creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학전문석사 학위 연구보고서

머신러닝 기반의 온라인 미디어 감성분석을 통한 자동차 제조사 주가 해석에 관한 연구

- 미디어 유형 및 제목과 내용의 차이점을 중심으로 -

2020 년 8 월

서울대학교 공학전문대학원

응용공학과 응용공학전공

박 민 수

머신러닝 기반의 온라인 미디어 감성분석을 통한 자동차 제조사 주가 해석에 관한 연구

- 미디어 유형 및 제목과 내용의 차이점을 중심으로 -

지도 교수 윤 성 로

이 프로젝트 리포트를 공학전문석사 학위
연구보고서로 제출함
2020 년 5 월

서울대학교 공학전문대학원
응용공학과 응용공학전공
박 민 수

박민수의 공학전문석사 학위 연구보고서를 인준함
2020 년 6 월

위 원 장 _____ 한 훈 _____ (인)

위 원 _____ 윤 성 로 _____ (인)

위 원 _____ 구 윤 모 _____ (인)

국문초록

화폐라는 가치 측정의 매개를 통하기 훨씬 이전부터 인류는 등가교환이라는 대전제 아래 다양한 유형 또는 무형의 재화들을 거래해왔으며, 이로부터 적정한 가치의 측정은 경제활동에 있어 필수요소가 되어왔다. 이러한 시장의 요구에 따라 다양한 재화에 대한 가치평가 방법들이 연구되고 있으며, 특히 1602년 네덜란드의 암스테르담에 세계 최초로 증권거래소가 세워진 이후로 기업의 가치평가 즉, 주식의 가치평가 분야는 폭발적 시장규모의 성장과 함께 가장 활발히 연구가 진행되고 있는 분야이다. 그러나 전통적인 시장경제하에서 발달하여온 주식 가치평가 방법의 대부분은 수치적 분석에 집중되어 왔으며, 이러한 수치적 해석이 많은 부분에 있어 적절하고 우수한 해석 결과를 제시하고 있음에도 불구하고 전통적 방법론이 다루지 못하는 정성적 데이터에 대한 부분은 수치적 해석의 한계로 지적되고 있다.

본 연구에서는 최근 급격하게 발달하고 있는 소프트웨어와 하드웨어 기술에 힘입어 다양하게 제시되고 있는 머신러닝 기반의 감성 분석(Sentiment Analysis) 방법을 활용하여 정성적 데이터를 수치화하고 이를 주가 해석에 활용하여 더 나은 해석 결과를 얻기 위한 방안을 제시하고자 한다. 이를 위하여 글로벌 자동차 제조사 중 6개사를 선정하여 2019년 한 해 동안 발생한 온라인 미디어 데이터를 키워드 기반으로 수집하였으며, 보다 세분화된 해석 특성을 파악하기 위하여 미디어 데이터를 전문매체, 일반매체, 사용자 그룹 및 기타의 4개 카테고리로 분리하고 각각의 미디어를 제목과 그 내용으로 구분하여 분석을 진행하였다. 수집된 데이터는 데이터 피쳐링(Data Featuring)을 통하여 구성을 파악하고 기본적인 전처리 과정을 거쳐 주가 예측에 대한 경향성을 파악하기 위하여 회귀 경로와 분류 경로로 나누어 선행 분석을 진행하였으며 회귀 경로에서는 Linear Regression과 Random Forest Regressor를 사용하여 기본이 되는 주식가격의 움직임 대비 RMSE(Root Mean Square Error)를 측정하는 방식으로, 분류 경로에서는 Logistic Regression과 Random Forest Classifier로 기존 주식 가격의 Null Model 대비 예측 정확도를 확인하는 방식으로 각각의 경향성을 파악하고자 하였다. 이후 실제적인 주가 해석을 위하여 VAR(Vector Autoregression) 모델을 이용한 예측 모델을 구축하고 각 미디어 유형별 그리고 제목과 그 내용을 나누어 예측 성능을 평가한 후

각각의 그룹이 가지는 해석 특성을 탐지하고자 하였다.

본 연구는 감성 분석을 통하여 주식시장의 움직임을 거시적으로 접근한 것이 아니라 산업 분야를 특정하고 해당 산업 분야에 속한 대표기업들의 미디어 데이터를 유형별로 구분하여 분석을 진행하였다는 점, 더 나아가 제목과 내용의 차이점을 함께 분석함으로써 향후 감성 분석을 통한 시장 해석 연구에 있어 데이터 선정의 방향성과 데이터 범주화의 필요성을 제시하였다.

주요어 : 주가해석; 감성분석; 자동차 제조사; 온라인 미디어 분류; 머신러닝
학 번 : 2018-22612

목 차

제 1 장 서론 1
제 1 절 연구 동기 1
제 2 절 연구 내용 3
제 3 절 연구 보고서 구성 5

제 2 장 배경 지식과 관련 연구 6
제 1 절 감성분석 6
2.1.1 감성분석의 유형 7
2.1.2 감성분석 방법론과 원리 7
제 2 절 시계열 예측 11
2.2.1 시계열 데이터의 구성요소 11
2.2.2 시계열 분석 방법론 12
2.2.3 벡터 자기 회귀(Vector Autoregression) 모형 12

제 3 장 데이터 분석 15
제 1 절 데이터 수집 및 전처리 15
3.1.1 대상 선정 및 데이터 수집 15
3.1.2 데이터 전처리(1) 16
3.1.2.1 중복 데이터 제거 16
3.1.2.2 주제 연관성 결여 데이터 분리 16
3.1.2.3 미디어 유형 구분 17
제 2 절 데이터 기본특성 분석 19
3.2.1 미디어 유형별 구성 19
3.2.2 데이터 발생빈도 특성 20
제 3 절 감성분석 22
3.3.1 데이터 범주화와 감성분석 프로세스 22
3.3.2 감성분석 결과 해석 23
3.3.2.1 제목과 내용, 미디어 유형에 따른 특성 23
3.3.2.2 Polarity vs. Scoring 28

제 4 장 주가 해석과 예측 31
제 1 절 분석 프로세스 31
제 2 절 경향성 분석과 결과 33
4.2.1 회귀 경로 34

4.2.2 분류 경로 36
4.2.3 상관 관계 38
제 3 절 예측성능 분석과 결과..... 40
4.3.1 데이터 전처리(2)..... 40
4.3.1.1 데이터 정상성(Data Stationary)..... 40
4.3.1.2 변수(Variables)..... 43
4.3.1.3 데이터 평활화(Data Smoothing) 44
4.3.2 미디어 유형 및 제목과 내용에 따른 주가예측..... 46

제 5 장 결론.....51

참고문헌.....53

부록 (Appendix).....56

Abstract.....67

표 목차

[표 3-1] 구글 트렌드를 활용한 제조사별 연관 키워드 선정 ..17
[표 3-2] 데이터 전처리 키워드 및 전후 비교.....18
[표 3-3] 제조사별 미디어 유형 분포 20
[표 3-4] 제조사 및 미디어 유형별 데이터 발생 빈도 상위 분포와 해당 시점의 이슈 사항 21
[표 3-5] 제목 및 내용에 따른 감성분석 결과 분포.....24
[표 3-6] 미디어 유형별 내용의 감성분석 결과 분포 26
[표 3-7] 미디어 유형별 제목의 감성분석 결과 분포 27
[표 3-8] 감성분석 결과 데이터 간 상관관계 및 표준편차 (FORD 社).....30
[표 4-1] 회귀 경로를 통한 교차 검증 결과.....35
[표 4-2] 분류 경로를 통한 교차 검증 결과.....37
[표 4-3] 데이터 정상화 방법과 VAR 모델 오차 성능.....42
[표 4-4] 제조사별 변수의 정상성과 ADF 검정결과.....42
[표 4-5] 추가 변수와 VAR 모델 오차 성능 44
[표 4-6] 데이터 평활화 기법과 VAR 모델 오차 성능.....46
[표 4-7] 예측 모형 실행 조건.....46
[표 4-8] 주식가격에 대한 VAR 모형 추정치 (FORD 社).....47
[표 4-9] 데이터 그룹별 주가예측 오차 성능 48

그림 목차

[그림 1-1] 글로벌 대표 자동차 제조사(6 개사) 및 회사개요 3
[그림 1-2] 감성분석을 통한 주가 해석 연구 흐름도 4
[그림 2-1] 머신러닝 기반 감성분석 기법의 학습 및 예측 프로세스.....9
[그림 2-2] 감성분석 접근방법의 유형.....10
[그림 2-3] 시계열 데이터 구성요소.....11
[그림 3-1] 수집 데이터 예시 15
[그림 3-2] 제조사별 미디어 유형 분포 그래프.....19
[그림 3-3] 제조사별 미디어 발생 시계열 데이터 20
[그림 3-4] 미디어 발생 빈도 상위 시점과 이벤트 (TESLA 社) 21
[그림 3-5] 데이터 범주화와 감성분석 프로세스 22
[그림 3-6] 제목 및 내용에 따른 감성분석 결과 분포.....24
[그림 3-7] 미디어 유형별 내용의 감성분석 결과 분포.....26
[그림 3-8] 미디어 유형별 제목의 감성분석 결과 분포.....27
[그림 3-9] Polarity/Scoring 방법에 따른 시계열 감성 분포 (FORD 社).....29
[그림 4-1] 주가 예측 모델 프로세스 32
[그림 4-2] 5-Fold 교차 검증 모식도.....34
[그림 4-3] 상관관계 분포 (FORD 社) 39
[그림 4-4] 주가 데이터 변환에 따른 정상성 (FORD 社) 41
[그림 4-5] 감성분석 데이터의 평활화 (FORD 社).....45
[그림 4-6] 제조사별 주가예측 그래프.....50

제 1 장 서 론

제 1 절 연구 동기

효율적 시장 가설(EMH: Efficient Market Hypothesis)에 따르면 주식시장은 정보의 반영에 있어 개별주식뿐만 아니라 주식시장 전체에 있어서도 극도로 효율적이기에 새로운 정보는 매우 빠른 속도로 확산되고 즉각적으로 주가에 반영된다. 따라서 과거의 주가를 통해 미래 가격을 예측하는 기술적 분석이나 재무 정보를 활용하여 저평가된 주식을 찾고자 하는 펀더멘탈 분석과 같은 기법으로는 시장수익을 상회하는 초과수익을 올릴 수 없다. 이와 연관하여 주가의 움직임을 설명하는 또 다른 이론으로 랜덤워크이론(RWT: Random Walk Theory)을 들 수 있는데, 이에 따르면 현재 주가의 움직임은 과거의 가격과는 무관하게 움직인다. 즉, 정보가 즉각적으로 주가에 반영된다고 할 때 미래의 주가는 미래의 정보만을 반영하여 움직이기 때문에 현재의 가격변동과는 무관하며, 새로운 정보는 본질적으로 예측할 수 없기에 그에 따른 주가의 변화 역시 예측할 수 없고 무작위적이라는 것이다.[1] 하지만 이와 같은 전통적 이론들이 이야기하고 있는 시장의 효율성과 예측 불가능성의 한계점은 이들이 전제하고 있는 주가 예측의 방법론들(기술적 분석, 펀더멘털분석 그리고 과거의 데이터에 기반한 예측)이 모두 정량적 데이터(Quantitative Data)에 기반하고 있으며, 정성적 데이터(Qualitative Data)에 대해서는 어떠한 활용이나 보완 장치를 보여주지 못하고 있다는 것이다. 이는 마치 인류의 절반을 구성하고 있는 남성 또는 여성 중 한쪽을 배제하고 다른 한쪽이 가지는 특성만을 가지고 인류를 정의하는 것과 같은 한계점을 지닌다.

그러나 시간의 흐름에 따른 기술의 발전은 이러한 양상을 바꾸어 놓고 있다. 최근 온라인 미디어의 발달에 따라 급격하게 증가하고 있는 데이터의 양과 이와 맞물려 빠른 속도로 발전하고 있는 소프트웨어와 하드웨어 기술은 우리가 이전에 생각하지 못했던 방식으로 다루지 못했던 데이터를 처리할 수 있게 해 주었다. 그중에서도 머신러닝(Machine Leaning)을 이용하여 인간의 언어를 컴퓨터가 이해할 수 있도록 해주는 자연어 처리(NLP: Natural Language Processing)는 그 대표적인 분야라 할 수 있다.[2] 이미 많은 연구에서 이러한 자연어

처리 방법을 활용하여 소셜네트워크서비스(SNS: Social Network Service)와 같은 온라인 미디어에서 나타나는 사회적 분위기나 대중의 심리가 주식시장의 움직임에 명백한 영향을 미치고 있음을 보여주고 있으며[3][4], 이는 기존의 정량적 데이터에 기반한 분석 모델의 한계점이라 여겨졌던 수치로 해석하기 어려운 주식의 이상 움직임에 대한 이론적 근거를 제시함으로써 이를 보완할 수 있는 방법으로 주목받고 있다.

이처럼 폭발적으로 증가하는 온라인 미디어는 이전에도 존재했으나 그 실체를 파악하기 어려웠던 심리와 같은 정성적 데이터를 활용 가능하게 해 주었으며, 정성적 데이터의 주식시장 영향력에 대한 다양한 연구들은 이러한 데이터의 활용에 그 의미와 중요성을 더하고 있다. 특히 이러한 정성적 데이터에 의한 영향은 B2B[i] 산업과 비교해 B2C[ii] 산업에서 더 크게 나타나는데 그 중 자동차 산업은 대표적인 B2C 산업의 하나로 소비심리가 미치는 영향이 크고 직접적인 분야라 할 수 있겠다. 이에 본 연구에서는 2019 년간 발생한 온라인 미디어 데이터를 자동차산업, 그중에서도 대표 자동차 제조사[iii]에 한정하여 수집하였으며, 미디어 데이터는 그 성격에 따라 4 가지 유형[iv]으로 구분하였다. 이렇게 수집된 온라인 미디어 데이터는 대표적인 회귀모델(Regression Model)인 Linear Regression 과 Random Forest Regressor, 그리고 분류모델(Classification Model)인 Logistics Regression 과 Random Forest Classifier 를 사용하여 1 차 분석을 진행함으로써 그 경향성을 알아본 후 VAR(Vector Auto-Regression) 기법을 사용한 주가 예측모델을 구축하여 실제 제조사별 주가의 예측 성능을 비교함으로써 미디어 유형별 해석 특성을 알아보고자 한다. 한 걸음 더 나아가 미디어의 제목과 내용을 분리하여 추가적인 분석을 진행함으로써 내용과 제목 간의 차이점에 대해서 탐색해보고자 한다.

---

[i] B2B: Business to Business.

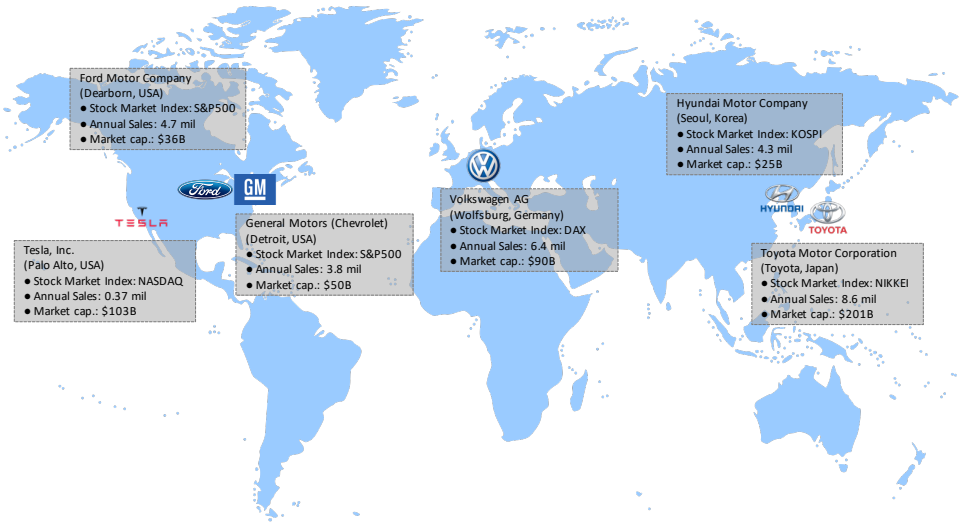
[ii] B2C: Business to Consumer.

[iii] 글로벌 자동차제조사 중 지역 및 제조 차량의 특성을 고려하여 대표성을 가지는 6개사 선별 - Volkswagen(독), Toyota(일), GM(미), Ford(미), Tesla(미), Hyundai Motor(한).

[iv] 전문매체(Professional Media), 일반매체(General Media), 사용자 그룹(Consumer Group), 기타매체(Others)

제 2 절 연구 내용

본 연구에서는 글로벌 자동차 제조사 중 지역 특성과 제조 차량 특성을 반영한 대표 제조사 6 개사를 선정하고 각 제조사와 관련하여 제조사 이름과 주요 모델을 키워드로 하여 2019 년 1 월 1 일부터 12 월 31 일 사이 1 년간 발생한 온라인 미디어 컨텐츠에 대하여 제목, 내용, 작성 시점 및 URL[v]주소 등의 수집을 진행하였으며 수집데이터는 감성분석의 용이성과 데이터의 객관성을 고려하여 영어로 작성된 미디어 데이터로 한정하였다. 이후 머신러닝 기반의 감성분석을 통하여 제조사별 주가와의 상관관계를 알아보고자 하였다.



[그림 1-1] 글로벌 대표 자동차 제조사(6 개사) 및 회사개요

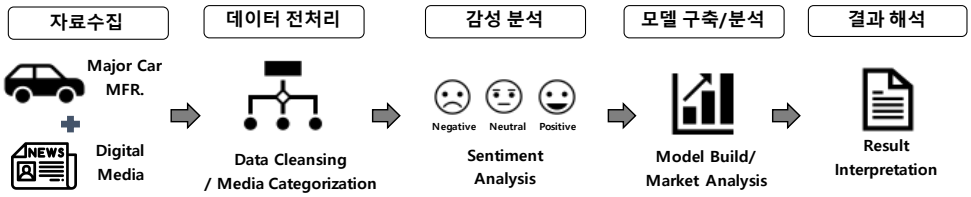
이후 수집 자료의 정합성을 높이기 위하여 키워드를 사용해 주제 연관성이 떨어져 감성분석 결과에 왜곡을 일으킬 수 있는 데이터를 분리하였고 수집된 데이터는 그 출처의 성격에 따라 4 가지 유형으로 구분하는 전 처리를 진행하였다. 전 처리된 데이터는 파이썬(Python)의 대표적인 자연어 처리 API[vi]인 Textblob 의 Sentiment Analyzer 를

---

[v] Uniform Resource Locator
[vi] Application Program Interface

이용하여 감성분석을 진행하였으며, 감성분석은 제목과 내용을 구분하여 진행함으로써 각각의 항목이 가지는 특성과 차이를 같이 확인하고자 하였다. 감성분석의 결과는 주가와의 Time Cycle 을 맞추기 위해 일 단위 평균값을 산출한 후 머신러닝 기반의 회귀, 분류 기법을 활용하여 주가와의 오차와 정확성을 일차적으로 분석함으로써 그 경향성을 찾고자 하였으며, 이후 VAR 기법을 이용한 예측모델을 통하여 예측값과 실제 값과의 오차를 확인함으로써 보다 실제적인 미디어의 특성에 따른 예측 성능을 탐색하고자 하였다.

본 연구에서는 데이터의 수집, 전처리, 감성분석, 분석 모델 구축 및 성능 검증 그리고 이의 결과 해석이라는 과정을 통하여 온라인 미디어에 나타난 심리가 실물 시장에 미치는 영향을 알아보고자 하였으며, 미디어 유형 및 제목과 내용을 구분하여 접근함으로써 각각의 미디어 유형이 가지는 특성을 확인할 수 있었다.



[그림 1-2] 감성분석을 통한 주가 해석 연구 흐름도

제 3 절 연구 보고서 구성

본 연구보고서는 총 5 개의 장으로 구성되어 있으며, 각 장에서 다루고 있는 내용은 다음과 같다.

1 장에서는 온라인 미디어의 감성 분석을 통하여 자동차 제조사의 주가와의 관계를 해석하고자 하는 연구 동기와 연구의 개략적인 구성에 대하여 살펴본다.

2 장에서는 본 연구의 수행을 위해 필요한 배경지식 및 기술에 대하여 알아본다.

3 장에서는 분석데이터의 기본적인 특성을 알아보고 해석 결과 향상을 위한 전처리 방법과 분석 절차에 관해 설명한다.

4 장에서는 예측 모델을 구축하고 이의 성능을 검증하며, 성능을 향상시킬 수 있는 방법을 알아본 후 미디어 유형별, 그리고 제목과 내용 간의 특성 분석을 수행한다.

5 장에서는 본 연구의 결과를 분석, 정리하고 후속 연구에 미칠 수 있는 영향과 의미에 대하여 논의한다.

제 2 장 배경 지식과 관련 연구

제 1 절 감성분석(Sentiment Analysis)

감성분석이란 '의견, 감정 또는 평가와 같은 것들을 긍정과 부정으로 구분하는 일'이라 할 수 있다.[5] 조금 더 구체적으로 말하자면 구절, 문장, 단락 또는 문서와 같이 텍스트 데이터 안에 녹아있는 감정의 극성(긍정, 부정 또는 중립적인)을 어떠한 분석 기술을 통하여 분류하거나 해석하는 일을 의미한다. 최근 온라인 미디어의 급속한 발달로 많은 사람들이 이를 통하여 그들의 느낌과 생각을 자연스럽게 표출하게 됨으로써 비지니스적인 측면에서 이들의 의견을 수집하고 분석하는 것인 매우 중요한 요소로 여겨지고 있다. 미국에서 실시된 조사에 따르면 81%의 인터넷 유저가 최소한 한 번 이상 제품에 대해 온라인으로 찾아본 경험이 있으며 나머지 20%의 사람들은 일반적으로 그렇게 한다고 답했다. 또한 온라인상의 리뷰를 읽는 독자들 중의 73~87%의 사람은 이것이 구매에 중요한 영향을 미쳤다고 답했으며, 이들은 컨슈머 리포트(Consumer Report)의 5-star-rated item 의 경우 4-star-rated item 보다 20~99%[vii]의 비용을 더 지불할 의사가 있다고 답했다. 그리고 32%의 사람들은 제품이나 서비스에 대한 등급을 남겼고 30%의 사람들은 제품에 대한 코멘트나 리뷰를 남긴다고 답했다. [6] 이렇듯 온라인상에서 다양한 형태로 표출되는 소비자 또는 잠재적 소비자들의 감성은 제품의 구매에 직접적이고 중대한 영향을 미치고 있으며 이는 결국 그 회사의 가치(주가)에도 직, 간접적인 영향을 주고 있다고 할 수 있다.

본 연구에서는 이러한 감성분석이라는 도구를 활용하여 온라인 미디어상에 나타난 자동차 제조사들의 텍스트 데이터에 나타나고 있는 감성을 분석하고 이를 이용하여 이러한 감성들의 형태가 해당 제조사의 주가[viii]에 미치는 영향을 알아봄으로써 두 지표와의 관계와 예측을 위한 지표로써 활용방안을 알아보고자 하였다.

[vii] 추가 지불의 비율(%)은 어떤 종류의 제품 또는 서비스인가에 따라 달라진다.

[viii] 주가는 해당 기업의 현황과 미래가치를 가장 즉각적으로 반영하고 있는 지표로 받아들여지고 있음.

2.1.1. 감성분석의 유형

감성분석이 어떻게 적용되고 작동되는지에 대하여 이해하기 위해서는 우선 이의 유형에 대한 이해가 필요하며, 일반적으로 감성분석은 다음과 같이 4 개의 유형으로 분류된다.

첫 번째 대표적인 유형인 Fine-grained Sentiment Analysis 는 텍스트 데이터의 극성(Polarity)을 분류하는 것으로 단순하게 긍정 또는 부정으로 구분하거나 조금 더 세분화된 형태인 매우 긍정, 긍정, 중립, 부정, 매우 부정과 같이 5 단계로 나타낼 수도 있으며, 경우에 따라 본 연구에서 사용한 방법과 같이 -1 에서 +1 사이의 숫자로 그 극성의 정도를 수치화 하여 표현해주기도 한다. 두 번째 유형인 Emotion Detection 의 경우 행복, 좌절, 분노, 슬픔 등과 같은 구체적인 감정의 유형을 찾아내는 것을 목표로 하며 이러한 감정을 감지하기 위하여 어휘 목록인 Lexicon[ix] 또는 머신러닝 알고리즘을 사용한다. 세 번째 유형은 Aspect-based Sentiment Analysis 로 보통의 감성분석이 텍스트 데이터 자체에 대하여 긍정, 부정, 중립과 같은 감성에 대하여 말해준다면 이 유형은 한 단계 더 들어가 특정 요소에 대한 의견을 확인한다. 예를 들어 자동차의 경우 단순 자동차에 대한 감성이 아닌 자동차의 실내 디자인이나 외형과 같이 부분을 특정하여 감성을 나타내어 주기 때문에 제품이나 서비스가 사용자 측면에서 어떻게 인식되는지 또는 강점과 약점은 무엇인지를 파악할 수 있어 제품 분석에 널리 활용되고 있다. 마지막 유형은 Intent Analysis 로 이는 모든 행동에 관한 것이다. 이 유형의 목적은 텍스트 데이터 안에 나타나 있는 메시지의 의도를 판단하는 것으로 비지니스에서는 일반적으로 고객지원 시스템의 작업 흐름 능률화에 주로 사용된다.

2.1.2. 감성분석 방법론과 원리

감성분석은 매우 다양한 자연어 처리 방법들과 알고리즘을 사용하는데 이 분야는 지금도 여전히 기존 방법의 성능을 개선하거나 새로운 방법론을 찾는 연구가 활발히 이루어지고 있다. 이러한 방법론은

[ix] 일반적으로 단어의 의미나 관계 등의 정보가 저장되어 있는 목록이나 저장소를 일컫는 말이나 감성분석에서는 이의 분석을 위하여 단어와 감성(긍정/부정 또는 여러 유형의 감정들)과의 관계가 사전 정의된 어휘 목록을 말한다.

사용 알고리즘(Algorithm)에 따라 머신러닝 방법론과 Lexicon 기반 접근 방법으로 대표되는 규칙 기반 접근법(RBA: Rule Based Approach) 방법론 그리고 이 두 가지 방법론을 혼용하는 하이브리드 방법론으로 나눌 수 있다.

