텍스트 마이닝 기법을 활용한 텍스트 종류 연구*

- 한국과 독일증시 마감보고서를 비교하면서 -

방경원**

목 차

〈국문초록〉

본 연구의 목적은 증시 마감보고서의 상황에 따라 특정 어휘군인 *오르다*와 *내리다*의 빈도수 순위에 차이가 있다는 것을 실증적으로 확인하는데 있다. 이는 어휘 단위에서도 텍스트종류를 특징짓기 위해서다. 어휘는 텍스트 내에서 실제 확인할 수 있는 중요한 자료임에도 그간 연구에서 소홀히 다루어졌다. 텍스트를 어휘 수준에서 분석하기에는, 시간과 노력에 한계가 있었기 때문이다. 빅데이터를 다루는 텍스트마이닝 기법은 이런 수고를 덜어줄 것이다. 그리고 앞으로도 어휘 단위에서의 텍스트 연구가 더욱 활성화 되리라 판단된다.

본 연구에서 사용된 자료는 독일편에서는 Handelsblatt의 온라인 판과 인터넷 매체인 Börse Online에서, 국내편에서는 Daum의 금융뉴스에서 발췌했다. 증권사 증시 월봉 차트를 참고하여 2017년 중반이후 연말까지 상승구간, 2018년 전 구간을 하락구간으로 정하였다. 단어구름과 통계자료 작업에는 R과 AntConc 프로그램이 사용되었다.

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A5B5A07057946)

** 한국외국어대학교

증시 마감보고서에 공통으로 등장하는 최다 빈도 어휘는 추세가 변해도 순위에 큰 차이를 보이지 않았으나, *오르다*와 *내리다*의 어휘군은 시간이 흐르면 빈도수 순위에 차이가 있음을 확인할 수 있었다. 따라서 상황 변동에 민감한 텍스트종류에서, 특정 어휘군의 빈도 순위는 상황 변화에 유동적이라고 판단된다. 이런 어휘는 텍스트종류를 특징짓는데도 유용할 것이다.

주제어 : 텍스트종류, 증시 마감보고서, 증시의 등락, 텍스트마이닝, 기계학습

1. 서론

본 연구의 목적은 증시 추세에 따라 증시 마감보고서에 나오는 특정 어휘군인 *오르다*와 *내리다*의 빈도 순위에 차이가 있음을 실증적으로 확인하는데 있다. 또한 이들 어휘가 핵심 어휘로 자리 잡게 되면 텍스트종류를 특징짓는데 기여할 것으로 판단된다.

주식 시장의 등락에 관한 연구는 주로 회계학의 관심 분야였다. 그리고 회계학 쪽에서의 연구는 정형적인 데이터에 기반 한 주가와 주가지표 간의 상관관계를 찾는 회기분석이 대부분이었다. 그러나 정보의 양이 커지면서 이를 수용해야 하는 필요성이 커지고, 빅데이터 처리기술도 개발되면서, 비정형 데이터인 뉴스와 주가와의 관계도 연구하게 되었다(김유신・김남규・정승렬, 2012). 빅데이터는 주로 텍스트 형태로 되어있기 때문에 이를 정형 데이터로 전환시켜야 하며, 이때 텍스트마이닝 기법을 사용하게 된다.

회계학에서도 텍스트마이닝 기법을 활용할 때, 자료를 모으고 이를 정제하여 정형 데이터로 만드는 과정은 텍스트를 빅데이터로 처리하는 모든 연구에 공통된다. 텍스트언어학에서도 텍스트마이닝 기법을 활용할 때 예외는 아니다. 이미 코퍼스언어학에서 이 과정에 대한 연구 성과가 축적되어 있기 때문에 오히려 유리하다. 다만, 단어와 단어빈도의 행렬로 된 정형 데이터로 전환하는 과정이 추가되는데, 이는 전산학자와 통계학자들의 몫이

기 때문에, 이들이 개발한 기술을 활용할 수 있으면 된다. 이 기술도 사용자에게 친숙하게 진화되고 있어, 여러 분야에서 자유롭게 이용되고 있다.

본 연구에서도 증시등락과 같은 주제를 다루지만, 회계학에서와는 다르게 주가 예측을 목표로 하지는 않는다. 이는 자료 선택에 있어서도 다양한 금융뉴스가 아니라 증시 마감보고서의 한 텍스트종류에서 알 수 있다. 그리고 선행적으로 주가를 예측하기 보다는, 이미 증시 등락에 따라 보고된 자료에 충실하게 서술하는 것을 목표로 한다. 이로써 텍스트종류의 특성을 찾아내는 데에 주력한다.

텍스트종류를 기술하는 모델로서 아담치크는 언어구조물, 상황적인 맥락, 주제, 기능의 차원으로 이루어진 다차원 모델을 소개한다(Adamzik, 2004: 59). 언어구조물은 상황적인 맥락(Wo), 주제(Was), 기능(Wozu)과 상호 작용을 하는 관계로, 각 차원 사이에 수직 계열화를 거부한다. 언어구조물 내의 각 기호 간의 결속성(Kohäsion)도 텍스트의 응집성(Kohärenz)과 명확히 구분되지 않는다. 이는 텍스트종류가 언어구조물로 실현될 때 각 차원에서 자유롭게 소통되는 것을 가정하기 때문이다. 어휘는 언어구조물을 구성하는 구체적인 단위로, 텍스트종류를 기술하는데 실체적인 기반이 된다.

본 논문은 서론에 이어, 2장 선행연구 소개, 3장 연구대상과 연구방법 설명, 4장 연구대상 분석, 5장 결론으로 구성된다.

2. 선행연구

텍스트마이닝 연구는 현재 여러 분야에서 시도되고 있다. 텍스트언어학 쪽에서도 텍스트마이닝 연구 기법을 활용할 수 있도록, 본 연구와 관련하여 비교적 활성화되어 있는 경제관련 텍스트 연구의 텍스트마이닝 연구기법을 소개하면서 상호 교류를 꾀하고자 한다.

