주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축방안

A domain-specific sentiment lexicon construction method for stock index directionality

저자
(Authors)

김재봉, 김형중
Jae-Bong Kim, Hyoung-Joong Kim

출처
(Source)

한국디지털콘텐츠학회 논문지 18(3), 2017.6, 585-592(8 pages)
Journal of Digital Contents Society 18(3), 2017.6, 585-592(8 pages)

발행처
(Publisher)

한국디지털콘텐츠학회
Digital Contents Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07222144

APA Style

김재봉, 김형중 (2017). 주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축방안. 한국디지털콘텐츠학회 논문지, 18(3), 585-592

JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 3, pp. 585-592, June. 2017

주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축방안

김 재 봉 · 김 형 중
고려대학교 빅데이터응용및보안학과

A domain-specific sentiment lexicon construction method for stock index directionality

Jae-Bong Kim · Hyoung-Joong Kim
Department of Big Data Application and Security, Korea University

[요 약]

개인용 디바이스의 발달로 개인들이 손쉽게 인터넷에 접속할 수 있게 되었으며, 소셜미디어를 통한 정보의 공유와 습득이 일반화 되고 있다. 특히 분야별 전문 커뮤니티가 발달하며 사회적 영향력을 행사하고 있어 기업과 정부는 이들의 의견을 반영하여 전략을 수립하는 일에 관심을 기울이고 있다. 온라인상의 다양한 텍스트로부터 대중의 의견을 읽어내는 것을 오피니언마이닝이라고 한다. 그 중 하나인 감성사전은 방대한 비정형데이터를 빠르게 파악하는 도구로 여러 분야에서 활용되고 있다.

주식시장은 사회의 여러 요인을 반영하여 변동한다. 최근에는 버즈량 분석 등 빅데이터를 기반으로 오피니언마이닝을 활용한 주식시장 연구가 시도되고 있다. 대표적인 예로 뉴스와 같은 텍스트 데이터 분석을 활용한 연구들이 발표되고 있다.

본 논문에서는 뉴스의 정제된 형식과 한정된 어휘를 사용한 기존연구를 보완하고자 증권전문 사이트 'Paxnet'의 게시 글을 분석대상으로 삼아 주식시장 맞춤형 감성사전을 구축하여 투자자들의 감성을 분석하는 데 기여했다.

[Abstract]

As development of personal devices have made everyday use of internet much easier than before, it is getting generalized to find information and share it through the social media. In particular, communities specialized in each field have become so powerful that they can significantly influence our society. Finally, businesses and governments pay attentions to reflecting their opinions in their strategies. The stock market fluctuates with various factors of society. In order to consider social trends, many studies have tried making use of bigdata analysis on stock market researches as well as traditional approaches using buzz amount. In the example at the top, the studies using text data such as newspaper articles are being published.

In this paper, we analyzed the post of 'Paxnet', a securities specialists' site, to supplement the limitation of the news. Based on this, we help researchers analyze the sentiment of investors by generating a domain-specific sentiment lexicon for the stock market.

색인어 : 오피니언마이닝, 감성분석, 말뭉치, 감성사전
Key word : Opinion Mining, Sentiment analysis, Corpus, Sentiment lexicon

http://dx.doi.org/10.9728/dcs.2017.18.3.585

Received 18 May 2017; **Revised** 05 June 2017
Accepted 25 June 2017

***Corresponding Author;** Hyoung-Joong Kim
Tel: +82-02-3290-4895, ※ 개인정보 표시제한
E-mail: khj-@korea.ac.kr

Ⅰ. 서 론

2010년 이후 스마트기기의 보급이 활발해지고 네트워크에 대한 접근환경이 개선되며 온라인상의 데이터가 기하급수적으로 증가하는 '빅데이터 시대'에 진입하게 되었다. 이러한 환경에서 소셜미디어는 생산자와 소비자 사이의 중간매개자 또는 정보허브로 기능하며 그 사회적 영향력이 갈수록 커지고 있다.[1]

빅데이터는 통상적으로 사용되는 데이터 수집, 관리 및 처리 소프트웨어의 수용 한계를 넘어서는 크기의 데이터를 의미한다[2]. 정형데이터와 비정형데이터를 모두 포함하는 개념이지만 통상적으로는 특정한 목적을 가지고 형식에 맞추어 구축된 '정형데이터'에 대비되는 개념으로 온라인상의 '비정형데이터'와 같은 의미로 통용되고 있다. 비정형데이터를 분석하는 방법은 정형데이터의 분석과는 다르다. 비정형데이터 분석방법으로는 크게 오피니언마이닝, 텍스트마이닝, 웹마이닝 등이 있다. 본 논문에서는 그 중에서도 오피니언 마이닝의 일종인 감성분석에 대해 다룬다.