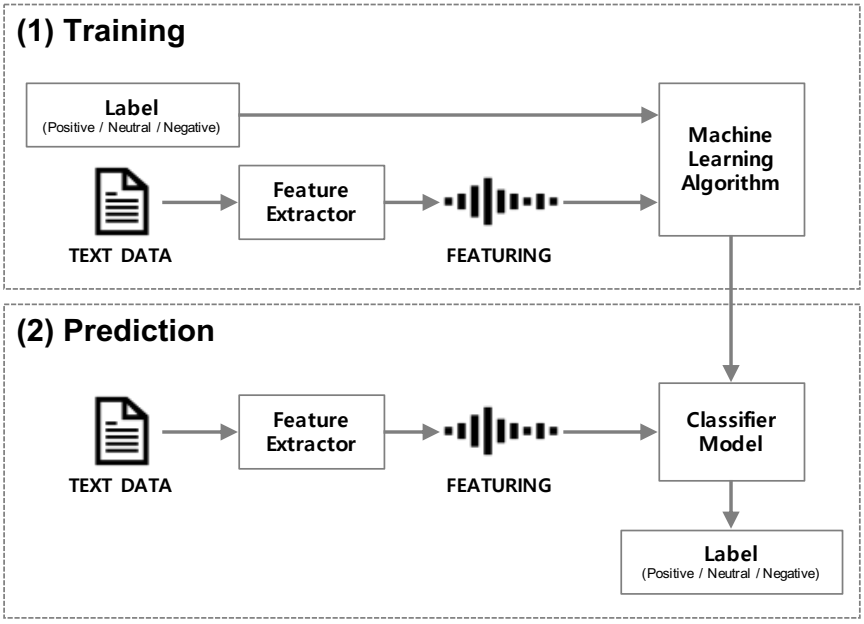
우선 머신러닝 기법은 기본적으로 모델을 구현하기 위하여 어떠한 입력값을 이용하여 알고리즘을 학습(Train)시키는 방법을 말하며 이는 학습하는 동안의 감독 형태에 따라 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 구분되는데 감성분석에서 주로 사용되는 지도 학습은 알고리즘에 주입하는 훈련 데이터에 레이블(Label)이라고 하는 답이 포함되며, 비지도 학습에는 훈련데이터에 레이블이 존재하지 않기에 시스템은 아무런 도움 없이 학습을 해야 한다.[7] 지도 학습은 일반적으로 분류(Classification)로 모델링 되며, 분류자(Classifier)는 레이블과 텍스트 데이터의 특성 데이터를 입력받아 감성분석 결괏값을 반환하는데 여기에서 텍스트 데이터의 특성(Feature)은 특성 추출기(Feature Extractor)를 사용하여 얻어진다. 이러한 특성 추출에는 전통적으로 빈도(Frequency)를 기반으로 하는 Bag of Words[x]나 Bag of n-grams[xi]의 방법을 사용하였으나 최근에는 단어 임베딩(Word Embedding) [xii]을 기반으로 하는 특성 추출 기법이 많이 사용되어 유사한 의미를 가진 단어들이 유사한 표현을 할 수 있게 하여 분류자(Classifier) 모델의 성능을 향상시킬 수 있다.

---

[x] Bag of Words란 텍스트 데이터를 분류하기 위한 방법 중 하나로 단어들의 순서를 고려하지 않고 출현 빈도에만 기반하는 수치화 표현법이다.

[xi] Bag of n-grams 기법은 연속한 n개의 단어를 하나의 분석 단위로 묶어 빈도를 기반으로 텍스트 데이터를 분석하는 기법이다.

[xii] Word Embedding은 텍스트를 구성하고 있는 각 단어를 개별적으로 수치화하는 방법으로 하나의 단어에 인덱스 정수를 할당하는 Bag of Words 방법과 달리 실수 벡터로 나타내기 때문에 각 단어의 벡터를 합치거나 더하는 방식으로 전체의 벡터를 표현할 수 있다.



[그림 2-1] 머신러닝 기반 감성분석 기법의 학습 및 예측 프로세스

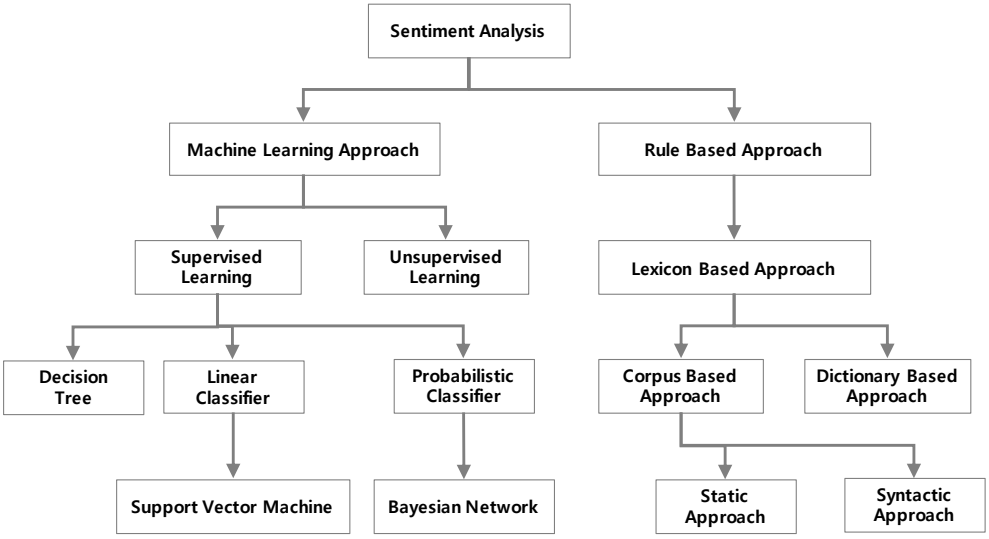
지도 학습의 프로세스 중 분류단계에서는 그 기법에 따라 Decision Tree, Support Vector Machines, Naïve Bayes 와 같은 통계적 모델들이 사용된다.[8]

규칙 기반 접근법의 경우 만들어진 규칙(답)을 가지고 분석이 이루어진다는 측면에서 머신러닝 기법의 지도학습법의 범주로 분류되기도 하나 본 연구에서의 규칙 기반 접근법은 어휘 목록(Lexicon)과 같이 사람이 만든 규칙 기반의 비지도 기법으로 분류하여 정의하였다. 이 접근 방식은 적은 양의 학습데이터로도 동작할 수 있으며 긍정적인 의견은 원하는 상태를, 부정적인 의견은 원하지 않는 상태를 다루는데 가난, 나쁨, 더러움과 같은 단어는 부정적인 의견을 나타내고, 부유한, 좋은, 아름다움과 같은 단어는 긍정적인 감성을 나타내는 것과 같은 방식이다. 이러한 의견을 나타내는 어휘 목록을 수집하는 방법에는 Manual Approach, Dictionary-based Approach[xiii] 그리고 Corpus-based Approach[xiv]가 있다.[9]

[xiii] Dictionary-based 접근법의 경우 텍스트가 전달하는 느낌을 측정하기위한 계산적 접근법으로 사전에 정의된 단어의 감성에 주로 의존하게 된다.

[xiv] Corpus-based 접근법은 텍스트로부터 추상적 규칙의 집합을 이끌어내는 데이터 기반 접근법으로 말뭉치의 동시 발생 확률과 구문적 패턴이 활용된다.[10]

본 연구에서 감성분석의 도구로 사용하고 있는 파이썬 API 인 Textblob 은 감성분석 기능을 가지고 있는 자연어 처리 라이브러리로, 패턴 라이브러리에 기반한 PatternAnalyzer 와 영화 감상평 말뭉치가 학습된 NLTK Classifier 인 NaiveBayesAnalyzer 의 두 가지 방법으로 감성을 분석을 실행한다.



[그림 2-2] 감성분석 접근방법의 유형

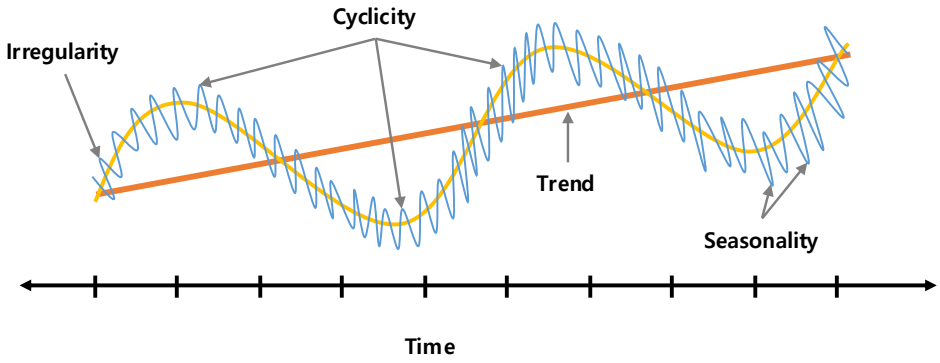
제 2 절 시계열 예측(Time Series Forecasting)

시간의 순서대로 배열된 일련의 관측값의 집합을 우리는 시계열(Time Series)이라고 부르며, 이러한 시계열은 다양한 분야에서 폭넓게 연구되고 있는데 특히 경제학 분야에서 시계열분석은 가격의 예측이나 이를 해석하는 가장 중요한 도구 중 하나로 인식되고 있다. [11] 시계열 분석에서는 과거에 관측된 데이터들을 분석하여 그 속에서 규칙적인 패턴을 찾고 이를 수식화 하여 만든 모델을 바탕으로 예측을 수행하는데, 보통 시계열 데이터는 그 속성에 따라 나타나는 양상이 다양하기 때문에 모든 데이터에 사용할 수 있는 규칙이나 모델은 있을 수 없으므로 각 데이터의 특성에 가장 적합한 모델을 적용할 필요가 있다.[12]

본 연구에서는 앞서 언급한 감성분석의 결과를 발생 시점을 기준으로 시계열 데이터화하고 이를 각 자동차 제조사의 주가와의 관계를 시계열 분석 방법을 통하여 알아보고자 한다. 이후 벡터 자기회귀(VAR) 모형을 바탕으로 한 주가 예측모델을 구축하여 각 데이터 그룹별 예측 성능을 비교하고 그 특성을 탐색하고자 하였다.

2.2.1. 시계열 데이터의 구성요소

시계열 데이터는 그 규칙성의 유무에 따라 Pattern 요소와 Irregularity(Random) 요소로 나누어지며 Pattern 요소의 경우 그 규칙성의 특성에 따라 다시 장기간에 걸쳐 데이터 전체의 증가 또는 감소를 나타내는 Trend, 데이터의 장기적 진동을 나타내는 Cyclicity 그리고 계절성 요인으로 인하여 단기적으로 나타나는 패턴인 Seasonality 로 구분할 수 있다.



[그림 2-3] 시계열 데이터 구성요소

2.2.2. 시계열 분석 방법론

시계열 분석에는 매우 다양한 종류의 방법들이 사용되고 있으며 적절한 모델의 선택이 분석 결과에 결정적인 영향을 미칠 수 있으므로 분석하고자 하는 시계열 데이터의 특성에 맞게 적절한 모델을 선정하는 것이 중요하다. 여러 접근법들 중에서 예측을 목적으로 주로 사용되는 방법에는 추세분석(Trend analysis), 평활법(Smoothing Method), 분해법(Decomposition Method), ARIMA(Auto-Regressive Integrated Moving Average) 그리고 ARCH(Auto-Regressive Conditional Heteroskedasticity) 등의 방법이 있으며, 그 외에도 적용 분야(Domain)에 따라 Spectral Analysis 나 Wavelet Analysis 와 같은 진동수 영역(Frequency Domain)에서의 분석법이나 Auto-correlation Function 과 같은 시간 영역(Time Domain)으로 구분되는 방법들도 있다.

이 중 1976 년 Box 와 Jenkins 에 의해 소개된 ARIMA 모형의 경우 과거의 특정한 규칙성이 현재 또는 미래에도 동일하게 유지된다는 가정하에 예측하는 방법으로 설정이 용이하고 기존 예측 모형들이 가지는 시간에 대한 경직성이나 주관성을 극복할 수 있다는 점에서 주목받았으나 다른 변수들과의 관계를 고려하지 않고 단변량(Univariate)분석이라는 한계를 보이게 된다.[13] 이러한 기존 시계열 분석 방법의 단점을 보완한 방법이 1980 년 Sims 가 고안한 VAR 모형인 것이다. [14]

2.2.3. 벡터 자기 회귀(Vector Autoregression) 모형

본 연구에서는 감성분석의 결과를 이용한 자동차 제조사 주가예측 모델 구축에 있어 AR 모형의 다변량(Multivariate) 형태로도 볼 수 있는 VAR 모형을 사용하였는데 VAR 모형은 연립 방정식 체계와 비슷하나 충격 반응 분석(Impulse Response Analysis)[xv]을 통해 다른 특정 변수가 내생 변수에 미치는 동태적 효과를 파악할 수 있다는 점, 그리고 분산 분해(Variance Composition) [xvi]을 통해 각 내생 변수의 변화 중에서 각 변수가 전체 변동에 영향을 끼친 부분의 상대적 효과를

[xv] 모형 내의 내생 변수들이 오차항에서의 단위 충격들에 대한 반응성을 추적함
[xvi] 모형 내의 내생 변수들에 미친 상대적 영향에 대한 정보를 제공함

분석할 수 있다는 분석상의 특징을 가진다.[14] 더불어 이와 같은 다변량 시계열 분석은 변수 그룹을 자체 과거 값과 다른 변수의 과거 값의 조합으로 나타냄으로써 변수 그룹 간의 상호작용 및 구성에 대한 통찰력을 제공하여 준다. VAR 모형에 의한 접근 방법은 모형 내의 모든 변수를 내생 변수[xvii]로 취급하기 때문에 다른 시계열 데이터와의 관계에 제약이나 임의 배제 없이 일반적인 형태로 정형화하여 동태적 관계를 이용하기 때문에 예측력을 높일 수 있어 시계열 분석 분야에서 폭넓게 사용된다. 수학적인 식으로 표현하면 다음과 같다.[15]

$$R_t = A \cdot R_{t-1} + \ldots + A' \cdot R_{t-k} + e_t$$

또는

$$\begin{bmatrix} R_t^1 \\ R_t^2 \\ R_t^3 \\ \ldots \\ R_t^n \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \ldots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \ldots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \ldots & A_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ A_{n1} & A_{n2} & A_{n3} & \ldots & A_{nn} \end{bmatrix} \begin{bmatrix} R_{t-1}^1 \\ R_{t-1}^2 \\ R_{t-1}^3 \\ \ldots \\ R_{t-1}^n \end{bmatrix} + \cdots + \begin{bmatrix} A'_{11} & A'_{12} & A'_{13} & \ldots & A'_{1n} \\ A'_{21} & A'_{22} & A'_{23} & \ldots & A'_{2n} \\ A'_{31} & A'_{32} & A'_{33} & \ldots & A'_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ A'_{n1} & A'_{n2} & A'_{n3} & \ldots & A'_{nn} \end{bmatrix} \begin{bmatrix} R_{t-k}^1 \\ R_{t-k}^2 \\ R_{t-k}^3 \\ \ldots \\ R_{t-k}^n \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ \ldots \\ e_{nt} \end{bmatrix}$$

n: 모형 내 고려되는 변수들의 수

k: 모형내 고려되는 시차들의 수 (VAR(k))

R_t, R_{t-1}, R_{t-k} *: 1 × n 변수 벡터*

A…A': n × n 계수 매트릭스

e_t*: 1 × n 혁신 벡터*

VAR 모형의 추정을 위해서는 오차항인 백색 잡음(White Noise)을 발생하기 위해 필요한 시차(Time Lag)를 결정해야 하는데 이러한 시차를 결정하는 기준으로는 Akaike(AIC), Hannan-Quin(HQIC) 그리고 Schwarz(Bayesian) 등이 있다.[16]

VAR 모형은 앞서 언급한 바와 같이 모형 내 모든 변수를 내생 변수로 간주하기 때문에 어떤 변수가 내생 변수이고 어떤 변수가 외생 변수[xviii]인지 구분할 필요가 없기에 모형이 단순해지며, 일반적으로

[xvii] 내생 변수란 모형 안에서 그 값이 결정되는 변수로 인과관계에서 결과가 되는 변수

[xviii] 외생 변수란 모형 밖에서 정해진 변수로 인과관계에서 원인이 되는 변수

복잡한 연립 방정식 형태의 모형에 비해 많은 경우보다 뛰어난 예측력을 보여준다는 장점이 있는 반면 모형을 설정하는 데 있어 다른 변수들과의 관계에 사전 규정을 하지 않기 때문에 무 이론적(a-theoretic)이라는 점과 구조적인 방정식의 형태가 아니기 때문에 개별 변수들의 예측값에 적절한 의미를 부여하기 힘들다는 점 등이 단점으로 꼽히고 있다.
그러나 이러한 단점에도 불구하고 VAR 모형은 모델의 단순성과 우수한 예측성능으로 다양한 분야에서 시계열 예측 모델로 사용되고 있으며 앞서 지적된 단점에 대해서도 그레인저 인과관계(Granger Causality)[xix]를 통해 변수 간의 인과관계를 파악하거나 충격 반응 분석을 통해 정책효과를 분석하는 방법 등을 통해 이를 극복하고자 하는 시도가 진행되고 있다.

[xix] 내생 변수란 모형 안에서 그 값이 결정되는 변수로 인과관계에서 결과가 되는 변수

제 3 장 데이터 분석

제 1 절 데이터 수집 및 전처리

3.1.1. 대상 선정 및 데이터 수집

본 연구에 사용된 데이터는 2019 년 1 월 1 일부터 2019 년 12 월 31 일까지 1 년 동안 웹(Web)상에 영어로 게재된 온라인 미디어 컨텐츠를 대상으로 글로벌 시장에서 대표성을 가지는 자동차 제조사 6 개사에 대하여 각 제조사와 관련된 키워드들을 기준으로 수집을 진행하였으며 수집된 자료는 제목(Title), 내용(Content), 해당 미디어 컨텐츠의 출처(URL) 및 게시 시간(Post Date) 등에 대한 정보를 포함한다. 6 개의 자동차 제조사는 글로벌 시장 점유율 상위 10 개사 중 지역 대표성을 가지는 제조사를 기본으로 수집 데이터의 언어 기반인 영어권 국가의 제조사를 추가하였으며 최근 자동차 업계에서 가장 주목받고 있는 전기차 제조사, 끝으로 국내 자동차 제조사를 포함하여 총 6 개 제조사를 선정하였다.[xx] 1 차로 총 121,237 개의 온라인 미디어 컨텐츠를 수집하였으며, 기본적인 전 처리를 통하여 최종적으로 64,703 개의 데이터가 본 연구의 분석에 사용되었다.

Title	Content	Site URL	Posting Date and Time
How to use Map App on Corolla 2020 SE?	I have an SE Corolla 2020, and when I select the Map button it says "Navigation app	https://www.reddit.com/r/Toyota/commen	2019.12.31 4:23
Prius Prime heater question	The 2020 Prime owner's manual states on page 510, "To prevent 12-volt battery disc	https://priuschat.com/threads/prius-prime	2019.12.31 3:06
Toyota RAV4 Hybrid Adds Juice To Nameplat	Toyota RAV4 TRD Off-Road version.TOYOTA Toyota just keeps improving its flagsl	https://www.forbes.com/sites/dalebuss/202	2019.12.31 3:05
Toyota RAV4 2019 Video Review	The Toyota RAV4 is regarded by many as the original Sports Utility Vehicle, and ove	https://www.motoring.com.au/toyota-rav4	2019.5.17 0:00
2019 Toyota Tacoma and RAV4 Just Receive	If you are looking to buy a new car, one of your most important factors could be how	https://www.torquenews.com/6626/toyota	2019.5.16 21:51
Japanese Aluminum For Tesla Batteries Excl	Tesla Gigafactory in Nevada gets a little boost as the U.S. Commerce Department ha	https://insideevs.com/news/356447/japane	2019.6.25 8:00
Tesla Semi Test Hints At Towing Capacity: Vi	One of the new videos of the Tesla Semi spotted on the road provides us with an opp	https://insideevs.com/news/356428/tesla-s	2019.6.25 7:56
New electric Ford Mustang with Tesla-inspire	Ford Mustang electric (Image: CHARGE) An all-electric Ford Mustang is set to debut	https://www.express.co.uk/life-style/cars/11	2019.6.25 6:45
Model 3 AC only works on "LO" temperature s	So I took my 6 months old model 3 MR that has (had and later updated to 4.1) 2019.2	https://teslamotorsclub.com/tmc/threads/r	2019.6.25 5:56
Tesla Model 3: A Tesla For The Masses, But	When Tesla first came to market, it did so with an expensive, two-seat electric sports	https://insideevs.com/reviews/356544/vide	2019.6.25 5:18
Tesla Model 3 Races Through Boring Compar	Sometimes boring is a good thing. Some may say this video is pretty boring, but we t	https://insideevs.com/news/356275/video-	2019.6.25 5:10
Tesla Gets Tariff Exemption For Japan-Sourc	Tesla has been granted a tariff exemption on aluminum it imports from Japan that is u	https://www.carscoops.com/2019/06/tesla-	2019.6.25 3:00
Fisker Electric SUV New Teaser Shows Turn	We also get to see the pop-out door handles. When you think of an electric SUV borr	https://www.motor1.com/news/354717/fisk	2019.6.14 16:21
I think I just saw a prototype for the redesigne	Maybe it has been a long day and I am just imagining things. Or maybe there is a mat	https://teslamotorsclub.com/tmc/threads/i	2019.6.14 15:43
sell my 2019 m3.	Please advise how to sell my new model 3. It's less than a month old. Great car, but	https://teslamotorsclub.com/tmc/threads/s	2019.6.14 15:02
New Easter Eggs Games Coming to Tesla	Tesla already has a few games in the cars right now. Asteroids, 2048, Missile Comm	https://www.torquenews.com/7908/new-ea	2019.6.14 14:55
Tesla is delivering 1,000 cars a day en route t	Tesla is trying to achieve record deliveries this quarter. It is counting on the North Am	https://electrek.co/2019/06/14/tesla-delive	2019.6.14 13:29
TeslaCam Captures The Sudden Fury And De	Last year Tesla made the Model S, Model X and Model 3 front cameras available for	https://insideevs.com/news/354707/tesla-c	2019.6.14 12:48
Next-gen Ford Ecosport Based On Ford Fiest	The Ecosport is among the best selling vehicles for Ford in India. The subcompact S	https://www.cartoq.com/ford-ecosport-ren	2019.2.25 0:00
Maruti Baleno Recalled In India — Free Servic	Get the Best Drivespark Stories Subscribe Now notifications 2019 Ford Figo Facelift	https://www.drivespark.com/four-wheelers	2019.2.25 0:00
X1 Racing E-Sports League Announced — A	Get the Best Drivespark Stories Subscribe Now notifications 2019 Ford Figo Facelift	https://www.drivespark.com/four-wheelers	2019.2.25 0:00
Vauxhall Corsa: 2019 supermini spied with pro	Vauxhall's all-new Corsa has been spotted testing with less disguise, a few weeks af	https://www.autocar.co.uk/car-news/indus	2019.2.25 0:00
Ford Kuga Vignale review	There is no doubt that the continuing popularity of premium SUVs has prompted Ford	https://www.carbuyer.co.uk/news/167418/f	2019.2.25 0:00
Ford probing possible problems with fuel econ	Ford Motor stated on Thursday it has employed outside experts to probe its vehicle fu	https://speedlux.com/ford-probing-possib	2019.2.25 0:00
...	...	...	...

[그림 3-1] 수집 데이터 예시

[xx] (1) 지역기준: 유럽-Volkswagen, 아시아-Toyota, 북미-GM
(2) 영어권 국자 제조사 추가 – Ford
(3) 전기차제조사 – Tesla
(4) 국내제조사 – Hyundai

3.1.2. 데이터 전처리(1)

수집된 데이터를 별도의 처리 과정을 거치지 않고 그대로 사용할 경우 데이터 세트가 가지는 잠재적 위험이 분석 결과의 왜곡 등을 야기할 수 있기에 분석에 적합한 형태로 Data Cleansing 작업이 필요하다.[17] 본 연구에서는 다음과 같이 3 단계에 걸쳐 기본적인 전처리 작업을 진행하였다.

3.1.2.1 중복 데이터 제거

온라인 미디어 컨텐츠의 경우 디지털 미디어가 가지는 근본적인 속성에 따라 재생산이 매우 용이하고 이에 따라 그 확산 속도가 매우 빠르기에[18] 동일한 컨텐츠가 중복적으로 만들어져 퍼져나가는 경우가 많다. 이는 본 연구에 사용하기 위하여 키워드를 통하여 수집된 데이터에서도 동일 양상을 보였으며, 이의 중복을 제거하기 위하여 제목을 비교하여 동일 데이터를 찾아서 제거하는 방법[xxi]을 사용하였다. 평균적으로 중복을 통하여 제거되는 데이터의 수는 전체 데이터의 10% 내외이다.

3.1.2.2 주제 연관성 결여 데이터 분리

직접적인 연관성이 떨어지는 데이터는 분석 결과의 왜곡의 원인이 된다. 웹상에서 텍스트 기반의 컨텐츠를 수집하는 기준이나 방법은 다양하지만 본 연구에서는 키워드 기반 자료수집을 진행하였는데 이 방법의 경우 전통적인 수집 방법과 비교해 걸리는 시간이 짧고 수집되는 데이터의 수가 적은 대신 상대적으로 특정 분야에 연관성이 높은 데이터를 수집할 수 있다는 장점이 있다.[19] 그러나 이러한 키워드 기반 수집장식에서도 주제 연관성이 떨어지는 데이터들이 포함되어 있는데 해당 키워드를 포함하고는 있으나 타 제조사와 관련된 주제에 단순 언급되는 형태이거나 여러 제조사나 브랜드를 단순 비교하는 것이

[xxi] Original 컨텐츠가 재생산 또는 전달된 경우 기본적으로 게시 시간이나 출처(URL) 비교로는 중복 판단이 어려우며, 내용의 경우 그 양이 방대하고 미세한 재생산 과정에서 미세한 차이가 발생할 경우 중복으로 인식되지 않는 경우가 있어 비교가 간단하고 변형의 확률이 적은 제목을 기준으로 하였다.

