주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안

저자
(Authors)

유은지, 김유신, 김남규, 정승렬

출처
(Source)

한국지능정보시스템학회 학술대회논문집 , 2012.12, 42-49 (8 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/Article/NODE02090331

APA Style

유은지, 김유신, 김남규, 정승렬 (2012). 주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안. 한국지능정보시스템학회 학술대회논문집, 42-49.

주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안

유은지, 김유신, 김남규, 정승렬

국민대학교 Business IT 전문대학원
서울특별시 성북구 정릉로 77
E-mail: nlj0123@naver.com, {trust, ngkim, srjeong}@kookmin.ac.kr

초록

최근 다양한 소셜미디어를 통해 생성되는 비정형 데이터의 양은 빠른 속도로 증가하고 있으며, 이를 저장, 가공, 분석하기 위한 도구의 개발도 이에 맞추어 활발하게 이루어지고 있다. 이러한 환경에서 다양한 분석 도구를 통해 텍스트 데이터를 분석함으로써, 기존의 정형 데이터 분석을 통해 해결하지 못했던 이슈들을 해결하기 위한 많은 시도가 이루어지고 있다. 뉴스에 대한 오피니언 마이닝을 통해 주가지수 등락 예측 모델을 제안한 최근의 연구는 이러한 시도의 대표적 예라고 할 수 있다. 이 연구를 비롯한 오피니언 마이닝을 다루는 많은 연구는, 각 어휘별로 긍정/부정의 극성을 규정해 놓은 감성사전을 사용하며, 한 문서에 나타난 어휘들의 극성 분포에 따라 해당 문서의 극성을 산출하는 방식을 채택한다. 하지만 특정 어휘의 극성은 한 가지로 고유하게 정해져 있지 않으며, 분석의 목적에 따라 그 극성이 상이하게 나타날 수도 있다. 본 연구는 특정 어휘의 극성은 한 가지로 고유하게 정해져 있지 않으며, 분석의 목적에 따라 그 극성이 상이하게 나타날 수도 있다는 인식에서 출발한다. 구체적으로는 주가지수의 상승이라는 한정된 주제에 대해 각 관련 어휘가 갖는 극성을 판별하여 주가지수 상승 예측을 위한 감성사전을 구축하고, 이를 기반으로 한 뉴스 분석을 통해 주가지수의 상승을 예측한 결과를 보이고자 한다.

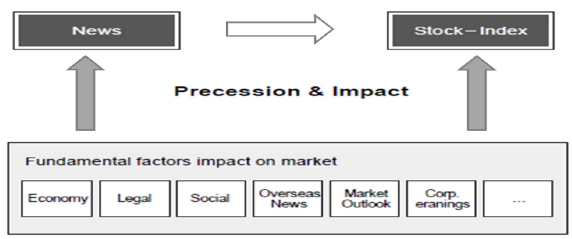
주제어:

감성사전 구축; 빅데이터 분석; 오피니언 마이닝; 주가지수 예측; 텍스트 마이닝

I. 서 론

최근 다양한 분야에서 무수히 많은 데이터가 실시간으로 발생하고 있다. 최근 여러 분야에 걸쳐서 화두가 되고 있는 빅데이터(Big Data)는 이와 같이 기존의 방법이나 도구로는 수집, 저장, 검색, 분석, 시각화가 어려운 정형 또는 비정형 데이터를 의미하며(Mckinsey, 2011), Gartner 그룹은 향후 유망 기술을 분석한 보고서(Gartner, 2011; Gartner, 2012)에서 2년 연속 빅데이터 관련 기술을 향후 2~5년 내에 IT분야에서 자리 잡을 주요 기술로 예상한 바 있다.

빅데이터에 대한 관심의 증가는, 다양한 분야에서 빅데이터를 분석하여 의미 있는 정보를 도출하기 위한 여러 시도들로 이어지고 있다. 이러한 시도의 대표적인 한 예로 뉴스에 대한 오피니언 마이닝(Opinion Mining)을 통해 주가지수의 등락을 예측하기 위한 지능형 투자의사결정 모형을 제안한 최근의 연구(김유신 외, 2012)를 들 수 있다. 기존의 연구(송치영, 2002)에서 밝혀진 바와 같이, 뉴스를 통해 현실 세계에서 일어나는 각종 현상에 대한 설명과 미래의 정치, 경제, 사회, 기업 등과 관련하여 앞으로 어떤 변화가 발생되고 진행되어 갈지 그에 대한 정보를 획득하는 것이 가능하기 때문에 뉴스와 주가는 밀접한 관계를 가진 것으로 볼 수 있으며, 뉴스를 통해 시장 참여자들은 주식 시장의 변동을 일부나마 예측할 수 있게 된다. 주가에 영향을 미치는 펀더멘털 요인(Fundamental Factors)들은 너무나도 다양하고 복잡하여 이러한 요인들이 뉴스와 주가에 영향을 미치고, 뉴스는 다시 주가에 영향을 미치는 <그림1>과 같은 식의 흐름이 발생하게 된다.

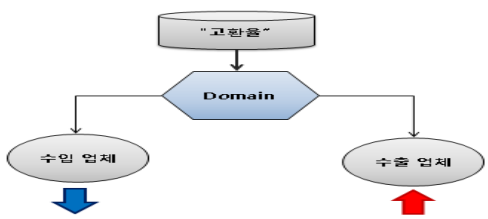


<그림1> 뉴스와 주가에 영향을 미치는 요인

<그림 1>에서 밝은 화살표의 흐름이 존재한다는 것은 뉴스의 분석을 통해 주가의 움직임을 예측할 수 있는 가능성이 존재함을 의미한다.

따라서 이러한 투사의사결정 과정을 체계화하고 일정 부분 자동화하기 위한 많은 연구가 수행되었다. 대표적인 연구로는 쉽게 판별이 가능한 특정 사건과 뉴스들을 위주로 그에 반응하는 주가 혹은 주가가 크게 변동되었을 때 이를 야기한 뉴스가 존재하였는지를 역으로 분석한 연구(송치영, 2002), 뉴스와 주가 변동성 간의 관계가 존재함을 확인한 연구(Mark, 1994)등을 들 수 있다.