감성분석은 텍스트가 담고 있는 부정, 긍정 등의 주관적 '감성'을 읽어내는 분석으로, 이를 활용하여 각 단어에 감성적 특성인 '극성'을 부여하는 감성사전을 구축할 수 있다. 감성사전은 다시 범용 감성사전과 도메인 맞춤형 감성사전으로 분류된다. 범용 감성사전은 텍스트가 사용된 분야나 상황에 관계없이 어느 맥락에나 적용이 가능한 극성 값이 부여된 사전으로, 단어의 사전적 의미에 기반하여 쉽고 빠르게 사전을 구축할 수 있으나 문맥에 따라 단어가 포함하는 감성이 달라질 경우 오류가 발생할 수 있다. 그러나 대부분의 경우, 텍스트를 해석하기 위해서는 글을 쓴 문맥을 반드시 고려하여야 한다. 이에 대응하여 개발되고 있는 것이 도메인별 맞춤형 사전이다. 맞춤형 사전의 경우에는 단순한 긍정, 부정의 '감성'뿐 아니라 해당 분야에서 긍정적이거나 부정적으로 받아들여지는 '사건'의 가능성에 대한 평가도 가능하다[3].

본 연구는 주식시장이라는 도메인에 특화된 감성사전 구축을 목적으로 한다. 여기서 긍정은 '주가상승'을, 부정은 '주가하락'을 의미한다. 데이터로는 증권정보공유에 특화된 웹사이트인 'Paxnet'의 게시 글을 활용한다. 주식시장의 방향성을 예측하기 위해 신문기사 데이터를 분석하여 감성사전을 구축한 연구 사례가 이미 있었으며[4][5][6][7], 정제된 형식과 한정적인 어휘만으로 이루어진 텍스트로 인한 한계가 지적되었다. 온라인 게시 글을 분석하여 감성사전을 구축한 선행연구를 통해 다수의 비전문가 의견을 반영한 감성사전의 성능이 전문가 의견을 반영한 감성사전이나 범용 감성사전에 비해 부족하지 않음을 확인하였으며[8][9], 주식분야 관련 웹사이트 중 가장 전문성, 신뢰성이 높고 전업투자로 시장에 참여하는 투자자들의 게시판 글을 기초로 많이 사용되는 증권 전문용어 중심으로 구성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 오피니언마이닝과 감성사전에 관련된 선행연구들을 검토하고, 3장에서는 감성분석을 주식시장에 적용한 기존 연구사례를 살펴보고, 온라인 게시 글을 분석하여 감성사전을 구축하는 연구를 제안한다. 기존 연구와의 차별점도 이 장에서 다룰 예정이다. 4장에서는 구축된 감성사전을 제시하고, 그 결과를 요약한다. 마지막으로 5장에서는 연구결과의 한계점, 향후 연구방안에 대해 논의하겠다.

Ⅱ. 오피니언마이닝 연구

2-1 오피니언마이닝 개념

산업분야에서 감성분석이라는 이름으로 불리는 오피니언마이닝은 비정형 텍스트데이터로부터 특정 상품이나 개념에 대한 사람들의 생각, 감정, 태도와 같은 주관적인 반응을 분석해내는 과정을 의미한다[10]. '감성분석'과 '오피니언마이닝'이라는 개념은 각각 2003년에 발표된 서로 다른 연구논문에서 처음 언급되었으며[11][12], 연구대상 및 방법론은 유사한 것으로 본다. 연구대상인 '감성' 또는 '의견(opinion)'은 일반적으로 긍정과 부정으로 분류되는 사람들의 감성을 의미하지만, 연구가 발전되면서 감성이 더욱 세분화되기도 하고, 상황에 따라서는 주관적인 감성을 넘어 그러한 감성을 유발하는 긍정적인 사건과 부정적인 사건으로 다루어지기도 한다.[13]

네트워크 인프라의 확충과 스마트기기의 보급 등으로 온라인상의 데이터가 기하급수적으로 증가하는 2010년 이후 빅데이터 분석에 대한 관심이 높아지며 오피니언마이닝 관련 연구도 활발히 이루어지고 있다. 다양한 비정형데이터 분석방법론 중에서도 오피니언마이닝은 가장 각광받는 분석법이다. 오피니언마이닝은 어느 도메인에나 적용이 가능하며, 대중의 의견으로부터 영향을 받는 정치, 경제 등 사회 각 분야의 의사결정에 유용한 정보를 제공하기 때문이다.

2-2 감성사전 연구

오피니언마이닝에서 가장 중요한 요인은 '감성어(sentiment words)'일 것이다. 감성어는 긍정적, 또는 부정적 감정을 표현하는 단어나 구절을 의미하며, 빅데이터 분석의 주요 관심인 대중의 감정을 보여주는 지표가 된다. 이러한 단어와 구절을 목록으로 정리한 것이 바로 '감성사전(sentiment lexicon)'이다[3].