그러한 경우이다. 수집된 데이터의 수가 충분하였기에 주제와 연관성이 떨어지는 데이터를 제거하는 방법이 아닌 주제 연관성이 높은 데이터를 추출하는 방식으로 전처리를 진행하였으며, 주제 연관성이 높은 데이터를 추출하기 위하여 제조사별 연관 키워드를 기준으로 컨텐츠의 직접적인 주제를 담고 있는 제목이 기준 키워드를 포함하는가 여부에 따라 데이터를 분리하였다. 제조사별 연관 키워드 선정은 구글에서 제공하는 검색어 통계 서비스인 구글 트렌드(Google Trend)를 활용하여 2019 년 1 년간 제조사명을 기준으로 한 상위 25 개의 연관검색어를 추출한 다음 해당 연관검색어가 포함하고 있는 제조사의 자동차 모델명을 분리하여 선정하는 방식을 취하였으며 이를 통하여 총 81 개의 키워드를 선정하였고 제조사별로 7~18 개의 키워드로 선별작업을 진행하였다. 이러한 방식은 일부 주제 연관성이 높은 데이터의 손실이 불가피하나 추출된 데이터에 주제 연관성이 높아진다는 장점이 있다.

TOP	TOYOTA	VOLKSWAGEN	TESLA	FORD	HYUNDAI	GENERAL MOTORS	CHEVROLET
1	toyota corolla	golf volkswagen	tesla model	ford focus	tucson hyundai	gm	chevy
2	toyota 2019	golf	model 3	ford fiesta	tucson	gm motors	silverado
3	camry	polo volkswagen	model 3 tesla	ford ranger	2019 hyundai	gm general motors	chevrolet silverado
4	toyota camry	polo	tesla stock	ford f150	elantra hyundai	ford	chevrolet cruze
5	rav4	vw	tesla truck	ford mustang	santa fe	ford general motors	ford
6	rav4 toyota	volkswagen passat	tesla price	mustang	santa fe hyundai	general motors car	chevrolet camaro
7	toyota yaris	passat	tesla model x	ford explorer	hyundai kona	general motors strike	camaro
8	toyota car	volkswagen jetta	tesla nikola	ford transit	kona	general motors stock	chevrolet car
9	honda	jetta	tesla car	ford car	hyundai car	chevrolet	chevrolet spark
10	toyota tacoma	volkswagen tiguan	model s	tom ford	hyundai accent	general motors company	chevrolet 2018
11	toyota hilux	tiguan	tesla model s	ford fusion	hyundai sonata	general motors chevrolet	chevrolet aveo
12	toyota hybrid	volkswagen car	model x	ford escape	hyundai i20	general motors cars	toyota
13	used toyota	volkswagen beetle	tesla cybertruck	ford ka	accent	general motors news	chevrolet captiva
14	toyota highlander	ford	cybertruck	ford edge	kia	general motors jobs	chevrolet corvette
15	toyota supra	audi	tesla 2019	ford mondeo	hyundai i30	general motor	nissan
16	supra	volkswagen up	tesla roadster	used ford	hyundai i10	general motors careers	gmc
17	toyota cars	toyota	tesla model y	raptor ford	hyundai venue	general electric motors	chevrolet cars
18	nissan	volkswagen cars	model y	ford st	toyota	toyota	equinox chevrolet
19	toyota tundra	volkswagen gti	tesla news	toyota	honda	chevy	onix chevrolet
20	bmw	volkswagen caddy	tesla cost	ford kuga	hyundai creta	gmc	chevrolet onix
21	ford	volkswagen touareg	new tesla	ford truck	ford	general motors canada	chevrolet impala
22	prius toyota	used volkswagen	tesla cars	ford cars	hyundai cars	general electric	blazer chevrolet
23	toyota 4runner	gol volkswagen	used tesla	ford gt	creta	general motors india	blazer
24	toyota chr	nissan	tesla stock price	ford ecosport	auto hyundai	general motors mexico	tahoe chevrolet
25	switch	volkswagen transporter	aktie tesla	ford f250	nissan	general motors michigan	chevrolet colorado
Key Word	toyota, corolla, camry, rav4, yaris, tacoma, hilux, highlander, supra, tundra, prius, 4runner, chr	volkswagen, vw, golf, polo, passat, jetta, tiguan, beetle, up, caddy, touareg, gol, transporter	tesla, model, model 3, model x, model s, cybertruck, model y	ford, focus, fiesta, ranger, f150, mustang, explorer, transit, fusion, escape, ka, edge, mondeo, raptor, kuga, ecosport, f250	hyundai, tucson, elantra, santa fe, kona, accent, sonata, i20, i30, i10, venue, creta, kia	general motors, gm, chevrolet, chevy, GMC	silverado, cruze, camaro, spark, aveo, captiva, corvette, equinox, onix, impala, blazer, tahoe, colorado
No.	13	13	7	17	13	5	13

[표 3-1] 구글 트렌드를 활용한 제조사별 연관 키워드 선정

3.1.2.3 미디어 유형 구분

마지막은 앞서 2 단계에 걸쳐 전 처리된 데이터에 대해 각각의 온라인 미디어 매체가 가지는 성격에 따라 이를 유형별로 구분하고

라벨링(Labeling)을 하는 것으로 본 연구에서는 자동차를 전문으로 다루는 잡지, 분석 기관 또는 채널에 해당하는 전문매체(Professional Media), 일반적인 대중매체로 종합 언론사, 일간지, 뉴스 사 등에 해당하는 일반 매체(General Media), 자동차와 관련된 동호회와 같은 사용자 그룹(Consumer Group) 그리고 앞서 언급한 3 가지 중 어느 유형에도 포함되지 않는 기타 매체(Others) 이렇게 4 가지로 그 유형을 구분하였다. 매체의 구분은 데이터 출처(Site URL)에서 상세 주소와 디렉터리 부분을 제외하고 순수 매체를 나타내는 주소 부분만을 분리하여 해당 매체의 성격을 기준으로 미디어 유형을 구분하였다. 본 연구에 사용된 전체 데이터에서 중복을 제외한 각 매체(주소) 수는 총 569 개로 제조사별로 258~420 개 정도의 개별 매체 소스를 가지는 것으로 나타났으며 이러한 전처리 과정을 통하여 최초 수집된 데이터는 총 121,237 개에서 64,703 개로 약 47% 감소하였다. 그 외에 주가와의 비교단계에서 결측값의 처리 또는 평활화(Smoothing)와 같은 데이터 처리는 분석을 진행하는 과정 중 필요단계에서 적용하였으며 이후 장에서 다루도록 한다.

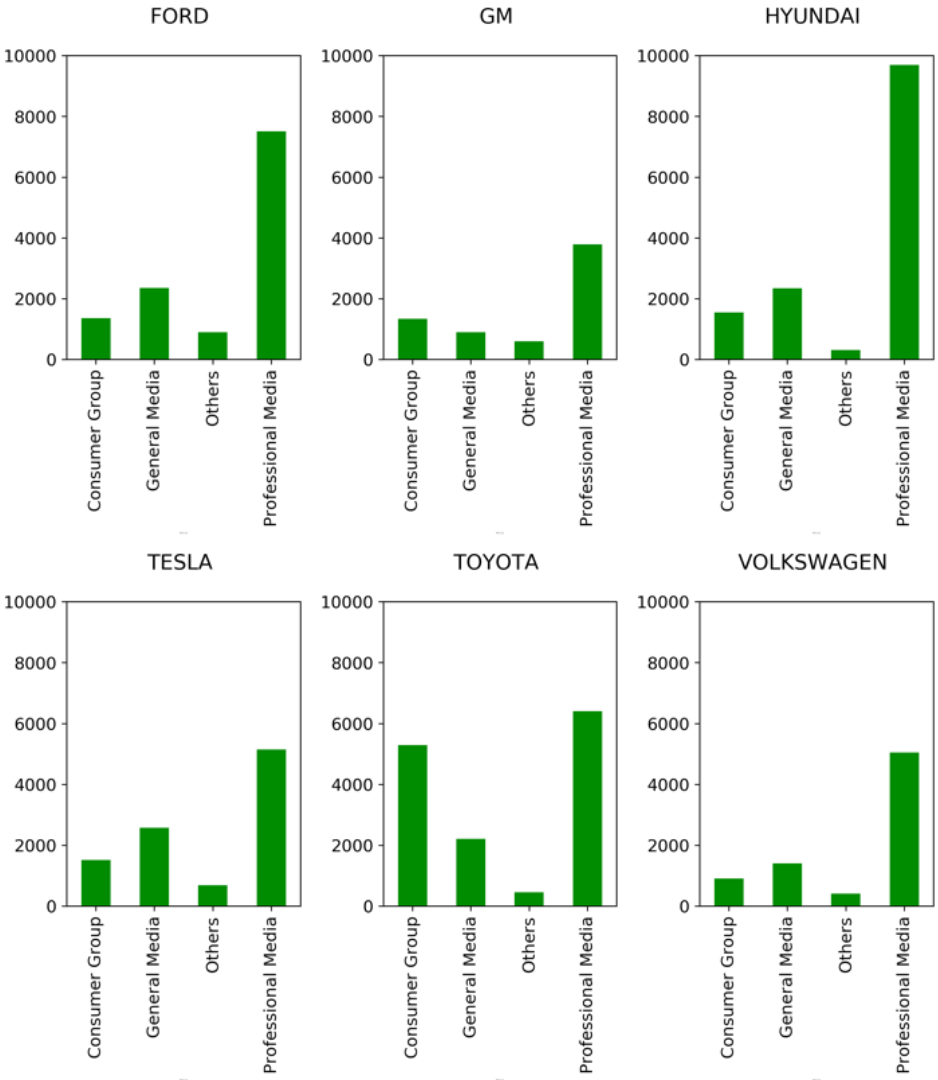
제조사		데이터 수 (수집)	데이터 수 (전처리)	키워드 수	키워드
Volkswagen (1)	세계 최대 자동차 제조 그룹	14,216	7,766	13	VW, Volkswagen, passat, jetta, golf, …
Toyota (2)	단일 브랜드 글로벌 1위	29,434	14,383	13	Toyota, Lexus, Camry, Collora, Yaris, …
General Motors (4)	북미 최대 자동차 제조사	13,105	6,605	18	GM, General Motors, Chevy, Chevrolet, …
Ford (6)	세계최대 전기자 제조업체	22,292	12,127	17	Ford, Mustang, Explorer, F-150, …
Hyundai (5)	글로벌 5위 자동차 제조 그룹	26,695	13,898	13	Hyundai, Kia, Sonata, Elentra, Auzera, …
Tesla (-)	세계최대 전기자 제조업체	15,495	9,924	7	Tesla, Model 3, Model S, Model X, …
*()안 숫자는 2019년 제조사 순위	합계	121,237	64,703	81	

[표 3-2] 데이터 전처리 키워드 및 전후 비교

제 2 절 데이터 기본특성 분석

3.2.1. 미디어 유형별 구성

앞서 기본적인 전 처리가 완료된 데이터 세트의 기본 구성을 살펴보면 미디어 유형별로 전문 매체의 비중이 전체의 절반을 넘는 58.1%를 차지하고 있으며 그 뒤를 이어 사용자 그룹과 일반 매체의 비중이 각각 18.5%, 18.2%로 유사한 비중을 차지하고 있음을 알 수 있다. 이를 통하여 뉴스나 기사 등 기관 매체에서 생성되는 컨텐츠의 양이 전체 데이터의 76.6%로 대부분을 차지하는 것으로 나타났다.



[그림 3-2] 제조사별 미디어 유형 분포 그래프

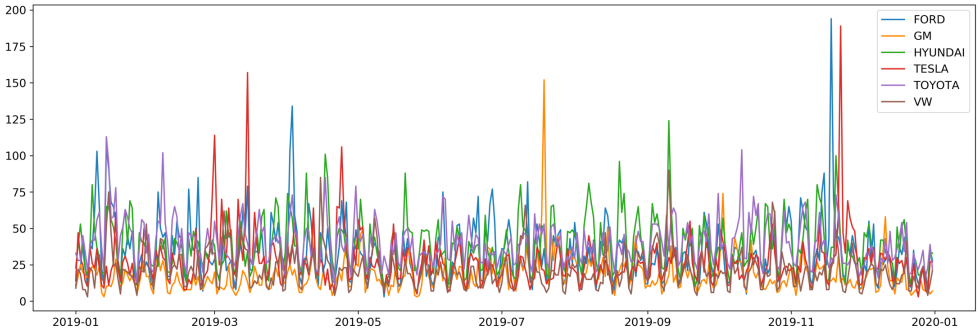
여기에 기타 매체로 분류된 대부분의 미디어 역시 대부분 기관 매체에서 생성되었다는 점으로 미루어 볼 때 그 비중은 81.8%로 더욱 높아진다고 할 수 있겠다.

	FORD		GM		HYUNDAI		TESLA		TOYOTA		VOLKSWAGEN		TOTAL	
Consumer Group	1362	11.2%	1333	20.2%	1546	11.1%	1525	15.4%	5304	36.9%	906	11.7%	11976	18.5%
General Media	2356	19.4%	892	13.5%	2347	16.9%	2570	25.9%	2214	15.4%	1407	18.1%	11786	18.2%
Others	899	7.4%	593	9.0%	310	2.2%	682	6.9%	457	3.2%	412	5.3%	3353	5.2%
Professional Media	7510	61.9%	3787	57.3%	9695	69.8%	5147	51.9%	6408	44.6%	5041	64.9%	37588	58.1%
TOTAL	12127	100.0%	6605	100.0%	13898	100.0%	9924	100.0%	14383	100.0%	7766	100.0%	64703	100.0%

[표 3-3] 제조사별 미디어 유형 분포

3.2.2. 데이터 발생빈도 특성

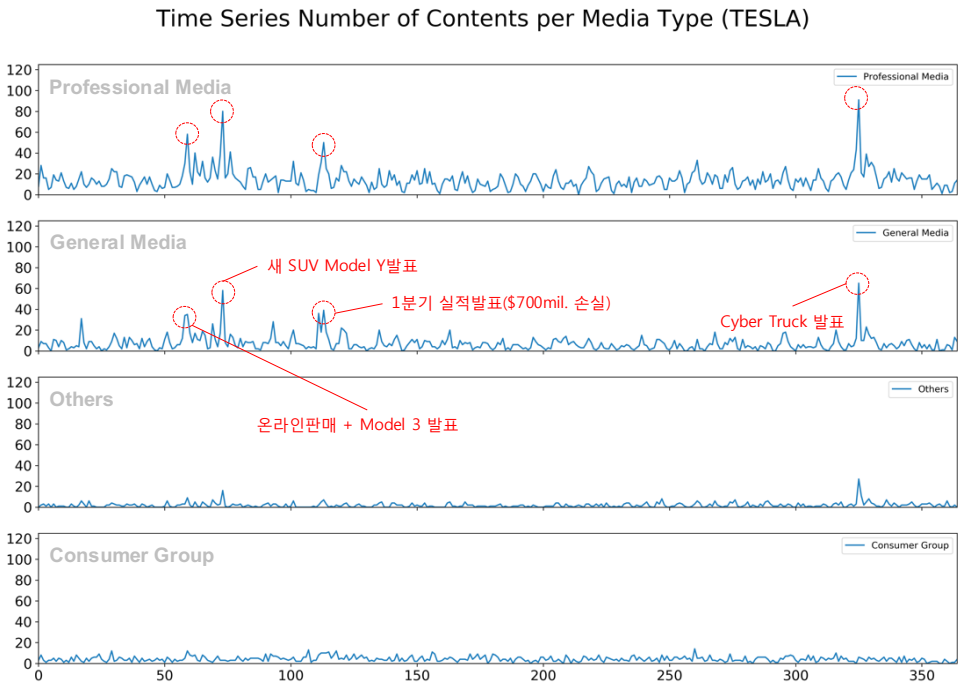
미디어 데이터를 유형별로 분류한 후 이를 시계열 형태로 배열하여 보면 전 제조사에 걸쳐 데이터 유형별로 특정 시점에 발생 빈도가 높게 나타나는 지점들이 발생함을 확인할 수 있는데, 이러한 영상은 전문 매체와 일반 매체에서 그 패턴이 매우 유사하게 나타남을 알 수 있다. 이러한 양상의 발생원인은 다음과 같이 추론하여 볼 수 있는데 기관 매체의 특성상 제조사별로 특정 이벤트가 발행하여 그에 대한 최초 게시물인 씨앗 뉴스가 생성되면 온라인 매체들은 의제 설정에 나서며 급격하게 확산되었거나[20], 신차발표회와 같이 홍보 등을 위하여 특정 시점에 발생한 이벤트의 목적 자체가 각종 매체들을 대상으로 미디어 컨텐츠의 대량 양산에 있는 경우이다.



[그림 3-3] 제조사별 미디어 발생 시계열 데이터

MFR.	Professional Media		General Media		Others		Consumer Group		Issue	
	Date	Count	Date	Count	Date	Count	Date	Count	Date	Issue
FORD	2019-11-18	120	2019-11-18	44	2019-11-18	23	2019-09-26	11	2019-11-18	Electric Mustang SUV Mach-E 공개
	2019-04-03	87	2019-04-03	34	2019-01-10	11	2019-02-05	10	2019-04-03	Ford China 2.0 발표
	2019-01-14	74	2019-07-12	33	2019-04-03	9	2019-03-21	10	2019-01-10	유럽 사업 재편(인원감축안) 발표
	2019-04-02	71	2019-06-06	30	2019-08-22	8	2019-08-07	10		
	2019-01-10	59	2019-01-10	29	2019-10-25	8	2019-05-09	9		
GM	2019-07-19	100	2019-07-19	28	2019-07-19	17	2019-03-29	12	2019-07-19	최초의 Mid-Engine Corvette 공개
	2019-10-03	46	2019-10-03	13	2019-10-03	11	2019-09-11	11	2019-10-03	3분기 실적발표 – 판매 소폭 증가
	2019-07-18	42	2019-04-12	11	2019-12-11	11	2019-05-17	10	2019-04-12	차세대 Corvette 공개 일정 발표
	2019-04-12	39	2019-07-22	11	2019-08-27	10	2019-07-06	10		
	2019-05-02	39	2019-05-03	10	2019-07-22	9	2019-03-14	10		
HYUNDAI	2019-09-10	94	2019-11-12	22	2019-09-10	7	2019-01-25	14	2019-04-17	소형 SUV Vanue 공개
	2019-04-17	75	2019-05-21	21	2019-01-15	7	2019-12-04	13		
	2019-11-20	74	2019-07-09	20	2019-11-20	7	2019-09-03	12		
	2019-08-20	73	2019-04-17	20	2019-01-22	5	2019-10-13	12		
	2019-04-09	66	2019-11-13	19	2019-12-05	5	2019-10-16	12		
TESLA	2019-11-22	91	2019-11-22	64	2019-11-22	22	2019-09-18	14	2019-11-22	사이버트럭 발표
	2019-03-15	79	2019-03-15	58	2019-03-15	16	2019-04-18	13	2019-03-15	새로운 SUV Model Y 발표
	2019-03-01	57	2019-04-24	39	2019-11-23	9	2019-01-30	12	2019-03-01	온라인 판매(오프라인 철수)/Model 3 가격 발표
	2019-04-24	50	2019-04-22	36	2019-03-01	9	2019-03-01	12	2019-04-24	1분기 실적발표 - 대규모 적자(700만불)
	2019-11-21	42	2019-03-01	34	2019-11-26	8	2019-04-29	12		
TOYOTA	2019-02-07	68	2019-10-11	18	2019-10-11	8	2019-01-14	29	2019-10-11	차세대 수소 전기차 Mirai 발표
	2019-01-15	68	2019-01-14	18	2019-11-20	6	2019-06-16	26	2019-01-14	Supra 공개
	2019-10-11	61	2019-03-14	17	2019-06-07	5	2019-02-05	25		
	2019-01-14	61	2019-04-03	17	2019-09-09	5	2019-09-11	24		
	2019-04-30	47	2019-09-27	15	2019-02-07	5	2019-03-27	24		
VOLKS WAGEN	2019-09-10	66	2019-04-15	21	2019-09-11	9	2019-02-27	8	2019-09-10	EV Crossover ID.4 공개
	2019-04-15	57	2019-09-10	20	2019-10-24	7	2019-07-07	8	2019-04-15	Diesel Scandal관련 전임 CEO 연관 기사 보도
	rhks	50	2019-07-12	16	2019-10-11	7	2019-07-10	7	2019-10-24	8세대 Golf 모델 공개
	2019-10-24	49	2019-01-15	16	2019-03-05	7	2019-05-08	7	2019-07-11	Beetle 모델 단종 발표
	2019-07-11	43	2019-07-11	15	2019-09-05	6	2019-10-25	7		

[표 3-4] 제조사 및 미디어 유형별 데이터 발생 빈도 상위 분포와 해당 시점의 이슈 사항

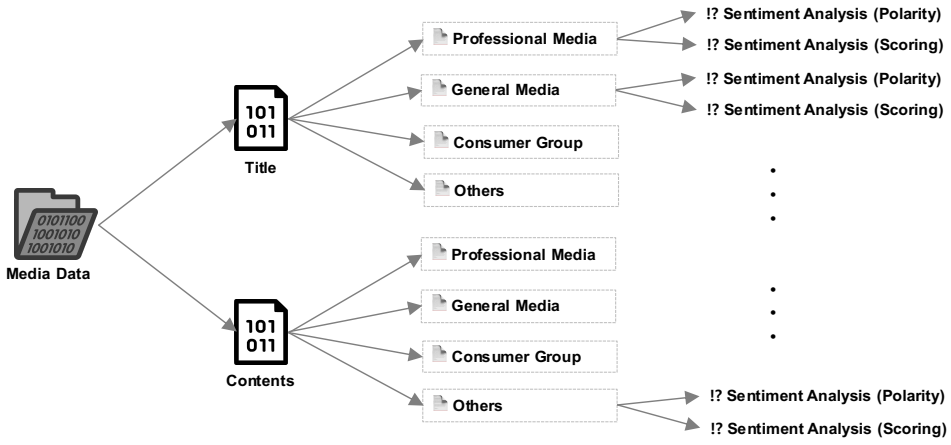


[그림 3-4] 미디어 발생 빈도 상위 시점과 이벤트 (TESLA 社)

제 3 절 감성분석

3.3.1. 데이터 범주화와 감성분석 프로세스

본 연구에서 사용되는 데이터의 경우 앞서 언급한 바와 같이 사전 선정한 글로벌 자동차 제조사 6 개사의 온라인 미디어를 기본으로 한다. 그러나 단순히 제조사별 데이터에 대한 감성분석을 진행하고 결과를 분석하는 것만으로는 차별적인 시사점을 도출하기 어렵기에 여기서는 감성분석 이전에 데이터 범주화(Categorization)를 통하여 데이터를 보다 세부적으로 나누고 각각의 데이터 그룹의 감성 특성과 주가 해석에 미치는 영향이 어떤 차이가 있는지 그리고 한 단계 더 나아가 어떠한 특성의 데이터가 주가 해석에 가장 적합하고 뛰어난 지표로서의 성능을 가지는지와 같은 새로운 시각을 제시하고자 하였다. 데이터 범주화는 다음 2 단계로 진행되게 되는데 1 단계는 제목과 내용의 분리로 함축적이며 간결하게 주제를 나타내는 제목보다 더 세부적이고 자세한 정보를 담고 있는 내용을 분리함으로써 성격이 다른 두 형태의 데이터 간의 차이점을 알아보고자 하였다. 2 단계는 앞장에서 설명한 바 있는 미디어 성격에 따른 4 가지 유형의 분리이다.



[그림 3-5] 데이터 범주화와 감성분석 프로세스

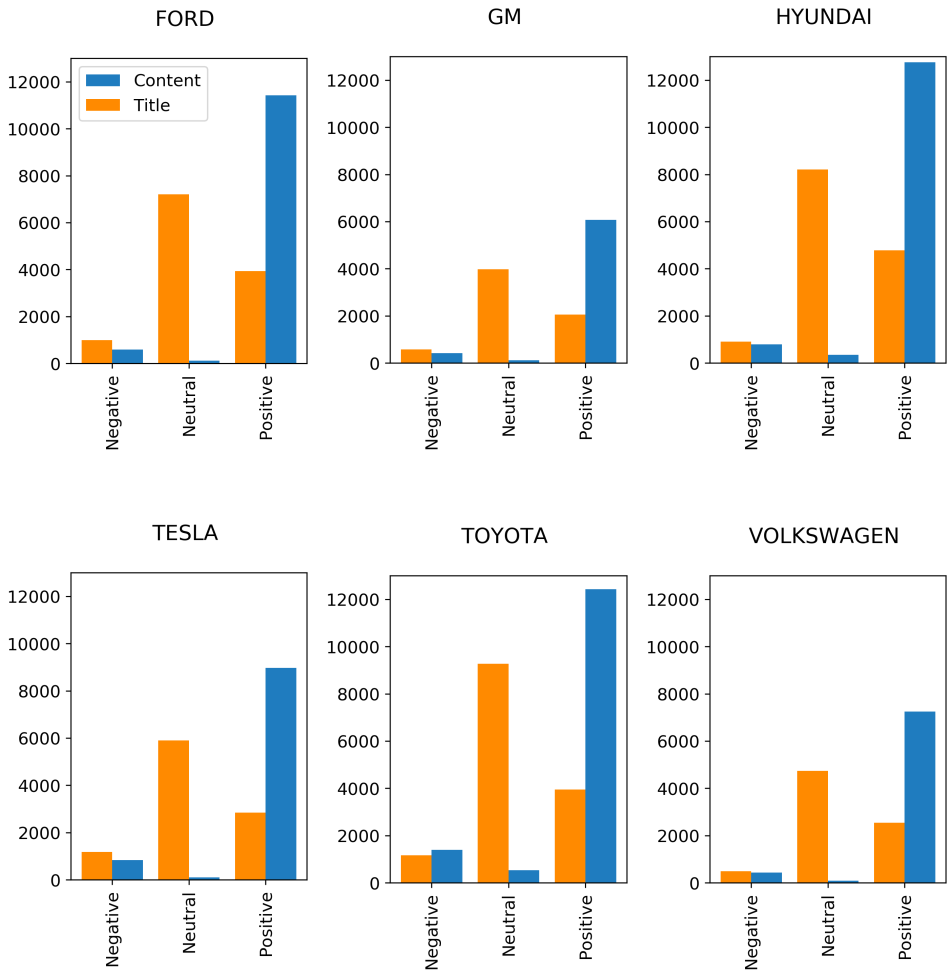
이렇게 범주화를 통하여 8 개로 나누어진 데이터 그룹을 바탕으로 이후 감성분석을 진행하게 된다. 감성분석의 도구로는 앞서 언급하였던 파이썬 자연어 처리 라이브러리인 TextBlob 을 사용하였고 여기에서는 긍정 또는 부정의 감성을 그 정도에 따라 -1 에서+1 사이의 숫자로 그

극성을 나타내어 반환하는데 이 경우 같은 긍정 또는 부정의 감성이라도 그 정도의 차이를 달리하여 보여줌으로써 보다 감성의 깊이를 세밀하게 표현할 수 있다는 장점이 있는 반면 -1 이나 +1 에 가까운 극단적인 감성을 가지는 데이터가 전체 감성의 평균에 미치는 영향이 커져 단위 기간 동안의 감성을 희석시킬 수 있다는 단점이 있다. 따라서 이를 보완하기 위한 방법으로 감성 표현을 긍정, 중립, 부정의 3 가지로만 나누고 구분된 데이터에 각각 +1, 0, -1 의 점수를 부여(Scoring)하는 보다 단순화된 방법으로 표현한 결과를 향후 분석에 사용하였다. 이 방법에서 감성 표현의 구분은 앞서 Polarity 에 의한 방법을 사용하여 도출한 결괏값을 바탕으로 이를 다시 분류하는 방법으로 진행하였는데 Polarity > 0 인 경우 긍정, Polarity = 0 인 경우 중립 그리고 Polarity < 0 인 경우 부정으로 분류하였다. 이러한 경우 극성으로 표현할 때와 같이 세밀한 감정의 깊이를 나타낼 수는 없으나 극단적인 감성에 따른 왜곡을 줄이고 단위 기간 동안 나타난 다수의 감성을 대변할 수 있다는 장점이 있다. 이와 같이 2 가지의 서로 상호보완적이면서도 다른 감성분석 방법을 통하여 각각의 차이를 확인하고자 하였다.