경제관련 자료는 주로 수치로 된 정형 데이터 중심이었으나, 정보가 폭발

적으로 증가하면서 비정형 데이터가 압도하게 되고, 이에 대한 연구수요도 증가하게 되었다. 신문기사를 텍스트마이닝하여 주가를 예측하고(안성원 · 조성배, 2010), 미국기업의 연차보고서를 텍스트마이닝하여 서비스화 동향 정보를 추출하는 것이(이지환 · 홍유석, 2013) 그 예이다. 최근에도 이런 시도는 계속 이어지고 있다. 신공항 신문기사를 텍스트마이닝하여 사회적 이슈를 파악하면서 정책에 반영한다든가(한무명초 · 김양석 · 이충권, 2017), 우정사업 관련 기사를 텍스트마이닝하여 전 세계 트렌드를 파악하고 이를 국내 우정정책에 반영하는 예이다(김홍림 · 김민진, 2019).

이런 시도는 주로 회계학과 경영학 쪽의 사례이지만, 텍스트자료에 기반했기 때문에 텍스트언어학과도 공유할 부분이 있다. 이 부분은 비정형 데이터를 정형 데이터로 전환시키는 작업이다. 이는 텍스트의 전처리 과정에 속하는데, 의미 있는 언어기호를 추출하기까지이다. 이렇게 추출된 기호는 단어빈도와 함께 행렬로 정돈될 때, 비로소 텍스트분석을 시작할 수 있다. 이후에는 연구자의 전문영역에 따라 분석방향이 달라진다. 한국은행 경제연구소에서는 경제 분석에 텍스트마이닝을 더욱 적극적으로 사용할 수 있도록, 연구용역을 통해 텍스트마이닝 방법론과 분석사례를 발표시키고 있다(김수현 외, 2019).

텍스트마이닝의 중심에는 단어가 있다. 텍스트언어학에서 다루는 텍스트는 여러 층위를 가정하고 있으며, 단어는 텍스트 내적인 요소로 구체적으로 확인할 수 있다. 텍스트종류 연구에서도, 텍스트마이닝 기법을 활용하여, 텍스트 내적인 요소로 단어를 중심으로 한 연구를 활성화시킬 수 있다. 유의미한 단어를 추출하는 텍스트 전처리 과정은 코퍼스언어학에서 축적된 연구 성과와도 크게 다르지 않기 때문이다. 문법에서 기능어로 구성된 불용어(stop words), 어간추출(stemming), 표제어추출(lemmatization), 품사태깅(POS tagging), 형태소 분석(morpheme analyzing), N-gram 등은 텍스트 전처리 과정에서 자주 등장하며, 코퍼스언어학에서도 이미 다루었던 분야이다. 텍스트 전처리 과정에서 추출된 유의미한 단어와 단어빈도로 구성

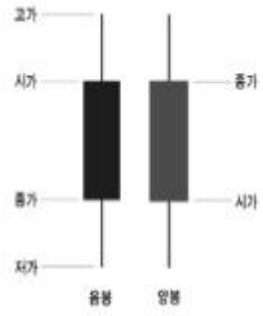
된 행렬(Matrix)은 컴퓨터가 계산할 수 있는 자료로 전환되어야 하며, 이에는 통계학자와 컴퓨터 전문가들이 개발해 놓은 수식과 알고리즘이 관여된다. 따라서 텍스트마이닝은 언어학, 전산학, 통계학 등이 함께 하는 다학제(multi-disciplinary) 연구 영역이다(김수현 외, 2019: 5).

텍스트마이닝의 다학제적 성격에서 볼 때, 본 연구에서 어휘에 집중하면서 텍스트마이닝 기법을 더 효과적으로 활용할 수 있는 계기가 될 것이다. 유의미한 어휘추출은 텍스트마이닝의 결과에 결정적인 영향을 미치기 때문이다. 또한, 어휘 연구는 텍스트의 내적요소를 유형화하는데 기여할 것이며, 이는 텍스트종류 연구의 일부로 자리매김할 수 있다. 증시마감보고서에 나오는 주가 등락에 관한 어휘는 주가를 예측하는 자료로 활용될 수 있지만, 본 연구에서는 텍스트종류를 특징짓는 데 한정한다. 주가가 결정되는 데에는 여러 요인이 있기 때문에, 주가등락 어휘만으로 주가를 예측한다는 것은 너무 성급한 결정이기 때문이다.

증시보고서의 언어학적 연구 사례로는 전문어로서의 증시언어 연구(Březina, 2014), 증시언어에 나타나는 은유 연구(Eitze, 2012; 방경원, 2015), 증시 마감보고서를 포괄적으로 텍스트종류로 특징짓는 연구(방경원, 2018)가 있었다. 기존의 어휘 연구는 주로 은유 표현이었다. 본 연구에서는 어휘를 중점적으로 다루면서 증시 마감보고서를 미시적으로 다루고자 한다.

3. 연구대상 및 연구방법

연구대상인 증시의 *오름*과 *내림*에 관련된 어휘를 분석하기 전에, 추세 구간을 확인하기 위해서 증권사의 월봉차트를 소개한다. 월봉차트의 이해를 돕기 위해 먼저 봉의 구조를 설명한다.



〈그림 1〉 음봉과 양봉의 구조

월봉은 봉의 몸통 기준으로 시초가인 월초의 지수 가격과 종가인 월말의 지수 가격을 표시해준다. <그림 1>에서 오른쪽 빨간 봉은 양봉으로, 시초가는 몸통의 아래를, 종가는 몸통의 위를, 왼쪽 파란 봉은 음봉으로, 시초가는 몸통의 위를, 종가는 몸통의 아래를 가리킨다. 상승 시에는 양봉이, 하락 시에는 음봉이 우세하다.

<그림 2>는 DAX30의 월봉 차트이고, <그림 3>은 Kospi200의 월봉 차트이다.

〈그림 2〉 DAX30 추이(2017~2018, 출처: 키움증권 해외증시차트)

독일 증시는 <그림 2>에서 보듯이 DAX30 기준으로 2017년 12월까지 상승추세이다가, 2018년 초 하락세로 전환되어, 2018년 11월까지 이어진다.