오피니언마이닝에 관한 연구가 활성화된 후 지난 몇 년간 감성사전 구축을 위한 다양한 알고리즘이 개발되었는데, 이것을 크게 세 가지 접근 방법으로 분류할 수 있다. 첫 번째는 단어의 극성을 직접 분류하는 수동 접근방법(manual approach)으로, 노동력과 시간이 많이 필요하기 때문에 주로 사후 검토

용으로만 사용하고 단독적으로 사전구축에 사용하지 않는다. 다른 두 개의 알고리즘은 사전구축에 주로 사용되는 자동화 알고리즘(automated algorithm)으로 단어의 사전적 의미에 기반하는 '사전기반 접근법(dictionary-based approach)'와 '말뭉치기반 접근법(corpus-based approach)'로 나눌 수 있다[3].

1) 사전기반 접근법(dictionary-based approach)

'사전기반 접근법(dictionary-based approach)'이란 감성사전이 아닌 일반적인 어학사전에 기반하여 감성사전을 구축하는 접근방법이다. 각 단어가 가지고 있는 사전적 의미에 기반하여 먼저 기초 감성어휘 목록을 작성하고, 동의어와 반의어를 기준으로 사전을 구축한다.

사전기반 사전구축방법에 관한 연구는 동의어와 반의어를 기준으로 단순 반복 작업(iteration)을 거쳐 사전을 확장하는 붓스트랩 방법(bootstrapping methods)[14]에서부터 확률론을 이용해 극성 값을 부여하는 연구[15], 확률보행(random walk)활용방법[16], PMI(pointwise mutual information)를 이용하는 방법[17] 등 다양하게 발전해 왔다. 국내에서는 최근 국립국어원 사전을 기반으로 집단지성을 이용한 한글 감성어 사전 구축을 연구한 사례가 있다[18].

사전기반 접근법은 쉽고 빠르게 사전을 구축할 수 있다는 장점이 있다. 그러나 사전의미에 기초한 감성사전은 언어가 사용되는 상황이나 환경에 대한 이해 없이 일반적인 의미에만 의존한다는 한계를 가지고 있다. 즉, 문맥을 고려하여야 하는 감성분석에는 적용하기 어렵다는 점이다. 그러나 많은 경우 감성분석에 앞서 반드시 상황에 대한 이해가 필요하기 때문에 사전기반 접근법과는 다른 접근방법이 개발되고 있다.

2) 말뭉치기반 접근법(corpus-based approach)

'말뭉치기반 접근법(corpus-based approach)'이란, 기초적인 감성어휘나 기존에 개발되어있는 범용감성사전을 텍스트 데이터에서 추출된 말뭉치에 적용하여 새로운 감성사전을 구축하는 방법이다[19]. 같은 단어라도 상황에 따라 의미가 달라질 수 있어 이러한 방법의 감성사전 구축은 매우 복잡하다.

이 접근법은 기초적인 감성 형용사에 언어학적 특성을 활용하여 문맥을 읽는 1997년 Hazivassiloglou와 Mckeown의 초기연구[20]에서 시작되어 추가적으로 발견되는 한계와 문제점을 해결해 나가며 발전되고 있다. 그렇지만 여전히 많은 어려움이 발견되고 있어 연구가 활발히 진행 중이다. 사전기반 접근법의 하나인 PMI에서 파생된 SO-PMI(semantic orientation using pointwise mutual information)는 말뭉치기반 접근법에 해당된다. 말뭉치기반의 사전구축은 매우 복잡한 과정임에도 불구하고, 언어의 의미를 해석할 때 문맥의 중요성을 무시할 수 없기 때문에 여전히 많은 전문가들이 이에 관한 연구를 활발히 진행하고 있다.

말뭉치기반 사전구축은 기본적으로 언어가 사용되고 있는 실제 텍스트를 활용하기 때문에 범용 감성사전을 구축하기 위해서는 모든 도메인의 텍스트를 분석해야 한다. 그렇지 않으면 편향된 감성사전이 만들어진다. 다양한 상황에서의 방대한 텍스트 데이터를 활용할 수 있다면 말뭉치를 분석하는 방법으로 범용 감성사전을 만들 수 있으나, 이는 여전히 매우 비효율적이다. 따라서 말뭉치기반 접근법은 주로 특화 감성사전을 구축하는데 활용되고, 범용 감성사전 구축에는 사전기반 접근법이 사용되는 것이 일반적이다.

Ⅲ. 주식시장 특화 감성사전 구축

3-1 관련 연구

참여와 개방, 공유를 지향하는 최근 인터넷 환경은 트위터, 인스타그램, 카카오톡 등 소셜 커뮤니티의 증가와 함께 다양한 주제에 대해서 다양한 사람들이 소통할 기회가 많아졌다.

최근에는 이런 소셜 커뮤니티의 내용들을 분석하여 사회, 정치, 경제 등 다양한 분야의 전략/마케팅 부문에서 활용하고 있다. 그러나 소셜 커뮤니티의 내용들은 양의 방대함과 형태의 다양성으로 인해 사람이 직접 일일이 보고 내용을 인지하는 데에는 현실적으로 한계가 있다. 때문에 이를 일정 기준에 따라 정형화할 필요성이 강조되었으며, 이를 위한 방법으로 감성분석이 활용되고 있다[21].