특히 뉴스 분석을 통해 주가의 움직임을 예측하기 위한 최근의 많은 연구들은, 텍스트 마이닝(Text Mining)과 오피니언 마이닝을 활용하여, 각 뉴스에 속한 어휘의 극성을 근거로 각 뉴스의 극성을 산출하는 방안을 제안하고 있다. 이에 속하는 많은 연구들은 각 어휘별로 극성을 미리 규정해 놓은 범용 감성사전을 사용하며, 한 문서에 나타난 어휘들의 극성 분포에 따라 해당 문서의 극성을 산출하는 방식을 채택한다. 이러한 방식은 특정 어휘의 극성은 한 가지로 고유하게 정해져 있다는 가정에 근거한다. 하지만 특정 어휘의 극성은 한 가지로 고유하게 정해져 있지 않으며, 분석의 목적에 따라 어휘의 극성이 상이하게 나타날 수도 있다. 예를 들어 “고환율” 이라는 어휘는 수출 중심의 업체에게는 긍정적으로, 수입 중심의 업체에게는 부정적으로 인식될 수 있다(그림 2).



<그림 2> 분석 목적에 따른 어휘 극성의 상이성

이러한 이유로 범용 감성사전을 사용하는 오피니언 마이닝의 한계가 여러 연구에서 지적되어 왔으며, 고객이 남긴 상품평을 분석하여 상품 특징 감성사전을 구축하는 연구(송조석·이수원, 2011) 등 특정 도메인에 특화된 감성사전 구축에 대한 연구가 최근 수행되고 있다. 하지만 이처럼 주제에 특화된 감성사전을 구축할 수 있는 분야는 비교적 한정되어 있는데, 이는 이러한 주제지향 감성사전을 구축하기 위해서는 해당 어휘를 담고 있는 문서와 함께 어휘의 극성을 판별하기 위한 목적 값(Target Value)이 준비되어 있어야 하기 때문이다. 예를 들어 송조석·이수원(2011)에서는 각 상품에 부여한 평가 점수를 목적 값으로 정의하고 상품평에 포함된 각 어휘들이 해당 목적 값에 어떠한 영향을 주는지 분석하였다. 이와 같이 주제지향 감성사전의 구축을 위해서는 목적 값을 적절하게 정의하고, 관련 어휘들이 해당 목적 값에 미치는 영향을 학습 과정을 통해 파악하는 과정이 반드시 필요하다.

이러한 측면에서 뉴스의 분석하여 주가의 등락을 예측하기 위한 오피니언 마이닝의 경우, 주식 시장에 특화된 감성사전을 구축하기에 매우 적합한 영역인 것으로 판단된다. 이는 분석 대상이 되는 문서 입력으로 비교적 정제된 어휘로 구성된 경제뉴스를 활용할 수 있을 뿐 아니라, 목적 값으로서 매 거래일의 주가지수 등락 여부를 명확하게 파악할 수 있기 때문이다. 본 연구에서는 뉴스에 나타난 어휘들의 극성을 주가지수의 등락에 미치는 영향에만 근거하여 판별하고, 이를 기반으로 주가지수 등락 관점에서의 관련 어휘들의 감성사전을 구축하고자 한다. 또한 이렇게 구축된 감성사전을 이용한 문서 분석을 수행하여, 범용 감성사전에 근거하여 주가지수의 등락을 예측한 기존의 연구의 한계를 보완하고자 한다.

본 논문의 이후 구성은 다음과 같다. 다음 절인 2절에서는 본 연구의 이론적 배경이 되는 뉴스를 활용한 주가 예측과 오피니언 마이닝에 대한 기존 연구들을 간략하게 소개하고, 3절에서는 지능형 투자 결정 모형의 구조 및 주식 시장에 특화된 감성사전의 구축 방안에 대해 간략하게 설명한다. 이에 대한 실험 결과는 4절에서 제시하며, 마지막 절인 5절에서는 본 연구의 기여 및 한계 그리고 향후 연구 방향을 제시한다.

II. 관련연구

2.1. 뉴스를 활용한 주가예측에 관한 연구

주식 시장의 분석과 예측은 경제 분야뿐 아니라 수학과 통계, 전산분야에 이르기까지 오랫동안 매우 중요한 연구과제로 인식되어 왔으며, 특히 최근 금융공학의 발전과 함께 과학적 방법을 통한 주가 예측과 활용에 대한 연구(안성원, 2010)가 활발하게 수행되고 있다. Schumaker(2010)는 이러한 주가 예측 방법론을 수학적 예측, 통계학적 예측, 그리고 인공지능적 예측의 세 가지로 분류한 바 있으며, 뉴스로부터 패턴을 추출하고 이를 주가예측에 적용함으로써 전통적인 주가 예측 방법론의 한계를 극복하기 위한 최근의 연구 동향을 소개하고 있다. 뉴스를 이용한 주가 예측은 수집된 뉴스에 대한 텍스트 분석을 통해 각 뉴스 내의 의미 있는 어휘 또는 말뭉치(Bag of words)를 추출하고, 이를 분석하여 해당 뉴스가 주가에 호재인지 악재인지 분류한 후 그 결과를 이용하여 주가에 대한 움직임을 예측하는 과정(Schumaker, 2010; 안성원, 2010)으로 요약된다.

뉴스와 주가에 관한 비교적 초기의 연구들은 뉴스와 주가 사이에 존재하는 관계 자체를 규명하고자 노력하였다. Mark Mitchell & Harold Mulherin(1994)는 뉴스의 개수 및 매체 특성이 거래 규모 및 수익률과 매우 강하게 연관되어 있음을 확인하였고, Lee Gillam et al.(2002)은 뉴스로부터 감지되는 시장 분위기와 주가의 시계열 움직임 사이의 관계를 규명하였다. Tak-chung Fu(2008)는 홍콩주식시장에 상장된 9 개 개별

주가를 선정하여 그들 주가의 흐름과 중국어 뉴스의 긍정/부정 감성 사이의 상관관계를 분석하였다. 또한 안희준(2010)은 남북관계에서 일어나는 특정 사건들과 주식시장과의 상관관계를 분석한 결과 뉴스가 주가에 유의미한 영향을 미친 것으로 분석하였으며, 신종화·이의철(2010)은 인터넷 뉴스 매체를 통해 실시간으로 전파되는 뉴스들이 시장에 매우 큰 영향을 미치고 있음을 사례를 통해 밝힌 바 있다.

