ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 June: 23(2): 39~70 http://dx.doi.org/10.13088/jiis.2017.23.2.039

잠재 토픽 기반의 제품 평판 마이닝*

박상민
군산대학교 산학융합공과대학
소프트웨어융합공학과
(*b1162@kunsan.ac.kr*)

온병원
군산대학교 산학융합공과대학
소프트웨어융합공학과
(*bwon@kunsan.ac.kr*)

최근 여론조사 분야에서 데이터에 기반을 둔 분석 기법이 널리 활용되고 있다. 기업에서는 최근 출시된 제품에 대한 선호도를 조사하기 위해 기존의 설문조사나 전문가의 의견을 단순 취합하는 것이 아니라, 온라인상에 존재하는 다양한 종류의 데이터를 수집하고 분석하여 제품에 대한 대중의 기호를 정확히 파악할 수 있는 방안을 필요로 한다. 기존의 주요 방안에서는 먼저 해당 분야에 대한 감성사전을 구축한다. 전문가들은 수집된 텍스트 문서들로부터 빈도가 높은 단어들을 정리하여 긍정, 부정, 중립을 판단한다. 특정 제품의 선호를 판별하기 위해, 제품에 대한 사용 후기 글을 수집하여 문장을 추출하고, 감성사전을 이용하여 문장들의 긍정, 부정, 중립을 판단하여 최종적으로 긍정과 부정인 문장의 개수를 통해 제품에 대한 선호도를 측정한다. 그리고 제품에 대한 긍·부정 내용을 자동으로 요약하여 제공한다. 이것은 문장들의 감성점수를 산출하여, 긍정과 부정점수가 높은 문장들을 추출한다. 본 연구에서는 일반 대중이 생산한 문서 속에 숨겨져 있는 토픽을 추출하여 주어진 제품의 선호도를 조사하고, 토픽의 긍·부정 내용을 요약하여 보여주는 제품 평판 마이닝 알고리즘을 제안한다. 기존 방식과 다르게, 토픽을 활용하여 쉽고 빠르게 감성사전을 구축할 수 있으며 추출된 토픽을 정제하여 제품의 선호도와 요약 결과의 정확도를 높인다. 실험을 통해, K5, SM5, 아반떼 등의 국내에서 생산된 자동차의 수많은 후기 글들을 수집하였고, 실험 자동차의 긍·부정 비율, 긍·부정 내용 요약, 통계 검정을 실시하여 제안방안의 효용성을 입증하였다.

주제어 : 토픽 모델, 오피니언 마이닝, 텍스트 요약, 데이터 분석, 여론조사

논문접수일 : 2017년 3월 8일 논문수정일 : 2017년 5월 4일 게재확정일 : 2017년 5월 15일
원고유형 : 일반논문 교신저자 : 온병원

1. 서론

2016년 아시안 리더십 컨퍼런스에 참여한 짐 클리프턴 갤럽 회장은 “모바일 기기의 확산으로 인해 갈수록 사람들은 여론조사에 덜 협조적이어서 여론조사로 앞일을 예측하는 것이 과거 그 어느 때보다 어렵다”라고 밝혔다. 그러면서, “설문조사를 통해 얻는 데이터 자체는 이제 큰 쓸모가 없고 신뢰하기도 어렵다. 빅데이터 분석을 통해 의미를 파악하고 새로운 발견이나 해법을 제공하는 것이 미래의 갤럽이 할 일”이라고 언급했다(On, H.Y., 2015). 이와 같이 조사원이 여론조

* 이 논문은 2016년도 한국정보처리학회 춘계학술발표대회에서 ‘빅데이터 분석 기반의 제품 평판 마이닝 알고리즘’의 제목으로 발표된 논문을 확장한 것임.
이 논문은 2016년도 정부(미래창조과학부)의 한국연구재단의 중견연구자지원사업(No. NRF-2016R1A2B1014843)의 연구비 지원으로 수행하였습니다.

사를 직접 수행하고 결과를 분석하는 기존의 전통적인 방식은 효율성이 크게 떨어지고 있다. 조사원과 통계 전문가를 활용하기 위해서는 많은 비용이 들어가고, 같은 내용이라도 설문 문항의 차이로 인해 다른 결과를 얻거나 설문 작성자의 주관적인 판단이 들어갈 위험성이 있다. 무엇보다 중요한 것은 표본의 크기가 작고 응답률이 높지 않으면 모집단을 추정하는데 많은 왜곡이 가해질 수 있어 궁극적으로 결과에 대한 신뢰를 얻을 수 없다. 게다가 여론조사를 신속하게 진행하는 것은 앞날을 예측하는데 매우 중요하지만, 사람에 의한 조사는 빠른 시간 내에 결과를 얻기란 쉽지 않는 것이 현 실정이다. 이러한 여러 제약을 해결하기 위해서 최근에는 빅데이터 분석 기법을 도입하여 인터넷상에 있는 정보를 수집하고 통계 분석을 통해 여론조사 결과를 도출하려는 흐름이 일고 있지만, 구체적인 방법론에 대한 심도 있는 연구가 학계 차원에서 이루어진 적은 없다.

본 논문에서는 다양한 여론조사 분야 중에서, 특정 제품에 대한 대중의 선호도를 파악하는 텍스트 마이닝 방안을 제안한다. 이러한 문제를 '제품 평판 마이닝'(Product Reputation Mining)이라고 명명하고, 구체적인 방법론과 특정 제품에 대한 사례 연구를 통하여 제안 방안의 효용성을 입증하고, 실험 결과로부터 의미 있는 정보를 추출하고자 한다. 먼저 사례 연구를 위해 국내에서 생산된 자동차를 분석 대상으로 선정하고, 본 논문에서 제안하는 방안을 적용하여 얻어진 결과를 분석하였다. K5, SM5, 아반떼 등 국내 자동차는 그 어떤 제품보다 대중의 관심이 많은 아이템이며, 지난 수십 년 동안 꾸준하게 팔린 베스트셀러이자 스테디셀러이다. 따라서 대중의 관심이 많고 인터넷상에 관련 정보도 쉽게 얻을 수 있다. 특히 대중의 관심이 바로 투영되어 나타나는 곳은 사용자 후기 게시판이다. 자동차를 구입하거나 관심이 많은 고객들은 후기 사이트에서 정보를 얻거나 제품의 단점을 지적하여 많은 사람의 여론을 환기시킨다. 본 연구에서는 국내 최대 자동차 후기 게시판으로 널리 알려진 '보배드림' 웹 사이트로부터 사용자 후기 게시판에 실린 텍스트 데이터를 수집하였다(Bobaedream, access in 2016). 제안 알고리즘을 보배드림과 같은 문서 코퍼스(document corpus)로부터 주요 토픽을 추출한다. 토픽의 예로서 자동차의 성능, 주행, 디자인 등을 둘 수 있다. 각 토픽에 대해 구체적으로 어떤 이야기들이 회자 되는지를 요약해서 보여주는 알고리즘이 필요하다. 예를 들면, 보배드림에서는 자동차의 디자인이라는 토픽에서 구체적으로 어떤 이야기들이 사람들에게 많이 언급되는지를 파악하는 것이다. 또한 토픽의 긍·부정을 판단하여 대중의 제품에 대한 선호도를 토픽별로 쉽게 알 수 있으며, 제품을 만든 회사 경영진에게는 중요한 정보가 될 것이다. 예를 들어, 아반떼 자동차의 디자인 토픽에 대한 긍·부정의 통계 수치를 계산하고, 어떤 점이 긍정이고 부정인지를 파악할 수 있다면, 제품을 개선하거나 제품 홍보를 하는데 크게 도움이 될 것이다.

이 논문에서 제안하는 방안의 우수성은 다음과 같다.

1) 문서에 숨어있는 토픽들을 추출하여 주어진 제품의 선호도를 조사하고, 토픽의 긍정적인 내용과 부정적인 내용을 요약하여 보여주는 제품 평판 마이닝 알고리즘을 처음으로 제안한다.
2) 일반적으로 특정 도메인의 감성사전 구축은 모든 문서들의 단어 빈도를 고려하여 긍·부정 단어를 식별한다. 반면 제안방안에

서는 토픽 내의 단어들의 확률 값(토픽 내의 중요도를 수치화한 값) 순서대로 긍·부정 단어를 식별하기 때문에 빠르게 감성사전을 구축할 수 있다.

3) 기존 토픽 모델링은 다른 텍스트 클러스터링 방법에 비해 우수한 성능을 보이지만, 추출된 토픽 결과들은 완벽하지 않기 때문에 정제 작업이 필요하다. 제안방안은 토픽 분할과 합병 방식을 통해 추출된 토픽들을 정제하여 제품의 선호도와 요약 결과의 정확도를 높인다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구의 기초가 되거나 관련 있는 연구를 정리하여 소개하였다. 3장 제안 방안에서는 제품의 평판 마이닝 알고리즘에 대해 구체적으로 설명한다. 4장에서는 실험 환경 및 결과에 대해 자세히 논의한다. 5장에서 결론 및 향후 연구 방향을 다룬다.

2. 관련연구

영상과 텍스트 등의 다양한 비정형 데이터로부터 소셜 이벤트를 분석하기 위해 MMTOM (Multi-modal Multi-view Topic-Opinion Model)이라는 방안을 제안하였다(Qian, S., Zhang, T., and C. Xu, 2016). 이 연구는 사회 이슈와 관련 있는 토픽(예: 시리아 내전)에 대해 각 집단(미국과 러시아 등)의 개별적 견해들과 공통적인 견해를 분류한다. 그리고 각 견해는 영상과 감성 키워드 위주로 출력된다. 이 연구는 제안방안과 비슷하게 토픽 모델 방법을 사용하지만 제품의 평판을 마이닝 하기 위해 토픽을 추출하고 긍·부정 비율을 측정하고, 감성 요약을 하는 방안과는 큰 차이가 있다.

온라인 상에 많이 존재하는 리뷰들은 사용자들의 제품 후기로 특정 제품에 대한 품평 정보가 담겨 있다. 이러한 리뷰에서 유용한 정보를 얻을 수 있다면 소비자와 제조회사에 큰 도움을 줄 수 있다. 온라인 소비자들의 의견이 담겨 있는 콘텐트의 품질 마이닝을 통해 유용한 리뷰와 그렇지 않은 리뷰를 자동으로 식별하는 연구를 수행하였다(Zeng, Y., Ku et al., 2014).

토픽 모델 방법을 통해 추출된 토픽의 레이블링(labeling)을 자동으로 수행하여 토픽의 의미를 쉽게 알 수 있는 방안을 제안하였다(Wan, X. and T. Wang, 2016).

많은 연구자들이 다양한 구매 프로세스를 연구하고 있으나 소비자 관점에서의 연구가 부족하다는 것을 파악하고, 각 구매 프로세스에서 고객들의 불편 사항을 구매자의 후기를 분석하여 해결한다. 오피니언 마이닝 방법을 적용하여 국내외 의류 제품에 대한 고객 평판 연구를 수행하였다(Lee, J.H. and H.G. Lee, 2015). 이 연구에서는 국내외 의류 제품의 상품 평을 수집하였고, 감성 사전을 이용하여 긍정 및 부정 요인들을 찾아 구매 프로세스에서의 문제점을 발견할 수 있었지만, 토픽을 이용한 감성분석은 시도되지 않았다.

사용자들의 의견을 담고 있는 비정형 텍스트를 마이닝하고 감성 분석하는 예측 모델을 제안하였다(Kim, S.W. and N.G. Kim, 2014). 이 연구는 학습 데이터에 대한 기본적인 자연어 처리를 통해 관심 용어를 추출하고 관심 용어가 긍정 문서와 부정 문서에서 각각 출현한 회수를 측정하여 각 용어의 감성 지수를 계산하여 감성 사전을 구축하였고, 이를 통해 긍정 또는 부정 문서를 식별하였지만, 제안방안처럼 토픽을 이용한 감

성분석은 시도되지 않았다.

특정 토픽에 대한 대중의 성향을 파악하여 토픽에 대한 긍정 및 부정 비율을 계산하였다. 또한 토픽에 대한 대중의 성향이 시간 흐름에 따라 강화되는지 쇠퇴하는지를 시각화 하여 보여준다 (Shim, H.M. and W.J. Kim, 2015). 이 연구는 토픽 모델을 사용하여 문서 코퍼스를 클러스터링 하고, 각 클러스터의 긍·부정 유무를 파악한다. 주로 각 클러스터가 긍정인지 부정인지를 판별한다. 본 논문에서 제안한 방안이 토픽의 연관어들과 확률 분포를 이용하여 감성사전을 구축하고 제품의 긍·부정 선호도 조사 및 감성 요약을 하는 방안을 제시한 반면, 심홍매와 김우주의 연구는 토픽 모델을 사용하여 문서 코퍼스를 클러스터링 하고, 각 클러스터의 긍·부정 성향을 클러스터에 있는 문서들의 저자들의 성향으로 판단한다. 제품 평판 마이닝 보다는 특정 이슈와 관련된 토픽들의 찬반 관계 흐름에 초점을 맞추었다.