실례로 버락 오바마 미국 전 대통령은 2012년 미국 대선에서 소셜 커뮤니티 분석을 통해 유권자들의 감성을 분석하여 이를 선거 전략 구축에 반영하여 재선에 성공하였으며[22], 스페인의 민간 금융 업체인 BBVA는 감성 분석을 통해 브랜드 이미지 제고와 고객관리 분야에서 성과를 창출했다[22].

감성분석에 대한 연구는 감성분석의 데이터가 되는 커뮤니티의 다양한 의견들을 일정한 기준에 의해 정형적으로 분석 결과를 도출할 수 있는 감성사전 구축으로 이어진다. 잘 구축된 감성사전은 수치적, 통계적인 방법 외에 추가적인 분석을 제공할 수 있기 때문에 감성사전 구축을 위한 연구가 활발하다. 이런 연구의 일환으로 상품 평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동구축 연구에서는 전자상거래에서 판매되는 상품들에 대해 고객들의 상품 평을 이용한 감성사전을 구축하였다. 이를 통해 평점 및 접속정보 등 각 상품에 대한 고객들의 평가 방법들 중 가장 효과적인 방법을 도출해 낼 수 있었다. 다만, 이 연구는 상품 평 반응 영역(도메인)에 특화된 감성사전을 구축하여 범용 감성사전을 사용했을 때보다 예측 정확도를 끌어 올렸지만, 서술어에 대해서만 감성사전을 구축한 한계점이 있었다고 언급되었다[23].

집단지성을 기반으로 단어들의 감성 태그 및 점수화를 활용한 연구[18]에서는 명사, 형용사, 동사, 부사를 대상으로 투표(집단지성+ 폭소노미)를 통해 단어별 긍정, 부정을 선정

하여 한글 감성사전을 구축하였다. 그러나 어미와 조사가 발달한 한글의 특성상 자연어 처리의 어려움이 있었으며, 자연어 처리를 위한 감성어 사전과 같은 자원 부족의 한계점이 나타났다. 또한 감성사전을 적용할 분야에 따라서 같은 단어라도 긍정과 부정이 다르게 인식되는 경우가 많아서 각 분야와 문맥에 따라 추가적으로 사용되는 단어들의 온톨로지 구축이 요구되었다.

주식시장 연구는 전통적으로 수치적, 통계적 접근법이 주류를 이루고 있지만, 최근 오피니언마이닝 기법의 발달에 힘입어 투자심리 등 주식시장에 영향을 미치는 다양한 요인들에 대한 감성분석이 활발하게 진행되고 있다.

주가예측과 관련된 감성분석들은 대체로 뉴스 데이터를 기반으로 이뤄지고 있다.

'뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정 모형'[5]에서는 형태소 분리 후 단어별 긍정/부정의 극성을 태깅한 후 인덱싱된 분류정보와 스코어링 룰을 이용하여 뉴스의 긍정/부정 극성을 판별하였다. 그러나 데이터가 충분하지 못했으며, 범용 감성 사전을 이용함으로써 극성 분류의 정확도가 다소 낮을 수 있는 한계점이 있었다.

'뉴스 감성분석과 SVM을 이용한 S&P500 주가지수 예측'[7] 또한 뉴스 데이터를 중심으로 하여 감성분석과 기술적 분석을 결합한 분석 방법을 제시하였으며, '주가지수 방향성 예측을 위한 주제지향 감성사전 구축방안'[6]에서는 주식시장에 특화된 단어에 적합한 극성을 부여하였다.

또한 '온라인 언급이 기업 성과에 미치는 영향 분석'[24]에서는 주식시장을 세분화하여 특정 산업 및 기업에 대한 감성사전 구축을 진행하였다.

그러나 앞선 연구들은 모두 품사가 명사에 한정되거나 뉴스 데이터에 의존하는 점, 그리고 데이터 양의 불충분이라는 한계점이 있었다.

3-2 감성사전 구축방안

본 연구에서는 증권 전문 소셜 커뮤니티의 다양한 게시 글을 토대로 주식시장 맞춤형 감성사전 구축하는 것을 목표로 한다.

1) 활용데이터 정의

본 연구에서는 주식 전문 사이트인 'Paxnet'에서 운영하는 게시판의 최근 3년 치의 게시 글을 연구 데이터로 선정하였으며, 이를 선정함에 있어서 다음의 고려사항을 반영하였다.

첫째, 데이터의 비정형성을 극대화하고 시각을 다양화하기 위해 누구나 자유롭게 참여하여 의견을 개진할 수 있는 온라인 게시판을 선정하였다. 게시판에 게재된 글은 뉴스에 비해 가감 없는 솔직한 의견이 표현되며, 사용되는 용어 또한 뉴스에 비해 훨씬 다양한 특징을 가졌다.

