엘라스틱 넷을 적용한 블로그 리뷰 감성사전 구축 및 극성 분류

김 승 범°, 권 수 정, 김 준 태
동국대학교 컴퓨터공학과
seungbum@dongguk.edu, sujeong_k@dongguk.edu, jkim@dongguk.edu

Building Sentiment Dictionary and Polarity Classification of Blog Review By Using Elastic Net

SeungBum Kim°, Sujeong Kwon, Juntae Kim
Department of Computer Science and Engineering, Dongguk University

요 약

오피니언 마이닝(Opinion Mining)은 소셜 데이터에서 특정 대상에 대해 느끼는 생각과 태도, 그리고 나아가 그 이유를 판별하여 의미 있는 정보로 변환하고 이를 의사결정에 활용하는 자연어처리 기술이다. 이러한 오피니언 마이닝을 다루는 연구들은 단어별로 긍/부정의 극성을 정해놓은 범용적인 감성사전을 사용하여 텍스트에 나타난 어휘들의 극성 분포에 따라 극성을 분류하는 방식을 사용한다. 하지만 같은 단어라고 하더라도 해당 단어의 극성이 한 가지로 고정되지 않고, 해석하는 사람에 따라 혹은 분석하는 의도에 따라 그 극성이 다르게 나타날 수 있다. 또한, 같은 긍정 혹은 부정을 의미하는 단어라 하더라도 해당 단어가 의미하는 긍정 혹은 부정의 강도가 다르다. 따라서 본 논문에서는 소셜 데이터를 수집하고, 회귀분석(Regression Analysis)의 방법인 엘라스틱넷(ElasticNet; Regularized Regression Method)을 사용하여 각 단어의 회귀계수를 구해 단어들의 가중치를 측정해 중요한 단어만 추출하는 방법을 통하여 감성사전 구축에 적용했다. 기존의 감성사전 구축으로 극성을 분류한 방법과 본 논문에서 제안하는 감성사전 구축으로 극성을 분류하는 방법을 비교한 결과, 본 논문에서 제안하는 방법이 더 우수한 정확도를 보였다.

1. 서 론

최근 무선 인터넷의 발달과 스마트 폰의 보급으로 시간, 장소에 구애받지 않고 서로 소통할 수 있는 소셜 미디어 서비스가 빠르게 성장하며 소셜 미디어를 통하여 생성되는 데이터의 양은 급속도로 증가하고 있다. 소셜 미디어의 데이터는 설문조사나 인터뷰처럼 인위적인 환경에서 생산된 것이 아닌, 자발적으로 표현한'날 것 그대로'의 데이터라는 점에서 사람들의 의견을 가장 잘 파악할 수 있는 분석 지표로 주목받고 있다. 이런 소셜 데이터에서 나타난 사람들의 생각과 태도, 성향과 같은 주관적인 데이터를 분석하는 방법을 오피니언 마이닝이라고 한다. 이러한 오피니언 마이닝을 하기 위해선 텍스트에서 형태소 분석을 통해 추출한 감정 서술어의 감정 방향을 비교하기 위해 필수적인 역할을 하기 때문에 감성사전은 매우 중요한 요소이다.

본 논문에서는 회귀분석의 방법인 엘라스틱넷을 사용하여 소셜 데이터를 수집하고, 각 단어들의 회귀계수를 구해 단어들의 가중치를 측정하여 중요한 단어만 추출하는 방법을 통하여 감성사전 구축에 적용을 하였다.

본 논문의 2장에서는 감성사전을 구축하여 사용한 관련 연구를 정리하고, 3장과 4장에서는 본 논문에서 제안하는 엘라스틱넷의 배경이 되는 이론들에 관해서 기술하고 이를 통해 감성사전을 구축하는 방법에 관해 설명한다. 5장에서는 실험의 내용과 결과를, 6장에서는 결론을 기술한다.