3.3.2. 감성분석 결과 해석

3.3.2.1 제목과 내용, 미디어 유형에 따른 특성

여기에서는 범주화된 데이터를 기준으로 감성분석을 진행하고 그 결과값을 수치화, 시각화 하여 결과의 분포에 대한 특징을 파악하고자 하였으며 각 제조사별 제목 데이터와 내용 데이터를 나누고 각각의 데이터 내에서 미디어 유형을 구분하였을 때 나타나는 감성분포를 확인하고자 하였다. 제목과 내용의 감성분석 결과는 그림 3-6 과 표 3-5 에 나타나 있는데 각 제조사별 분포를 보면 뚜렷한 경향성을 확인 할 수 있다. 내용 데이터의 경우 긍정(Positive) 감성을 나타내는 데이터가 전체 데이터의 91%를 차지하며 압도적으로 높은 비율을 나타내었으며 부정(Negative) 감성이 6.9%, 중립(Neutral) 감성이 2.1%를 차지하며 그 뒤를 이었다. 이러한 감정의 비율과 순서는 제조사 간에 약간의 차이만 있을 뿐 동일한 경향을 띠고 있음을 알 수 있다.



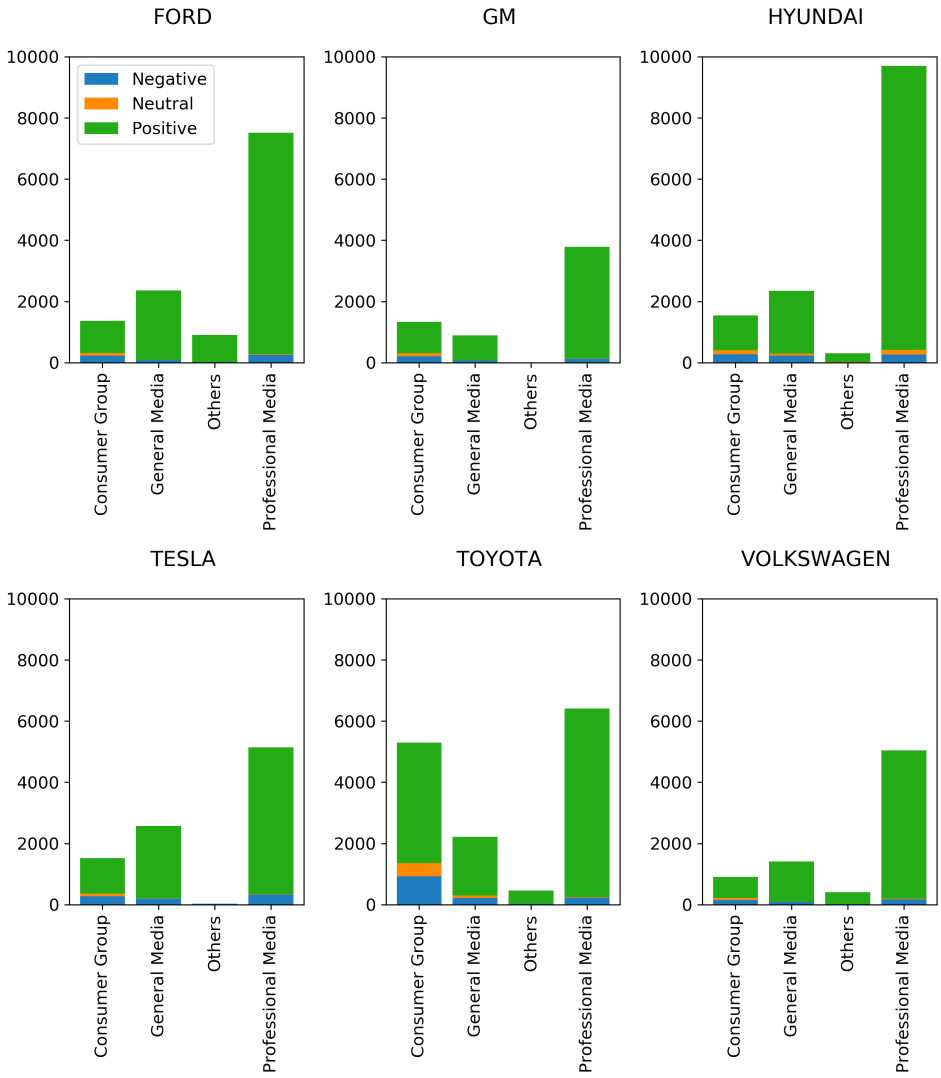
[그림 3-6] 제목 및 내용에 따른 감성분석 결과 분포

	FORD				GM				HYUNDAI			
	Content		Title		Content		Title		Content		Title	
Negative	587	4.8%	994	8.2%	424	6.4%	571	8.6%	788	5.7%	909	6.5%
Neutral	120	1.0%	7199	59.4%	115	1.7%	3972	60.1%	347	2.5%	8215	59.1%
Positive	11420	94.2%	3934	32.4%	6066	91.8%	2062	31.2%	12763	91.8%	4774	34.4%
TOTAL	12127	100.0%	12127	100.0%	6605	100.0%	6605	100.0%	13898	100.0%	13898	100.0%
	TESLA				**TOYOTA**				**VOLKSWAGEN**			
	Content		Title		Content		Title		Content		Title	
Negative	842	8.5%	1178	11.9%	1396	9.7%	1170	8.1%	429	5.5%	484	6.2%
Neutral	110	1.1%	5895	59.4%	553	3.8%	9268	64.4%	93	1.2%	4739	61.0%
Positive	8972	90.4%	2851	28.7%	12434	86.4%	3945	27.4%	7244	93.3%	2543	32.7%
TOTAL	9924	100.0%	9924	100.0%	14383	100.0%	14383	100.0%	7766	100.0%	7766	100.0%

[표 3-5] 제목 및 내용에 따른 감성분석 결과 분포

그러나 이러한 패턴은 제목 데이터에서는 전혀 다르게 나타나는데 중립 감성의 데이터가 전체의 60.7%를 차지하며 가장 높은 비율을 차지하였고 그 뒤로 긍정 감성이 31.1%로 두 번째, 부정 감성은 8.2% 가장 적은 비중을 차지하는 것으로 나타났다. 제조사 간에는 내용 데이터에서와같이 정도의 차이만 있을 뿐 비율과 순서는 동일하게 나타났다.

이러한 감성의 패턴은 미디어 유형별로 범주화된 데이터에서도 동일하게 나타나는데 내용 데이터의 경우 GM 사의 기타 매체 분포를 제외하고는 미디어의 유형과 관계없이 나타나는 압도적으로 높은 긍정 감성의 데이터 비율과 그 뒤를 이어 부정과 중립 감성의 순으로 나타나는 양상을 동일하게 보이는 것을 알 수 있다. 제목 데이터의 경우도 중립 감성의 비중이 가장 높게 나타나고 그다음이 긍정 감성과 부정 감성의 순으로 감소하는 형태가 미디어의 유형과 상관없이 동일하게 나타나고 있으나 VOLKSWAGEN 사의 전문 매체에서만 부정 감성의 비중이 긍정 감성보다 높게 나타났다. 미디어의 유형을 구분해보면 미디어 유형에 따라 나타나는 특성도 확인할 수 있는데 가장 두드러지는 부분은 사용자 그룹 미디어에서 나타나는 것으로 감성 결과의 분포에서 긍정 감성의 비율이 타 미디어 유형인 전문 매체나 일반 매체와 비교하여 상대적으로 낮게 나타난다는 것이다. 이는 표현이 비교적 자유롭고 제약이 없는 사용자 그룹 컨텐츠의 특성을 반영한다고 볼 수 있다

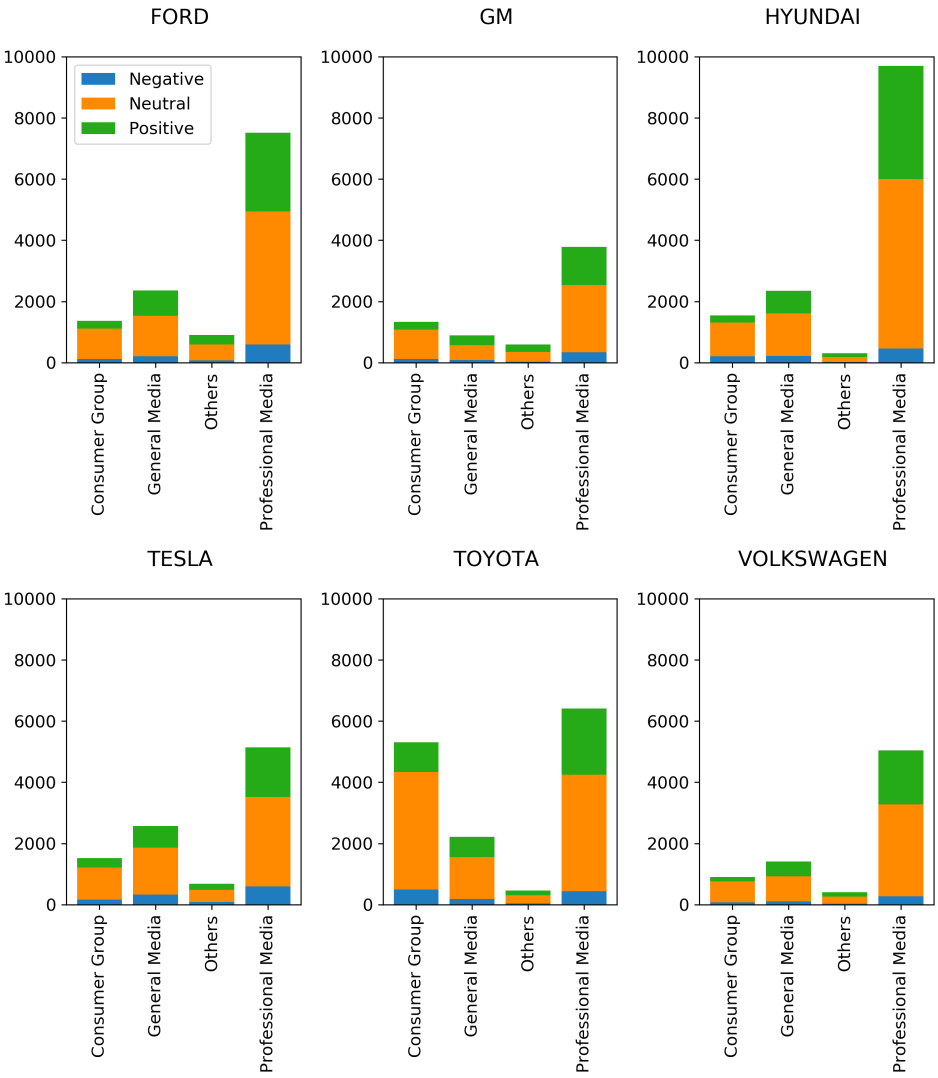


[그림 3-7] 미디어 유형별 내용의 감성분석 결과 분포

	FORD				GM				HYUNDAI			
	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media
Positive	1045	2263	874	7238	1022	819	11	3643	1134	2054	295	9280
Neutral	89	7	2	21	101	2	0	12	133	62	5	147
Negative	228	85	23	251	210	71	582	132	279	231	10	268
TOTAL	1362	2355	899	7510	1333	892	593	3787	1546	2347	310	9695

	TESLA				TOYOTA				VOLKSWAGEN			
	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media
Positive	1155	2361	649	4807	3937	1915	427	6155	684	1316	395	4849
Neutral	85	9	0	10	433	81	4	23	67	8	2	16
Negative	279	200	33	330	922	218	26	230	155	83	15	176
TOTAL	1519	2570	682	5147	5292	2214	457	6408	906	1407	412	5041

[표 3-6] 미디어 유형별 내용의 감성분석 결과 분포



[그림 3-8] 미디어 유형별 제목의 감성분석 결과 분포

	FORD				GM				HYUNDAI			
	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media
Positive	246	820	303	2565	256	316	236	1254	234	732	120	3688
Neutral	999	1332	521	4347	953	494	327	2198	1107	1396	165	5547
Negative	117	204	75	598	124	82	30	335	205	219	25	460
TOTAL	1362	2356	899	7510	1333	892	593	3787	1546	2347	310	9695

	TESLA				TOYOTA				VOLKSWAGEN			
	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media	Consumer Group	General Media	Others	Professional Media
Positive	308	710	202	1631	966	661	155	2163	146	480	153	272
Neutral	1048	1527	396	2924	3843	1362	259	3804	679	823	232	3005
Negative	169	333	84	592	495	191	43	441	81	104	27	1764
TOTAL	1525	2570	682	5147	5304	2214	457	6408	906	1407	412	5041

[표 3-7] 미디어 유형별 제목의 감성분석 결과 분포

본 감성분석을 통해 공통적으로 나타나는 몇 가지 특징을 도출하면 다음과 같은데 첫째는 제목과 내용 그리고 미디어 유형과 관계없이 긍정 감성의 데이터 비중이 부정 감성의 데이터 비중에 비해 높게 나타난다는 것이다. 그 원인은 미디어 유형에 따른 성격과 그 관계 특수성에서 찾을 수 있는데 우선 미디어의 유형 측면에서 앞 장에서 언급한 바와 같이 수집된 데이터의 76.6%가 기관 매체라는 점에 주목할 필요가 있다. 이 중 전문 매체의 비중이 특히 높은데 이러한 전문 매체에서 경우 전체 수입의 절반을 넘는 부분을 광고 수익에 의존하고 있으며[21] 이 광고의 주요 고객은 바로 자동차 제조사이다. 이와 같은 관계 특수성은 기관 매체로 하여금 그들의 광고 판매처에 대해 부정적인 감성의 미디어 컨텐츠를 많이 생산할 수 없는 구조를 만들고 있으며[22] 이로 인하여 긍정 또는 중립 감성이 부정 감성과 비교해 상대적으로 높은 비율을 나타내고 있다고 볼 수 있다. 두 번째는 제목 데이터의 경우 감성 판단이 되지 않는 중립 감성의 데이터 비중이 매우 높게 나타난다는 것인데 그 원인은 두 가지로 추론해 볼 수 있다. 하나는 데이터의 길이이다. 제목의 경우 대부분 하나의 문장 길어야 두 문장을 넘지 않기 때문에 감성 분석을 위한 재료가 충분하지 않을 수 있다. 또 다른 원인은 제목이 가지는 근본적 특성인데 제목은 그 내용이 말하고자 하는 바를 함축적으로 표현하기 때문에 비유법, 은유법, 상징 등의 수사법들이 사용되는 경우가 많고 특히 온라인 미디어의 경우 사용자의 관심을 유발하기 위해 선정적이거나 자극적인 또는 내용과 관련성이 떨어지는 형태로 작성되는 경우가 많기 때문이다.[23][24]

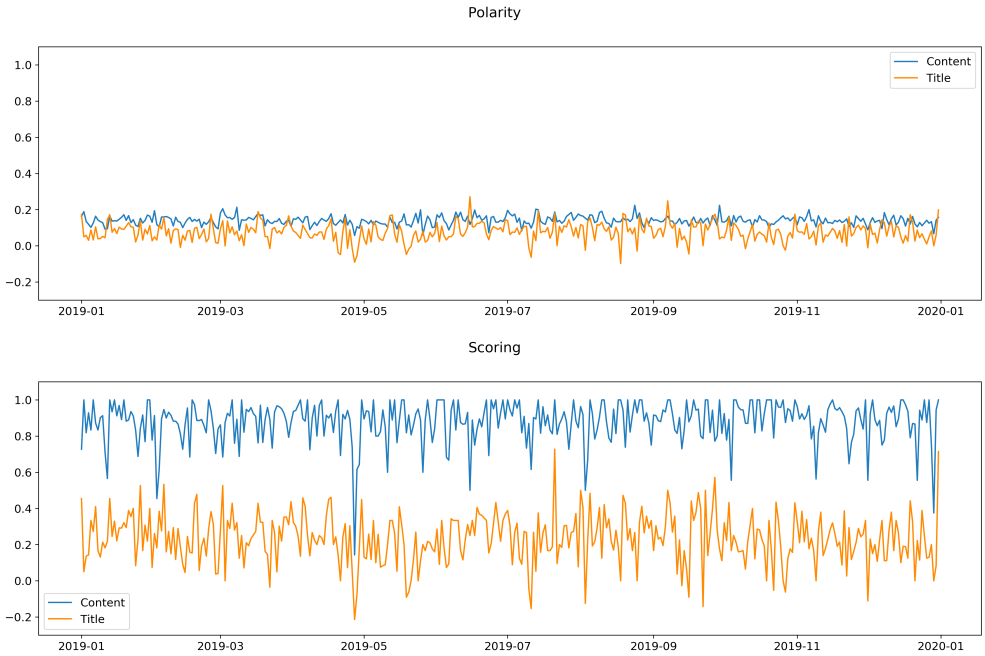
3.3.2.2 Polarity vs. Scoring

그림 3-9 는 감성분석 방법을 Polarity 와 Scoring 으로 달리하였을 때 나타나는 감성의 시계열 분포에 대해 Volkswagen 사의 예를 나타낸 것이다. 수집된 데이터에서는 데이터 간의 시간 간격이나, 분포가 일정하지 않아 단순 시계열로 배열 시 그 편차가 너무 커지는 것을 방지하고 이후 주가와 같이 해석 시 주가 데이터와의 시간 단위를 맞추기 위하여 여기에서는 단위 시간의 간격을 1 일로 정하고 해당 일의 감성 값을 그날 동안 발생한 모든 데이터의 Polarity 와 Scoring 감성 값을 각각의 평균값으로 정하여 사용하였다.

$$S_d = \frac{Y_{1,d} + Y_{2,d} + Y_{3,d} + \ \ldots + Y_{n,d}}{n}$$

S_d: d 일의 감성 값
$Y_{n,d}$: d 일의 n 번째 데이터의 감성분석 값
n: 해당 일의 데이터 총 수량

분포를 살펴보면 Polarity 의 경우 감성의 단위가 매우 세밀하여 각 데이터 간의 편차가 크지 않고 제목과 내용 사이도 그 간격이 크게 나타나지 않지만, Scoring 의 경우 감성 간의 간격이 상대적으로 매우 크기에 편찻값도 크고 제목과 내용 사이의 간격도 그 차이가 뚜렷하게 나타남을 알 수 있다.



[그림 3-9] Polarity/Scoring 방법에 따른 시계열 감성분포 (FORD 社)

추가적으로 각 감성분석 결과 데이터 간의 상관관계를 알아보기 위하여 연속형 자료의 선형적 상관관계를 나타낼 때 주로 사용하는 Pearson Correlation Coefficient 법으로 다음 식과 같이 도출한 결과를 표 3-8 에 각 데이터의 표준편차와 함께 나타내었다.

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{m}(X_i-\mu_X)(Y_i-\mu_Y)}{m}}{\sigma_X \sigma_Y}$$

μ_X: *모집단 X 의 평균*
μ_Y: *모집단 Y 의 평균*
σ_X: *모집단 X 의 표준편차*
σ_Y: *모집단 Y 의 표준편차*
m: *모집단의 개체 수*

Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.450392	0.316759	0.289415	0.025339
	Scoring	0.450392	1	0.180023	0.320906	0.114689
Title	Polarity	0.316759	0.180023	1	0.801896	0.051276
	Scoring	0.289415	0.320906	0.801896	1	0.138513

[표 3-8] 감성분석 결과 데이터 간 상관관계 및 표준편차 (FORD 社)

일반적으로 Pearson Correlation Coefficient 에서 상관계수인 r 값이 +0.1 과 +0.3 사이이면 약한 양의 선형 관계, +0.3 과 +0.7 사이이면 뚜렷한 양의 선형 관계 그리고 +0.7 과 +1.0 사이이면 강한 양의 선형 관계가 있는 것으로 해석하는데 이를 바탕으로 하였을 때 각 데이터 간에는 대체로 일정 수준 이상의 양의 선형 관계를 보이고 있으며 이는 두 가지 감성분석 방법들 그리고 주제와 내용 데이터 간의 감성의 방향에 상관성이 있음을 나타낸다고 할 수 있겠다. 제조사별 데이터에서도 큰 경향성은 유사하게 나타났으며 그 결과는 부록에 수록하였다.

제 4 장 주가 해석과 예측

제 1 절 분석 프로세스

본 연구에서 감성분석 결과를 활용한 주가 분석을 두 단계로 나누어 진행하였는데 첫 번째는 경향성 분석으로 제조사별 20 가지

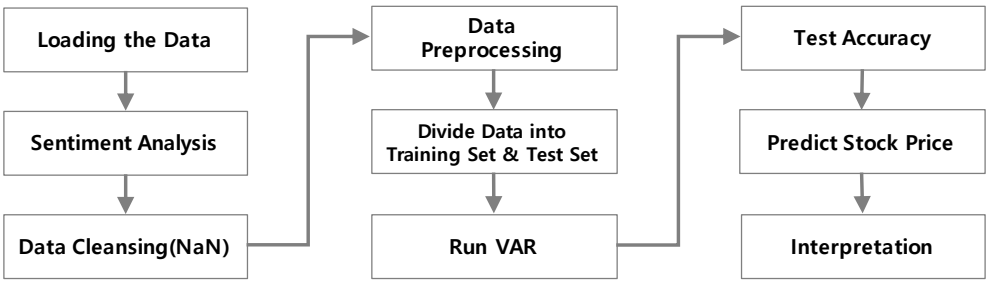
[xxii]로 범주화된 데이터 그룹의 감성분석 결과와 각 제조사의 주가 사이의 연관성을 회귀 경로와 분류 경로로 나누어 검증을 진행함으로써 각각의 데이터 그룹이 가지는 주가 해석의 경향성을 확인하고자 하였다. 회귀 경로에서는 검증 도구로 Linear Regression 과 Random Forest Regressor 를 사용하였으며, 분류경로에서는 Logistic Regression 과 Random Forest Classifier 를 사용하였고 검증 방법으로는 교차 검증(Cross Validation) 방법을 사용하여 진행하였다. 여기서 사용한 주가는 온라인 미디어 데이터 수집 기간과 동일한 2019 년 동안 해당 제조사의 주가 중 일일 종가를 기준으로 사용하였으며 각 제조사의 주식이 상장된 주식시장의 Index[xxiii]의 일일 종가도 이후 비교 검증을 위하여 동일 기간의 데이터를 수집, 사용하였다.

두 번째 단계는 감성분석 결과를 이용하여 제조사별 주가 예측 성능을 확인해 보는 것으로 우선 VAR 모형 기반의 주가 예측모델을 구축하고 난 후 이 모델에 각각의 범주화된 데이터의 감성분석 결과를 학습시키면 예측모델이 만들어지는데 이 모델을 통하여 실제 주가를 예측하여 보고 예측 정확도를 바탕으로 그 특성을 탐색해 보고자 한다. 예측 모델을 이용한 특성 파악 프로세스는 그림 4-1 과 같이 진행되는데 우선 3 번째 단계인 Data Cleansing 단계까지는 앞장에서

[xxii] 20가지 범주화된 데이터 그룹은 다음과 같다.
1) 데이터 범주(2): 제목/내용
2) 미디어 유형 범주(4+1): 전문매체/일반매체/사용자그룹/기타+전체
3) 감성분석 방법(2): Polarity/Scoring

[xxiii] 각 제조사별 상장 주식시장과 사용한 Market Index는 다음과 같다.
[제조사 / 주식시장 / Market Index]
1) FORD / NYSE / S&P 500
2) GENERAL MOTORS / NYSE / S&P 500
3) HYUNDAI / KSE / KOSPI
4) TESLA / NASDAQ / NASDAQ Composite
5) TOYOTA / TSE / NIKKEI 225
6) VOLKSWAGEN / XETR(FWB) / GDAXI

다루었으며 이후 모델 적합성을 위하여 결측값 처리나 데이터 평활화와 같은 추가적인 데이터 전처리를 진행한다. 이후 입력 데이터를 학습데이터와 검증데이터로 나누어 모델에 학습시키고 그 모델의 예측 결괏값으로 실제 값과의 비교를 통하여 정확도를 산출하는 방식으로 진행하였다.



