IDG
Custom Content Services

IT WORLD CIO

IDG Tech Report

"글에서 감정을 읽다" 감성 분석의 이해

소셜 미디어는 이제 단순한 소통의 도구가 아닌, 분석의 가치가 있는 새로운 정보의 바다다. 소셜 미디어의 방대한 데이터를 분석하는 감성 분석은 대중의 기분을 감지해 주가나 선거 결과를 예측할 수 있을 뿐만 아니라, '누가 무엇을 왜' 좋아하는지에 대한 인사이트도 제공한다. 기존의 소셜 분석은 단순히 페이스북의 '좋아요' 횟수에만 주목했다. 그러나 이에 반해 감성 분석은 더 심도 있는 소셜 마케팅을 가능하게 한다. 이제 마케터에게 있어 감성 분석은 '고려의 대상'이 아닌, '필수적으로' 도입할 기술이 됐다. 감성 분석이 무엇인지, 그리고 그 활용 방안과 발전 과제를 살펴보자.

- 새로운 마케팅 리서치 수단
- 마케팅 팀이 감성 분석 툴에 주목해야 하는 이유
- Top 6 웹 분석 서비스 업체 : 포레스터 리서치
- 감성 분석의 3단계
- 감성 분석 기술의 현주소
- 인공 지능 연구의 최전방

"글에서 감정을 읽다" 감성 분석의 이해

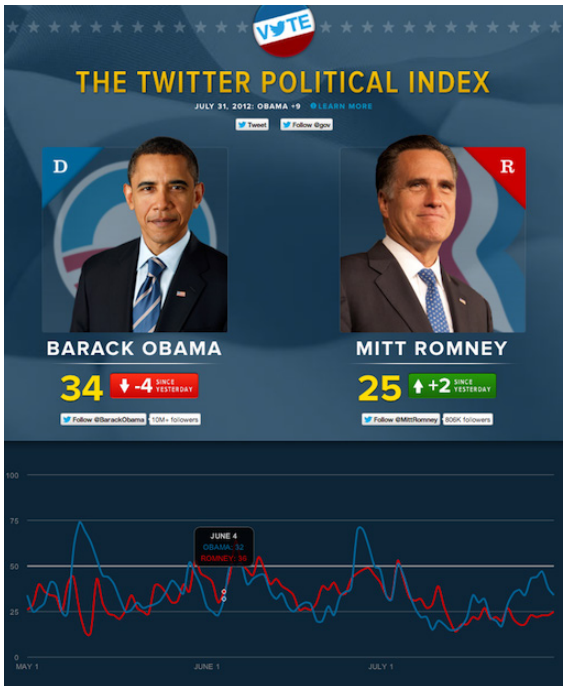
신수정 기자 | ITWorld

포브스가 2012년 발표한 '10년 전에는 존재하지 않았던 유망직업 10선'에는 CLO(Chief Listening Officer)가 포함돼 있다. CLO는 트위터와 페이스북 등, 각종 소셜 미디어에서 고객이 자사의 제품이나 브랜드에 대해 어떤 이야기를 하는지 귀 기울이며 도움이 될 만한 정보를 가려내는 사람이다.

소셜 미디어의 데이터는 설문조사나 인터뷰 자료처럼 인위적으로 제어된 환경에서 생산된 것이 아닌, 소비자가 자발적으로 표현한 '날 것 그대로'의 데이터라는 점에서 소비자를 가장 잘 파악할 수 있는 마케팅 지표로 각광받고 있다. 이러한 소셜 데이터에서 소비자가 특정 대상에 대해 느끼는 좋고 싫음, 그리고 나아가 그 이유를 분석해 주는 '감성 분석(Sentiment Analysis)'은 소셜 마케터들의 관심을 사로잡고 있다.

새로운 마케팅 리서치 수단

감성 분석(Sentiment Analysis)은 '오피니언 마이닝(Opinion Mining)'으로도 불리는데, 이는 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술이다. 원문인 'sentiment'를 번역해 '감성 분석', 혹은 '감정 분석'으로도 불리나 기사에서는 현재 널리 쓰이고 있는 '감성 분석'이라는 용어를 사용한다.



감성 분석은 빅데이터 분석의 일환으로 이해할 수 있는데, 이는 감성 데이터가 빅데이터 중에서도 사람들의 주관적인 의견이 드러난 것이기 때문이다. 감성 분석이 성공적으로 활용된 가장 대표적인 예는 '빅데이터 선거'라고도 불렸던 2012년 미국 대선에서 찾을 수 있다.

