감성분석 연구 동향

이정훈*
*경기대학교 응용통계학과
e-mail:vhrehfdl7tp@naver.com

Sentimental Analysis Research Trends

Jung-Hoon Lee*
*Dept. of Application Statistics, Kyonggi University

요 약

비정형 데이터 증가로 텍스트 마이닝을 사용해 데이터를 분석하는 연구가 주목받고 있다. 감성분석은 단어와 문맥을 분석하여 텍스트의 감정을 파악하는 기술이다. 본 논문에서는 감성분석 연구 동향, 적용분야, 방법론에 관해 분석하고 기술하려 한다. 감성분석은 2001년 채팅의 감정을 분석하면서 시작되었고, 2008년부터 본격적으로 연구가 진행되었다. 감성분석은 SNS, 상품 후기, 영화평, 뉴스 기사 등 다양한 데이터에 적용되고 있으며, 사회이슈 찬반 분석과 장소 선호도 분석 등 다양한 연구에서 사용되었다. 감성분석 방법은 감성사전을 이용하는 방식과 기계학습을 사용하는 방식으로 나누어지며 분석 방법을 발전시키기 위한 연구가 진행되고 있다.

1. 서론

스마트폰 대중화와 SNS 발전으로 비정형 데이터 규모가 증가하고 있다.[1,2] 과거에는 수작업으로 분석할 수 있던 작업이 현재는 방대해진 데이터 때문에 불가능하게 되었다. 따라서 문서를 자동으로 분석, 분류하는 기술인 텍스트 마이닝에 관한 관심과 필요성이 증가하고 있다.[3]

감성분석은 단어와 문맥을 분석하여 텍스트의 감정을 파악하는 기술이다. 감성분석을 설문조사와 같이 수작업으로 시행하면 많은 시간과 비용이 발생하게 된다.[4,5] 그러나 소프트웨어를 사용해 자동화된 분석 기법을 적용하면 비용과 시간이 적게 들어 즉각적인 분석이 가능해진다.[6,7] 기업은 소비자 반응을 파악해 발전된 제품과 서비스를 제공할 수 있고, 사회는 특정 이슈에 관한 대중 여론을 파악해 정책에 반영할 수 있는 장점이 있다.[4,8]

해외에서는 이미 감성분석과 관련된 연구들이 많이 진행되었으며 실제 서비스에 적용되고 있는 단계이다.[7,9] 영어는 한국어보다 오랫동안 연구되었기 때문에 감성분석의 핵심인 감성사전이 방대하고 정교하게 구축돼있다.[3] 최근 국내에서도 많은 한국어 감성분석 연구를 진행하였지만, 아직 영어보다 부족한 수준이다.[10,11]

본 논문에서는 한국어와 관련된 감성분석 연구 동향을 파악하고자 한다. 과거에서 현재까지 등재된 국내 논문을 비교, 분석하여 감성분석 연구 동향, 적용분야, 분석방법을 기술하려 한다.

2. 감성분석 연구 동향

한국에서 감성이란 키워드가 처음 언급된 논문은 1993년에 감성공학 정의를 설명하고 연구방법과 활용을 제시한 논문이다.[12] 이 논문에서는 감성공학을 인간의 감정을 공학적으로 연구하고 활용하는 학문으로 정의하였다.

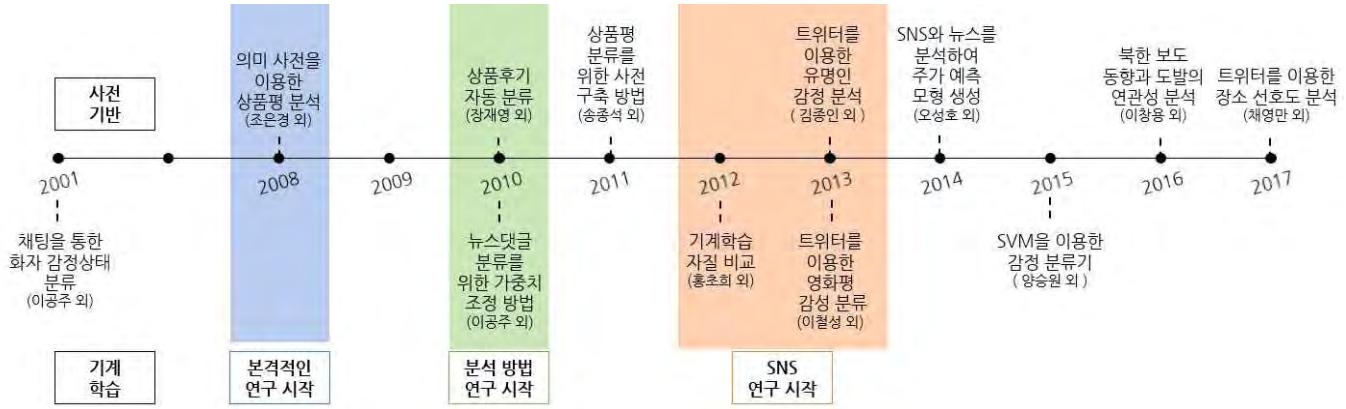
2001년에는 HMM 알고리즘을 사용해 채팅을 긍정, 질문, 미소, 부정, 고민 등 총 10가지 감정으로 분류하는 논문이 발표되었다.[13] 이 논문에서는 문장 내 단어를 어절로 분리하고 기계학습을 적용해 TexMo라는 감성분석 소프트웨어를 구현하였다.

