 적용하여 타 분석법 대비 향상된 예측 정확도를 보였다.

해운 분야에서도 텍스트마이닝, 감성분석 등을 활용하여 매월 발행되는 관련 산업 기사와 SNS 자료 등을 활용하여 시장 현황이나 참여자들의 심리, 태도 등을 판별할 수 있으며 텍스트를 활용하여 시황을 예측할 수 있다.

전술한 바와 같이 해운산업에 영향을 주는 위험은 가격 변동성 및 노출 크기가 큰 영향을 미친다. 가격의 변동성은 유가, 용선료, 운임 등 외부에

3) Shen, Jiang, and Zhang(2012), pp. 1-5.

서 주어지는 변수이며 일반적으로 통제가 불가능한 요소들이다. 노출의 크기는 선박 발주, 용선 시점과 기간 등 해운선사의 의사결정으로 조정이 가능한 영역이다. 기업의 의사결정은 기업 지배구조, 의사결정 과정, 시장 분석 역량에 따라 상이하며 의사결정에 따라 통제가 불가능한 리스크를 최소화하거나 회피 또는 전가할 수 있다.[4)]

해운산업의 위험 관리에 필요한 의사결정은 기간, 시점 등이 중요한 변수로 여겨진다. 의사결정을 올바르게 할 수 있는 능력이 기업의 재무 건전성과 경영 안정을 도모할 수 있으며, 해운산업의 국가 경쟁력과도 연결될 수 있다. 따라서 잘못된 의사결정으로 인한 위험을 최소화하기 위해서 정확도가 높은 분석 능력을 갖춘 정밀한 해운시장 예측 연구가 필요하다.

2. 연구 목적

해운시황을 예측하는 대표적인 방법론으로는 시장의 수요·공급과 관련된 시계열 자료와 같은 정량적인 데이터를 기반으로 예측을 하는 시계열분석 연구가 전통적인 방법론이라고 할 수 있다. 또한 최근 인공지능(AI: Artificial Intelligence)의 발전으로 데이터를 분석하고 학습한 정보를 바탕으로 의사결정을 내리는 머신러닝(Machine Learning)과 알고리즘을 계층으로 구성하고 자체적인 학습 및 결정을 내리는 딥러닝(Deep Learning)과 같은 새로운 방법론이 해운시황 예측 연구에 적용되어 다수 발표되고 있다. 하지만 이러한 연구 방법론들은 정량적 자료를 활용한 방법론으로서 실시간으로 자료를 수집하는 것이 불가능하기 때문에 외부에서 발생하는 충격으로 인한 이벤트 발생 시 단기적인 시장 흐름을 파악할 수 없는 단점이 있다.

4) KMI 동향분석(2018), 빅데이터 분석은 해운에서 어떻게 활용되는가, 통권 제90호

본 연구는 전통적인 정량적 데이터를 활용한 계량분석 방법에서 벗어나 기존에 활용되지 않은 비정형 데이터(보고서, 기사 등)를 이용해 해운시황 예측을 실시하고자 한다. 이를 통해 예측의 정확도 제고 및 비정형 데이터를 활용한 예측 방법론의 적용 가능성을 살펴볼 것이다. 특히 코로나19로 인해 해운시황에 예상치 못한 큰 변화가 발생해 해운시황의 변동성이 커짐에 따라 단기적인 시장 흐름을 파악하는 데 있어 전문가들의 견해가 담긴 비정형 데이터의 가치는 더욱 높아졌을 것이라 예상된다.

아울러 '시장 심리'가 잘 반영된 해운 업계 전문가들의 비정형 데이터(기사, 리포트 등) 활용은 계량 자료가 가지는 분석 시점과 활용 자료의 시간 차이가 발생하는 단점을 보완할 수 있을 것으로 기대된다. 이를 위해 SVM을 활용한 인공지능 예측 방법을 도입하여 기사 및 보고서 등 해운시황 전문가들의 의견이 반영된 비정형 데이터를 활용한 해운시황 예측 연구를 하는 데 목적을 두고 있다.

제2절 연구 내용 및 방법

1. 연구 범위 및 주요 내용

본 연구의 주요 목적은 '시장 심리' 예측 기법을 도입해 변동성이 높은 해운시황 예측의 정확도를 제고하는 데 있다. 이를 위해 전통적인 계량분석 방법에서 벗어나 기존에 활용되지 않은 자료(보고서, 기사 등)를 이용한 해운시황 예측 방법을 실시한다.

본 연구는 텍스트마이닝 기법을 적용한 해운시황 예측 신뢰도 제고를 위해 SVM을 활용한 비정형 데이터의 지도 학습 기반 인공지능 예측 방법을 도입한다. 이를 통해 기사 및 보고서 등 해운시장 전문가 의견, 시장 현황 등이 실제 해운시황 예측에 잘 반영되는지 성능 테스트를 진행한다.

국적 선사를 대상으로 더 정확한 해운시황 예측 정보를 제공하기 위해 본 연구를 통해 고도화된 감성분석 방법론을 지속적으로 보완·발전시켜 해운선사를 비롯한 관계 기관에 단기 해운시황 정보를 제공하고자 한다.

주요 연구 내용을 살펴보면 제2장에서는 해운시장이 반영된 비정형 자료에 대한 구분과 개요를 정리하였으며, 감성분석과 SVM을 활용한 국내외 선행연구들에 대해 살펴본다. 제3장에서는 데이터마이닝과 텍스트마이닝, 감석분석과 SVM의 정의, 특징을 정리한다. 제4장에서는 감성분석 및 SVM에 적용할 해운시장 분석 자료와 분석 방법을 설명하고 분석 결과에 따라 해운시황 예측에 활용 가능한지를 검토한다.

마지막 제5장에서는 앞선 분석 결과를 바탕으로 SVM 기법을 활용한 해운시황 예측 연구와 관련된 결론과 시사점을 도출한다.

2. 연구 방법

본 연구에서는 국내외의 활용 가능한 해운 관련 문헌 조사 수집을 1차적으로 수행한다. 로이즈리스트, PR News, JOC(Journal of Commerce), 블룸버그 등 비정형 자료에 대한 조사를 진행하였으며, 그중 분석에 적합하다고 판단되는 해운 전문 기사 자료(로이즈리스트, PR News)의 텍스트 데이터를 분석에 활용하였다.

다음으로, 감성분석과 SVM을 활용한 해운시황 예측 모형 개발을 위해 수집된 자료를 뜻을 지닌 최소 단위인 형태소로 분리하였으며, 품사로 태깅(tagging)하여 정리하였다. 전치사나 관사와 같이 특정한 의미가 없는 어휘와 특수문자 및 기호는 필터링하여 토큰화(tokenization)하였다.

마지막으로, 전처리 작업을 거친 비정형 데이터를 활용해 감성분석과 SVM 기법을 통한 해운시황 예측을 실시하여 감성분석과 SVM의 예측 성능을 비교한다.

02
선행연구 및 해운시장 비정형 자료

제1절 선행연구 검토

IBM 보고서(2015)에 의하면 전 세계 데이터 약 80%가 텍스트(text)와 같이 별도로 정의된 구조가 없이 정형화되지 않은 비정형 데이터(unstructured data)로 구성되어 있다.[5)]

우리가 매일 접하는 언어, 텍스트 등에는 광범위한 정보가 포함되어 있으나, 단시간에 수많은 문서가 생산되는 현대사회에서는 단순히 인지능력만으로 광범위한 텍스트 자료를 처리하는 데 한계가 있다. 이제 우리는 텍스트마이닝을 이용하여 방대한 텍스트 데이터(text data)를 자동으로 분석할 수 있게 되면서 효율적으로 유용한 정보를 취득할 수 있다.[6)]

텍스트 데이터는 여러 분야의 연구에도 활용되고 있는데, 대표적으로 문헌정보학 분야에서 오래전부터 연구 대상이 되어 왔다. 또한 정치학, 사회학이나 공학, 의학계열 등 여러 학문에서도 다양한 분야에서 활용되고 있으며, 특히 공학분야에서는 문서 집합의 추상적인 주제(topic)를 분석하기

5) IBM(검색일: 2021.10. 9)
6) 김수현 외(2020년), p. 1.

위한 통계적 모델 기법인 토픽 모델링(topic modelling)으로 과거와 최신 연구동향을 파악하여 향후 진행할 연구 주제를 탐색하기도 한다.[7]

텍스트마이닝은 이미 우리 생활과 아주 밀접한 관계를 맺고 있다. 논문을 특정 주제어로 검색하거나 자료, 도서 등을 검색할 때에도 텍스트마이닝으로 정보가 요약된 데이터베이스를 사용하고 있다.

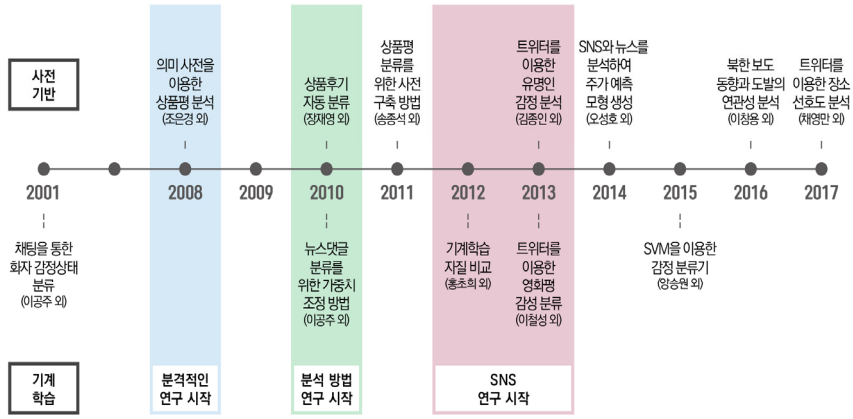
이 외에도 신제품을 출시할 때, 고객 리뷰의 감성분석을 통해 신제품의 초기 반응이나 품질을 검토하기도 한다. 전화, 인터넷 등 유무선 통신으로 이용 서비스나 상품 질에 대한 고객 불만 등이 제기될 때에도 해당 자료를 분석하여 향후 서비스 품질 개선에 활용한다. 소셜 미디어에서는 이용자들의 댓글이나 게시글 데이터 분석을 통해 잠재적으로 발생 가능성이 있는 사이버 범죄(cyber crime)를 예방할 수도 있다. 또한 다수의 텍스트 데이터 분석 결과를 마케팅이나 홍보에 활용할 수도 있다. 일부 소셜 미디어(social media)에서는 과거에 찾은 검색어나 주제 관련 광고가 특정되어 제공되기도 하는데, 이러한 광고 방식 또한 텍스트 데이터를 활용한 검색어, 연관검색어를 통해 밀접한 서비스나 상품을 선별해주는 알고리즘에 기반한 것이다.[8]

이처럼 텍스트마이닝은 문서 요약(sammarization), 문서분류(classification), 문서군집(clustering), 특성추출(feature extraction) 등의 기능을 가지고 있어 정보 검색과 추출, 시각화, 패턴 분석 등 여러 분야에서의 연구 동향을 파악할 수 있는 유익한 수단으로 여겨지며 다양한 분야에서 활용되고 있다.

이 보고서에서 활용하고자 하는 텍스트마이닝의 한 종류인 감성분석과 관련된 연구 또한 1990년대 후반부터 본격적으로 시작되어 현재에 이르기까지 다양한 방면에서 중요한 연구 분야로 자리 잡고 있다.

7) 노설현(2020), pp. 75~87.
8) 이재윤 외(2007), pp. 346~369.

〈그림 2-1〉 연도별 감성분석 연구 동향



자료: 이정훈(2018), p. 359

한국에서 감성이라는 키워드가 처음 언급된 논문은 이구형(1993)의 논문으로 감성공학 정의를 설명하고 연구 방법 및 활용 방안을 제시한 논문이었으며, 여기에서는 감성공학을 인간의 감정을 공학적으로 연구하고 활용하는 학문으로 정의하였다.[9)]

감성분석은 2001년 인터넷 채팅(chatting)의 감정 분석을 시도하면서 시작되었으며, 이후 2008년부터 본격적으로 감성분석에 대한 연구가 진행되었다. 감성분석은 SNS, 뉴스기사, 영화평, 상품 후기 등 여러 분야의 데이터에 적용되며, 장소에 대한 선호도 분석, 사회 이슈에 대한 찬반 분석 등 다양한 연구에 사용되고 있다. 해외에서는 이미 감성분석과 관련된 연구가 많이 진행되었으며, 실제 서비스에 적용되고 있는 단계이다. 영어는 한국어보다 오랜 기간 동안 연구되었기 때문에 감성분석의 핵심인 감성사전이 방대하고 정교하게 구축되어 있다.[10)]

9) 이구형(1993), pp. 59~63.
10) 이정훈(2018), pp. 358~361.

현재 감성분석은 금융시장, 문화관광산업, 스포츠산업, 마케팅 등 다양한 분야에 걸쳐 널리 활용되고 있다.

윤원중 외(2009)는 금융위기 발생 이전에 위험성을 알려주는 장치인 조기경보시스템에 대하여 연구하였다. 1997년 외환위기 기간의 코스피(KOSPI)를 대상으로 신호처리기법을 적용하여 의미 있는 특징들을 추출한 후 SVM을 이용하여 학습하였다. 학습된 SVM 분류기에 1994년부터 2019년까지의 코스피를 적용하여 실험한 결과, 주식시장이 영향을 받았던 금융위기 구간을 잘 검출해 내는 모습을 확인할 수 있었다.[11]

박선 외(2011)는 적조에 대한 기존 연구가 대부분 화학적 특성이나 생물학적 원인 규명에 초점이 맞춰 연구되고 있는 데 반해, 역전파 신경망과 회귀신경, SVM에 10일간의 수온, 기온, 강수량을 적용하여 미래에 발생할 적조를 예측하는 새로운 방법을 제안하였다. 분석 결과, SVM이 평균 86.9%의 정확률을 기록하며 역전파 신경망이나 회귀신경망에 더 정확한 결과를 나타내는 것이 검증되었다. 해당 연구는 이를 통해 적조 발생을 미리 예측하여 적조에 대한 피해를 최소화하는 데 목적을 두었다.[12]

정지선 외(2015)는 인터넷상에 나타나는 특정 기업과 관련된 뉴스 데이터 수집을 통해 감성분석을 실시하여 주가 등락에 대한 예측을 시도하였으며, 평균적으로 약 56%의 예측률을 보이는 것으로 분석되었다. 이로써 개별 기업에 대한 뉴스 정보의 분석 결과가 해당 기업 주가 변동에 영향을 미치는 요인임이 밝혀졌으며, 향후 적절한 데이터 분석을 통해 주가 변동 예측에 유용하게 활용될 것으로 기대되었다.[13]

홍진주 외(2016)는 인터넷 게시판 악성 댓글 문제를 해결하기 위해 변형

11) 윤원중 외(2019), pp. 238~240.
12) 박선 외(2011), pp. 39~45.
13) 정지선·김동성·김종우(2015), pp. 37~51.

된 악성 어휘를 인식하여 한계점을 개선하는 악성 댓글 탐지 기법을 제안하였다. SVM을 사용해 학습 데이터 500개를 분석하여 악성 댓글을 탐지할 수 있는 식을 도출하였으며, 평가 데이터 100개를 활용하여 악성 댓글 탐지 기능에 대한 성능을 평가하였으며, 87.8%의 높은 정확도를 나타내었다.[14]

오세경 외(2017)는 기업의 부도를 예측하는 과정에서 계량화한 뉴스의 텍스트 데이터를 새로운 정보의 원천으로 활용 가능한 방법을 제시했다.[15] 기존에 존재하던 정보 원천과 더불어 텍스트 정보를 내재하고 있는 인공지능 기반의 예측 방법론을 제시하여 기존 방법론과 인공지능의 예측력을 비교·분석하였다. 본 연구에서는 첫 번째, 기존 연구를 기반으로 뉴스의 텍스트 데이터와 같은 새로운 정보 적용에 따라 부도 예측 성능을 높일 수 있는지, 두 번째, 인공지능 기법 도입과 같은 신규 방법론이 기존 방법론 대비 예측력이 향상되었는지 등을 중점적으로 연구하였다.

〈표 2-1〉 기업 부도 예측 방법론 요약

예측 모형		
분류	방법론	특징
이진 분류 방법	로지스틱 회귀분석 (Logit)	전통적(대표적)인 이진 분류 모형
	의사결정 나무 (Decision Tree)	대표적인 데이터마이닝 기반 이진 분류 방법론
생존분석	콕스 비례-위험모형 (Cox-PH Hazard)	공변량의 특성에 따른 생존 기간 예측 모형
인공지능 (머신러닝, 딥러닝)	랜덤 포레스트(RF: Random-Forest)	랜덤 포레스트 여러 개의 의사결정 나무(Decision Tree)들을 임의적으로 반복 학습하여 추정하는 앙상블 기법을 활용한 예측 방법론
	SVM	데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모형을 만들어 예측하는 방법론

14) 홍진주 외(2016), pp. 260~267.
15) 오세경·최정원·장재원(2017), pp. 1~4.

예측 모형		
분류	방법론	특징
	심층 신경망(DNN: Deep Neural Network)	인공신경망의 숨겨진 층(Hidden Layer) 층을 겹겹이(Deep) 설계한 방법론
	순환 신경망(RNN: Recurrent Neural Network)	DNN의 숨겨진 층(Hidden Layer) 설계 시 변수 간의 시간 순서(Sequence)를 고려하여 설계하여 학습 과정에 활용한 딥러닝 방법론
시장 정보	KMV 모형	옵션 가격 결정 모형을 기반으로 주가 변동에 따른 부도 확률을 산출하는 방법론

자료: 오세경·최정원·장재원(2017), pp. ix-x

연구 결과, 인공지능을 이용한 다른 방법론들이 전반적으로 기존의 전통적인 예측 방법보다 예측력이 우수한 것으로 나타났다. 또한 추가적으로 뉴스 텍스트를 정보의 원천으로 추정한 예측 모형과 시장 정보를 기반으로 하는 예측 모형인 KMV 모형은 상호 유사한 결론을 도출할 수 있는 것으로 분석되었다.[16]

Piao, 변정용(2017)는 국민 건강정보 데이터를 활용하여 SVM을 이용한 진료과목 예측에 관한 연구를 진행하였다. 개인 맞춤형 의료에 대한 요구가 높아짐에 따라 여러 의료 정보가 기반이 되는 질병 진단 관련 연구가 많이 진행되어 왔다. 그러나 기존의 연구들은 특정 질환과의 관련 데이터를 활용하여 특정 질환 예측만을 위한 것으로, 진료과목 예측을 시도한 연구는 없다는 한계가 있었다. 일반 건강검진 결과의 데이터를 이용하여 진료과목을 예측한 결과는 평균 약 80% 이상의 정확도를 나타내었으며, SVM은 다른 예측 알고리즘보다 뛰어난 결과를 보여 주었다.[17]

Yoosin Kim, Mingon Kang, and Seung Ryul Jeong(2018)은 한국

16) 오세경·최정원·장재원(2017), pp. 73~74.
17) Minghao Piao·변정용(2017), pp. 303~308.

어에 대한 자연어 처리 알고리즘과 감성분석 어휘집을 개발하여 온라인 커뮤니케이션의 영화 평가에 대한 감성과 흥행 간 관계를 분석하였다. 영화 147편에 대한 23만 3,631개의 리뷰를 R 프로그램의 XML 크롤링 패키지(crawling package)를 통해 분석한 결과, 부정어 706개, 긍정어 617개 등 약 81.69%의 정확도를 보여 주었다.[18)]

김은미(2021)는 비트코인 수익률의 변동성 예측을 위해 딥러닝 기법과 뉴스 감성 정보를 적용하였다. SVM, 인공신경망, LSTM, 로짓을 적용한 단일 예측 모형을 구축하였고, 변동성 예측의 성능 향상을 위해 통합 모형을 제안하였다. 뉴스 감성 정보를 반영한 예측 모형과 과거의 비트코인 가격 정보를 기반으로 구축한 예측 모형의 성능 비교 결과, 뉴스의 감성 정보를 반영한 예측 모형의 성과가 56.7%로 나타났다. 이러한 연구 결과를 통해 무분별한 투자 리스크를 예방하고 현명한 투자로 이어질 수 있도록 유익한 정보를 제공하고자 하였다.[19)]

이 외 감성분석과 관련된 해운 분야 연구로 박성화·권장한·김태일(2018)은 뉴스 빅데이터 분석 방법을 사용함으로써 해운산업 신용 위험 측정의 어려움을 보완하여 우리나라 해운산업의 신용 위험을 측정하고자 하였다. 분석 결과, 해운뉴스지수는 우리나라의 경기 수축기 외에도 주요 해운선사 위기에도 큰 폭으로 상승하는 경기 역행적인 패턴을 보였다. 특히 매출지수, 업황과 통계적으로 유의미한 상관관계를 나타내었다. 뉴스 빅데이터 분석을 활용한 해운뉴스지수가 해운시장의 리스크를 감지하고 예측력을 높이는데 유익한 정보를 내재하고 있음을 확인할 수 있었다.[20)]

최건우·윤희성·김동원(2019)은 로이즈리스트의 기사를 수집하여 감성분석을 통해 컨테이너 운임시장과 비교 분석 하였다. NRC 감성사전에서는

18) Yoosin Kim, Mingon Kang·Seung and Ryul Jeong(2018), pp. 4090-4102.
19) 김은미(2021), pp. 19~32.
20) 박성화·권장한·김태일(2018), pp. 281~298.