오바마 캠프는 대선 2년 전부터 선거 캠페인 팀에 빅데이터 팀을 설치하고 수집할 수 있는 모든 정보를 수집, 수치화해 의사결정의 근거로 활용했다. 이들이 수집한 데이터에는 나이, 성별, 지역별 인구분포와 같은 일반적인 통계 데이터뿐만 아니라 개인의 정치적 성향이나 취미, 관심분야 등이 나타난 소셜 데이터도 있

었다.

오바마 캠프는 부동층을 공략하기 위해 이러한 소셜 데이터를 수집해 유권자 개인을 대상으로 한 맞춤 선거 전략을 펼쳤다. 예를 들어, 이들은 소셜 데이터를 바탕으로 헐리우드에서 열리는 정치헌금 디너파티에 참가해 기부금을 낼 가능성이 가장 높은 유권자 그룹이 40대 여성이라는 것을 알아냈다. 또한, 이러한 유권자 집단에 가장 어필할 수 있는 배우가 조지 클루니라는 분석 결과까지 활용해 디너파티에 클루니를 초대했고, 성공적으로 헌금을 모금할 수 있었다.

이처럼 감성 분석은 오바마의 선거 전략과 같이 여론 분석의 도구로 활용될 수도 있으나, 소비자가 느끼는 바를 직접적으로, 별도의 조사 과정 없이 파악할 수 있다는 점에서 기업의 마케팅 도구로도 활용될 수 있다. 물론, 소비자의 감성을 측정하기 위해 굳이 감성 분석 기술을 사용하지 않고도 '별점'과 같은 간단한 수치 자료를 이용할 수도 있기는 하다.

그러나 리뷰와 별점이 일치하지 않거나, 의도적으로 별점을 높게 측정해 신뢰성이 떨어지는 리뷰가 존재한다는 문제점이 있다. 이와 같은 '신뢰도'의 문제를 해결하기 위해서는 실제 소비자가 쓴 리뷰 텍스트 자체를 분석해 의견의 긍정 또는 부정 여부를 파악해야 한다. 감성 분석을 활용하면 방대한 데이터를 실시간으로 수집하고 분석하기 때문에 신뢰도를 높일 수 있을 뿐만 아니라, 조사 과정에 있어 시간 차에 따른 오류를 최소화할 수 있는 장점이 있다. 바로 이것이 감성 분석 기술이 등장한 배경이다.

텍스트와 멀티미디어와 같은 비정형 데이터의 수집, 분석, 관리 서비스를 제공하는 코난테크놀로지의 김문희 마케팅 부장은 "이러한 소셜 미디어의 분석과 관련한 수요는 지난해와 비교했을 때 5~6배 정도 늘어났다. 특히 PR, 마케팅 쪽에서 소셜 데이터 분석의 필요성을 확실히 느끼는 것 같다. 지난해까지만 해도 과연 감성 분석이 효과적일지 반신반의하는 분위기였으나 올해부터는 확연히 달라진 업계의 반응을 느낄 수 있었다"고 말했다.

김문희 부장은 이어, "금융, 방송 통신 분야뿐만 아니라 여태까지 IT와는 접점이 별로 없었던, 식품 제조업체와 같이 소비자 브랜드에 크게 의존하는 업체들의 수요가 증가했다. 소셜 데이터의 분석을 통해 소비자의 의견이나 목소리를 듣는 것으로 보다 고객을 잘 이해할 수 있기 때문이다"라고 덧붙였다.

마케팅 팀이 감성 분석 툴에 주목해야 하는 이유

모든 마케팅이 그러하지만 특히 소셜 마케팅이 성공하기 위해서는 다양한 소셜 플랫폼에 올라오는 사용자의 활동이나 각종 피드백 데이터를 수집한 뒤, 체계적으로 '수치화'하는 작업이 필요하다. 이 가운데서도 긍정적, 부정적 의견은 물론 심지어는 중립적인 의견을 해석할 수 있어야 한다.

그러나 개개인의 감성이나 의견은 연령이나 성별, 나이 등의 데이터와는 달리 숫자로는 쉽게 표현될 수 없는 것이다. 과거 마케터들은 오직 직관에만 의존해야 했다. 때문에 소셜 채널에 출판하는 각종 뉴스나 마케팅 자료가 사용자들

에게 얼마나 호응을 이끌어내는지, 사용자의 참여가 저조했던 이유가 과연 무엇이었는지에 대해 수치화된 이유를 제시할 수 없었으며, 따라서 소셜 마케팅에 대한 제대로 된 피드백을 받을 수 없었다.