뉴스와 주가에 관한 최근의 연구들은 뉴스에 대한 분석을 통해 주가를 예측하고자 하는 방향으로 집중되어 있다. Mittermayer(2006)에 따르면 이러한 분야의 대부분 연구들은 텍스트 자동 분류기술에 기반하여 뉴스를 분석하고 이를 통해 주가, 환율 등의 가격 추이를 예측하는 것으로 목적으로 하고 있으며, 일반적으로 긍정, 부정, 그리고 중립의 세 가지 유형으로 뉴스를 구분하고 있는 것으로 파악되었다. Schumaker(2010)는 경제 뉴스에서 여러 번 반복되는 고유명사만을 추출하여 뉴스가 배포된 시점의 주가와 20 분 후의 주가에 대한 기계학습을 통해 20 분 후의 주가를 예측하는 기계학습시스템 AZFinText(Arizona Financial Text System)를 제안하였으며, 시뮬레이션을 통해 AZFinText 가 시장 평균 및 유사 펀드보다 높은 수익률을 달성함을 확인하였다. 또 다른 연구로는 웹에 공개된 개인투자자들의 의견들을 분석하여 감성 예측을 실시하고 각 저자에 대한 신뢰성을 측정한 후 기계학습을 통해 개별 주가의 움직임을 예측한 연구(Vivek Sehgal & Charles Song, 2007)가 있다. 또한 Woojin Paik(2007)은 개별 기업에 대한 긍정/부정 뉴스가 주가의 상승/하락에 영향을 미칠 것이라는 가정에 기반하여, 뉴스 텍스트로부터 추출한 긍정/부정 어휘패턴을 학습하여 주가의 방향성을 예측하는 실험을 수행하였다. 실험 결과 주가상승 시 69%, 주가하락 시 64%의 높은 예측 정확도를 보였으나, 최초의 긍정/부정 어휘패턴 추출을 사람이 수행하였다는 한계를 갖는다. 안성원 조성배(2010)는 뉴스가 발행한 이후의 가격이 이전 가격에 비해 ±2%로 변동이 생긴 경우 이 뉴스가 주가에 영향을 줄 수 있는 뉴스인 것으로 판단하였으며, 해당 뉴스 텍스트에서 추출한 특성을 주가의 변동폭과 대비하여 긍정/부정으로 분류한 뒤 이를 이용하여 개별 종목 주가의 상승/하락을 예측하는 모형을 제시하였다.

이와 같은 선행연구들은 뉴스가 어떠한 형태로건 주가에 영향을 미치고 있으며, 뉴스를 이용해 주가를 예측하고자 하는 시도들이 실제 투자 성과로 나타날 수 있다는 가능성을 보이고 있다. 이러한 오피니언을 파악함으로써 주가의 방향성을 예측하는 투자의사결정에 활용할 수 있음을 시사하고 있다.

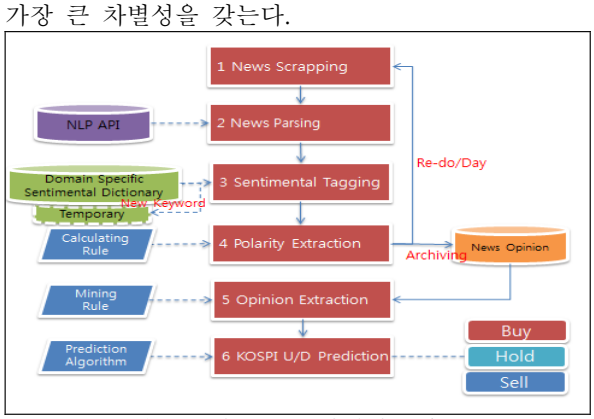
2.2. 오피니언 마이닝에 관한 연구

오피니언 마이닝은 온라인 뉴스와 소셜 미디어의 코멘트 등 사용자가 다양한 컨텐츠를 통해 표출한 의견을 추출, 분류, 이해, 자산화하는 과정을 의미하며, 감성분석(Sentimental Analysis)(Bing Liu, 2010; Hsinchun Chen & Divid Zimbra, 2010)의 다양한 기술을 활용하여 수행된다. 문서에 포함된 의견 정보를 오피니언 마이닝을 통해 추출하고 분류하는 연구는, 크게 의견의 의미방향을 분류하는 연구와 언어적 자원을 구축하는 연구로 나누어질 수 있다. 언어적 자원을 구축하는 연구는 개별 언어의 내용을 대상으로 어휘 또는 어의 수준에서 중립/긍정/부정 평가를 태깅해 놓은 감성사전을 사용하며, WordNet 의 각 어휘에 중립/긍정/부정 값을 태깅한 SentiWordNet 기반 연구가 대표적이다. 분석 대상 텍스트로부터 오피니언 또는 감성을 정확하게 추출하기 위해서는, 미리 구축된 감성사전에 각 어휘의 감성이 명확히 정의되어 있어야 한다. 하지만 언어의 사용 과정에서 발생하는 다양한 문맥적 의미 변이와 동태적 활용, 동음이의어의 존재 등으로 인해, 각 어휘의 감성을 명확히 식별하고 이에 대해 긍정 점수를 부여하는 작업은 매우 어려운 작업으로 인식되고 있다. 따라서 이러한 복잡한 상황을 고려하여 더욱 정교한 감성사전을 구축하고, 이를 통해 감성 평가의 정확성을 향상시키는 시도가 다양한 연구(김명규 외, 2009; 윤홍준, 2010; 정유철, 2008; 김진옥, 2011)에서 이루어진 바 있다. 특히 송종석(2011)은 어휘의 일반적인 의미를 기준으로 감성사전을 구축하고 긍정 점수를 부여하는 방식보다는, 도메인의 특징을 고려하여 주제에 특화된 감성사전을 구축하였을 때 감성 평가의 정확도가 더욱 향상되는 현상을 확인하였다. 오피니언 마이닝을 활용한 연구 중 본 논문에서 다루고자 하는 주제와 직접 관련된 논의는 김유신 외(2012)에서 찾을 수 있다. 김유신 외(2012)는 개별 기업의 주가가 아닌 주가 지수의 방향성을 예측하기 위해 오피니언 마이닝을 활용하였다는 점에서 주가 예측 관련된 기존의 연구와 차별성을 갖는다. 하지만 이 연구는 문서의 극성 판별에 가장 큰 영향을 미치는 감성사전의 구축에 대해서 명확하게 언급하고 있지 않다는 한계를 갖는다. 즉 이 연구는 주식 시장 영역에 특화된 감성사전이 아닌 범용 감성사전을 이용하여 오피니언 마이닝을 수행하였으며, 이로 인해 동일한 어휘가 도메인에 따라 긍정/부정으로 상이하게 인식될 수 있는 현상을 적절하게 반영하지 못했다는 한계를 갖는다. 이러한 한계를 극복하고 주가지수 방향 예측의 정확도를 향상시키기 위해서는, 주식 시장 영역에 특화된 감성사전을 구축하는 방안에 대한 연구가 반드시 수행되어야 한다.