60% 이상, Afinn은 약 51%의 정확도를 기록했다. 해당 연구에서는 시장 심리에 대한 과학적인 접근이 해운회사의 성과 향상에 긍정적인 영향을 미칠 것으로 예측하였다. 이 같은 방법론을 통해 해운산업의 심리(sentiment)를 측정하여 불확실성을 최소화할 수 있을 것으로 판단하였다.[21]

〈표 2-2〉 감성분석에 대한 주요 선행연구

선행연구명	연구 목적/방법론	주요 연구 내용
• 윤원중 외(2019) • 금융위기 조기경보시스템에 관한 연구	• (목적) 금융위기 발생 직전 조기 경보시스템 작동을 위한 SVM 활용 • (방법론)신호처리기법, SVM	• SVM을 통해 1994~2019년까지의 코스피(KOSPI)를 사용하여 주식시장이 영향을 받았던 금융위기 구간 검출
• 박선 외(2011) • 신경망과 SVM을 이용한 적조 발생 예측	• (목적) SVM 기법을 적용하여 적도 발생 예측 정확도 측정 • (방법론) 역전파 신경망, 회귀신경, SVM	• 역전파 신경망과 회귀신경, SVM에 10일간의 수온, 기온, 강수량을 적용하여 미래에 발생할 적조 예측
• 정지선 외(2015) • 온라인 언급이 기업 성과에 미치는 영향 분석: 뉴스 감성분석을 통한 기업별 주가 예측	• (목적) 감성분석을 통한 주가 등락 예측 • (방법론) 뉴스 데이터, 감성분석	• 인터넷상 특정 기업의 뉴스 정보 수집 및 분석을 통해 주가 변동 예측 활용도 검토
• 홍진주 외(2016) • 감성분석과 SVM을 이용한 인터넷 악성 댓글 탐지 기법	• (목적) 인터넷 악성 댓글 탐지 위한 SVM과 감성분석 활용 • (방법론) 감성분석, SVM	• SVM을 사용하여 평가 데이터 100개를 이용하여 인터넷 게시판 악성 댓글 탐지에 대한 성능을 평가
• 오세경 외(2017) • 빅데이터를 이용한 딥러닝 기반의 기업 부도 예측 연구	• (목적) 비정형 데이터인 뉴스 데이터를 계량화하여 기업 부도 예측력 향상 • (방법론) 로짓분석(Logit), 의사결정나무모형(Decision Tree), 랜덤 포레스(Random-Forest), 서포트벡터머신(SVM), 심층신경망(DNN), 순환신경망(RNN), KMV	• 문자 정보를 포함한 인공지능 기반 예측 방법을 제시하고 기존 방법론과 성능 비교 분석

21) 최건우·윤희성·김동원(2019), pp. 274~286.

선행연구명	연구 목적/방법론	주요 연구 내용
• Piao, Minghao 외(2017) • SVM을 이용한 건강검진 정보 기반 진료과목 예측	• (목적) SVM을 활용하여 개인 맞춤형 의료에 대한 요구에 따른 진료과목 예측 • (방법론) 데이터마이닝, SVM	• SVM 및 진료과목 예측 모델링을 통해 국민건강정보 데이터를 이용하여 진료과목 예측
• Yoosin Kim 외(2018) • Text Mining and Sentiment Analysis for Predicting Box Office Success	• (목적) 리뷰를 기반으로 한 감성분석을 통한 영화 평가에 대한 감성과 흥행 분석 • (방법론) 감성분석	• 한국어에 대한 자연어 처리 알고리즘과 감성 분석 어휘집을 개발하여 온라인 커뮤니케이션의 영화 평가에 대한 감성과 흥행 간 상호관계 연관성 검증
• 김은미(2021) • 감성분석을 이용한 뉴스 정보와 딥러닝 기반의 암호화폐 수익률 변동 예측을 위한 통합 모형	• (목적) 뉴스의 감성 정보와 딥러닝을 적용하여 비트코인 수익률을 예측함으로써 무분별한 투자 예방 • (방법론) 로짓, 인공신경망, LSTM, SVM	• 가격 정보 기반 예측 모형과 뉴스 감성 정보를 활용한 예측모형 성능 비교
• 박성화 외(2018) • 뉴스 빅데이터를 이용한 해운 뉴스지수 개발	• (목적) 뉴스 빅데이터를 이용하여 해운산업 신용 위험 측정의 어려움을 보완하여 우리나라 해운산업 신용 위험 측정 • (방법론) 텍스트마이닝, 예측 회귀분석	• 뉴스 빅데이터 분석을 활용한 해운뉴스지수를 통해 해운시장 위험도 예측
• 최건우 외 (2019) • 감성분석을 활용한 컨테이너 시장 예측	• (목적) 감성분석을 통한 컨테이너 운임시장 분석을 통해 해운산업 심리를 측정하여 산업 불확실성 최소화 • (방법론) 감성분석	• 로이즈리스트 기사를 기반으로 한 감성분석을 통해 컨테이너 운임시장 비교 분석

이 외에도 최근 감성분석 관련 연구들은 텍스트의 감성을 단순한 분류로 국한하지 않고 세밀한 단위들로 간주하여 분석하는 경향을 나타내기도 한다.[22] Roy et al.(2018)은 엔론(Enron)의 이메일 코퍼스(Email Corpus)

22) Bollen, Mao and Zeng(2011), pp. 1-8.

를 테스트 사례로 활용하고 데이터마이닝 기술을 적용하여 인적요인 패턴과 추세를 파악하였다. 해당 연구에서는 감성분석을 사용하여 제안된 인적요인 온톨로지(ontology)에 대한 여러 구성 요소들의 평가가 가능한 잠재적인 방법을 소개하였다.[23]

또한 감성분석은 특히 금융시장을 중심으로 주식지수, 시황 예측이나 전망 등에 다양하게 활용되고 있다. 주식시장 예측에 관한 기존 연구들은 대부분 주가의 변동성을 이용하여 가격 패턴을 발굴하는 방식으로, 대부분 기술적 분석(technical analysis)지표를 활용하여 주가지수를 예측하는 연구가 진행되어 왔다.[24]

하지만 주식시장은 기업 내외부 요인의 상호관계가 주가 형성에 영향을 미치는 복잡하고 특수한 가격 결정 메커니즘(mechanism)으로 주식 가격의 변동 요인을 명료하게 설명하기 어려운 경우가 빈번하게 발생한다.[25] 주가 변동의 외부 요인은 현실에서 발생하는 다양한 현상에 대한 설명, 각종 정보들을 내재하고 있는 뉴스 정보 등이 중요한 원인으로 작용할 수 있다.[26] 이처럼 주가와 뉴스는 상화 밀접한 관계를 형성하고 있으며, 시장 참여자들은 뉴스 정보를 통해 주식시장의 변동성을 예측할 수 있게 되는 것이다.[27]

Kalyani, Bharathi, Jyothi(2016)는 뉴스 감정 분석을 활용한 주식 동향 예측을 위해 머신러닝 SVM을 통한 감정 분류를 활용하여 주식 추세를 약 80%의 높은 정확도로 예측하였다.[28]

다오두안홍·안현철(2014)은 인공지능 기법을 활용하여 주식시장의 등락

23) Roy et al.(2018), pp. 1~22.
24) 김유일·신은경·홍태호(2004), pp. 221~243.
25) Chen et al(1986), pp. 383-403.
26) 홍태호·김은미·차은정(2017), pp. 23~36.
27) 송치영(2002), pp.1~34.
28) Joshi, Bharathi and Jyothi(2016), pp. 67~77.

을 예측하고 매매 거래를 수행하는 트레이딩 시스템의 시장 예측 알고리즘으로 활용될 수 있는 새로운 주식시장 등락 예측 모형을 제시하였다. 2004년부터 2013년까지의 10년치 국내 주식시장 데이터를 기반으로 한 코스피(KOSPI) 200 지수의 등락 예측에 제안 모형을 적용하였으며, 비교 모형으로 로지스틱 회귀모형, 다중판별분석, 의사결정나무, 인공신경망, SVM, 퍼지SVM 등도 같이 적용시켜 성과를 정밀하게 검증하였다. 그 결과, 제안 모형이 예측 정확도는 물론 투자수익률(Return on Investment) 측면에서도 다른 모든 비교 모형들에 비해 월등히 우수한 성능을 보임을 확인하였다.[29)]

Yancong Xie and Hongxun Jiang(2017) 또한 중국의 온라인 금융 뉴스를 기반으로 비정형 데이터 수집과 텍스트마이닝을 통한 전처리 및 SVM을 활용한 주식시장 예측 모델을 개발하여 주식시장을 예측하고자 했다. 연구 결과, 뉴스가 증시 동향과 주가에 상당한 영향을 미치며, 정확도가 높은 증시 예측 모델이 될 가능성을 확인하였다.[30)]

Ming-Feng Tsai and Chuan-Ju Wang(2017)은 기업 간 재무 위험을 분석하기 위해 재무 보고서에 대한 텍스트 정보를 사용하여 금융 감정 어휘와 재무 위험 사이의 상관관계를 연구하였다. 금융 감정 어휘와 금융 리스크의 관계 분석을 위해 금융 감정 어휘집을 활용하였으며, 위험 예측과 금융 감정 어휘가 높은 상관관계가 있는 것으로 나타나 상호 간 관련 중요성이 도출되었다.[31)]

허준영·양진용(2015)은 회사 재무 정보를 기반으로 SVM을 이용하여 주식 가격의 예측력을 검증하였으며, 이를 통해 회사의 내재 가치를 나타내는 재무 정보가 주식 가격 예측에 얼마나 효과적인지를 평가하였다. 다른

29) 다오두안홍·안현철(2014), pp. 43~58.
30) Yancong Xie and Hongxun Jiang(2017), pp. 500-510.
31) Ming-FengTsai and Chuan-Ju Wang(2017), pp. 243-250.

기법과의 비교를 위해 전문가 점수와 기계학습 방법인 인공신경망, 결정트리, 적응형 부스팅을 통한 예측 결과와 비교하였으며, 비교 결과 SVM의 성능이 57.5%의 예측 정확도를 나타내며 실행 시간이나 예측력 면에서 모두 우수한 것이 입증되었다.[32]

이우식(2017)은 기술적 분석 지표와 딥러닝 분석을 결합하여 코스피 지수 예측 모형을 개발하였으며, 제시한 모형들 간의 예측력을 비교·분석하였다. 분석 결과 정확도 70% 이상을 나타내었으며, 기술적 분석 지표에 딥러닝 알고리즘을 결합한 모형이 주가지수 방향성 예측에 응용될 수 있을 것으로 파악되었다.[33]

김재봉·김형중(2017)은 기존 주식시장의 명사 또는 뉴스 분석을 통한 감성사전의 한계점을 보완, 개선하고 주식시장이라는 도메인에 특화된 감성사전 구축을 목적으로 뉴스 텍스트 데이터를 분석하여 주가지수의 방향성을 예측하고자 했다. 감성을 표현하는 다양한 글과 의견을 반영하기 위하여 주식 전문 커뮤니티의 게시글을 지정하였고, 형용사를 포함한 말뭉치별로 극성값을 부여하는 방식으로 사전을 구축하여 유의성이 확인되었다.[34]

홍태호·김은미·차은정(2017)은 주식시장의 추세를 확인하기 위하여 기술적 분석을 활용하고 시장의 외부 요소를 반영하기 위하여 뉴스의 감성분석을 활용하였다. 또한 기술적 지표 중 시계열분석 지표인 상대강도지수(RSI: Relative Strength Index)를 활용하여 주식시장 예측에 시계열분석 방법을 반영하고 뉴스를 감성분석한 결과를 결합하여 데이터마이닝 기법인 로짓과 SVM을 적용하였다. 분석 결과, S&P 500을 적용한 예측성과는 감성분석을 결합한 SS_LOGITS이 62.77%의 예측률로, 기술적 분석만 실행했을 때보다 약 5% 향상된 효과를 보여 주었다. 다우존스 지수를 적용

32) 허준영·양진용(2015), pp. 167~172.
33) 이우식(2017), pp. 287~295.
34) 김재봉·김형중(2017), pp. 585~592.

한 예측 성과 또한 SS_SVMD가 약 62.83%의 예측률을 나타내, 기술적 분석만 실행했을 때보다 감성분석 결과가 결합될 때 향상된 예측 성과를 도출할 수 있었다. 해당 연구에서는 다양한 현상을 반영하는 뉴스 정보를 감성분석을 활용한 예측 모형에 적용함으로써 예측하기 어려운 주식시장의 변동성을 일부나마 예측할 수 있도록 하여 투자자의 의사결정에 더욱 유용한 정보를 제공하고자 하였다.[35]

〈표 2-3〉 감성분석을 활용한 주식시장 관련 선행연구

선행연구명	연구 목적/방법론	주요 연구 내용
• Joshi Kalyani 외(2016) • Stock trend prediction using news sentiment analysis	• (목적) 뉴스 감정 분석을 활용한 주식 동향 예측 • (방법론) 감성분석, SVM	• 뉴스 기반 데이터의 감정 분석을 이용하여 주식 동향을 예측하기 위해 머신러닝 SVM을 통한 감정 분류
• 다오두안홍 외(2014) • 주식시장 예측을 위한 π-퍼지 논리와 SVM의 최적 결합	• (목적) 인공지능 기법을 활용한 주식시장 등락 예측 • (방법론) 로지스틱회귀모형, 의사결정나무, 다중판별분석, SVM, 퍼지SVM	• 인공지능 기법을 활용한 주식시장 등락 예측 모형 제시
• Yancong Xie 외(2017) • Stock Market Forecasting Based on Text Mining Technology	• (목적) 중국 온라인 금융 뉴스 데이터를 활용한 주식시장 예측 • (방법론) 텍스트마이닝, SVM	• 온라인 금융 뉴스를 기반으로 비정형 데이터 수집과 텍스트마이닝을 통한 전처리, SVM을 활용하여 주식시장 예측 모델 개발
• Ming-Feng Tsai 외(2017) • On the risk prediction and analysis of soft information in finance reports	• (목적) 재무 보고서를 활용한 기업 간 재무 위험 분석 • (방법론) 감성분석, SVM, bag-of-words	• 기업 간 재무 위험을 분석하기 위해 재무 보고서에 대한 텍스트 정보를 사용하여 금융 감정 어휘와 재무 위험 사이의 상관관계를 연구

35) 홍태호·김은미·차은정(2017), pp. 23~36.

선행연구명	연구 목적/방법론	주요 연구 내용
• 허준영 외(2015) • SVM 기반의 재무 정보를 이용한 주가 예측	• (목적) SVM 방법을 활용한 주식 가격 예측력 검증 • (방법론) SVM, 인공신경망, 의사결정나무, 적응형 부스팅	• 회사 재무 정보와 SVM 방법을 활용하여 주식 가격 예측력 검증
• 이우식(2017) • 기술적 분석 지표와 딥러닝 분석을 이용한 코스피 지수 방향성 예측	• (목적) 인공지능을 포함한 기계학습을 통한 한국 코스피지수의 예측력 분석 • (방법론) 의시결정나무,감성분석, SVM, 딥러닝분석	• 딥러닝(deep learning) 모형과 기술적 분석 (technical analysis) 지표를 결합하여 코스피 지수를 예측 모형 개발, 제시한 모형들의 예측력을 비교· 분석
• 김재봉 외(2017) • 주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축 방안	• (목적) 감성분석을 통한 주가지수 방향성 예측 • (방법론) 감성분석	• 증권 전문 사이트의 게시글을 대상으로 주식시장 맞춤형 감성사전을 구축하여 투자자 감성분석 실시
• 홍태호 외(2017) • 뉴스 감성분석과 SVM을 이용한 다우존스 지수와 S&P 500 지수 예측	• (목적) 주식시장 추세 확인을 위한 뉴스 감성분석 활용 • (방법론) 시계열분석, 감성분석	• 기술적 분석과 시장의 외부요소를 반영하기 위한 뉴스의 감성분석을 활용하여 주식시장 변동성 예측

전술한 바와 같이 현재 감성분석은 금융시장을 중심으로 주식지수, 시황 예측이나 전망 등에 활발하게 활용되고 있으나 아직까지 해운 분야에 적용된 사례는 미흡한 단계라 할 수 있다. 현재 해운 분야에서는 인공신경망을 결합하여 물동량이나 지수를 예측하는 등 인공지능 관련 연구를 지속적으로 진행하고 있다.

임상섭·윤희성(2018a)은 케이프선 운임을 대상으로 단계적 회귀모형과 머신러닝 기반의 랜덤 포레스트 모델을 이용하여 중요 영향 요인을 분석하

였다, 연구 결과, 랜덤 포레스트가 모든 성능 지표를 기준으로 우수한 것으로 드러났으며. 이는 케이프선 운임이 랜덤 포레스트 모델이 제시한 공급 요인들에 민감하게 반응하기 때문인 것으로 해석되었다.[36]

임상섭·윤희성(2018b)는 단기 예측의 중요성이 비교적 더 강조되는 탱커시장의 시황 예측에 기계학습 모델(ANN, SVM)과 계량 모델(ARIMA)을 결합한 하이브리드 모델을 적용하였다. 분석 결과, ARIMA-SVM 모델이 다른 모델들에 비해 성능이 우수한 것으로 나타났다. 해당 연구는 발틱원유유조선운임지수(BDTI)와 발틱석유제품선 운임지수(BCTI)를 대상으로 하여 전체 탱커시장을 포괄할 뿐만 아니라 탱커시장에서 비중이 높은 스팟 거래 기간과 부합하는 결과를 제시하였다.[37]

Mostafa(2004)는 월별 순톤수 자료를 이용하여 전통적 ARIMA 모델과 인공신경망 모형을 이용하여 수에즈 운하 이용률을 예측하였다. 이를 통해 인공신경망 모형의 RMSE값이 ARIMA 모델보다 낮은 것으로 분석되면서 모형의 우수성이 검증되었다.[38]

Guan et al.(2016)는 2011년 1월부터 2014년 11월까지 주간 BSI 데이터를 사용하여 SVM을 기반으로 한 하이브리드 다단계 예측 모델이 BSI 예측에서 높은 정확도를 가지고 있다는 것을 보여 주었다.[39]

Tsai and Huang(2017)은 다층 신경망(MLP)을 활용하여 아시아 주요 10개 항만을 대상으로 컨테이너 물동량 예측 연구를 진행하였다. 환율, 국내총생산(GDP), 경제성장률, 산업생산지수, 수출입 물동량 등을 활용변수로 사용하였으며 그 결과, 홍콩항에서의 컨테이너 수출입 물동량 예측치 오차가 가장 적은 것으로 나타나 향후 활용 가능성이 높을 것으로 파악되

36) 임상섭·윤희성(2018a), pp. 539~545.
37) 임상섭·윤희성(2018b), pp. 79~98.
38) Mostafa(2004), pp.139-156.
39) Guan et al.(2016), pp. 219-232.

었다.[40]

Wu et al.(2019)은 인공지능 알고리즘 KNN(K-nearest neighbors)과 SVM을 기반으로 한 예측 모델을 통해 네 개 노선의 BPI(Baltic Panamax Index)와 선사별 의사결정을 분석하였다. 분석 결과, 제안된 예측 모델이 효과적으로 선사 수익을 증가시킬 수 있으며, 효율적으로 선단 관리의 의사 결정을 하는 데 도움이 될 것으로 확인되었다.[41]

Kamal et al.(2019)은 BDI 예측을 위해 벙커유 가격, 항로별 용선료, BCI(Baltic Capesize Index), BPI(Baltic Panamax Index), BSI(Baltic Supramax Index) 등의 주별 데이터를 활용하여 DNN과 LSTM 연구를 진행하였다. DNN의 RMSE가 LSTM보다 낮은 것으로 나타나 모델의 우수성이 증명되었으며, 앙상블 방법과 SVM 기법을 활용한 다양한 모델 개발의 필요성이 언급되었다.[42]

김동균·최정석(2021)은 기계학습 모델인 간소화된 장단기 메모리 구조의 GRU(Gated Recurrent Units)와 장단기 메모리를 활용하여 2003년 3월부터 2020년 5월까지의 중국 컨테이너 운임지수(CCFI) 데이터를 학습에 사용하여 컨테이너선의 해상 운임 예측 모델을 제안하였다. 컨테이너 운임의 시계열 데이터를 이용하여 운임 예측 가능 모델을 제시했으며, 전통적인 기존의 시계열 예측 분석 방법인 ARIMA 모델과의 비교·분석을 통해 제시한 운임 예측 모델의 우수한 성능을 증명하였다. 이를 통해서 딥러닝 기법을 이용하여 생성된 모델이 컨테이너선의 운임 예측 신뢰도를 향상시킬 수 있을 것으로 판단되었다.[43]

40) Tsai and Huang(2017), pp. 5001-5010.
41) Wu et al.(2019), pp. 136604-136615.
42) Kamal et al.(2019), pp. 427-434.
43) 김동균·최정석(2021), pp. 574~583.

〈표 2-4〉 인공지능 기법을 적용한 해운시장 주요 선행연구

선행연구명	연구 목적/방법론	주요 연구 내용
• 임상섭 외(2018) • 케이프선 시장 운임의 결정 요인 및 운임 예측 모형 분석	• (목적) 인공지능을 이용한 중요 요인 분석을 통한 케이프선 시장 운임 예측 • (방법론) 단계적 회귀모형, 랜덤 포레스트(Random Forest) 모델	• 케이프선 운임을 대상으로 단계적 회귀모형과 머신러닝 기반의 랜덤 포레스트 모델을 이용하여 중요 영향 요인 분석 및 운임 예측 모형 분석
• 임상섭 외(2018) • ARIMA-SVM 하이브리드 모델을 활용한 탱커 운임지수 예측	• (목적) 탱커 운임지수 예측을 위해 인공지능 하이브리드 모델을 활용 • (방법론) ANN, ARIMA, SVM	• 탱커시장의 시황 예측에 기계학습 모델(ANN과 SVM)과 계량 모델(ARIMA)을 결합한 하이브리드 모델을 적용하여 탱커시장 운임 예측
• Mostafa, M.(2004) • Forecasting the Suez Canal traffic: a neural network, analysis	• (목적) 수에즈 운하 이용률 예측을 위한 인공신경망 모형 활용 • (방법론) ARIMA, RSME	• 월별 순톤수 자료를 활용하여 전통적 ARIMA 모델과 인공신경망 모형 예측 성능 비교
• Author Guan 외(2016) • Multi-Step hybrid prediction model of Baltic Supermax Index based on Support Vector Machine	• (목적) SVM을 기반으로 한 예측 모델을 통해 BSI 예측 • (방법론) SVM	• 주간 BSI 데이터를 사용하여 SVM을 기반으로 한 하이브리드 다단계 예측 모델로 BSI 예측 정확도 검증
• 김동균 외 2명(2021) • 딥러닝 기법을 활용한 컨테이너선 운임 예측 모델	• (목적) 딥러닝과 ARIMA 모델과의 비교 분석을 통한 컨테이너 해상 운임 예측 • (방법론) 딥러닝, ARIMA	• 시계열 데이터 및 ARIMA 모델을 활용한 컨테이너선 해상운임 예측 모형 제안
• Tsai, Huang 외(2017) • Using artificial neural networks to predict container flows between the major ports of Asia.	• (목적) 항만 컨테이너 물동량 예측을 위한 인공지능 기법 활용 • (방법론) 다층신경망 (MLP)	• 환율, GDP, 경제성장률, 산업생산지수, 수출입 물동량 등을 활용변수로 사용하고 MLP를 활용한 아시아 주요 항만 컨테이너 물동량 예측

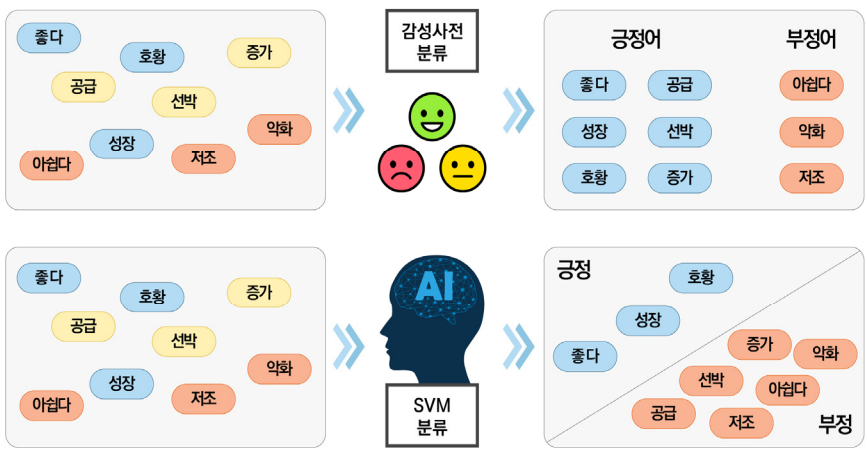
선행연구명	연구 목적/방법론	주요 연구 내용
• Wu, Lan, Yu 외(2019) • Prediction for the Region Disposition of Panama Dry Bulk Fleet Management	• (목적) 효율적인 선사 경영을 위한 인공지능 알고리즘과 SVM 예측 모델 개발 • (방법론) KNN (K-nearest neighbors), SVM	• 인공지능 알고리즘 KNN과 SVM을 기반으로 한 예측 모델을 통해 4개 노선의 BPI(Baltic Panamax Index)와 선사별 의사결정 분석
• Kamal 외(2019) • Forecasting high-dimensional multivariate regression of Baltic Dry Index(BDI) using Deep Neural Networks(DNN)	• (목적) DNN, LSTM을 적용한 BDI 예측 • (방법론) DNN, LSTM, RMSE, SVM	• BDI 예측을 위해 벙커유 가격, 항로별 용선료, BCI (Baltic Capesize Index), BPI(Baltic Panamax Index), BSI(Baltic Supramax Index) 등의 주별 데이터를 활용하여 DNN과 LSTM 분석

본 연구에서는 기사, 보고서 등의 비정형 데이터를 인공지능 지도학습을 통해 분류하는 방법을 실시한다. 기존의 감성분석 분류 방법은 연구자에 의해 사전에 정의된 분류 방법으로 특정 상황의 심리를 정확히 반영하지 못하는 한계가 있다.

