기계학습 기반 환경이슈 감성분류기 개발
: 기후변화 중심으로

2018. 10. 04.

KEI 사회환경연구본부
김도연

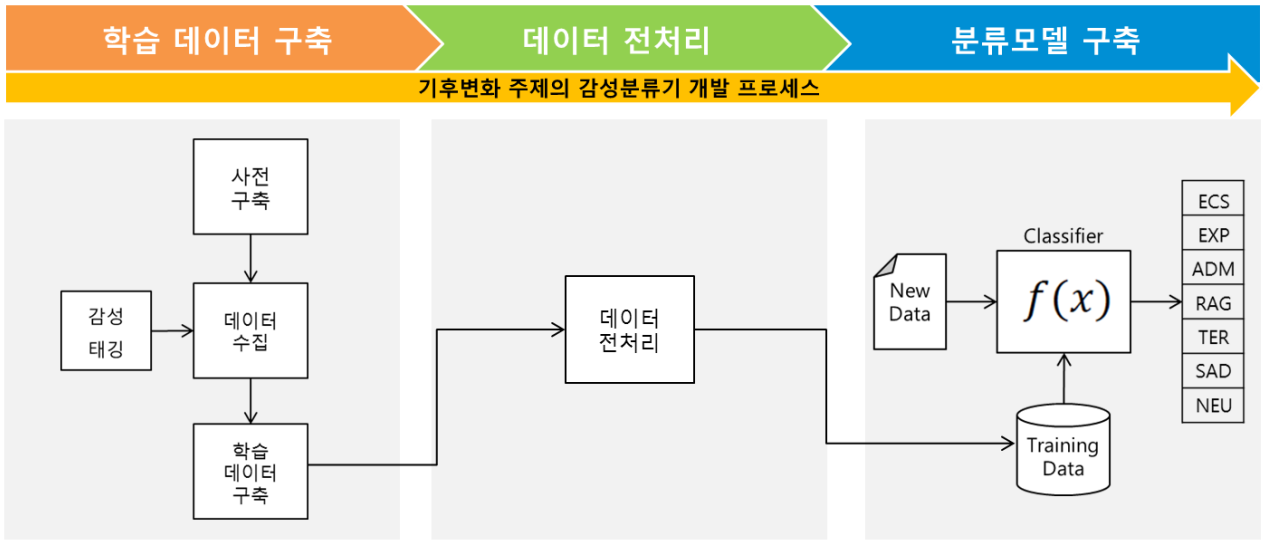
연구 범위 및 흐름도

'17년도 연구

- 연구 주제 : 텍스트 마이닝을 이용한 KEI 연구동향 분석
  - KEI 연구보고서와 환경뉴스 데이터를 이용한 분석 결과 '기후변화'가 주요 키워드로 나타남
  - 특히 환경뉴스에서 기후변화 세부 현상(폭염, 한파, 폭우, 태풍 등)에 대한 관심이 높아지고 있음

'18년도 연구

- 연구 주제 : 기계학습 기반 환경이슈 감성분류기 개발 - 기후변화 중심으로
  - 기후변화 세부 현상에 대한 환경정책 수요자인 국민의 인식 분석 필요 있음
  - SNS 및 댓글로 부터 기후변화 이슈에 대한 국민 인식(감성)을 파악하기 위한 감성분류기 개발



- 기후변화 사전구축
- SNS 및 댓글 데이터 수집
- 7가지 감성 클래스 구축
- 학습데이터에 감성 태깅
- 감성 태깅 크로스 체크

- 이모지 한글화
- 형태소 분석
- 정규화
- 말뭉치 생성
- Sparse Terms 및 low TF-IDF 삭제

- 학습 데이터 및 검증 데이터 분류
- 기계학습 기반의 분류분석
  - SVM, Naïve Bayes
  - CNN, RNN(LSTM, GRU)
- 분류모델 성능 비교 분석