〈그림 3〉 Kospi200 추이(2017~2018, 출처: 키움증권 업종종합차트)

한국 증시도 <그림 3>에서 보듯이 Kospi200 기준으로 2017년 12월까지 상승추세이다가, 2018년 1월 하락세로 전환되었으며, 2018년 9월까지 지속되었다. 본 연구에서는 상승과 하락구간 중 자료가 확보된 2017년 8월 초부터 2018년 말까지의 기간을 조사 기간으로 정했다. 따라서 큰 추세로 2017년은 상승구간으로 2018년은 하락구간으로 나누어 자료를 분석했다. 그러나 상승추세 중에도 눌림 구간이 하락추세 중에도 반등 구간이 있기 때문에, 세밀한 관찰을 위해 추세의 최고점과 최저점 기준의 월 단위 분석도 병행한다. <그림 2>에서 알 수 있듯이 독일 증시에서 최고점은 2018년 1월이고, 최저점은 2018년 12월이다. <그림 3> 한국 증시에서도 최고점은 2018년 1월이고, 최저점은 2018년 10월이다.

한국 증시 마감보고서는 Daum의 금융뉴스에서[1], 독일 증시 마감보고서는 Handelsblatt 디지털판과[2] 인터넷 증권매체인 Börse Online에서[3] 발췌

1) http://finance.daum.net/news/news_list.daum?limit=30&section=&type=market

2) Handelsblatt Digital 판 :

했다. 수집된 자료가 임의적이지만 텍스트종류로서 요건은 갖추고 있기 때문에 대표성은 유지된다.

연구방법은 텍스트마이닝 작업 순서에 따라, 원시 데이터를 정제하여 유의미한 단어를 추출하고, 이 단어들의 빈도를 중심으로 유용한 정보를 유추하는 과정으로 구성된다. 분석도구로는 R과 AntConc 프로그램이 이용되며, R을 통하여 단어구름(Word Cloud)을 그려내고, AntConc를 통하여 단어의 빈도를 분석한다.[4]

4. 연구대상 분석

독일 증시와 한국 증시에서 자주 나타나는 어휘로 분리하여 분석하면서 비교한다. 먼저 단어구름을 통하여 상승과 하락추세 전체를 개관하고, 다음 자세한 관찰을 위해 최다 빈도 상위 어휘군의 통계치를 막대그래프로 시각화하면서, 상승추세와 하락추세 전체 구간과 최고점과 최저점으로 나누어 비교해 보았다.

4.1. 독일증시에서 자주 나타나는 어휘

단어구름으로부터[5] 시작하는 것은 언어 통계를 한 눈에 요약할 수 있는 시각적인 효과가 있기 때문이다.

---

http://www.handelsblatt.com/finanzen/maerkte/?navi=FINANZEN_SUBNAV_M%C3%84RKTE_1980480

3) 인터넷 증권매체 Börse Online : https://www.boerse-online.de/maerkte/marktberichte

4) 독일어나 한국어는 특수문자로 취급되어, R에서는 여러 정제 과정을 거치는 중에 종종 깨어지기 때문에, 이를 보완하는데, 다국어를 안정적으로 처리하는 AntConc 프로그램이 이용된다. 또한 AntConc에서는 단어만 추출되기 때문에 숫자, 문장부호, 특수기호 등을 따로 제거할 필요가 없다.

5) R프로그램인 tm-패키지를 사용하였다.

2017년 상승추세 2018년 하락추세

〈그림 4〉 독일 증시 2017년 상승추세와 2018년 하락추세 간 단어구름 비교

<그림 4>는 조사 대상인 2017년과 2018년 독일 증시 마감보고서의 단어 자료를 사용하여 작성된 것이다. <그림 4>의 왼쪽은 상승추세인 2017년 8월부터 12월까지의 자료 중, 정제 과정을 거쳐 유의미한 단어를 추출한 다음 100회 이상 빈도수를 가진 단어 형태로 구성되었고, <그림 4>의 오른쪽은 하락추세인 2018년 1월부터 12월까지 자료 중 같은 과정을 거쳐 100회 이상 빈도수를 가진 단어 형태로 이루어졌다. 중앙에 있는 글자 크기 별로 빈도수가 높은 것을 알 수 있다.

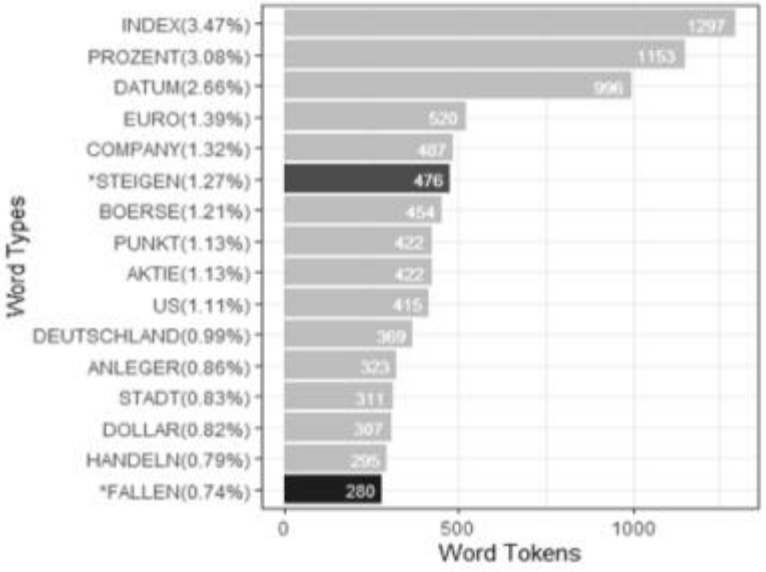
단어구름 만으로는 양 추세 간 차이를 구분하기가 쉽지 않다. 이를 보완하기 위해서 다음에서는 단어 빈도를 더 구체적으로 분석했다.[6] 통계 수치는 시각화를 통해 이해도를 높였다.[7] 단어구름에서는 개별 단어들의 빈도수만 처리했지만, 다음 그림에서는 유사한 의미를 갖는 어휘를 함께 묶어서 다루었기 때문에, 순위에서 약간의 차이가 있을 수 있다. 예를 들어, 단어구름에서는 Prozent가 빈도수가 제일 커서 중앙에 위치하고 있으나, <그림 5>에서는 INDEX에게 순위가 밀려나고 있다. 이는 INDEX에는 index뿐만 아니라 dax, mdax, sdax, tecdax, leitindex 등의 빈도수도 합쳤기 때문이다.