예를 들어 컨테이너 기사 중 '혼잡(congestion)'이라는 단어는 감성사전에 부정으로 분류되겠지만, 해운시황에서는 항만의 혼잡으로 인해 컨테이너 공급에 문제가 발생하여 운임이 상승될 가능성이 높다. 이처럼 감성사전 기반의 분류는 기사 전체의 문맥을 파악하지 못해 잘못된 분류를 하여 예측의 오차가 발생할 수 있는 가능성이 존재한다.

하지만 SVM 분류 방법은 최적의 분리경계면(maximum-margin hyperplane)을 기반으로 분류하며, 긍정·부정 분류와 같이 2개의 그룹에 대한 분류 작업 시 의사결정 나무(Decision Tree)나 신경망(Neural Network) 모델보다 분류 정확성이 높은 장점이 있으며, 어떠한 입력 자료(Input Data)를 학습시키는지에 따라 특정 상황의 심리를 반영한 분류가 가능하다는 장점이 있다.

〈그림 2-2〉 감성사전 및 SVM 기반 분류 방법



자료: 저자 작성

제2절 해운시장 비정형 자료 개요

산업에 관한 뉴스는 해당 산업을 파악할 수 있는 가장 신속한 정보 중 하나로서 산업에 연관된 여러 중요한 정보를 다루고 있다. 따라서 뉴스 정보는 해당 산업의 위기 징후를 사전적으로 인지할 수 있는 추가적(additional)이고 대체(alternative) 가능한 정보의 근원으로서 충분한 가치를 지닌 것으로 볼 수 있다.

텍스트마이닝은 본 연구의 데이터로 활용하는 뉴스 정보 외에도 산업 공시 데이터, 신문, 간행물, SNS 등 광범위한 매체까지 확장할 수 있다. 텍스트마이닝은 인터넷 웹, 문서 등의 텍스트 정보를 데이터베이스로 수집한 후 데이터로 정제하는 과정을 포괄하는 개념을 의미한다. 정보 처리 기술과 관련 인프라의 발전에 따라 최근 다양한 분야에 걸쳐 급격하게 활용도가 높아지고 있다.

텍스트를 이용한 정확한 예측 모형을 구축하기 위해서는 대량의 텍스트 정보를 확보해야 하는데, 텍스트 데이터를 수작업(scrap)으로 취합하게 되는 경우에는 분석 표본(sampling) 데이터를 수집하는 범위의 한계가 나타난다. 따라서 텍스트마이닝 과정을 진행할 때는 효율적으로 광범위한 데이터를 확보하고 관리하기 위해서는 직접 웹 데이터베이스(Data Base)에 접근해 데이터를 확보하는 방식인 웹 크롤링(web crawling)을 주로 이용한다. 이는 수작업에 의존하던 기존 텍스트 데이터 수집 방법에 비해 활용 가능한 데이터의 범위를 크게 확장할 수 있고 분석자의 실수 등의 오류(error)를 줄이고 편의성을 확보할 수 있다.[44]

SVM을 활용한 해운시장 분석을 위해서는 현재 발행되고 있는 해운시장

44) 오세경 외(2017), p. 3.

관련 기사, 보고서 등 비정형 자료에 대한 분석이 선행되어야 한다. 기관마다 발행 주기, 형태 등이 상이하므로 이에 대한 조사가 필요하며, 이를 어떻게 구성하여 시장 예측 정확도를 높일 수 있는지에 대한 연구가 진행되어야 한다.

해운시장에 대한 비정형 자료를 발행하는 기업, 언론사, 해운 관련 연구소, 선박중개회사 등 다양한 집단에서 일간지, 산업전문지, 보고서 등을 활용하여 시장에 대한 분석을 하고 있다. 해운시황을 예측하기 위한 텍스트 마이닝 연구를 위해서는 크게 기사를 발행하는 해운 관련 전문지, 해운 관련 연구소, 해운 중개 기관에서 발간하는 리포트 등에서 텍스트 정보를 획득할 수 있다.

1. 기사 및 정기간행물

세계적으로 해운시황에 관한 정보를 정기적으로 발행하는 기관은 영국 로이즈리스트(Lloydslist), 미국 JOC(The Journal Of Commerce), 노르웨이의 트레이드윈즈(Tradewinds) 등이 있다.

1) 로이즈리스트

로이즈리스트(Lloydslist)는 1734년부터 신문 지면 발간을 시작으로 현재까지 지속해서 발행되고 있는 해운산업 정기간행물이다. 영국의 해상무역이 전성기를 맞이했던 17세기, 현재와 같이 통신을 활용한 정보 공유가 원활하지 않아 선주, 화주, 해상보험업자 등 영국의 해운 관련 종사자들은 템스강 인근 '로이즈 커피하우스(Lloyd's Coffee House)'라는 커피숍에 모여 정보를 교환하였다. 당시 커피숍의 주인 에드워드 로이드가 해상무역

과 관련된 다양한 정보들을 정리해 '로이즈 리스트(Lloyd's List)'라고 일컫는 신문 형식의 문서를 만들어 배포한 것이 오늘날 로이즈리스트의 유래가 되었다.

로이즈리스트는 신문 지면을 구독하는 고객이 2% 미만으로 감소하면서 2013년 이후 디지털 포맷의 온라인 서비스로 재편하였다. 또한 로이즈리스트는 컨테이너선, 드라이벌크선, 탱커와 가스선, 기타 등으로 구분하여 관련 기사를 게재하고 있다. 관련 섹터를 중심으로 기사를 매일 업데이트 할 뿐만 아니라 정기적으로 시장 전망에 대한 보고서도 발간하고 있다. 구독자들에게 해운시장 전망, 올바른 의사결정, 위험 회피, 기회 제공을 하는 것을 목표로 한다.

〈표 2-5〉 로이즈리스트 정보 제공 목록

구 분		내 용
시장 보고서	컨테이너	컨테이너 시장 주요 이슈, 전문가 인터뷰, 운임 현황 및 전망
	벌크	벌크시장 시장 주요 이슈, 전문가 인터뷰, 운임 현황 및 전망
	탱커와 가스	탱커 시장 주요 이슈, 전문가 인터뷰, 운임 현황 및 전망
	기타	해운금융, 물류 분야 주요 이슈 및 현황
	정기 보고서	일별 보고서
	시장 전망	해운시장 전망(정기/부정기)
	전문가 인터뷰	해운 전문가 인터뷰(주 단위로 작성)
	주요 이슈	정기 기사 외 특별 기고문, 심층 취재 등

자료: 로이즈리스트(검색일: 2021. 8. 22)

2) PR News

영국에 본사를 둔 PR News Service Ltd는 1997년 설립되었다. PR News는 전 세계를 대상으로 하는 컨테이너 운송 뉴스 제공 업체로 주요 물류회사, 해운선사, 항만 관련 회사 등이 주로 구독하는 디지털 뉴스 매체이다.

〈그림 2-3〉 PR News 정보 제공 서비스 구성



자료: PR News(검색일: 2021. 8. 22)

PR News의 정보 제공 서비스는 크게 뉴스와 편의정보 제공(Comfort) 두 가지로 구분된다. 뉴스에서는 컨테이너산업에 대한 뉴스를 실시간으로 제공하는 것이 주요 특징이며, 전 세계 해운시장 최신 뉴스, 해운 주요국 동향과 전망, 전문가 칼럼 등을 제공하고 있다.

또한 편의정보 제공 부분에서는 해운시장과 밀접한 관계를 맺고 있는 항만에 관한 정보도 제공하고 있는데, 주요 컨테이너 항로 정보, 주간/연간 서비스(Weekly/annual Service capacity), 서비스 빈도(Service frequency), 운항 기간(voyage duration), 운항자와 슬롯 구매자(operators and slot purchasers), 전체 항만 회전율(Full port rotations) 등이다.

〈표 2-6〉 PR News 정보 제공 서비스 목록

구 분	내 용
뉴스 (News)	전 세계 해운시장 최신 뉴스
	미국, 유럽, 중국 등 주요 해운국 주요 이슈 및 동향, 전망
	전문가 칼럼, 특별 기고문 등
편의정보 제공 (Comfort)	주요 컨테이너 항로 정보 제공
	주간 연간 서비스(Weekly/annual Service capacity)
	서비스 빈도(Service frequency), 운항 기간(voyage duration), 운항자와 슬롯 구매자(operators and slot purchasers)
	전체 항만 회전율(Full port rotations)

자료: PR News(검색일: 2021. 10. 23)

3) 기타

트레이드윈즈(Trade Winds)는 해운 뉴스를 제공하는 전문 언론사다. 1990년에 창립되었으며, 약 30년간 운영되고 있다. 본사는 노르웨이 오슬로에 있으며 영국 런던, 중국 상하이, 싱가포르 등 거점 국가들에 지사를 설립하여 전 세계에 해운시장과 관련된 정보를 전달하고 있다.

트레이드윈즈는 컨테이너, 벌크, 탱커, 가스 금융 등 해운산업과 관련된 광범위한 영역을 다룬다. 크루즈선까지 포함하여 선종별 주요 이슈와 운임시장 동향을 전달하고 있으며, 선박보험, 해운 관련 규제, 해양 사고 등과 같이 선박 운항과 관련된 주요 정보도 폭넓게 제공하고 있다.

〈그림 2-4〉 트레이드윈즈 제공 정보 구성



자료: 트레이드윈즈(검색일: 2021. 10. 23)

〈표 2-7〉 트레이드윈즈 정보 제공 목록

구 분	내 용
금융(Finance)	해운기업의 재무 동향, 정부 지원 등
유조선 시장(Tankers)	유류시장, 탱커선 시장 동향 등
가스 시장(Gas)	LNG 선박 건조, 용선 등 관련 조선소 동향
건화물 시장(Dry Cargo)	벌크선 운임, 시장 동향
컨테이너선 시장(Box ships)	컨테이너 운임, 시장 동향
오프쇼어 시장(Offshore)	해양 플랜트, 조선산업 등
여객선 시장(Passenger ships)	크루즈, 카페리 시장 동향
선박 매매시장(ship sales)	선박 발주 정보, 스크러버 장착 등
선박 보험시장(Insurance)	선박 보험, 세미나 등
해상안전(Safety)	해상 사고, 안전 규제, 재난 등
해사법률 및 정책(Legal)	경제 제재, 해운 규제 등

자료: 트레이드윈즈(검색일: 2021. 10. 23)

JOC는 1827년 9월부터 서비스를 시작했으며 해운, 항만 분야 외에도 연관 산업인 철도, 물류 분야까지 다양한 기사를 제공하고 있다. 특히 해운 산업과 관련된 블록체인, 인공지능, 사물인터넷과 같은 신기술에 대한 기사 섹션이 별도로 있으며, 대형 선사에서 소형 선사에 이르기까지 다양한 데이터를 제공하는 것이 특징이다.

〈그림 2-5〉 JOC 제공 정보 구성



자료: JOC(검색일: 2021. 8. 22)

JOC는 또한 스페셜 리포트를 통해 해운·항만 관련 산업 종사자를 위한 노동시장 보고서, 국가별 물류산업 시장 분석 보고서 등을 발행하고 있으며, 온라인을 통해 국가별 무역 및 해운 물류 관련 지식센터를 운영하고 있다. 이 외에도 해운산업 규제, 관련 정책을 분석하고 논의하는 별도의 행사 등을 정기적으로 개최하여 해운 관련 산업 동향과 주요 이슈 등을 정기적으로 제공하고 있다.

〈표 2-8〉 JOC 정보 제공 목록

구 분	내 용
특별 주제(Special Topics)	노동시장, 산업, 규제에 대한 특별 보고서
항만(Ports)	미국 항만, 항만 생산성, 항만 운영 등
운항계획(Sailings)	컨테이너 선박에 대한 주요 운항 계획
해사 전반(Maritime)	컨테이너 정기 서비스, 조선, 운임 등
다목적 선박 시장(BreakBulk)	다목적 선박 동향 등
육상 운송 시장(Trucking)	내륙 운송 시장 정보(미국 중심)
물류시장(Logistics)	항공, 물류기술 등
복합운송시장(Rail&Intermodal)	복합운송, 내륙항만 등
규제 및 정책(Regulation&Policy)	무역 규제, 운송정책 등
기술 시장(Technology)	신기술(4차 산업혁명)

자료: JOC(검색일: 2021. 8. 22)

2. 해운시장 관련 연구기관

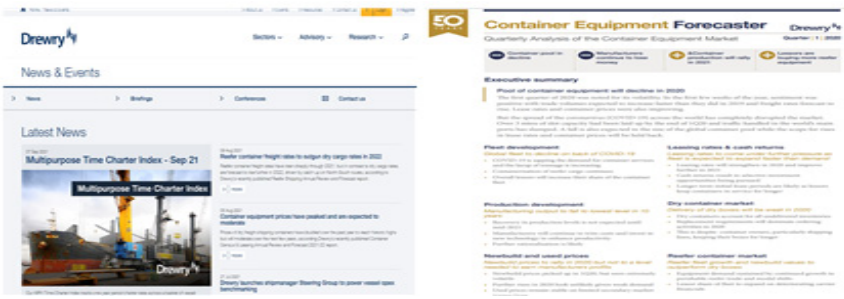
글로벌 해운시장을 분석, 연구하는 대표적인 해운시장 분석 전문기관으로는 영국 해운조사기관 드류리(Drewry)와 클락슨(Clarkson), 영국 해운조사기관 IHS Markit, 프랑스 해운조사기관 알파라이너(Alphaliner) 등이 있다.

1) 드류리

드류리(Drewry)는 영국 런던, 인도 델리, 싱가포르, 중국 상하이 등 총 4개의 지사를 통해 여러 국가에 해운 관련 정보를 제공하고 있다. 드류리는 크게 해운 관련 연구(Maritime Research), 해운 관련 자문(Maritime Advisors), 공급망 관련 자문 및 선박 금융 연구(Supply Chain Advisors and Maritime Financial Research) 등 네 개의 사업부를 통해 정보를

제공하고 있으며, 400명 이상의 전문가를 통해 월평균 282건의 관련 기사를 제공하고 있다. 또한 드류리는 컨테이너, 드라이벌크, 탱커, 항만 및 터미널, LNG/LPG, 선박 운항, 물류 분야에 대한 월·분기·연간 보고서를 발간하고 있다.

〈그림 2-6〉 드류리 발행 정기 보고서



자료: 드류리(검색일: 2021. 8. 22)

2) 알파라이너

알파라이너(Alphaliner)는 컨테이너 정기 서비스, 선대 운항 정보 발주량, 선박 해체량, 해적 정보, 선박자동식별장치(AIS) 등 해운시장의 전반적인 정보를 제공하고 있다. 알파라이너는 총선대 규모, 신조선 발주량, 인도량 등 선박 운영 주요 통계를 전달하는 월간 모니터링 보고서, 선사나 항만 관련 주요 이슈를 다루는 주간 보고서를 정기적으로 발행하고 있다.

월간 보고서는 선박 운영 주요 통계 자료 외에도 선사별, 항로별 선대 투입 규모, 시장 점유율, 화물 처리량 등 세부적인 정보를 제공하고 있다. 주간 보고서는 1주 단위로 발생한 해운 주요 이슈, 선박 사이즈별 용선시장 정보, 정기 서비스 개설, 폐쇄 등에 대한 정보를 전달하고 있다.

〈그림 2-7〉 알파라이너 발행 월간 보고서 및 주간 보고서



자료: 알파라이너(검색일: 2021. 8. 22)

3) 클락슨

클락슨(Clarkson)은 해운 브로커, 금융, 연구 등 해운산업의 전반적인 영역을 다루고 있다. 1852년부터 운영되었으며 세계 22개 국가, 48개 지사를 운영 중이다. 전 선종과 선형에 대한 브로킹을 실시하고 있으며 선사 및 선박에 대한 위험 관리도 한다.

또한 포워딩 등 항만 관련 서비스와 플랜트산업에 대한 시장 지원 등 여러 분야에 걸쳐 다양한 사업을 영위하고 있으며 해운산업의 기초 통계 관련 정보 등을 제공하고 있다. 약 8만 8천 척의 선박에 대한 위치 추적 실시, 약 2만 4천 종류의 관련 통계를 생산하여 제공하고 있다. 해운산업 분야 외에도 600여 개의 세계 조선소에 대한 자료를 해운 정보망(Shipping Intelligence Network)라는 통계 전용 웹사이트를 통해 제공하고 있다.

3. 해운중개업체 리포트

1) 발틱거래소

발틱거래소(Baltic Exchange)는 650개의 회원사로 구성된 해운 운송시장 정보 제공 기업이다. 발틱거래소는 1744년부터 시작되었으며 1916년 비공개 유한회사로 설립되었다. 이후 2016년 11월 싱가포르 기업에 매각되었으며, 현재 본사는 영국 런던에 소재하고 있다. 이 기업은 해운거래 선물시장뿐만 아니라 BDI(Baltic Dry Index)를 통해 세계 운임 추이 정보를 제공한다. BDI 외에도 BCI(Baltic Capesize IndexI), BPI(Baltic Panamax Index), BSI(Baltic Supramax Index), BDTI(Baltic Dirty Tanker Index), BCTI(Baltic Clean Tanker Index)를 제공하며, 2018년부터 세계 컨테이너 화물 운임지수인(FBX)와 LNG 관련 정보도 제공하고 있다.

발틱거래소는 또한 현물 운임시장 정보를 제공하고 있으며, 매일 드라이 벌크 시장에 대한 자료(Baltic Briefing)를 발간한다. 해당 자료에는 운임지수에 대한 내용과 함께 선형별, 지역별 이슈 정보가 수록되어 있다. 이외에 멤버십에 한정하여 이슈 보고서를 제공하고 있으며, 경제 관련 매거진도 발간하고 있다.

〈그림 2-8〉 발틱거래소 정보 제공

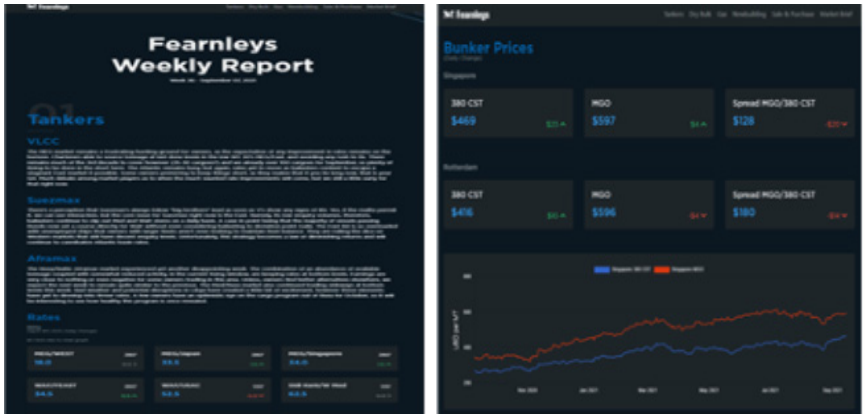


자료: 발틱거래소(검색일: 2021. 8. 22)

2) 펀리스

펀리스(Fearnleys)는 1869년부터 해운산업에 대한 연구를 시작한 기관으로 탱커 분야에 강점이 있다. 매주 주간 서비스를 제공하여 유가 동향 및 관련 선형 자료를 제공한다. 또한 선박 공급 및 항만 체선 등 운임에 영향을 미치는 다양한 정보를 제공한다. 선박 판매 및 구매 업무도 병행하며, 탱커에 강점이 있지만 카페리, LNG 선박에 대한 연구도 진행하고 있다.

〈그림 2-9〉 펀리스 정보 제공 주간 보고서



자료: 펀리스(Fearnleys), (검색일: 2021. 8. 22)

03
감성분석과 서포트벡터머신(SVM)

데이터마이닝은 저장된 방대한 양의 데이터에서 체계적이고 통계적인 규칙이나 패턴을 찾아내는 데이터베이스 파생 기술을 뜻하며, 텍스트마이닝은 고정된 필드에 저장되지 않은 SNS, 기사, 보고서 등의 비정형화된 자료의 데이터를 분석하여 가치 있는 정보를 추출하는 분석 방법이다.[45]

제3장 제1절에서는 이러한 데이터마이닝과 텍스트마이닝의 정의 및 차이점에 대해 자세히 서술하였다. 다음으로, 제3장 제2절에서는 감성분석의 정의와 분석 방법에 대해 설명하였다. 감성분석은 비정형 자료를 활용하여 시장 참여자에 대한 심리를 정량적인 수치로 객관화할 수 있어 해운시장에 대한 전문가들의 전망이나 인식을 측정할 수 있다는 장점이 있다.