2. 관련 연구

오피니언 마이닝은 소셜 데이터에서 특정 대상에 대해 느끼는 생각과 태도, 그리고 나아가 그 이유를 판별하여 의미 있는 정보로 변환하고 이를 의사결정에 활용하는 자연어처리 기술이다. 오피니언 마이닝에서 감성사전은 텍스트에서 형태소 분석을 통해 추출한 감정 서술어의 감정 방향을 비교하기 위해 필수적인 역할을 한다. 이러한 감성사전 구축과 관련된 연구들은 다음과 같다.

안정국 등은 국립국어원의 사전을 기반으로 개인적인 선입견이 들어가지 않는 범용적인 감성사전을 구축하였다[1]. 이종혁 등은 희로애락을 기준으로 기쁨, 분노, 슬픔, 즐거움의 동의어로 감성사전을 구축하였다[2]. 유은지 등은 상품 평을 기반으로 영화나 의류, 모바일 등에서 사용자의 감정을 분석할 수 있는 감성사전을 자동으로 구축하는 시스템을 구현하였다[3]. 송종석 등은 주가지수의 상승이라는 한정된 주제에 대해 각 어휘가 갖는 극성을 판별하여 주가지수 상승예측을 위한 감성사전을 구축하였다[4]. 서정렬 등은 PMI를 기반으로 SentiWordNet과 비교하여 PMI 값과 SentiWordNet의 값이 긍정으로 일치하면 긍정어로, 부정으로 일치하면 부정어로 저장하는 방식으로 사전을 구축하였다[5].

본 논문에서는 회귀분석의 방법인 엘라스틱넷을 사용하여 각 단어의 회귀계수를 구해 단어들의 가중치를 측정하는 방식으로 감성사전을 구축하고, 극성을 분류하여 정확도를 측정하였다.

3. 회귀분석(Regression Analysis)

회귀분석은 한 변수가 다른 변수에 대해 미치는 영향을 추정할 수 있는 통계기법을 말한다. 이러한 회귀분석 방법에는 선형 회귀분석과 능형 회귀분석, 라쏘 회귀분석, 엘라스틱넷 등이 있다.

·선형(Linear)회귀분석

선형회귀분석은 최소자승법(Method of Least Squares)

을 사용하여 회귀계수(Coefficient)의 추정량을 구한다. 하지만 최소자승법은 설명 변수의 개수가 늘어나게 되면 어떤 설명 변수가 회귀계수의 추정량을 구할 때 큰 영향을 미치는지에 대한 해석이 어려워지고, 공분산 행렬의 행렬식이 0에 가까운 값이 되어 회귀 계수의 추정량이 매우 나빠지게 되는 다중공선성(Multicollinearity)의 문제가 발생하여 안 좋은 추정량을 갖게 된다[6].

·라쏘(LASSO)회귀분석

라쏘 회귀분석은 λ_1제약 조건을 사용하여 영향력이 없는 변수의 회귀계수를 0으로 만들어 예측에 필요한 중요한 변수만을 선택해 차원을 축소하고, 변수 선택이 가능해 예측모형의 해석력을 증가시켜준다. 하지만 라쏘 회귀분석은 설명변수 간에 높은 상관관계가 존재할 때는 다중공선성의 문제가 발생한다[7].

·능형(Ridge)회귀분석

능형 회귀분석은 λ_2제약 조건을 이용하여 회귀계수의 크기를 축소함으로써 다중공선성의 문제를 해결하고, 예측정확도를 높인 방법이다. 하지만 능형 회귀분석은 차원축소가 불가능하므로 때문에 모든 변수가 모형에 포함되어 모형에 대한 해석력이 떨어진다는 단점이 존재한다[8].