[그림 4-1] 주가 예측 모델 프로세스

제 2 절 경향성 분석과 결과

3 장에서는 수집된 제조사별 데이터를 미디어 유형별로 나누고 이를 Polarity 와 Scoring 두 가지 방법을 통하여 감성분석을 진행하였다. 그리고 이 결과치를 일 일평균 값으로 변환하는 작업으로 주가 데이터와의 시간 단위를 맞추었다. 이 절에서는 범주화된 감성분석 데이터와 제조사별 주가와의 관계를 회귀 경로와 분류 경로를 통하여 검증하여보고 그 경향성을 탐색함과 동시에 각 제조사가 속해있는 주식시장의 Market Index 데이터를 추가로 활용하여 각 데이터 그룹 간의 상관관계를 알아보고자 한다.

하지만 분석에서 적절한 결과를 도출하기 위해서는 이에 앞서 추가적으로 몇 가지 데이터 전처리 작업이 필요한데 첫 번째는 결측값(NA: Not Available)의 처리이다. 감성분석 데이터의 경우 미디어 데이터이기 때문에 조사기간동안 데이터가 비는 결측값이 거의 존재하지 않는다. 미디어를 유형별로 나누었을 때 기타 매체와 같이 절대적인 도수 값이 작은 경우 제조사별로 드물게 나타날 뿐이다. 그러나 주가 데이터의 경우는 이와 다르다. 기본적으로 주말이나 공휴일과 같이 주식시장이 열리지 않는 날은 데이터가 없고 이러한 날이 국가별 또는 주식시장의 상황에 따라 제조사별로 다르게 나타나는데 약간씩의 차이가 있지만, 평균 일 년 365 일을 기준으로 약 110 일 전후의 결측값이 발생하므로 주가에 있어서 전체 데이터의 약 1/3 에 해당하는 결측값의 처리는 결과에 많은 영향을 미친다고 할 수 있다. 결측값을 처리하는 방법에는 크게 결측값을 제거(Deletion)하는 방법과 이를 채우는 방법(Imputation)이 있는데 앞서 언급한 바와 같이 주가 데이터의 경우 결측값의 수가 전체 데이터에서 차지하는 비중이 크기에 이를 삭제처리 할 경우 정보의 손실에 따른 해석력의 저하를 가져올 수 있어 여기에서는 고려하지 않았다. Imputation 의 경우에도 이전 값, 이후 값 또는 특정 값을 채우는 방법 등 다양한 방법들이 존재하나 주가의 선형성을 고려하여 본 연구에서는 보간법(Interpolation) [xxiv]을 사용하여 결측값을 처리하였다.[25] 두 번째 처리 방법은 감성 데이터의 시간 축 이동(Time Shift)이다. 효율적 시장 가설에 따르면 시장에 공개되는 데이터는 즉각적으로 시장의 가격에 반영된다고 가정하고 있으나 이는 실제 현실에서는 조금 다르게 나타난다. 정보가 발생한 이후 개인이나

[xxiv] DataFrame 값에 선형으로 비례하는 방식을 통하여 결측값을 채우는 방법

기관의 능력과 상황에 따라 정보를 획득하는 시간에 편차가 존재하며 또한 정보의 종류 또는 미디어의 종류에 따라 그 확산 속도를 달리한다. 그러나 일반적으로 정도의 차이가 있을 뿐 과거 시점의 데이터가 미래 시점의 주가에 반영된다는 점은 동일하다. 따라서 여기서는 1 일 전(d-1)의 감성 데이터가 그다음 날(d)의 주가에 반영된다고 가정하고 회귀 분석 시 감성 데이터의 시간을 -1 일만큼 이동하여 주가의 시간 축과 맞추고 분석을 진행하였다.

4.2.1. 회귀 경로

회귀 경로에서는 Linear Regression 과 Random Forest Regressor 두 가지 모델을 통하여 관계를 분석하였는데 일 단위 평균을 통하여 적어진 데이터의 수를 고려하여 기본적인 검증 방법은 K=5 인 K-Fold 교차검증법을 사용하였다. K-Fold 교차검증법은 데이터의 수가 적을 때 데이터 분리해 검증데이터의 수를 증가 시켜 학습이 원활하게 이루어질 수 있도록 하는 방법으로 그림 5-1 에 나타난 것처럼 데이터 세트를 K 개의 부분집합으로 나눈 후 하나의 검증용 데이터(1)를 제외한 학습용 데이터(K-1) 세트로 회귀분석 모형을 만들고 이를 검증하는 것을 반복한다. 이때 각 데이터의 부분집합이 검증용 데이터가 되기 때문에 총 K 개의 모형과 같은 수의 교차검증성능이 도출되며 최종적으로 이를 평균하여 성능을 계산하는 방법이다.

[그림 4-2] 5-Fold 교차 검증 모식도

[표 4-1] 회귀 경로를 통한 교차 검증 결과[xxv]

Subject			RMSE of Linear Regression						RMSE of Rnadom Forest					
			FORD	GM	HYUNDAI	TESLA	TOYOTA	VW	FORD	GM	HYUNDAI	TESLA	TOYOTA	VW
Number of Post			0.685	1.806	7371.134	64.320	514.463	15.605	0.775	1.991	7946.044	68.861	547.862	16.389
Title	Polarity	All	0.700	1.739	7347.995	65.210	510.306	15.490	0.806	2.053	8448.923	71.544	556.049	18.230
		Professional Media	0.696	1.749	**7326.759**	64.526	509.499	15.445	0.804	2.156	8444.154	71.635	583.627	17.022
		General Media	0.682	1.735	7340.340	64.967	509.256	15.481	0.714	1.940	8073.200	75.473	585.777	16.676
		Consumer Group	0.687	1.733	7367.968	64.448	509.031	15.445	0.754	1.996	7772.213	69.508	581.487	16.779
		Others	0.679	1.726	7429.974	**63.624**	**503.197**	15.452	0.759	1.971	8118.274	68.637	530.641	16.930
	Score	All	0.692	1.743	7422.294	65.093	509.089	**15.435**	0.732	2.087	8400.752	66.835	566.334	16.336
		Professional Media	0.691	1.742	7414.255	64.505	507.756	15.460	0.795	1.890	8634.589	71.037	546.277	16.336
		General Media	**0.675**	1.737	7361.726	64.613	510.208	15.517	0.706	**1.744**	7702.803	68.166	544.225	16.231
		Consumer Group	0.686	1.737	7348.356	64.462	509.840	15.454	0.709	1.821	7599.767	**65.461**	518.653	**15.854**
		Others	0.678	**1.694**	7388.080	63.728	503.493	15.489	**0.675**	1.847	**7493.880**	65.951	**509.481**	15.908
Content	Polarity	All	0.701	**1.717**	7307.249	65.471	510.546	15.449	0.781	2.030	8157.660	69.663	574.215	17.612
		Professional Media	0.701	1.729	7285.088	64.902	508.309	15.380	0.833	2.021	8059.667	72.318	594.538	16.869
		General Media	0.685	1.726	7308.653	64.958	506.399	**15.359**	0.810	2.105	8082.576	74.223	594.853	18.128
		Consumer Group	0.688	1.728	7320.520	64.139	510.904	15.443	0.819	1.952	8523.109	73.231	555.170	17.308
		Others	0.697	1.736	7356.191	64.330	**503.503**	15.437	0.829	1.962	8283.090	71.120	571.636	18.144
	Score	All	0.687	1.742	7321.663	64.854	507.617	15.486	0.751	1.948	8167.057	71.828	572.053	16.317
		Professional Media	0.693	1.731	**7266.265**	64.918	507.107	15.564	0.735	1.836	7820.380	70.799	530.337	16.433
		General Media	**0.683**	1.725	7356.183	64.574	505.873	15.468	0.704	1.777	7563.443	67.036	533.397	15.633
		Consumer Group	0.688	1.726	7344.633	64.292	510.805	15.518	0.699	**1.762**	7563.613	66.662	566.808	**15.527**
		Others	0.689	1.721	7527.929	**63.965**	508.470	15.607	**0.697**	1.790	**7460.907**	**65.011**	**515.416**	15.797
RMSE of Stock Price			0.593	1.666	6395.007	53.597	418.416	12.823	0.593	1.666	6395.007	53.597	418.416	12.823

[xxv] 해당 표는 세로축을 기준으로 제목과 내용 범주 아래 Polarity과 Scoring을 구분하여 4개의 소그룹을 만들고 1)각 소그룹 내에서 가장 성능이 좋은 미디어 유형에 음영 처리한 후 2)각 소그룹 내에서 음영 처리된 미디어 유형을 제목과 내용 안에서 각각 비교하여 그 중 성능이 뛰어난 것에 Bold 처리하고 3)최종적으로 제조사 열에서 가장 뛰어난 결과 값에 밑줄을 그어 구분 하였다.

회귀 경로 교차검증에서 감성과 주가 데이터의 관계 정합성은 RMSE(Root Mean Square Error) 값을 도출하는 방식으로 측정하였는데 이는 실제 값(y_i)에서 예측 값(y'_i)을 뺀 후 이의 제곱의 합을 더하여 구한다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2}$$

표 4-1 은 각 범주화 데이터의 회귀분석 교차검증 결과를 감성 분석 방법에 따라 제조사별로 구분하여 나타낸 것으로 그 경향성을 살펴보면 회귀 방법에 있어 Linear Regression 의 RMSE 가 Random Forest Regressor 와 비교해 모든 항목에서 적게 나타나고 있음을 알 수 있으며, 감성 분석 방법에서는 Polarity 보다는 Scoring 이 전반적으로 나은 성능을 보이고 있다. 미디어 유형에 따른 경향은 기타 매체가 범주화 데이터 그룹 내에서 상대적으로 높은 경향을 보이는 경우가 많았고 그 뒤를 이어 일반 매체와 사용자 그룹의 순으로 뛰어난 성능을 보이는 것으로 나타났다. [xxvi]

4.2.2. 분류 경로

분류 경로에서도 기본적인 검증 방법은 회귀 경로와 동일한 K-Fold 교차검증법을 사용하였으며 분류자(Classifier)로 Logistic Regression 과 Random Forest Classifier 를 사용하였다. 다만, 분류의 경우 그 특성상 분류의 기준이 되는 범주를 만들어 주어야 하는데 여기서는 일 단위 주가의 변동 폭을 기준으로 1%를 넘게 움직이는 경우를 기준으로 하여 범주를 구분하였다.

$$Stock\ Price\ Change_d = \frac{Close\ Price_d - Close\ Price_{d-1}}{Close\ Price_{d-1}}$$

[xxvi] 주가 자체의 움직임을 나타내는 Null Model의 RMSE(표준 편차) 값과 비교하여 보았을 때 회귀 경로의 모델들은 이를 상회하는 것으로 나타나 회귀 경로가 이를 해석하는데 적합한 모델이 아니라고 볼 수도 있으나, 1차 검증의 목적은 데이터 그룹별 경향성을 찾는 것에 있으므로 추가적인 해석은 하지 않도록 한다.

Subject			Accuracy of Rnadom Forest						Accuracy of Logistics Regression					
			FORD	GM	HYUNDAI	TESLA	TOYOTA	VW	FORD	GM	HYUNDAI	TESLA	TOYOTA	VW
Number of Post			0.735	0.661	0.725	0.543	0.839	0.672	0.758	0.738	0.758	0.567	0.878	0.730
Title	Polarity	All	0.645	0.611	0.611	0.468	0.773	0.606	0.747	0.741	**0.766**	0.545	**0.878**	**0.727**
		Professional Media	0.625	0.642	0.653	**0.562**	0.778	0.623	0.747	0.741	**0.766**	**0.548**	**0.878**	**0.727**
		General Media	0.633	0.630	0.650	0.455	0.806	0.634	0.749	0.739	**0.766**	0.540	**0.878**	**0.727**
		Consumer Group	0.663	0.636	0.711	0.551	0.784	0.672	0.746	0.741	**0.766**	**0.548**	**0.878**	**0.727**
		Others	0.662	0.663	0.708	0.494	0.838	0.642	**0.751**	**0.743**	0.764	0.536	0.877	**0.727**
	Score	All	0.628	0.689	0.703	0.477	0.825	0.617	0.747	0.741	**0.766**	0.543	**0.878**	**0.727**
		Professional Media	0.694	0.667	0.678	0.543	0.809	0.713	0.747	0.741	**0.766**	0.526	**0.878**	**0.727**
		General Media	**0.735**	0.714	**0.769**	0.551	0.839	0.711	0.749	0.739	**0.766**	0.545	**0.878**	**0.727**
		Consumer Group	**0.735**	**0.747**	0.755	0.512	0.847	0.700	0.746	0.741	**0.766**	0.534	**0.878**	**0.727**
		Others	0.731	0.738	0.708	0.483	**0.877**	**0.725**	**0.751**	**0.743**	0.764	0.528	0.877	**0.727**
Content	Polarity	All	0.661	0.628	0.664	0.493	0.759	0.540	0.747	0.741	**0.766**	**0.545**	**0.878**	**0.727**
		Professional Media	0.628	0.576	0.600	0.471	0.784	0.592	0.747	0.741	**0.766**	**0.545**	**0.878**	**0.727**
		General Media	0.660	0.600	0.656	0.501	0.814	0.559	0.749	0.739	**0.766**	**0.545**	**0.878**	**0.727**
		Consumer Group	0.644	0.631	0.691	0.540	0.754	0.680	0.746	0.741	**0.766**	**0.545**	**0.878**	**0.727**
		Others	0.643	0.641	0.617	0.495	0.782	0.620	**0.751**	**0.743**	0.764	0.544	0.877	**0.727**
	Score	All	0.716	0.653	0.670	0.507	0.775	0.683	0.747	0.741	**0.766**	**0.545**	**0.878**	**0.727**
		Professional Media	0.697	0.719	0.708	0.534	0.859	0.669	0.747	0.741	**0.766**	**0.545**	**0.878**	**0.727**
		General Media	0.707	**0.825**	0.755	**0.562**	0.864	0.719	0.746	0.736	**0.766**	**0.534**	**0.878**	**0.727**
		Consumer Group	0.724	0.713	0.744	0.540	0.823	0.719	0.746	0.741	**0.766**	**0.545**	**0.878**	**0.727**
		Others	**0.742**	0.738	**0.767**	0.511	**0.866**	**0.722**	**0.751**	**0.743**	0.764	0.539	0.877	**0.727**
Accuracy of Null Model			0.747	0.741	0.766	0.545	0.878	0.727	0.747	0.741	0.766	0.545	0.878	0.727

[표 4-2] 분류 경로를 통한 교차 검증 결과[xxvii]

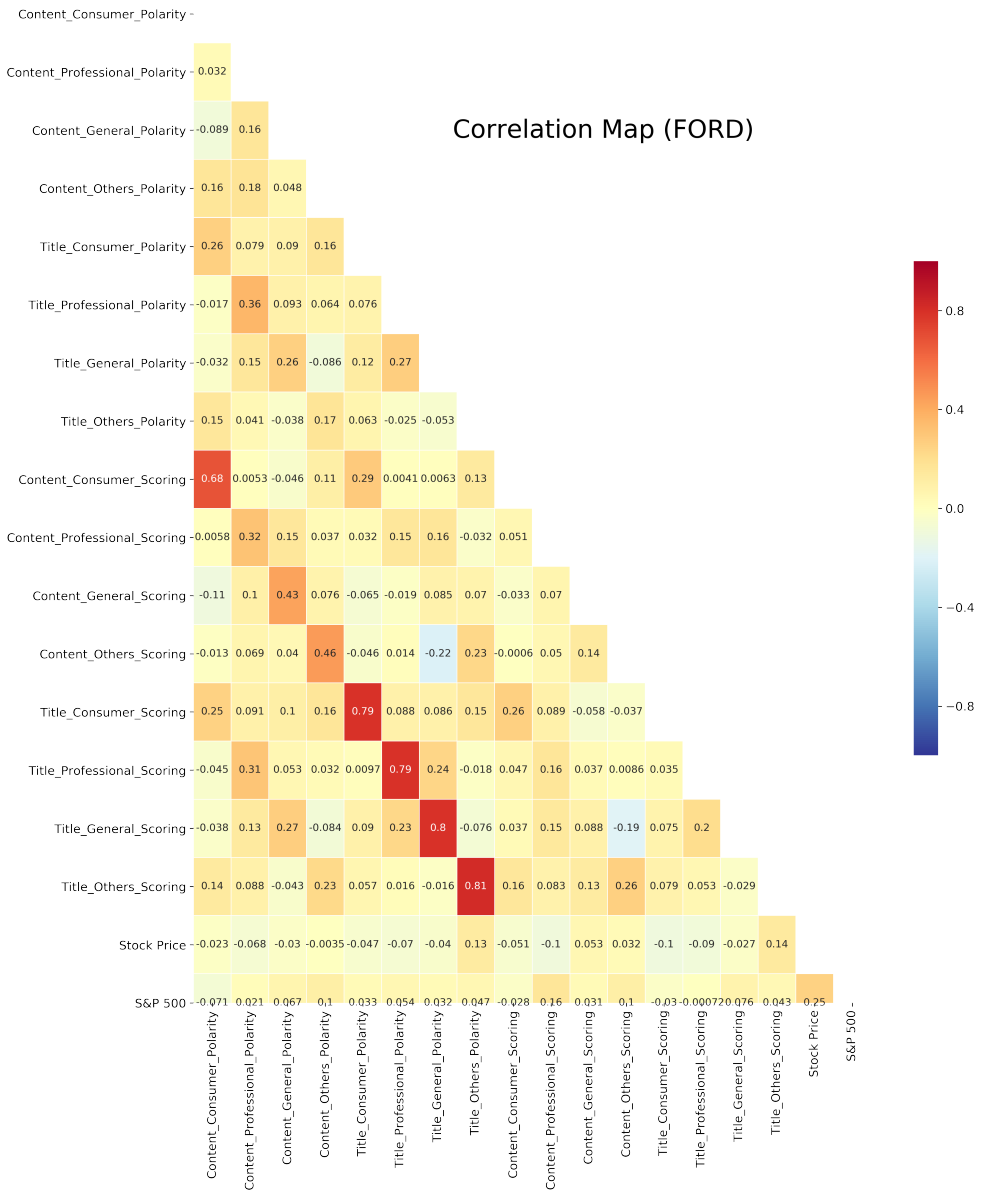
[xxvii] 분류 경로에서 Logistic Regression의 경우 각 범주데이터별로 동일 값을 나타내는 경우가 많아 이 경우 비교 시 동일 값 모두를 동일하게 구분 표기 처리하였다.

분류경로에서는 회귀 경로와는 달리 오차(Error)값이 아닌 정확도(Accuracy)를 산출하여 그 성능을 평가하였는데 그 결과는 표 4-2 와 같다. 모델 간의 비교에서는 Logistic Regression 모델이 모든 데이터 그룹에서 Random Forest Classifier 보다 우수한 성능을 나타내는 것으로 나타났으나, 감성 분석 방법이나 제목, 내용 간에는 유의미한 차이를 찾을 수 없었다. 미디어 유형에서는 사용자 그룹이 타 미디어 유형 대비 상대적으로 뛰어난 성능을 보였으며 일반 매체와 기타 매체가 그다음으로 유사한 수준의 성능을 보였다. 더불어 분류 경로 검증의 경우 각 방법에서의 정확도가 Null Model 의 정확도와 비교하여 높게 형성되어 절대적 성능에서도 그 의미가 있다 할 수 있겠다.

이번 분석의 목적은 언급한 바와 같이 절대적인 모델 성능에 있는 것이 아니라 데이터 그룹에 따른 경향성을 파악하는 데 있다. 이러한 관점에서 접근하였을 때 감성 분석 방법에서는 Scoring 을 통한 방법이 높은 성능을 보여주었으며, 제목과 내용의 데이터 그룹에서는 미세하지만, 제목의 성능이 높게 나타났다. 미디어 유형에서는 기타 매체, 사용자 그룹 그리고 일반 매체의 순으로 성능이 높은 것으로 나타났는데 이러한 경향성에 대한 해석은 다음 장에서 다루도록 한다.

4.2.3. 상관 관계

경향성 분석의 마지막으로 제목과 내용 그리고 각 미디어 유형별로 범주화된 감성 데이터와 제조사별 주가 그리고 해당 제조사가 속해 있는 주식시장의 Market Index 간의 상관관계(Correlation)를 확인해 보았다. 그림 4-3 은 FORD 사의 데이터를 도식화한 것으로 해당 분포에서도 확인할 수 있듯이 기본적으로 Polarity 를 바탕으로 하는 구분된 Scoring 은 대부분의 미디어 유형에서 강한 양의 상관관계를 나타내고 있으며, 각 미디어 유형의 내용과 제목 사이에도 전반적인 양의 상관관계를 보이고 있음을 알 수 있다. 다만 주가와의 연동성에서는 특정 매체 또는 데이터 그룹이 일관적으로 높은 상관관계를 나타내는 경향성은 보이지 않았고 제조사에 따라 다르게 나타남을 알 수 있었다. 이에 특정적인 미디어 그룹 또는 데이터 범주를 일관적으로 사용하는 것보다는 제조사나 데이터의 특성에 맞게 선별해서 사용하는 것이 향후 보다 정도 있는 결과를 도출하는데 더 합리적이라 할 수 있겠다. 타 제조사의 상관관계 분포는 부록에 수록하였다.



[그림 4-3] 상관관계 분포 (FORD 社)

제 3 절 예측성능 분석과 결과

4.3.1. 데이터 전처리(2)

4.3.1.1 데이터 정상성(Data Stationary)

VAR 모델을 활용한 주가 예측에 앞서 본 연구에서는 모델에 가장 적합한 Data Set 을 만들기 위하여 몇 가지 사항을 고려하였는데 첫 번째는 데이터 정상성(Data Stationary)이다. 기본적으로 시계열 분석에 사용되는 데이터의 경우 정상성을 가지는 것을 전제로 하는데 이는 정상성을 가지는 시계열 데이터의 특징이 시간과 무관하기 때문이며 반대로 정상성을 가지고 있지 않은 시계열 데이터의 경우 장기적인 관점에서 보았을 때 예측 가능한 패턴을 가지고 있지 않는다는 것이 된다. 따라서 시계열 예측에서 변수로 사용되는 데이터에 대해서 정상성 유무를 사전에 판단해 볼 필요가 있다. 정상성 판단은 시계열 데이터를 그래프와 같이 시각화하여 판단하는 방법과 Augmented Dickey Fuller(ADF) 검정[xxviii]과 같은 단위근 검정 방법이 많이 사용된다. 여기에서 사용된 ADF 검정은 Dickey Fuller(DF) 검정을 일반화 한 것으로 회귀 분석 결과로 나온 계수를 검정 통계량으로 사용하며 다음과 같은 회귀 방정식을 가진다.

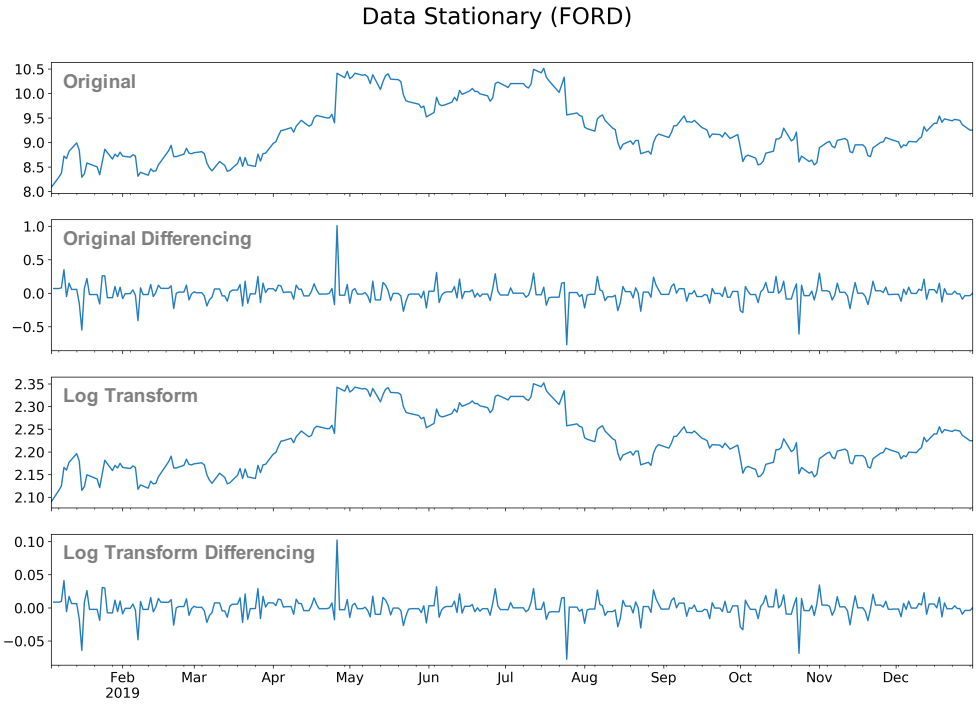
$$\nabla Y_t = \alpha + \beta t + \gamma Y_{t-1} + \phi_1 \nabla Y_{t-1} + \phi_2 \nabla Y_{t-2} + \cdots + \phi_p \nabla Y_{t-p} + e_t$$

이 식에서 γ값은 검정 통계량으로 귀무가설이 성립하는 경우 $\gamma=0$ 이며, $\nabla Y_t = Y_t - Y_{t-1}$ 를 종속변수로, 원 시계열 Y_t와 차분 값의 지연 값들 ∇Y_{t-1}, ∇Y_{t-2}, $\cdots$, ∇Y_{t-k} 을 독립변수로 하는 회귀 분석을 통하여 γ 값을 구할 수 있다.

감성분석 결과 데이터의 경우 정해진 상한값(+1)과 하한값(-1)을 벗어날 수 없고 앞서 3 장에서 다루었던 시계열 그래프를 참조하여 보더라도 전형적인 정상성을 가지는 데이터의 배열을 보인다. ADF 검정

[xxviii] 데이터의 정상성(Stationary)를 알아보기 위한 단위근(unit root test) 검정 방법으로, 자료에 단위근이 존재한다는 귀무가설(H_0)과 해당 자료가 정상성을 가진다는 대립가설(H_1)에서 p-value로 데이터의 정상성을 판단하는 방법

결과 역시 p-value 가 매우 작은 값[xxix]을 나타내며 정상성을 나타냄을 확인할 수 있었다. 그러나 주가의 경우 추세를 보이며 대표적인 비 정상성(Non-Stationary) 데이터로 ADF 검정 결과 제조사를 막론하고 높은 p-value 를 나타냈다. 여기서는 비정상성을 가지는 주가 데이터를 정상성을 가지는 데이터로 변환하기 위하여 주가에 대해 차분(Differencing), 로그 변환(Log Transformation) 그리고 로그 변환 후 차분(Log Transformation and Differencing)의 방법을 통하여 정상화 여부를 확인하고자 하였다. 그림 4-4 는 FORD 社의 주가 데이터에 대하여 위의 방법들을 적용한 후 그래프로 시각화한 것으로 시각적으로 보았을 때 차분을 거쳤을 때 정상성을 가질 것으로 예측할 수 있으며 실제 ADF 검정에서도 원래 주가와 로그 변환에서는 p-value > 0.05 로 나타나 정상성이 없음이 확인되었고 차분과 로그 변환 후 차분에서는 p-value 가 매우 작게 나타나 정상성을 가짐이 확인되었다.