최근 국내에서는 텍스트 마이닝 기법을 이용한 다양한 적용 사례들이 나타나고 있다. 소셜 네트워크 서비스와 감성 분석 기반의 검색엔진인 '짐고' 는 페이스북과 트위터 등의 소셜 네트워크 서비스에 게재된 댓글을 작성한 사용자의 감정을 분석하여 대중과 기업에 제공해주는 서비스이다(Kim, M.S., 2017). 그러나 짐고는 본 연구에서 제안하는 방안과는 달리 단순한 오피니언 마이닝 방법을 적용하였다.

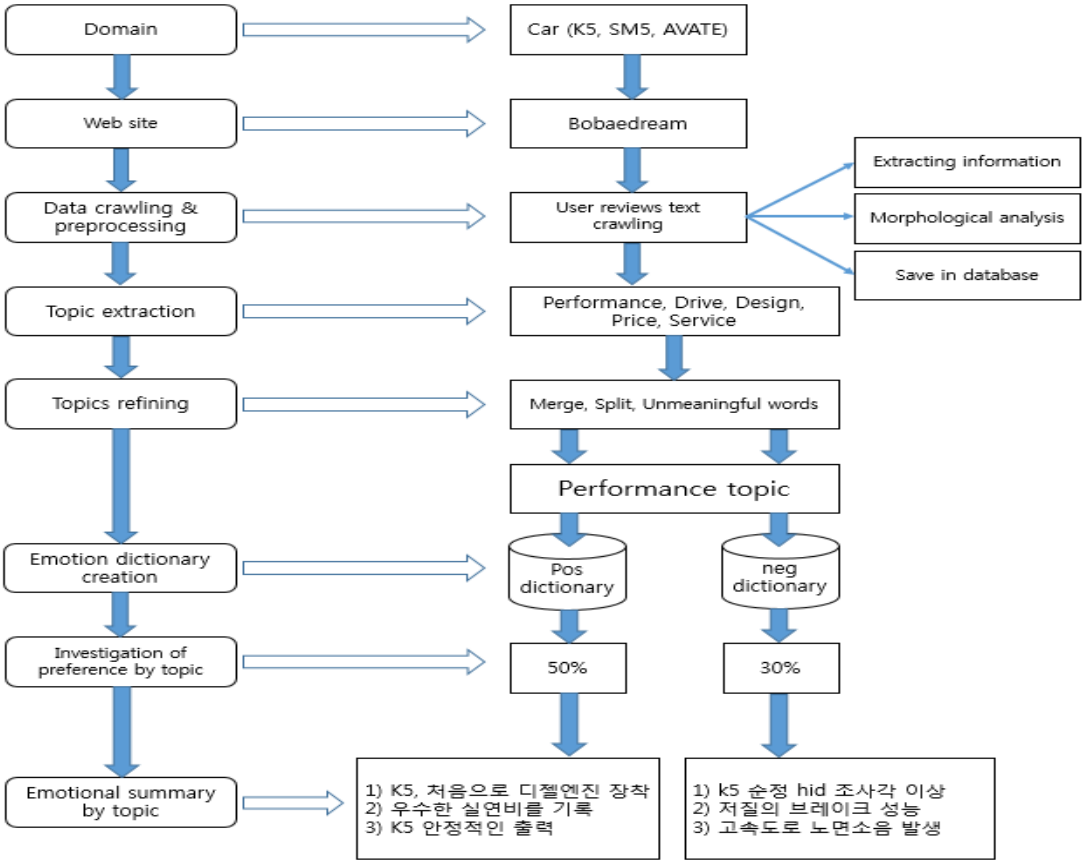
3. 제안방안

그림 1은 본 논문에서 제안한 제품 평판 마이닝 알고리즘을 도식화하여 보여준다. 자동차와 같은 특정 도메인에서 일반 대중이나 고객들이 많이 찾는 보배드림 사이트의 사용자 후기 텍스트들을 자동으로 수집하고 정제한다.

이러한 사용자 후기 텍스트 문서들을 입력으로 하는 토픽 모델 알고리즘을 실행하여 사용자 후기 게시판 문서에서 많이 회자되는 토픽들을 추출한다. 각 토픽은 연관어와 확률 값으로 구성된다. 예를 들면, 토픽1={(컬러감, 0.7), (유선형, 0.2), (참신한, 0.1)}. 토픽1의 연관어들을 고려할 때, 토픽1은 '디자인'에 관련된 토픽임을 알 수 있다. 토픽 추출 알고리즘은 3.1절의 알고리즘 1에서 자세히 다룬다. 토픽 모델링 알고리즘은 통계적인 방법을 사용하기 때문에 가장 성능이 좋은 토픽 모델링 알고리즘 조차도 실제 데이터에 적용했을 때 토픽 결과들이 완벽하지 않다. 이를테면, 토픽1과 토픽2가 모두 '디자인'에 대한 것일수도 있고, 한 토픽 내에 의미적으로 상이한 2개 이상의 연관어 집합들이 존재할 수 있다. 본 논문에서 전자를 합병 문제(merge problem), 후자를 분할 문제(split problem)라 부른다. 이러한 토픽들을 후처리(post-processing) 함으로써 정확한 토픽들을 출력한다. 예를 들면, 보배드림의 후기 게시판에서 많이 다루어지는 토픽으로는 '성능', '주행', '디자인' 등이 있다.

다음, 각 토픽에 대해 긍·부정 감성사전을 구축하고, 토픽 별 선호도를 조사한다. 예를 들면, 아반떼 자동차의 '디자인' 토픽에서 긍정 50%와 부정 30% 의 선호도를 측정한다. 토픽의 긍·부정 비율을 측정하는 알고리즘은 3.4절의 알고리즘 2에서 자세히 다룬다.

마지막으로 자동차의 '디자인' 토픽에서 긍정 내용과 부정 내용을 요약하여 보여줌으로써 고객들이 특정 자동차의 '디자인' 토픽에서 무엇을 긍정적으로 보고 있는지, 무엇을 부정적으로 여



〈Figure 1〉 Flow chart of the proposal product reputation mining algorithm

기는지를 일목요연하게 알 수 있다. 토픽의 감성 요약 알고리즘은 3.5절의 알고리즘 3에서 자세히 다룬다.

3.1 토픽 추출

보배드림에서 수집한 사용자 후기 게시판 K5 자동차(시간 t_0) 문서를 r_1, r_2, r_3, r_4, r_5라 가정하면 $t_0 = \{r_1, r_2, r_3, r_4, r_5\}$로 표현할 수 있다. 이 단계에서는 토픽 모델 알고리즘(topic modeling algorithm)을 사용하여 t_0에 있는 문서들을 클러스터링(clustering)하여 클러스터 세트(cluster set)를 생성한다. 예를 들어, t_0에 있는 K5문서들로부터 다음 2개의 클러스터 세트(C_1과 C_2)들을 얻었다고 가정하자. 이를 테면, $C_1 = \{r_1, r_4, r_5\}$와 $C_2 = \{r_2, r_3\}$라고 한다면, C_1에 속하는 r_1, r_4, r_5 문서들은 내용이 서로 유사해야 하며, C_2에 속하는 문서들과 다르면, 클러스터링이 효과적으로 수행되었다고 할 수 있다. 또한 C_1에 속한 사용자 후기 문서들을 분석하여, 토픽들(a set of topics)과 토픽들의 확률 분포(probability distribution)를 추출할 수 있다면, r_1, r_4, r_5 문서들이 실제로 K5

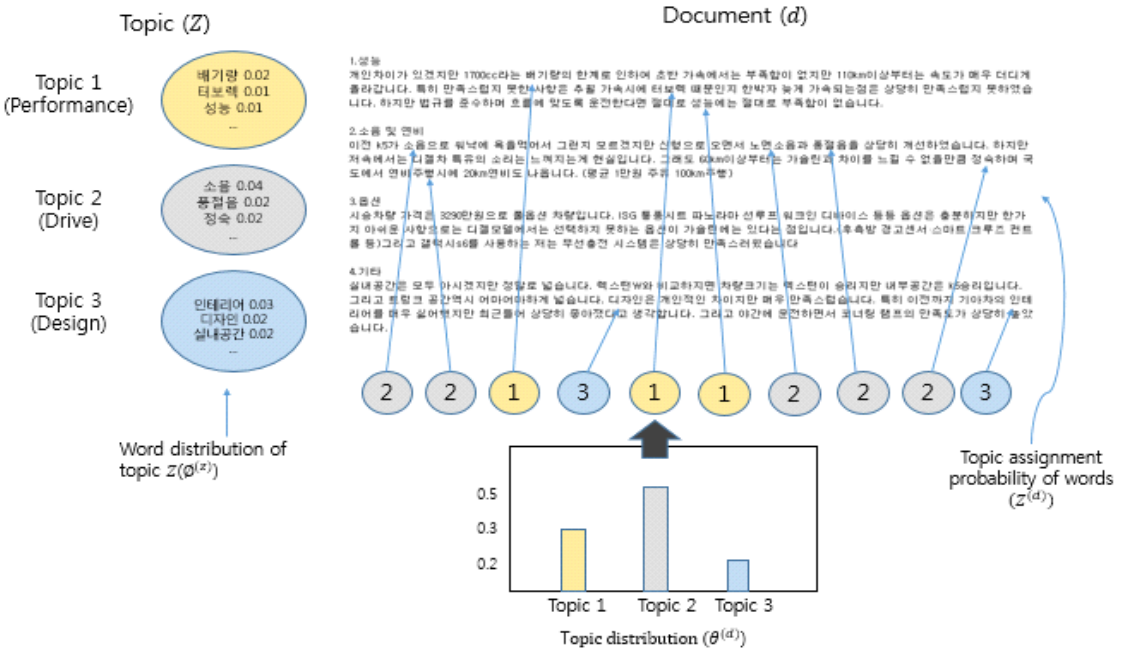
의 어떤 내용을 다루고 있는지를 파악할 수 있게 된다. 예를 들면, 해당 문서에서 K5의 성능과 주행에 대한 내용이 주로 회자 된다면, K5 주요 토픽(주제)들은 '성능'과 '주행'이 될 것이다. 또한 해당 문서들 중에서 성능에 대해 60%, 주행에 대해 40% 정도의 비율로 이야기되고 있다면, 토픽들에 대한 확률 분포는 각각 0.6과 0.4임을 알 수 있다. 즉, $t_0 = \{r_1, r_2, r_3, r_4, r_5\}$이 입력으로 주어지면, 토픽 모델링 알고리즘을 사용하여 해당 사용자 후기 문서들의 토픽들을 추출할 수 있다. 또한 각 토픽은 단어(명사)들의 집합이고, 단어들은 서로 연관 관계를 가진다. 따라서 토픽 내의 단어들을 분석하면 토픽이 실제로 무엇인지를 알 수 있다. 또한 각 단어는 확률 값을 가지고 있어, 토픽 내에 그 단어의 중요성을 쉽게 파악할 수 있다.

연구에서는 토픽 모델링 알고리즘을 위해 Latent Dirichlet Allocation (LDA) 방식을 사용하여 사용자 후기 게시판 문서의 토픽 세트와 확률 분포를 추출한다. LDA는 문서들의 토픽(주제)를 알기 위해 원본 텍스트 내의 단어를 분석하는 통계적인 방법이다. LDA는 확률 그래프 모델 중의 하나로 Dirichlet 분포를 이용하여 텍스트 문서 내의 단어들이 어떤 특정 토픽에 포함될 확률을 계산하는 모델이다.

텍스트 문서에는 여러 토픽들이 혼합되어 있다는 가정에서 LDA는 출발한다. LDA는 토픽 레이어(layer)를 통해 문서 레이어와 단어 레이어로 연결된다. 이와 같은 그래프를 3부 그래프(tripartite graph)라고 하는데, 문서, 토픽, 단어 등 3개의 다른 형태의 노드들로 구성된다. 동일 그룹의 노드들은 서로 연결되지 않고, 다른 타입의 노드들과 연결된다. 예를 들면, 문서1은 토픽1, 토픽2, 토픽3에 연결되고, 연결된 선분의 중요도는 확률 값이다. 만일 문서1과 토픽1, 토픽2, 토픽3 각각의 확률 값이 0.6, 0.4, 0.0이라고 하면, 문서1은 토픽1(성능)과 토픽2(주행)로 구성된다. 그리고 성능과 주행의 확률 분포는 0.6과 0.4이다. 또한 LDA는 어떤 텍스트 문서도 생성되기 전에 이미 토픽 구조가 존재하며, 숨겨져 있는 토픽 구조(hidden topic structure)에 의해서 문서가 생성된다는 생성확률모델(generative model)이다. 즉, 숨겨져 있는 토픽 구조를 미리 가정하고, 현재 관찰 가능한 문서 내의 단어들은 이로부터 생성되었다는 가정에서 출발한다. LDA에 숨겨져 있는 토픽 구조(파라미터)는 다음과 같다.