·엘라스틱넷(ElasticNet)

엘라스틱넷은 식(1)처럼 라쏘 회귀분석의 λ_1제약조건과 능형 회귀분석의 λ_2제약조건을 결합한 방법으로, 식(2)에서 $\alpha=0$이면 능형 회귀분석, $\alpha=1$이면 라쏘 회귀분석과 같으므로, 엘라스틱넷은 0에서 1 사이의 α에 해당하며, 라쏘 회귀분석과 능형 회귀분석의 볼록 결합(Convex Combination)에 해당하기 때문에 능형 회귀분석과 라쏘 회귀분석의 비율을 조정할 수 있다.

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{P}\beta_i x_{ij})^2+\lambda_1\sum_{j=1}^{P}|\beta_j|+\lambda_2\sum_{j=1}^{P}|\beta_j|^2 \quad (1)$$

$$(1-\alpha)\sum_{j=1}^{P}|\beta_j|+\alpha\sum_{j=1}^{P}|\beta_j|^2 \le t \quad (2)$$

따라서 라쏘 회귀분석의 영향력이 없는 변수의 회귀계수를 0으로 만들어 차원을 축소해 변수 선택이 가능한 장점과 능형 회귀분석의 전체적인 회귀 계수의 크기를 축소함으로써 관련성이 높은 설명 변수가 있을 때 변수들을 그룹화하여 다중공선성의 문제를 해결한 장점 두 가지를 동시에 만족하는 방법이다[9].

따라서 본 논문에서는 엘라스틱넷을 사용하여 사전을 구축하여 각 단어의 회귀계수를 구해 극성을 분류하는데 필요 없는 단어들은 제외하고, 중요한 단어들만을 선별하여 사전을 구축하였다.

4. 감성사전 구축 및 극성 분류

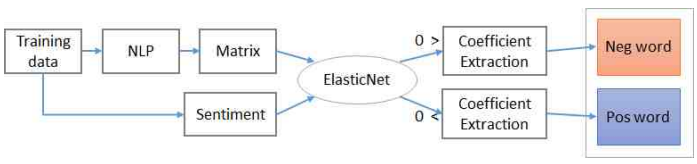

그림 1 감성사전 구축 모형

오피니언 마이닝에서 감성사전의 중요성은 매우 크다. 텍스트에서 형태소 분석을 통해 추출한 감정 서술어의 감정 방향을 비교하기 위해 필수적인 역할을 한다.

본 논문에서는 감성사전을 회귀분석의 방법인 엘라스틱넷을 사용하여 구축한다. 그림 1은 감성사전을 구축하는 방식에 대한 모형이다.

우선 감성 사전 구축을 위해 수집한 블로그 리뷰 데이터를 말뭉치(Corpus) 단위로 분리하여 저장한 후 구두점과 숫자, 불용어(stopword)를 제거하는 등의 자연어 처리를 한다. 자연어 처리를 한 단어들을 문서x단어의 행렬(Matrix)로 만든 후 희소행렬로 변환한다. 엘라스틱넷을 사용하기 위한 독립변수 X와 설명변수 Y는 각각 문서x단어의 행렬과 텍스트의 극성을 의미하는 감정값(Sentiment)으로 한다. 엘라스틱넷을 수행할 때 10-묶음 교차 검증법(10-fold Cross Validation)을 통하여 오차가 가장 작을 때의 최소λ를 추출하여 해당 λ일 때의 단어와 각 단어의 회귀계수를 저장하여 0보다 작은 값을 가지는 단어는 부정어로, 0보다 큰 값을 가지는 단어는 긍정어로 하여 사전을 구축한다.

본 논문에서는 이렇게 만들어진 사전의 단어와 회귀계수를 사용하여 각 단어의 빈도수뿐 아니라 단어가 의미하는 긍정 혹은 부정의 강도까지 고려해 극성을 분류하고, 정확도를 측정해 단순히 단어의 출현 빈도수로 극성을 분류하는 방법과의 비교 보았다.