III. 주제 지향 감성사전 기반의 주가지수 상승 예측 모형

3.1. 예측 모형 구조

본 연구에서 제안하는 지능형 투자의사결정 모형은 <그림 3>에 나타나있다. <그림 3>은 김유신 외(2012)에서 제안한 시스템을 수정 보완한 것으로, 기존 시스템이 범용 감성사전을 사용한 것에 비해 새로 제안하는 시스템은 주식시장 도메인에 특화된 감성사전을 별도로 구축하여 사용한다는 점에서 가장 큰 차별성을 갖는다.



<그림 3> 지능형 투자의사결정 모형

<그림 3>의 전체 모형의 각 단계를 설명하면 다음과 같다. 모형의 첫 단계에서는 분석 대상 뉴스를 수집한다. 이를 위해 국내 모든 경제관련 매체의 경제 뉴스가 모여있는 포털 사이트의 증권 섹션을 스크래핑 하는 방법을 사용하였다. 두 번째 뉴스 파싱 단계에서는 뉴스의 텍스트를 형태소 분석기 모듈에 의해 분해하며, 분해된 어휘에 대해 감성사전을 참조하여 극성을 부여하는 과정은 세 번째 단계인 감성 태깅 단계에서 수행된다.

감성사전 구축 과정은 다음 부절인 3.2 절에서 소개하기로 한다. 이후의 단계는 태깅된 어휘들의 극성에 따라 문서의 극성을 산출한 후, 그 결과에 따라 주가지수의 상승에 대한 확률을 제시하는 과정을 나타낸다.

3.2. 감성사전 구축

오피니언 마이닝에서 감성 어휘 사전의 역할은 매우 중요하다. 특히 경제 뉴스에 대한 분석을 통해 익일 주가지수의 상승여부를 판단하기 위해서는, 뉴스 기사를 구성하고 있는 각 어휘의 긍정/부정 영향에 대한 정확한 판단이 반드시 필요하다. 하지만 각 어휘의 긍정 점수는 어휘 자체의 고유 특성으로 부여되기 보다는, 사용되는 도메인과의 관계에서 유연하게 부여되는 것으로 파악하는 것이 바람직하다. 즉 일반적으로 부정적으로 인식되는 어휘가 주식시장에서는 긍정적으로 인식되는 경우도 있을 수 있다. 예를 들어 “ 금리 하락”의 경우 경기 침체를 나타내는 징표가 될 수도 있지만, 주식시장에서의 “ 금리 하락”은 일반적으로 호재로 인식된다. 따라서 본 연구에서는 뉴스로부터 주가지수의 움직임과 관련된 오피니언을 보다 정확하게 도출하기 위해, 주식 시장에 특화된 감성 어휘 사전을 구축하는 방안을 제시하고자 한다. 또한 이러한 감성 어휘 사전을 구축함에 있어서 사람의 인위적 판단을 최대한 배제하고 명확한 기준과 분석 과정을 제시함으로써, 유사 분야에서의 주제 지향적 감성사전 구축에 실마리를 제공하고자 한다.

<그림 4> 주가지수 등락 예측을 위한 감성사전

본 연구의 핵심인 주가지수 등락 예측에 필요한 감성사전 구축을 위해서는 주가지수 관련 어휘, 각 어휘가 출현한 뉴스의 식별자, 그리고 각 뉴스가 보도된 직후 주가지수의 움직임의 세 가지 정보가 필요하다. 이러한 세 가지 정보에 근거해서 주가지수 등락 예측을 위한 감성사전을 구축하는 전체 과정이 <그림 4>에 도식화되어 있다. 우선 대상 뉴스로부터 형태소 분석을 통해 명사만을 추출하고, 이들 어휘 중 Start List, 최소 빈도수 등의 기준을 사용하여 필터링을 통과한 어휘에 대해 긍정 지수를 계산한다. 각 어휘의 긍정 지수는 해당 어휘가 출현한 전체 뉴스 수 대비 해당 어휘가 출현한 뉴스 보도 후 주가지수가 상승한 회수의 비율로 정의되며 구체적인 산출 식은 다음 식과 같다. 하나의 뉴스의 전체 긍정 지수는 해당 뉴스에 출현한 어휘들의 긍정 지수를 평균한 값으로 산출된다.

$$Term(i,j) = \begin{cases} 1 & (Doc(j)\text{가 } Term(i)\text{를 포함하고, } Doc(j)\text{의 발행 익일 주가지수가 상승한 경우}) \\ 0 & (\text{그 외의 경우}) \end{cases}$$

$$Term(i).NumDocs = Term(i)\text{를 포함하고 있는 뉴스의 개수}$$

$$Term(i).Score = \frac{\sum_{j=1}^{n} Term(i,j)}{Term(i).NumDocs}$$

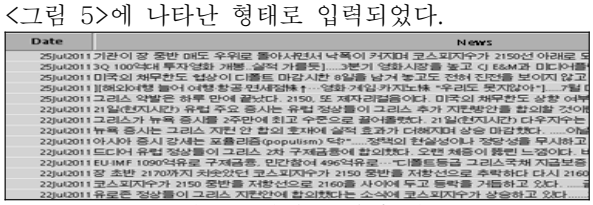
각 뉴스의 극성에 따라 주가지수의 상승 및 하락을 예측하기 위해서는 뉴스의 긍정/부정 구분을 위한 임계값의 설정이 필요하다. 본 연구에서는 판단 기준을 위해 특정 임계값을 제시하는 대신, 임계값의 변화에 따라 주가지수 등락의 예측 정확도가 변하는 양상을 보이고자 하며, 자세한 내용은 다음 절인 4절에서 소개하도록 한다.