이 뿐만 아니라 기업의 규모가 커지고 시장 규모가 방대해 질수록 소비자의 의견 각각을 추적하는 것 자체가 매우 어렵다는 문제가 있다. 특히 구글이나 애플, 삼성과 같은 거대 기업과 같은 경우 뉴스, 소셜 미디어, 블로그를 통틀어 자사명이 언급되는 정도는 하루에 수천, 수만 건이 넘어가며 이를 인력을 투입해 일일이 추적하기란 사실상 불가능에 가깝다. 이러한 방대한 작업을 처리하기 위해 수많은 기업이 소셜 데이터를 자동으로 처리하는 감성 분석 기술에 주목하며 효과적인 분석 솔루션을 차례로 제시하고 있다.

예를 들어, 어도비 시스템즈는 지난해 4월, 어도비 디지털 마케팅 컨퍼런스에서 '어도비 소셜(Adobe Social)'의 '예측 퍼블리싱(Predictive Publishing)' 기능을 발표한 바 있다. 어도비의 예측 퍼블리싱 솔루션은 페이스북과 같은 소셜 네트워크에서 사용자의 행동과 반응을 분석, 기업 콘텐츠에 대한 참여를 예측해 참여도를 높일 수 있는 최적의 타이밍을 제시하는 기술이다. 이와 같은 감성 분석 서비스나 툴을 이용하면 다음과 같은 마케팅 리서치를 효율적으로 실행할 수 있다.

- 제품 및 서비스에 대한 사용자의 의견이나 평가, 별점 조사
- 업체에 대한 부정적인 의견이나 이슈의 실시간 모니터링
- 시장 현황이나 경쟁 업체의 활동, 소비자 트랜드 추적
- 업체의 활동이나 관련 이슈에 대한 대중의 반응 측정

특히, 웹 전체가 자사에 대해 어떤 말(긍정/부정)을 하고 있는지 실시간으로 추적이 가능하다는 점에서 감성 분석은 소비자 모니터링 과정에서 빠른 대응을 가능하게 하는 전략적인 툴이 될 것으로 보인다. 일례로 스페인의 민간 금융 업체인 BBVA(Banco Bilbao Vizcaya Argentaria, S.A.)는 IBM의 감성 분석 서비스인 IBM 소셜 애널리틱스(Social Analytics)를 도입해 브랜드 이미지 및 고객 관리 분야에서 큰 성과를 이룬 바 있다.

BBVA는 감성 분석 기술을 통해 현재 온라인에서 가장 중요한 주제가 무엇인지 파악할 수 있게 되었을 뿐만 아니라 이에 대한 대응방안을 즉각적으로 계획할 수 있어 고객의 불만사항을 신속하게 해결할 수 있었다. 실제로 IBM의 소셜 애널리틱스를 도입한 결과, 2011년 상반기에 BBVA는 자사에 대한 긍정적인 피

드백이 전년도 대비 1% 이상 증가했으며 부정적인 피드백은 1.5% 이상 감소했다는 사실을 발견했다.

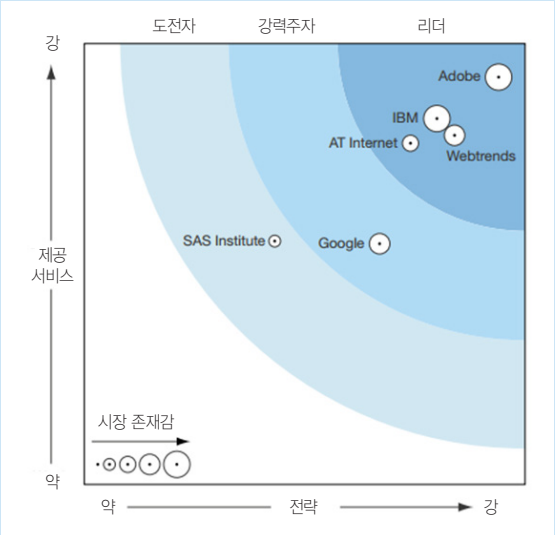
SK플래닛의 경우, 2007년 OMS(Opinion Mining System) 개발을 시작으로 2010년에는 OMS 2.0 시스템을 완성했고, 이를 바탕으로 SK텔레콤은 B2B를 대상으로 'Smart Insight'라는 서비스를 제공하고 있다.