2008년은 감성분석과 관련된 연구가 본격적으로 시작된 시기이다.[14] 감성사전을 이용해 상품평 분석 연구를 한 논문에서는 형태소 분석을 하고 감성사전을 구축해 상품 분류 모델을 구현하였다.[15]

2010년에는 분석 정확도를 높이기 위해 감성분석 방법과 관련된 연구들이 주로 시행되었다. 자동화된 감성사전을 만들기 위한 연구와 기계학습에 적합한 학습요소를 발견하는 연구 등 여러 가지 방법을 적용하였다.[16,17]

2012년에는 트위터, 페이스북 등 SNS가 한국에서 서비스하고 유행하게 되면서 많은 비정형 데이터를 얻을 수 있게 되었다.[18] 이후 SNS와 관련된 연구들이 많이 시행되고 있으며 주로 트위터를 이용한 연구가 진행되고 있다.

이후 감성분석은 SNS, 상품 후기, 영화평 등 다양한 분야에 활발하게 적용되고 있다. 특히 2016년에는 북한의 보도 동향과 같은 생소한 데이터를 사용하였다. 또한, 분석 정확도를 높이기 위한 연구도 진행되고 있다.



(그림 1) 년도별 감성분석 연구 동향

3. 감성분석 적용분야

3.1 SNS

여러 감성분석 적용분야 중 대표적으로 SNS와 관련된 연구가 많이 시행되었다. SNS를 분석하여 유명인 감정상태 파악, 기상청 만족도 분석, 사회이슈 찬반 파악, 실시간 장소 추천, 주가예측 모형 생성, 장소 선호도 분석 등 다양한 연구를 진행했다.[1,5,8,19,20,21] 여러 SNS 중 트위터를 분석한 연구가 많이 이루어졌다. 트위터는 사용자가 많고 오픈 API를 지원하기 때문에 간편하게 대규모 데이터를 수집할 수 있다.[1,21] 그러나 트위터는 제한된 글자 수로 빠르게 전송하기 때문에 오탈자와 축약어가 많아 정확한 의미분석이 힘든 문제가 있다.[18]

3.2 상품 후기

상품 후기는 소비자 의사결정에 직접적인 영향을 미치는 중요한 요소이다.[4] 상품 후기를 분석하여 상품평 자동분류, 상품평 의미분석, 소셜커머스와 오픈마켓 이용경험 비교 등 여러 연구를 진행하였다.[4,15,23] 상품 후기는 상품 종류에 따라 단어의 감정이 달라지는 특성이 있다. 예를 들어 과일에서 '꿀'이란 단어는 긍정을 뜻하지만, 의류에서 '꿀'은 관련 없는 단어이다. 이러한 특성으로 상품 후기 분석은 주제별 감성사전을 만들어 사용하는 연구가 진행되고 있다.[8,20,23]

3.3 영화평

영화평에도 감성분석을 적용하여 감성분석 사례분석 연구, 영화 흥행 예측, 감정 키워드에 따른 영화 검색 시스템 구현을 진행하였다.[14,22,24] 여러 분야 중 영화와 관련된 연구가 주로 이루어지는 이유는 영화와 관련된 평점과 후기들이 많기 때문이다.

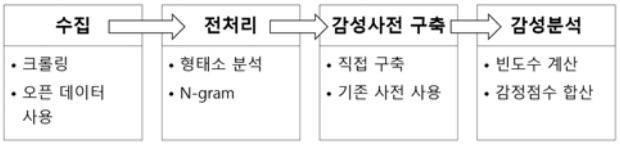
3.4 뉴스

감성분석은 텍스트가 존재하는 모든 분야에 적용할 수 있다. 북한의 보도자료를 분석하여 북한 도발과 연관성을 비교한 연구와 북한과 관련된 뉴스를 분석하여 평화지수를 분석한 연구가 있다.[2,25]

4. 감성분석 방법론

4.1 사전 기반 분석

사전 기반 분석 방법은 데이터에서 감성사전과 일치하는 단어가 존재하면 감성점수를 부여해 감정을 분류하는 방법이다. 사전 기반 방법은 수집, 전처리, 감성사전 구축, 감성분석 과정으로 이루어진다.