5. 실험

5.1 실험 데이터

keyword	page	contents	sentiment
음식점 맛있는	1	페북에서만 보던 비쥬얼 드디어 만나보았음 인천사는 사몽이가 추천해준 맛집 코	1
음식점 맛있는	1	홍천식당 음식점 시동골에서는 빙수 비쥬얼 콩국수와 왕돈까스까지 가능하답니다	1
음식점 맛있는	1	지난 주말- 오랫만에 강남역에 갔어요 ㅎㅎ 저녁 약속이 있었던지라 사람 많은 토	1
음식점 맛있는	1	대학 칭구들과 이태원 프라하에서 망년회를 하게되었어요^^ 처음가본 체코음식	1
음식점 맛있는	1	이수역 맛집 추천 - 뽕잎사랑 샤브칼국수 퓔빙 음식점!충신대입구역 (이수역) 13번	1
음식점 맛있는	2	안녕하세요 우리 잇님들제가 보기보단 음이 참 부실해졌나봐요.아니면 아들 둘 낳	1
음식점 맛있는	2	테이스티로드 맛집 환상 비쥬얼 뽐내는 고인돌갈비얼마 전에 테이스티로드 맛집	1
음식점 맛있는	2	와우~~~독특한 제주현지인맛집 찾았어요~~~^^지인이 가보고 맛있다고 알려줘	1
음식점 맛있는	2	삼성역 맛집 착한음식점 가격대비최고 남해안식당 후기삼성역 3번 출구에서 가까	1
음식점 맛있는	2	이태원 태국음식점 :) 스파이스마켓이태원은 정말 어느곳 못지 않게 맛집이 많은	1
음식점 맛있는	3	안녕하세요 친구가 분평동 맛집이라고 귀가 따갑게 말했던 왕비해물문어보쌈!드	1
음식점 불만족	3	홍익돈까스 솔플 수성구 로스까스 약간불만족 음식 때깔 좋아보이게하려고그런지	0
음식점 불만족	1	오호~ 오늘은 오랜만에안추천 음식점을 들고 왔네염~~안추천 음식점을 먹게된	0
음식점 불만족	1	최근에 음식점 포스팅을 안하고 있다.현찬도 안받고 있을 뿐더러.. 외식도 별로 안	0
음식점 불만족	1	원래는 쭈꾸미달인 가고싶었는데문이 닫혀있었던 관계로 ㅠ^ㅠ(왜 우리가 찾아	0
음식점 불만족	1	2015년 6월 23일 그동안 음식점 리뷰도 좋고 해서 꼭 한번 가보고 싶었던 어다리	0
음식점 불만족	1	사실 여기는 맛집이 아니라 비추하는 ㅋㅋㅋㅋㅋㅋ개망 식당.앙코르톰의 코끼리테	0
음식점 불만족	1	[얼큰이's 신혼여행 : 하이난(海南) 싼야(三)] #006 _ 명성에 비해 불만족스러웠던	0
음식점 불만족	2	젤렌 실망스러운 이태원 불가리아음식점오랜만이에요^-^ 블태기가 와서 한동안	0
음식점 불만족	2	식구 밥상점심이 정말 부실하게 나왔던 날.언니와 저는 미련없이 회사식당을 나와	0
음식점 불만족	2	늘봄흙돼지가 제주도 맛집이라고 소문이 자자했다. 제주도 노형동에 위치한 이 곳.	0
음식점 불만족	3	누가 그러던데 ㅋㅋ 전주에는 아무 음식점이나 들어가도 맛있다고..슬직히 그건 뻥	0

그림 2 블로그 크롤링 데이터

본 절에서는 음식점을 다녀온 후 블로그에 남긴 리뷰 2만 건을 대상으로 실험을 수행한다. 블로그 리뷰 2만 건 중 1만 건을 훈련데이터로, 1천 건을 검증 데이터로 임의 추출하여 사용하였다. 감정값(Sentiment)은 긍정어로 크롤링한 문서는 긍정으로 간주하여 1로, 부정어로 크롤링한 문서는 부정으로 간주하여 0으로 분류하였다.