SK플래닛은 여기서 더 나아가 하둡(Hadoop) 기반의 병렬/분산 처리 환경에서 비정형 빅데이터를 대상으로 이슈 키워드, 연관키워드 분석과 같은 일반적인 텍스트 마이닝은 물론, 오피니언 마이닝 플랫폼(Opinion Mining Platform)을 구축하는 작업을 추진해 왔다. 2011년부터는 한국어와 영어 자연언어 처리 기술을 포함해 OMS를 구성하는 핵심기반 기술을 전면적으로 개선시키는 등 다양

Top 6 웹 분석 서비스 업체: 포레스터 리서치

2014년 포레스터 웨이브 웹 애널리틱스(Forrester Wave: Web Analytics) 보고서는 총 6개의 주요 웹 분석 서비스를 소개했다. 이들 6개 업체들은 텍스트마이닝, 감성 분석 기술을 포함한 웹 분석 서비스를 제공하고 있다.

업계 선두를 달리는 '리더(leader)' 업체로는 어도비, IBM, 웹트렌즈(Webtrends), AT 인터넷(AT Internet)이 선정됐으며 그 뒤를 따르는 '강력 주자(Strong Performer)'로는 구글이 지목됐다. 마지막으로 웹 분석 시장에 새로 뛰어든 업체들 가운데 두각을 나타내는 '도전자(Contender)'로는 SAS(SAS Institute)가 소개됐다.



출처 : The Forrester Wave™: Web Analytics,Q2 2014

업체	평가 제품	제품 버전(출시일)
어도비(Adobe)	어도비 애널리틱스 프리미엄(Adobe Analytics Premium)	2014 봄
AT 인터넷 (AT Internet)	애널라이저 III(Analyzer III)	2012. 10
구글(Google)	구글 애널리틱스 프리미엄(Google Analytics Premium)	상시 업데이트 중
IBM	IBM 디지털 애널리틱스(IBM Digital Analytics)	버전 14(2014. 1)
SAS	어드반스드 커스토머 익스피리언스(Advanced Customer Experience)	버전 6.1(2013. 4)
웹트렌즈(Webtrends)	애널리틱스 온 디맨드(Analytics On Demand) 세그먼트(Segments) 스트림즈(Streams) 액션 센터(Action Center) 웹트렌즈 익스플로러(Webtrends Explorer)	2013. 10 2013. 10 2013. 4 2013. 7 2014. 1

출처 : The Forrester Wave™: Web Analytics,Q2 2014

한 기술을 개발하고 있다.

감성 분석의 3단계

감성 분석은 총 3단계로 이뤄진다. 첫 번째는 각종 소셜 미디어 매체에서 정보를 수집하는 '데이터 수집' 단계다. 두 번째는 이렇게 총체적으로 수집된 정보에서 사용자의 주관이 드러난 부분만을 걸러내는 '주관성 탐지' 과정이다. 마지막 세 번째 단계에서는 '극성 탐지' 작업이 이뤄지는데, 이는 추출한 감성 데이터를 '좋음'과 '싫음'의 양 극단으로 분류하는 과정이다.

1. 데이터 수집(Data Collection)

인터넷 상에 있는 방대한 양의 UCC 데이터는 감성 분석이 진정한 빛을 발할 수 있게 하는 핵심 요소다. 각종 블로그나 게시판, 제품 평가란 등 공개적인 데이터 소스뿐만 아니라 트위터나 페이스북과 같은 개인적인 소셜 네트워크 사이트에서도 데이터를 끌어올 수 있다.

감성 분석 기술을 적용할 데이터를 수집하기 위해서는 보통 검색 엔진을 활용한다. 검색 엔진은 사용자의 질의어를 입력받아 질의어가 포함된 모든 문서, 즉 각종 별점이나 리뷰, 코멘트를 포함한 관련 데이터를 수집한다. 예를 들어 SK 플래닛의 오피니언 마이닝 플랫폼과 같은 경우, 특정 사이트만을 대상으로 하기보다는 제휴 정보를 포함해 뉴스, 블로그, 카페, 전문 커뮤니티, SNS 등 채널을 다양화해 국내 외 수집 가능한 사이트들로부터 데이터를 수집한다.

그러나 이렇게 총체적으로 수집된 데이터는 사용자의 '감성'과는 관련이 없는 부분도 포함하고 있기에 데이터 수집을 마친 뒤에는 감성 분석을 적용할 부분만을 추출하는 '주관성 탐지' 작업이 필요하다.

2. 주관성 탐지(Subjectivity Detection)

필요한 텍스트를 수집하고 난 뒤에는 감성 분석에 사용될 텍스트 요소만을 분리, 분류하는 작업이 필요하다. 일반적으로 웹에서 수집된 텍스트는 문장 내에서 '감성'과는 관련이 없다고 판단되는, 주관성이 없는 부분도 제외시킨다. 또한, 지나친 정보 수집에서 따르는 문제를 피하기 위해 텍스트 저자의 이름, 성별과 같은 개인 정보를 걸러내는 과정을 거친다.