6) AntConc 프로그램을 사용하였다.

7) 시각화에는 R의 ggmap이나 ggplot2와 같은 패키지가 활용되었다.

그러나 전체 큰 흐름을 파악하는 데에는 변함이 없다.

<그림 5>와 <그림 6>에서 기초한 언어자료는 2017년 8월에서 12월까지 독일경제지 Handelsblatt 디지털판에서 수집되었다. 이 자료로부터 전체 67,162 Word Tokens(이하 토큰으로 명명함)와 8,996 Word Types(이하 타입으로 명명함)가 조사되었다. 이들 중에서, 불용어들을 제외한 유의미한 부분만 택하면 37,391 토큰과 7,653 타입이 추출된다.[8)]



〈그림 5〉 상승추세인 2017년도 하반기 독일 증시 최다 빈도 언어통계

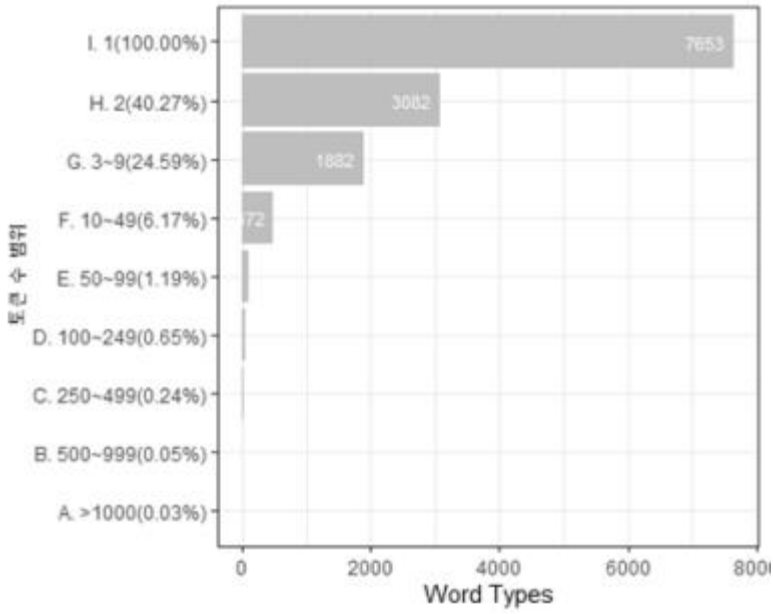
<그림 5>에 표시된 최다 빈도 상위 타입 각각에는 개개 타입 뿐만 아니라 문법적인 여러 형태, 복합어 및 유사한 의미로 사용되는 타입도 함께 포

8) AntConc에서 Tool Preferences → Word List → Word List Range에 본 연구 목적에 맞게 작성한 stopword list 파일을 첨부한 후 Word List를 실행하여 해당 수치를 얻었다. 불용어목록(stopword list)은 관사, 전치사, 대명사 등, 내용이 없기 때문에 검색 시 배제할 목적으로 작성되었다.

함시켰다. 본 연구 목적에 맞게 lemma list를 작성하여 AntConc에서 Tool Preferences → Word List → Lemma List에서 작성된 lemma list를 올리고 Word List를 실행하여 의도한 정보를 추출하였다. 예를 들어 STEIGEN (476회)에는 steigen(11), steigt(9), stieg(64), stiegen(49), gestiegen(19), gestiegene(2), gestiegenen(5), steigende(11), steigenden(3), steigender(5), steigern(5), steigerte(6), klettert(4), kletterte(23), kletterten(11), geklettert (4), klettertour (1), oben(61), plus(163), positive(20) 등이 포함된다. 다른 대문자 타입에도 같은 요령으로 합산시켰다. 이런 합산에는 주관적인 판단에 기초하기 때문에 어느 정도 의견의 차이는 있을 수 있지만, 큰 흐름은 계속 유지된다고 판단된다.

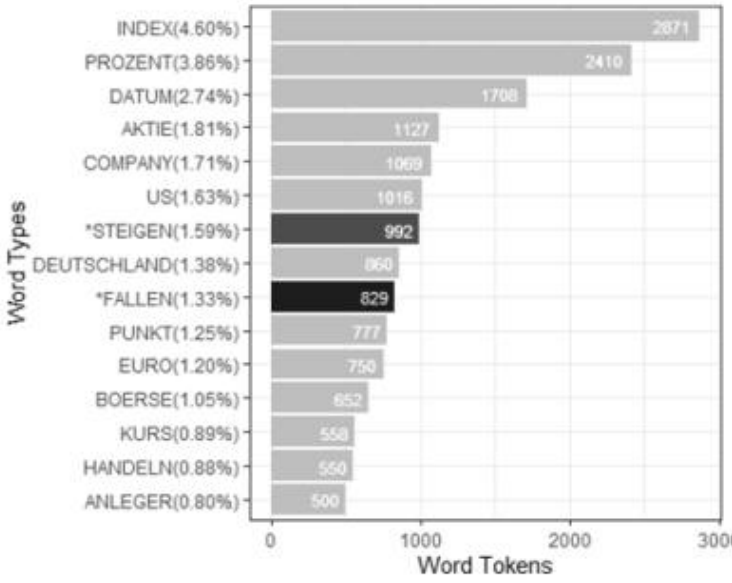
타입 명칭 옆 괄호 안의 비율 표시는 불용어를 처리한 후의 전체 토큰 수에 대한 비교이다. 이들 최다 빈도 상위 16개 타입의 내용을 살펴보면 등락(STEIGEN, FALLEN), 지수 (INDEX), 측정치(PROZENT, PUNKT), 환율(EURO, DOLLAR), 주식(BÖRSE, AKTIE), 기업(COMPANY), 투자자(ANLEGER), 거래(HANDELN), 시간(DATUM), 지역(US, DEUTSCHLAND, STADT) 등이 주를 이룬다. 이는 주식시장 마감시점에서 지수종류, 지수수치, 환율, 주식, 기업, 투자자, 시간, 지역 등 지수 등락과 관련된 정보임을 알 수 있다. 따라서 이들 어휘는 독일 주식시장 마감보고서의 특징이라고 판단된다.