연구 내용 및 성과

- 연구 목적: 기후변화 주제의 SNS 및 주요 포털 댓글 데이터 기반 감성분류 알고리즘 개발

- 연구 내용: 기후변화 사전 구축, 감성 분류 학습 데이터 구축, 감성분류 알고리듬 개발
  - **기후변화 사전** : 기후변화에 따른 현상을 4개의 범주(온도, 강수, 토지, 해양) 분류 후 구축
    - 환경관련 문서에 워드 임베딩 방법(LDA, Word2Vec) 적용 후보군 추출
    - 전문가(최희선,명수정) 및 SNS 이용자 의견 반영
  - **감성분류 기준표** : 기후변화 현상에 자주 나타나는 7개 감성 클래스 구축
    - 7개 감성 카테고리 : 황홀/기쁨, 기대/관심, 감탄/존경, 분노/짜증, 두려움/공포, 슬픔/수심, 중립
  - **감성분류 학습데이터** : 약 5만 건 단문 데이터에 감성을 수작업으로 파악
    - 기후변화 사전 기준 5만건을 수집하여 7개의 감성 클래스 태깅
  - **감성분류 알고리즘** : 다양한 기계학습 기반 분류 알고리즘 구축
    - SVM, Naïve bayes, ~~CNN (진행 중)~~

- 연구 성과: 기후변화 주제의 감성분류기 구축
  - SVM을 이용한 모델의 분류 정확도가 높게 나타남
    - 7개 감성 분류 정확도 : **77.96%** / 3개 감성 분류 정확도: **87.25%**

기후변화 사전 및 감성 클래스

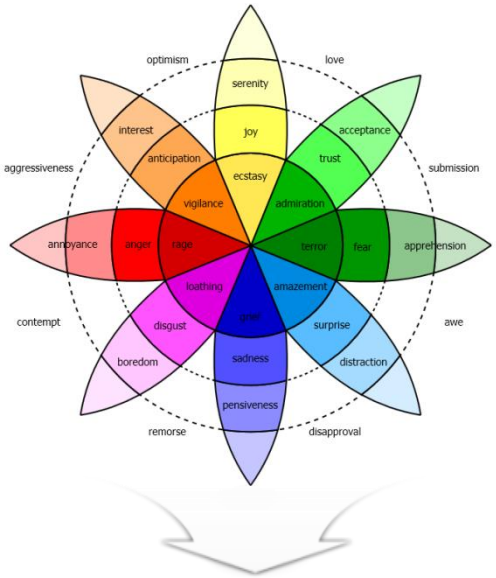
기후변화에 따른 현상 사전

- 기후변화 관련 텍스트 데이터 파악 기준

구분	순번	온도	강수	토지	해양
전문가	1	강추위	대설	가뭄	녹조
	2	결빙	산성비	사막화	라니냐
	3	무더위	우박	산불	쓰나미
	4	열대야	장마	산사태	엘니뇨
	5	열섬	적설	열대림파괴	적조
	6	열파	집중강우	지진	침수
	7	온난	집중호우	토지황폐화	파랑
	8	온실가스	폭설	화산폭발	풍랑
	9	이상고온	폭우	-	풍수해
	10	이상기온	홍수	-	해랑
	11	이상저온	황사비	-	해수면
	12	폭염	-	-	해일
	13	한파	-	-	-
	14	혹서	-	-	-
	15	혹한	-	-	-
비전문가	16	짱춥	눈난리	갈라진땅	괴물파도
	17	짱덥	눈쓰레기	메마른땅	큰파도
	18	졸춥	눈폭탄	산폭팔	-
	19	졸덥	물난리	찢어진땅	-
	20	넘덥	비폭탄	흔들리는땅	-
	21	넘춥	흙비	-	-
	22	너무춥	흙탕물비	-	-
	23	너무덥	-	-	-
	24	개춥	-	-	-
	25	개덥	-	-	-