예를 들어, 코난테크놀로지의 '텍스트 애널리틱스(Text Analytics)'는 구조화되어 있지 않은 대량의 텍스트에서 의미있는 정보를 추출하는 정보분석기술이다. 텍스트 애널리틱스는 주어진 텍스트의 구성 요소들을 '긍정', '부정', '중립', '객관'이라는 4가지 분야로 구분한다. '오늘 스마트폰을 새로 샀다'와 같이 사실만을 진술하는 문장은 가치판단이 전혀 개입되지 않은 '객관'으로 분류돼 분석 대상에서 제외되는 것이다.

3. 극성 탐지(Polarity Detection)

극성 탐지 단계에서는 주어진 데이터가 '긍정'인지, 혹은 '부정'인지를 판단하는 '극성 분석(Polarity Detection)' 작업이 이뤄진다. 컴퓨터는 텍스트 안에 있는 긍정적, 부정적인 단어를 탐지, 이를 정량화 한 뒤 통계적 기법을 적용한다. 예를 들어, 문서에서 각 단어가 나타나는 '빈도'나, 긍정이나 부정과 같은 '속성'에 따라 점수나 가중치를 부여한 뒤, 각 단어가 나타내는 점수의 총합이나 평균을 구해 전체 텍스트가 과연 긍정적인지 혹은 부정적인지 알아내는 것이다.

감성 분석에는 크게 '문서' 단위의 극성 분석, '속성' 단위의 극성 분석, 그리고 감성어 사전(lexicon)을 이용한 분석 방법이 있다. 문서 단위의 극성 분석은 주로 기계 학습을 이용하는데, 여기에는 '훈련'과 '분류'의 두 과정이 포함된다.

기계 학습이란 인공지능의 한 분야로, 주어진 데이터에 존재하는 특정한 패턴을 공식화 한 뒤 이를 바탕으로 컴퓨터가 다른 유사한 데이터를 해석할 수 있도록 학습시키는 것이다. 인간의 학습 절차를 모방한 모델을 통해 컴퓨터는 새로운 데이터를 해석하거나 기상현상이나 주가의 상승/하락 여부와 같은 가까운 미래를 예측할 수 있다.

기계 학습의 훈련 단계에서 사용자는 수작업으로 긍정 혹은 부정으로 분류된 문서에서 특정 값을 추출해 '훈련 데이터(Training Data)'를 생성한다. 컴퓨터는 그 다음, 일정한 학습 모델을 통해 문서 전체가 긍정인지 부정인지를 '분류(classify)'하게 된다.

그러나 문서 전체의 긍정/부정 여부에만 집중하게 되면 여러 문제가 생길 수 있다. 다음과 같은 노트북 리뷰가 있다고 가정해보자.

(1) 이 노트북은 최첨단 프로세서와 메모리를 갖추고 있어 뛰어난 성능을 자랑한다.
(2) 스크린도 크고 선명해 오랫동안 사용해도 눈이 피로하지 않다는 장점이 있다.
(3) 또한, 태블릿 PC로서 스크린의 터치감도 이전 모델보다 확연히 좋아졌다.
(4) 그러나 이 노트북은 결정적으로 지나치게 무거워 구매가 망설여진다.

문장 (4)를 제외한 문장 (1)~(3)번은 모두 노트북에 대해 긍정적으로 말하고 있다. 그러나 우리는 앞의 긍정적인 세 문장에도 불구하고 마지막 문장 하나때문에 이 글이 호평과는 거리가 멀다는 것을 알고 있다.

그러나 컴퓨터는 왜 이 글이 결과적으로 부정적인지 알지 못한다. 컴퓨터는 단순히 해당 텍스트가 '긍정(+)' 3개와 '부정(–)' 1개로 이루어져 있다고 분석한 뒤, 전체적으로 봤을 때 '긍정' 2개로 이루어진 '긍정적인 텍스트'라고 판단할 뿐이다. 즉, 단순히 긍정/부정 단어가 몇 차례 나왔는지 기계적으로 횟수를 세는 것만으로는 텍스트 전체의 의미를 잘못 해석할 가능성이 있는 것이다.