마지막으로, 제3장 제3절에서는 SVM에 대한 정의와 특징에 대해 서술하였다. 감성분석은 단어 분류를 특정 연구자에 의해 사전에 정의된 감성사전을 기반으로 분류하는 것으로, 중의적 의미를 띠는 단어 분류에 오류가 발생할 수 있다. SVM은 비정형 데이터와 해운시황을 매칭시킨 학습을 통해 특정 단어가 해운에 긍정 또는 부정의 의미를 나타내는지 직접 분류하여 이러한 오류의 격차를 일부 해소할 수 있다.

45) 국가기록원(2017), p. 359.

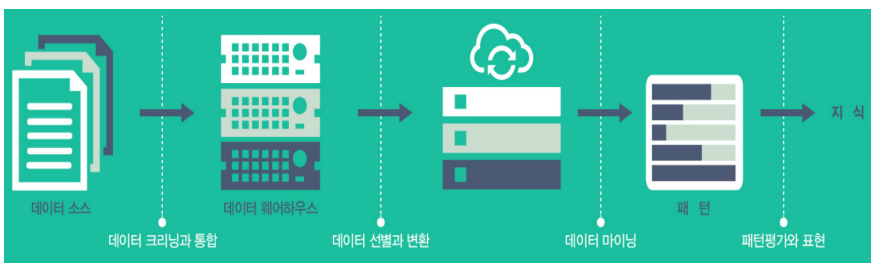
제1절 데이터마이닝과 텍스트마이닝

1. 데이터마이닝

1) 데이터마이닝 정의

데이터마이닝(Data mining)은 대용량의 데이터베이스에서 숨겨져 있는(hidden) 예측 가능하고(predictive) 유용한 정보를 추출하는 것을 말한다. 데이터마이닝 수행 시에는 인공지능과 다양한 수학적·통계적 기법 등을 이용한 패턴 인식기술 등을 활용하여 데이터에 내재되어 있는 유의미한 패턴, 관계 등과 관련된 규칙을 발견하게 된다. 〈그림 3-1〉과 같이 데이터 소스에서 데이터를 클리닝(cleaning)하여 통합 과정을 거친 후, 데이터를 선별하여 변환하고 데이터마이닝 과정을 진행함으로써 패턴을 발굴하고 표현하는 것이다.[46] 또한 기존의 전통적 통계분석과 데이터마이닝은 다음과 같은 특징을 지닌다.

〈그림 3-1〉 데이터마이닝 과정



자료: 국립중앙과학관 빅데이터(검색일: 2021. 9. 12)

46) 국립중앙과학관 빅데이터(검색일: 2021. 9. 12)

〈표 3-1〉 전통적 통계분석과 데이터마이닝

전통적 통계분석	데이터마이닝
대상 집단이 있으며, 모집단의 분포 혹은 모형 등 여러 가지 가정을 전제로 하게 됨. 이 전제조건하에서 분석을 실시, 표본(Sample) 관찰을 통해 모수(Population) 전체를 추론(Inference)하는 과정.	표본조사·실험에서 필연적으로 수반되는 분포라든가 모형에 대한 전제조건이 필요하지 않으며 대용량 자료여야 한다는 전제조건 있음. 모집단의 전체 자료를 이용하여 필요한 정보·지식을 추출하는 과정.

자료: 국립중앙과학관 빅데이터(검색일: 2021. 9. 12)

데이터마이닝이 지정된 대량의 데이터 내에서 체계적이고 자동으로 일정한 패턴이나 통계적 규칙 등을 발굴하는 파생 기술이라면 기존의 전통적인 통계분석은 이미 정립된 모형이나 가설에 의거, 데이터를 검증하거나 요약·보고를 하는 것에 초점이 맞춰져 있는 것이 특징이다.

데이터마이닝은 목표지향(defined), 목표불명(undefined) 등 크게 두 가지로 구분할 수 있다. 몇 개를 특정 대상(target field)으로 구분 혹은 설명 하는 것이 목표지향 데이터마이닝을 의미한다. 이와는 달리 목표불명 데이터마이닝은 데이터에 단순히 존재할 법한 유사성, 패턴 등을 발견하는 것으로, 주로 소규모 그룹에 대한 구분으로 이해할 수 있다.[47] 목표지향과는 달리 특정 대상과 같은 것을 사용하지는 않는다.

2) 데이터마이닝 분석

데이터마이닝을 수행하는 기법은 크게 분류(classification), 추정(estimation), 예측(prediction), 유사 행태 집단화(affinity grouping), 군집화(clustering), 서술·설명 등의 기초분석(profiling) 등 여섯 가지로 분류된다.

47) 김용환(2016)

데이터마이닝의 주요 여섯 가지 기법 중에서 분류, 추정, 예측 등이 목표지향 방법이며, 유사 행태 집단화나 연관규칙 분석, 군집화 등은 목표불명 방법이다. 마지막으로 서술·설명 등의 기초분석은 목표지향, 목표불명 등에 모두 해당하는 것으로 볼 수 있다.

첫 번째, 데이터마이닝의 가장 기본적인 기법인 '분류'는 인종, 성별 등 무리가 생활하면서 흔하게 구분하는 방법이다. 이는 범주형 자료(categorical data)나 이산형 자료(discrete data)에 활용된다. 그 예로 신용 평점자들을 대상으로 고신용자, 중간신용자, 저신용자 등과 같은 분류, 인터넷 웹에 게시할 콘텐츠, 산업코드 지정 등의 업무에서 이산형이나 범주형으로 구분된다.

두 번째는 '추정'으로, 수치형(numerical), 연속형(continuous)으로 결과를 규정한다는 점이 분류와 구분된다. 이러한 추정 기법은 추정값에 순위 등이 적용 가능한 장점이 있으며, 이는 고객관계관리, 즉 CRM(Customer Relationship Management)에서 주로 사용된다.

세 번째는 '예측기법'으로, 미래의 행동, 미래 추정치의 예측에 따라 구분되는 점을 제외한다면 분류나 추정의 기법과 동일하다. 즉, 예측기법은 입력변수에 대한 관계나 타깃변수에 대한 예측을 기반으로 예측모델링(Predictive Modeling)을 하는 것이다. 예로 현재 시점 이후 6개월 동안 이탈하거나 추가 서비스를 신청할 가능성이 있는 고객들의 행동을 예측하는 데 사용될 수 있다.

네 번째는 '유사 행태 집단화'로 유사 집단화로도 불리며, 연관규칙으로 통용된다. 유사 집단화의 간단한 접근 방법은 데이터로부터 규칙을 만드는 것이다. 만약 개 사료와 'KMI 마트'의 제품이 있을 때, 개 사료를 구입하고 개 제품을 구입하는 확률을 P1, 개 제품을 구입하고 개 사료를 구입하는 확률을 P2로 가정하여 그 확률을 통해 연관규칙을 도출하는 것이다. 이러

한 기법을 통해 소매점은 교차판매, 효율적인 매장 진열 방식, 다양한 패키지 상품 개발 등의 전략을 수행할 수 있게 된다.

다섯 번째, '군집화'는 여러 군의 집단을 같은 성질의 소그룹으로 구분하는 것으로, 앞서 언급한 바 있는 '분류'는 이미 정의된 집단으로 구분하므로 군집화와는 다른 체계를 지닌다. 군집화는 우선 개별적으로 유사한 성질의 집단으로 구분이 되며, 이로 인해 구분된 집단의 성질이 결정된다. 기업 고객 홍보 전략 설정 시, 전체 고객을 대상으로 고객 각각을 같은 성질의 고객 그룹으로 분류한 후, 이 중 상위 집단만을 선별하여 프로모션, 마케팅 등을 진행하거나 개별 집단의 특성을 반영한 이벤트를 실시하는 등의 전략을 수립할 수 있다.[48]

마지막인 여섯 번째는 서술 혹은 설명 등을 사용하는 기초분석으로 복잡하게 구성되어 있는 데이터베이스를 간단하게 묘사하고 서술하는 기법이다. 즉, 고객이나 상품 등에 대한 이해도를 증가시켜 원하는 제품과 서비스의 종류나 비율을 분석하고 이를 통해 고객의 욕구와 니즈를 파악하여 상품 개발과 효율적인 정보 전달 방식을 고려하는 것이다. 이때, 의사결정 나무나 연관규칙 등은 고객 관련 프로필 구성에 효율적인 도구가 될 수 있다.

2. 텍스트마이닝

1) 텍스트마이닝 정의

텍스트마이닝(Text mining)은 빅데이터를 활용한 비정형 데이터 분석 기법 중 하나이다. '텍스트 기반의 지식 발견(knowledge discovery from text)'이라고도 불리며, 이는 자연어로 구성되어 있는 비정형화된 문서

48) 김용호(2016)

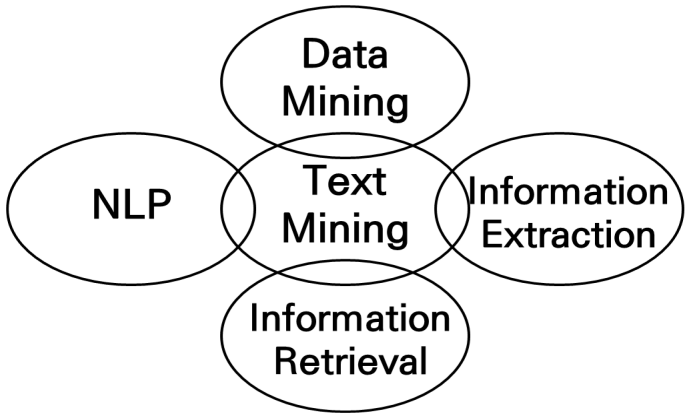
(unstructured data)를 원천으로 내재되어 있는 패턴이나 관계를 파악함으로써, 가치있고 유의미한 지식이나 정보를 발견하 분석 기법이다.[49] 텍스트마이닝은 발굴한 정보를 통한 새로운 지식 확보, 텍스트 분류 등에 활용되고 있다.[50]

텍스트마이닝이 분석하는 대상은 텍스트, 그림, 영상과 같은 비정형 데이터로서 형태가 복잡하고 일정하지 않기 때문에 분석하기 어렵다. 따라서 컴퓨터가 인간의 언어를 인식하여 처리해 주는 자연어 처리(NLP: Natural Language Processing) 기술을 기반으로 한다. 자연어 처리 기술은 비정형 데이터(텍스트)를 정형화(수치)시켜 데이터를 분석 가능한 형태로 만들어 준다[51]. 〈그림 3-2〉와 같이 텍스트마이닝은 데이터마이닝, 정보 검색, 통계학, 기계학습(machine learning), 컴퓨터 언어학(computational linguistics) 등이 결합 된 학문이기 때문에 객관화된 분석이 가능하다.[52]

텍스트마이닝은 다음 그림과 같이 자연어 처리 기술(NLP)과 정보 검색(information retrieval), 정보 추출(information extraction) 방법 그리고 기존의 데이터마이닝과 결합하여 비·반정형 빅데이터로부터 유용한 정보를 추출하고 가공하는 목적으로 사용된다[53]. 방대한 텍스트 데이터에서 유의미한 정보를 추출하고, 해당 정보와 연계된 정보를 파악하는 등 단순히 텍스트가 가지고 있는 정보 이상의 유익한 결과를 얻어 낼 수 있다는 것이 장점이다.[54]

49) Hotho et al.,(2005), pp. 1~37.
50) Dorre et al., (1999), pp. 398~401.
51) 전채남(2016), pp. 57~72.
52) 오하영· 박정식(2017), pp. 20~21.
53) 윤희성 외(2017), pp, 62~63.
54) elec4, 빅데이터 시장동향과 분석기법(검색일: 2021. 8. 23)

〈그림 3-2〉 텍스트마이닝 Areas



자료: Miner et al.(2012), p. 31

2) 텍스트마이닝 분석

텍스트마이닝 과정에서 비정형 데이터의 자연어 처리에는 다양한 방법이 있다. 하지만 기본적인 처리 과정은 비정형 데이터 수집, 데이터 전처리, 정보 추출, 정보 분석의 네 단계로 진행된다.

비정형 데이터의 수집은 연구 수행자가 도출하고자 하는 유용한 정보가 포함되어 있는 대용량의 텍스트 데이터를 대상으로 하며, 연구의 방향에 따라 웹 검색, 기업에서 사용하고 있는 일부 데이터 검색, 개인정보 검색 등 다양한 출처를 기반으로 이루어진다.[55]

데이터에 대한 전처리를 수행할 경우, 수집된 텍스트 데이터의 공백과 구두점을 구분자로 하여 어휘를 분할하는 토큰화(tokenization) 과정과 품사 태깅(POS tagging) 등으로 데이터를 가공한다. 정보 추출 단계는 수학 모델 또는 알고리즘을 활용하여 연구에 필요한 키워드의 추출을 실행하는

55) 정근웅(2018), p. 12.

것이다.

마지막 단계인 정보 분석에서는 추출된 키워드를 연구 목적에 따라 기계학습이나 통계학을 적용하여 주요 키워드를 분류하여 결과를 도출한다.[56] 텍스트마이닝은 다음 〈표 3-2〉와 같이 5단계로 수행이 된다.

〈표 3-2〉 텍스트마이닝의 단계 및 내용

단계	내용
1. 텍스트 전처리 (Text Preprocessing)	• 텍스트 클린 업(Text clean up): 데이터를 8비트로 구조화하기 위해 사용하는 바이너리(binary) 형식으로 변환된 텍스트를 정규화하거나 수식 및 표, 그림을 처리하는 등 원치 않거나 불필요한 정보를 제거함
	• 토큰화(Tokenization): 전 단계에서 확인된 약어에 포함되지 않는 공백이나 문장부호에 텍스트를 분할함
2. 텍스트 변환 (Text Transformation)	• 문서에서 추출된 텍스트의 분포와 관계를 고려하여 단어의 집합을 저장, 관리함 • 벡터 스페이스 모델(Vector space model), 백 오브 워드(Bag of words) 등의 방법을 이용해 문서를 벡터로 공간에 표현하는 단계를 의미함
3. 특징 선택 (Feature Selection)	• 데이터마이닝(Data mining)에서 변수 선택(modify)에 해당하는 단계임 • 정보 획득(Information gain), 카이제곱 통계량(chi-square statistic), 엔트로피 기반 순위(Entropy-based ranking), 문서 빈도(Document frequency), Term strength 등과 같은 방법이 존재 • 전술한 특징 선택 방법들은 수차례의 반복 수행 및 적절한 방법의 활용을 통해 최적 특징을 선택하거, 모델 품질과 분석 결과는 선택된 특징에 종속됨
4. 분석(Analysis) 및 모델링(Modeling)	• 군집화(Clustering) 단어가 가지는 유사성을 평가한 후 비슷한 성격을 가지는 단어들끼리 묶는 분석 방법으로, 확률적 클러스터링 알고리즘(Probabilistic clustering algorithm), K-평균 알고리즘 (K-means clustering algorithm), 계층형 클러스터링 알고리즘 (Hierarchical clustering algorithm) 등의 방법이 존재함

56) Miner(2012), pp. 151-180.

단계	내용
	• 사회연결망 분석(Social network analysis): 텍스트 간의 상관관계를 노드(node)와 링크(link)로 모델링하는 방법으로 텍스트 간 내재된 관계를 표현하고 도출함
	• 감성분석(Sentiment analysis): 언어에 나타난 주관성 요소를 탐지하며 단어가 가지는 긍정 및 부정의 정도를 판별하고 정량화하는 분석 방법으로 지식기반 방법(Knowledge-based method), 통계적 방법(Statistical method) 등의 방법이 존재함
5. 모델링(Modeling)	• 범주화(Classification): 나이브 베이즈 분류기(Naive Bayes Classifier), 최근접 분류자(Nearest Neighbor Classifier), 서포트벡터머신(Support Vector Machines), 의사결정나무 분류(Decision Tree classifiers) 등이 존재함
	• 주제 모델과 질의응답 모델(Topic model and Question answering model): 디리클레 할당(Latent Dirichlet Allocation), 확률적 잠재의미 분석(Probabilistic Latent Semantic Analysis) 등이 존재함

자료: 윤희성 외(2017), pp. 63~64

3. 데이터마이닝과 텍스트마이닝의 차이

데이터마이닝과 텍스트마이닝은 분석대상이 텍스트 데이터와 수치 데이터라는 점에서 차별성을 지니며, 텍스트마이닝은 데이터마이닝과는 달리 이를 체계적으로 구조화시키는 과정이 필요하다. 텍스트마이닝의 구조화 과정에서 가장 일반적으로 활용되고 있는 것이 문서색인이다.[57]

데이터마이닝이 주로 구조적인 데이터를 대상으로 유용하고 잠재적인 패턴을 끌어내는 방식이라면, 텍스트마이닝은 주로 자연어로 구성된 비구조적인 텍스트 안에서 패턴이나 관계를 추출하여 유의미한 정보를 도출하는

57) 조태호(2001), pp. 76~85.

것으로, 주로 텍스트의 자동 분류 작업, 새로운 지식 생성 작업에 활용되고 있다. 현재 우리가 사용하는 대부분의 정보는 구조화가 명확하게 되지 않은 텍스트 형태로 존재하므로 자연어로 된 텍스트 문서의 지능적인 분석은 매우 어려운 작업으로 간주할 수 있다.[58]

텍스트마이닝과는 달리 데이터마이닝은 데이터 간 관계나 패턴을 탐색하고 모형화하여 기업 의사결정에 이용하는 것을 목적으로 적용되며, 대부분 일반적인 데이터베이스와 같은 구조화된 자료를 이용한다. 따라서 데이터마이닝 작업 수행을 위해서는 적용할 데이터가 표준화되고 명확해야 하며, 구조화가 잘 이루어진 후에 적용할 수 있다. 데이터마이닝 기법을 이용한 텍스트 분석을 위해 최근 XML 형식과 같은 웹 문서를 구조화하는 시도가 이루어지고 있으나 아직 텍스트 문서들은 대다수가 비구조적인 형태를 구성하고 있다.[59]

〈표 3-3〉 데이터마이닝과 텍스트마이닝 비교

비교 내용	데이터마이닝	텍스트마이닝
대상 데이터	범주화·수치 데이터	텍스트
데이터 구조	관계형 데이터베이스	비정형 텍스트
목표	미래의 결과 예견 및 예측	적합한 정보를 획득하여 내재된 의미를 정제하고 범주화
방법	기계학습	기계학습 포함 신경망 처리, 온톨로지, 인덱싱, 자연어 처리 등에 적용 가능

자료: 안동진(2020), p.10을 바탕으로 저자 재구성

58) 최윤정·박승수(2002), p. 34.
59) 구미숙 외(2006), pp, 299~308.

제2절 감성분석(Sentiment Analysis)

1. 감성분석 정의

감성분석(Sentiment Analysis)이란 텍스트에 내재되어 있는 의견이나 평가, 태도, 감성 등의 주관적 정보들을 컴퓨터를 통해 분석하는 과정을 의미한다. 자연어 데이터의 감성분석은 예전부터 연구되어 왔으나 언어 특유의 속성인 모호성으로 인해 쉽지 않았다.

감성분석이 사용되는 곳은 주변에서 다양하게 관찰할 수 있다. 기업 내부적으로는 콜센터, 고객 의견 피드백 등과 같은 데이터를 활용하여 분석할 수 있으며 외부적으로는 기업 관련 언론 보도나 SNS의 홍보 자료 등에 달린 이용자의 게시글이나 댓글을 통해 긍정적·부정적 반응을 판단하는 등 다양한 방면에 이용되고 있다. 개인 측면에서는 영화 감상 전 리뷰 참고 등과 같이 특정 제품, 서비스 등의 이용 여부에 대한 의사를 결정하는 데 활용할 수 있다.

2. 감성분석 방법

감성분석은 크게 두 가지 단계로 이루어져 있다. 첫 번째는 문서의 어떤 부분에 의견이 내재 되어 있는지를 정의(Opinion definition)한다. 다음 단계로는 첫 번째 과정을 통해 취합된 의견을 요약(Opinion summerization)하게 된다.

정의에서 가장 먼저 수행해야 하는 것은 분석에 필요한 네 가지 요소를 검색하는 것이다. 네 가지 요소는 분석 대상 개체(Entity)나 개체특성

(Aspect/Feature), 개체 관련 의견에 내재되어 있는 감성(Sentiment), 의견 표현 주체(Opinion holders), 발화 시점(Time)이다.[60]

일반적으로 일반 의견(Regular opinions)과 비교 의견(Comparative opinions)의 두 가지로 분류할 수 있다. 전자는 단일 개체나 개체 특성에 관한 의견을 표현한 것이다. 일반 의견은 또 두 가지로 분류할 수 있다. "A 제품은 좋다" 같은 직접적 의견(Direct opinions)과 "A 제품은 잘 작동한다"와 같은 간접적 의견(Indirect opinions)으로 분류할 수 있다. 비교 의견은 이처럼 대상에 대한 두 개 이상의 언급을 통해 상호 간 의견을 나타내는 것이다.

비교 의견은 기본적으로 2개 이상의 대상이 언급되어야 하기 때문에 분석이 복잡한 측면이 있다. 일반 의견에 한정하여 감성분석 수행에 필요한 요소들을 살펴보면, 일반 의견은 대상(Target)이 하나이기 때문에 대상, 감성, 주체, 시점 등 네 가지 요소를 아래와 같이 나타낼 수 있다.

$$(g_i, so_{ijk}, h_j, t_k)$$

위 표현에서 분석 대상은 g_i, 발화 주체는 h_j, 발화 시점은 t_k로 나타낸다. so_{ijk}는 i에 대해서 k시점에 j가 표현한 감성이다.

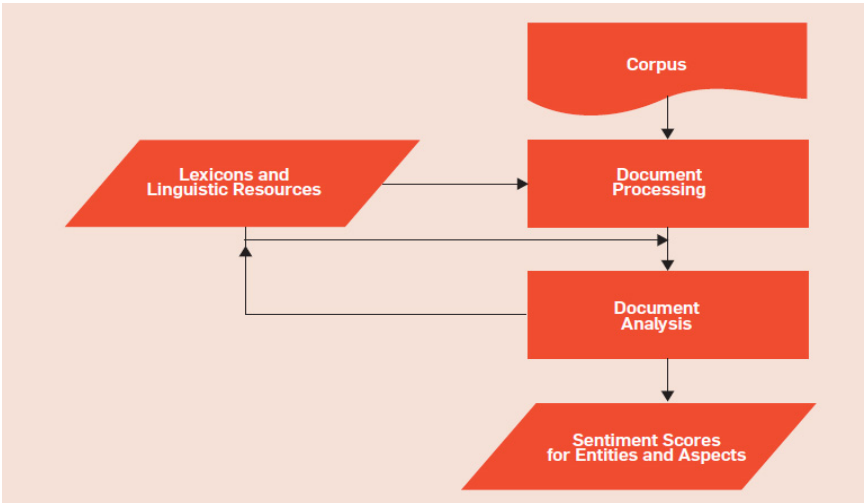
분석 대상이 하나인 경우에도 개체에 대한 서술인지, 개체 특성에 대한 서술인지를 구분하고자 g를 e(entity)와 a(aspect/feature)로 구분하여 나타내기도 한다. 이때, 문장에서 의견을 정의하는 경우 아래의 다섯 가지 요소를 분석 대상으로 나타내게 된다.

$$(e_i, a_{jl}, so_{ijkl}, h_j, t_k)$$

60) Data Science 블로그(검색일: 2021. 9. 12)

위에 언급된 다섯 가지 요소를 활용하면 문장에서 의견이 나타나는 부분을 구조화할 수 있다. 문장 외에 문서와 마찬가지인 긴 텍스트에 내재된 감정 분류 시에도 동일한 요소를 추출하여 나타낼 수 있다. 일반적인 감성분석 시스템의 순서도는 아래와 같다.