[그림 4-4] 주가 데이터 변환에 따른 정상성 (FORD 社)

표 4-3 은 차분, 로그 변환, 로그 변환 후 차분의 데이터 처리 방법을 적용하고 제목과 내용에 대해 Polarity 와 Scoring 의 감성분석 방법을

[xxix] p-value < 0.00001

구분하여 도출한 감성분석 결과 값과 해당 제조사의 주가 데이터에 대하여 VAR 모델을 통하여 오차 성능을 구해본 것이다. 이 결과를 보면 제조사별 오차가 가장 적게 나타나는 경우의 비율은 차분을 통하여 정상성을 가진 데이터의 경우에서 가장 높음을 관찰할 수 있다. 다만 차분을 적용한 두 가지 경우 중 로그 변환 후 차분의 경우 샘플 사이즈에 의해 모델 구현이 원활하지 않은 경우가 발생한 바 이 연구에서는 원래 주가 값을 차분한 데이터를 기준으로 예측모델을 구현하고자 한다. 다만 결괏값에서도 나타나듯이 제조사별 또는 범주화된 데이터의 구분에 따라 그 경향성이 약간씩 다르게 나타나고 있기에 데이터 그룹이 한정될 경우 적절한 변환 방법 선정을 다시 고려하는 것이 적합하다 할 수 있겠다.

			FORD	GM	HYUNDAI	TESLA	TOYOTA	VOLKSWAGEN
CONTENT	Polarity	None	0.30621%	0.36155%	0.32421%	0.85892%	0.22197%	0.08173%
		Diff	0.36451%	0.30736%	0.37194%	0.79916%	0.26675%	0.17250%
		Log	0.30796%	0.36645%	0.33594%	0.86118%	0.22296%	0.09918%
		Log Diff	0.31841%	0.30749%	0.36204%	0.83029%	0.26171%	0.22227%
	Scoring	None	0.37007%	0.35258%	0.44287%	0.88115%	0.22901%	0.11735%
		Diff	0.44973%	0.33623%	0.44029%	0.82388%	0.24393%	0.22622%
		Log	0.34981%	-	0.44423%	0.88032%	0.22957%	0.11983%
		Log Diff	0.41469%	-	0.42439%	0.84945%	0.24623%	0.22902%
TITLE	Polarity	None	0.27360%	0.34689%	0.38799%	0.88110%	0.22402%	0.05525%
		Diff	0.38625%	0.31497%	0.44708%	0.86328%	0.22760%	0.16351%
		Log	-	-	0.44496%	-	0.24040%	-
		Log Diff	-	-	0.32232%	-	0.25738%	-
	Scoring	None	0.28067%	0.32666%	0.40777%	0.85381%	0.23228%	0.04316%
		Diff	0.38379%	0.47204%	0.38949%	0.81838%	0.21062%	0.16992%
		Log	-	-	-	-	-	-
		Log Diff	0.30621%	0.36155%	0.32421%	0.85892%	0.22197%	0.08173%

[표 4-3] 데이터 정상화 방법과 VAR 모델 오차 성능

표 4-4 는 각 제조사별로 최종적으로 모델에 사용하게 될 3 가지 변수들인 감성분석결과, 제조사 주사 그리고 Market Index 에 대해 ADF 검정을 실시한 결과이며 원 시계열 데이터가 정상성을 가지지 않았던 제조사 주가와 Market Index 의 경우 역시 차분을 통하여 정상성이 확보되었음을 알 수 있다.

		FORD		GM		HYUNDAI		TESLA		TOYOTA		VOLKSWAGEN	
		Original	Differencing	Original	Differencing	Original	Differencing	Original	Differencing	Original	Differencing	Original	Differencing
Sentiment	p-value	< 0.000001		< 0.000001		< 0.000001		< 0.000001		< 0.000001		< 0.000001	
Stock	ADF Statistic	-2.68681	-17.590375	-3.177699	-16.021518	-2.367394	-6.048073	0.048386	-14.47701	-1.180203	-15.143005	-1.18487	-14.718279
	p-value	0.076349	< 0.000001	0.021312	< 0.000001	0.151151	< 0.000001	0.962355	< 0.000001	0.682054	< 0.000001	0.68007	< 0.000001
Market Index	ADF Statistic	-1.828461	-17.320809	-1.828461	-17.320809	-2.081469	-9.468772	-1.842229	-12.791341	-1.575688	-16.118672	-1.707687	-12.116455
	p-value	0.366426	< 0.000001	0.366426	< 0.000001	0.252031	< 0.000001	0.359707	< 0.000001	0.495794	< 0.000001	0.427156	< 0.000001

▶ Critical Values (1%) : -3.457 | (5%): -2.873 | (10%): -2.573

[표 4-4] 제조사별 변수의 정상성과 ADF 검정결과

4.3.1.2 변수(Variables)

두 번째 고려 사항은 추가 변수이다. 기본적으로 주가의 예측을 위한 모델에서 사용되는 변수는 제조사의 주가 데이터와 각 데이터 그룹의 감성분석 결과 두 가지 이다. 여기에 더하여 본 연구에서는 적절한 외생 변수를 추가함으로써 해석 능력을 향상시킬 수 있는지를 탐색해 보고 적절한 추가 변수를 선정 함으로써 성능향상의 방법으로 활용하고자 하였다. 추가적인 변수로 고려한 것은 3 가지로 제조사의 주식시장 Market Index, 주가의 거래량(Volume) 그리고 미디어 데이터의 발생 수량(No of Contents)이다. 적절한 추가 변수를 선정하기 위하여 앞에서와 같이 각 변수의 오차 성능을 확인해 보았는데 우선 앞서 데이터 정상성을 획득하기 위한 방법으로 원래 값의 차분을 적용하는 것을 전제로 하여 주가와 감성분석 데이터에 각 추가 변수를 하나 또는 두 개를 조합하는 형식으로 적용하여 전체 경우의 수에 대한 성능을 검증하였다. 표 4-5 는 이의 오차 성능을 검증한 것으로 이를 통하여 확인할 수 있는 것은 다음과 같다. 첫째, 기본적으로 고려한 주가와 감성분석 데이터 2 가지 변수만을 적용했을 때 보다 추가 변수를 반영하였을 때 오차가 상대적으로 낮게 나타나는 경향을 보였다. 둘째, 추가 변수의 수가 늘어난다고 해서 해석 능력이 올라가는 것은 아니다. 이는 추가 변수가 1 개일 때가 2 개일 때보다 전반적으로 오차가 낮게 나타남을 통해 알 수 있다. 끝으로 추가 변수의 적용에 대하여 해석 능력의 향상 경향성은 있으나 적절한 변수의 종류와 수량은 제조사별 데이터의 특성에 따라 달라질 수 있음을 알 수 있다. 이후 범주화된 데이터의 특성에 따른 분석을 위하여 여기에서는 추가 변수 중 상대적으로 가장 높은 해석 능력을 보여준 Market Index 한가지로 한정하여 적용하기로 한다.

Data		Variable 1	Variable 2	FORD	GM	HYUNDAI	TESLA	TOYOTA	VW
CONTENT	Polarity	-	-	0.36451%	0.30736%	0.37194%	0.79916%	0.26675%	0.17250%
		Index	-	0.33368%	0.31362%	0.30647%	0.78472%	0.27484%	0.23637%
		NoC	-	0.37327%	0.30311%	0.38469%	0.73620%	0.32283%	0.22216%
		Volume	-	0.30243%	0.43812%	0.36206%	0.86624%	0.28805%	0.20668%
		Index	NoC	0.38315%	0.30863%	0.40762%	0.74481%	0.29603%	0.20719%
		Index	Volume	0.35142%	0.25925%	0.33727%	0.89305%	0.26280%	0.29700%
		NoC	Volume	0.34847%	0.34044%	0.35610%	0.78941%	0.33237%	0.23344%
	Scoring	-	-	0.44973%	0.33623%	0.44029%	0.82388%	0.24393%	0.22622%
		Index	-	0.41263%	0.29014%	0.43418%	0.79790%	0.24227%	0.20678%
		NoC	-	0.36503%	0.31023%	0.44814%	0.73947%	0.24890%	0.28823%
		Volume	-	0.37747%	0.37004%	0.40923%	0.85741%	0.26359%	0.22774%
		Index	NoC	0.38228%	0.30245%	0.44654%	0.74126%	0.24179%	0.27205%
		Index	Volume	0.44291%	0.32711%	0.45014%	0.87243%	0.25263%	0.24839%
		NoC	Volume	0.40191%	0.33259%	0.44297%	0.78706%	0.26265%	0.28691%
TITLE	Polarity	-	-	0.38625%	0.31497%	0.44708%	0.86328%	0.22760%	0.16351%
		Index	-	0.39940%	0.46250%	0.41324%	0.80112%	0.25138%	0.11261%
		NoC	-	0.42630%	0.34862%	0.45820%	0.80806%	0.26086%	0.18209%
		Volume	-	0.36232%	0.35232%	0.46051%	0.91281%	0.25925%	0.15576%
		Index	NoC	0.44119%	0.48552%	0.48459%	0.75887%	0.25465%	0.17065%
		Index	Volume	0.37875%	0.34104%	0.45045%	0.86750%	0.26284%	0.20191%
		NoC	Volume	0.39800%	0.46434%	0.46586%	0.84801%	0.27074%	0.18240%
	Scoring	-	-	0.38379%	0.47204%	0.38949%	0.81838%	0.21062%	0.16992%
		Index	-	0.37766%	0.41546%	0.37993%	0.86495%	0.24776%	0.18211%
		NoC	-	0.40148%	0.48795%	0.41306%	0.78960%	0.28008%	0.16762%
		Volume	-	0.38210%	0.42864%	0.41209%	0.86467%	0.23159%	0.14766%
		Index	NoC	0.41057%	0.62662%	0.43084%	0.78036%	0.22852%	0.20302%
		Index	Volume	0.38504%	0.43431%	0.38550%	0.90861%	0.26617%	0.16796%
		NoC	Volume	0.41365%	0.55553%	0.44081%	0.83912%	0.26992%	0.15833%

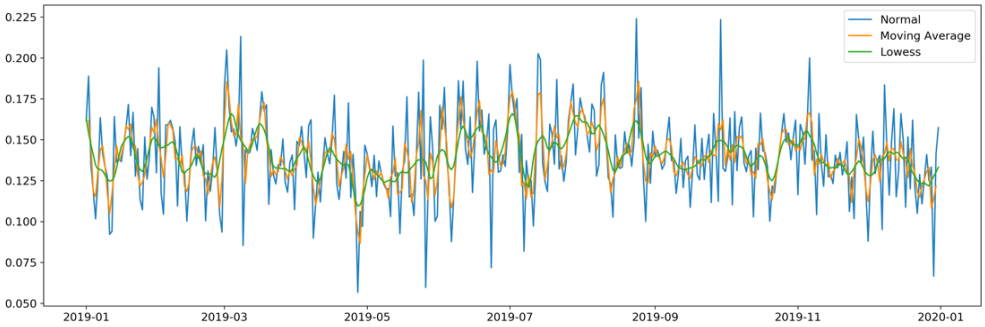
[표 4-5] 추가 변수와 VAR 모델 오차 성능

4.3.1.3 데이터 평활화(Data Smoothing)

마지막 고려사항은 데이터 평활화(Data Smoothing)이다. 평활화는 데이터 내에 존재하는 노이즈로 인하여 발생할 수 있는 경향성의 훼손 등을 줄이기 위한 방법으로 사용된다. 그림 4-5 는 감성분석 결과 데이터를 이동 평균(Moving Average, day = 3)법[xxx]과 국소 가중 산점도 평활화(Locally Weighted Scatter Plot Smoothing, fraction =

[xxx] 이동평균법(Moving Average): 과거 일정 기간 동안의 이를 평균하여 추세를 파악하고 미래를 예측하는 방법으로 일반적인 평균을 구하는 단순이동평균법, 특정부분에 가중치를 주어 계산하는 가중이동평균법 그리고 지수이동평균법 등이 있다.
$\Rightarrow F_{n+1} = \sum Z_t/m,\ (t=n-m+1) = (Z_n + Z_{n+1} + Z_{n+2} + \ldots + Z_{n-m+1})/m$

0.03)법[xxxi]으로 평활화한 결과를 보여주고 있다. 그래프의 형태에서도 알 수 있듯이 원래 데이터에 포함되어 있는 노이즈와 특이점(Outlier)들을 평활화하여 데이터의 변동성을 단순화하고 움직임의 경향을 보다 명확하게 나타내고 있음을 알 수 있다.



[그림 4-5] 감성분석 데이터의 평활화 (FORD 社)

표 4-6 는 감성분석 결과에 이동 평균과 가중 회귀분석의 데이터 평활화 방법을 적용하여 주가예측 특성을 알아본 것이다. 일관 되다고는 볼 수 없으나 전반적으로 데이터 평활화를 적용한 데이터 그룹이 적용하지 않은 그룹에 비해 상대적인 오차가 적게 나타남을 확인할 수 있으며, 그중에서도 국소 가중 산점도 평활화 방법을 적용한 경우가 이동 평균을 적용한 경우보다 약간이지만 예측 오차가 낮음을 알 수 있다. 다음에서 다루게 될 범주화된 데이터 그룹별 주가예측에서는 감성분석 결과에 가중 회귀분석의 데이터 평활화 기법을 적용하기로 한다.

[xxxi] 국소 가중 산점도 평활화(Locally Weighted Scatterplot Smoothing): 산점도 등에서 추세를 파악하기 위한 방법으로 이변량 자료를 나누어 각각의 구간에서 가중선형회귀(Weighted Linear Regression)를 사용하여 회귀곡선을 구하는 방법

DATA		Smoothing	FORD	GM	HYUNDAI	TESLA	TOYOTA	VW
CONTENT	Polarity	None	0.33368%	0.31362%	0.30647%	0.78472%	0.27484%	0.23637%
		Moving Avg.	0.35071%	0.17479%	0.29094%	0.81894%	0.22703%	0.21825%
		LOWESS	0.33537%	0.22165%	0.31241%	0.83362%	0.25911%	0.18022%
	Scoring	None	0.41263%	0.29014%	0.43418%	0.79790%	0.24227%	0.20678%
		Moving Avg.	0.42019%	0.25226%	0.44683%	0.83933%	0.24536%	0.20207%
		LOWESS	0.48367%	0.20820%	0.34135%	0.87728%	0.25191%	0.17480%
TITLE	Polarity	None	0.39940%	0.46250%	0.41324%	0.80112%	0.25138%	0.11261%
		Moving Avg.	0.39858%	0.55292%	0.35117%	0.79764%	0.22067%	0.17409%
		LOWESS	0.41792%	0.28274%	0.31498%	0.87126%	0.27954%	0.13296%
	Scoring	None	0.37766%	0.41546%	0.37993%	0.86495%	0.24776%	0.18211%
		Moving Avg.	0.40490%	0.53148%	0.34845%	0.82159%	0.20207%	0.15330%
		LOWESS	0.44579%	0.23775%	0.36267%	0.79586%	0.25455%	0.12618%

[표 4-6] 데이터 평활화 기법과 VAR 모델 오차 성능

4.3.2. 미디어 유형 및 제목과 내용에 따른 주가예측

미디어 유형 및 제목과 내용에 따른 주가 예측에 앞서 추가적인 데이터 전처리 작업을 통하여 예측모델 실행 조건들을 세분화하고 기준을 정립하였으며 그 내용은 표 4-7 에 나타난 바와 같다.

구분	적용기준	비고
예측 모델 (Model)	벡터자기회귀(VAR) 모형	
적용 변수 (Variable)	감성분석 결과, 제조사 주가, Market Index	
결측값 처리 (Not Available)	보간법(Interpolation)	보간법 적용 후 최종 NaN 삭제
데이터 정상화 (Stationary)	차분(Differencing)	적용: 비정상화 데이터 (제조사 주가, Market Index)
데이터 평활화 (Smoothing)	가중회귀(LOWESS / Fraction = 0.03)	적용: 감성분석 결과
학습/검증 데이터 셋 (Training/Test Data Set)	N / N-10	N: 전체 데이터 수
오차 계산 (Error)	(예측값-실제값)/실제값 x 100 (%)	

[표 4-7] 예측 모형 실행 조건

본 연구에서 사용된 벡터자기회귀모형(Vector Autoregressive Model)은 다변량 시계열 데이터들의 선형적 상관관계를 분석하는 모형이다. 따라서 모형 내 서로 인과관계가 있는 모든 변수의 현재의 관측값을 종속변수로, 모든 시차변수(과거 관측치들)를 독립변수로 구성한 선형회귀 방정식을 통하여 시계열 프로세스를 추정해내는 방법이다.[26]

상기 예측모델 실행 조건을 기준으로 주식가격을 종속변수로 하는 모형의 추정치에 대한 FORD社의 예시는 표 4-8에 나타난 바와 같다.

	Coefficient	Std. Error	T-stat	Probability
Constant	0.156319	0.117745	1.328	0.184
Stock Close	-0.062312	0.061833	-1.008	0.314
Market Index	0.000039	0.000477	0.081	0.935
Sentiment	8.700759	4.937043	1.762	0.078

► Dependent Variable: Stock Close / Lag of Order: 5

[표 4-8] 주식가격에 대한 VAR 모형 추정치 (FORD社)

여기에서는 모델의 적정 차수(Lag of Order)를 결정하는 방법으로 AIC(Akaike Information Criterion)방법을 사용하였으며, 이를 계산하는 수식은 아래와 같다. 아래 기준을 최소로 만드는 p를 최적 차수로 결정한다. [27]

$$\mathrm{A}IC = \log(|\widetilde{\sum}|) + \frac{2pm^2}{n}$$

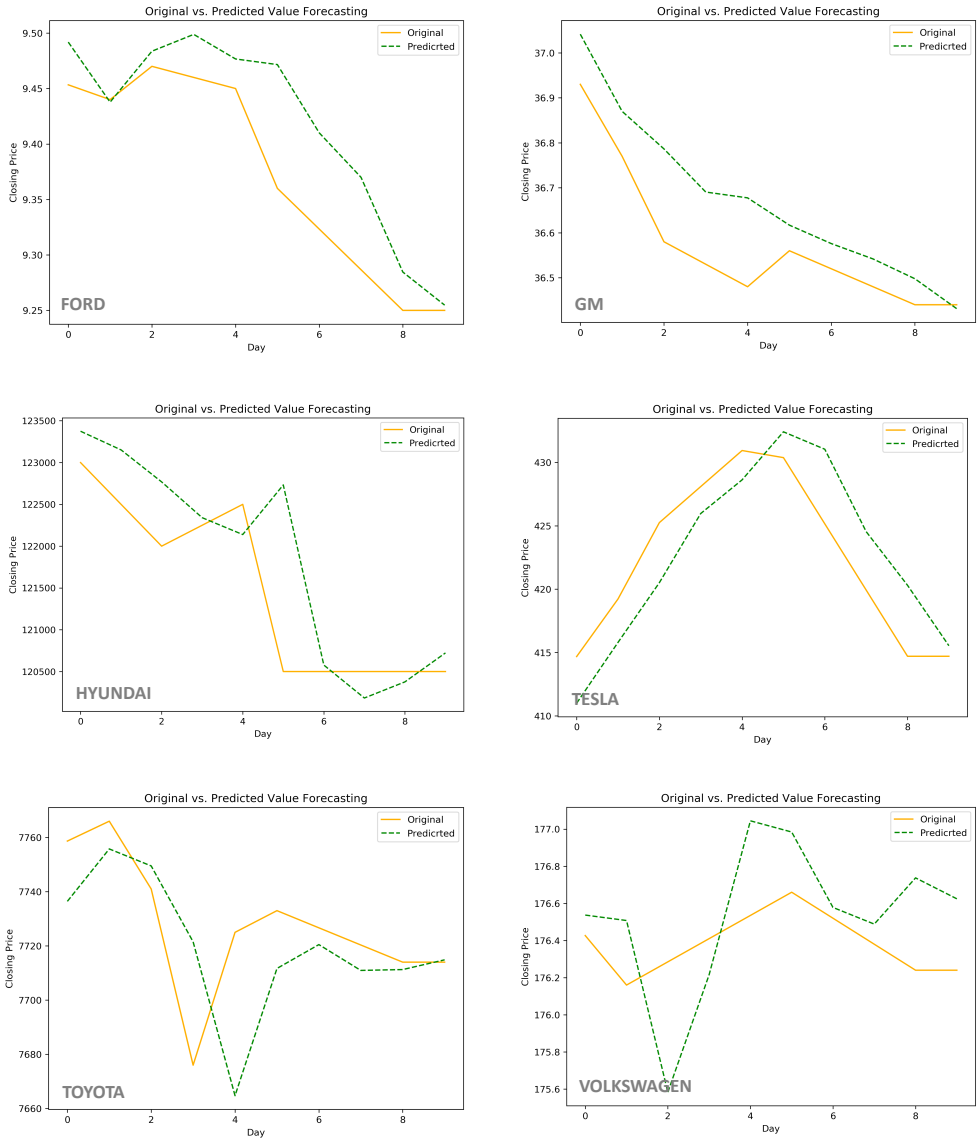
m: 벡터 시계열 차원 수
p: 차수
n: 유효한 샘플 사이즈

다음에서는 정립된 예측모델 실행 조건을 기준으로 최종적으로 제조사별 범주화된 데이터 그룹 간의 예측 오차를 확인하여 보고 그 결과를 통하여 데이터 그룹별 해석 특성을 확인한 후 이를 통하여 유의미한 경향성을 탐색하고자 하였다.

[표 4-9] 데이터 그룹별 주가예측 오차 성능

Subject			VAR Model Error (%)								
			FORD	GM	HYUNDAI	TESLA	TOYOTA	VW	Mean		
Title	Polarity	All	0.43605%	0.23747%	0.43982%	0.87407%	0.22817%	0.20201%	0.40293%	0.37274%	0.37103%
		Professional Media	0.39748%	0.25282%	0.35768%	0.81462%	0.27348%	0.13485%	0.37182%		
		General Media	0.37447%	0.25482%	0.34243%	0.82935%	0.27633%	0.10102%	0.36307%		
		Consumer Group	0.38679%	0.25563%	0.35812%	0.79965%	0.27373%	0.09720%	0.36185%		
		Others	0.37689%	0.27095%	0.35372%	0.80191%	0.27741%	0.10333%	0.36404%		
		Mean	**0.39433%**	**0.25434%**	**0.37035%**	**0.82392%**	**0.26582%**	**0.12768%**	**0.37274%**		
	Scoring	All	0.49006%	0.24340%	0.31477%	0.86191%	0.22918%	0.17244%	0.38529%	0.36931%	
		Professional Media	0.37899%	0.23092%	0.36066%	0.87778%	0.25896%	0.11895%	0.37104%		
		General Media	0.38138%	0.23728%	0.34977%	0.87040%	0.25816%	0.08578%	0.36379%		
		Consumer Group	0.39620%	0.24208%	0.34308%	0.81873%	0.27347%	0.10135%	0.36248%		
		Others	0.38687%	0.24619%	0.35596%	0.82525%	0.28007%	0.08928%	0.36394%		
		Mean	**0.40670%**	**0.23997%**	**0.34485%**	**0.85081%**	**0.25997%**	**0.11356%**	**0.36931%**		
Content	Polarity	All	0.47114%	0.27798%	0.43022%	0.83632%	0.24249%	0.18351%	0.40694%	0.36946%	0.36623%
		Professional Media	0.44463%	0.27941%	0.34486%	0.78423%	0.25732%	0.07989%	0.36506%		
		General Media	0.38665%	0.25619%	0.36747%	0.78284%	0.26210%	0.07973%	0.35583%		
		Consumer Group	0.39691%	0.27959%	0.37609%	0.76588%	0.26777%	0.07951%	0.36096%		
		Others	0.39428%	0.28312%	0.35828%	0.77094%	0.26103%	0.08337%	0.35850%		
		Mean	**0.41872%**	**0.27526%**	**0.37538%**	**0.78804%**	**0.25814%**	**0.10120%**	**0.36946%**		
	Scoring	All	0.54627%	0.35032%	0.36780%	0.86051%	0.23515%	0.18897%	0.42483%	0.36301%	
		Professional Media	0.34722%	0.17259%	0.35444%	0.81980%	0.26653%	0.10789%	0.34474%		
		General Media	0.33362%	0.21504%	0.35101%	0.81643%	0.25519%	0.10111%	0.34540%		
		Consumer Group	0.33313%	0.21660%	0.38159%	0.81976%	0.25975%	0.11523%	0.35434%		
		Others	0.33429%	0.22045%	0.35751%	0.80747%	0.25447%	0.10007%	0.34571%		
		Mean	**0.37891%**	**0.23500%**	**0.36247%**	**0.82479%**	**0.25422%**	**0.12265%**	**0.36301%**		
Mean			**0.39967%**	**0.25114%**	**0.36326%**	**0.82189%**	**0.25954%**	**0.11628%**	**0.36863%**	**0.36863%**	**0.36863%**

표 4-9 은 VAR 모형을 기반으로 미디어 유형 및 제목과 내용에 따른 주가 예측 오차 성능의 결과이다. 우선 제조사별로는 예측 오차의 범위는 0.12%~0.82% 수준으로 절대적인 기준에서 큰 편차를 보이지 않는데 이는 이번 분석의 대상이 되는 제조사들이 모두 글로벌 기업으로 매우 다양한 미디어 컨텐츠의 소스가 되고 직간접적으로 기업의 주가에 영향을 받는 정도의 편차 역시 크지 않음을 알 수 있다. 미디어 유형에 따른 경향을 보면 일반매체와 사용자그룹 매체의 경우에서 상대적으로 예측 오차가 적게 나타나는 것을 관찰할 수 있는데 컨텐츠의 절대적인 수는 작지만, 상대적으로 압도적으로 많은 구독자를 바탕으로 높은 영향력을 미치는 일반매체 특성과 시장(소비자)의 분위기를 제도적 제약이나 특수관계에서 오는 외부의 영향 없이 가감 없이 전달하는 사용자그룹 미디어의 특성을 고려하여 본다면 이러한 경향을 이해 할 수 있을 것이다. 다음으로 감성분석의 방법을 달리하였을 때 나타나는 해석 성능의 차이도 알아볼 수 있었는데 Scoring 으로 분석한 결과가 Polarity 로 분석한 결과와 비교하여 더욱 높은 성능을 나타내었다. 이는 앞서 2 절에서 다루었던 교차검증에서의 양상과 동일한 것으로 상대적으로 감성 간의 변위가 세밀한 Polarity 방법 보다는 감성을 카테고리화하여 그 방향성을 보다 확실하게 보여주는 Scoring 방법의 결과가 높았다. 끝으로 제목과 내용 간의 해석 성능은 내용을 바탕으로 한 경우 그 오차가 더 작은 것으로 확인되었는데 이를 통하여 상대적으로 감성분석의 재료가 풍부한 내용의 경우가 분석의 재료가 한정적이고 다양한 수사법으로 감성의 방향이 희석될 수 있는 제목의 경우보다 높은 해석력을 제공함을 알 수 있었다. 이러한 데이터 그룹별 분석 결과의 비교를 통하여 절대적인 기준이나 경향을 제시할 수는 없지만, 상대적인 경향성과 각 데이터 그룹이 가지는 특성의 차이를 이해할 수 있었다. 그림 4-6 은 제조사별로 내용 데이터의 Polarity 감성분석의 결과를 가지고 학습한 모델의 테스트 데이터(n=10) 예측 결과를 도식화 한 것이다. 전반적으로 모델의 예측값들이 실제 값들을 잘 추정하여 나타내고 있음을 확인할 수 있다.