이러한 문제를 해결하기 위해서는 문서 전체의 긍정/부정 여부가 아니라 문장 단위로 텍스트를 분석하는 '속성' 단위의 극성 분석이 필요하다. 여기서 '속성'이란, 노트북의 성능, 스크린, 무게와 같은 대상의 여러 특징을 의미한다. 속성 단

위의 극성 분석을 하기 위해서는 우선 대상이 어떤 속성으로 구성돼 있는지, '속성명'을 추출한 다음, 각각의 속성에 따른 감성어를 찾은 뒤 그 감성어가 긍정에 해당하는지, 부정에 해당하는지 분류하는 과정이 필요하다.

이제 기업들은 단순히 '좋아요'가 몇 개인지 아는 것에서 나아가 '무엇을' '어떤 이유로' 좋아하고 있는지도 알고 싶어 한다. 예를 들어 '이 스마트폰은 카메라 기능이 좋긴 하지만 너무 무겁다'라는 텍스트의 경우, 문서 자체의 긍정/부정 여부를 따지는 것이 아니라 문장 단위로 들어가 사용자가 구체적으로 스마트폰의 어떤 점을 좋아하고 싫어하는지에 대한 분석을 할 수 있다.

코난테크놀로지의 김문희 부장은 "속성 분석을 하게 되면 해당 스마트폰의 장점은 '카메라 기능'이고, 단점은 '무게'라는 것을 분석할 수 있다. 우리는 보다 심층적인 자료를 요구하는 고객에게 이와 같은 기술을 제공하고 있다"고 말했다.

이처럼 자동 감성 분석 과정에서는 긍정, 부정의 어휘를 분류한 후 문서 내에서 이런 극성 어휘들의 빈도를 연산하는 방법이 주를 이뤄 왔다. 그러나 컴퓨터가 포괄적으로 감성 표현을 학습하기 위해서는 실제 감성 표현들이 주석된 대량의 '코퍼스(corpus)', 즉 감정어 사전이 구축되어야 한다. 영어의 경우, 약 만 개의 문장 안에 나타나는 감성 표현들로 구축된 사전, MPQA(Multi-perspective Question Answering)가 있다.

서울대학교 언어학과 신효필 교수를 포함한 서울대학교 연구진은 MPQA와 같은 역할을 하는 한국어 감성 코퍼스, KOSAC(Korean Sentiment Analysis Corpus)을 2011년부터 2013년간 2년에 걸쳐 구축했는데, 여기에는 332개 신문기사, 7,744개 문장을 주석 대상으로 삼아 총 만 7,582개의 감성 표현을 수록돼 있어 한국어 감성 분석 연구에 중요한 역할을 하고 있다. 감성어 사전을 구축하기 위해서는 컴퓨터가 동의어나 반의어를 자동적으로 수집하는 방법도 있으나, 정확성을 높이기 위해서는 사람이 직접 단어를 선별하는 과정이 필요하다. 여기서 중요한 것은, 사전이 얼마나 많은 감성어를 포함하고 있는지에 따라 분석의 질이 달라질 수 있다는 점이다.

이와 같은 감성의 극성 분석을 이해하기 위해서는 컴퓨터가 사용하는 기본적인 언어 모형과 기계학습 알고리즘을 이해할 필요가 있다.

감성 분석 기술의 현주소: 자연어 처리의 한계

감성 분석 기술은 컴퓨터에게 인간의 언어 자체를 이해시키는 과정이므로로 단순한 '지식의 축적'과는 거리가 있다. 미국의 퀴즈쇼 '제퍼디(Jeopardy)'에서 IBM의 슈퍼컴퓨터 '왓슨(Watson)'이 인간 경쟁자들을 제치고 두 배가 넘는 점수차이로 우승한 것과는 다른 문제라고 할 수 있다. 만약 왓슨에게 토론 대회에서 각 참가자의 찬반여부 및 그 내용을 분석하라고 한다면 퀴즈쇼에서 정답을 말하는 것처럼 쉽지는 않을 것이다.

신효필 교수는 "감성 분석과 그 기반이 되는 자연어 처

리 연구는 현재 문서 단위 분석, 속성 분석에서 넘어가 이제는 '담화 분석(Discourse Analysis)' 범위까지 확대되고 있다"고 말했다. 담화 분석이란, 여러 문장 사이의 연과 관계와 문맥을 고려해 문장들의 의미관계를 분석하는 기술이다. 예를 들어, 문장에서 '이것, 그것, 저것'이 가리키는 것이 구체적으로 무엇인지, '그, 그녀, 그들'과 같은 대용어까지 파악하는 것이다.