IV. 실험

4.1 실험 데이터 소개

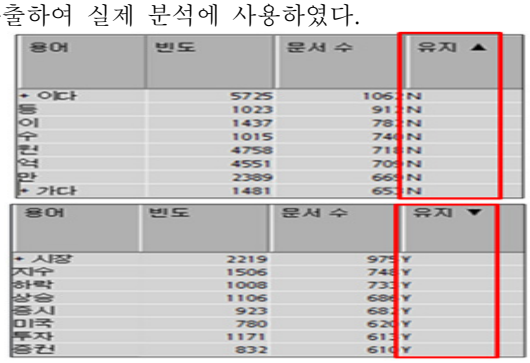
본 절에서는 3절에서 제안한 감성사전 구축 방법론 및 주가지수 등락 예측 모형의 예측력을 평가하기 위해, 포털 사이트 네이버의 증권정보 필드의 주요뉴스에 게재된 기사 중 M사와 E사의 기사 2011년 7월부터 9월까지 3개월간 1,072건을 대상으로 실험을 수행하고자 한다. 예측 대상이 되는 목표 변수는 3개월 동안 주식시장이 개장된 63일의 주가지수 등락 여부이며, 이 변수는 감성사전 구축 과정에서 가장 중요한 기준으로 사용된다. 뉴스 텍스트의 형태소 분석은 일부는 서울대학교 IDS 연구실에서 제공하는 꼬꼬마 형태소 분석기 API를 수정 보완한 것을 사용하고, 일부는 SAS Enterprise Miner를 사용하였다.

수집한 3개월 데이터의 기간 중 주식 시장은 휴일을 제외한 63일이 개장되었고, 그 중 상승 33일 하락 30일의 변동이 있었으며 뉴스 텍스트는 <그림 5>에 나타난 형태로 입력되었다.



<그림 5> 입력 데이터 일부

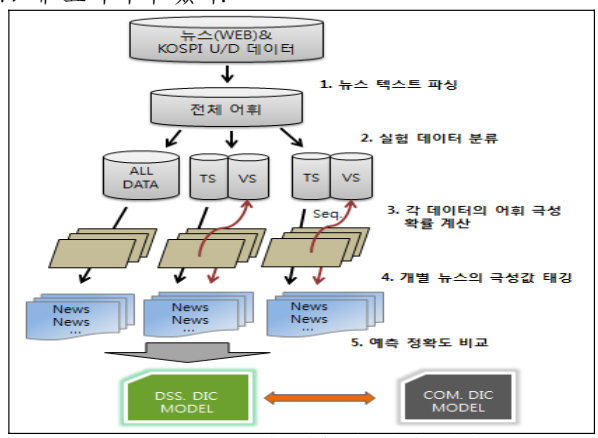
입력된 뉴스 텍스트 1,072건을 파싱하였다. 이 때, 파싱된 어휘 결과에서 명사를 제외한 모든 품사, 숫자, 한 음절 어휘를 제거하였고 이를 Start List로 정의하여 다시 파싱 과정을 수행하였다. SAS Enterprise Miner에서 Start List는 유지할 어휘들을 기록한 Data Set을 제공함으로써 구현된다. 그 결과 <그림 6>과 같이 "Y"로 나타난 어휘는 추후 분석 과정에 사용되며 "N"으로 나타난 어휘는 제외된다. 이러한 과정을 통해 파싱된 결과에서 최소 빈도수 1%를 만족하는 어휘만을 추출하여 실제 분석에 사용하였다.



<그림 6> Start List를 적용한 파싱 결과의 예

4.2 실험 모형

본 연구에서 수행한 실험의 전체 과정은 <그림 7>에 요약되어 있다.

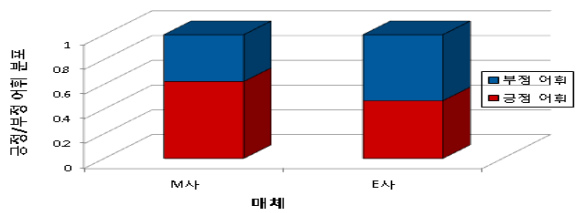


<그림 7> 제안 모형의 예측력 평가 실험 개요

전체적으로 범용 감성사전을 사용하여 구현한 예측 모형과, 주식시장 도메인에 특화된 감성사전을 구축하고 이를 통해 구현된 예측 모형과의 예측 정확도를 비교하는 실험을 수행하였다. 또한 문서의 극성에 따라 주가지수의 등락을 예측하는 경우 그 기준이 되는 임계값의 설정이 매우 중요하므로, 임계값의 변화에 따라 주가지수 등락의 예측 정확도가 변하는 양상을 보이고자 하였다. 유사한 실험을 세 종류의 세트로 수행하였는데, 이는 학습 데이터(Training Data)와 검증 데이터(Validation Data)의 분할 방법에 따른 것이다.

4.3 결과 분석

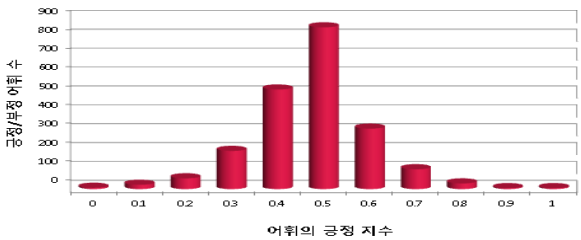
본 실험에 사용된 뉴스에 대해 제공하는 매체별 긍정/부정 어휘의 분포를 살펴본 결과가 <그림 8>에 나타나있다.