<그림 6>에서는 토큰 수 범위 별 누적 Word Types를 보여주고 있다. 1.19%인 91개 타입이 50개 이상의 토큰을, 75% 가량의 타입은 1~2개 토큰이어서, 자주 사용되는 타입은 1% 가량임을 알 수 있다. 토큰 수 범위에서 괄호 안의 비율은 불용어를 제외한 전체 타입 수 7,653개에 대한 각각의 누적 Word Types의 상대 수치이다. 이로써 누적 합계의 비율은 100%로 된다.



〈그림 6〉 2017년도 하반기 독일 증시 토큰 수 범위별 누적 Word Types 통계

2017년도 상승 추세에 맞게 <그림 5>에서 STEIGEN(1.27%)은 FALLEN(0.74%)보다 높다. 참고로 FALLEN(280)에는 fallen(12), fallenden(4), fiel(56), fielen(31), sank(16), sinken(2), abwärts(9), abwärtssog(3), abwärtsstrudel(1), abwärtstrend(2), tief(16), tiefer(17), tiefsten(7), minus(93), unten(11) 등이 포함되었다.



〈그림 7〉 하락추세인 2018년도 독일 증시 최다 빈도 언어통계

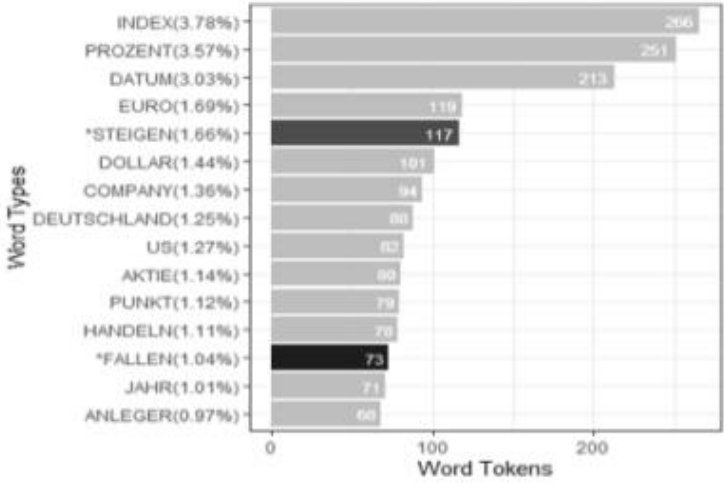
<그림 7>에서는 2018년의 자료 중, 자료 접근의 제약으로 4월까지는 독일경제지 Handelsblatt 디지털판에서, 5월부터 12월까지는 Börse-Online에서 수집했다. 이 자료로부터 전체 111,523 토큰과 10,644 타입이 조사되었다. 이중 불용어들을 제외하면 62,538 토큰과 9,203 타입이 추출되었다.

2017년도와 비교하여 최다 빈도 상위 어휘에서 STADT와 DOLLAR가 빠지고, KURS가 추가되었다. 이런 순위 다툼은 최다 빈도 상위 범위를 조정하면 완화되리라고 본다.

2018년은 하락추세이지만 <그림 7>에서는 상승 추세였던 2017년과 같이 STEIGEN(1.59%)이 FALLEN(1.33%)보다 높다. 그러나 등락을 나타내는 이 어휘군 사이의 차이는 2017년 하반기에는 0.53%포인트, 2018년에는 0.26%포인트로 하락 추세에서는 차이가 줄어든 것을 알 수 있다. 추세를 파악하는데 이런 수치의 차이가 절대적인 기준이 될 수는 없지만, 추세 전환

에 대한 생각의 단초를 줄 수는 있다.

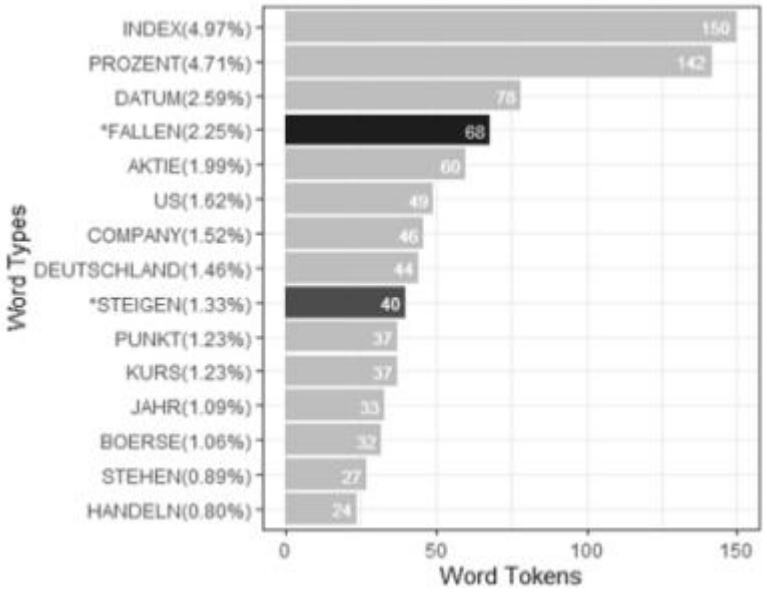
추세와 등락에 관한 어휘군 사이의 관계를 좀 더 확실히 확인하기 위해서 최고점인 2018년 1월과 최저점인 2018년 12월을 비교했다.



〈그림 8〉 최고점인 2018년 1월 독일 증시 최다 빈도

<그림 8>에서는 독일경제지 Handelsblatt 디지털판에서 전체 12,804 토큰과 3,061 타입이 조사되었고, 이중 불용어들을 제외하면 7,038 토큰과 2,301 타입이 추출되었다.

<그림 9>에서는 Börse-Online에서 전체 5,221 토큰과 1,482 타입이 조사되었고, 이중 불용어들을 제외하면 3,017 토큰과 1,051 타입이 추출되었다.