〈그림 3-3〉 감성분석 시스템 순서도



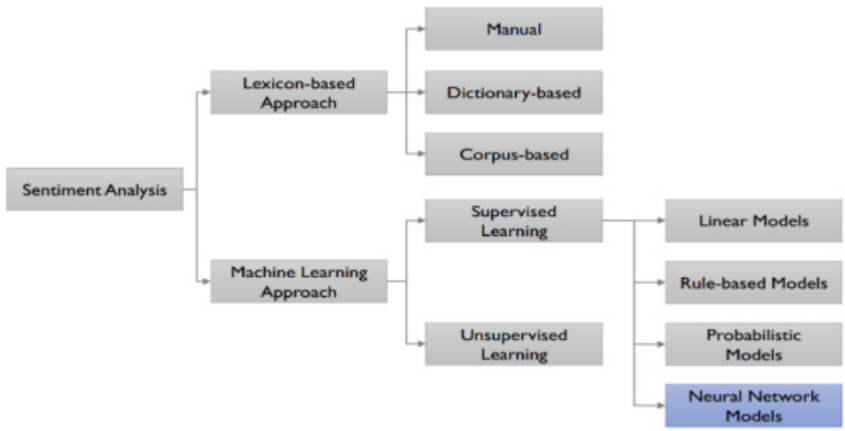
자료: ceine(검색일: 2021. 9. 12)

실제로 감성분석을 수행할 때에는 요소의 개수를 줄이기 위해 "하나의 문서 내에서는 한 사람이 한 개체에 대해서만 즉시 평가한다"라고 가정한다. 주요 분석 대상인 고객 피드백, 리뷰 등은 발화자 본인이 이용한 제품에 대해 평가하는 것이기 때문에 위의 가정을 대부분 만족하게 된다. 해당 가정을 사용하면 다섯 개 요소에서 대상(e,a), 발화자(h), 발화 시점(t) 등을 무시할 수 있게 되며, 문서에서 나타나는 감성 표현만을 적절하게 추출해도 분석 가능하다.[61]

61) Data Science 블로그(검색일: 2021. 9. 12)

감성분석의 방법론은 〈그림 3-4〉와 같이 구분된다. 큰 범주는 어휘 기반(Lexicon-based) 감성분석과 머신러닝 기반(ML-based) 감성분석으로 분류할 수 있다. 각각의 방법론은 세부적인 방법들로 재분류할 수 있으며, 최근 BERT(Bidirectional Encoder Representations from Transformers)와 같은 트랜스포머 계열의 인공신경망 모델이 자주 사용되는 추세이다.

〈그림 3-4〉 감성분석 방법론 구분



자료: Data Science 블로그(검색일: 2021.8.22)

감성분석은 분석 대상인 텍스트의 범위를 어떻게 설정하느냐에 따라 종류를 크게 세 가지로 나눌 수 있는데, 문서 단위 감성분석(document-level), 문장 단위 감성분석(sentiment-level), 속성 단위 감성분석(aspect-level)이 있다.[62] 또한 감성분석은 사용한 방법론에 따라 머신러닝 기반의 기계학습 기법(machine-learning approach)과 감성사전을 기반으로 한 사전 기반 기법(lexicon-based approach) 등 두 가지로 분류된다.

62) 최건우·윤희성·김동원(2019), p. 277.

감성사전은 긍정 의미의 어휘와 부정 의미의 어휘를 구분해 놓은 어휘 사전으로, 일반적으로 활용되는 사전은 감성어휘사전(Opinion Lexicon)'[63]과 오픈소스(Open sourse)로 출시되어 많은 사용자들에 의해 활용되고 있는 센티워드넷(SentiWordNet)이 있다.[64] 감성어휘사전(Opinion Lexicon)은 긍정 어휘 2,006개와 부정 어휘 4,783개로 구성되어 있으며 어휘를 단순하게 긍정 어휘와 부정 어휘로 분류한 감성사전이다.

그러나 센티워드넷은 어휘를 품사별, 즉 명사, 동사, 형용사, 부사 등으로 구분하였으며 품사마다 긍정, 부정과 더불어 중립의 감성까지도 분류하여 어휘별로 극성에 대한 점수를 부여하여 보여 준다. 센티워드넷의 품사별 어휘의 개수는 명사 8만 2,115개, 동사 1만 3,767개, 형용사 1만 8,576개, 부사 3,621개로 구성되어 있으며, 품사별 감성값을 이용하여 어휘가 속한 문장의 극성을 계산할 수 있다.[65]

각 어휘들의 극성 점수는 감성이 긍정의 극성을 띨 경우, 0에서 1사이의 값으로 계산되며 감성이 부정의 극성을 띨 경우에는 -1에서 0사이의 값이 설정되어 계산된다. 값이 1에 가까울수록 긍정 성향을 나타내고 -1에 가까울수록 부정적 성향을 나타낸다. 어휘의 감성 극성값이 0의 값이면 어휘의 감성이 중립임을 의미한다. 각각의 어휘 집단은 동의어나 유의어가 같은 그룹에 포함되어 있으며, 기계학습 방법에 의해 구축된 센티워드넷은 어휘 목록(WordNet)의 어휘 데이터를 기반으로 구축되었다.[66] 일반적으로 감성사전 기반의 감성값의 합계, 평균 또는 빈도수 등을 이용하여 감성분석을 수행하며 문장이나 문서의 특성에 적합한 가중치를 부여하여 분석하기도 한다.

63) Hu and Liu(2004), pp. 168~177.
64) Esuli and Sebastiani(2006), pp. 417~422.
65) 정근웅·김종욱(2018), p. 148.
66) Baccianella et al.(2010), pp. 2200~2204.

〈표 3-4〉 감성에 따른 단어(예)

도메인(Domain)	기본 감성 어휘(Basic Sentiment Words)
슬픔(Sadness)	슬픔(sadness), 외로움(loneliness), 불행(unhappiness), 우울(depression)
부끄러움(Shame)	부끄러움(shame), 죄책감(guilt), 후회(regret), 당혹감(embarrassment)
분노(Anger)	분노(anger), 짜증(irritation), 혐오(disgust), 분노(rage)
공포(Fear)	공포(fear), 걱정(worry), 염려(anxiety), 초조함(nervousness)
좌절(Frustration)	좌절(frustration), 체념(resignation), 무력(powerlessness), 절망(despair)
기쁨(Delight)	기쁨(delight), 즐거움(pleasure), 환희(joy), 행복(happiness)
만족(Satisfaction)	만족(Satisfaction), 희열(gratification), 성취감(fulfillment)

자료: 김동원(2017) p. 27, 재인용: Tronvoll(2011), Esuli(2006)[67]

제3절 서포트벡터머신

1. 서포트벡터머신 정의

서포트벡터머신(이하 SVM)은 널리 사용되고 있는 신경망보다 사용이 간결하며 분류나 회귀 분석 모두 사용이 가능하지만 분류에서 사용하게 되는 기법으로 초평면(Support Vector Hyperplane)을 이용해 분류를 수행하게 되는 알고리즘이다. 평면 상태에서 초평면(Hyperplane)으로 경계를 나눠 주어 분류하는 방식을 의미하는 것이다. 이때 어느 한쪽으로 치우치지

67) 김동원(2017), p. 27, 재인용: Tronvoll(2011), pp. 111-134, Esuli(2006), pp. 417-422.

않게 분류해야 하며, 공통되는 빈 공간의 경우에도 양쪽 데이터와의 균등한 위치에 분류 기준을 제시해야 한다.

직관적으로, 초평면이 가장 가까운 학습 데이터 점과 큰 차이를 가지면 분류 오차(classifier error)가 작으므로 좋은 분류를 하기 위해서는 분류된 점에 대해 가장 가까운 학습 데이터와 가장 먼 거리의 초평면을 탐색해야 한다. 일반적으로 초기 문제가 유한 차원의 공간에서 다루어지게 되는데, 종종 데이터의 선형 구분이 이루어지지 않는 문제가 발생한다.[68]

이같은 문제 해결을 위해 초기에 발생하는 문제의 유한 차원에서 더 고차원으로 대응시켜 분리를 편리하게 하는 방법이 제시되었다. 해당 분리과정에서 계산량이 증가하는 것을 방지하기 위해 각각의 문제들에 적절한 커널 함수$k(x,y)$를 정의한 SVM 구조를 설계함으로써 초기 문제의 변수들을 이용하여 내적 연산을 효율적으로 계산하도록 한다.[69] 높은 차원 공간에서의 초평면은 상수 벡터의 내적 연산 및 점들의 집합으로 정의할 수 있다. 초평면에 정의된 벡터는 데이터베이스 내에서 표현되는 이미지 벡터 매개변수들과 선형적 결합이 가능하도록 선택된다. 이렇게 선택된 초평면에서 대응된 점 x는 아래의 커널식과 같은 관계가 성립 할 수 있다.[70]

$$\sum a_i k(x_i, x) = constant$$

만약 $k(x,y)$가 x와 y가 멀어질수록 작아지게 된다면, 각각의 합산은 테스트 점 x와 대응되는 데이터 점 x_i 와의 근접성 정도를 표현하게 된다. 이같은 방식으로, 앞서 언급한 커널식의 합은 구별하고 싶은 집합 내의 데이터 점과 테스트 점 간에 상대적인 근접성을 측정할 수 있다. 초기 공간

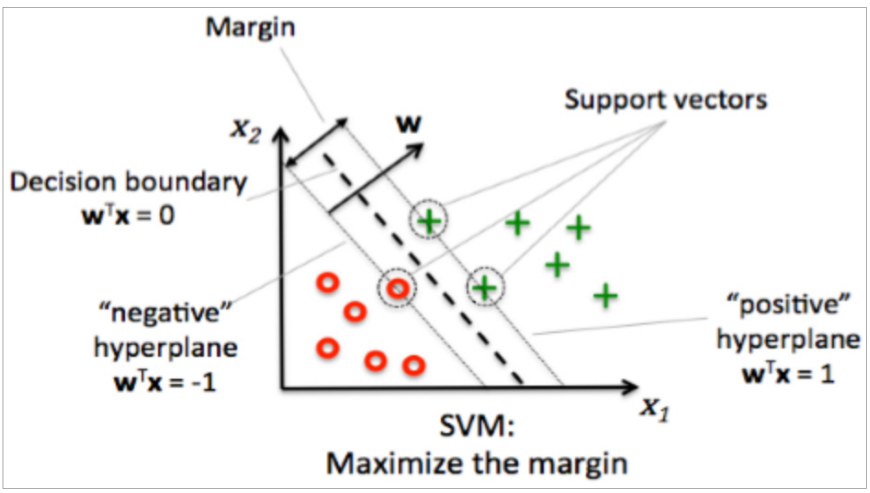
68) 삼성 소프트웨어 아카데미(SSAFY) 서포트백터머신(검색일: 2021. 08.22)
69) William H. Teukolsky, Saul A. Vetterling, and William T(2007), “Numerical Recipes: The Art of Scientific Computing”, 3th ed., Cambridge University Press, pp. 890-899.
70) wikipedia 서포트벡터머신(검색일 : 2021. 7. 16)자료를 바탕으로 저자 재장리

내 볼록하지 않은 집합의 점 x가 높은 차원으로 대응될 시 더 복잡해질 수 있어 주의해야 한다.[71]

또한 SVM은 기계학습의 한 방법으로서 지도학습(supervised learning) 모델 중 가장 많이 사용되고 있다. SVM은 퍼셉트론(perceptron)의 개념을 확장한 알고리즘이나, 퍼셉트론은 분류 오류를 최소화한 데 반해 SVM은 마진(margin)을 최대로 하므로 차이가 있다. 퍼셉트론은 항시 결정경계가 선형이어야 하기 때문에 범용 함수 근사기가 될 수 없다. 그러나 SVM은 선형으로 분리할 수도 있는 고차원으로 효율적으로 매핑이 가능한 커널을 사용하기 때문에 퍼셉트론의 한계이자 단점을 극복할 수 있는 것이다.

예로 2차원에서 하나의 기준이 되는 선이 마진이며 계량에서 일컫는 일종의 최소제곱법으로 이해할 수 있다. 〈그림 3-5〉의 그래프를 보면 빨간색 원과 초록색 더하기로 분류된 집단이 있다.

〈그림 3-5〉 서포트벡터머신



자료: 장삼용 머신러닝 온라인 학습자료(2016), (검색일: 2021. 8. 13)

71) 삼성 소프트웨어 아카데미(SSAFY) 서포트백터머신(검색일: 2021. 08. 22)

SVM은 두 집단을 분류하는 경계선 중 마진과 가장 멀리 떨어져 있는 경계선을 찾아내는 알고리즘이다. 이 알고리즘은 경계선 가까이에 있는 서포트벡터를 지나는 선과 거리가 최대가 되는 경계선을 구하는 과정이다. SVM이 퍼셉트론의 개념으로부터 출발하고 확장되었으므로 입력되는 트레이닝 데이터와 입력값과 곱해지는 가중치와 관련된 수식들은 다음과 같이 정리될 수 있다.[72]

$$\frac{1}{2}\sum_{j} w_j^2 + c\sum_{i}\xi^{(i)}$$

j는 트레이닝 데이터 특성값의 개수가 범위가 되며, i는 트레이닝 데이터의 개수가 범위가 된다. c는 상수와 비슷하며, ξ는 여유변수(slack variables)로, 이 변수는 분류 오류가 생기는 데이터에 적절한 제약(페널티)을 부여하여 비선형적으로 분리되는 모델을 완화시켜 최적화된 수렴값을 가지도록 하는 선형제약값이다.[73]

2. SVM 특징 및 장단점

SVM은 전술한 바와 같이 다양한 특징을 보이는데, SVM은 텍스트와 하이퍼텍스트의 분류에 있어 학습 데이터양을 상당히 감소할 수 있게 해 준다. 또한 이미지를 분류하는 작업에서 SVM을 사용할 수 있으며, SVM은 기존 쿼리(Query) 개량 구조 대비 월등히 정확도가 높은 검색률을 보인다. SVM은 이 외에도 분류된 화합물로부터 단백질을 90% 수준까지 구분하는 의학 분야에 유용하게 사용되기도 하며, 손글씨의 특징 또한 인지할 수 있다.

72) 장삼용 머신러닝 온라인 학습자료(2016), (검색일: 2021. 8. 13)
73) 위의 자료(검색일: 2021. 8.13)

SVM의 장점은 강력한 모델로 다양한 데이터 세트에서 잘 작용하여 범주나 수치 예측 문제에 사용할 수 있으며, 오류 데이터의 영향을 받지 않는 데 있다. 또한 과적합이 되는 경우가 적고 신경망보다 사용하기 쉽다는 장점이 있다. 데이터 특성이 얼마 되지 않을지라도 복잡한 결정경계를 찾는 것이 가능하며, 특히나 저차원과 고차원의 데이터(특성이 적을 때와 많을 때)에서 잘 작동한다.[74]

반면 SVM의 단점은 다수의 테스트 조합이 필요하다는 것인데, 최적의 모델을 찾기 위해 커널과 모델에서의 다양한 테스트가 필요하다. 샘플이 방대할수록 속도가 느려지고 메모리 할당이 커지며, 성능이 제대로 발현되기 어렵다. 따라서 여러 가지 연산이 필요하며 입력 데이터 세트가 방대할 경우에는 학습 속도가 느리다. 또한 데이터 전처리와 매개 변수 설정에 따라 정확도가 상이하며, 신경망과 같이 해석이 복잡하고 어려운 블랙박스 형태를 구성하고 있다.

〈표 3-5〉 SVM의 장단점

구분	특징
SVM 장점	범주, 수치 예측 문제 사용 가능 오류 데이터의 영향이 적음 과적합이 발생하는 경우가 적음 신경망 대비 사용하기 쉽고 용이함
SVM 단점	여러 개의 조합 테스트 필요 학습 속도가 느림 신경망과 같이 어렵고 복잡한 블랙박스 형태를 지님

자료: 저자 작성

74) O'REILLY(검색일: 2021. 8. 13)

04
SVM을 활용한 해운시장 분석

제1절 자료 및 분석 방법

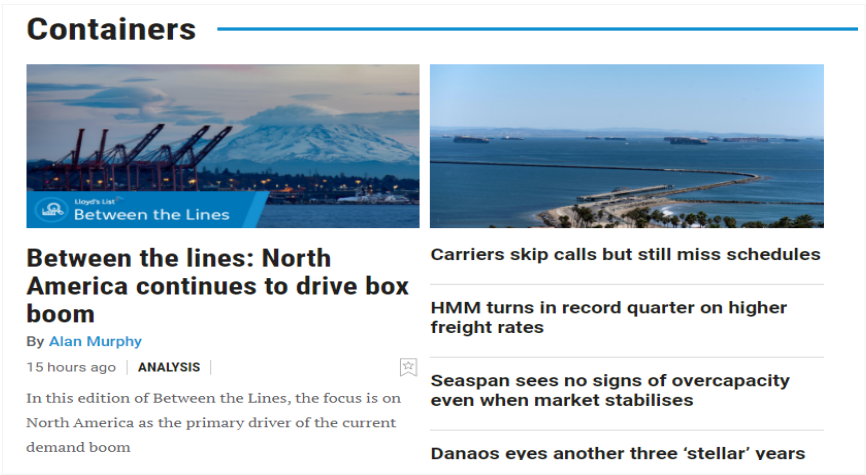
1. 자료

비정형 데이터를 활용한 해운시장 분석 방법은 학습 데이터에서 등장하지 않은 새로운 단어가 등장할 때 이를 분석하지 못하고 제외하는 단점이 존재한다. 따라서 학습 데이터와 테스트 데이터 간격이 길어질 경우 새로운 단어의 빈도수가 증가하여 충분한 데이터를 읽어 내지 못해 예측 성능이 하락할 가능성이 발생한다. 이에 따라 본 연구에서는 코로나19라는 팬데믹이 발생한 시점인 2020~2021년 자료를 활용하여 특정 단어들이 해운시황에 어떤 영향을 미치는지 분석하였다.

본 연구에서 활용된 로이즈리스트와 PR News를 선정한 배경은 첫째, 로이즈리스트 기사는 컨테이너, 건화물 등 선종별 분류가 잘되어 있으며 자료의 양이 풍부하기 때문이다. 비정형 데이터를 활용한 해운시장 분석을 위해서는 연구 목적에 맞는 충분한 양의 데이터를 확보하는 것이 핵심이다. 블룸버그 데이터는 자료의 양이 방대한 반면 특정 키워드(컨테이너 또는

건화물) 검색 시 목적(해운시황과)과 연관성이 떨어지는 기사들이 함께 다량 검색되어 정보를 가공할 때 연구자의 주관이 다수 개입되는 문제가 발생한다. 또한 JOC나 트레이드윈즈 등은 선종별 분류는 잘되어 있지만 로이즈리스트 대비 자료의 양이 상대적으로 적어 주 단위로 분류할 때 충분한 양의 데이터를 축적할 수 없다. 둘째, PR News는 컨테이너 정보만 제공하여 건화물 시황을 예측 실험할 수 없는 단점이 있지만 자료의 신속성, 정확성 및 명료성이 높아 비정형 데이터 가공이 편리하여 예측 실험 데이터로 선정하였다.

〈그림 4-1〉 로이즈리스트 컨테이너 기사



자료: 로이즈리스트(검색일 : 2021.10.23)

1) 로이즈리스트

로이즈리스트는 구독자들에게 시장에 대한 전망, 올바른 의사결정 등을 목표로 서비스를 제공하고 있다. 과거에는 종이신문 발간이 중심이었으나 2013년 이후 온라인 서비스로 재편되었다. 현재 해운시장을 컨테이너, 건

화물, 탱커와 가스, 기타로 구분하여 기사를 게재하고 있으며, 관련 섹터를 중심으로 기사를 매일 업데이트할 뿐만 아니라 정기적으로 시장 전망에 대한 보고서를 발간하고 있다.

본 연구에 활용된 로이즈리스트의 컨테이너 시황 자료는 2020년 1월~2021년 7월 기간의 자료를 수집하여 주 단위로 분류한 자료이다. 총 79주의 데이터를 수집하였으며, 이 중 2020년 1월~12월 데이터(50주)를 학습을 위한 데이터로 활용하였고, 2021년 1월~7월 데이터(29주)를 성능 평가를 위한 테스트 데이터로 활용하였다. 예측은 컨테이너 시황의 대표적인 지표 중 하나인 상하이 컨테이너 운임지수(SCFI: Shanghai Containerized Freight Index)를 활용하였으며, 특정 항로에 대한 기사의 편향성을 확인하기 위해 대표 항로인 중국~미주 서안, 중국~유럽 항로를 추가로 예측 실험하였다.

로이즈리스트의 건화물 시황 자료는 2020년 1월~2021년 8월 기간 데이터를 수집하여 2020년 1월~12월 데이터(52주)를 학습 데이터로 활용하였고, 2021년 1월~8월 데이터(35주)를 성능 평가를 위한 테스트 데이터로 활용하였다. 건화물 기사 또한 특정 선종의 편향성을 확인하기 위해 영국의 발틱 해운거래소(Baltic exchange)가 발표하는 건화물 종합지수인 BDI를 비롯하여 세부 선종별 지표인 BCI, BPI 및 BSI를 대상으로 예측 실험을 실시하였다.

2) PR News

PR News는 컨테이너 뉴스를 서비스 하고 있다. 이 서비스는 컨테이너 서비스에 국한되며 메일링을 통해 실시간으로 제공된다. 글로벌 컨테이너 선사들의 동향을 실시간으로 제공하기 때문에 텍스트마이닝에 활용될 가능성이 높다. 이 서비스에서 정보는 크게 네 가지로 구분되는데, 첫째는 선사

들의 운영 측면에서 변화가 있을 경우 정보를 제공하는 것이다. 예를 들어 임시결항(blank sailing)을 비롯한 항만 통과·변경, 신규 노선 개설, 서비스 중단 및 연장 등에 관련된 내용이다. 둘째, 컨테이너 선사의 정책에 관한 내용으로 운임 인상, 신조선 발주, 중고선 매입 등에 관련된 정보이다. 셋째, 항만 관련 정보로 항만 정체 및 개발에 관련된 정보를 제공한다. 마지막으로, 편집자의 코멘트로 현재의 컨테이너 시황 상황이나 이슈에 대해 간략히 정리하여 정보를 제공한다.