<그림 8> 매체별 긍정/부정 어휘 분포

긍정 지수는 최대 1.0, 최소 0.0의 값을 가지며 1.0인 경우 해당 어휘의 극성이 완전 긍정으로, 0.0인 경우 해당 어휘의 극성이 완전 부정으로 해석된다. 두 매체 모두 긍정/부정 어휘의 분포가 비교적 균등하게 나타났지만, 매체 M사의 뉴스에서는 부정적 어휘보다 긍정적 어휘의 분포가 높게 나타났으며 매체 E사에서는 부정적 어휘가 긍정적 어휘의 분포보다 다소 높게 나타났다.

각 어휘의 긍정 지수에 따른 분포는 <그림 9>에서 확인할 수 있다. 분포에 따르면 어휘의 긍정 지수는 극단적으로 긍정 혹은 부정을 나타내는 경우는 매우 적으며 0.3에서 0.7 사이의 긍정 지수를 갖는 어휘가 대부분인 것을 확인할 수 있다. 특히, <그림 9>에서 0.5의 긍정 지수를 가진 어휘가 많은 것을 알 수 있는데 이는 문서의 극성 구분 시 임계값의 설정이 매우 중요함을 암시하고 있다. 이와 같은 분포를 따르는 어휘에 대해서 긍정 극성, 부정 극성, 그리고 중립성을 갖는 대표적인 어휘의 예가 <그림 10>에 나타나있다.



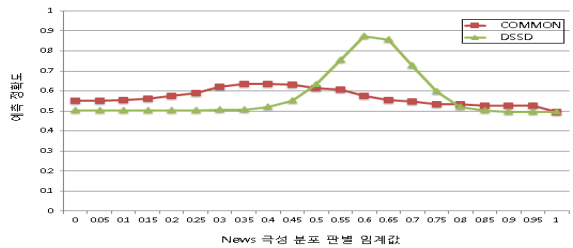
<그림 9> 어휘의 긍정지수 분포

긍정 어휘		중립 어휘		부정 어휘	
어휘	긍정 지수	어휘	긍정 지수	어휘	긍정 지수
진출	0.7	가능	0.5	역부족	0.1
안도	0.7	가능	0.5	급락세	0.2
급반등	0.7	가량	0.5	난항	0.2
안도감	0.7	개월	0.5	내림세	0.2
이익률	0.7	거래일	0.5	두려움	0.2
상승률	0.7	경우	0.5	심각	0.2
낙관론	0.7	기간	0.5	하락률	0.2
낙관적	0.7	너스	0.5	먹구름	0.3
희망	0.6	년간	0.5	불투명	0.3
호재	0.6	년래	0.5	쇼크	0.3
호전	0.6	누구	0.5	직격탄	0.3
호조	0.6	다음	0.5	추락	0.3
호평	0.6	다음날	0.5	침체	0.3
확장	0.6	모습	0.5	감소	0.4
상승세	0.6	사실	0.5	부정적	0.4
성공	0.6	사실상	0.5	불안	0.4
긍정적	0.6	수년	0.5	불안감	0.4
기대감	0.6	시각	0.5	위축	0.4
오름세	0.6	시간	0.5	위험	0.4

<그림 10> 주제지향 감성사전의 일부

이렇게 구축된 주제지향 감성사전을 통해 세 가지 실험을 수행하였다. 첫 번째 실험은 학습 데이터와 검증 데이터의 구분 없이 하나의 데이터 세트에 대한 예측 정확도 평가 실험이다. 뉴스 문서로부터 텍스트 형태소 분석을 통해 명사들을 추출하고, Start List, 최소 빈도수 필터링을 통해 대상어휘들을 선정한다. 이렇게 선정된 어휘들의 긍정 지수는 해당 거래일의 주가등락 여부를 근거로 산출된다. 또한 각 뉴스 문서의 긍정 지수는 포함된 어휘들의 긍정 지수의 평균 값으로 계산된다. 이렇게 도출된 문서의 긍정 지수가 정해진 임계값보다 큰 경우 익일 거래일의 주가지수를 상승으로 예측하고, 그렇지 않은 경우 하락으로 예측하였다. 그리고 전체 예측 건수 중 상승/하락을 정확하게 예측한 비율을 예측 정확도로 정의하였다.

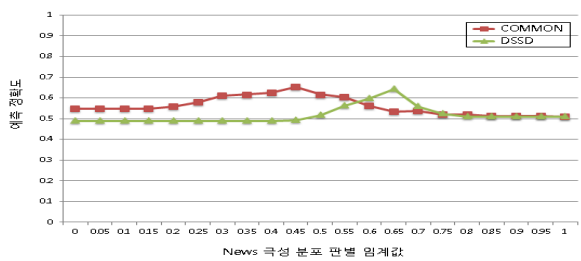
이렇게 수행된 실험의 결과가 <그림 11>에 나타나있다.



<그림 11> 하나의 데이터 세트에 대한 예측 정확도 비교 실험

<그림 11>에서 범용 감성사전을 사용한 경우는 문서의 판별 임계값의 변화에 따라 예측 정확도가 크게 영향 받지 않았으며, 문서의 극성이 0.4를 넘는 경우를 상승으로 예측한 경우 정확도가 가장 높게 나타났다. 본 연구에서 구축한 주제지향 감성사전을 사용한 경우는 판별 임계값의 변화에 따라 예측 정확도가 비교적 큰 폭으로 변화했으며, 임계값을 0.6으로 사용한 경우 가장 높은 예측력을 나타냈다. 하지만 이 실험은 동일한 데이터 세트에 대해 학습과 검증을 수행하므로 과적합화(overfitting)의 가능성이 존재하게 된다. 따라서 학습과 검증 데이터 세트를 구분하여 수행한 실험의 결과가 <그림 12>에 나타나있다.