(그림 2) 사전 기반 분석 흐름

4.1.1 수집

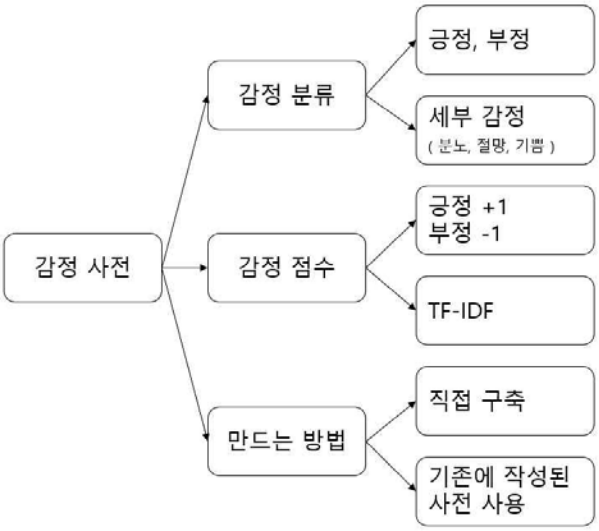
첫 번째로 수집은 크롤링을 이용해 직접 수집하거나 혹은 공공기관이나 상용 서비스에서 제공해주는 데이터를 사용한다.[1,3,26] 수집은 분석의 기초적인 단계로 데이터가 있어야 분석을 시작할 수 있으며 많은 데이터를 수집하는 것이 중요하다.

4.1.2 전처리

두 번째로 전처리는 분류 정확도를 높이기 위해 수집한 데이터를 가공하는 작업이며 대표적으로 형태소 분석과 N-gram 방식이 사용되고 있다. 형태소는 더는 분해할 수 없는 단어를 뜻한다. 형태소 분석은 다양한 형태의 문장과 문서에 유연하게 대응하기 위해 사용된다. 특히 어미와 조사 조합에 따라 단어 형태의 변화가 큰 한국어는 단어 원형을 복원하는 작업이 필요하다. 형태소 분석 후 명사, 형용사, 동사 등 핵심적인 품사를 추출한다.

N-gram 방법은 문장을 2음절 또는 3음절 단위로 끊어 단어를 찾는 방법이다.[17] 품사 추출 후 정확한 분석을 위해 데이터 특성에 맞춰 추가 전처리 작업을 해준다. 예를 들어 SNS는 특성상 오탈자, 띄어쓰기 오류, 축약이 많으므로 띄어쓰기 교정기와 맞춤법 검사기를 사용해 전처리해준다.

4.1.3 감성사전 구축



(그림 3) 감성사전 구축 방법

세 번째로 감성사전을 구축한다. 감성사전은 감정단어, 감정분류, 감정점수로 구축되어 있으며 연구 목적에 따라 구성이 달라진다. 사전은 감정분류를 긍정, 부정과 같이 극성으로 분류한 사전과 기쁨, 분노, 절망 등 세부 감정으로 분류한 사전이 있다.

감정점수를 부여하는 간단한 방법은 긍정이면 +1점 부정이면 -1점을 주는 것이다. 하지만 이 방법은 감정의 강도를 반영할 수 없다는 문제가 있다. TF-IDF 방법은 단어의 빈도를 계산해 점수를 계산하는 방법이다. TF는 한 문서에서 단어가 언급된 빈도수이고 IDF는 전체 문서에서 해당 단어가 언급된 문서의 빈도수이다. TF와 IDF를 곱해 감정점수를 계산하면 감정의 강도를 표현할 수 있고 더욱 정교한 분석이 가능해진다.[22]

<표 1> 감성사전 구조

감정단어	감정분류	감정점수
행복	긍정	1.13
우울	부정	2.23
절망	슬픔	2.12

감성사전은 사전을 직접 만들거나 기존에 구축된 감성사전을 이용하는 방법이 있다. 사전을 만드는 방법은 수작업으로 만드는 방법과 자동으로 만드는 방법으로 나뉜다. 수작업으로 사전을 만들면 정교한 구축이 가능하지만 많은 시간과 비용이 소모된다.[16] 반면 자동화된 방법은 적은 비용으로 방대한 규모의 사전을 구축할 수 있지만, 정확도가 떨어질 수 있다.[26]