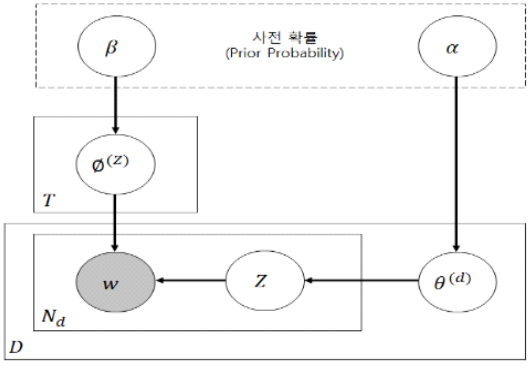
1) 토픽 개수 (number of topics)
2) 문서 d의 토픽 분포도 ($\theta^{(d)}$)
3) 문서 d의 각 단어의 특정 토픽 배정 확률 ($Z^{(d)}$)
4) 토픽 Z의 단어 분포 ($\varnothing^{(z)}$)

그림 2는 LDA를 설명하기 위한 예제 그림이다. 그림에서 보면, 문서들의 집합(a collection of documents)는 크게 3개의 토픽들로 구성된다. 그림의 왼편에는 3개의 토픽들이 있는데 성능(노란색), 주행(회색), 디자인(파란색)에 관한 토픽들이다. 각 토픽은 단어와 단어의 확률 값으로 구성된다. 즉, 성능 토픽에는 배기량, 터보렉, 성능 등의 단어들과 토픽 내 단어들의 확률 분포로 구성된다. 예를 들면, 토픽 Z(예: 성능)의 단어 분포는 $\varnothing^{(z)}$이다. 그림에서 원 도형은 토픽을 나타내며, 히스토그램($\theta^{(d)}$)은 문서 별 토픽 분포도를 나타낸다. 또한 화살표는 문서 d 내의 각 단어의 토픽 지정을 의미한다. 즉 문서 d에 있는 각 단어가 특정 토픽에 있는 단어들의 확률 분포에 의해 생성됨을 의미하고, $Z^{(d)}$로 표현한다. 결론적으로, 주어진 각 문서는 본래 성능, 주행, 디자인이라는 숨겨진 토픽들의 확률 분포에 의해 만들어지



〈Figure 2〉 Topic distribution chart (Blei, D., 2012, p.77-84)

며, 그 문서에 있는 각 단어는 토픽 내의 그 단어의 확률 분포에 의해 문서 내에 생성된다는 것이 LDA의 핵심이다.



〈Figure 3〉 LDA graphical model (Wagner, C., accessed in 2010)

이러한 개념은 그림 3과 같이 그래픽 모델로 표현하여 쉽게 이해할 수 있고, 다음과 같이 수식으로 나타낼 수 있다.

$$P(d,w) = P(d)P(\theta^{(d)}|\alpha)\sum_{z} P(\emptyset^{(z)}|\beta)P(w|z,\emptyset^{(z)})P(z|\theta^{(d)})$$

<수식 1> 문서 d와 단어 w에 대한 사후확률

위의 수식에서 P(d, w)는 문서 d와 단어 w에 대한 사후확률(posterior probability)이다. 또한 α와 β는 사전확률(prior probability)을 나타내고, α는 문서 내의 토픽 분포, β는 토픽 내의 단어가 어떤 확률로 분포되는지를 나타내는 사전확률이다. 그러나 수식 1을 사용하여 모집단(a collection of documents)의 모든 단어들을 고려하

여 사후 확률을 계산하는 것은 불가능하다. 따라서 간접 방식으로 사후확률을 추정하게 되며, 주로 깁스 샘플링(Gibbs sampling)을 사용한다. 깁스 샘플링은 2개 이상의 변수들의 결합확률분포로부터 연속적인 표본을 샘플링을 한다. 결합분포가 명확히 알려져 있지 않으나, 각 변수의 조건부 분포는 알려져 있을 경우에 적용 가능하며, 추정하고자 하는 변수의 나머지 변수에 대한 조건부 확률분포에 의존하여 교대로 표본을 채취하는 방법으로 구현이 가능하다.

알고리즘 1: 사후 확률 추정을 위한 깁스 샘플링

입력: 문서 컬렉션, 사전확률 α와 β, 토픽 개수

1단계: 임의의 샘플링

2단계: for 단어 w in 문서 d

$$P(z_i = j|z_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_ij}^{WT} + \beta}{\sum_{w=1}^{W} C_{w_j}^{WT} + W\beta} \frac{C_{d_ij}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_it}^{DT} + T\alpha} \quad \text{<수식 2>}$$

3단계: $P(z_i = j|z_{-i}, w_i, d_i, \cdot)$에 의하여 단어 w를 토픽 t에 배정

깁스 샘플링에 있는 수식 2의 사후확률은 P(t|d)와 P(w|t)의 곱(product)으로 구해진다. 깁스 샘플링 알고리즘을 자세히 설명하기 위해 다음과 같은 예제를 고려한다.

$$D_1 = \{a(T_2), b(T_2), c(T_1), b(T_2)\}$$
$$D_2 = \{a(T_1), b(T_2), b(T_1), d(T_2)\}$$
$$D_3 = \{d(T_1), b(T_2), e(T_2), b(T_1)\}$$

위의 예제에서 D_1, D_2, D_3는 각각 보배드림 사용자 후기 게시판 문서를 나타내며, D_1은 a,b,c,b 단어들로 구성된다. 또한 D_2은 a,b,b,d 단어들로 구성되고, D_3은 d,b,e,b 단어들로 구성된다. 깁스 샘플링 알고리즘이 실행되기 전에 각 단어는 임의로 (randomly) 토픽 T_1이나 T_2에 배정(assign)된다. 예를 들면, D_1의 단어 a를 임의로 토픽 T_2에 배정되고, 이 과정은 알고리즘의 1단계에 해당된다. 또한 2단계에서는 C^{WT}와 C^{DT} 테이블이 다음과 같이 만들어 진다.

〈Table 1〉 C^{WT} table

C^{WT}	T_1	T_2
a	1	1
b	2	4
c	1	0
d	1	1
e	0	1

C^{WT}을 통해 a 단어는 T_1과 T_2 토픽에 각각 한 번씩 배정 되었음을 알 수 있다. 또한 C^{DT} 테이블을 통해 D_1 문서가 T_1과 T_2 토픽에 1번, 3번 배정 됨을 알 수 있다.

〈Table 2〉 C^{DT} table

C^{DT}	D_1	D_2	D_3
T_1	1	2	2
T_2	3	2	2

위와 같이 C^{WT}와 C^{DT} 테이블이 생성된 후에 각 단어의 확률 값을 구하고 확률 값이 큰 토픽에 다시 배정된다. 예를 들면, D_1의 a 단어의 확률 값을 구하기 위해, 먼저 $C^{WT}(a,T_2) = C^{WT}(a,T_2) - 1$

을 하고, $C^{DT}(a,T_2) = C^{DT}(T_2,D_1) - 1$을 한다. 그리고 a 단어가 각 토픽에 속할 확률을 구한다.

$$\begin{aligned} &\mathrm{P}(Z_i = T_1 | z_{-i}, a, d_i, \cdot) \\ &\quad = \frac{1+0.01}{4+5\times 0.01} \times \frac{1+25}{2+2\times 25} \\ &\quad = 0.13 \end{aligned}$$

$$\begin{aligned} &\mathrm{P}(Z_i = T_2 | z_{-i}, a, d_i, \cdot) \\ &\quad = \frac{0+0.01}{6+5\times 0.01} \times \frac{2+25}{1+2\times 25} \\ &\quad = 0.00088 \end{aligned}$$

w는 단어의 개수를 말하며, T는 토픽들의 수이다. α와 β는 사전확률로 $\alpha = \frac{50}{T} = 25$와 β = 0.01의 값을 사용한다. a 단어가 T_1에 속할 확률이 크므로 T_1에 배정한다. 마지막으로 $C^{WT}(a,T_2) = C^{WT}(a,T_2) + 1$을 하고, $C^{DT}(a,T_2) = C^{DT}(T_2,D_1) + 1$을 한다. 위와 같은 방식으로 각 단어의 토픽을 구하게 된다.

이러한 깁스 샘플링 방식을 통해 문서들의 토픽 세트가 추출된다. 예를 들면, 그림 2에서 보이는 성능, 주행, 디자인이라는 3개의 토픽들이 추출된다. 각 토픽은 유사한 의미를 지니는 단어들과 그 토픽 내의 단어들의 확률 분포가 출력되어, 특정 단어는 그 토픽 내의 중요도를 판단할 수 있다. 이와 같이 깁스 샘플링 알고리즘을 통해 추출된 토픽으로부터 분할 및 합병 연산을 수행한다.

여러 토픽 모델 알고리즘 중에서 LDA는 성능이 우수한 것으로 알려져 있다. 대량의 문서 컬렉션을 기계적으로 빠르고 정확하게 처리할 수 있어, 현재 널리 사용되고 있다. 하지만, LDA는 텍스트 문서 요약(text summarization)을 하는 통계적인 방법론이기 때문에 어떤 텍스트 도메인이 입력으로 주어지면, 단지 토픽 세트(a set of topics)만 결과로 준다. 각 토픽의 레이블(label)은 도메인 전문가가 토픽에 속하는 단어들을 분석하여 수작업으로 토픽의 레이블을 정해야 하는 단점이 있다. 예를 들면, 대용량 뉴스 기사로부터 LDA를 사용하여 자동으로 토픽 세트를 추출하였다고 가정하자. 각 토픽은 연관성 있는 단어들과 확률 값으로 구성된다. 그리고 도메인 전문가는 토픽을 분석하여, 각 토픽의 레이블을 결정하게 된다. 이를 테면, 토픽 = {(실업, 0.4), (해고, 0.3), (회사, 0.3)}에서 토픽 단어들은 실업문제와 관련 있으며, 도메인 전문가는 LDA로부터 추출된 그 토픽을 '실업문제'로 레이블링(labeling)하게 된다.

3.2 토픽 정제

기존의 토픽 추출 알고리즘은 통계적인 방법으로, 단어의 확률 분포에 따라 기계적으로 토픽을 추출하여 제공하기 때문에, 의미(semantic meaning)에 맞지 않게 중복된 토픽들을 제공하게 된다. 이를 테면 토픽 1과 토픽 2는 상당히 유사한 단어들을 서로 포함하고 있기 때문에 동일한 내용을 다루는 것으로 판단되며, 결과적으로 토픽 추출 알고리즘의 정확도를 높이기 위해서는 하나의 토픽으로 합병되어야 한다.

또한 하나의 토픽에 여러 개의 의미가 포함될 가능성이 존재한다. 아래의 토픽 3을 보면 토픽 3에는 '주행'에 관련된 의미와 '디자인'에 관련된 의미가 동시에 포함되어 있다는 것을 알 수 있다.

〈Table 3〉 Topic merge problem

Topic 1	
엔진	0.0091
모터	0.0041
브레이크	0.0026
기본기	0.0009
연동	0.0006
스마트	0.0006
스피커	0.0006
공기역학적	0.0006
교환	0.0003
스위치	0.0003
Topic 2	
터보	0.0045
엔진	0.0029
순정	0.0011
상태	0.0008
교환	0.0006
발생	0.0006
불안	0.0006
엔진음	0.0003
습기	0.0003
실망	0.0003

〈Table 4〉 Topic split problem

Topic 3	
브레이크	0.0026
마력	0.0008
소음	0.0007
내부	**0.0006**
코너링	0.0003
연비	0.0003
외관	**0.0003**
세련된	**0.0003**
깔끔	**0.0003**
날렵	**0.0003**

따라서 이러한 기존 토픽 추출 알고리즘의 문제점을 해결하기 위해 토픽 합병 및 분할 연산을 수행해야 한다. 이를 테면 토픽 1과 토픽 2에 존재하는 유사한 단어들을 하나의 토픽으로 합병하여 토픽 4를 추출하였으며, 토픽 3에 존재하는 '주행'과 '디자인'에 관련된 토픽을 분할하여 토픽 5와 토픽 6을 추출한다.

〈Table 5〉 Results of merge and split operations

Topic 4 (merge)	Topic 5 (split)	Topic 6 (split)
엔진	브레이크	내부
모터	마력	외관
터보	소음	세련된
순정	코너링	깔끔
기본기	연비	날렵
연동		
스마트		
상태		
교환		
발생		
공기역학적		
불안		
엔진음		
실망		

또한 어떤 토픽은 토픽 내 연관어들이 의미 없는 단어들로 이루어져있다. 이러한 토픽을 불용 토픽이라 부른다. 예를 들면, 토픽1={(그, 0.3), (이다, 0.2), (그래서, 0.2), (이면, 0.2), (거기서, 0.1)}에서 토픽 단어들이 의미가 없는 단어들로 구성되는 경우이며, 이러한 불용 토픽은 필터링 된다.