둘째, 3년의 충분한 기간 동안 축적된 데이터를 통해 단어의 고유 극성의 정확도를 높이고자 하였으며, Paxnet과의 정식 약정을 통해 게시판 데이터를 직접 받음으로써 데이터의 손실을 없앴다.

마지막으로, 다양한 주식투자자들(전문가-초보)이 이용하는 증권 전용 전업투자자 게시판을 활용하여, 증권업에서 주요 사용되는 전문용어들을 통해 전문성을 높이고자 하였다.

2) 선행연구와의 차별 점

이번 연구는 기존의 연구와 다음의 차별성을 갖고 진행한다.

첫째, 주가방향성 예측을 위한 감성사전 구축에 관한 기존의 연구는 뉴스 데이터를 기반으로 진행되었다. 뉴스 데이터는 정제된 표현이 많으며, 한정된 어휘로 만들어졌기 때문에 연구의 제약사항으로 언급되었다. 이번 연구에서는 인터넷 게시판 글을 통해 시각의 다양화와 데이터의 비정형성을 극대화했다.

둘째, 주가방향성 예측을 위한 감성사전 구축에 관한 기존의 연구는 명사 또는 동사를 분석했지만, 이번 연구에서는 형용사까지 활용하였으며, 특히 단어 중심의 기존연구와 달리 데이터상의 말뭉치를 추출하거나 자주 쓰이는 단어들을 조합한 말뭉치를 활용하여 주식시장의 특성을 반영하였다.

3) 연구방법

주식시장에 특화된 감성사전 구축은 그림 1과 같은 순서로 진행하였다.

연구에 제한 목적으로 팍스넷 게시판의 게시 글들을 파일형태로 수령하여 데이터를 확보하였다. 동일한 단어가 시황별, 문맥상 의미, 다른 단어들과의 조합에 따라 반대의 감성으로 사용될 수 있으므로 극성의 정확한 상황별 적용을 위하여 충분한 기간 내 활용 사항을 분석 적용하기 위해서는 3년 내외의 기간이 적절한 점과 팍스넷 내 다수의 게시판 중에서 가장 조회 수가 많고 전문성이 뛰어난 점을 이유로 전업투자자 게시판을 기초로 활용하였으며 연구대상에서 제외되는 광고성 글과 이미지들을 제거하고자 하였다.

이후 그림 2와 같이 플레인 텍스트만을 추출하여 순수 게시 글 원형을 유지하였으며, 각 게시물 내 글들을 문장 단위로 구 별하고 문장 내 형태소 별로 분리하여 품사를 부착시켰다. 추출된 어휘들을 품사별로 재정리하고 조사, 감탄사, 접속사, 관형사 등 불용어를 제외하고. 명사, 형용사, 동사를 대상으로 R프로그램을 이용하여 출현빈도를 정리하였다. 빈도 높은 순으로 정리한 단어 중 사람, 생각, 오늘, 마음, 방법, 의견 등 주가 방향성 예측과 무관한 단어들을 제외하고, 투자, 주식, 종목, 시장, 돈, 매수 등 주식시장 감성 단어 후보들을 추출하였다.

게시판 참여자들이 많이 언급하는 추출된 단어들을 기초로 위 단어와 함께 같은 문장 내에서 자주 언급되어 조합을 이루는 단어들의 말뭉치 패턴을 분석하고 각 단어별 확률지

수를 파악하기 위해 PMI를 사용하였다. PMI는 확률론에 기초한 방법으로 두 확률변수의 연관성을 나타내는 지표이다. 즉 분석하고자 하는 두 단어의 의미극성이 비슷할 경우, 같은 문서 내에서 나타날 확률이 높다는 가정 하에 계산된다[25].

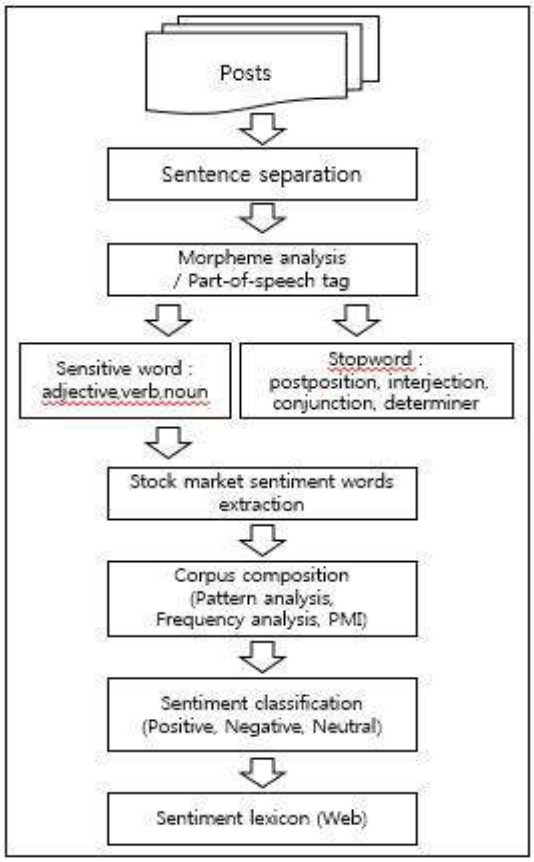

그림 1. 연구 순서도
Fig. 1. Research Flowchart

두 단어 간의 연관성을 계산하는 PMI의 계산식은 식 (1)과 같다.