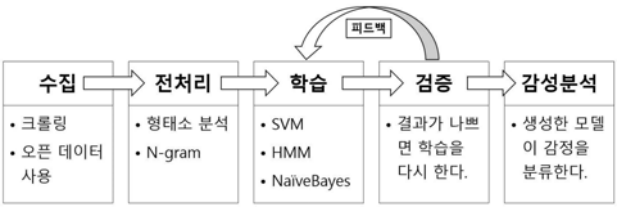
사전을 직접 만들지 않고 기존에 제작된 감성사전을 사용하는 방법도 있다.[26] 영어는 SentiwordNet 이라는 대규모의 감성사전이 구축되어 있고 한국어는 서울대에서 만든 KOSAC과 다른 연구에서 제작한 사전들이 있다. 그러나 현재 한국어 사전은 영어와 비교하면 양과 질이 빈약한 문제가 있다.[3,10] 한국어 감성사전의 한계를 해결하기 위해 감정단어를 한국어->일어->영어로 번역하여 분석을 시도한 연구도 진행되었다.[3]

4.1.4 감성분석

마지막으로 감성사전을 사용해 감성분석을 한다. 분석은 감정단어 빈도를 계산하는 방법과 감정점수를 합산하는 방법이 있다. 빈도수 체크방법은 긍정/부정 단어와 일치하는 빈도수를 계산해 감정을 분류하는 방법이다.[3] 점수 합산 방법은 사전과 일치하는 단어의 감정점수를 계산하여 분류하는 방법이다. 이 방법은 강조 부사와 문장 위치에 가중치를 부여하여 문맥을 반영한 분석이 가능하다.

4.2 기계학습 기반 분석

기계학습을 이용한 분류 방법은 수집, 전처리, 학습, 검정, 감성분석으로 이루어진다. 수집과 전처리는 사전 기반 분석과 같다.



(그림 4) 기계학습 감성분석 흐름도

4.2.1 학습

학습은 전처리 된 데이터를 기계학습 알고리즘에 적용해 분류 모델을 만드는 과정이다. 기계학습에 사용되는 알고리즘 종류는 다양한데 그 중 분석에 자주 사용되는 알고리즘은 SVM, HMM, Naive-Bayes 등이 있다.

SVM은 데이터를 구분하는 분류 선을 찾는 기법이며 감성분석 연구에 많이 사용되고 있다. SVM은 학습 데이터에 대해 과적합 되지 않고 kernel을 사용해 비선형 데이터에 대해서도 분류 가능 하다는 장점이 있다.[11] SVM을 사용해 신문기사의 댓글 극성파악과 트위터의 감성분석 등 여러 연구가 진행되었다.[11,17,18]

4.2.2 검증

학습이 끝나면 테스트 데이터를 사용해서 모델이 정확하게 감정을 분류하는지 검증을 해야 한다. 검증 후 문제가 있으면 피드백을 해 좋은 분류 모델을 생성할 때까지 학습과 검증을 반복한다.

기계학습 방법의 단점은 학습시킬 때 데이터가 사전에 분류되어 있어야 한다는 것이다. 만약 수집한 데이터 규모가 크고 분류되어 있지 않다면 학습데이터를 직접 분류해야 한다. 수동으로 분류하는 과정 중 잘못된 분류를 하게 되면 학습 성능에 악영향을 미칠 수 있다.[27]

5. 결론

본 논문에서는 국내에서 연구된 감성분석 연구 동향, 활용분야, 방법론에 관해 비교, 분석하여 기술하였다. 감성분석은 2001년에 시작되어 최근까지 진행되고 있으며 초기에는 상품 후기에 주로 적용되었지만, 점차 다양한 형태 데이터에 분석을 적용하였다. 분석방법은 크게 사전 기반 분석과 기계학습 기반 분석으로 나뉘며 각각의 방식으로 발전하고 있다. 앞으로 감성분석은 데이터가 많아질수록 필요성이 증가하며 주목받게 될 것이다.