(a) 학습 데이터 세트에서의 예측 정확도

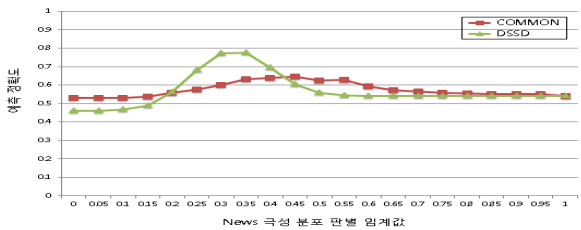


(b) 검증 데이터 세트에서의 예측 정확도
<그림 12> 학습(70%)/검증(30%) 데이터 세트에 대한 예측 정확도 비교 실험

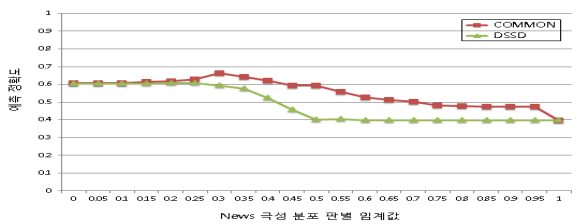
<그림 12>는 무작위 추출을 통해 학습을 위한 데이터로 전체 데이터의 70%를, 검증용 데이터로 나머지 30%를 사용한 실험의 결과를 보여준다. 실험 결과 학습 데이터 세트에 대한 예측 정확도는 <그림 11>의 결과와 유사하게 나타났다. 검증 데이터 세트에 대한 실험의 경우 범용 감성사전을

사용한 모형과 주제지향 감성사전을 사용한 모형에서 최대 정확도를 갖는 임계값이 서로 다르게 나타났을 뿐, 최대 정확도는 서로 유사하게 나타났다. 즉 주제지향 감성사전을 사용한 경우 학습 데이터 세트에서의 예측 정확도에 비해 검증 데이터 세트에서의 예측 정확도가 현저히 저하되는 현상이 발견되었다. 이는 범용 감성사전을 사용한 경우는 각 어휘에 대해 이미 극성이 부여되어 있으므로 과적합화로 인한 부작용이 나타나지 않는 반면, 주제지향 감성사전의 경우 분석 대상 뉴스에 따라 어휘의 극성이 달라지므로, 상대적으로 과적합화로 인한 부작용이 크게 나타났기 때문인 것으로 판단된다.

이러한 현상은 세 번째 실험 결과인 <그림 13>에서 더욱 크게 나타난다. 주가지수 등락 예측은 그 도메인의 특성상 과거의 정보를 통해 모형을 수립하고, 해당 모형을 현재 주어진 정보에 적용함으로써 직후의 미래를 빠르게 예측해야 할 필요가 있다. 세 번째 실험은 이를 반영한 것으로, 전체 뉴스 중 시기적으로 앞선 70%의 뉴스를 학습 데이터 세트로, 나머지 30%의 뉴스를 검증 데이터 세트로 사용한 실험이다.



(a) 학습 데이터 세트에서의 예측 정확도



(b) 검증 데이터 세트에서의 예측 정확도

<그림 13> 시간에 따른 학습/검증 데이터 세트 분할 실험 결과

<그림 13>의 경우 학습 데이터 세트에 대한 실험 결과조차 <그림 11>의 전체 데이터 세트에 대한 결과와 다르게 나타났다. 또한 주제지향 감성사전을 사용한 모형의 검증 데이터 세트에서의 정확도는 더욱 저하된 것으로 나타났다. 이러한 결과는 학습 데이터 세트와 검증 데이터 세트의 이질성이 매우 크게 존재했기 때문인 것으로 파악된다. 즉 학습 데이터 세트에서 긍정의 의미로 파악된 어휘가 검증 데이터 세트에서 부정의 의미를 갖는 경우, 또는 그 반대의 경우가 존재함으로써, 검증 데이터 세트에서의 정확도가 낮게 나타난 것으로 판단된다. 예를 들어 학습 데이터 세트의 수집 기간에는 '그리스' 라는 어휘가 긍정의 의미를 나타내었으나, 이후 그리스의 경제 상황이 악화되어 검증 데이터 세트의 수집 기간에는 동일한 어휘가 강한 부정의 의미를 나타내는 경우가 충분히 존재할 수 있을 것으로 판단된다.

위의 세 가지 실험을 통해 나타난 결과를 요약하면 다음과 같다. 우선 동일 데이터 세트에 대한 실험에서는 범용 감성사전에 비해 주제지향 감성사전을 사용한 예측 모델의 정확도가 다소 높게 나타났는데, 이는 동일한 어휘의 감성지수가 주제에 따라 서로 다르게 나타날 수 있는 현상에 기인한 것으로 파악된다. 하지만 제안한 모델의 정확도는 학습 데이터와 검증 데이터를 구분하여 수행한 실험에서 매우 낮게 나타나는데, 이는 제안 모델에서 과적합화 현상으로 인한 부작용이 매우 크게 나타났기 때문인 것으로 판단된다. 본 실험은 3개월이라는 다소 짧은 기간의 데이터를 사용하였기 때문에, 이러한 부작용이 더욱 크게 나타난 것으로 판단되며, 향후 충분히 긴 기간의 많은 데이터에 대한 실험을 통해 이러한 현상을 재검증할 필요가 있다. 또한 시기에 따라 학습 데이터와 검증 데이터를 구분하여 수행한 실험에서는 이러한 부작용이 더욱 크게 나타났는데, 이는 두 데이터 세트 간에 시간 간격이 존재함으로써 동일한 어휘가 두 세트에서 갖는 긍정 지수가 서로 다르게 나타났기 때문인 것으로 파악된다. 이러한 부작용 역시 충분히 긴 기간의 데이터 분석을 통해 다소 완화될 수 있을 것으로 기대된다.