$$PMI(w_1,w_2) = \log\frac{p(w_1,w_2)}{p(w_1)p(w_2)} \quad (1)$$

그림 2. 게시글 문장단위 구분
Fig. 2. Separate postings by sentence

여기서 w_1과 w_2는 분석하고자 하는 두 단어를 나타내며, PMI(w_1,w_2)는 두 단어가 같은 문서 내에서 나타날 확률을 나타낸다. PMI 값이 양수이면 두 단어가 같은 문서에 나타날 확률이 높아 비슷한 의미극성을 가진다는 뜻이며, 음수일 경우에는 그 확률이 낮아 다른 의미극성을 나타낸다고 볼 수 있다.

추출된 단어와 PMI 값이 양수를 가진 빈도가 높은 단어를 중심으로 주식시장 변동성을 예측하는 말뭉치를 구성하였다. 일례로, "종목"이라는 핵심어와 PMI값이 높게 나온 "가다"라는 단어로 말뭉치를 구성한 후, "뜨다" "오르다"처럼 유사한 의미로 많이 사용되는 말뭉치를 포함하고, "가지 못하다" "뜨지 못하다" "오르지 못하다"처럼 반대 의미를 가지는 말뭉치들도 포함하였다. 각각의 말뭉치에 주가 상승의 의미는 '긍정', 주가하락의 의미는 '부정', 방향성이 없는 경우는 '중립'의 감성 분류를 적용하는 과정을 반복하였고, 향후 지속적 연구의 용이성과 활용의 편의를 위하여 웹 기반으로 감성사전을 구축하였다.

그림 3은 '소비'라는 추출된 단어를 중심으로 한 말뭉치와 극성 값을 부여한 예이다.

No	단어	단어	단어	단어	단어	감성
1932	소비	기대	못미치			부정
1931	소비	증가	안			부정
1930	소비	개선	안			부정
1929	소비	회복	안			부정
1928	소비	저조				부정
881	소비	나빠지				부정
880	소비	좋아지				긍정
879	소비	지지부진				부정
878	소비	악화				부정
877	소비	개선				긍정
876	소비	실망				부정
875	소비	기대				긍정
874	소비	심리				중립
873	소비	지수				중립
872	소비	회복				긍정
871	소비	위축				부정
870	소비	감소				부정
869	소비	증가				긍정
868	소비					중립

그림 3. 주식시장 방향성 예측 말뭉치 구성
Fig. 3. Construct corpus to predict direction of stock market

Ⅳ. 주식시장 특화 감성사전 구축 결과

주식시장 방향성 예측을 위한 도메인 맞춤형 감성사전 구축에 관한 본 연구 과정에서 감성사전을 핵심적으로 등장하는 단어를 중심으로 말뭉치(corpus)방식으로 구축하였으며, 297개의 핵심어와 1,614개의 말뭉치를 포함하고 있다. 1,614개의 말뭉치는 주가상승을 의미하는 '긍정'이 731개, 주가하락을 의미라는 '부정'이 654개, '중립' 229개의 극성 값으로 구성되었다. 구축된 감성사전을 각 게시 글에 대입하여 긍정인지 부정인지 정도를 표시하고자 긍정의 극성 값을 가지는 말뭉치의 비율인 긍정지수로 산출하였으며, 그 계산식은 식 (2)와 같다.

$$\text{긍정지수} = \frac{\text{긍정 말뭉치의 수}}{\text{긍정 말뭉치의 수} + \text{부정 말뭉치의 수}} \quad (2)$$

긍정지수는 0에서 1의 값을 가지며 이 수치는 그 게시 글에 내재되어 있는 주식시장 방향성에 대한 감성을 나타낸다. 긍정과 부정의 말뭉치의 수가 동일한 경우인 0.5는 중립을 의미하고 0.5보다 크고 1까지는 긍정의 감성을 표시하며, 0.5보다 작고 0까지는 부정의 감성을 표시한다.

사전 구축을 위하여 활용된 전업투자자 게시판 3,809개의 글들에 주가 방향성 예측을 위하여 구축한 감성사전을 적용하였다. 매매일지의 작성과 그래프에 대한 설명, 종목의 나열 등 734개의 글에서는 말뭉치가 나타나지 않았으며, 게시 내용 문장 수가 작거나 시장 상황보다 심리 상황이 많은 투자자의 애로 사항 등의 글에는 말뭉치가 충분히 나타나지 않았다. 말뭉치의 빈도수가 작아 나타날 수 있는 감성 예측의 오류 가능성을 배제하고 검증하기 위하여 출현한 말뭉치의 수가 5개 보다 적은 경우를 제외한 1,828개의 게시 글에 대하여 감성사전을 적용하였다.