감성분류 기준표

- 기후변화 주제에 적합한 감성 선택



감성 구분		감성 태그
긍정	황홀/기쁨	ECS
	기대/관심	EXP
	감탄/존경	ADM
부정	분노/짜증	RAG
	두려움/공포	TER
	슬픔/수심	SAD
중립		NEU

학습 데이터 구축: 8개 감성으로 분류된 5만건

기후변화 현상 사전

기후변화에 따른 '현상' 키워드					
구분	no.	온도	강수	토지	해양
전문가	1	강추위	대설	가뭄	녹조
	2	결빙	산성비	사막화	라니냐
	3	무더위	우박	산불	쓰나미
	4	열대야	장마	산사태	엘니뇨
	5	열섬	적설	열대림파괴	적조
	6	열파	집중강우	지진	침수
	7	온난	집중호우	토지황폐화	파랑
	8	온실가스	폭설	화산폭발	풍랑
	9	이상고온	폭우		풍수해
	10	이상기온	홍수		해랑
	11	이상저온	황사비		해수면
	12	폭염			해일
	13	한파			
	14	혹서			
	15	혹한			
비전문가	16	짱춥	눈난리	갈라진땅	괴물파도
	17	짱덥	눈쓰레기	메마른땅	큰파도
	18	졸춥	눈폭탄	산폭팔	
	19	졸덥	물난리	찢어진땅	
	20	넙덥	비폭탄	흔들리는땅	
	21	넘춥	흙비		
	22	너무춥	흙탕물비		
	23	너무덥			
	24	개춥			
	25	개덥			

		감성 태그
긍정	황홀/기쁨	ECS
	기대/관심	EXP
	감탄/존경	ADM
부정	분노/짜증	RAG
	두려움/공포	TER
	슬픔/수심	SAD
중립		NEU

Chanel	Keyword	Keyword1	Content	Tag
뉴스 댓글	온도	강추위	난 추울때 겨울 냄새가 넘 좋아	ECS
뉴스 댓글	온도	무더위	밖에서 일하시는 이 세상에 아버지 어머니들 힘내세요!!!	ADM
뉴스 댓글	온도	무더위	더 뜨거워지면 좋겠다	EXP
뉴스 댓글	온도	강추위	동요에 찬바람 불어도 괜찮아요 가사있죠? 이 날씨에 괜찮은 사람 없을듯	RAG
뉴스 댓글	온도	무더위	하루하루가 더위때문에 넘 힘드네요..내일은 얼마나 또더울까..이런생각에..	TER
뉴스 댓글	온도	너무춥	의정부 -13 너무 춥다..ㅠㅠ	SAD
뉴스 댓글	온도	강추위	백수라 춥든말든....	NEU
Facebook	온도	열대야	샤워하고먹는아이스아메리카노가 진리임 열대야극복완료 단순해ㅋㅋㅋ	ECS
Facebook	온도	무더위	밖에 내다 놓은 화초들이 무더위를 견디다니..허허..튼튼한것들	ADM
Facebook	온도	열대야	열대야가 이번주가 끝이래요!!!!	EXP
Facebook	온도	결빙	이젠 눈이 싫어 지네요	RAG
Facebook	온도	결빙	역대급추위.....ㄷㄷㄷ올 첫 영산강 결빙현상 발생	TER
Facebook	온도	무더위	응? 무더위 본격적 시작이라는데 ,, 그리고 10월까지 덥다는데 실화냐 ㅠㅠ	SAD
Facebook	온도	강추위	이거봐 매일매일이 강추위임	NEU