V. 결론

본 논문에서는 주식관련 뉴스에 대한 오피니언 마이닝을 통해 주가지수의 등락을 예측하기 위해, 주식 도메인에 특화된 주제지향 감성사전을 구축하고 이를 활용하는 방안을 제시하였다. 구체적으로는 주식관련 뉴스에 포함된 어휘를 입력으로 사용하고 익일의 주가지수 등락 여부를 목적 변수로 사용하여 각 어휘의 극성을 도출한 주식 도메인에서의 감성지수 사전을 구축하였다. 또한 이렇게 구축된 사전을 통해 주가지수의 등락을 예측하는 모형을 수립하고, 그 예측 정확도를 범용 감성사전을 사용한 모형의 예측 정확도와 비교 평가하는 실험을 수행하였다. 실험 결과 하나의 데이터 세트에 대한 실험에서는 제안 모형의 정확도가 우수하게 나타났으나, 학습 데이터와 검증 데이터를 구분하여 수행한 실험에서는 제안 모형이 과적합화 문제로 인해 정확도가 오히려 낮게 나타나는 현상을 발견하였다.

본 연구의 가장 큰 기여는 주식 도메인에 특화된 감성사전을 구축하기 위한 하나의 기준을 제시하였다는 점에서 찾을 수 있다. 즉 각 어휘의

감정지수를 도출하기 위해서는 긍정/부정 여부를 판단하기 위한 목적 변수가 반드시 필요한데, 이러한 목적 변수로 주가지수의 등락 여부를 사용할 수 있는 가능성을 제시하였다. 한편 본 연구는 제안 모형의 평가를 위한 실험 부분에서 많은 한계를 보이고 있다. 우선 주식관련 뉴스의 특성상 수집 시기에 따라 각 어휘의 극성이 크게 달라질 수 있으므로, 예측 정확도를 높이기 위해서는 충분히 오랜 기간에 걸쳐 수집된 많은 뉴스에 대한 실험이 필요하다. 또한 분석 대상 어휘의 선정 과정에서 명사만을 추출하여 사용했는데, 명사 이외의 다른 품사도 고려한 실험이 수행되어야 한다. 이와 관련하여 하나의 어휘가 아닌 말뭉치, 또는 동시 출현 어휘에 대한 감성지수를 평가하는 방안도 모색되어야 한다. 마지막으로 주식 도메인에서 비교적 일관된 감성지수를 갖는 어휘와 짧은 기간 간격으로 변화하는 감성지수를 갖는 어휘를 식별하여, 문서의 극성 도출 시 이를 차별적으로 반영하기 위한 연구가 반드시 수행되어야 한다.

참고문헌

[1] 김유신, 김남규, 정승렬, "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형," 지능정보연구, 18(2), 2012, pp. 143-156.

[2] 김명규, 김정호, 차명훈, 채수환, "텍스트 문서 기반의 감성 인식 시스템," 감성과학, 12(4), 2009, pp. 433-442.

[3] 김진옥, 이선숙, 용환승, "한글 텍스트의 오피니언 분류 자동화 기법," 정보과학회논문지: 데이터베이스, 38(6), 2011, pp. 423-428.

[4] 송종석, 이수원, "상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축," 정보과학회논문지: 소프트웨어 및 응용, 38(3), 2011, pp. 115-177.

[5] 송치영, "뉴스가 금융시장에 미치는 영향에 관한 연구," 국제경제연구, 8(3), 2002, pp. 1-34.

[6] 신종화, 이의철, "한국형 경제 뉴스 속보가 금융시장에 미친 영향," 한국경영학회 통합학술대회, 2010, pp. 79-109.

[7] 안성원, 조성배, "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측," 한국컴퓨터종합학술대회 논문집, 27(1), 2010, pp. 364-369.

[8] 안희준, 전승표, 최종범, "남북관계 관련 뉴스가 주식시장에 미치는 영향," 한국금융연구원: 한국경제의분석, 16(2), 2010, pp. 199-231.

[9] 윤홍준, 김한준, 장재영, "오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법," 정보과학회논문지: 컴퓨팅의 실제 및 레터, 16(2), 2010, pp. 135-259.

[10] 정유철, 최윤정, 맹성현, "감정 기반 블로그 문서 분류를 위한 부정어 처리 및 단어 가중치 적용 기법의 효과에 대한 연구," 인지과학, 19(4), 2008, pp. 477-497.

[11] B. Liu, "Opinion Mining," Department of Computer Science University of Illinois at Chicago, 2010.

[12] F. L. Chung, "Chak-man Ng: Discovering the Correlation between Stock Time Series and Financial News," Web Intelligence, 1, 2008, pp. 880-883.

[13] Gartner, "Gartner identifies the top10 strategic technologies for 2011," 2010.

[14] Gartner, "2012 Hype Cycle for Emerging Technologies," 2012.

[15] G. Lee. "Economic News and Stock Market Correlation: A Study of the UK Market," Conference on Terminology and Knowledge Engineering, 2002.

[16] H. Chen and D. Zimbra, "AI and Opinion Mining," IEEE Intelligent Systems, 25(3), 2010, pp. 74-80.

[17] M. A. Mittermayer, G. Knolmayer, "Text Mining Systems for Market Response to News: A Survey," The Institute of Information Systems Working Papers, 2006.

[18] M. L. Mitchell and J. H. Mulherin, "The Impact of Public Information on the Stock Market," The Journal of Finance, 49(3), 1994, pp. 923-950.

[19] Mckinsey and Company, "Big Data: The next Frontier for Innovation, Competition, and Productivity," 2011.

[20] R. P. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System," ACM Transactions on Information Systems, 27(2), 2009.

[21] T. C. Fu, K. K. Lee, D. C. M. Sze, and F. L. Chung, "Chak-man Ng: Discovering the Correlation between Stock Time Series and Financial News," Web Intelligence, 1, 2008, pp. 9-12.

[22] W. Paik, M. H. Kyoung, K. S. Min, H. R. Oh, C. Lim, and M. S. Shin, "Multi-stage News Classification System for Predicting Stock Price Changes," 한국정보관리학회: 정보관리학회지, 24(2), 2007 pp. 123-141.

[23] V. Sehgal and C. Song, "SOPS: Stock Prediction using Web Sentiment Department of Computer Science University of Maryland College Park, Maryland, USA," Seventh IEEE International Conference on Data Mining: Workshops", 2007, pp. 21-26.