형태소 분석은 R 언어에서 지원하는 카이스트 SWRC연구소에서 개발한 한 나눔 형태소 분석기를 사용하였고, 엘라스틱넷 또한 R로 구현하였다. 감성사전 구축에 사용된 총 1만 건의 문서 중 총 107,512개의 단어를 찾을 수 있었고, 총 문서에서 10번 이상 나오는 단어만을 추려, 2,822개의 단어를 추출해 사용하였다.

5.2 실험 모형

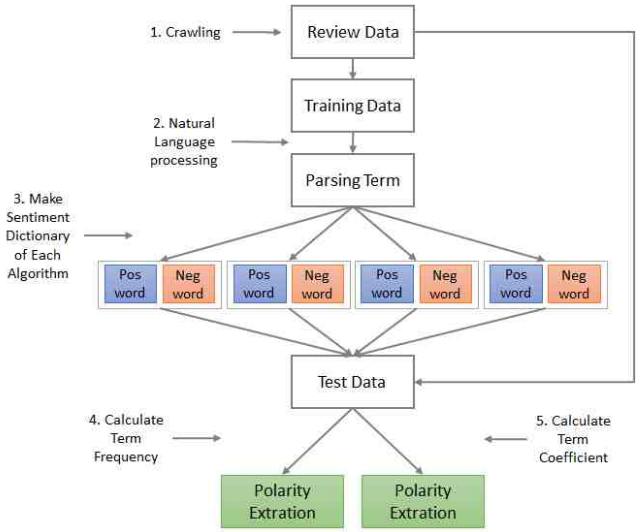

그림 3 전체 실험 모형 구성

본 연구에서 수행한 실험의 과정은 그림3과 같다. 전체 과정의 각 단계를 설명하면, 첫 번째 단계에서는 분석하려는 대상 도메인의 리뷰를 수집한 후 훈련 데이터와 검증 데이터로 구분한다. 두 번째 단계에서는 리뷰의 텍스트를 형태소 분석기로 분해하여 저장하고, 세 번째 단계에서는 선형 회귀분석, 라쏘 회귀분석, 능형 회귀분석, 엘라스틱넷, 4가지 알고리즘으로 각 단어의 회귀계수를 측정하여 감성사전을 구축한다. 네 번째 단계에서는 4가지 감성사전으로 검증 데이터의 극성을 추출한다. 이때 단어의 출현 빈도를 이용하는 극성 계산과 단어의 회귀계수를 이용하는 극성계산 두 가지 방법을 사용하여 정확도를 비교하였다.

5.3 실험 결과

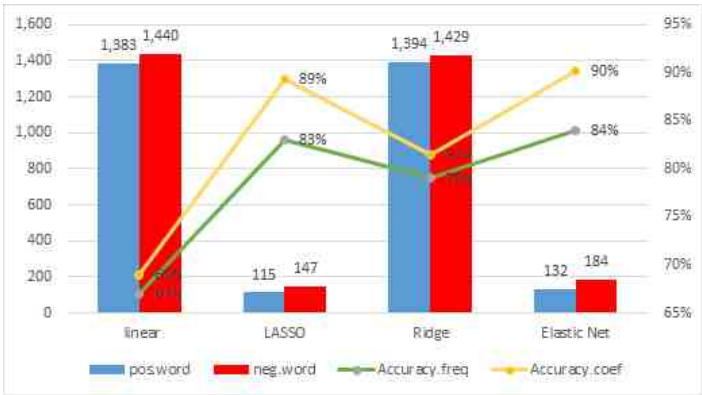

그림 4 감성사전 별 예측 정확도

그림 4에서 좌측의 파란색 막대는 각 회귀분석 방법으로 구축한 감성사전의 긍정 단어의 숫자, 우측의 빨간색 막대는 부정단어의 숫자를 의미하며, 초록색 선과 노란색 선은 단어의 빈도수만으로 예측한 정확도, 단어의 회귀계수를 이용한 예측의 정확도를 의미한다.