한편, 영어와 비교했을 때 한국어 감성 분석 기술과 자연어 처리 연구는 다

감성 분석에 사용되는 언어 모형 및 알고리즘

컴퓨터가 사용하는 가장 대표적인 언어 모형이 바로 '좋다(good)', '나쁘다(bad)'와 같이 하나의 단어를 다루는 '유니그램(Unigram)'이다. 유니그램 모형은 어떤 단어가 모여 어떤 문장을 이룰 지에 대한 확률을 측정하는 방법이다. 이 모델을 사용하면 각 단어가 사용되는 빈도를 분석할 수 있으며, 감성어의 등장 횟수에 기반해 통계적인 수치를 제시할 수 있다.

유니그램은 통계적으로 정확한 모델을 제공하며 직관적인 해석이 가능하기에 비정형 데이터를 가장 간단하게 수치화할 수 있다는 장점이 있다. 그러나 단순히 단어의 확률 분포만을 측정하는 유니그램은 단어의 순서와 같은 문법 사항을 고려하지 않는다. 예를 들어, 유니그램만으로는 '좋지 않다'와 같이 부정어로 인해 의미가 반대가 되는 경우나 '매우 좋다'처럼 감성의 강도가 더 센, 두 단어가 하나의 의미를 이루는 감성어를 해석할 수 없다.

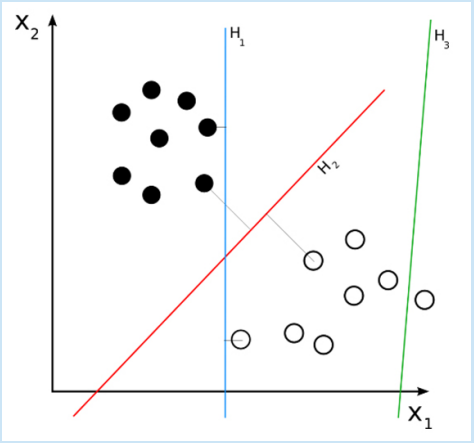
이와 같은 유니그램의 한계를 보완하는 것이 바로 '바이그램(Bigram)'인데, 이는 부정어나 강조 부사와 같이 두 단어로 이뤄진 표현을 인식하는 모형이다. 2개의 단어를 대상으로 하는 바이그램에서 더 나아가면 이제 문장 내에 있는 n개의 연속적인 단어의 순서와 상관관계를 파악하는 N-그램 모형이 있다.

감성 데이터를 '좋음'과 '싫음'의 양 극단으로 나누는 분류 방법에는 '나이브 베이즈(Naïve Bayes)'와 '서포트 벡터 머신(Support Vector Machine, SVM)'과 같은 기계학습 알고리즘이 있다. 나이브 베이즈는 감성 분석에 사용되는 가장 기본적인 분류 알고리즘으로, 주어진 텍스트가 각각 긍정과 부정 항목에 포함될 확률을 구하는 방법이다. 나이브 베이즈 분류 방법은 텍스트를 순서가 없는 단순한 단어의 집합으로 가정한다. 이를 'Bag-of-Words(BoW)' 방식이라고 한다. '가방'에 단어 조각들이 무작위적으로 있다는 뜻이다.

단순한 문서 카테고리 분류와 같은 작업의 경우, 나이브 베이즈는 80%~90%까지, 상당한 정확도를 보일 수도 있다. 그러나 감성 분석은 단어의 문법적인 연관성을 정확하게 판별할 수 있어야 제대로 된 분석을 할 수 있다. 그렇기 때문에 단순한 BoW 방식을 넘어서 속성 분석까지로 분석 영역이 확대되고 있다.

한편, SVM은 두 개의 다른 항목에 속해있는 점들을 분류하는 수많은 평면들 가운데 최대한 두 클래스의 점들과 동일한 거리를 유지하는 평면을 찾아내는 알고리즘이라 할 수 있다. 즉, 감성 분석에 있어 SVM은 즉 얼마나 긍정과 부정의 '경계'를 잘 설정하는가'의 문제다. 이 경계를 더 정확히 설정할수록 컴퓨터는 긍정적인 텍스트와 부정적인 텍스트를 분류할 수 있다.

그림 2 | 키-값 쌍 데이터베이스



출처 : Wikipedia

소 더디게 진행되고 있다. 이는 바로 한국어라는 언어 자체의 복잡성 때문이다.

신효필 교수는 "감성 분석이라는 기술 자체도 어려운 일이지만 특히 영어에 비해 한국어는 문제가 더 복잡하다. 우선 주어, 동사, 목적어 등 어순이 비교적 고정돼 있는 영어와 달리 한국어는 어순이 자유로워 구문의 구조와 의미를 분석하는 작업이 더욱 복잡하다. 뿐만 아니라 한국어는 주어나 명사의 생략이 많아 모호한 문장이 많으며 매우 복잡한 형태소 구조로 이뤄져 있다"고 설명했다.