〈그림 9〉 최저점인 2018년 12월 독일 증시 최다 빈도 언어통계

<그림 8>에 비해 <그림 9>에서 자료 규모가 절반 축소된 것은 자료가 추출된 언론사의 변경으로 인한 것이다. 그러나 텍스트종류의 특징은 작성 언론사에 특화된 것이 아니기 때문에 크게 문제되지 않는다. 이는 <그림 8>과 <그림 9>의 타입들이 대부분 일치하는 것에서도 알 수 있다. 이 두 그림에서 STEIGEN(*상승*)과 FALLEN(*하락*)은 2018년 1월 최고점에서는 *상승*이 *하락*보다 우세하며, 2018년 12월 최저점에서는 역으로 *하락*이 *상승*보다 우세한 것을 알 수 있다.

4.2. 한국증시에서 자주 나타나는 어휘

독일증시 관찰에서와 같이 통계수치의 시각화를 위해 단어구름으로부터 시작한다.

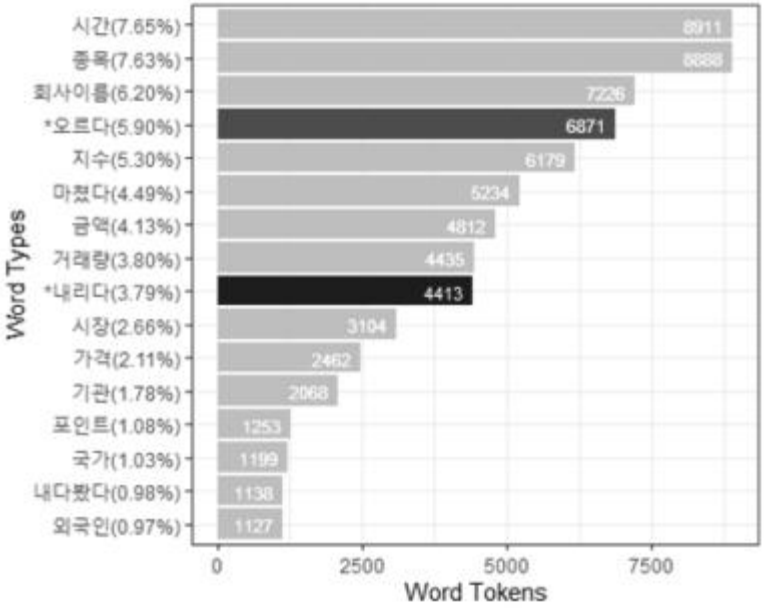
2017년 상승추세 2018년 하락추세

〈그림 10〉 한국 증시 2017년과 2018년 추세에서 단어구름 비교

<그림 10>은 조사 대상인 2017년과 2018년 한국 증시 마감보고서의 자료를 사용하여 작성된 것이다. <그림 10>의 왼쪽은 상승추세인 2017년 8월부터 12월까지의 자료 중 정제 과정을 거쳐 유의미한 단어를 추출한 다음 100개 토큰 이상의 타입으로 구성되었고, <그림 10>의 오른쪽은 하락추세인 2018년 1월부터 12월까지 자료 중 같은 정제 과정을 거쳐 100개 토큰 이상의 타입으로 이루어졌다. 중앙에 있는 글자 크기 별로 빈도수가 높은 것을 알 수 있다.

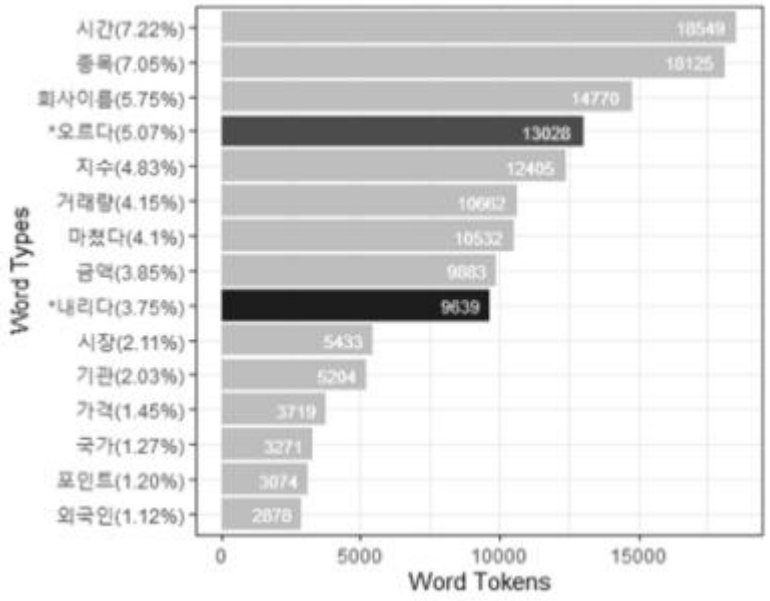
독일 증시 마감보고서와 같이 다음에서는 AntConc 프로그램을 사용하여 토큰과 타입의 분포를 더 구체적으로 분석했다.

<그림 11>에서는 2017년 8월에서 12월까지 Daum의 금융뉴스에서 자료를 수집했다. 이 자료로부터 전체 143,403 토큰과 15,759 타입이 조사되었다. 이중 불용어들을 제외하면 116,482 토큰과 13,859타입이 추출된다. <그림 11>에 표시된 최다 빈도 상위 15개 타입의 내용을 살펴보면 등락(*오르다, 내리다*), 지수(*가격*), 측정치(*포인트*), 주식(*종목*), 기업(*회사이름*), 투자자(*기관, 국가*), 시장(*금액, 거래량, 마쳤다*), 시간, 전망(*내다봤다*) 등이어서 독일 증시 마감보고서와 비슷하기 때문에 텍스트종류로서의 특징이 더욱 일반화된다고 판단된다.