감성사전에 쓰인 단어의 숫자를 보면 라쏘 회귀분석과 엘라스틱넷으로 만든 감성사전은 극성 분류를 하는 데 중요한 역할을 하는 단어만 사용, 엘라스틱넷은 라쏘 회귀분석의 다중공선성의 문제를 해결하여 감성사전의 단어 수가 조금 더 많은 것을 알 수 있다.

긍정/부정 단어의 출현 빈도수로 예측한 정확도를 살펴보면 엘라스틱넷으로 만든 감성사전이 84%로서 새로운 검증데이터를 가장 잘 나타내는 것을 알 수 있었다. 단어의 회귀계수를 이용한 예측 정확도 또한 엘라스틱넷으로 만든 감성사전의 경우 90%의 정확도를 보여 긍정/부정 단어의 출현 빈도수로 예측한 정확도보다 약 6% 더 높은 것을 확인할 수 있다.

6. 결론

본 논문에서는 회귀분석의 방법인 엘라스틱넷을 사용하여 특정한 도메인의 소셜 데이터를 수집하여 각 단어의 회귀계수를 구해 단어들의 가중치를 측정하고, 중요한 단어만 추출하는 방법을 통하여 감성사전 구축에 적용하였다.

선형 회귀분석, 라쏘 회귀분석, 능형 회귀분석, 엘라스틱넷, 4가지 알고리즘으로 각 단어의 회귀계수를 측정하여 감성사전을 구축한 후, 4가지 감성사전으로 검증 데이터의 극성을 추출하였다. 이때 단어의 출현 빈도를 이용하는 극성 계산과 단어의 회귀계수를 이용하는 극성계산 두 가지 방법을 사용하여 정확도를 비교 실험해본 결과 엘라스틱넷으로 만든 감성사전이 가장 높은 정확도를 보였고, 회귀계수를 이용하는 방법의 경우 출현 빈도수만을 이용하는 방법보다 정확도가 6% 더 높은 것을 확인할 수 있었다. 또한, 기존의 논문에서 연구한 실험 결과[3][4]에 비해서도 높은 정확도를 보여 더 향상된 결과를 보였다.

Reference

[1] 안정국, 김희웅, “한글 감성어 사전 API 구축 및 자연어 처리의 활용”, 한국지능정보시스템학회, 177-182, 2014.11
[2] 이종혁, 김원상, 박제원, 최재현, “오피니언 마이닝을 활용한 블로그의 극성 분류 기법”, 한국디지털콘텐츠학회논문지, 제 15권, 제 4호, 599-568, 2014.8
[3] 송종석, 이수원, “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축”, 정보과학회논문지: 소프트웨어 및 응용, 제 38권, 제 3호, 157-168, 2011.3
[4] 유은지, 김유신, 김남규, 정승렬, “주가지수 방향성 예측을 위한 주제지향 감성사전 구축 방안”, 지능정보연구, 제 19권, 제 1호, 95-110, 2013.3
[5] 서정렬, 고찬, “감성 분석에 의한 Big Data 분석”, 융복합지식학회논문지, 제 2권, 제 1호, 15-21, 2014.1
[6] Tibshirani, R. ,“Regression shrinkage and selection via the lasso”, J. Royal. Statist. Soc B., Vol.58, No.1, 267-288, 1996.
[7] Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso”, Journal ofthe Royal Statistical Society. Series B (Methodological), Vol.58, No.1, 267-288, 1996.
[8] Arthur E. Hoerl and Robert W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, Technometrics, Vol.12, No.1, 55-67, 1970.
[9] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”, J. R. Statist. Soc. B. , Vol.67, No.2, 301-320, 2005.