3.3 감성사전 구축

표 6은 3.1과 3.2 절의 과정을 통해 얻은 토픽의 예시이다. 이러한 연관어와 확률 값을 이용한다면 추출된 토픽의 긍·부정 비율을 계산 할 수 있다. 각 토픽에 대해 사용자들의 선호도를 판별할 수 있는 단어를 지정하여 감성사전을 구축한다. 감성사전은 각 토픽마다의 연관어를 통해 문서의 감정을 판단할 수 있는 감성단어로 구분하여 구축한다. 표 7은 이와 같은 방법으로 구축한 K5 자동차의 성능에 대한 감성사전이다.

〈Table 6〉 Example of topic words using LDA

Word of topic (w)	Probability value (P(w))
엔진	0.0046
하이브리드	0.0044
연비	0.0035
주행	0.0034
영업	0.0049
가솔린	0.0021
에어백	0.0017
시스템	0.0017
브레이크	0.0016
...	...

〈Table 7〉 Example of sentimental dictionary in the 'performance' aspect of k5

	Word of topic	Probability value
Positive	연동	0.0007
	굳	0.0007
	스마트	0.0007
	공기역학적	0.0007
	통과	0.0007
Negative	흉기	0.0004
	께름직	0.0004
	껍질	0.0004
	저질	0.0004
	부족함	0.0007

3.4 토픽 선호도 측정

토픽의 긍정과 부정을 측정하기 위해 3.3절에서 구축한 감성사전을 이용한다. 정확한 값을 측정하기 위해 알고리즘 2를 제안한다.

알고리즘 2: 토픽 선호도 측정

입력: 문서 컬렉션, 토픽 집합 T, 집합 $S = \emptyset$
출력: 토픽 긍정 00%, 토픽 부정 00%

1단계: 각 문서에서 문장 리스트$(s_1, \dots, s_n)$ 추출
2단계: $S = S \cup \{s_1, \dots, s_n\}$
3단계: $pos = 0,\ neg = 0$
4단계: for 토픽단어 $t \in T$
for 문장 $s \in S$
<수식 6>을 사용하여 $score(s,t)$ 계산

$$\alpha \cdot \frac{\sum_{w' \in tokens(s) \wedge w'=t} P(t)}{|tokens(s)|} + (1-\alpha) \cdot \frac{|긍정단어\ w' \in tokens(s)| - |부정단어\ w' \in tokens(s)|}{|tokens(s)|}$$

<수식 6>

if $(score(s,t) > 0)$ $pos++$
else $neg++$

5단계: 토픽 긍정 $= \frac{pos}{pos+neg} \times 100\%$, 토픽 부정 $= \frac{neg}{pos+neg} \times 100\%$

알고리즘 2가 실행되면, 각 문서에서 문장들이 추출되고 집합 S에 저장된다. S는 문장들의 집합을 나타낸다. 수식 6에서 $tokens(s)$은 문장 s을 구성하는 단어들의 집합이고, $|tokens(s)|$은 $tokens(s)$ 집합의 단어 개수이다. $\sum_{w' \in tokens(s) \wedge w'=t} P(t)$는 $tokens(s)$ 집합에 있는 단어 w'가 토픽 T 내의 단어 t와 같을 때, t의 확률 값 P(t)들을 모두 더한 값이다. 즉, 수식 6의 첫 번째 항은 문장 s에 존재하는 단어들의 토픽 내 확률 값을 더해서 단어 개수만큼 나눈 것이다. 첫 번째 항이 높을수록 문장 s와 토픽 T는 연관성이 커진다. 두 번째 항에서는

$tokens(s)$ 집합에 있는 긍정 단어의 개수와 부정 단어의 개수의 차를 단어 개수로 나눈 것이며, 문장 s에 긍정 단어 개수가 많으면 양의 값을, 부정 단어 개수가 많아지면 음수 값을 가지게 된다. α은 첫 번째와 두 번째 항의 가중평균을 구하는 파라미터이다. 본 연구에서는 $\alpha = 0.5$을 사용하여 첫 번째와 두 번째 항을 균등하게 고려한다.

3.5 토픽 감성 요약

주어진 토픽의 긍정 내용과 부정 내용을 요약하여 보여주는 알고리즘 3을 제안한다. 3.4절의 알고리즘 2와 유사하며, 수식 6의 $score(s,t)$ 값에 의해 내림차순으로 정렬한 다음, 상위 k개의 긍정인 문장과 부정인 문장을 출력한다.

알고리즘 3: 토픽 감성 요약

입력: 문서 컬렉션, 토픽 집합 T, 집합 $S = 0$
출력: Top-k 긍정 문장 리스트, Top-k 부정 문장 리스트

1단계: 각 문서에서 문장$(s_1, \ldots, s_i)$ 추출
2단계: $S = S \cup \{s_1, \ldots, s_i\}$
3단계: for 토픽단어 $t \in T$
for 문장 $s \in S$
<수식 6>을 사용하여 $score(s,t)$ 계산

$$\alpha \cdot \frac{\sum_{w' \in tokens(s) \wedge w'=t} P(t)}{|tokens(s)|} + (1-\alpha) \cdot \frac{|\text{긍정단어 } w' \in tokens(s)| - |\text{부정단어 } w' \in tokens(s)|}{|tokens(s)|}$$

<수식 6>
4단계: $score(s,t)$에 의해 문장들을 내림차순으로 정렬
5단계: Top-k 긍정 및 부정 문장 출력

---

4. 실험 환경 및 실험 결과

4.1 실험 환경

본 연구의 실험을 위해 국내 최대 자동차 후기 사이트인 보배드림에서 총 4,321개의 사용자 후기 텍스트를 자동으로 수집하였다. 표 8은 보배드림에서 수집한 사용자 후기 문서의 통계를 나타낸다.

〈Table 8〉 Data characteristics

Type of Car	Number of documents	Document creation period
K5	1,585	2012년 ~ 2016년
SM5	1,379	2006년 ~ 2016년
아반떼	1,357	2006년 ~ 2016년

각 사용자 후기 텍스트 문서에서 제목, 날짜, 본문 텍스트를 추출한 후에 카이스트 SWRC연구소에서 개발한 한나눔 한국어 형태소 분석기(Lee, S.W., 24 october, 2016)를 사용하여 형태소 분석을 수행하고, 데이터베이스에 저장하였다. 토픽 추출을 위해 오픈 소스인 JGibbLDA(Phan, X.H. and C.T Nguyen, 08 November, 2016)를 사용하였다. 토픽 정제와 감성사전 구축은 수작업으로 진행하였고, 토픽 선호도 측정과 감성 요약은 Java 프로그래밍 언어를 사용하여 구현하였다. LDA의 사전확률인 α와 β의 최적 값을 구하기 위해 Python프로그래밍 언어를 사용하여 토픽 일관성을 측정하였다. 또한 제안방안의 우수성을 입증하기 위해 30명에게 설문조사를 하였고 통계적인 검증을 위해 SPSS 21의 Bonferroni 사후 검정을 실시하였다.

실험에 사용된 컴퓨터는 Intel(R) Core(TM)

i7-4790으로 CPU 성능은 3.60GHz이고, 리눅스 계열의 Ubuntu 15.10에서 제안방안을 실행하였다.

4.2 실험 결과

4.2.1 토픽 모델링을 위한 사전확률 최적화

LDA는 베이지안(Bayes) 확률모델을 사용하기 때문에 사전확률인 α와 β가 입력으로 주어진다. α값이 커질수록 문서 내 토픽의 개수는 많아지고 β값이 커질수록 토픽 내 특정 소수의 연관어들의 확률 값이 커진다. 최적의 α와 β값을 얻기 위해 토픽 일관성(topic coherence)를 고려한다. 토픽 일관성은 토픽을 구성하는 연관어들이 얼마나 유사한 의미를 가지고 있는지를 측정한다(Alters, N. and M. Stevenson, 2013).

$$\frac{\sum_{1\le i\le n-1}^{n} \underset{i+1\le j\le n}{sim(w_i,w_j)}}{\binom{n}{2}} \qquad \text{<수식 7>}$$

수식 7은 주어진 토픽 내 단어 쌍들의 유사도의 총합에서 평균을 구한 것으로 분자는 주어진 토픽 내 단어들의 각 쌍의 유사도의 총합을 의미하며, 분모는 분자의 평균을 내기 위한 것이다. w_i 또는 w_j는 토픽 내 연관어를 의미한다. $\mathrm{sim}(w_i,w_j)$는 토픽 내 연관어간의 워드넷(WordNet) 값을 의미한다. 워드넷은 영어 단어의 의미 관계를 파악하는데 유용한 데이터베이스이다(Wikipedia, accessed in 2017). 예를 들면 'cost'와 'price'는 철자는 다르지만, 비슷한 의미를 지니고 있다. 따라서 워드넷을 사용하면 'cost'와 'price'의 유사도(similarity)는 높게 나타난다. 표 9는 다양한 α와 β값에 따라 모든 토픽의 수식 7의 평균 값을 보여준다. 실험 결과에 따르면, α = 0.1와 β = 0.3일 때 토픽 일관성이 가장 높음을 알 수 있다. 따라서 제안방안의 모든 실험은 사전확률은 α = 0.1와 β = 0.3을 사용하였다.

〈Table 9〉 Results of prior probabilities

	α=0.1, β=0.3	α=0.3, β=0.2	α=0.5, β=0.1
Topic coherence	0.074	0.073	0.067

4.2.2 토픽 추출 결과

표 10은 K5, SM5, 아반떼 자동차의 '성능' 토픽에서 상위 10개의 연관어들과 확률 값들을 나타낸다. 연관어와 확률 값을 통해 토픽에서 주로 언급되는 단어와 토픽 내 단어들 중에서 중요한 정도를 유추할 수 있다. K5 자동차의 연관어 중 '모터'의 확률 값은 0.0041로 가장 높다. 이것은 K5 자동차의 성능과 관련하여 일반 대중들이 가장 많이 언급한 단어는 '모터'임을 알 수 있다. 반면에 SM5 자동차는 '엔진', 아반떼 자동차는 '에어벡'이라는 단어가 가장 많이 언급되었다. 흥미로운 점은 K5, SM5, 아반떼 자동차의 토픽 내 연관어들이 대부분 서로 다르다는 것이다. 이것은 각 자동차의 성능에 대해 대중은 자동차마다 다른 이야기를 하고 있다는 것이다. 반면에 K5, SM5, 아반떼 자동차에서 동일하게 '엔진'이라는 단어는 성능 토픽에서 공통으로 언급되는 단어이다. 이것은 모든 자동차에서 '엔진'이라는 단어를 공통적으로 이야기하고 있으며, 자동차의 성능에서 '엔진'은 가장 중요한 이슈인 것을 알 수 있다. 실험을 통해 '성능' 토픽 이외에 '주행', '디자인', '가격', '서비스' 등의 토픽을 얻을 수 있었다. 그러나 '가

〈Table 10〉 Example of topics labeled by ‘performance’

Ranking	K5		SM5		AVANTE	
	Word of topic	Probability distribution	Word of topic	Probability distribution	Word of topic	Probability distribution
1	모터	0.0041	**엔진**	0.0034	에어백	0.0086
2	가솔린	0.0038	성능	0.0029	서스펜션[1]	0.0046
3	모드	0.0038	엑셀레이터	0.0026	정면충돌	0.0042
4	스프링	0.0038	CVT[2]	0.0025	NHTSA[3]	0.0036
5	EV[4]	0.0036	리콜	0.0019	**엔진**	0.0036
6	느낌	0.0035	기능	0.0018	센서	0.0035
7	스티어링휠[5]	0.0033	문제	0.0015	테스트	0.0031
8	배터리	0.0032	전자식	0.0014	결과	0.0031
9	피스톤[6]	0.0029	출력	0.0013	셋업	0.0026
10	**엔진**	0.0029	장착	0.0013	이상	0.0024

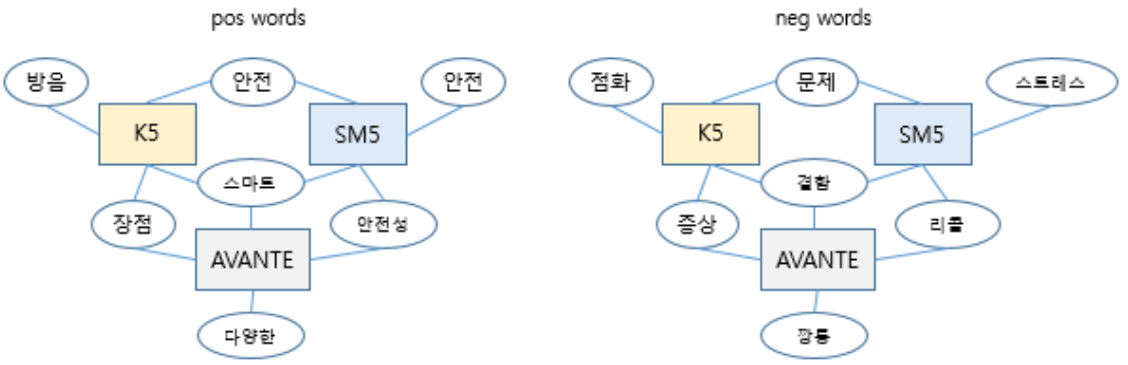
1) 서스펜션: 노면의 충격이 차제에 전달되지 않게 충격을 흡수하는 장치
2) CVT: 무단 변속기
3) NHTSA: 미국교통안전국, 자동차 충돌 테스트 제공
4) EV: 전기자동차
5) 스티어링휠: 차량의 바퀴를 통해 움직임을 바꾸는 조향장치(Doopedia, Accessed in 2017)
6) 피스톤: 왕복엔진, 왕복 운동하는 펌프/압축기에서 사용하는 부품

격’과 ‘서비스’ 토픽 결과는 미미하기 때문에 본 논문에서는 그 결과를 생략한다. 다만 이러한 결과를 통해 보배드림 사이트에서 가격과 서비스는 거의 이야기 되지 않고 있음을 간접적으로 알 수 있다. 또한 ‘주행’과 ‘디자인’ 토픽은 부록에 일목요연하게 정리하였다.

