감정분석 연구와 감정 말뭉치 학습 데이터셋

국민대학교 인공지능학부
강 승 식

목차

- 감정의 정의 및 유형 분류
- 감정분석 말뭉치 구축 기법
- 감정분석 방법론
- 문장벡터 생성 및 토큰화 기법
- 감정분석 말뭉치 구축 사례

Definition of a sentiment and emotion

- **Sentiment** : a *positive* or *negative* feeling underlying the opinion
  - Sentiment analysis is interpreted synonymously to opinion mining
    - Primarily text-oriented, but there are multimodal approaches
  - 감성 컴퓨팅(affective computing): 기계가 인간의 감성을 인지, 해석하고 처리할 수 있게 하는 기술

- **Emotion** : an *integrated feeling state*
  - Involving physiological changes, motor-preparedness, cognitions about action, and inner experiences that emerges from an appraisal of the self or situation
  - "어떤 현상이나 일에 대하여 일어나는 마음이나 느끼는 기분"
  - Emotions involve a set of expressive(생각-감정의 표현), behavioral(행동에 관한), physiological(생리학적), and phenomenological(현상학적) features.
  - *E. Kim and R. Klinger, "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies," 2018.* https://arxiv.org/ftp/arxiv/papers/1808/1808.03137.pdf

M. Munezero, C. Montero, E. Sutinen, J. Pajunen, "Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text," IEEE Trans. on Affective Computing, pp.101-111, 2014.

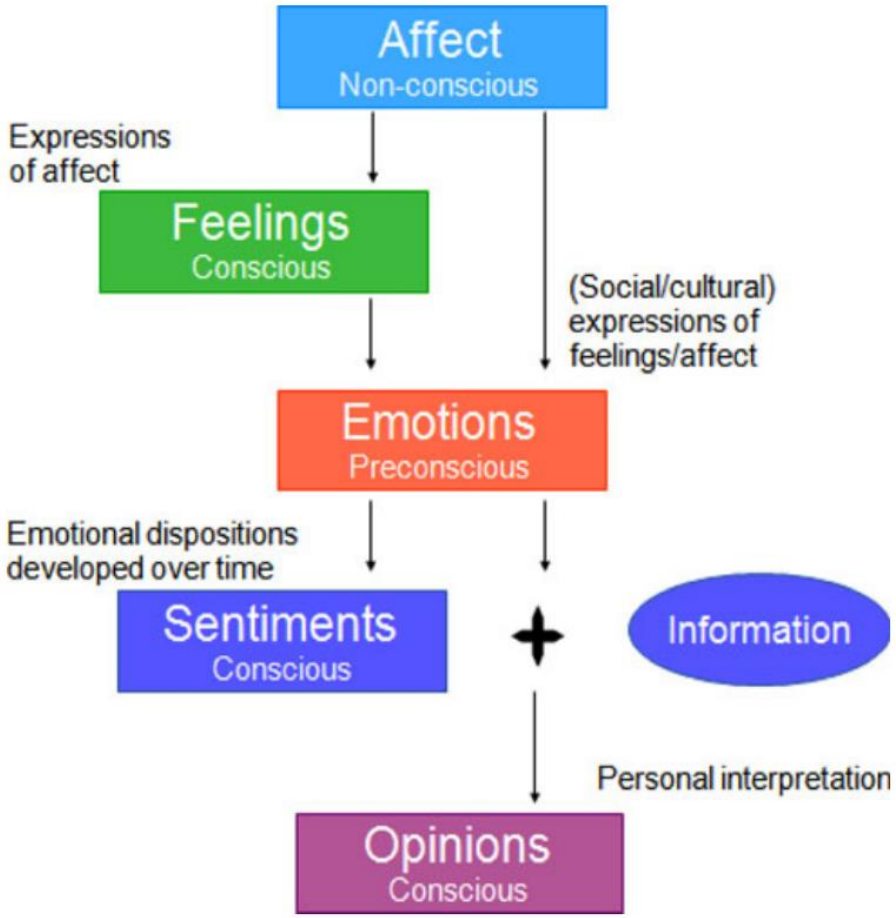


Fig. 2. Differentiating factors between affect, feelings, emotions, sentiments and opinions.