참고문헌

[1] 김종인, 김경록, 문남미 “트위터 타임라인을 이용한 감정 상태 분석” 제40회 한국정보처리학회 추계학술발표대회 논문집 제20권 2호 (2013. 11)

[2] 이창용, 문호석 “텍스트마이닝을 이용한 북한 보도동향과 북한 도발과의 연관성 분석” 국방연구 2016년 12월 제59권 제4호, pp. 103-124

[3] 김영민, 정석재, 이석준 “소셜 미디어 감성분석을 통한 주가 등락 예측에 관한 연구” Entrue Journal of Information Technology December 2014 / Vol.13, No.3

[4] 장재영 “온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘” 한국전자거래학회지 14권 제4호

[5] 김인겸, 김혜민, 임병환, 이기광 “감성분석 결과와 사용자 만족도와의 관계 -기상청 사례를 중심으로-” 한국콘텐츠학회 논문지 '16 Vol. 16 No. 10

[6] 김상도, 박성배, 박세영, 이상조, 김권양 “음절 커널 기반 영화평 감성분류” 한국지능시스템학회 논문지 2010, Vol. 20, NO. 2, pp. 202-207

[7] 장필식 “소셜 데이터의 주된 감성분석에 대한 연구” Journal of The Korea Society of Computer and Information Vol. 19, No. 12, December 2014

[8] 강선아, 김유신, 최상현 “텍스트마이닝을 이용한 사회이슈 찬반 분류에 관한 연구” 한국데이터정보과학회지 2015, 26(5), 1167-1173

[9] 이철성, 최동희, 김성순, 강재우 “한글 마이크로블로그 텍스트의 감정 분류 및 분석” 정보과학회논문지 데이터베이스 제40권 제 3호(2013.6)

[10] 안주영, 배정환, 한남기, 송민 “텍스트 마이닝을 이용한 감정 유발 요인 'Emotion Trigger'에 관한 연구 "Bibliographic info: J Intell Inform Syst 2015 June: 21(2): 69~92

[11] 임좌상, 김진만 “한국어 트위터의 감정분류를 위한 기계학습의 실증적 비교” 멀티미디어학회 논문지 Vol.17, No. 2, February 2014 (pp. 232-239)

[12] 이구형 “감성공학의 연구와 제품개발에의 응용” 대한인간공학회 1993년 춘계학술대회논문집

[13] 문현구, 장병탁 “HMM을 이용한 채팅 텍스트로부터의 화자 감정상태 분석” 한국정보과학회 가을 학술발표논문집Vol. 28. No. 2

[14] 조은경 “감성 분석 연구의 현황과 말뭉치에 기반한 사례 분석” 언어과학연구 제60집

[15] 명재석, 이동주, 이상구 “반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템” 정보과학회 논문지 소프트웨어 및 응용 제35권 제 6호 (2008.06)

[16] 송종석, 이수원 “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축” 정보과학회논문지 소프트웨어 및응용 제38권 제 3호, 2011.3, 157-168

[17] 이공주, 김재훈, 서형원, 류길수 “뉴스 댓글의 감정 분류를 위한 자질 가중치 설정” 한국마린엔지니어링학회지 제34권 제6호, pp.871~879, 2010. 9

[18] 홍초희, 김학수 “트윗 감정 분류를 위한 다양한 기계학습 자질에 대한 비교 연구” 한국콘텐츠학회 논문지 12 Vol. 12 No. 12

[19] 오평화, 황병연 “트위터의 감정 분석을 통한 실시간 장소 추천 시스템” The Journal of Society for e-Business Studies Vol.21, No.3, August 2016, pp.15-28

[20] 김동영, 박제원, 최재현 “SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형 비교 연구” Journal of Information Technology Services 제13권 제3호 2014년 9월

[21] 채인영, 이영민, 유기윤, 김지영 “소셜 미디어 텍스트를 이용한 장소 선호도 분석 기법” 한국지형공간정보학회지 Vol.25 No.4 December 2017 pp.55-64

[22] 문성민, 하효지, 이경원 “영화의 흥행 성과와 리뷰 감정어휘와의 관계 분석” Design Convergence Study 53 Vol.14. no.4 (2015.8)

[23] 채승훈, 임재익, 강주영 “사용자 리뷰를 통한 소셜커스와 오픈마켓의 비교분석” Bibliographic info: J Intell Inform Syst 2015 December: 21(4): 53~77

[24] 오성호, 강신재 “사용자 영화평의 감정어휘 분석을 통한 영화검색시스템” 한국산학기술학회 논문지, vol14, No. 3, pp. 1422-1427, 2013.

[25] 권오병, 박다솔, 최지혜, 이재윤 “비정형자료로부터의 평화지수 분석을 통한 한반도 정세 파악 방법” 한국IT서비스 학회지 Vol.12, No.4 , 2013.12

[26] 조정태, 최상현 “영화리뷰 감성 분석을 통한 평점 예측 연구” 대한경영정보학회 제34권 제3호 2015년 9월 pp. 161-177

[27] 홍소라, 정연오, 이지형 “대용량 소셜 미디어 감성분석을 위한 반감독 학습기법” Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 5, October 2014, pp. 482-488