4.2.3 토픽 감성사전 결과

그림 4는 자동차의 성능 토픽에서 구축된 감성사전 단어들의 예시이다. 그림에서 사각형은 K5, SM5, 아반떼 자동차의 성능 토픽을 나타낸다. 각 원에 있는 단어는 감성 단어를 의미하며, 사각형과 선분으로 연결되어 있다면, 단어가 사각형의 성능 토픽 내에 출현한 단어를 의미한다. 예를 들면, 그림에서 ‘방음’이라는 단어는 K5와 연결되어 있으며, ‘방음’ 단어가 K5 자동차의 성능 토픽 내에 존재하는 긍정 단어를 나타낸다.

그림을 통해 각 자동차의 성능 토픽에서 어떤 긍정 또는 부정 단어들이 존재하고, 어떤 단어들이 다른 자동차의 성능 토픽에도 출현하는지를 쉽게 알 수 있다. 그림에서 ‘스마트’라는 단어는 K5, SM5, 아반떼에서 모두 나타난다. 이것은 각 자동차의 성능 토픽에서 ‘스마트’라는 긍정 단어가 있고 사용자들은 각 자동차에 공통적으로 어떤 부분을 긍정적으로 보고 있는지에 대해 알 수 있다. 또한 K5자동차에서는 긍정 단어에 ‘방음’이라는 단어가 독립적으로 출현하고, 다른 자동차들에서는 나타나지 않기 때문에 ‘방음’이라는 장점은 K5 자동차만 유일하게 갖고 있는 장점임을 알 수 있다. 부정 단어에서는 자동차마다 ‘결함’이라는 단어가 출현하는데, 이것은 자동차마다 특정 부분에 관한 ‘결함’이 공통적으로 존재



〈Figure 4〉 Graph of positive and negative words in all topics labeled by 'performance'

한다는 것을 알 수 있다. 또한 K5자동차에서는 다른 자동차와 달리 '점화'라는 부정 단어가 출현하는데 이를 통해서 K5자동차만이 갖는 문제점을 파악할 수 있어 K5자동차가 갖는 문제점과 개선해야 할 점을 쉽게 알 수 있다.

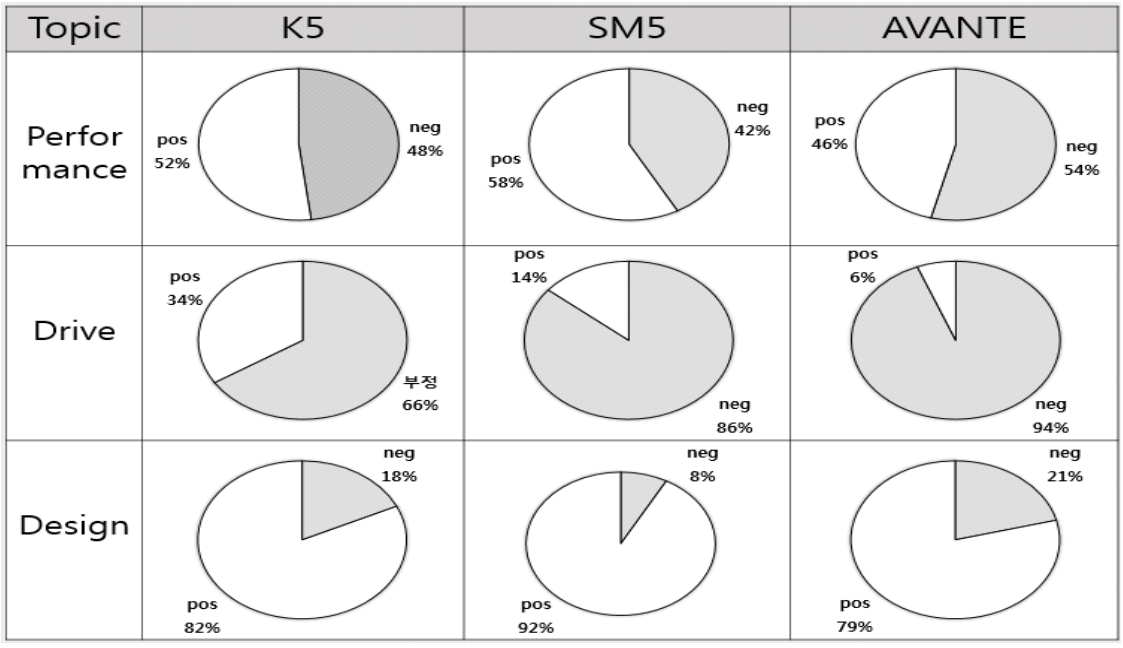
4.2.4 자동차 토픽의 긍·부정 비율 결과

그림 5는 각 토픽에서 보배드림 후기 게시판에서 얼마나 긍정적인 내용과 부정적인 내용을 다루는지에 대한 비율을 나타낸 파이차트 이다. 이러한 차트를 확인하면 각 자동차의 성능, 주행, 디자인에서 사용자들이 긍정적으로 말하는지, 아니면 부정적으로 말하는지를 직관적으로 알 수 있다. 그림에서 각 자동차의 성능 토픽에 대한 긍·부정 비율이 다르다. K5자동차의 성능에서 사용자들은 52%의 긍정적인 내용을 이야기하였고, SM5자동차는 58%, 아반떼 자동차는 46%였다. 따라서 자동차의 성능에서는 SM5, K5, 아반떼 자동차 순으로 긍정 비율이 높았다. 특히 아반떼 자동차는 긍정 비율보다 부정 비율이 높았다. 주행의 경우에는 대체로 부정적인 의견이 주를 이루었다. K5 자동차는 66%, SM5 자동차는 86%, 아반떼 자동차는 94%로 부정 비율이 높았다. 반면에 자동차 디자인의 경우에는 평균 84%로 긍정적인 비율이 압도적으로 높았다. 전체적으로 자동차의 성능과 디자인에서는 긍정 비율이 높고, 주행의 경우에는 부정 비율이 높음을 알 수 있다. 또한 K5, SM5, 아반떼 자동차 중에서 SM5 자동차가 성능과 디자인에서 다른 자동차들보다 긍정 비율이 높았지만 주행에서 K5 자동차에 비해 높은 부정 비율을 보였다.

4.2.5 자동차 토픽의 감성 요약 결과

그림 5를 통해 자동차의 성능, 주행, 디자인에서 K5, SM5, 아반떼 자동차의 긍·부정 비율을 확인하였으나, 각 토픽에서 구체적인 내용을 파악하는 것이 좀 더 필요하다. 예를 들면, SM5 자동차의 주행에서 부정 비율이 매우 높다. 어떤 이유 때문에 부정 비율이 높은지를 알 수 있다면 SM5를 제조하는 회사는 자동차를 개발할 때 참고를 할 수 있을 것이다.

표 11은 K5 자동차 성능 토픽의 감성 요약 결



〈Figure 5〉 Pros and cons ratios of experimental vehicles

〈Table 11〉 Positive and negative summaries in the 'performance' aspect of K5

Positive		
Ranking	Content	Score
1	기아자동차가 'K5'에 처음으로 디젤엔진을 장착했다. 기존 2.0리터, 2.4리터 가솔린엔진 외에 1.7리터 디젤엔진을 새롭게 추가했다.	0.50003
2	김 서림을 자동으로 막아주는 오토 디포그[1)]	0.50002
3	이 배기시스템을 장착하면 마력은 4마력, 토크는 0.8 오른답니다.	0.50001
4	'K5 터보 GDi'는 배기량 대비 높은 출력을 실현하는 엔진 다운사이징을 통해 국내 경쟁 차종은 물론 수입차를 압도하는 최고출력 271ps, 최대토크 37.2kgm의 막강한 동력성능과 우수한 연비 경제성을 동시에 확보했다.	0.50000
5	브렘보 380mm 4피스톤에 빛나는 굵직굵직한 야수 발톱을 장착!	0.50000
Negative		
Ranking	Content	Score
1	k5 순정hid[2)]	-0.50000
2	사제 달 걸 그랬습니다 로체이노베이션 동호회 오면 결함개선코너에 결함 있는 거 장난 아니게 있습니다	-0.50000
3	가장 문제가 되는 제로백은 제가 봐도 좀 이상하다 싶었습니다	-0.49999
4	기가 차고 어의가 없었습니다. 핸들이 녹고 있더군요..... 녹고만 있는 게 아니라.. 지독한 고무타는 냄새와 함께	-0.49999
5	브레이크 = 기대에 부흥하는 저질의 브레이크 성능	-0.49999

1) 디포그: 습기 제거 장치
2) Hid: 발광 관내의 방전에 의해 빛을 발산하는 램프(A shopping dictionary, accessed in 2017)

과를 보여준다. 알고리즘 3을 적용하여 성능 토픽과 가장 관련 있는 상위 5개의 긍·부정 문장을 출력한 결과이다. 주요 긍정 내용은 엔진이나 배기시스템의 출력을 높인 것이고, 부정 내용은 핸들, 제로백, 브레이크 등 자동차의 각종 부품의 결함에 대한 내용이다. 지면 관계상 상위 5개의 긍·부정 문장들의 요약 결과를 보여주지만, 대체로 긍정과 부정 내용이 정확히 일치함을 알 수 있다. 지면 관계상 다른 자동차에 대한 토픽 결과는 부록에 정리하였다.

4.2.6 자동차 간 감성사전 비교 결과

표 12는 K5자동차와 아반떼 자동차 간의 성능 토픽에 해당하는 감성사전들의 단어 출현을 비교한 것이다. $\frac{|A \cap B|}{|A \cup B|}$ 는 K5자동차와 아반떼 자동차에 동시에 출현한 단어들을 나타낸 확률 값이고, $\frac{|A-B|}{|A|}$ 는 아반떼 자동차에서만 출현한 단어의 확률 값, $\frac{|B-A|}{|B|}$ 는 K5자동차에서만 출현한 단어의 확률 값이다. 표 12를 보면, K5 자동차와 아반떼 자동차 사이에 단어들은 많이 중복되지 않는다. 또한 긍정 단어보다는 부정 단어가 중복이 높은 것을 알 수 있다. 긍정 단어의 수는 K5 자동차가 많고 부정 단어의 수는 아반떼 자동차가 많음을 알 수 있다.