즉, 영어에서의 형태소 분석의 경우, 하나의 단어가 하나의 형태소에 해당하므로 단어를 분리할 일이 없다. 그러나 한국어의 경우 각 단어에서 형태소를 분리한 후, 그 각각의 품사까지 결정해야 하기 때문에 분석 과정이 훨씬 더 복잡해진다. 이에 대해 신효필 교수는 "보통 자연어 처리는 '형태소 분석', '구문 분석', 그리고 '의미 분석' 순으로 이뤄진다. 현재 영어에서는 구문 분석까지 진도가 많이 나간 상태다. 이에 비해 한국어는 아직도 형태소 분석 과정에서 많은 어려움을 겪고 있다"고 말했다.

인공 지능 연구의 최전방

그러나 이와 같은 어려움에도 불구하고 감성 분석, 특히 여기에 필요한 한국어 자연어 처리 기술에 대한 연구는 소프트웨어 분야에서 국가 혁신 기술 개발에도 직결되는 문제이기 때문에 정부 차원에서도 활발하게 지원되고 있다. 미래창조과학부는 2013년 5월, '엑소브레인(Exobrain)'이라는 한국 인공지능 컴퓨터 개발계획에 총 1,070억 원의 대규모 투자에 나섰다.

'몸 밖에 있는 뇌'라는 뜻의 엑소브레인 프로젝트는 컴퓨터가 인간과 의사소통을 하고 인간의 감성을 이해하기 위해서는 자연어 처리에 대한 연구가 중요하다고 보고 있다. 자동통역 인공지능 연구센터장이자 엑소브레인 과제 총괄책임자인 ETRI의 박상규 박사는 "엑소브레인 프로젝트는 자연어를 이해하여 지식을 자가 학습하며 전문 직종에 취업 가능 수준의 인간과 기계의 지식소통이 가능한, 지식과 지능이 진화하는 소프트웨어 개발을 목표로 하고 있다"고 설명했다.

연구기간은 3단계에 걸쳐 총 10년 동안 진행될 예정이며, 2017년까지 진행될 1단계 연구에 투입되는 연구비는 428억 원이다. 이는 소프트웨어 분야에서 이정표가 될 만한 대형 프로젝트로, ETRI, 솔트룩스, KAIST, 포항공대 등 연간 26개의 연구기관 366명의 연구원이 참여하게 된다.

프로젝트 1단계가 종료될 2017년에는 IBM의 왓슨을 따라잡고 2단계부터는 컴퓨터가 스스로 지식을 학습해 세계 최고 수준의 지능 진화형 기술을 선보인다는 목표다. 특히, 컴퓨터가 자연어의 의미를 이해하고 학습하여, 전문가 수준의 의사소통이 가능한 '지능진화형 질의응답(Wise Question Answering)' 소프트웨어 개발이 연구 1단계에서 진행할 예정이며, 구체적인 연구 개요는 다음과 같다.

– 통계적 언어지식 및 언어 지식베이스 연계 자연어 의미분석을 위한 자연어 심층이해 기술
– 자연어의 진화 및 특정 도메인에 대한 언어지식 학습을 위한 지속적 언어지식 학습 및 추출 기술
– 질문 도메인/유형/정답유형/난이도를 고려한 자연어 질문분석 및 이해 기술 개발

이처럼 감성 분석과 그 기반이 되는 자연어 처리 기술은 차세대 선거 전략뿐만 아니라 기업의 이윤 창출을 배가할 효과적인 마케팅 수단으로, 나아가 인공지능 연구의 일환으로까지 각광받고 있다.

사람들이 '스테이크'에 대해 인터넷에 적어놓은 모든 글을 분석할 수 있다고 가정해보자. 컴퓨터는 몇 명의 사람들이 어떤 굽기의 스테이크를 얼마나 먹었는지에 대한 단순한 수치만을 제시할 뿐이다. 예를 들어 레어, 미디엄, 웰던을 먹은 사람들의 수가 각각 100명이라는 사실은 쉽게 알 수 있다. 그러나 그들이 '왜' 특정 굽기를 선호했는지는 사람들의 말을 실제로 이해할 수 있어야만 가능하다. '왜'야말로 사람들이 가장 궁금해하는 것이다. 감성 분석은 이 육하원칙 가운데 가장 중요한 마지막 항목에 대한 고찰을 가능하게 하며 이는 궁극적으로는 인간의 '보편적인 사고'에 대한 해답을 제공해 줄 수 있는 기술이 될 것이다. ITWORLD