〈그림 11〉 2017년도 하반기 상승추세에서 한국 증시 최다 빈도 언어통계

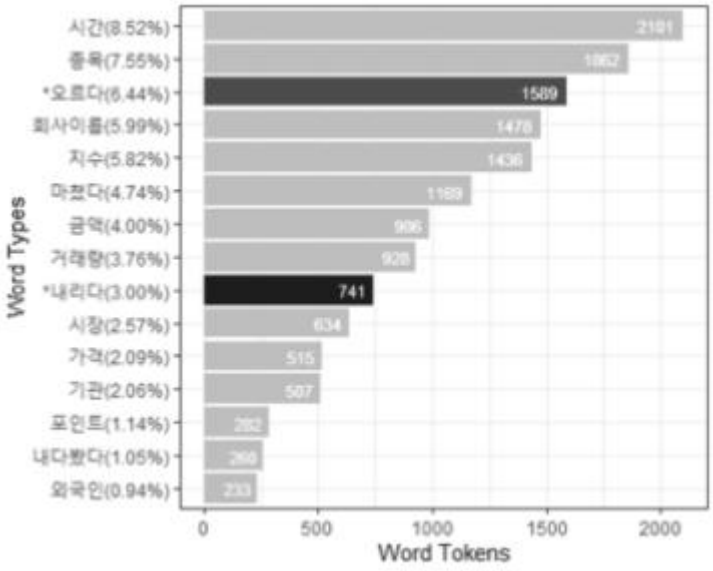
<그림 11>에서 볼 수 있듯이 2017년도 상승 추세에 맞게 *오르다*(5.90%)가 *내리다*(3.79%)보다 더 자주 사용되었다. *오르다*(6871)에 함께 한 어휘군은 *상승(2447), 올랐다(2371), 강세(538), 상위(334), 급등(287), 우위(214), 돌파(123), 높다(115), 상한가(109), 위(98), 반등(69), 끌어올리다(51), 넘다(33), 치솟다(31), 가파르다(28), 웃돌다(8), 상회(4), 고공행진(3), 떠받치다(3), 훈풍(3), 후광(1), 강타(1)*가 있으며, 내리다(4413)에 함께한 어휘군에는 *하락(1967), 내리다(1349), 약세(453), 떨어지다(147), 하한가(140), 급락(114), 낙폭(74), 밀리다(71), 낮다(37), 아래(28), 후퇴(24), 내주다(4), 급감(2), 하회(2), 후폭풍(1)*이 있다. 이 어휘군에는 파생어, 복합어들이 다수 포함되었다. 타입면에서는 *내리다*에 비해 *오르다*가 압도적으로 많다.



〈그림 12〉 2018년도 하락추세에서 한국 증시 최다 빈도 언어통계

<그림 12>에서는 Daum의 금융뉴스에서 수집된 2018년 자료에 기초했다. 이 자료로부터 전체 306,700 토큰과 25,318 타입이 조사되었다. 이중 불용어들을 제외하면 256,951 토큰과 23,974 타입이 추출된다. 2017년과 비교하여 최다 빈도 상위 타입의 내용은 대체로 비슷하다. 2018년에는 하락추세임에도 <그림 12>에서 보는 바와 같이 *오르다*(5.07%)가 *내리다*(3.75%)보다 상승추세였던 2017년과 같이 높다. 그러나 등락을 나타내는 이 어휘군 사이의 차이는 2017년 하반기에는 2.11%포인트, 2018년에는 1.32%포인트로 하락추세에서는 상대적으로 줄어든 것을 알 수 있다.

추세와 등락에 관한 어휘군 사이의 빈도 차이를 좀 더 확실히 확인하기 위해서, 한국 증시에서도 최고점인 2018년 1월과 최저점인 2018년 10월을 비교했다.

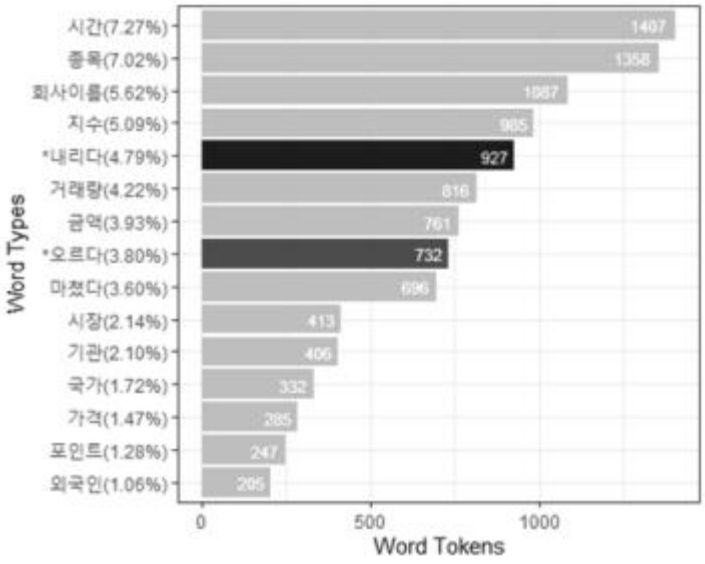


〈그림 13〉 최고점인 2018년 1월 한국 증시 최다 빈도 언어통계

<그림 13>에서는 전체 29,536 토큰과 5,351 타입이 조사되었고, 이중 불용어들을 제외하면 24,661 토큰과 4,397타입이 추출되었다.

<그림 14>에서는 전체 23,131 토큰과 4,868 타입이 조사되었고, 이중 불용어들을 제외하면 19,344 토큰과 3,994 타입이 추출되었다.

최다 빈도 상위 집단에 속하는 내용은 대체로 비슷하기 때문에, 어휘 면에서도 텍스트종류로 특징지을 수 있다고 판단된다. 다만, 이 두 그림에서 2018년 1월 최고점에서는 *오르다*가 *내리다*보다 우세하며 2018년 10월 최저점에서는 역으로 *내리다*가 *오르다*보다 우세한 것을 알 수 있다.



〈그림 14〉 최저점인 2018년 10월 한국 증시 최다 빈도 언어통계

독일 증시와 한국 증시의 마감보고서를 텍스트종류로 특징짓는데 특정 어휘군도 참고 될 수 있음을 실증 연구로 확인할 수 있었다. 더 나아가 특정 어휘는 상황에 따라 사용 빈도에 차이가 날 수 있음도 확인할 수 있었다.

5. 결론

증시 마감보고서에 자주 나오는 어휘의 빈도 순위는 상황 변화에 상관없이 대체로 유지되었으나, 등락에 관한 어휘 빈도의 순위에는 변동이 있었다. 추세 기간이 일 년이라도 월 단위로 나누어 관찰하면 이런 변동이 좀 더 동적으로 관찰될 수 있으리라 판단된다. 이는 후속 연구과제로 미룬다.