PR News는 2020년 2월~2021년 8월 기간 총 79주의 데이터를 수집하였으며, 2020년 2월~12월 데이터(44주)를 학습 데이터로 활용하였고, 2021년 1월~8월 데이터(35주)를 예측 성능 평가를 위한 테스트 데이터로 활용하였다. 실험 대상은 로이즈리스트와 동일하게 종합 SCFI 지수 및 SCFI 세부 항로 지표인 중국~미주 서안, 중국~유럽 항로에 대해 실험을 실시하였다. 하지만 PR News는 컨테이너 정보만 수집이 가능한 한계가 있어 건화물 시황 예측 실험은 실시하지 못하였다.

〈표 4-1〉 실험 테스트 및 예측 샘플 수

단위: 개(주)

구분	총데이터 수	학습 데이터 수	테스트(예측) 데이터 수
로이즈리스트 (컨테이너)	79	50	29
로이즈리스트 (건화물)	87	52	35
PR News (컨테이너)	79	44	35

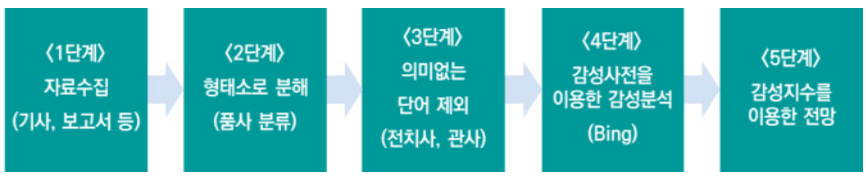
자료: 저자 작성

2. 분석 방법

1) 감성분석

감성분석 방법은 우선 비정형 데이터 자료를 수집한 후 〈그림 4-2〉와 같이 기사, 보고서 등 비정형 데이터의 의미 없는 단어를 제외하는 전처리 작업을 실시한다. 다음으로 전처리 작업을 거쳐 형태소로 분해된 단어들을 감성사전과 매칭하여 최종적으로 '긍정' 또는 '부정'의 빈도수를 활용해 해운시장을 예측하는 방식이다. 감성사전 기반의 분류는 감성사전을 구축한 연구자에 의해 '긍정' 및 '부정' 등으로 사전에 정의되어 있으며 'Bing', 'Afinn', 'NRC' 등 감성사전 종류에 따라 분류에 차이가 있다.

〈그림 4-2〉 감성분석 흐름도



자료: 저자 작성

본 연구에서는 R에서 제공하는 감성사전 중 단어에 대해 긍정(positive)과 부정(negative)으로 구분된 'Bing'을 사용하여 테스트를 실시하였다. 'Bing'은 총 6,778단어에 대해 '긍정'과 '부정'으로 구분되어 있으며, '긍정'과 '부정'의 빈도 수를 통해 자료가 어떠한 감성을 지니고 있는지 파악한다.

R을 이용하여 감성분석을 실시하는 코드는 〈표 4-2〉와 같다. 첫째, 작업 디렉토리를 지정하고 'dplyr'와 같이 감성분석에 필요한 관련 함수를 불러온다. 둘째, 'tidytext'를 이용하여 문장을 최소 단위인 형태소로 분리한다. 셋째, 쪼개진 형태소에서 일반적으로 필요 없는 단어(전치사, 관사)를 제외

한다. 또한 단어 외에 /, “, ?와 같이 의미 없는 기호도 제외한다. 마지막으로, 감성사전을 통한 빈도수를 활용해 예측을 실시한다.

〈표 4-2〉 감성분석 R 코드

구분	코드
작업 디렉토리 설정	# set working directory to where the corpus is stored on c drive setwd("D:/R_senti/data") text <- readLines("21-1-1.txt")
관련 함수 로딩	# set working directory to where the corpus is stored on c drive library(dplyr) text_df <- data_frame(line = 1:928, text = text)
형태소로 분리	library(tidytext) text_df %>% unnest_tokens(word, text) tidy_books <- text_df %>% unnest_tokens(word, text)
필요 없는 단어 제외	data(stop_words) tidy_books <- tidy_books %>% anti_join(stop_words) # Replacing “/”, “@” and “|” with space: reports <- tm_map(reports, toSpace, "/") reports <- tm_map(reports, toSpace, "@") reports <- tm_map(reports, toSpace, " \ |") reports <- tm_map(reports, toSpace, "/") reports <- tm_map(reports, content_transformer(tolower)) reports <- tm_map(reports, removeNumbers) reports <- tm_map(reports, removeWords, stopwords("english")) reports <- tm_map(reports, removeWords, c("blabla1", "blabla2")) reports <- tm_map(reports, removePunctuation) reports[[1]]$content

구분	코드
'Bing' 어휘집 적용	library(tidyr) tidy_books %>% inner_join(get_sentiments("bing")) %>% count(sentiment) %>% spread(sentiment, n, fill = 0) %>% mutate(sentiment = positive - negative)
워드클라우드	library(wordcloud) tidy_books %>% anti_join(stop_words) %>% count(word) %>% with(wordcloud(word, n, max.words = 100)) custom_stop_words
부정과 긍정 워드 분리하는 워드클라우드	library(reshape2) tidy_books %>% inner_join(get_sentiments("bing")) %>% count(word, sentiment, sort = TRUE) %>% acast(word ~ sentiment, value.var = "n", fill = 0) %>% comparison.cloud(colors = c("gray20", "gray80"), max.words = 100)

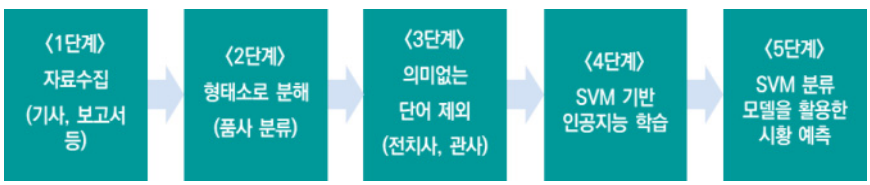
자료: 저자 작성

2) 서포트벡터머신(SVM)

본 연구에서는 〈그림 4-3〉과 같은 프로세스로 SVM을 활용한 해운시황 예측을 실시한다. 첫째, 해운 전문 기관 또는 전문가들이 발행하는 비정형 데이터를 수집하고 이를 최소 형태인 형태소로 분리하는 품사 분류를 실시한다. 둘째, 분석에 의미 없는 관사, 전치사 등을 제외하는 작업을 실시한다. 마지막으로, 전처리 작업을 거친 최종 단어들은 SVM을 활용한 인공지능 기반 학습 및 분류 기법을 통해 해운시장에 공시되는 컨테이너 및 건화

물 지수 SCFI, BDI 등과 매칭 학습시켜 '긍정' 또는 '부정'으로 분류 및 예측을 실시한다. SVM을 활용한 '긍정' 및 '부정' 분류는 사전에 정의된 감성사전 방법과 달리 특정한 단어가 SCFI, BDI와 같은 해운시황 지표 '상승' 및 '하락'에 어떠한 영향을 미치는지 학습을 통해 분류된다. 예로 'Supply'라는 단어가 등장했을 때 해운시황 지표가 상승했다면 'Supply'는 '긍정'의 값을 갖게 된다.

〈그림 4-3〉 SVM 흐름도



자료: 저자 작성

기사 및 보고서 등의 비정형 데이터를 분석하기 위한 비정형 데이터의 전처리 작업은 감성사전과 동일하게 R을 활용하여 실시하였다. 첫째, 작업 디렉토리를 설정하고 'tidyr'과 같이 단어 전처리에 필요한 함수를 읽어 온다. 둘째, 'word_tmp' 함수를 사용하여 문장을 최소 단위인 형태소로 분리한다. 셋째, 분리된 형태소에서 분석에 필요 없는 단어(관사, 전치사) 및 '/', '?', '!', 등과 같이 의미 없는 기호들을 제외한다. 넷째, 지금까지의 과정을 거쳐 토큰화된 단어들을 해운시장 지표와 매칭하여 분석을 실시한다. 이를 통해 해운시황에 '긍정' 또는 '부정'적인 영향을 미친 단어로 분류할 수 있으며, 이후 예측을 위한 테스트 데이터를 불러왔을 때 기존 학습 데이터를 활용해 분류된 단어들과 매칭하여 해운시황를 예측한다.

〈표 4-3〉 SVM 분석 R 코드

구분	코드
작업 디렉토리 설정 및 파일 불러오기	setwd('D://R_work') file_list = list.files(path='./data/',pattern='.txt') scfi = read_xlsx('./data/scfi.xlsx',sheet=2) names(scfi) <- c('scfi','name','num','date','updown')
형태소 분리	df <- NULL for(file in file_list){ if(substring(file,1,(nchar(file)-4)) %in% scfi$name){ path <- './data/' file_name <- paste0(path,file) word_tmp <- readLines(file_name) word_tmp <- tokenize_words(word_tmp) word_tmp <- unlist(word_tmp) word_tmp <- paste(word_tmp,collapse=',') title_tmp <- substring(file,1,(nchar(file)-4)) score_tmp <- scfi[scfi$name==title_tmp,'updown'] data_tmp <- data.frame(titel = title_tmp, content = word_tmp, scfi =score_tmp)
단어 전처리 및 의미 없는 단어 제외 (소문자 변경, 숫자 제거, 문장부호 제거, 의미 없는 단어 제거)	corpus <- Corpus(VectorSource(temp.train.df)) corpus <- tm_map(corpus, content_transformer(gsub), pattern="c(",replacement="(",fixed=TRUE) corpus <- tm_map(corpus, content_transformer(tolower)) corpus <- tm_map(corpus, removeNumbers) corpus <- tm_map(corpus, removePunctuation) corpus <- tm_map(corpus, removeWords, c("the", "and", stopwords("english")))
학습 및 예측 범위 설정	predict <- function(train_range,target_week){ train.df = df[train_range,] test.df = df[target_week,] temp.train.df <- strsplit(as.character(train.df$content),",") temp.test.df <- strsplit(as.character(test.df$content),",")
예측 결과 기록	result_class <- data.frame(updown = ifelse(df$updown[51:79]=="down","down","up"), pred = ifelse(hat.21==1,"up","down")) write.csv(result_class,"result_class.csv") write.csv(hat.21,"cc.csv“

구분	코드
	word_freq <- apply(X = X.train,MARGIN = 2,FUN = sum) words_coef <- data.frame(freq = c(0,word_freq), coef = cf) write.csv(words_coef,"words_coef.csv")
예측 정확도 평가	lconfusionMatrix(result_class$pred,result_class$updown)

자료: 저자 작성

제2절 분석 결과

감성분석 및 SVM의 해운시황 예측 정확도 비교에 활용된 자료는 〈그림 4-4〉와 같이 컨테이너 시황은 상하이발 컨테이너 운임지수인 SCFI를 사용하였으며, 건화물 시황은 발틱에서 발표하는 건화물 종합지수인 BDI를 사용하였다. 감성분석과 SVM을 통해 도출된 예측값을 SCFI 및 BDI 지표의 전주 대비 '상승' 및 '하락' 일치도를 비교해 예측 성능을 확인한다.

〈그림 4-4〉 SCFI 및 BDI(주 단위)

SCFI	txt name	num	date	updown
1,636.00	20-1-1	1	03-Jan-20	up
1,543.00	20-1-2	2	10-Jan-20	down
1,562.00	20-1-3	3	17-Jan-20	up
1,545.00	20-1-4	4	24-Jan-20	down
1,423.00	20-2-2	5	14-Feb-20	down
1,367.00	20-2-3	6	21-Feb-20	down
1,394.00	20-2-4	7	28-Feb-20	up
1,361.00	20-3-1	8	06-Mar-20	down
1,610.00	20-3-2	9	13-Mar-20	up
1,549.00	20-3-3	10	20-Mar-20	down
1,515.00	20-3-4	11	27-Mar-20	down
1,662.00	20-4-1	12	03-Apr-20	up
1,630.00	20-4-2	13	10-Apr-20	down
1,528.00	20-4-3	14	17-Apr-20	down
1,495.00	20-4-4	15	24-Apr-20	down

BDI	txt name	num	date	updown
941.50	20-1-1	1	03-Jan-2020	down
790.80	20-1-2	2	10-Jan-2020	down
763.60	20-1-3	3	17-Jan-2020	down
634.80	20-1-4	4	24-Jan-2020	down
519.00	20-1-5	5	31-Jan-2020	down
439.00	20-2-1	6	07-Feb-2020	down
419.20	20-2-2	7	14-Feb-2020	down
465.20	20-2-3	8	21-Feb-2020	up
519.00	20-2-4	9	28-Feb-2020	up
573.20	20-3-1	10	06-Mar-2020	up
627.60	20-3-2	11	13-Mar-2020	up
623.80	20-3-3	12	20-Mar-2020	down
585.40	20-3-4	13	27-Mar-2020	down
607.60	20-4-1	14	03-Apr-2020	up
610.50	20-4-2	15	10-Apr-2020	up

자료: 저자 작성

1. 감성분석

사전 기반 기법의 감성분석은 감성사전 혹은 특정 연구자가 제작한 사전을 이용하여 감성분석을 수행한다. 본 연구에서 사용된 'Bing'을 활용한 감성분석 또한 감성사전 기반의 연구 수행 방법이며, 〈표 4-4〉의 예와 같이 감성사전 'Bing'의 '긍정(positive)' 또는 '부정(negative)'으로 분류되어 정의된 사전을 활용하여 해운시황 예측 실험을 실시하였다.

〈표 4-4〉 Bing 감성사전(예)

단어	의미
2-faced	부정(negative)
2-faces	부정(negative)
a+	긍정(positive)
abnormal	부정(negative)
abolish	부정(negative)
abominable	부정(negative)
abominably	부정(negative)
abominate	부정(negative)
abomination	부정(negative)
abort	부정(negative)

자료: R 감성어휘집

1) 로이즈리스트

(1) 컨테이너

로이즈리스트 컨테이너 시장 보고서 기반 감성분석 예측은 종합 SCFI 지표와 특정 항로에 대한 편향성을 파악하기 위해 대표적인 컨테이너 항로라고 할 수 있는 중국~미국 서안 및 중국~유럽 항로의 SCFI 세부 지표를

활용하여 각각 실험하였다. 예측 결과 〈표 4-5〉와 같이 SCFI 예측은 2021년 1월~7월 데이터(29주)의 예측 샘플 중 15개가 예측 성공을 기록하여 51.7%의 예측 성공 확률을 기록하였다. SCFI 세부 지표인 중국~미국 서안 및 중국~유럽 항로 예측 성공률은 동일하게 58.6%로 SCFI 종합 지표 대비 소폭 높은 예측 성공률을 기록하였다.

〈표 4-5〉 감성분석을 활용한 컨테이너 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총예측 횟수	예측 성공	예측 실패
SCFI(종합)	29(100.0)	15(51.7)	14(48.3)
SCFI(중국~미국 서안)	29(100.0)	17(58.6)	12(41.4)
SCFI(중국~유럽)	29(100.0)	17(58.6)	12(41.4)

자료: 저자 작성

(2) 건화물

로이즈리스트 건화물 시장 보고서 기반 감성분석 예측은 발틱 건화물 종합 지표인 BDI와 세부 선종에 대한 편향성을 파악하기 위해 BCI, BPI 및 BSI를 추가하여 예측 실험을 하였다. 실험 결과, 〈표 4-6〉과 같이 BDI 지표는 2021년 1월~8월 데이터(35주)의 예측 샘플 중 21개가 예측 성공하여 60.0%의 예측 성공률을 기록하였다. 세부 선종 지표인 BCI, BPI 및 BSI의 예측 성공률은 각각 54.3%, 65.7% 및 68.6%를 기록하며, 중소형 건화물선 지표인 BPI 및 BSI의 예측 성공률이 종합 건화물선 지수인 BDI보다 소폭 높은 것으로 나타났다.

〈표 4-6〉 감성분석을 활용한 건화물 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총 예측 횟수	예측 성공	예측 실패
BDI	35(100.0)	21(60.0)	14(40.0)
BCI	35(100.0)	19(54.3)	16(45.7)
BPI	35(100.0)	23(65.7)	12(34.3)
BSI	35(100.0)	24(68.6)	11(31.4)

자료: 저자 작성

2) PR News

(1) 컨테이너

PR News 컨테이너 시장 보고서 기반 감성분석 예측은 로이즈리스트와 동일하게 SCFI 종합 지표와 중국~미국 서안 및 중국~유럽 세부 항로 지표를 각각 예측 실험하였다. 실험 결과, 〈표 4-7〉과 같이 종합 SCFI 지표는 2021년 1월~8월 데이터(35주)의 예측 샘플 중 13개가 예측 성공하여 37.5%의 낮은 예측 성공률을 기록하였다. SCFI 세부 지표 중국~미국 서안 및 중국~유럽 항로의 예측 성공률은 각각 48.6%, 45.7%를 기록하며 종합 SCFI 지표 예측 성공률 대비 약 10%p 높은 것으로 나타났다. 결론적으로 감성분석을 활용한 해운시황 예측 방법은 로이즈리스트의 데이터가 PR News 대비 상대적으로 활용하기에 적합한 것으로 파악된다.

〈표 4-7〉 감성분석을 활용한 컨테이너 시황 예측 결과(PR News)

단위: 개, %

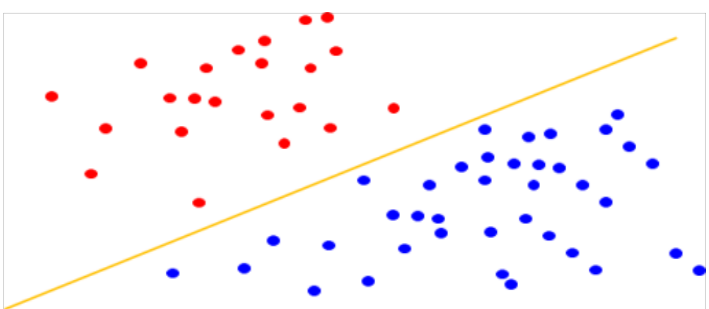
구분	총 예측 횟수	예측 성공	예측 실패
SCFI(종합)	35(100.0)	13(37.5)	22(62.5)
SCFI(중국~미국 서안)	35(100.0)	17(48.6)	18(51.4)
SCFI(중국~유럽)	35(100.0)	16(45.7)	19(54.3)

자료: 저자 작성

2. SVM 분석

SVM은 분류 과제에 사용할 수 있는 머신러닝 지도학습 모델이며 〈그림 4-5〉와 같이 결정 경계(Decision Boundary), 즉 분류를 위한 기준선을 정의하는 모델이다. 특정 연구자에 의해 분류된 감성사전을 이용한 감성분석 방법과 달리 SVM은 주어진 데이터들을 학습하여 두 개의 클래스로 분류하고 새로운 데이터가 두 클래스 중 어느 곳에 속하는지 결정한다. 따라서 SVM 분석에서는 감성사전 기법과 달리 사전에 정의되지 않은 비정형 데이터 학습을 통해 '긍정(positive)'과 '부정(negative)'으로 분류하고 예측 실험을 실시한다.

〈그림 4-5〉 SVM 결정경계



자료: 저자 작성

1) 로이즈리스트

(1) 컨테이너

2020년 1월~12월 총 1년간 로이즈리스트의 컨테이너 시황 기사들을 수집하여 문자 전처리 작업을 수행한 결과, 〈표 4-8〉과 같이 9,569개(중복 단어 제외)의 단어가 수집되었으며, 이 중 SCFI 지표 상승에 긍정적인 영향을 미친 단어는 5,112개, 부정적인 영향을 미친 단어는 4,457개로 분류되었다.

〈표 4-8〉 로이즈리스트 문자 전처리 결과(컨테이너 기사)

단위: 개

분류된 총단어	긍정어	부정어
9,569	5,112	4,457

자료: 저자 작성

컨테이너 시황 관련 기사에 SVM을 활용한 '긍정' 및 '부정'으로 분류한 단어는 〈표 4-9〉와 같다. '긍정'으로 분류된 단어는 일반적으로 긍정적인 의미라고 표현할 수 있는 'increase', 'positive' 등이 해운시황 지표에 긍정적(상승)인 영향을 미치는 단어로 분류되었으며, 'pandemic', 'congestion'과 같이 컨테이너 공급에 부정적인 영향을 미치는 단어들 또한 해운시황 지표에(상승)는 긍정적인 영향을 미치는 단어로 분류되었다. '부정'으로 분류된 단어는 하락 또는 부정적인 의미를 뜻하는 'disruption', 'drop', 'decline', 'reduce', 'outbreak' 등이 해운시황 지표에 부정적(하락)인 영향을 미치는 단어로 분류되었다.

〈표 4-9〉 SVM을 통해 분류된 긍정어 및 부정어(컨테이너 기사)

긍정어		부정어	
supply	increase	disruption	drop
cargo	transhipment	capacity	decline
demand	positive	executive	reduce
pandemic	customer	added	return
congestion	import	outbreak	carriers

자료: 저자 작성

로이즈리스트 컨테이너 기사 기반 SVM 기법을 활용한 컨테이너 시황 예측은 감성분석 방법과 동일하게 종합 SCFI 지표, 중국~미국 서안 및 중국~유럽 세부 항로 지표를 각각 예측 실험하였다. 예측 결과 〈표 4-10〉과 같이 종합 SCFI 지표는 2021년 1월~8월 데이터(29주)의 예측 샘플 중 17개의 샘플이 예측 성공을 기록하며 58.6%의 예측 성공률을 기록하였다. SCFI 세부 지표인 중국~미국 서안 및 중국~유럽 항로 예측 성공률은 각각 51.7%, 48.2%로 SCFI 종합 지표 예측 성공률 대비 상대적으로 소폭 낮은 예측 성공률을 기록하였다.

〈표 4-10〉 SVM 기법을 활용한 컨테이너 시황 예측 결과(로이즈리스트)

단위: 개, %

구분	총예측 횟수	예측 성공	예측 실패
SCFI(종합)	29(100.0)	17(58.6)	12(41.4)
SCFI(중국~미국 서안)	29(100.0)	15(51.7)	14(58.3)
SCFI(중국~유럽)	29(100.0)	14(48.2)	15(51.8)

자료: 저자 작성

(2) 건화물

2020년 1월~12월 총 1년간의 로이즈리스트 건화물 시황 관련 기사들을 수집하여 전처리 작업을 수행한 결과 〈표 4-11〉과 같이 8,741개(중복 단어 제외)의 단어가 수집되었으며, 이 중 건화물 지표에 긍정적인 영향을 미친 단어는 4,342개, 부정적인 영향을 미친 단어는 4,399개로 분류되었다.