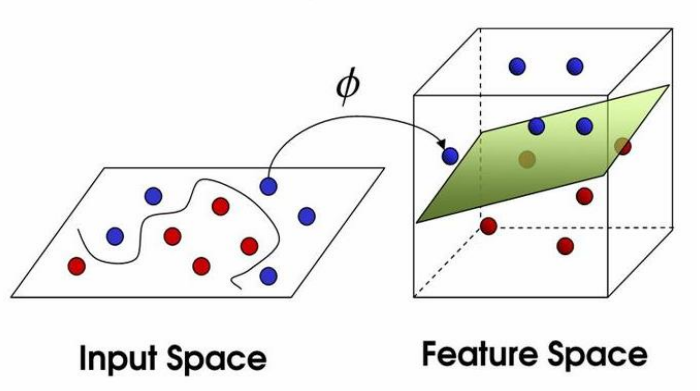
학습 데이터 전처리

<table>
<tr><th>전처리 단계</th><th>전처리 내용</th></tr>
<tr><td>1) 이모지 한글로 변환
2) 이모티콘(특수문자) 전처리
3) 형태소 분석
4) ID 삭제
5) 정규화 : 함축어, 신조어, 은어 등</td><td>- SNS 특성을 반영한 전처리 단계
- 형태소 분석기 : 은전한닢-Mecab 이용함
- 이모지 전처리: 약 1,200개 이모지 한글로 변환
예) 😍😱</td></tr>
<tr><td>6) Document Term Matrix(DTM) 생성
7) 말뭉치(Corpus) 생성 : 단어길이 최소 2글자 이상
8) Sparse Terms 삭제 : 출현빈도가 매우 낮은 단어 삭제
9) Low TF-IDF 삭제</td><td>- DTM :
<table>
<tr><th></th><th>Term1</th><th>Term2</th><th>...</th><th>TermM</th></tr>
<tr><td>Doc1</td><td>2</td><td>1</td><td>...</td><td>0</td></tr>
<tr><td>Doc2</td><td>0</td><td>4</td><td>...</td><td>2</td></tr>
<tr><td>...</td><td></td><td></td><td></td><td></td></tr>
<tr><td>DocN</td><td>3</td><td>1</td><td>...</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>10) 데이터 프레임(Data frame) 형태로 변환</td><td>- 기계학습 분석에 적합한 형태로 변환</td></tr>
</table>

분류 모델 구축

SVM Modeling

- 커널(Kernel) 트릭을 이용한 비선형 데이터 분류
- 커널 종류 및 파라미터
  1) 선형 커널(Linear Kernel) : Cost, Gamma
  2) RBF 커널(Radial Basis Function Kernel : Cost, Gamma
  3) 시그모이드 커널(Sigmoid Kernel) : Cost, Gamma, Coefficient
  4) 다항식 커널(Polynomial Kernel) : Cost, Gamma, Coefficient, Degree



$$Linear\ Kernel: K(x_n, x_i) = (x_n, x_i)$$

$$RBF\ Kernel: K(x_n, x_i) = \exp(-\gamma \parallel x_n - x_i \parallel^2 + C)$$

$$Sigmoid\ \ Kernel: K(x_n, x_i) = \tanh(\gamma(x_n, x_i) + r)$$

$$Polynomial\ Kernel: K(x_n, x_i) = (\gamma(x_n, x_i) + r)^d$$

분류 모델 구축

NaiveBayes Modeling

- 파라미터 : 라플라스 추정기(fL)

  - 각 예측 범주에서 발생 확률이 0이 되지 않도록 하기 위해 기본적으로 주는 작은 값
  (최소한 1의 값을 줌)
  - 라플라스 추정기는 어떠한 값이든 설정 가능하나 실제로 충분히 큰 Training data를 가지고 있다면
  가장 작은 단위의 값을 설정하면 됨

감성분류 분석 : 교차표 (일부)

SVM_Model_1: 3개 감성분류 교차표

	1.POS	2.NEG	3.NEU	Row Total
1.POS	6467	276	511	7254
	0.892	0.038	0.070	0.372
	0.917	0.034	0.120	
	0.331	0.014	0.026	
2.NEG	263	7603	778	8644
	0.030	0.880	0.090	0.443
	0.037	0.926	0.183	
	0.013	0.389	0.040	
3.NEU	325	335	2962	3622
	0.090	0.092	0.818	0.186
	0.046	0.041	0.697	
	0.017	0.017	0.152	
Column Total	7055	8214	4251	19520
	0.361	0.421	0.218	