모든 텍스트종류에서 같은 현상을 관찰하기는 힘들지만, 상황 변화에 민감한 영역일수록 특정 어휘 사용의 빈도에 관심을 가질 수 있다. 이런 어휘들은

핵심 어휘일 수 있으며, 텍스트종류를 특징지을 수도 있다고 판단된다.

텍스트마이닝에서 텍스트는 비정형데이터로, 이를 작업하기 위해서는 정형데이터로의 전환이 필수적이다. 텍스트마이닝에서 텍스트는 어휘의 집합으로 간주되기 때문에, 텍스트작업은 이런 어휘들의 정형적인 집합을 공학적으로 유추해 내는데 주된 관심을 쏟는다. 이런 점에서 텍스트마이닝과 텍스트종류 연구는 상호 접점을 찾을 수 있다. 텍스트종류라는 명칭도 독일어권에서는 일상적으로 광범위하게 통용되고 있어 폐기할 수는 없지만, 텍스트종류를 개념적으로 규정짓고자 할 때 항상 논의 되는 것이 정형성이다. 어휘의 집합에서 정형성을 찾으려는 시도는 텍스트마이닝 뿐만 아니라 텍스트언어학에서도 주요한 과제이다.

텍스트마이닝 기법을 효과적으로 활용하기 위해서는 미리 준비해야할 작업이 많다. 감성도를 측정하기 위해서는 감성 단어 목록이 잘 구비되어야 하듯이, 주식시장의 *오름*과 *내림*의 분위기를 측정하기 위해서, 사전에 측정기준에 대한 논의가 활성화되어야 한다. *오름*이냐 *내림*이냐와 같이 이분법적인 결정을 요구하는 상황은 디지털 기계에서 작업하기 좋은 환경이기도 하다. 인공지능이나 기계학습이 초기에 시행착오를 많이 하지만, 학습기간이 길수록 실수의 횟수도 점차 줄어든다. 본 연구도 많은 시행착오를 거치면서, 더욱 성숙한 결과물에 도달할 것이라 판단된다.

참고문헌

1. 자료

Daum 금융뉴스, http://finance.daum.net/news/news_list.daum?limit=30&section=&type=market (접근: 2017, 2018)

Börse Online, https://www.boerse-online.de/maerkte/marktberichte (접근: 2018)

Handelsblatt, http://www.handelsblatt.com/finanzen/maerkte/?navi=FINANZEN_SUBNAV_M%C3%84RKTE_1980480(접근: 2017, 2018)

2. 저서 및 논문

김수현 외,『경제 분석을 위한 텍스트 마이닝』. Bank of Korea WP(한국은행경제연구원), 2019. (Available at SSRN: https://ssrn.com/abstract=3405781 or http://dx.doi.org/10.2139/ssrn.3405781)

김유신 · 김남규 · 정승렬,「뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형」,『지능정보연구』18(2), 지능정보연구, 2012, 143~156면.

김홍림 · 김민진,「텍스트마이닝을 통해 살펴 본 2019년 우정사업 트렌드」,『우정정보』겨울, 정보통신정책연구원, 2019, 57~69면.

방경원,「증시 마감시황 보고에 나타난 은유표현 - 영어, 한국어, 독일어 사용지역 비교」,『독일언어문학』70, 한국독일언어문학회, 2015, 67~86면.

_____,「독일증시 마감보고서의 언어학적 기술 - 전문 텍스트종류 구성요소를 중심으로」,『인문학연구』55, 조선대학교 인문학연구원, 2018, 217~244면.

안성원 · 조성배,「뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측」,『한국정보과학회 학술발표논문집』37(1C), 한국정보과학회, 2010, 364~369면.

이지환 · 홍유석,「10-K 연차보고서의 텍스트마이닝을 통한 서비스화 동향추출 및 분석방법론 개발」,『대한산업공학회 춘계공동학술대회 논문집』, 대한산업공학회 2013, 1631~1637면.

한무명초 · 김양석 · 이충권,「텍스트 마이닝 기법을 활용한 동남권 신공항 신문기사 분석」,『스마트미디어저널』제6권 제1호, 한국스마트미디어학회, 2017, 47~53면.

Adamzik, Kirsten, *Textlinguistik. Eine einführende Darstellung*. Tübingen: Max Niemeyer Verlag, 2004.

Březina, Jaroslav, “Phänomen Fachsprachen – Börsensprache unter der Lupe” in: PhiN(Philologie im Netz) 70, 2014, pp.17~38.

Eitze, Kathrin, *Metaphern in der Börsenfachsprache. Eine kontrastive Analyse des Spanischen und Deutschen*. Hamburg, 2012.

Abstract

Text Type Research Using Text Mining Technique

- Comparing Korean and German Stock Closing Reports -

Bang, Kyung-Won*

The purpose of this study is to empirically confirm that there is a difference in the frequency of UP and DOWN specific vocabulary groups depending on the situation of stock closing report. This is to characterize text types in lexical units. Although vocabulary is the only entity in the text, it has been neglected in previous research. This is because time and effort were limited in analyzing text at the lexical level. Text mining techniques that deal with big data will reduce this effort. In addition, it is expected that the study of texts in the lexical unit wii be activated in the future.

The data used in this study were taken from the online edition of Handelsblatt and from Börse Online, the internet media in Germany, and from Daum's financial news in Korea. The second half of 2017 is on the uptrend, based on securities firms' stock market charts. The 2018 interval is set for the downtrend. R and AntConc programs were used to work with word cloud and language statistics.

The most frequent vocabulary that appeared in the closing report did not show a big difference in the ranking even if the trend changed. However, the vocabulary groups of the Up and Down were found to have different frequency rankings as time changed. Therefore, it is judged that certain vocabulary groups can be fluidly observed depending on the text types. This kind of vocabulary will also be useful for characterizing text types.

Key Words : Text type, Stock closing report, The up and down of the stock market, Text mining, Machine Learning

* Hankuk University of Foreign Studies

<필자소개>
이름 : 방경원
소속 : 한국외국어대학교
전자우편 : kwonbang@hanmail.net

논문투고일 : 2020년 2월 4일
심사완료일 : 2020년 2월 24일
게재확정일 : 2020년 2월 24일