〈표 4-11〉 로이즈리스트 문자 전처리 결과(건화물 기사)

단위: 개

분류된 총단어	긍정어	부정어
8,741	4,342	4,399

자료: 저자 작성

건화물 기사의 SVM을 활용한 두 클래스('긍정' 및 '부정')로 분류된 단어는 〈표 4-12〉와 같다. '긍정'으로 분류된 단어는 건화물 물동량에 큰 비중을 차지하는 품목인 'coal', 'iron' 등이 건화물선 지표에 긍정적(상승)인 영향을 미치는 것으로 분류되었으며, 컨테이너 시황과 동일하게 공급사슬에 부정적인 영향을 미치는 'pandemic'이라는 단어도 건화물선 지표에서는 긍정적(상승)인 영향을 미치는 단어로 분류되었다. 부정적(하락)인 단어로 분류된 단어들은 'emissions', 'imo', 'scrubber', 'decarbonisation'등과 같이 환경 규제와 관련된 단어들이 건화물선 지표에 부정적(하락)인 영향을 미치는 단어로 분류되었다.

〈표 4-12〉 SVM을 통해 분류된 긍정어 및 부정어(건화물 기사)

긍정어		부정어	
coal	positive	emissions	imo
exports	increase	risk	carbon
iron	largest	negative	scrubber
pandemic	contracts	safety	decarbonisation
develop	trade	australia	pressure

자료: 저자 작성

로이즈리스트 건화물 기사 기반 SVM 기법을 활용한 건화물 시황 예측 또한 감성분석 방법과 동일하게 발틱의 건화물선 종합 지표인 BDI와 세부 선종별 지표인 BCI, BPI 및 BSI를 각각 예측 실험하였다. BDI 예측 결과 〈표 4-13〉과 같이 2021년 1월~8월(35주)의 예측 샘플 중 15개의 샘플이 예측 성공을 기록하며 42.9%의 성공률을 나타내었다. 세부 선종인 BCI, BPI 및 BSI 예측 결과 각각 42.9%, 45.7%, 42.9%를 기록하며 BDI와 유사한 수준의 예측 성공률을 기록하며 뚜렷한 차이가 나타나지 않았다.

〈표 4-13〉 SVM 기법을 활용한 건화물 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총예측 횟수	예측 성공	예측 실패
BDI	35(100.0)	15(42.9)	20(57.1)
BCI	35(100.0)	15(42.9)	20(57.1)
BPI	35(100.0)	16(45.7)	19(44.3)
BSI	35(100.0)	15(42.9)	20(57.1)

자료: 저자 작성

2) PR News

(1) 컨테이너

2020년 2월~12월 총 11개월의 PR News 컨테이너 시황 관련 정보들을 수집하여 전처리 작업을 수행한 결과 〈표 4-14〉와 같이 총 5,565개의 단어(중복 단어 제외)가 수집되었으며, 이 중 컨테이너 지표에 긍정적인(상승) 영향을 미친 단어는 3,253개, 부정적인 영향을 미친 단어는 2,212개로 분류되었다. PR News는 정보를 요약해서 신속하게 발송하는 특징이 있기 때문에 로이즈리스트 자료 대비 상대적으로 작은 단어가 수집되었다.

〈표 4-14〉 PR News 문자 전처리 결과(컨테이너)

단위: 개

분류된 총단어	긍정어	부정어
5,565	3,253	2,212

자료: 저자 작성

PR News의 SVM을 활용한 두 클래스('긍정' 및 '부정')로 분류한 단어는 〈표 4-15〉와 같다. '긍정'으로 분류된 단어는 'access', 'add', 'addition', 'success' 등과 같이 일반적인 의미가 긍정으로 해석되거나 물동량 상승이 연상되는 단어들이 컨테이너 시황 지표에 긍정적인(상승) 영향을 미치는 것으로 나타났다. '부정'으로 분류된 단어들은 'drop', 'lockdown', 'regular', 'competition' 등과 같이 하락을 뜻하거나 규제와 관련된 단어들이 컨테이너 시황 지표에 부정적인(하락) 영향을 미치는 단어로 분류되었다.

〈표 4-15〉 SVM을 통해 분류된 긍정어 및 부정어(PR News)

긍정어		부정어	
access	calling	consumption	return
add	addition	drop	decline
correct	connect	suffered	regular
success	involved	competition	concern
great	extend	lockdown	reduces

자료: 저자 작성

PR News 또한 감성분석을 활용한 SCFI 예측 실험과 동일하게 종합 SCFI 지표 및 세부 항로 지표인 중국~미국 서안 및 중국~유럽 항로를 각각 예측 실험하였다. 실험 결과 〈표 4-16〉과 같이 종합 SCFI 지표 예측은 2021년 1월~8월(35주)의 예측 샘플 중 26개의 샘플이 예측 성공하여 74.3%의 높은 예측 성공률을 기록하였으며, 컨테이너 세부 항로 지표인 중국~미국 서안 및 중국~유럽 예측 성공률은 각각 65.7% 및 60.0%를 기록하며 종합 SCFI 지표 대비 소폭 낮은 예측 성공률을 기록하였다.

〈표 4-16〉 SVM 기법을 활용한 컨테이너 시황 예측 결과(PR News)

단위: 개, (%)

구분	총예측 횟수	예측 성공	예측 실패
SCFI(종합)	35(100.0)	26(74.3)	9(25.7)
SCFI(중국~미국 서안)	35(100.0)	23(65.7)	12(34.3)
SCFI(중국~유럽)	35(100.0)	21(60.0)	14(40.0)

자료: 저자 작성

제3절 감성분석 및 SVM 예측 성공률 비교

1. 로이즈리스트

1) 컨테이너

로이즈리스트 컨테이너 시황의 감성분석 및 SVM 분석 방법을 활용한 예측 실험 결과, SCFI 종합 지표는 감성분석 예측이 51.7%, SVM 예측이 58.6%를 기록하며 SVM 방법론이 감성분석 방법론 대비 약 7%p 높은 예측 성공률을 기록하였다. 반면 SCFI 세부 항로 지표인 중국~미국 서안 및 중국~유럽 항로 지표에 대한 예측 성공률은 감성분석이 58.6%를 기록하며 로이즈리스트를 활용한 예측 결과에서는 가장 높은 성공률을 나타내었다. SVM을 활용한 예측 방법은 세부 항로 지표 모두 SCFI 종합 지표 예측 성공률 대비 낮게 나타났다. 결론적으로 로이즈리스트 컨테이너 시황에 대한 감성분석 및 SVM 분석을 통한 예측 결과, SCFI 종합 지표는 SVM 기법 활용이 더 적합한 것으로 판단되며, 세부 항로 지표는 감성분석 방법이 더 적합한 것으로 파악된다.

〈표 4-17〉 컨테이너 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총예측 횟수	감성분석 예측 성공 횟수(성공률)	SVM 예측 성공 횟수(성공률)
SCFI(종합)	29(100.0)	15(51.7)	17(58.6)
SCFI(중국~미국 서안)	29(100.0)	17(58.6)	15(51.7)
SCFI(중국~유럽)	29(100.0)	17(58.6)	14(48.2)

자료: 저자 작성

2) 건화물

로이즈리스트 건화물 시황의 감성분석 및 SVM 분석 방법을 활용한 예측 실험 결과는 〈표 4-18〉과 같이 감성분석 예측 성공률이 SVM 예측 성공률 대비 모두 높은 것으로 나타났다. 먼저, BDI 예측 결과 감성분석이 60.0%를 기록하며, 42.9%의 예측 성공률을 기록한 SVM 대비 약 17%p 높은 예측 성공 확률을 기록하였다. 세부 선종별 지표인 BCI, BPI 및 BSI 예측 결과 역시 감성분석을 활용한 예측 성공률이 SVM 대비 모두 높게 나타났다.

특히 수프라막스선 지표인 BSI 예측 결과는 감성분석이 68.6%를 기록하며 SVM 대비 약 26%p 높은 성공률을 기록하였다. 결론적으로 로이즈리스트 건화물 시황의 비정형 데이터 분석은 감성분석이 SVM 방법론과 비교하여 더 적합한 모델이라고 판단할 수 있다.

〈표 4-18〉 건화물 시황 예측 결과(로이즈리스트)

단위: 개, (%)

구분	총예측 횟수	감성분석 예측 성공 횟수(성공률)	SVM 예측 성공 횟수(성공률)
BDI	35(100.0)	21(60.0)	15(42.9)
BCI	35(100.0)	19(54.3)	15(42.9)
BPI	35(100.0)	23(65.7)	16(45.7)
BSI	35(100.0)	24(68.6)	15(42.9)

자료: 저자 작성

2) PR News

PR News 컨테이너 시황의 감성분석 및 SVM 분석 방법을 활용한 예측 실험 결과는 〈표 4-19〉와 같이 SCFI 종합 지표에서 감성분석 예측이 35개의 샘플 중 13개가 예측 성공하며 37.5%의 성공률을 기록하였으며, SVM 방법론을 활용한 예측 방법은 26개의 예측 성공을 기록하며 74.3%로 감성분석 대비 약 두 배 높은 예측 성공률을 기록하였다.

SCFI 세부 항로 지표인 중국~미국 서안 및 중국~유럽 항로에 대한 감성분석을 활용한 예측 성공률은 각각 48.6%, 45.7%를 기록하였으며, SVM을 활용한 예측 성공률은 각각 65.7%, 60.0%를 기록하며 모두 SVM을 활용한 예측 결과가 높게 나타났다. 결론적으로 PR News를 활용한 컨테이너 시황 예측은 SVM 방법론이 감성분석 대비 높은 예측 성공률을 기록하며 상대적으로 더 적합한 모델이라고 판단할 수 있다.

〈표 4-19〉 컨테이너 시황 예측 결과(PR News)

단위: 개, (%)

구분	총예측 횟수	감성분석 예측 성공 횟수(성공률)	SVM 예측 성공 횟수(성공률)
SCFI(종합)	35(100.0)	13(37.5)	26(74.3)
SCFI(중국~미국 서안)	35(100.0)	17(48.6)	23(65.7)
SCFI(중국~유럽)	35(100.0)	16(45.7)	21(60.0)

자료: 저자 작성

05
결론 및 제언

제1절 결론

1. 해운시황 예측에 비정형 데이터 분석 방법론 적용

해운산업에서는 〈표 5-1〉에서 나타난 바와 같이 여전히 업계 전문가들의 경험으로 쌓인 노하우를 통한 직관에 의해 많은 거래가 이루어지고 있다. 물론 업계 전문가들의 직관이 과학적인 분석 방법보다 우수할 수 있겠지만 돌발적인 사건이 발생했을 때 시장을 바라보는 관점에 대한 차이가 나타날 수 있으므로 시장을 객관화할 수 있는 정량적인 방법론이 필요하다.

기존 정량적(정형) 데이터를 활용한 분석 방법은 해운시장에 대한 수요와 공급을 기반으로 과거와 현재의 추이를 분석하여 미래를 예측한다. 하지만 코로나19 이후 해운시장에서 발생하는 변동성이 큰 상황에서는 정량적 데이터를 활용한 분석 방법은 급격히 변화하는 단기적인 시장 흐름을 파악할 수 없는 한계가 발생할 수 있다. 따라서 본 연구에서는 코로나19 발생 이후 시점인 2020~2021년의 비정형 데이터를 활용해 시장 심리를 정량화할 수 있는 텍스트마이닝 기법 중 하나인 SVM 방법론을 적용하여 해운시황을 예측

하는 새로운 분석 방법의 활용 가능성을 살펴본 것에 의미가 있다.

〈표 5-1〉 기업 규모에 따른 선박투자 의사결정에 미치는 요인 중요도

저자	대기업	중소기업
과학적 분석	0.574	0.491
휴리스틱(직관)	0.345	0.293
시장 정서	0.081	0.216
합계	1.000	1.000

자료: 김경수·박은식·우수한(2014), p. 183

2. SVM에 활용된 비정형 데이터는 '양'보다 '질'이 중요

SVM 기법을 활용한 컨테이너 시황 예측은 로이즈리스트와 PR News 두 가지를 사용하였다. 〈표 5-2〉와 같이 전처리 작업을 거친 비정형 데이터의 단어 수(중복 단어 제외)는 로이즈리스트 및 PR News 각각 9,569개, 5,565개로 로이즈리스트 기사의 총 단어 수가 PR News 대비 약 4천 개 많은 것으로 나타났다. 하지만 예측 성공률은 PR News를 활용한 예측이 74.3%로, 로이즈리스트를 활용한 예측 성공률 대비 두 배 수준의 예측 성공률을 나타냈다. 결론적으로 SVM에 활용된 비정형 데이터는 자료의 '양'보다 시황에 적합한 데이터를 찾아 적용하는 것이 중요한 것으로 파악된다.

〈표 5-2〉 SVM 기법을 활용한 예측 결과(로이즈리스트 vs PR News)

단위: 개, %

구분	분류된 단어 수 (중복 단어 제외)	SCFI 예측 성공률
로이즈리스트	9,569	37.5
PR News	5,565	74.3

자료: 저자 작성

3. 시황에 적합한 비정형 데이터 활용 필요

제4장에서 수행한 감성분석 및 SVM을 활용한 예측 실험 결과에서 나타났듯이 특정한 방법론이 모든 시황 지표에서 우수한 예측 결과를 기록하지 못했다. 〈표 5-3〉과 같이 로이즈리스트를 활용한 SCFI 종합 지표 예측에서는 감성분석과 SVM 예측 성능이 비슷한 결과를 기록했지만, 건화물선 지표인 BDI 및 PR News를 활용한 예측 성능 결과는 극명한 차이가 나타났다. 따라서 해운시황 예측 성능을 높이기 위해서는 특정 시황에 적합한 비정형 데이터 발굴 및 방법론 적용이 필요하다.

〈표 5-3〉 감성분석 vs SVM

단위: %

예측 대상	감성분석 예측 성공률	SVM 예측 성공률
SCFI (로이즈리스트)	51.7	58.6
SCFI (PR News)	37.5	74.3
BDI (로이즈리스트)	60.0	42.9

자료: 저자 작성

4. 단어의 중의성 해소

감성분석은 감성사전 또는 특정 연구자에 의해 사전적으로 분류되어 특정 시점의 중의적인 단어 해석 및 문맥 파악에 한계가 있다. 하지만 SVM 방법론을 활용한 분류 방법은 특정 시점의 비정형 데이터와 해운시황을 매칭 학습하여 중의적인 단어가 해운시황에 '긍정적(positive)' 또는 '부정적

(negative)'인 영향을 미치는지 파악하여 두 클래스로 분류한다. 예로 'pandemic', 'congestion'과 같은 단어는 일반적으로 부정적 의미로 해석되지만, 본 연구에서 수행한 SVM을 활용한 분류 결과 'pandemic', 'congestion'과 같은 단어가 등장했을때 해운 지표가 상승하는 것으로 학습되어 해운 지표 상승에는 '긍정적(positive)'인 단어로 분류되었다.

결론적으로 SVM을 활용한 분류 방법은 감성사전 또는 특정 연구자에 의해 사전적으로 분류되어 있지 않은 특정 시황에 적합한 맞춤형 분류 사전 구축이 가능할 것으로 기대된다.

제2절 제언

1. 시장 심리를 측정한 해운시황 예측 방법 도입 및 서비스 제공

기존 해운시황을 예측하는 연구는 정량적 데이터를 활용한 전통적인 시계열 방법이나 인공지능 기법(LSTM, DNN 등)을 통해 예측하는 방법이 대부분이다. 이와 같은 방법론은 과거 및 현재의 정량적 데이터 특징이나 패턴을 분석하여 미래의 해운시황을 예측한다. 그러나 이러한 방법론은 돌발적인 이벤트 발생 시 정량적 데이터 추세의 특징이나 패턴에만 의존하여 예측하므로 단기적으로 시장 참여자의 시장 심리를 파악하지 못한다. 따라서 본 연구에서는 비정형 데이터를 정량화(객관화)할 수 있는 SVM 기법을 통해 시장 심리를 측정한 해운시황 예측 방법을 도입한 것에 의미가 있다.

〈표 5-4〉 이론적 및 산업적 기여 내용

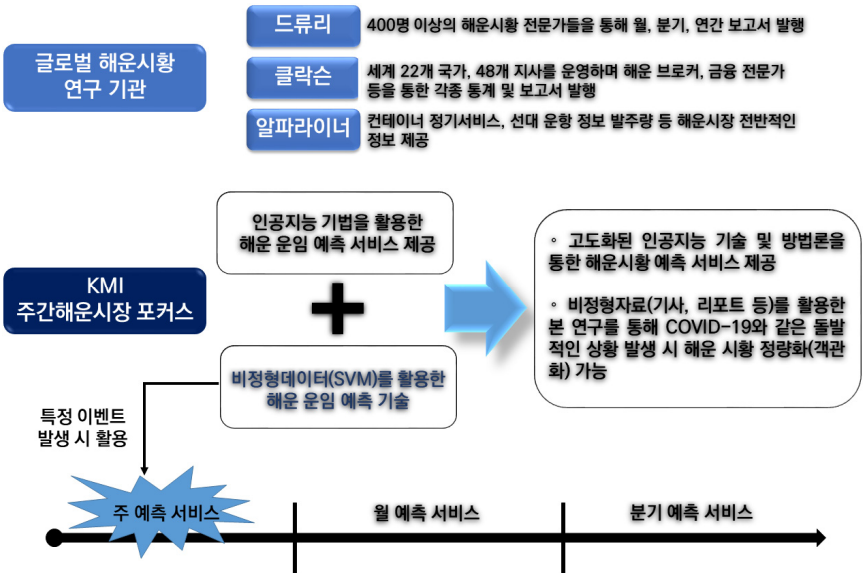
구분	내용
이론적 기여	시장 심리(비정형 데이터)를 객관화하여 해운 시황을 예측하는 방법론 도입
산업적 기여	고도화된 해운 운임 예측 서비스를 통해 관련 산업에서 비즈니스 경쟁우위를 점할 수 있도록 지원

자료: 저자 작성

대표적인 글로벌 해운시황 연구기관은 드류리, 클락슨, 알파라이너 등이 있으며 월, 분기, 연간 단위의 보고서, 리포트 등을 통해 해운시황 전망 서비스를 제공하고 있다. 대부분의 서비스는 관련 업계 전문가들이 통계 자료나 시장 상황 등을 종합한 오랜 노하우를 통해 직관적인 전망 분석이 이루어진다. 하지만 이러한 분석은 시장을 객관화(정량화)할 수 없기 때문에 해운시황을 전망하는 주체가 누구냐에 따라 시장을 바라보는 관점에 차이가 나타날 수 있어 주장의 설득력이 떨어질 수 있다.

하지만 현대사회는 인공지능 및 빅데이터 등 디지털 기술의 발전으로 이 같은 새로운 방법론을 활용한 연구가 고도화되고 있으며, 한국해양수산개발원에서는 『KMI 주간해운시장 포커스』를 통해 인공지능 기법을 활용한 건화물선 주간 운임 예측 서비스를 제공하고 있다. 본 연구 결과에서 도출된 시장 심리를 정량화(객관화)하는 기법은 인공지능 예측 서비스에 하나의 변수로 활용할 수 있다. 특히 코로나19와 같이 돌발적인 이벤트 발생 시 실시간으로 배포되는 비정형 데이터(기사, 리포트 등)의 정량화(객관화)는 단기적인 예측 서비스의 질적 향상에 큰 도움이 될 것으로 기대된다.

〈그림 5-1〉 연구 결과 활용



자료: 저자 작성

2. 다차원·다양한 연구 실험 필요

본 연구에서는 컨테이너 시황의 경우 총 세 가지(종합 SCFI 지표, 중국~미국 서안, 중국~유럽) 예측 테스트를 실시하였으며, 건화물 시황은 총 네 가지(BDI, BCI, BPI, BSI) 테스트를 수행하였다. 하지만 SVM의 단점인 실험 시간이 길다는 한계로 인해 다차원적인 실험을 수행하지 못했다. SVM 방법에서 두 클래스로 분류하는 결정계수를 활용하는 방법은 본 연구에서 사용된 평면(1차원)적 방법뿐만 아니라 다차원적으로 분석하는 방법이 있다. 또한 두 클래스 사이의 마진(결정계수와 입력값의 거리)을 최대로 하는 방법과 분류 오류를 최소화하기 위한 마진을 최소화하는 방법 등 최적의 파라미터값을 찾는 다양한 연구 실험을 수행할 필요가 있다.

3. 체계적인 비정형 데이터 DB 구축 필요

비정형 데이터 분석에서 가장 대표적인 문제점 중 하나는 데이터 수집 문제이다. 전통적인 예측 방법인 계량적 분석을 위한 정형 데이터와 달리 텍스트마이닝을 위한 비정형 데이터는 특정한 형식의 구조를 가지고 있지 않기 때문에 체계적인 수집 및 축적에 한계가 있다. 온라인상에는 빅데이터라고 할 수 있는 SNS, 기사 및 보고서 등 다양한 비정형 데이터가 존재하지만 대부분의 가치 있는 자료들은 보안이나 지적재산권의 이유로 웹페이지에 존재하는 자료를 색인하고 저장하는 웹크롤링(web crawling) 작업이 불가능하다. 따라서 비정형 데이터를 활용한 예측 기법의 성능을 향상시키기 위해서는 체계적인 비정형 데이터 DB 구축이 필요하다.

4. 복합어 처리에 대한 해결 방안 필요

비정형 데이터를 활용한 분석 방법에서 단어 전처리 과정은 핵심이라고 할 수 있을 만큼 중요한 과정이다. 영어 원문의 특성상 'Supply Chain'과 같은 단어들이 합쳐져 하나의 뜻을 이루는 문자들이 존재하지만 본 연구에서는 이러한 문자를 'Supply'와 'Chain'으로 각각 분리하여 분석한다. 이러한 전처리 과정에서 발생된 오류들이 쌓이는 문제로 분석을 하였을 때 성능이 떨어질 가능성이 크므로 단어 전처리 과정에서 오류를 줄일 수 있는 명확한 기준을 제시할 필요가 있다.

참고문헌

국내 문헌

김경수·박은식·우수한(2014),「해운기업의 선박투자 의사결정 방식에 관한 연구」,『물류학회지』, 24(2), pp. 183.

김동원(2017),「The Business Impact of Social Media」, 한국해양대학교 대학원 박사학위 논문, p.27.

김동균·최정석(2021),「딥러닝 기법을 활용한 컨테이너선 운임 예측 모델」,『해양환경안전학회지』, 27(5), pp. 574~583.

김수현 외(2020),「경제분석을 위한 텍스트 마이닝」, BOK 경제연구 제2019-18호, 한국은행 경제연구원, p.1.