긍정지수 히스토그램은 그림 4와 같다.

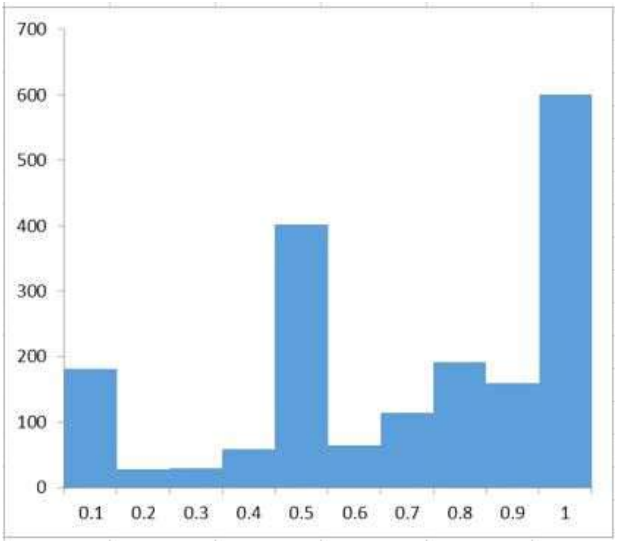

그림 4. 게시글 긍정지수 히스토그램
Fig. 4. Positive index Histogram

긍정 게시 글의 수는 1,128개, 부정 또는 중립의 게시 글 수는 700개 이며, 평균 긍정지수는 0.66998이며 메디안은 0.714로 나타났으며, 긍정지수가 긍정과 중립 및 부정 등으로 구분 분포되어 있어 긍정지수는 게시 글 별로 유의미한 차별성이 있음을 확인 하였다.

본 연구는 주식시장에 집중된 맞춤형 특화 사전을 구축함에 있어 말뭉치를 기반으로 하여 감성 분석을 시도하고, 기존 주식시장의 명사 또는 뉴스 분석을 통한 감성사전의 한계점을 보완 개선하려는 목적으로 진행하였다. 감성을 표현하는 다양한 글과 의견을 반영하기 위하여 주식 전문 커뮤니티의 게시 글을 지정하였고, 형용사를 포함한 말뭉치 별로 극성 값을 부여하는 방식으로 사전을 구축하여 유의성을 확인 하였다.

주가 방향성 예측에 시장 참여자들의 투자 심리와 감성 파악을 통한 참고 지표로서 의미를 확인하였으며, 주식시장의 다른 지표들과 복합적으로 분석된다면 방향성 예측에 더 크게 활용이 가능할 것이라 판단된다.

Ⅴ. 연구의 한계 및 향후 연구 방향

본 연구에서 제시한 주식시장 특화 감성사전의 한계점은 다음과 같다. 첫째, 유사어 사전의 필요성이다. 수많은 사람이 참여하는 주식시장에서 하나의 움직임에 대해 다양한 표현들이 존재하며, 이는 동일한 상황이라도 많은 다른 어휘들로 표현되는 한글의 특성이며 의사표현이 자유로운 인터넷 커뮤니티의 경우에는 더욱 다양하다. 이러한 표현들을 모두 아우르기 위해서는 한글 유사어 연구가 병행되어야 할 필요성이 있다. 두 번째로 연구 대상 범위의 확대 필요성이다. 인터넷 커뮤니티의 표현들 외에 전문적인 표현들도 포함될 필요가 있다. 다양성의 확보에서 한발 더 나아가 증권사 리서치 자료 등 전문적인 용어와 표현을 추가하면 더욱 정교한 감성사전을 만들 수 있을 것이다.

향후 연구에서는 상기 한계점을 보완하여 감성사전의 주가지수 방향성 예측력을 향상시키고, 나아가 실제 투자에 도움을 줄 수 있는 보조지표로 활용할 수 있는 감성사전을 구축할 수 있을 것으로 기대된다.

참고문헌

[1] C. Han, and K. Kim, "Twitter's impact on the election of TV debates," *Journal of Digital Contents Society*, Vol. 14, No.2, p.207-214, 2013

[2] C. Snijders, U. Matzat, and U. Reips, "Big data: Big gaps of knowledge in the field of Internet science," *International Journal of Internet Science*, Vol. 7, No. 1, pp. 1-5, 2012.

[3] B. Liu, Sentiment Analysis and Opinion Mining, *Morgan &*

Claypool, 2012.

[4] S. Ahn and S. B. Cho, “Stock prediction using news text mining and time series analysis,” *Proceedings of Korea Intelligent Information Systems Society Conference*, pp. 364-369, 2010.

[5] Y. Kim, N. Kim, and S. R. Jeong, "Stock-index invest model using news big data opinion mining," *KIIS Journal of Intelligence and Information Systems*, Vol. 18, No. 2, pp. 143-156, 2012.