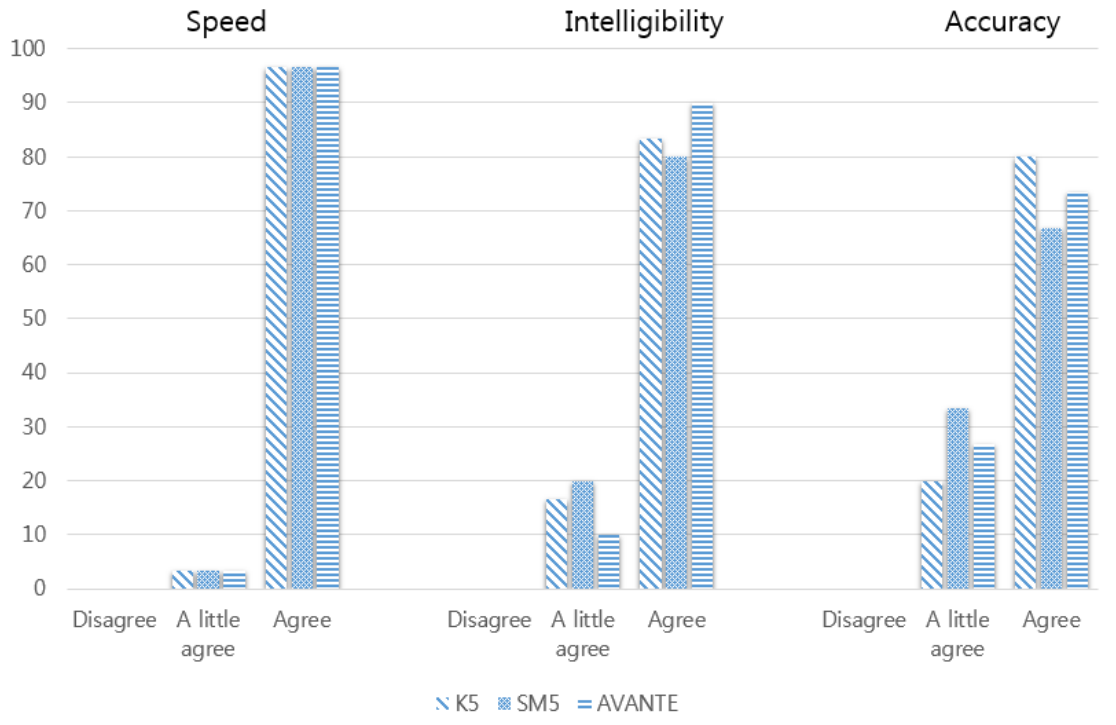
〈Table 12〉 Comparison of topic words between K5 and AVANTE
(A: AVANTE's sentimental dictionary,
B: K5's sentimental dictionary)

	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	$\frac{\lvert A - B \rvert}{\lvert A \rvert}$	$\frac{\lvert B - A \rvert}{\lvert B \rvert}$
All words	0.0839	0.8717	0.805
Pos words	0.0843	0.8409	0.8679
Neg words	0.125	0.8462	0.7778

4.2.7 제안방안의 효용성에 대한 설문조사 및 통계적 검정

그림 6은 제안방안을 설문 조사하여 얻은 결과이다. 그림에서 신속성은 제안방안이 전통적으로 사람에 의한 설문조사 방법과 비교했을 때 빠르게 결과를 얻을 수 있는지에 대한 응답이다. 이해도는 제안방안의 결과를 쉽게 이해할 수 있는지에 대한 응답이며, 정확성은 제안방안의 긍·부정 비율 측정과 요약 결과가 정확하며 신뢰할 수 있는지에 대한 응답이다. 신속성에 있어 '그렇다'라는 응답이 평균 96.70%이다. 이것은 설문 응답자의 대부분이 제안방안이 기존 설문조사 방법보다 빠르게 결과를 얻을 수 있다고 응답한 것이다. 이해도의 경우에는 K5, 아반떼, SM5 자동차가 각각 83.3%, 80%, 90%로 '그렇다'라는 응답을 보였다. 마지막으로 정확성에서는 설문 응답자 중의 80%, 66.7%, 73.3%가 정확하다고 응답하였다. 설문조사 결과에서 신속성과 이해도에 비해 정확성은 '그렇다'에서 약간 낮은 확률을 보였는데, 그 원인은 아반떼와 SM5 자동차의 주행 토픽의 감성 분석 긍정 결과에서 아반떼 자동차의 긍정적인 측면에 대한 내용이라고 보기 애매한 문장들이 있었다. 예를 들면, "확실히 소음 좀 있더군요. 근데 전 그게 굉장히 소리가 듣기가 좋았습니다"라는 문장이 한 예시이다. 그러나 '그렇다'라는 대답이 '보통이다'라는 대답보다 두 배 이상 많음을 알 수 있다. 설문조사는 20세 이상 성인 30명을 대상으로 진행되었다.

그림 6의 설문조사 결과를 통계적으로 검증하기 위해 ANOVA 분석과 사후 검정을 실시하였



〈Figure 6〉 Survey results of the proposed scheme

다(Kim, H.C et al, 2009). 표 13, 14는 ANOVA 분석과 사후검정에 대한 결과이다. 종속변수로는 신속성, 이해도, 정확성이며, 요인은 각 제품인 K5, SM5, 아반떼 자동차이다. 유의수준은 0.05이고, 귀무가설은 '자동차 간의 설문 응답의 평균은 같다'이며 대립가설은 '자동차 간의 설문 응답의 평균은 같지 않다'이다. ANOVA 분석에서 신속성은 1.000, 이해도는 0.921, 정확성은 0.226의 유의확률로 이에 따라, '자동차 간의 설문 응답의 평균은 같다'라는 귀무가설을 채택한다고 할 수 있다. 허나 자동차 간의 설문 응답의 평균이 어떻게 같은 지 더 심도 있게 알아보기 위해 사후 검정을 실시하였다. 사후 검정 중에도 일반적으로 많이 쓰이는 Bonferroni 사후 검정 방법을 택하여 사후 검정을 실시하였다. 신속성과 이해도는 각각 종속변수간 1.000의 유의확률로 귀무가설을 채택하였다. 정확도에서는 K5 자동차와 아반떼 자동차, 아반떼 자동차와 SM5 자동차 간의 유의확률이 각각 0.252과 0.740로 다른 종속변수간 유의확률 보다 낮은 값을 보였지만, 종속변수간 유의확률이 0.05를 넘었기에 귀무가설을 채택한다. 즉, 자동차와 토픽의 종류에 상관 없이 제안방안의 효용성에 대한 설문 응답자들의 일관성을 엿볼 수 있다.

〈Table 13〉 ANOVA analysis results

		Sum of squares	df	Mean square	F	Significance probability
Speed	Intergroup	.000	2	.000	.000	1.000
	Within-group	2.900	87	.033		
	Total	2.900	89			
Intelligibility	Intergroup	.022	2	.011	.082	.921
	Within-group	11.800	87	.136		
	Total	11.822	89			
Accuracy	Intergroup	.622	2	.311	1.515	.226
	Within-group	17.867	87	.205		
	Total	18.489	89			

〈Table 14〉 Statistical test results

	Type of car (i)	Type of car (j)	Mean difference (i–j)	Standard error	Significance probability	95% confidence interval	
						Threshold value	Upper value
Speed	K5	아반떼	.000	.047	1.000	-.12	.12
		SM5	.000	.047	1.000	-.12	.12
	AVANTE	K5	.000	.047	1.000	-.12	.12
		SM5	.000	.047	1.000	-.12	.12
	SM5	K5	.000	.047	1.000	-.12	.12
		아반떼	.000	.047	1.000	-.12	.12
Intelligibility	K5	아반떼	.033	.095	1.000	-.20	.27
		SM5	.033	.095	1.000	-.20	.27
	AVANTE	K5	-.033	.095	1.000	-.27	.20
		SM5	.000	.095	1.000	-.23	.23
	SM5	K5	-.033	.095	1.000	-.27	.20
		아반떼	.000	.095	1.000	-.23	.23
Accuracy	K5	아반떼	.200	.117	.273	-.09	.49
		SM5	.067	.117	1.000	-.22	.35
	AVANTE	K5	-.200	.117	.273	-.49	.09
		SM5	-.133	.117	.773	-.42	.15
	SM5	K5	-0.67	.117	1.000	-.35	.22
		아반떼	.133	.117	.773	-.15	.42

5. 결론 및 향후 연구

본 논문에서는 제품의 평판 마이닝을 위한 구체적인 방안을 제시하였다. 보배드림의 사용자 후기 게시판으로부터 K5, SM5, 아반떼 자동차에 관한 사용자 후기 문서들을 수집하고 토픽 모델 방법을 사용하여 사용자 후기 게시글들에 숨겨져 있는 토픽들을 추출한 후 분할과 합병 연산을 수행하여 토픽들을 정제한다. 그리고 토픽에 대한 감성사전을 구축하여 긍·부정 비율을 측정하고 토픽과 관련된 구체적인 내용을 요약해서 보여줌으로써 그 토픽에 대한 구체적인 긍·부정 내용을 알 수 있다.

본 연구의 결과물은 다양한 제품을 생산하는 산업계에 쉽게 응용 가능하며, 장기적으로 여론조사를 대체할 것으로 예상된다. 또한 신제품의 문제가 무엇인지를 파악하여 제품을 개선하거나, 타사 제품과의 비교를 통해 어떤 방향으로 제품을 홍보할 지, 어떤 방향으로 유지, 보수 해야 할지에 대한 기초 자료로도 활용이 가능하다.

향후 연구로는 실험 데이터의 크기를 늘려 빅데이터급 규모의 실험을 진행할 것이다. 또한 제안 방안을 일반화 시킴으로써 다양한 제품에 적용 가능한지에 대해 실험할 것이다. 끝으로 본 연구 결과물을 시연한 프로토타입 시스템을 개발함으로써 기술 수요가 큰 기업체에 기술 이전을 할 계획이다.

참고문헌(References)

A shopping dictionary, http://terms.naver.com/entry.nhn?docId=2464217&cid=51399&categoryId=51399, (Accessed 2017)

Aletras, N. and M. Stevenson, “Evaluating topic coherence using distributional semantics”, *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, Potsdam, Germany, 2013

Blei, D., “Probabilistic topic models,” *Communications of the ACM*, Vol.55, No.4, (2012), 77-84

Bobaedream, http://www.bobaedream.co.kr/ (Accessed 2016)

Das, R., M. Zaheer, and C. Dyer, “Gaussian LDA for topic models with word embedding”, *Proceedings of Conference of the Association for Computational Linguistics (ACL)*, Beijing, China, 2015

Doopedia, http://terms.naver.com/entry.nhn?docId=1234816&cid=40942&categoryId=32359 (Accessed 2017)

Lee, S.W. et al, HannanumKorean morphological analyzer Version 0.8.4, https://kldp.net/hannanum/ (Downloaded 24 October, 2016)

Jeong, D.M., J.S. Kim, G.N. Kim, J.W. Hu, B.W. On, and M.J. Kang, “A Proposal of a keyword extraction system for detecting social issues”, *Journal of Intelligence and Information Systems*, Vol.19, No.3, (2013), 1-23

Jo, T.M. and J.H. Lee, “Latent keyphrase extraction using LDA model”, *Korea Intelligent Information Systems Society*, Vol.25, No.2, (2015), 180-185

Kim, H.C., J.C. Oh, B.I. Yoon and K.M. Jeong, “Analysis of variance”, *Statistical understanding of Kyungmoon*, (2009) 194~209

Kim, M.S., “SNS search engine based on opinion

analysis ‘ZimGo’’, http://news.donga.com/3/all/20170114/82372402/ (Accessed 2017)

Kim, S.W. and N.G. Kim, “A Study on the Effect of using sentiment lexicon in opinion classification”, *Journal of Intelligence and Information Systems*, Vol.20, No.1, (2014), 133-148

Lee, J.H. and H.G. Lee, “A Study on customer reviews about domestic and imported clothes products through opinion mining”, *Korea Intelligent Information Systems Society*, (2015), 223-234

Liu, B., “Sentiment analysis and opinion mining”, *Morgan& Claypool Publishers*, Vol.5, No.1, (2012), 1-167

On, H.S., “Now, let’s go to the polls and cook big data!”, *Chosub Biz*, http://biz.chosun.com/site/data/html_dir/2015/06/05/2015060501615.html?Dep0=twitter (Accessed 2016)

Phan, X.H. and C.T. Nguyen, *JGibbLDA - A Java implementation of Latent Dirichlet Allocation (LDA) Version 1.0*, http://jgibblda.sourceforge.net/ (Downloaded 08 November, 2016)

Qian, S., T. Zhang,, and C. Xu, “Multi-modal multi-view topic-opinion mining for social event analysis”, *Proceedings of ACM Multimedia Conference (ACMMM)*, Amsterdam, Netherlands, 2016

Shim, H.M. and , W.J. Kim, “A Study of topic sentiment propensity analysis using big data”, *Journal of Intelligence and Information Systems*, Vol.20, No.20, (2015)

Wagner, C., “Topic models,” http://www.slideshare.net/clauwa/topic-models/5274169 (Accessed 2010)

Wan, X. and T. Wang, “Automatic labeling of topic models using text summaries”, *Proceedings of Conference of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016

Wikipedia, https://wikipedia.org/wiki/wordnet (Accessed 2017)

Zeng, Y., T. Ku, S. Wu, L. Chen, and G. Chen, “Modeling the helpful opinion mining of online consumer reviews as a classification problem”, *International Journal of Computation Linguistics & Chinese Language Processing*, Vol.19, No.2, (2014), 17-31

〈Table 15〉 Positive and negative summaries in the ‘driving’ aspect K5

Positive		
Ranking	Content	Score
1	특히 시승차는 공장에서 막 출고된 신차임에도 시내주행에서 최고 10km/ l 의 우수한 실연비를 기록했다. 최고출력은 165마력으로 기존과 같다. 향후 후속 모델에서는 출력 개선도 함께 이뤄졌으면 하는 바람이다.	0.50002
2	K5 터보 터빈업 차량이 안정적으로 출력을 올릴 수 있는 비결이 스테이지2 프로그레시브 메탄올 인젝션 시스템입니다.	0.50001
3	K5 하이브리드는 지난달 26일부터 이달 10일까지 16일 동안 하이브리드 차량 최초로 미국 48개주(州) 전역(알래스카, 하와이 제외)을 일주하며 최고 연비를 달성하는 기네스 기록에 도전했다.	0.50001
4	일단 주행성능은 엄청 만족스럽다고 할 수는 없지만 충분하다 정도로 말할 수 있겠습니다.	0.50001
5	이전의 소렌토에 비해 만족합니다 과속을 즐기는 스타일이 아니구요 적당히 밟고 다니는데 괜찮네요 안정성도 SUV에 비해 더 좋습니다.	0.50001
Negative		
Ranking	Content	Score
1	진짜 더시끄러워요... 소음에 민감한편은 아니지만 스트레스 받을정도로..	-0.49999
2	아스팔트에선 정숙성이 좋았는데 고속도로 노면에서는 소음이 좀 올라오네요	-0.49999
3	생각외로 차가 안나가는듯한....뒷바퀴쪽 소음...노면 타는 소리도 쫌올라오고 근데 핸들이 묵직하고 운전 포지션은 쫌 나오는 듯. 터보 나오면 휠 서스만 하면 공도 타기 딱 안성맞춤일듯해요	-0.49999
4	힘이 부칠경우 가속도가 떨어지는 단점이 있습니다.	-0.49999
5	하지만 기아차 내구성에 보면 정말 환장하실겁니다 주행하다 보면 여러잡소리에 소음에 정말 장난아닙니다	-0.49998

표 15은 K5 자동차 주행 토픽의 감성 요약 결과를 보여준다. 알고리즘 3을 적용하여 주행 토픽과 가장 관련 있는 상위 5개의 긍·부정 문장을 출력한 결과이다. 주요 긍정 내용은 우수한 실연비, 안정적인 출력, 충분한 주행성능에 대한 것이고, 부정 내용은 시끄러운 소음, 떨어지는 가속도, 내구성 등에 대한 내용이다.