김은미(2021),「감성분석을 이용한 뉴스정보와 딥러닝 기반의 암호화폐 수익률 변동 예측을 위한 통합모형」,『한국지식경영학회지』, 22(2), pp. 19~32.

김유일·신은경·홍태호(2004),「신경망과 SVM을 이용한 주가지수예측의 비교」,『인터넷전자상거래연구』, 제4권 제3호, pp. 221~243.

김용환(2016),「통계전문교재 개발 및 개선」, 통계청 통계교육원 교육자료.

김재봉·김형중(2017),「주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축 방안」,『한국디지털콘텐츠학회 논문지』, 18(3), pp. 585~592.

구미숙 외(2006),「데이터 마이닝 기법을 이용한 XML 문서의 온톨로지 반자동 생성」,『정보처리학회 논문지』, 제13권 3호, pp. 585~592.

국가기록원(2017),「차세대 기록관리 모델 재설계 연구 개발」, p. 359.

노설현(2020),「토픽모델링을 활용한 인공지능 관련 이슈 분석」,『디지털융복합연구』,Vol. 18. No. 5, pp. 75~87.

다오두안홍·안현철(2014),「주식 시장 예측을 위한 π-퍼지 논리와 SVM의 최적 결합」,『한국지능정보시스템학회지』, 20(4), pp. 43~58.

박선 외(2011),「신경망과 SVM을 이용한 적조 발생 예측」,『대한전자공학회지』, 2011. 9., pp. 39~45.

박성화·권장한·김태일(2018),「뉴스 빅데이터를 이용한 해운뉴스지수 개발」,『해양정책연구』, 33(1), pp. 281~298.

송치영(2002),「뉴스가 금융시장에 미치는 영향에 관한 연구」,『국제경제연구』, 8권 3호, pp. 1~34.

안동진(2020),「텍스트마이닝을 적용한 한국과 중국의 청소년 비만 연구동향 분석」, 한국교원대학교 체육교육 박사학위 논문

오세경·최정원·장재원(2017),「빅데이터를 이용한 딥러닝 기반의 기업 부도예측 연구」, 한국금융연구원,『KIF Working Paper』, pp. 1~4, 73~74.

오하영·박정식(2017),「컴퓨터와 R을 활용한 텍스트 데이터 분석 기초」, 휴먼 싸이언스, pp. 20~21.

윤원중 외(2009),「금융위기 조기경보시스템에 관한 연구」,『한국지능시스템학회』, 제 19권 제2호, pp. 238~240.

이구형(1993),「감성공학의 연구와 제품개발에의 응용」,『대한인간공학회 춘계학술대회논문집』, pp. 59~63.

이우식(2017),「딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측」,『한국데이터정보과학지』, 28(2), pp. 287~295.

이정훈(2018),「감성분석 연구동향」,『한국정보처리학회지』, 2018년도 춘계학습발표대회 2018 May 11, pp. 358~361.

이재윤·문주영·김희정(2007),「텍스트 마이닝을 이용한 국내 기록관리학 분야 지적 구조 분석」,『한국문헌정보학회지』, 제41권 제1호, p.347.

이태원·홍태호(2017),「SVM+MTL을 이용한 감성분류 통합모형: 온라인 쇼핑몰의 고객리뷰를 중심으로」,『Entrue Journal of Information Technology』, 16(1), pp. 81~95.

이형용(2008),「한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측

기법 결합모형」, 『대한산업공학회지』, 7(13), pp. 33~43.

임상섭·윤희성(2018a), 「케이프선 시장 운임의 결정요인 및 운임예측 모형 분석」, 『한국항해항만학회지』, 42(6), pp. 539~545.

_____(2018b), 「ARIMA-SVM 하이브리드 모델을 활용한 탱커 운임지수 예측」, 『한국금융공학회지』, Volume 17, pp. 79~98.

윤희성 외(2017), 「빅데이터를 활용한 해운시장분석 연구-빅데이터 파악 및 확보방안」, 한국해양수산개발원, 연구개발적립금사업 2017-01, pp.63~64.

전채남(2016), 「언어 자료로 세상 보기 : 산업 분야의 언어 처리와 세종 말뭉치 운용」, 『새국어생활』, 26권 2호, pp. 57~72.

정근웅(2018), 「감성분석을 통한 온라인 구전과 판매 성과에 대한 요인 연구 : 웹사이트 스팀을 중심으로」, 성균관대학교 대학원 석사학위 논문, p.12.

정근웅·김종욱(2018), 「감성분석 기반의 게임소비자 온라인 구전효과」, 『한국디지털정책학회논문지』, Vol.16. No.3, p.148.

정지선·김동성·김종우(2015), 「온라인 언급이 기업 성과에 미치는 영향 분석: 뉴스 감성분석을 통한 기업별 주가 예측」, 『지능정보연구』, 21(4), pp. 37~51.

조태호(2001), 「텍스트 마이닝의 개념과 응용」, 『지식정보인프라』, pp. 76~85.

최건우·윤희성·김동원(2019), 「감성분석을 활용한 컨테이너 시장 예측」, 『해양정책연구』, 34(2), pp. 274~286.

최윤정·박승수(2002), 「웹 컨텐츠의 분류를 위한 텍스트마이닝과 데이터마이닝의 통합 방법 연구」, 『인지과학』, p. 34.

허준영·양진용(2015), 「SVM 기반의 재무 정보를 이용한 주가 예측」, 『한국정보과학회지』, 제21권 제3호, pp. 167~172.

홍진주 외(2016), 「감성분석과 SVM을 이용한 인터넷 악성댓글 탐지 기법」, 『한국정보통신학회논문지』, 20(2), pp. 260~267.

홍태호·김은미·차은정(2017), 「뉴스 감성분석과 SVM을 이용한 다우존스 지수와 S&P500 지수 예측」, 『한국인터넷전자상거래학회지』, 17(1), pp. 23~36.

KMI 동향분석(2018), 「빅데이터 분석은 해운에서 어떻게 활용되는가」, 통권 제90호.

Minghao Piao·변정용(2017), 「SVM을 이용한 건강검진정보 기반 진료과목 예측」, 『정보처리학회논문지』, 6(6), pp. 303~308.

국외 문헌

Andreas Hotho, Andreas Nurnberger & Gerhard Paaß(2005), "A Brief Survey of Text Mining", The Journal for Language Technology and Computational Linguistics, pp. 398~401.

Bollen, J., Mao, H. & Zeng, X.(2011), "Twitter mood predicts the stock market", Journal of Computational Science, 2, pp. 1~8.

Chen, Nai-Fu, Richard Roll & Stepen A. Ross(1986), "Economic forces and the stock marꟾket," Journal of Business, 59(3), pp. 383~403.

Chen, H. H., Chen, M. & Chiu, C. C.(2016), "The integration of artificial neu- ral networks and text mining to forecast gold futures prices", Comꟾmunications in Statistics-Simulation and Computation, 45(4), pp. 1213~1225.

Diener, E., Smith, H & Fujita, F(1995), " The personality structure of affect", Journal of Personality and Social Psychology, 69, p. 130.

Esuli, A. & F. Sebastiani(2006), "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," Proceedings of the 5th International Conference on Language Resources and Evaluation, Vol.6, pp. 417~422.

Elisa Claire Aleman Carreon, Hirofumi Nonaka & Toru Hiraoka(2018), "Relation Analysis between Hotel Review Rating Scores and Sentiment Analysis of Reviews by Chinese Tourists Visiting Japan", *The japanese journal of the institute of industrial applications engineers*, 6(2), pp. 95~99.

Guan, Feng. et al.(2016), "Multi-Step hybrid prediction model of Baltic Supermax Index based on support vector machine", *Journal Neural network world*, 26 (3), pp. 219~232.

G.Miner. et al.(2012), "Practical text mining and statistical analysis for non-structured text data applications", Academic Press, pp. 2~31.

Hu, M. & B. Liu(2004), "Mining and Summarizing Customer Reviews", *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp.168~177.

Jochen Dörre, Peter Gerstl & Roland Seiffert(1999), "Text mining: finding nuggets in mountains of textual data", *Proceedings of the fifth ACM SIGKDD interna tional conference on Knowledge discovery and data mining,* Association for Computing Machinery,pp. 398~401.

Kalyani Joshi, N. Bharathi & Rao Jyothi(2016), "Stock trend prediction using news sentiment analysis", *International Journal of Computer Science & Informati on Technology*, 2016(8-3), pp. 67~77.

Kamal, I. M. et al.(2019), "Forecasting high-dimensional multivariate regression of Baltic Dry Index (BDI) using Deep Neural Networks (DNN)", *ICIC Express Letters,* 13(5), pp. 427~434.

Miner, G.(2012), "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications", Massachusetts: Academic Press, pp. 151~180.

Ming-Feng, Tsai & Chuan-Ju Wang(2017), "On the risk prediction and analysis of soft information in finance reports", *European Journal of Operational Rese arch,* 257(1), pp. 243~250.

Moraes, R. Valiati & J.F. Gaviao Neto(2013), Document-level sentiment classificatio n: An empirical comparison between SVM and ANN, Expert systems with applications, v.40 no.2, pp. 621 – 633.

Mostafa, M.(2004), "Forecasting the Suez Canal traffic: a neural network, analysis", Maritime Policy & Management, 31(2), pp. 139~156.

Elisa Claire Alemán Carreón, Hirofumi Nonaka & Toru Hiraoka(2018), "Relation An alysis between Hotel Review Rating Scores and Sentiment Analysis of Revie ws by Chinese Tourists Visiting Japan", The Japanese Journal of the Institute of Industrial Applications Engineers (JJIIAE), Vol. 6, No. 2. pp. 95~99.

Roy, G. et al.(2018), "Sentiment analysis in organizational work: Towards an ontolog y of people analytics", *Expert System*, 35(5), pp. 1~22.

Shen, S., Jiang, H. & Zhang, T.(2012), "Stock market forecasting using machine learn ing algorithm", *Semantic Scholar*, pp. 1~5.

Stefan Albertijn et al.(2011), "Financing Shipping Companies and Shipping Operati ons: A Risk-Management Perspective", *Journal of applied corporate financ e*, v.23 no.4, pp. 70~82.

Stefano Baccianella, Andrea Esuli & Fabrizio Sebastiani(2010), "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", *European Language Resources Association*, pp. 2200~2204.

Tronvoll, B(2011), "Negative emotions and their effect on customer complaint beha vior", *Journal of Service Management*, 22, pp. 111~134.

Tsai, F. M. & Huang, L. J.(2017), "Using artificial neural networks to predict containe r flows between the major ports of Asia", *International Journal of Production Research*, 55(17), pp. 5001~5010.

Yancong Xie & Hongxun Jiang(2017), "Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method", *J.Comp*, 12, pp. 500~510.

Yanran Wu a. et al.(2018), "Optimistic bias of analysts' earnings forecasts: Does inve stor sentiment matter in China?", *Pacific-Basin Finance Journal*, 49, pp. 14 7~163.

Yoosin Kim, Mingon Kang & Seung Ryul Jeong(2018), "Text Mining and Sentiment Analysis for Predicting Box Office Success", *KSII TRANSACTIONS ON INTE RNET AND INFORMATION SYSTEMS*, 12(8), pp. 4090~4102.

William H. Teukolsky, Saul A. Vetterling & William T(2007), "Numerical Recipes: The Art of Scientific Computing", 3th ed., Cambridge University Press, pp.8 90~899.

Wu, Lan et al.(2019), "Prediction for the Region Disposition of Panama Dry Bulk Fleet Management", *IEEE access: practical research*, open solutions Info v. 7, pp. 136604~136615.

Zhou, Yusheng. et al.(2021), "Sustainability disclosure for container shipping: A tex t-mining approach", *Transport policy*, 110, pp. 465~477.

인터넷 자료

국립중앙과학관 빅데이터, https://terms.naver.com/list.naver?cid=58370&categoryId=58370/(검색일: 2021. 9. 12.)

드류리, https://www.drewry.co.uk/news-events(검색일: 2021. 8. 22.)

로이즈리스트, https://lloydslist.com(검색일: 2021. 8. 22.)

발틱거래소, https://www.balticexchange.com(검색일: 2021. 8. 22.)

삼성소프트웨어아카데미, https://lab.ssafy.com(검색일: 2021. 10. 23.)

알파라이너, https://www.alphaliner.com/(검색일: 2021. 8. 22.)

장삼용 머신러닝 인터넷 학습자료(2016),「초보자를 위한 파이썬 200제」, http://blog.naver.com/PostView.nhn?blogId=samsjang&logNo=220969601609&categoryNo=78&parentCategoryNo=49&viewDate=&currentPage=1&postListTopCurrentPage=1&from=postView(검색일: 2021. 8. 13.)

클락슨, https://sin.clarksons.net/Timeseries(검색일: 2021. 10. 23.)

트레이드윈즈, https://www.tradewinds.com(검색일: 2021. 10. 23.)

해양수산부 블로그, https://blog.naver.com/koreamof/222482524247(검색일: 2021. 8. 21.)

Data Science 블로그, https://yngie-c.github.io/nlp/2020/07/31/sentiment_analysis/(검색일: 2021. 9. 12.)

elec4(2015. 8. 3),「빅데이터 시장동향과 분석기법」, https://elec4.co.kr/article/articleView.asp?idx=10458(검색일: 2021. 8. 23.)

Fearnleys, https://fearnleys.com(검색일: 2021. 8. 22.)

IBM, https://www.ibm.com/blogs/watson(검색일: 2021. 10. 9.)

JOC, https://www.joc.com/(검색일: 2021. 8. 22.)

O'REILLY,「파이썬 라이브러리를 활용한 머신러닝」, https://kolikim.tistory.com/24(검색일: 2021. 8. 13.)

PR News, https://www.prnews.com(검색일: 2021. 8. 22.)

Ingenieria Industrial,「Techniques and Applications for Sentiment Analysis」, https:

//www.ceine.cl/techniques-and-applications-for-sentiment-analysis/ (검색일: 2021. 9. 12.)

https://yngie-c.github.io/nlp/2020/07/31/sentiment_analysis/(검색일: 2021. 8. 12.)

wikipedia, https://ko.wikipedia.org/wiki/%EC%84%9C%ED%8F%AC%ED%8A%B8_%EB%B2%A1%ED%84%B0_%EB%A8%B8%EC%8B%A (검색일: 2021. 7. 16)

기본연구보고서 발간목록

2021년

01	시민참여형 해안돌봄 정책 도입방안 연구	정지호
02	해양 플라스틱 쓰레기가 선박 운항에 미치는 영향분석 연구	김보람
03	여성어업인 노동정책 방향 연구	홍혜수
04	순환경제 시스템을 활용한 어업폐기물의 자원화 방안 연구	고동훈
05	1conomy 시대, HMR 시장 확대에 따른 수산부문 대응전략 수립 연구	마창모
06	AIS 데이터 기반 해상교통 안전 평가모델 개발 연구	황선일
07	항만 에너지 관리시스템 도입을 위한 로드맵 구축 연구 - 부산항 신항 전기에너지 소비를 중심으로 -	김근섭
08	해양분야 리빙랩 활성화 방안 연구	좌미라
09	빅데이터의 연계·활용을 통한 선박의 배출량 산정체계 고도화 방안 연구	안용성
10	해양공간계획 집행체계 고도화 방안 연구	최희정
11	수산식품산업 푸드테크 환경분석 및 적용전략 연구	이상건
12	AIS데이터기반 해상물동량 추정 연구	황수진
13	항만부문 정책의사결정 지원 빅데이터 플랫폼 구축 방안 연구	이기열
14	항만 컨테이너 물류 프로세스 디지털화 모형 연구	서정용
15	해양수산 분야 국제산업연관표 구축기반 연구	정수빈
16	해양수산 과학기술 정책평가모형 연구 - 해양수산 R&D를 중심으로 -	전형모
17	글로벌 경제위기와 해운산업 대응체계 연구	박성화
18	항만산업 경기진단체계 구축 연구 - 경기동행지수 개발을 중심으로 -	김성아
19	전국 무역항 부두 분류 체계 개선방안 연구	이수영
20	IMO 규제기반 해사산업의 글로벌 지속발전방안 연구 - 新해사산업의 육성 및 지원을 위한 법제화 연구(5차년도) -	박한선
21	수소에너지 거점 구축을 위한 항만의 대응방안 연구	신수용

2020년

01	해양수산업의 지역 간 연관구조 분석	장정인
02	해양수산 분야 기술 대외의존도 분석연구 - 스마트항만을 중심으로 -	전형모
03	지역 해양수산 오픈 플랫폼 구축방안 연구	최지연
04	갯벌 거버넌스 개선방안에 관한 연구	육근형
05	해양환경정책의 능동적 추진을 위한 재원체계 개선 방안 연구	박수진
06	항만 대기환경 관리 표준 및 평가모형 연구	안용성
07	해양수산분야 사회문제해결형 R&D 기반 구축 연구	좌미라
08	해양 유입 하천쓰레기 관리체계 개선방안	이윤정
09	수산식품 품질·안전관리 제도 개선방안 연구	이헌동
10	국제법 변화에 대응한 어선원 안전 및 권리 제고방안 연구	한덕훈
11	스마트 양식 클러스터 추진 방안 수립 연구	이상철
12	해양포유류 보호에 관한 수산업 대응 방안 연구	정명화
13	수산물 수급통계 개선 방안 연구	김수현
14	IMO 온실가스 규제 대응 정책방향 연구	박한선
15	퇴직전문가 활용을 통한 해운업 경쟁력 강화방안 연구	안영균
16	글로벌 선사들의 물류통합화 전략에 대한 국적선사의 대응방안	전형진
17	내항여객운송항로 정책 발전방안 연구	김태일
18	블록체인 기술기반 식품콜드체인 체계 구축 연구	조지성
19	항만자동화 도입 관련 노무 갈등 해소 방안 연구	김찬호
20	스마트항만과 스마트도시 연계 발전 방안 연구	이언경
21	항만의 회복탄력성 측정 모형 구축에 관한 연구	김성기
22	IMO 규제기반 해사산업의 글로벌 지속발전방안 연구 - 新해사산업의경제적파급효과분석연구(4차년도) -	박한선
23	국내 항만연계 산업의 가치사슬 및 공급사슬 연계성 강화방안 - 자동차 산업을 중심으로	신수용

수시연구보고서 발간목록

2021년

01	해양환경산업 육성 방안 연구	한기원
02	선원의 인권 및 근로여건 향상을 위한 선원근로감독관 제도 개선 방안 연구	허성례
03	포스트 코로나 시대를 대비하는 지역 해양축제 활성화 연구	최일선
04	회복탄력성 개념을 적용한 항만 위기관리 시스템 구축 연구	김성기
05	중대재해처벌법 시행에 따른 항만에서의 대응방안 연구	최상균
06	크루즈산업 COVID-19 방역체계 구축방안 연구	황진회
07	한국형 선주사의 최적 운영방안 연구	고병욱
08	어촌형 생활서비스 전달체계 개선 연구	이호림
09	원양산업의 ESG 도입 기초 연구	윤미경
10	수상레저활동 관리체계 개선방안 연구	홍장원
11	연안지역 자연성 회복을 위한 정책 방향_간척지·담수호 중심으로	정지호
12	스마트항만 구축에 따른 탄소저감 효과 연구 - 동력전환을 중심으로 -	김가현

2020년

01	포스트 코로나 19 해양수산 분야 정책방안	박광서
02	생분해성 어구 사용 활성화 방안 연구	심성현
03	해양법 전문인력 양성 방안 연구	박영길
04	무인도서 해양주권 강화와 이용 활성화를 위한 제도 개선방안	정지호
05	지역 해양수산 재정분권 대응방향	황재희
06	데이터 3법 개정에 따른 항만·물류 데이터 활용도 제고 방안	이기열
07	양식 활어 유통 효율화 방안 연구	마창모
08	해양레저관광사업 추진을 위한 제도정비 방안	홍장원
09	해양바이오기업의 규제 정비 방안 연구	최석문
10	비상체제 시 선원의 안전을 위한 선박-항만-항공 이동경로 구축방안	이혜진

일반연구보고서 발간목록

2021년

01	해양수산 정책영향평가를 위한 기초 연구	김주현
02	해양교육의 사회·경제적 가치 평가 연구	이슬기
03	해양수산업 경기진단체계 기초연구 - 해운업을 중심으로	권장한
04	선박 기술진보를 고려한 탄소 배출량 추정 연구	최건우
05	서포트벡터머신(SVM) 기법을 활용한 해운시황 예측 연구	김병주
06	크루즈 여객 수요 전망 모형 구축 연구	안승현
07	디지털 공급사슬 물류정보통합 구축전략 연구(Ⅲ)	서정용
08	수출용 전복의 유통경로 분석 연구	이정필
09	인공지능기반 해상운임 예측 연구(3차년도)	황수진
10	인공신경망모형을 이용한 양식수산물 단수 전망에 관한 연구 - 김 양식을 중심으로 -	천성훈
11	해양수산분야 미래 리스크 발굴 및 파급효과 분석 연구	박광서

2020년

01	지역 해양관광 경쟁력 지수 체계화 연구	최일선
02	AIS 기반 글로벌 선박 배기가스 배출량 분석 연구	강무홍
03	김 중기 수급전망모형「Gim-MFoS」구축 연구	허수진
04	중앙 북극 공해 비규제어업방지협정 이행방안 연구	김민수
05	동해 평화관광구역 조성 방안 연구	윤인주
06	디지털 공급사슬 물류정보통합 구축전략 연구(Ⅱ) - 일반 수출입 컨테이너 정보교환방식 중심 -	이언경
07	스트레스 테스트를 통한 우리나라 해운·조선 기업의 안정성 분석	박성화
08	인공지능기반 해상운임예측 연구	황수진
09	국제물류주선업 실태분석 및 경쟁력 제고 방향 연구	최나영환
10	디지털화에 따른 개별 직업의 대체 가능성 추정 연구	박희대
11	시계열 분석을 통한 해운시장 분석 및 예측 연구	고병욱
12	해양수산분야 글로벌 경제효과 분석모형(KMI-GEM) 시범 구축 연구	임병호
13	해운항만물류 인력양성사업 관리운영 제도 개선 방안 연구	이자연

일반연구 2021-05

서포트벡터머신(SVM) 기법을 활용한 해운시황 예측 연구

인 쇄 2021년 12월 29일
발 행 2021년 12월 31일

발 행 인 김 종 덕
발 행 처 한국해양수산개발원
주 소 49111 부산시 영도구 해양로 301번길 26(동삼동)
연 락 처 051-797-4800 (FAX 051-797-4810)
등 록 1984년 8월 6일 제313-1984-1호

조판·인쇄 갭스앤디 컴퍼니 (051-911-9890)

판매 및 보급: 정부간행물판매센터 Tel: 02-394-0337
정가 6,000원