[6] E. Yu, Y. Kim, N. Kim, and S. R. Jeong, "Predicting the direction of the stock index by using a domain-specific sentiment dictionary," *KIIS Journal of Intelligence and Information Systems,* vol. 19, No. 1, pp. 95-110, 2013.

[7] E. Cha and T. Hong, "S&P500 Stock price index prediction using news emotion analysis and SVM," *Proceedings of Korea Society of Management Information Systems Conference*, pp. 173-178, 2016.

[8] D. Kim, T. Cho and J. H. Lee, "A domain adaptive sentiment dictionary construction method for domain sentiment analysis," *Proceedings of the Korean Society of Computer Information Conference*, Vol. 23, No. 1, pp. 15-18, 2015.

[9] S. H. Lee, J. Cui and J. W. Kim, “Sentiment Analysis on Movie Review Through Building Modified Sentiment Dictionary by Movie Genre," *KIIS Journal of Intelligence and Information Systems*, Vol. 22, No. 2, pp. 97-113, 2016.

[10] B. Pang and L. Lee, Foundations and Trends® in Information Retrieval, Vol. 2, *now Publishers Inc*, 2008.

[11] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing," in *Proceeding of the 2nd International Conference on Knowledge Capture,* Sanibel Island, FL, USA, pp. 70-77, 2003.

[12] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in *Proceeding of the 12th International Conference on World Wide Web*, Budapest, Hungary, pp. 519-528, 2003.

[13] J.Lee, W. Lee, J. Park and j. Choi, “The Blog Polarity Classification Technique using Opinion Mining,” *Journal of Digital Contents Society,* Vol. 15, No 4, p.559-568, 2014

[14] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," *AAAI journal of American Association for the Artificial Intelligence*, Vol. 4, No. 4, pp. 755-760, 2004.

[15] S. M. Kim and E. Hovy, "Determining the Sentiment of Opinions," in *Proceeding of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics*, Geneva, Switzerland, No. 1367, 2004.

[16] A. Hassan and D. Radev, "Identifying Text Polarity Using Random Walks," in *Proceeding of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, Uppsala, Sweden, pp. 395-403, 2010.

[17] P. D. Turney and M. L. Littman(2002, May). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *arXiv preprint*[Online], cs/0212012, NRC-44929, pp. 1-9, Available: https://arxiv.org/ftp/cs/papers/0212/0212012.pdf

[18] J. An and H. W. Kim, "Building a Korean Sentiment Lexicon Using Collective Intelligence," *KIIS Journal of Intelligence and Information Systems*, Vol. 21, No. 2, pp. 49-67, 2015.

[19] E. Riloff and J. Shepherd, "A Corpus-based Approach for Building Semantic Lexicons," *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing,* pp. 117-124, 1997.

[20] V. Hatzivassiloglou and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives," in *Proceeding of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, pp. 174-181, 1997.

[21] H. Chen and D. Zimbra(2010, June). AI and Opinion Mining. *IEEE Journals ans Mazagines*[Online]. 25(3), pp. 74-80, 2010. Available: http://ieeexplore.ieee.org/abstract/document/5475086/

[22] S. Shin, Read Emotions in the Article! Understanding Emotional Analysis, *IDG Korea*, pp. 1-11, 2014.

[23] J. Song and S. Lee, “Automatic Construction of Positive/Negative Feature-predicate Dictionary for Polarity Classification of Product Reviews” *Journal of KISS: Software and Applications*, Vol. 38, No. 3, pp. 157-168, 2011.

[24] J. S. Jeong, D. S. Kim and J. W. Kim. "Influence Analysis of Internet Buzz to Corporate Performance : Individual Stock Price Prediction Using Sentiment Analysis of Online News," *KIIS Journal of Intelligence and Information Systems*, Vol. 21, No. 4, pp. 37-51, 2015.

[25] S. Song, D. Lee and S. Lee. “Identifying Sentiment Polarity of Korean Vocabulary Using PMI,” *Journal of Korea Information Science Society*, Vol. 37, No. 1(C), pp. 260-265, 2010.

김재봉(Jae-Bong Kim)

1990년 : 경희대학교 경영학과 학사
2015년~현 재: 고려대학교 빅데이터응용 및 보안학과(석사과정)

1990년~2016년: 현대증권
2017년~현 재: KB증권 디지털고객본부장
※관심분야 : 빅데이터분석, 오피니언마이닝, 머신러닝 등

김 형 중(Hyoung-Joong Kim)

1978년 : 서울대학교 전기공학과 학사
1986년 : 서울대학교 제어계측공학과(공학석사)
1989년 : 서울대학교 제어계측공학과(공학박사)

1989년~2006년: 강원대학교 교수
2006년~현 재: 고려대학교 정보보호대학원 교수
※관심분야 : 컴퓨터보안, 패턴인식, 가역정보은닉, 머신러닝, 빅데이터분석 등