〈Table 16〉 Positive and negative summaries in the 'design' aspect K5

Positive		
Ranking	Content	Score
1	K5의 매력은 뭐니뭐니해도 디자인이다. 특히나 군더더기 없는 측면라인은 인상적이다. 쏘나타의 유려한 곡선으로 만들어낸 측면라인과는 다른 모습의 디자인은 선택의 갈림길에서 가장 큰 영향을 미치는 요소이다. 일단 높은 벨트라인은 쏘나타나 동급 세단들이 추구하는 스포티한 이미지를 구축하는 동일한 요소이지만 헤드램프와 앞 팬더에서 시작되어 뒷팬더까지 한번에 이어지는 강인한 직선을 통해 전혀 다른 분위기를 연출하고 있다. 보닛과 앞유리의 각을 줄이고 루프에서 트렁크 리드로 흐르는 라인을 길게 설정해 쿠페라이크한 디자인을 보이는 것도 K7에서 이어져온 모습이다.	0.50011
2	곧 F/L 된다고 하는 소식이 들려오는 시점입니다. 6월정도를 예상하는 분위기던데 개인적으로는 k9, k7, k3 패밀리룩과 유사하게 F/L 될 것 같다고 생각됩니다. 개인 취향이겠으나 전 지금 디자인에 만족하고 있습니다.	0.50003
3	안을 들여다 보면 K7과 스포티지R에서 보았던 좌우를 크게 가로지르는 센터페시아의 모습을 볼 수 있다. 여기에 K7이 빛을 활용해 실내공간을 더욱 고급스럽게 연출해 독창성을 보였고 스포티지R이 일체형 패널로 중량감을 살렸다면 K5는 운전자를 향하고 있는 센터페시아를 통해 분위기를 살리고 있다. 운전자가 더욱 중시되는 중형차의 특성을 보여주는 부분이다.	0.50001
4	일단 외관은 기존 k5와 비슷하지만 좀도 세련된 느낌입니다. 후방등도 뒷휀다까지 이어지면서 세련되 보이고요	0.50000
5	여기에 피터슈라이어룩이라 불리는 라지에타 그릴과 날렵한 해드라이트는 상당히 공격적이지만 YF의 그것처럼 과하지 않은, 그래서 적절하게 역동적인 앞모습을 연출합니다.	0.50000
Negative		
Ranking	Content	Score
1	가죽의 재질은 고급/외제차의 그것을 기대하기 힘들지만 많이 발전한 느낌. 다만 주름이 생기는 고질적인 문제는 나타날 것 같음	-0.50000
2	k5실내 별로 같다고들 많이 하시더라구요 ??	-0.49999
3	트렁크쪽에 마감재를 바르다만 느낌이네요.....	-0.49997

표 16은 K5 자동차 디자인 토픽의 감성 요약 결과를 보여준다. 알고리즘 3을 적용하여 주행 토픽과 가장 관련 있는 상위 5개의 긍·부정 문장을 출력한 결과이다. 허나 부정의 부분에 보면 3개의 문장만 출력된 것을 알 수 있다. 이러한 원인은 K5 문서 집합 내에 디자인에 대해 언급된 내용이 적었고, 그림 5에서 나타났듯이 이러한 디자인 토픽에서도 부정에 대한 내용이 18%로 낮은 비율을 나타내고 있기에 3개의 문장만이 추출된 것이다. 주요 긍정 내용은 측면라인, 고급스러운 실내공간, 세련된 느낌, 역동적인 앞모습이며 주요 부정내용은 가죽의 주름, 실내, 트렁크 쪽의 마감재에 대한 내용이다.

〈Table 17〉 Positive and negative summaries in the ‘performance’ aspect SM5

Postive		
Ranking	Content	Score
1	변속기가 기존의 6단변속기가 아닌 CVT가 장착되었습니다. CVT의 주행특성은 위의 영상에서 속도계를 보지 마시고, 엔진 회전계를 보시면 됩니다. 알피엠이 꾸준히 올라가도 변속이 되지 않고 그냥 그대로 가속이 됩니다.	0.50002
2	그래도 수동모드시 엔진브레이크 걸리는 즐거움을 CVT 미션에서도 느낄 수 있다는 것은 장점입니다.	0.50001
3	열선은 앞좌석만 장착되어있고 2단으로 조절가능합니다.	0.50001
4	르노삼성자동차 (대표이사: 프랑수아 프로보)는 국내 가솔린 2000cc 중형차 중 최고 연비효율인 14.1Km/L의 연비와 가속성능이 향상된 프리미엄 웰빙 패밀리세단 SM5 에코-임프레션(Eco-Impresion)을 출시했다고 2일 밝혔다.	0.50001
5	2.0lpi엔진에 cvt미션 장착입니다. 기어가 수동모드가 지원되길래 6단자동인줄알았는데 cvt네요	0.50001
Negative		
Ranking	Content	Score
1	육기통임에도 이상하게 안 나가서 엄청 실망...	-0.50000
2	일단 연비 부문에서도 타사 대비 그렇게 뛰어난 점을 어필하지 못한 점 승차감 위주 세팅으로 인해 다이내믹한 운전은 포기해야 한다는 점 여성운전자에게만 추천을 한다라는 식의 표현은 정말 ... '3세대는 실패작이야' 라는 소리로 들리더군요.	-0.50000
3	다른차량에비해 연비가 엄청잘나오는거 같은데.. 그래도 연비가 안좋아서 티코를 하나 구입할예정..	-0.50000
4	엔단이나 파킹에 기어놓고 가만히잇으면 알피엠이 왓다갓다 하는데 뭐가 문제일까요? 그리고 핸들 완전꺽고 기어 r이나d놧을때 본넷쪽 진동은 뭐일까요? 잘아시는분잇으신가요?	-0.49999
5	'SM5 뉴임프레션 LPLi'를 모는 부산의 개인 택시기사들이 엔진 등에 꾸준히 문제가 발생하고 있다며 리콜 등 근본적인 대책을 마련해 달라고 르노삼성자동차에 요구했다.	-0.49996

표 17은 SM5 자동차 성능 토픽의 감성 요약 결과를 보여준다. 성능 토픽과 가장 관련 있는 상위 5개의 긍·부정 문장을 출력한 결과이다. 주요 긍정 내용은 CVT 변속기, 향상된 가속성능, 신차에 관한 내용이며, 부정 내용은 승차감 위주의 세팅, 좋지 못한 연비성능, 알피엠의 불안정, 엔진의 문제에 대한 내용이다.

〈Table 18〉 Positive and negative summaries in the ‘driving’ aspect SM5

Positive		
Ranking	Content	Score
1	르노삼성은 기존에 갖고 있는 품질이라는 이미지에 웰빙 드라이빙이라는 마케팅을 도입했다. 지금의 시장은 친환경과는 별도로 스포티가 대세지만 그보다는 편안함을 내세운 것이다. 이런 전략은 실내의 편의성에서 여실히 나타나고 동력 성능에도 반영된다. 하체도 승차감 위주를 지향한다. 기본적으로 뉴 SM3가 보여줬던 전략을 SM5에도 도입했다고 할 수 있다.	0.500021
2	서스펜션, 스티어링 모두 편안함 위주로 셋팅된게 차를 타서 달리자 마자 느껴집니다. 노면 충격도 대부분 하체가 흡수하구요	0.500004
3	현대차는 그래두 부앙앙앙 하고 좀 나가는 그런맛(?) 이 있는대 삼성차는 조용하고 편안함에 비중을뒀는지	0.500004
4	승차감 - 순정일시 승차감 정말 좋죠..맥퍼슨 스트럿과 멀티링크 코일스프링 조합으로 상당히 안정적인 승차감을 제공합니다.	0.49900
5	SM5 에코-임프레션(Eco-Impresion)은 유가상승으로 경제성있는 차량을 선호하는 고객의 요구에 맞추어 국내 가솔린 2000cc 동급 최고 연비 효율인 14.1Km/L 실현하였고, 업그레이드된 최첨단 뉴 엑스트로닉(New X-tronic) 변속기 적용 및 엔진성능을 최적으로 튜닝하여 가속 성능 및 승차감을 개선 하였으며, 최고 수준의 안전성과 웰빙(Well-Being)사양을 적용하여 한층 더 업그레이드된 경제적인 프리미엄 웰빙 패밀리세단으로 재탄생하였다.	0.40001
Negative		
Ranking	Content	Score
1	진짜 차 안나간다는 생각뿐이 안났습니다.. cvt미션특성상 동력손실도 많다고 알고 있었을 뿐더러	-0.50000
2	악셀밟아대면 RPM은 쭉쭉올라간다 . 근데 차가 안나간다	-0.50000
3	직접타보니 승차감개판에 타이어소리 엄청 들어오고 시끄럽던데	-0.50000
4	엔진쪽에서 다다다 거리는 소음이 발생합니다	-0.49999
5	그리고 엔진 소음은 밖에서 들으면 디젤 특유의 덜덜거리는 소리납니다.. 실내에서는 휘발유 차보다는 좀더 소리나는데 신경쓰이고 시끄러운 정도는 아닙니다.	-0.49999

표 18은 SM5 자동차 주행 토픽의 감성 요약 결과를 보여준다. 주행 토픽과 가장 관련 있는 상위 5개의 긍·부정 문장을 출력한 결과이다. 주요 긍정 내용은 웰빙 드라이빙, 편안함 위주의 세팅, 안정성인 승차감, 안정성과 웰빙 사양을 적용에 관한 내용이며, 부정 내용은 높은 동력손실, 주행성능, 좋지 못한 승차감, 주행 소음에 대한 내용이다.

〈Table 19〉 Positive and negative summaries in the ‘design’ aspect SM5

Positive		
Ranking	Content	Score
1	SM5의 초대 모델이 데뷔한 것은 1998년 3월 삼성자동차 때였다. 당시는 닛산의 맥시마를 베이스로 했었다. 2003년 9월 페이스리프트를 했고 2005년 1월 2세대로 진화했다. 1세대와는 달리 닛산 티아나를 르노삼성 버전으로 모디파이한 모델이었다. RSM만의 독창성을 주장하기에는 한계가 있었지만 참신한 스타일링 디자인과 마케팅으로 나름대로의 입지를 확대해 나갔다.	0.50003
2	개인적으로 매장방문후 무난한 디자인(YF소나타의 억지스런 디자인이 싫어서)과 고급스럽고 깔끔한 실내가 이뻐서 호감있게 보고 있었는데	0.50002
3	디자인도 얍쌀한게 유선형으로 잘 빠진편이고...	0.50000
4	외관 평 : 확실히 네이비 새로 나온 색 컬러감은 정말 좋았습니다. 광빨도 살아있어 자태가 아름답다고나 할까요... 요새 나오는 중형 추세이기도 하지만 크기도 나름 웅장하니 중형임에도 꽤나 큼지막 합니다. 사진엔 안나오는데 타이어 휠도... 순정인데도 너무 이쁘더라구요.... 러블리~	0.50000
5	실내디자인 - 깔끔한 이미지입니다. 당시엔 실내부분도 정말 괜찮다라고 생각했는데	0.50000
Negative		
Ranking	Content	Score
1	도어트림 엑센트를 보는듯하다. 디자인이란 없다. 정말 싼티난다.	-0.50000
2	촌스런 색깔이라고 느꼈습니다.	-0.49999

표 19은 SM5 자동차 디자인 토픽의 감성 요약 결과를 보여준다. 디자인 토픽과 가장 관련 있는 상위 5개의 긍·부정 문장을 출력한 결과이다. 부정에 대한 결과가 2개만 추출 된 이유는 표 16의 이유와 유사하다. 주요 긍정 내용은 참신한 스타일링, 고급스럽고 깔끔한 실내, 컬리감에 관한 내용이며, 부정 내용은 싼 티, 촌스러운 색깔에 대한 내용이다.