[그림 4-6] 제조사별 주가예측 그래프

제 5 장 결론

우리가 처리할 수 있는 데이터의 양보다 더 빠른 속도로 데이터는 증가하고 있으며, 이러한 상황 속에서 내가 필요로 하는 데이터를 선별하고 나아가 주어진 데이터의 가공을 통하여 더욱 가치 있는 데이터로 만들고 활용하고자 하는 요구들은 지속적으로 증가하고 있다. 이는 그 부가가치가 가장 큰 금융시장에서 가장 활발하게 이루어지고 있는데 그 분야 중 하나가 비정형적 데이터를 이용한 주가의 해석이다. 이를 위하여 본 연구에서는 2019 년 1 월 1 일부터 12 월 31 일까지 1 년간 주요 자동차 제조사들과 관련된 온라인 미디어 컨텐츠를 수집하였고 이를 감성분석 기법을 통하여 수치화하고 시계열 데이터화하여 주가 해석의 요소로써 예측에 활용하였다.

수집된 데이터의 모델 적합성을 향상 시키기 위하여 중복데이터와 주제 연관성이 결여된 데이터를 분리하여 감성분석의 희석을 방지하였으며 이전 연구들과 차별성을 위하여 수집된 데이터를 미디어 유형별로 구분하고 더 나아가 제목과 내용을 구분하여 범주화함으로써 각 데이터 그룹이 가지는 감성의 분포나 구상 특성을 확인할 수 있었을 뿐만 아니라 주가 해석의 경향성을 확인할 수 있었다. 또한, 감성분석의 방법으로 편차가 Polarity 와 Scoring 이라는 두 가지의 다른 방법을 도입하여 적용함으로써 세분화된 그룹을 구성하고 결과를 관찰하였다. 온라인 미디어 데이터와 주가와의 관계 분석을 통하여 일차적으로 시행한 교차 검증에서는 회귀 경로와 분류 경로에서 각각의 검증 방법을 달리하여 오차와 정확도를 확인함으로써 데이터 그룹 간의 해석 특성을 도출하였으며 각 변수별 상관관계 분포를 시각화함으로써 그 양상을 탐색하고자 하였다. 2 차로 시행한 예측성능 분석은 비교적 간단하면서도 성능이 우수한 VAR 모형을 기반으로 구축한 모델을 활용하였으며 예측 수행에 앞서 성능향상을 위하여 데이터 정상화, 보조 변수의 활용 그리고 데이터 평활화 기법을 통한 전처리 작업을 수행하였다. 이 모델을 통한 주가 예측에서는 오차 기준 최소 0.08%에서 최대 0.88%의 높은 예측 성능을 보여주었으며, 각 데이터 그룹이 가지는 특성을 확인함으로써 향후 미디어 데이터와 감성분석을 활용한 주가 예측에 있어 방향성을 제시하였다.

본 연구는 앞서 언급한 바와 같이 비정형적 데이터의 활용에 있어 그 범주를 세분화하였을 뿐만 아니라 제목과 내용을 구분하여 접근 방법을 다양화하였으며, 감성분석의 방법에서도 결과의 편차가 다르게 나타나는 방식으로 이원화함으로써 기존의 연구들이 접근하지 않았던 방법을 시도하여 차별성을 추구하였다.

이후의 연구에서는 예측모델을 다양화하여 각 데이터 그룹이 가지는 특성과 예측 모델과의 상관관계를 분석함으로써 예측성능 향상을 위한 조합에 대한 고찰이 필요하며 데이터의 범주를 다양화하여 데이터의 감성이 편향되지 않고 시장을 더욱 잘 대변할 수 있는 구성이 필요하다. 또한 본 연구에서는 다량의 데이터를 수집하였음에도 불구하고 주가와의 시간 배열을 위하여 일 단위 평균을 활용함으로써 표본의 수가 급격하게 줄어들어 모델 생성을 위한 학습 데이터의 양이 다소 충분하지 못하였으나 다음부터는 이를 시간 단위로 구분하거나 시계열 데이터 수집 기간을 늘려 데이터 세트의 크기를 키워 충분한 학습데이터를 확보한 후 모델을 적용한다면 보다 개선된 결과를 얻을 수 있을 것이다.

끝으로 본 연구에서는 이전에 시도되지 않았던 데이터 범주화라는 방법을 도입하여 비정형적 데이터를 활용한 주가 해석을 시도함으로써 향후 이러한 연구의 방향성을 제시함과 동시에 근거를 마련하였으며 이를 바탕으로 다양한 분야의 후속 유사 연구에 활용될 수 있을 것이다. 특히 특정 데이터 그룹이 가지는 특성을 바탕으로 예측 성능을 어떻게 높일 수 있는가에 대한 예시를 보여줌으로써 효율적인 연구수행에 도움이 될 것으로 생각한다.

참 고 문 헌

[1] Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1), 59-82.

[2] Chopra, D., Joshi, N., & Mathur, I. (2016). *Mastering natural language processing with python.* Packt Publishing Ltd.

[3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1-8.

[4] Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PloS one*, 10(9).

[5] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 347-354).

[6] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1-135.

[7] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: *Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media.

[8] Rokade, P. P., & Aruna, K. D. (2019). Business intelligence analytics using sentiment analysis-a survey. *International Journal of Electrical and Computer Engineering*, 9(1), 613.

[9] Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675-684).

[10] Darwich, M., Noah, S. A. M., Omar, N., & Osman, N. A. (2019). Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. *Journal of Digital Information Management*, 17(5), 297.

[11] Parzen, E. (1961). An approach to time series analysis. *The Annals of Mathematical Statistics*, *32*(4), 951-989.

[12] 강창구. (2006). 시계열 예측기법에 대한 비교 분석. *계간 국민계정*, 2006 년 제 3 호, 80-105

[13] Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE transactions on power systems*, *18*(3), 1014-1020.

[14] 문권순. (1997). 벡터자기회귀 (VAR) 모형의 이해. *통계분석연구*, *2*(1), 23-56.

[15] Tsay, R. S. (2005). *Analysis of financial time series* (Vol. 543). John wiley & sons.

[16] Hatemi-j, A. (2003). A new method to choose optimal lag order in stable and unstable VAR models. *Applied Economics Letters*, *10*(3), 135-137.

[17] Maletic, J. I., & Marcus, A. (2009). Data cleansing: A prelude to knowledge discovery. In *Data mining and knowledge discovery handbook* (pp. 19-32). Springer, Boston, MA.

[18] Lindgren, S. (2017). *Digital media and society*. Sage.

[19] Agre, G. H., & Mahajan, N. V. (2015, February). Keyword focused web crawler. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*(pp. 1089-1092). IEEE.

[20] 김경모. (2012). 온라인 뉴스 확산과 여론 형성. *언론과학연구*, *12*(4), 35-72.

[21] Biagi, S. (2012). *Media impact: An introduction to mass media*. Cengage Learning.

[22] 배정근. (2012). 대기업 광고주가 자사 신문기사에 미치는 영향에 대한 기자 인식 연구: 종합일간지 경제· 산업부 기자들을 중심으로. *한국언론학보*, *56*(5), 373-396.

[23] 최영, & 박창신. (2009). 온라인 뉴스이용에 관한 연구: 조선닷컴 뉴스의 연성화 및 제목 선정성과 조회수간의 상관관계. *커뮤니케이션학 연구*, *17*(1), 31-53.

[24] 최영, 박창신, & 고민경. (2010). 온라인 뉴스의 제목달기 분석: 네이버 뉴스캐스트의 헤드라인 특성을 중심으로. *커뮤니케이션학 연구*, *18*(1), 115-140.

[25] Pourahmadi, M. (1989). Estimation and interpolation of missing values of a stationary time series. *Journal of Time Series Analysis*, *10*(2), 149-169.

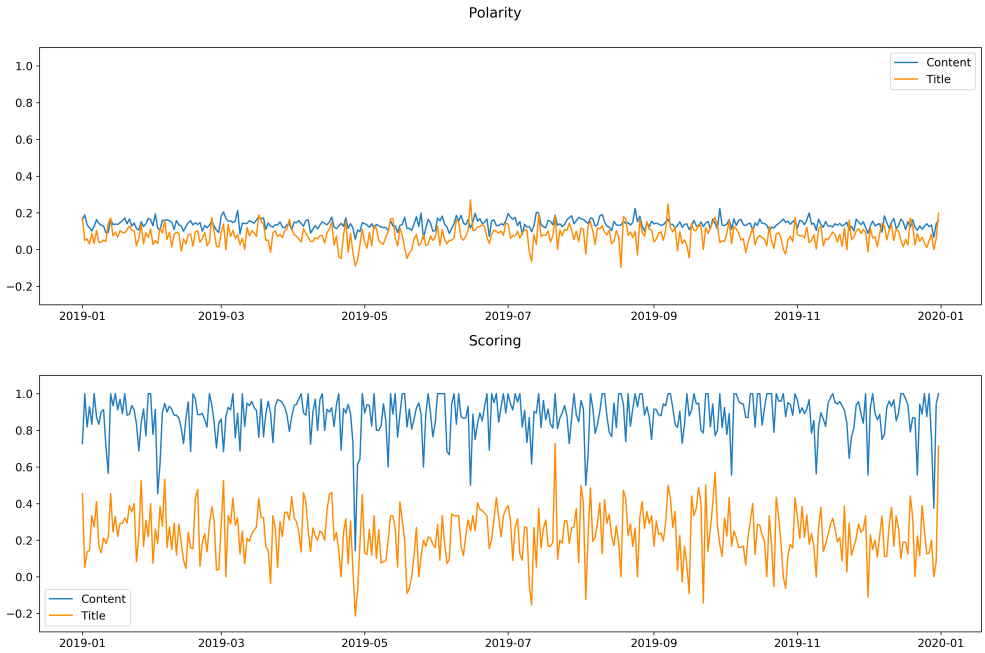
[26] 박수민, 이일주, & 이용길. (2017). 국제유가 변동이 가계의 세부 소비구조 변화에 미치는 영향에 대한 고찰: 벡터자기회귀모형 (VAR) 을 이용하여. 한국혁신학회지, 12(1), 143-166.

[27] 윤지성, 허남균, 김삼용, & 허희영. (2010). 계절형 다변량 시계열 모형을 이용한 국제항공 여객 및 화물 수요예측에 관한 연구. 한국통계학회논문집, 17(3), 473-481.

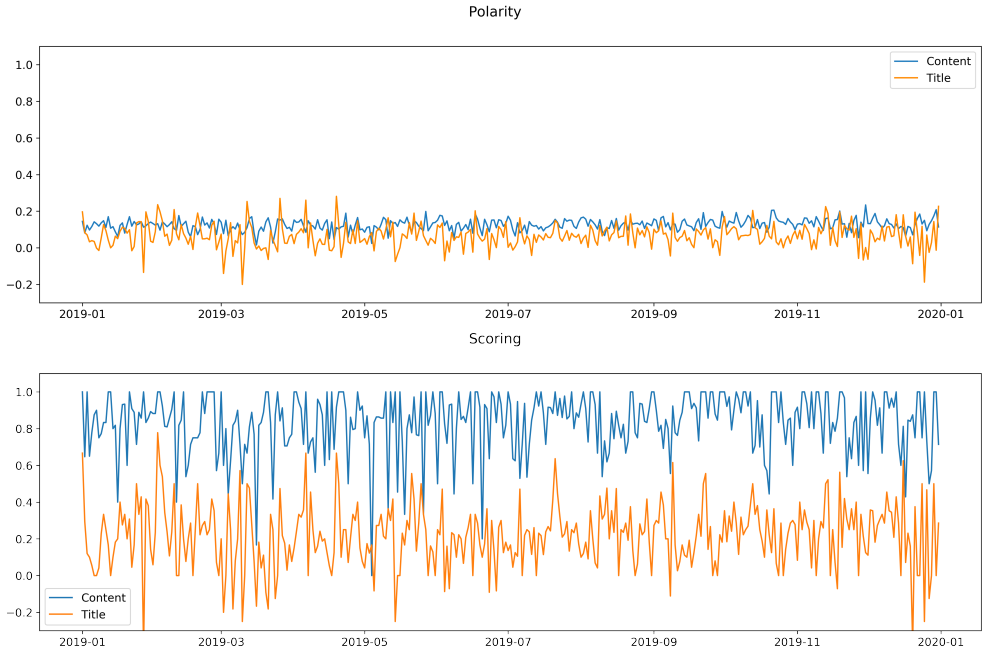
부록 (Appendix)

A1. 제조사별 Polarity/Scoring 방법에 따른 시계열 감성분포

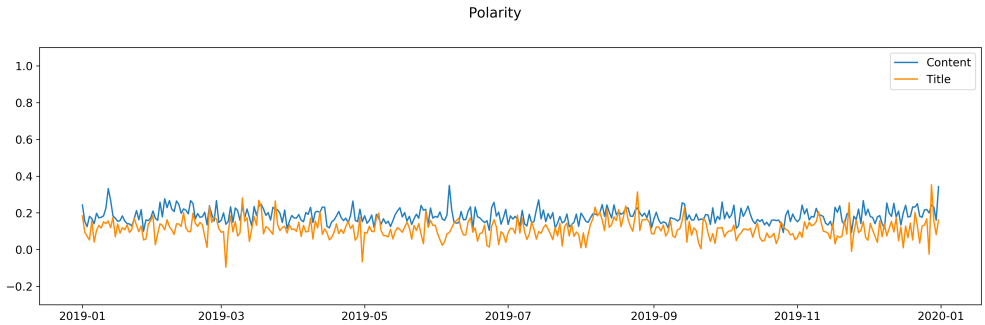
1) FORD

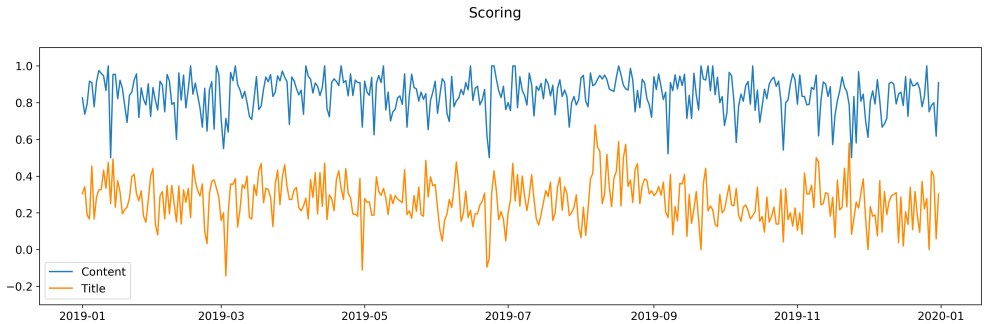


2) GM

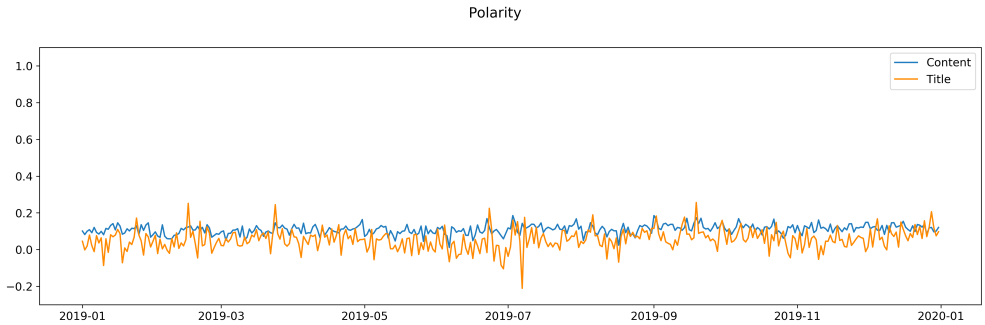


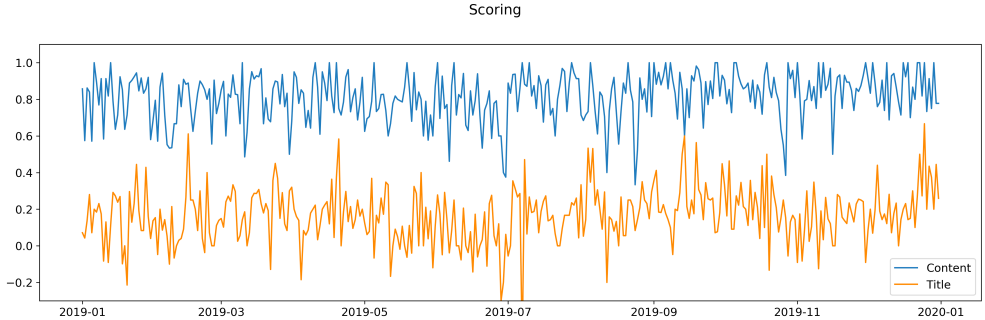
3) HYUNDAI



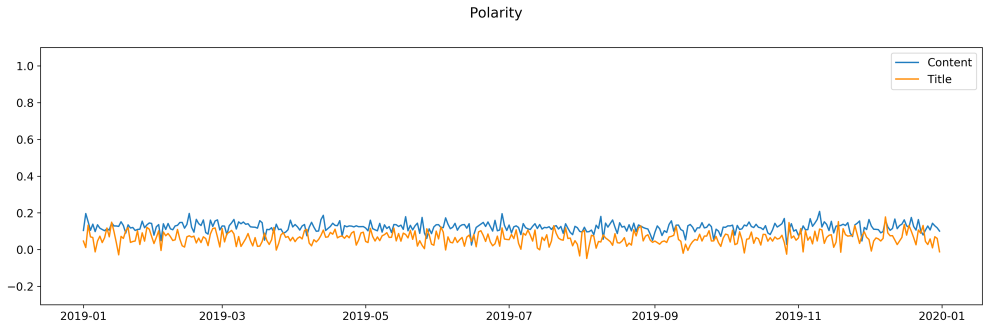


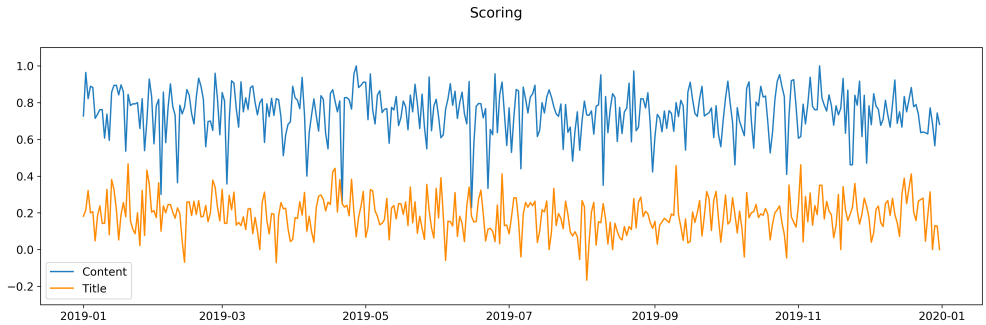
4) TESLA



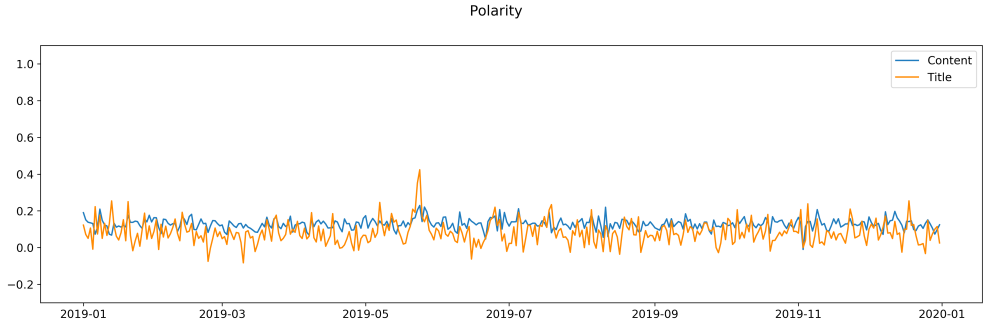


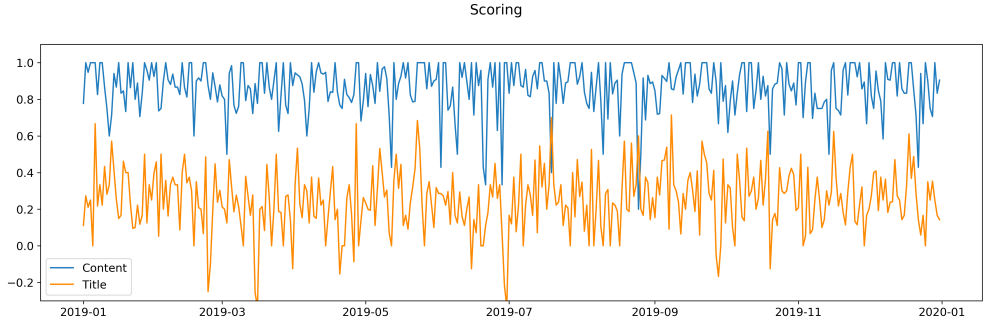
5) TOYOTA


Polarity
Content
Title
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01


Scoring
Content
Title
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01

6) VW


Polarity
Content
Title
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01


Scoring
Content
Title
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01

A2. 제조사별 감성분석결과 데이터 간 상관관계 및 표준편차

1) FORD

Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.450392	0.316759	0.289415	0.025339
	Scoring	0.450392	1	0.180023	0.320906	0.114689
Title	Polarity	0.316759	0.180023	1	0.801896	0.051276
	Scoring	0.289415	0.320906	0.801896	1	0.138513

2) GM

Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.596481	0.159431	0.133885	0.031709
	Scoring	0.596481	1	0.14489	0.15362	0.165219
Title	Polarity	0.159431	0.14489	1	0.826144	0.066017
	Scoring	0.133885	0.15362	0.826144	1	0.175271

3) HYUNDAI

Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.265144	0.368107	0.211838	0.039499
	Scoring	0.265144	1	0.172376	0.317593	0.09997
Title	Polarity	0.368107	0.172376	1	0.756798	0.052784
	Scoring	0.211838	0.317593	0.756798	1	0.118271

4) TESLA

Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.632692	0.311676	0.253701	0.025551
	Scoring	0.632692	1	0.185831	0.214331	0.132551
Title	Polarity	0.311676	0.185831	1	0.805064	0.055927
	Scoring	0.253701	0.214331	0.805064	1	0.153163

5) TOYOTA

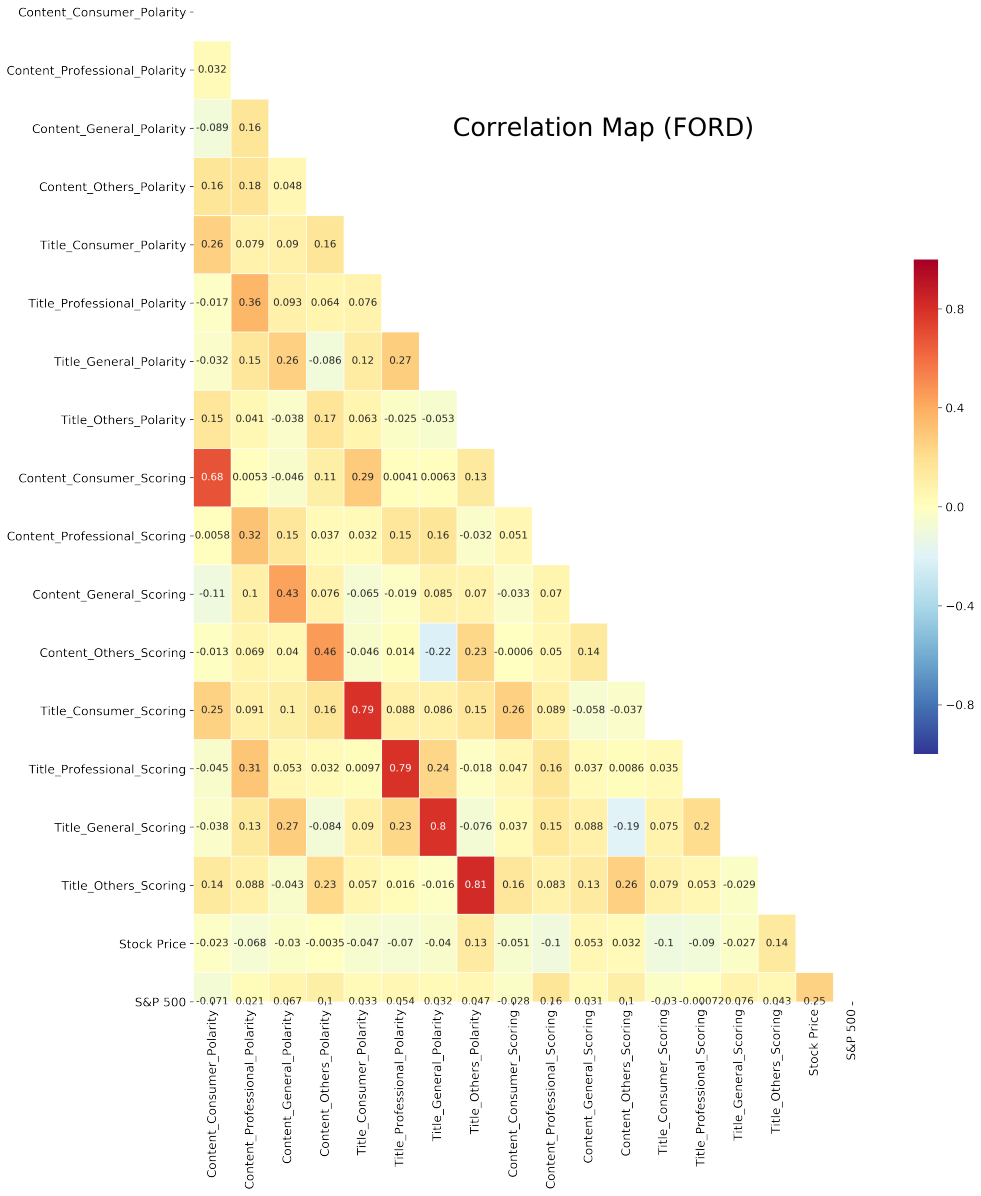
Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.634478	0.176339	0.175712	0.026003
	Scoring	0.634478	1	0.124141	0.189623	0.127747
Title	Polarity	0.176339	0.124141	1	0.780614	0.034376
	Scoring	0.175712	0.189623	0.780614	1	0.100644