SVM_Model_1: 7개 감성분류 교차표

	1.ECS	2.EXP	3.ADM	4.RAG	5.TER	6.SAD	7.NEU	Row Total
1.ECS	1864	54	197	41	21	15	341	2533
	0.736	0.021	0.078	0.016	0.008	0.006	0.135	0.130
	0.796	0.026	0.075	0.015	0.007	0.006	0.079	
	0.095	0.003	0.010	0.002	0.001	0.001	0.017	
2.EXP	69	1736	70	26	89	17	146	2153
	0.032	0.806	0.033	0.012	0.041	0.008	0.068	0.110
	0.029	0.831	0.027	0.010	0.032	0.006	0.034	
	0.004	0.089	0.004	0.001	0.005	0.001	0.007	
3.ADM	152	104	2148	26	34	16	64	2544
	0.060	0.041	0.844	0.010	0.013	0.006	0.025	0.130
	0.065	0.050	0.823	0.010	0.012	0.006	0.015	
	0.008	0.005	0.110	0.001	0.002	0.001	0.003	
4.RAG	61	23	45	2110	207	90	435	2971
	0.021	0.008	0.015	0.710	0.070	0.030	0.146	0.152
	0.026	0.011	0.017	0.794	0.073	0.034	0.101	
	0.003	0.001	0.002	0.108	0.011	0.005	0.022	
5.TER	27	35	24	203	2061	159	115	2624
	0.010	0.013	0.009	0.077	0.785	0.061	0.044	0.134
	0.012	0.017	0.009	0.076	0.730	0.059	0.027	
	0.001	0.002	0.001	0.010	0.106	0.008	0.006	
6.SAD	40	32	21	79	297	2348	268	3085
	0.013	0.010	0.007	0.026	0.096	0.761	0.087	0.158
	0.017	0.015	0.008	0.030	0.105	0.877	0.062	
	0.002	0.002	0.001	0.004	0.015	0.120	0.014	
7.NEU	128	106	105	172	116	32	2951	3610
	0.035	0.029	0.029	0.048	0.032	0.009	0.817	0.185
	0.055	0.051	0.040	0.065	0.041	0.012	0.683	
	0.007	0.005	0.005	0.009	0.006	0.002	0.151	
Column Total	2341	2090	2610	2657	2825	2677	4320	19520
	0.120	0.107	0.134	0.136	0.145	0.137	0.221	

전체 데이터: 감성분류 정확도

(단위: %)

Model	Sentiment Class	
	7 Class	3 Class
svm_Model_1	**77.96**	**87.25**
svm_Model_2	68.71	76.25
svm_Model_3	76.85	86.31
svm_Model_4	75.04	84.08
nb_Model	50.57	65.73

매체별 데이터: 감성분류 정확도

(단위: %)

Chanel	Model	Sentiment Class	
		7 Class	3 Class
Facebook	svm_Model_1	80.70	**88.53**
	svm_Model_2	53.94	69.94
	svm_Model_3	77.59	86.41
	svm_Model_4	**81.43**	87.69
	nb_Model	35.25	57.89
Twitter	svm_Model_1	78.79	**87.47**
	svm_Model_2	52.23	70.44
	svm_Model_3	79.86	86.43
	svm_Model_4	**79.94**	86.48
	nb_Model	38.60	63.09
Instagram	svm_Model_1	75.19	**86.37**
	svm_Model_2	41.10	71.33
	svm_Model_3	73.68	84.44
	svm_Model_4	**75.30**	85.41
	nb_Model	42.30	54.06
News comment	svm_Model_1	77.05	88.43
	svm_Model_2	42.47	66.28
	svm_Model_3	76.99	86.50
	svm_Model_4	**79.73**	**88.93**

향후 계획

- 딥러닝 기반 분류기 구축
  - Pytorch를 이용한 CNN 기반 분류 알고리듬 구축
  - 분류 정확도 성능 평가를 통해 최종 분류모델 선정

- 감성분류를 위한 CNN 모델