〈Table 20〉 Positive and negative summaries in the ‘performance’ aspect AVANTE

Positive		
Ranking	Content	Score
1	대한민국 준중형을 대표하는 아반떼는 한국자동차전문기자협회가 뽑은 ‘2016년 올해의 차’로 선정된 바 있는데요. 아반떼는 이에 안주하지 않고 아반떼 스포츠로 디자인에서 동력성능까지 모든 것을 새롭게 완성했습니다. 아반떼 스포츠만의 차별화된 디자인, 가솔린 1.6 터보 엔진이 만들어내는 폭발적인 힘과 리어 멀티링크 서스펜션의 안정적인 성능의 조합으로 스포티한 드라이빙을 새롭게 정의하고 있는데요. 더 많은 보통의 사람들이 더 강력한 파워와 스피드를 경험할 수 있도록 탄생한 아반떼 스포츠의 모든 것을 지금부터 공개합니다.	0.500083
2	테스트 결과 크루즈는 정면충돌과 측면충돌 별 다섯, 전복부문 별 넷을 받아 종합 안전도에서 최고점수(별 다섯개)를 얻었다.	0.500082
3	아반떼 쿠페는 당초 지난 연말에 국내에서 출시될 예정이었으나 마땅한 엔진이 없어 양산이 미뤄졌으며, 국내용의 경우, 1.8리터급 누우엔진이 탑재되는 미국용과 달리 성능이 개선된 2.0리터급 엔진이 장착될 예정인 것으로 알려지고 있다.	0.500068
4	신형 아반떼 스포츠의 세부 구성은 구체적으로 알려진 게 없으나, 쏘나타 1.6 터보의 파워트레인을 탑재한 차종으로 추정된다. 감마 1.6 터보 GDi 가솔린 엔진과 건식 7단 DCT 자동 변속기를 적용한다는 의미다.	0.500046
5	계기반의 구성도 좋고, 시인성도 좋고, 각종 스위치류의 조작 편의성이나 조작감 등등 훌륭합니다.	0.500033
Negative		
Ranking	Content	Score
1	육기통임에도 이상하게 안 나가서 엄청 실망...	-0.50000
2	지난해 1월에도 비슷한 사고가 있었다. 당시 운전자도 고속도로 1차선에서 차량의 쏠림 현상이 갑자기 나타나 비상도로에 세워놓고 보니 운전석 뒷바퀴가 빠져 버린 것이었다.	-0.5
3	안녕하세요 회원님들! 작년md신차구입해서 이제15000키로 탔는데 얼마전부터 시동걸면 핸들잠기면서 vdc등도 같이들어오더군요 시동껐다가 다시 시동걸면 증상 사라지고요. 이거 왜 이런현상인지 아시는분 있으세요? 조만간 as센터함 들릴려구요	-0.5
4	뻥트립 맞네요. 실제 연비와 트립 연비 2.5km/l 차이 납니다.	-0.49999
5	브레이크 또한 이상하게 한박자 느리다는 생각을 하였습니다 악셀과 비슷하게 어느정도 밟아줘야 선다는 느낌?	-0.49997

표 20은 아반떼 자동차 성능 토픽의 감성 요약 결과를 보여준다. 주요 긍정 내용은 차별화된 기술, 높은 안전도, 조작의 편의성이며 부정내용은 쏠림 현상, 부품 내구성, 잠기는 핸들, 브레이크가 있다.

〈Table 21〉 Positive and negative summaries in the ‘driving’ aspect AVANTE

Positive		
Ranking	Content	Score
1	디젤차와는 비교 할 수 없는 조용함	0.50000
2	깜박이 켰다. 우와~~깜박이 소리가 매우 좋다. 내차랑 산타페와 비교가 안된다. 함 들어보시라..고급차에 적용되는 소리다.	0.49999
3	공인연비만큼은 나오지 않지만 그래도 lpg라는걸 감안하면 정말 만족합니다. 고속 80 시내20 주행했을때 20000원이면 380키로정도 탑니다. 고속도로만 타면 20000원에 400키로는 그냥 탑니다.	0.49999
4	하체는 처음에는 시내주행시 너무 딱딱하고 방지턱에서 통통 튀는 느낌이였으나 고속에선 오히려 안정감 있고 잘잡아주는 느낌이여서 좋았습니다.	0.49999
Negative		
Ranking	Content	Score
1	물론 하이브리드 시스템의 내구성에 대한 의문이나 오토스탑이 해제되어 시동이 걸릴때의 진동의 발생등 앞으로도 개선해 나아가야 할 부분도 보입니다.	-0.50000
2	정숙성은 녹색불에서는 봐줄만하나 조금만 더 밟아도 소음이 조금 있더군요 고속도로에서도 120이상 좀 심하구요	-0.50000
3	1. 엔진소음 - GDI 특유의 엔진소음 정말 거슬립니다. 거짓말 더 보태서 디젤 엔진소음이랑 비슷합니다.	-0.49999
4	md기름덜먹고 무난하게 타기엔좋은데 핸들, 기어노브도 금방까지고 잡소리에..현대 품질에 좀더 신경써줬으면..	-0.49999
5	부드러운 주행중엔 안나고,,, 엔진에 부하가 걸릴때 진동이 커지면 덜덜덜덜 뭔가가 떨리면서 부딪히는 듯한 소리가 나네요	-0.49999

표 21은 아반떼 자동차 주행 토픽의 감성 요약 결과를 보여준다. 긍정에 대한 결과가 4개만 추출된 이유는 표 16의 이유와 유사하다. 주요 긍정 내용은 주행 중 정숙성, 좋은 소리, 만족스러운 연비, 고속에서의 안정감이 있고, 부정 내용으로는 주행 중 진동의 발생, 엔진 부하 시 소음이 있다.

〈Table 22〉 Positive and negative summaries in the ‘design’ aspect AVANTE

Positive		
Ranking	Content	Score
1	전면과 후면범퍼는 공기저항계수를 낮추기 위해 역동적인 에어로파츠 디자인을 채용했으며 옆면 아래쪽에는 사이드실 몰딩을 장착하여 연비를 향상시켰다고 합니다. 또한 트렁크 끝부분에는 리어스포일러가 장착되어 최대한 공기저항을 줄이기 위한 노력이 가미되었음을 볼수 있습니다. (실제로 아반떼 LPI 하이브리드 모델의 경우 기존 아반떼보다 공기저항계수가 0.03cd가 낮아졌습니다)	0.50001
2	역동적으로 한층 더 진화된 디자인	0.50001
3	‘아반떼 스포츠’로 명명된 신차는 1.6터보 GDi 엔진을 최초 탑재하고 내외관 디자인을 보다 역동적으로 꾸몄다.	0.50001
4	디자인이 날렵하게 너무 세련되고 스포츠카 느낌이라고나 할까..	0.50001
5	내장도 제네시스 비슷한 고급스러운 디자인이라 하더군요.	0.50000
Negative		
Ranking	Content	Score
1	기분 나뻣던점. 1. 싼티 좔좔 흐르는 내부와 세세한 부분의 배려따위 없는 인테리어.(운전석 조수석에 등이 안달려 있고 선글라스 수납함도 없어요~~)	-0.50000
2	튜익스 옵션도 들어가 있었네요. 튜익스 서스펜션 나쁘지 않았습니다. 가속 좋습니다. 쭉쭉 잘나가네요 인테리어는 싼티 느낌이 납니다.	-0.50000
3	이부분은 인조가죽 + 플라스틱으로 되어있는데 싼티가 펄펄 넘칩니다. 그래도 싸니까뭐....	-0.50000
4	수동 깡통인데 인테리어 싼티나는 플라스틱 도배	-0.50000
5	외관 : 구형티가 제법나죠. 갠적으로 궁딩이는 아직 이쁘다고 생각합니다	-0.49999

표 22은 아반떼 자동차 디자인 토픽의 감성 요약 결과를 보여준다. 주요 긍정 내용은 에어로파츠 디자인, 역동적인 디자인, 날렵하고 세련된 디자인, 고급스러움이 있고, 주요 부정 내용은 싼 티 나는 내부, 플라스틱 도배, 인조가죽, 구형 티가 나는 외관이 있다.

Abstract

Latent topics-based product reputation mining

Sang-Min Park* · Byung-Won On**

Data-drive analytics techniques have been recently applied to public surveys. Instead of simply gathering survey results or expert opinions to research the preference for a recently launched product, enterprises need a way to collect and analyze various types of online data and then accurately figure out customer preferences.

In the main concept of existing data-based survey methods, the sentiment lexicon for a particular domain is first constructed by domain experts who usually judge the positive, neutral, or negative meanings of the frequently used words from the collected text documents. In order to research the preference for a particular product, the existing approach collects (1) review posts, which are related to the product, from several product review web sites; (2) extracts sentences (or phrases) in the collection after the pre-processing step such as stemming and removal of stop words is performed; (3) classifies the polarity (either positive or negative sense) of each sentence (or phrase) based on the sentiment lexicon; and (4) estimates the positive and negative ratios of the product by dividing the total numbers of the positive and negative sentences (or phrases) by the total number of the sentences (or phrases) in the collection. Furthermore, the existing approach automatically finds important sentences (or phrases) including the positive and negative meaning to/against the product. As a motivated example, given a product like Sonata made by Hyundai Motors, customers often want to see the summary note including what positive points are in the ‘car design’ aspect as well as what negative points are in thesame aspect. They also want to gain more useful information regarding other aspects such as ‘car quality’, ‘car performance’, and ‘car service.’ Such an information will enable customers to make good choice when they attempt to purchase brand-new vehicles. In addition, automobile makers will be able to figure out the preference and positive/negative points for new models on market. In the near future, the weak points of the models will be improved by the sentiment analysis. For this, the existing approach computes the sentiment score of each

* Department of Software Convergence Engineering, Kunsan National University
** Corresponding Author: Byung-Won On,
Department of Software Convergence Engineering,
Kunsan National University, 558 Daehak-ro, Gunsan-si, Jeollabuk-do 54150, Korea,
Tel: +82-63-469-8913, Fax: +82-63-469-7423, E-mail: bwon@kunsan.ac.kr

sentence (or phrase) and then selects top-k sentences (or phrases) with the highest positive and negative scores.

However, the existing approach has several shortcomings and is limited to apply to real applications. The main disadvantages of the existing approach is as follows:

(1) The main aspects (e.g., car design, quality, performance, and service) to a product (e.g., Hyundai Sonata) are not considered. Through the sentiment analysis without considering aspects, as a result, the summary note including the positive and negative ratios of the product and top-k sentences (or phrases) with the highest sentiment scores in the entire corpus is just reported to customers and car makers. This approach is not enough and main aspects of the target product need to be considered in the sentiment analysis.

(2) In general, since the same word has different meanings across different domains, the sentiment lexicon which is proper to each domain needs to be constructed. The efficient way to construct the sentiment lexicon per domain is required because the sentiment lexicon construction is labor intensive and time consuming.

To address the above problems, in this article, we propose a novel product reputation mining algorithm that (1) extracts topics hidden in review documents written by customers; (2) mines main aspects based on the extracted topics; (3) measures the positive and negative ratios of the product using the aspects; and (4) presents the digest in which a few important sentences with the positive and negative meanings are listed in each aspect. Unlike the existing approach, using hidden topics makes experts construct the sentimental lexicon easily and quickly. Furthermore, reinforcing topic semantics, we can improve the accuracy of the product reputation mining algorithms more largely than that of the existing approach. In the experiments, we collected large review documents to the domestic vehicles such as K5, SM5, and Avante; measured the positive and negative ratios of the three cars; showed top-k positive and negative summaries per aspect; and conducted statistical analysis. Our experimental results clearly show the effectiveness of the proposed method, compared with the existing method.

Key Words : topic model, opinion mining, text summarization, data analytics, public survey

Received : March 8, 2017 Revised : May 4, 2017 Accepted : May 15, 2017
Publication Type : Regular Paper Corresponding Author : Byung-Won On

저 자 소 개

박 상 민
군산대학교 소프트웨어융합공학과의 학부 4학년 재학 중이며 대학원에 진학할 예정이다. 연구 분야는 빅데이터 기반의 여론조사 소프트웨어 개발이며, 관심 분야는 데이터 마이닝과 인공지능이다.

온 병 원
2007년, 미국 펜실베이니아 주립대학교의 컴퓨터공학과에서 박사학위를 취득한 후, 캐나다 브리티시 컬럼비아 대학교에서 박사 후 연구원으로 재직하였다. 2010년, 미국 일리노이 대학교의 차세대디지털과학센터에서 선임연구원으로 근무하였고, 서울대학교 차세대융합기술연구원에서 연구교수를 역임하였다. 현재는 군산대학교 소프트웨어융합공학과의 조교수로 재직 중이다. 주요 연구 분야로는 데이터 마이닝, 정보검색, 빅데이터, 인공지능 등이다.