6) VW

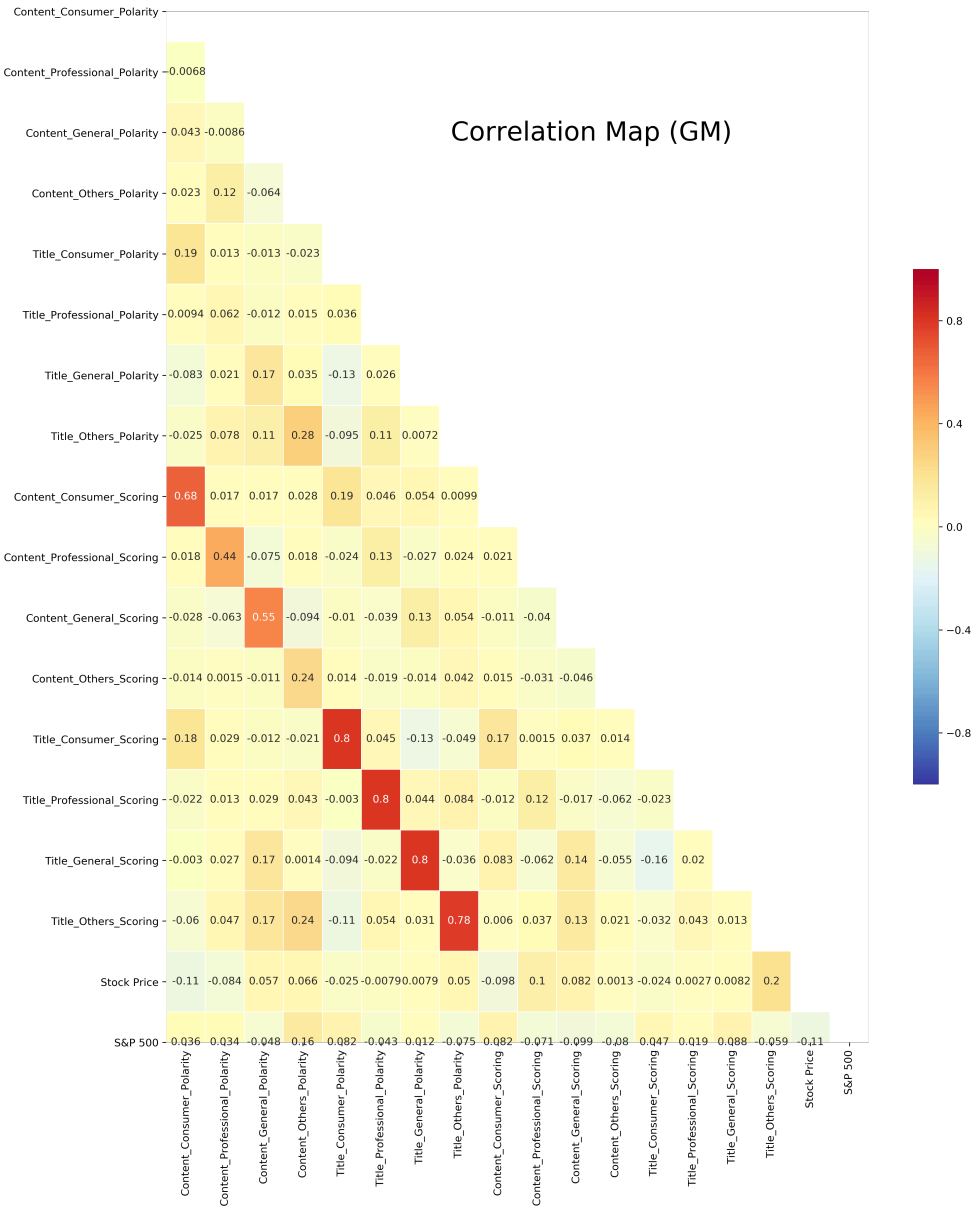
Correlation		Content		Title		Standard Deviation
		Polarity	Scoring	Polarity	Scoring	
Content	Polarity	1	0.511905	0.399849	0.215964	0.02962
	Scoring	0.511905	1	0.152527	0.130139	0.136028
Title	Polarity	0.399849	0.152527	1	0.731764	0.062594
	Scoring	0.215964	0.130139	0.731764	1	0.171084

A3. 제조사별 감성분석과 주가 데이터 간 상관관계 분포

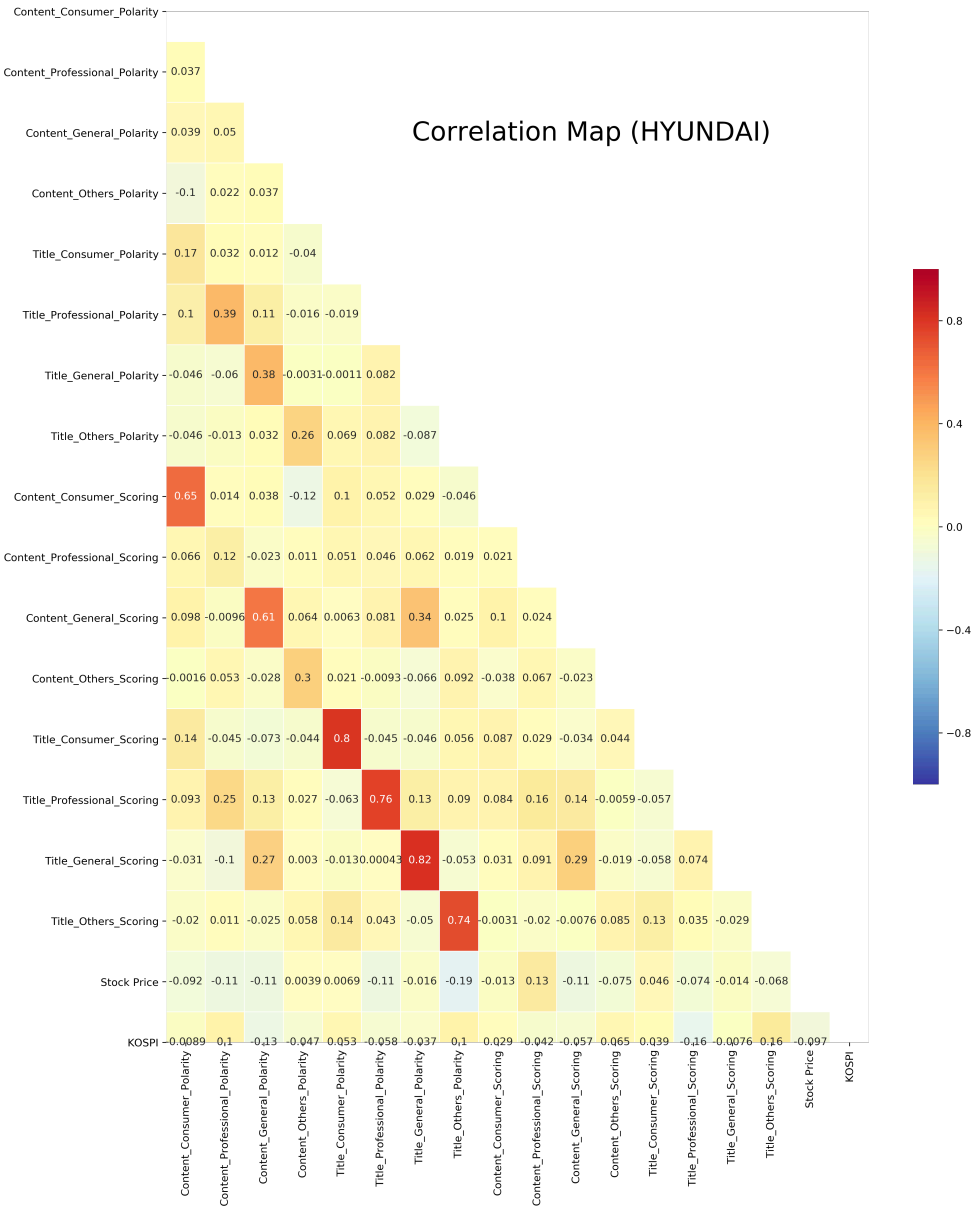
1) FORD



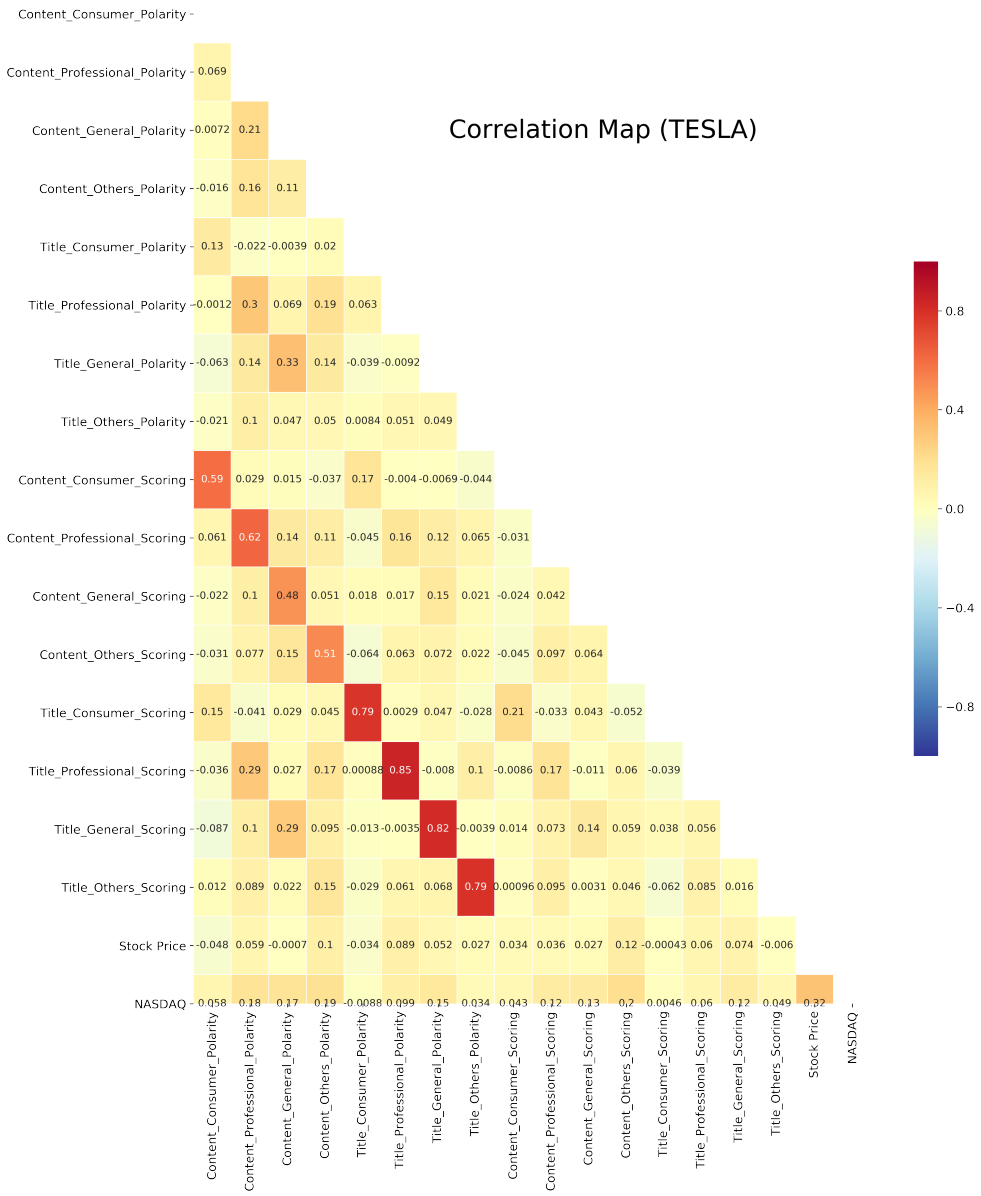
2）GM


Correlation Map (GM)
Content_Consumer_Polarity
Content_Professional_Polarity
Content_General_Polarity
Content_Others_Polarity
Title_Consumer_Polarity
Title_Professional_Polarity
Title_General_Polarity
Title_Others_Polarity
Content_Consumer_Scoring
Content_Professional_Scoring
Content_General_Scoring
Content_Others_Scoring
Title_Consumer_Scoring
Title_Professional_Scoring
Title_General_Scoring
Title_Others_Scoring
Stock Price
S&P 500
0.8
0.4
0.0
−0.4
−0.8

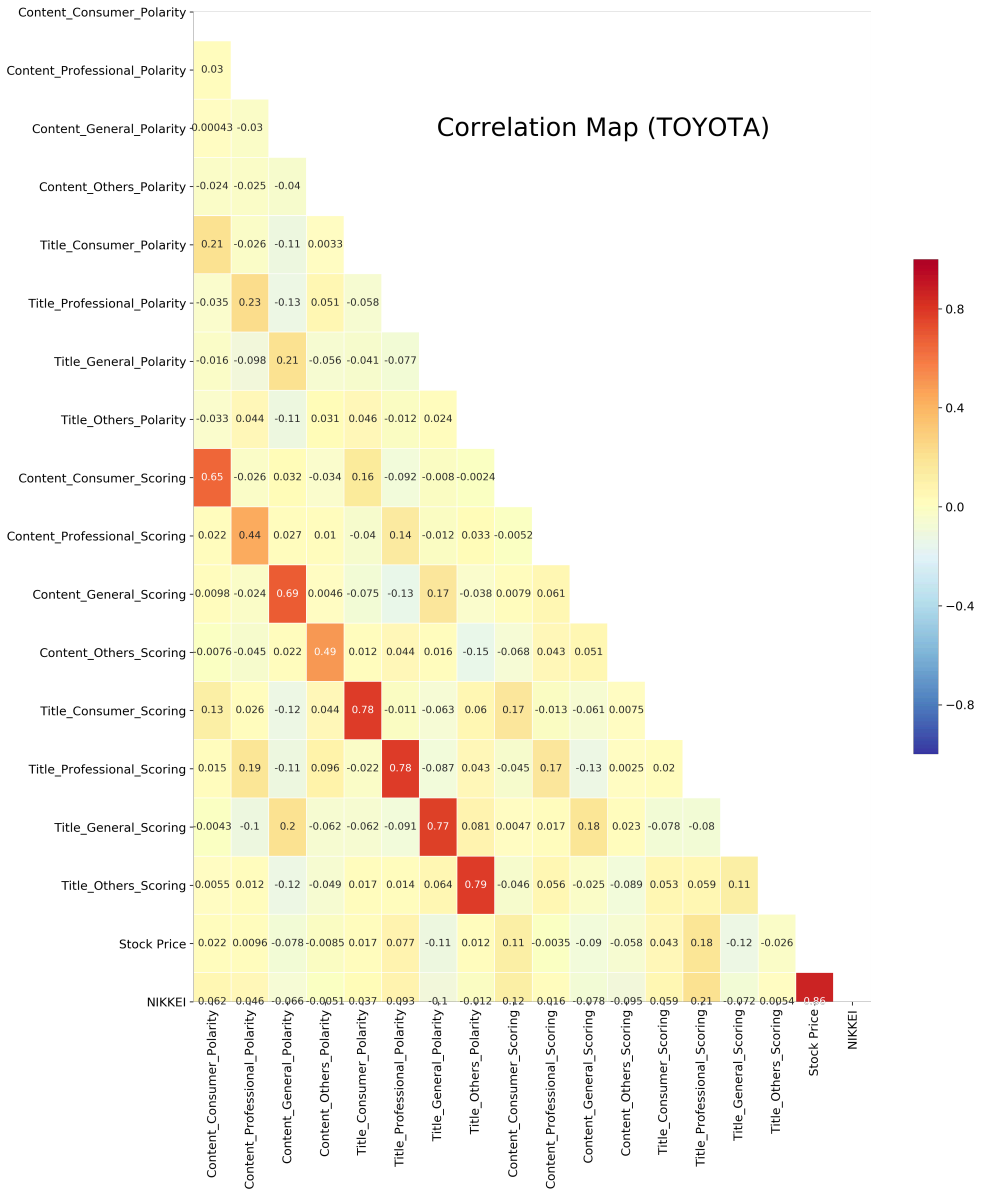
3) HYUNDAI


Correlation Map (HYUNDAI)
Content_Consumer_Polarity
Content_Professional_Polarity
Content_General_Polarity
Content_Others_Polarity
Title_Consumer_Polarity
Title_Professional_Polarity
Title_General_Polarity
Title_Others_Polarity
Content_Consumer_Scoring
Content_Professional_Scoring
Content_General_Scoring
Content_Others_Scoring
Title_Consumer_Scoring
Title_Professional_Scoring
Title_General_Scoring
Title_Others_Scoring
Stock Price
KOSPI
0.8
0.4
0.0
−0.4
−0.8

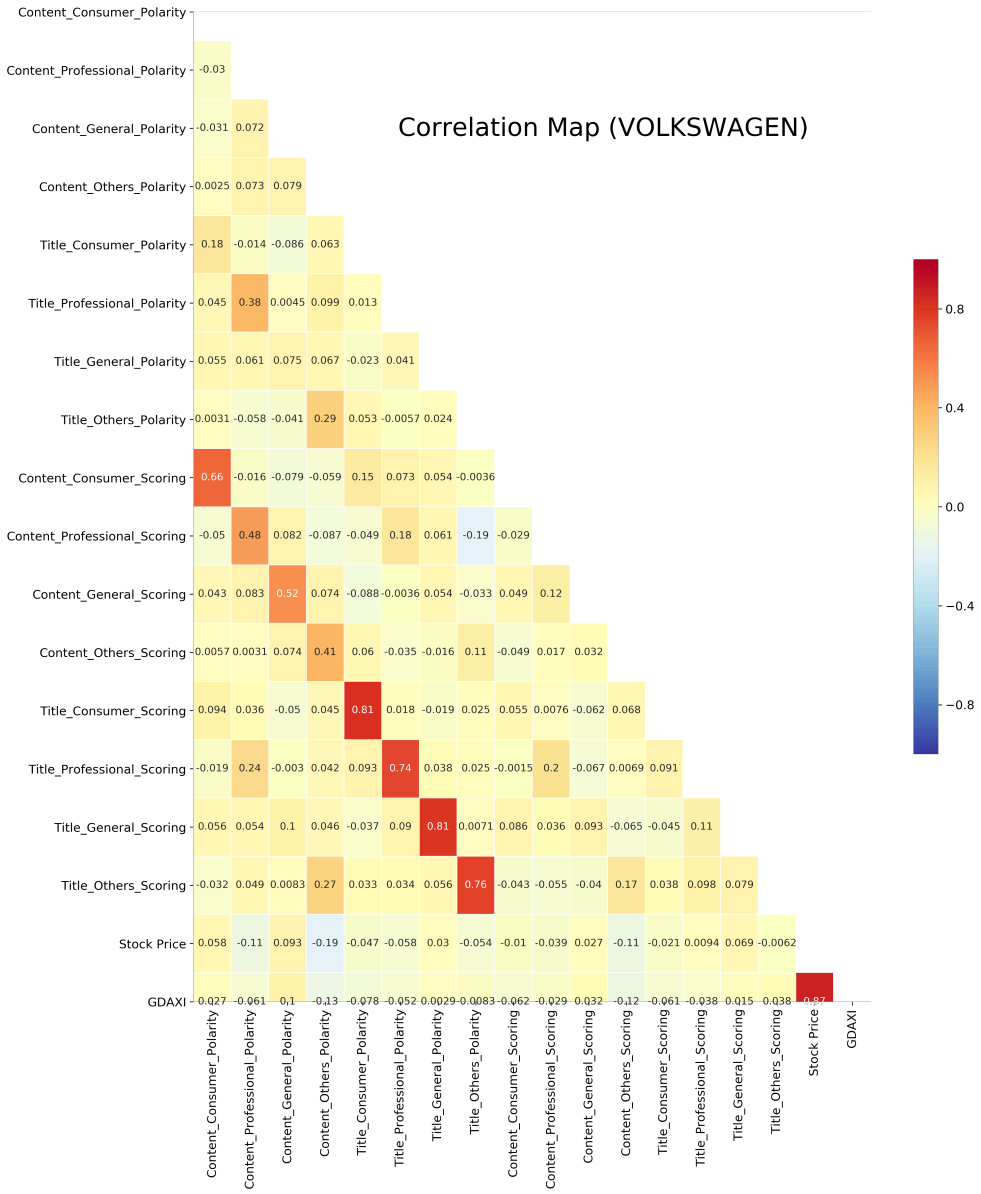
4) TESLA


Correlation Map (TESLA)
Content_Consumer_Polarity
Content_Professional_Polarity
Content_General_Polarity
Content_Others_Polarity
Title_Consumer_Polarity
Title_Professional_Polarity
Title_General_Polarity
Title_Others_Polarity
Content_Consumer_Scoring
Content_Professional_Scoring
Content_General_Scoring
Content_Others_Scoring
Title_Consumer_Scoring
Title_Professional_Scoring
Title_General_Scoring
Title_Others_Scoring
Stock Price
NASDAQ
0.8
0.4
0.0
−0.4
−0.8

5) TOYOTA


Correlation Map (TOYOTA)
Content_Consumer_Polarity
Content_Professional_Polarity
Content_General_Polarity
Content_Others_Polarity
Title_Consumer_Polarity
Title_Professional_Polarity
Title_General_Polarity
Title_Others_Polarity
Content_Consumer_Scoring
Content_Professional_Scoring
Content_General_Scoring
Content_Others_Scoring
Title_Consumer_Scoring
Title_Professional_Scoring
Title_General_Scoring
Title_Others_Scoring
Stock Price
NIKKEI
0.8
0.4
0.0
−0.4
−0.8

6) VW


Correlation Map (VOLKSWAGEN)
Content_Consumer_Polarity
Content_Professional_Polarity
Content_General_Polarity
Content_Others_Polarity
Title_Consumer_Polarity
Title_Professional_Polarity
Title_General_Polarity
Title_Others_Polarity
Content_Consumer_Scoring
Content_Professional_Scoring
Content_General_Scoring
Content_Others_Scoring
Title_Consumer_Scoring
Title_Professional_Scoring
Title_General_Scoring
Title_Others_Scoring
Stock Price
GDAXI
0.8
0.4
0.0
−0.4
−0.8

Abstract

Machine Learning Based Car Manufacturer Stock Price Analysis by Online Media Sentiment Analysis

Minsoo Park
Department of Engineering Practice
Graduate School of Engineering Practice
Seoul National University

Long before implementing the intermediation of value measurement called currency, mankind has been trading various types of goods on the major premise of the exchange of equivalent, and from that, the measurement of the proper value has been an essential element of economic activity. According to these market demands, value evaluation methods for various goods have been researched, in particular, since the world's first stock exchange was established in Amsterdam, the Netherlands in 1602, the valuation of companies, that is, the valuation of stocks has become one of the most actively researched field with the explosive growth in the stock market. However, most stock valuation methods that have been developed in the traditional market economy system are focused on numerical analysis, and although many of these numerical interpretations provide adequate and excellent analysis results, they cannot deal with qualitative data and this is pointed out as a limitation of numerical interpretation.

The purpose of this study is to provide a method for better analytic results with the implementation of machine learning-based

sentiment analysis made possible by the advances in software and hardware technologies and to digitize qualitative data and use this get more accurate stock analysis. To this end, six global automakers were selected to collect online media data generated during the year of 2019 based on keywords, and to understand more detailed analysis media data was collected in and divided into four categories of professional media, general media, consumer group, and others, and each media was analyzed by dividing it into titles and contents. The characteristics of collected media materials were identified through data featuring and after basic pretreatment they were divided into regression path and classification path to identify the analysis trends in stock value prediction. In the regression path, linear regression and random forest regressors are used to measure RMSE (Root Mean Square Error) against the movement of the base stock price. In the classification path, the logistic regression and random forest classifier are used to compare the null model of the existing stock price to grasp each trend by checking the prediction accuracy. Afterward, a prediction model using a vector autoregression (VAR) model is constructed for practical stock price analysis, and the prediction performance is evaluated by classifying each media type and the title and its contents. After which the group or category's characteristics were analyzed.

In this study, rather than comprehensively approaching the movement of the stock market through sentiment analysis, we analyzed the media data by classifying the media data of representative companies belonging to the industry, and further analyzing the differences between the title and content. By analyzing them together, we suggested the direction of data selection and the necessity of data categorization in future market analysis research through emotional analysis.

Keywords : Stock Price Analysis, Machine Learning, Sentiment Analysis, Car Manufacturer, Online Media Category
Student Number : 2018-22612