TABLE 1
Definitions Provided by Merriam-Webster Online Dictionary [6]

Subjectivity Term	Definition	Synonym
Affect	The conscious subjective aspect of an emotion considered apart from bodily changes; also a set of observable manifestations of a subjectively experienced emotion	Feeling
Feeling	An emotion state or reaction; often unreasoned opinion or belief	Sentiment, Emotion
Emotion	Excitement; the affective aspect of consciousness; a state of feeling; a conscious mental reaction (as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body	Feeling, Sentiment
Sentiment	An attitude, thought, or judgement prompted by feeling; a specific view or notion	Feeling, Emotion, Opinion
Opinion	A view, judgement, or appraisal formed in the mind about a particular matter; A belief stronger than impression and less strong than positive knowledge	Feeling, Sentiment

Soo-Min and Edward Hoby, "Determining the sentiment of opinions," COLING'04, 2004.

- Word sentiment classifier
- Sentence sentiment classifier

Fig. 1. Example of an opinion (based on Kim and Hovy [24]).

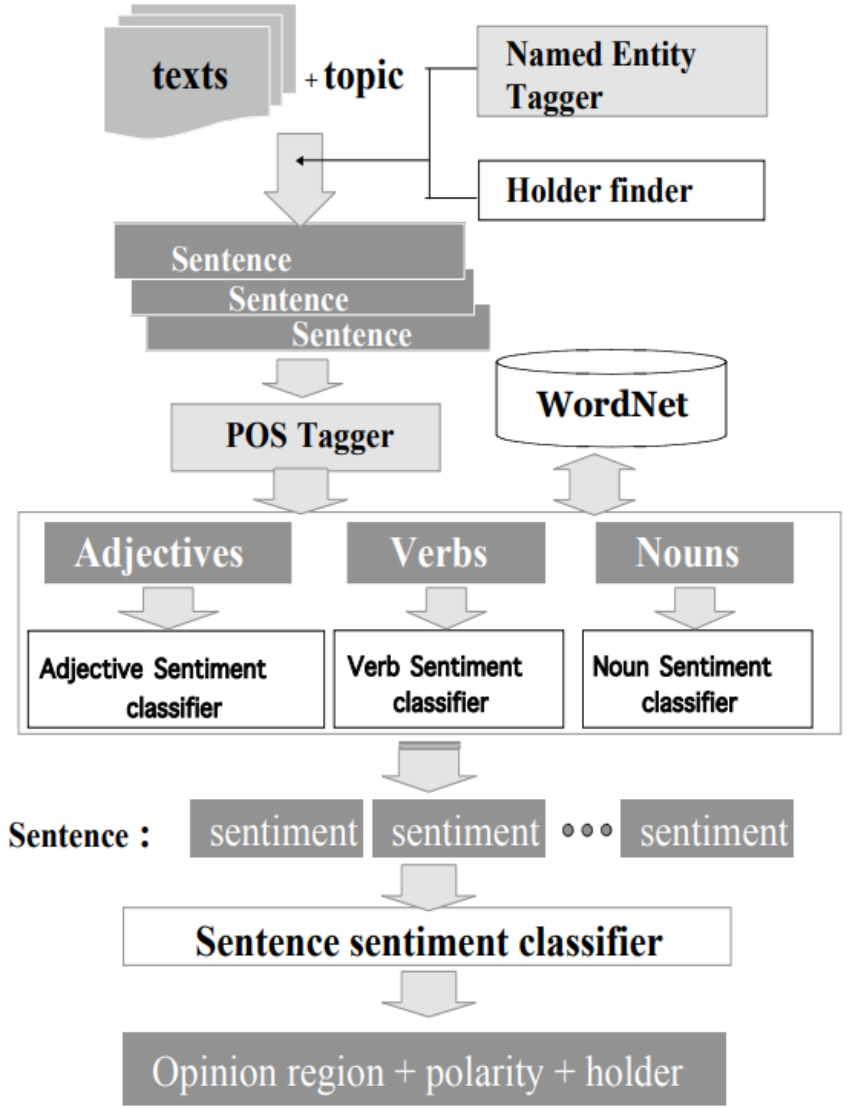


Figure 1: System architecture.

감정의 유형 분류

- 2진 분류: positive, negative (neutral)
  - Movie review, product review

- Ekman(1960), Plutchik(1980), Russell(1980)
  - Ekman’s Theory of Basic Emotions

    Emotions should be considered discrete categories rather than continuous.
  - Plutchik’s Wheel of Emotions

    Emotions are mixed and derived from the various combinations of basic ones
  - Russel’s Circumplex Model

- Ekman’s basic emotions: 감정 분석 레이블 유형
  - Discrete label sets: joy, anger, fear, sadness
  - Continuous values: valence(가치), arousal(각성,흥분)

- Plutchik’s 8 basic emotions in 4 opposite pairs
  - Joy-sadness, anger-fear, trust-disgust, and anticipation-surprise

Plutchik and Russel

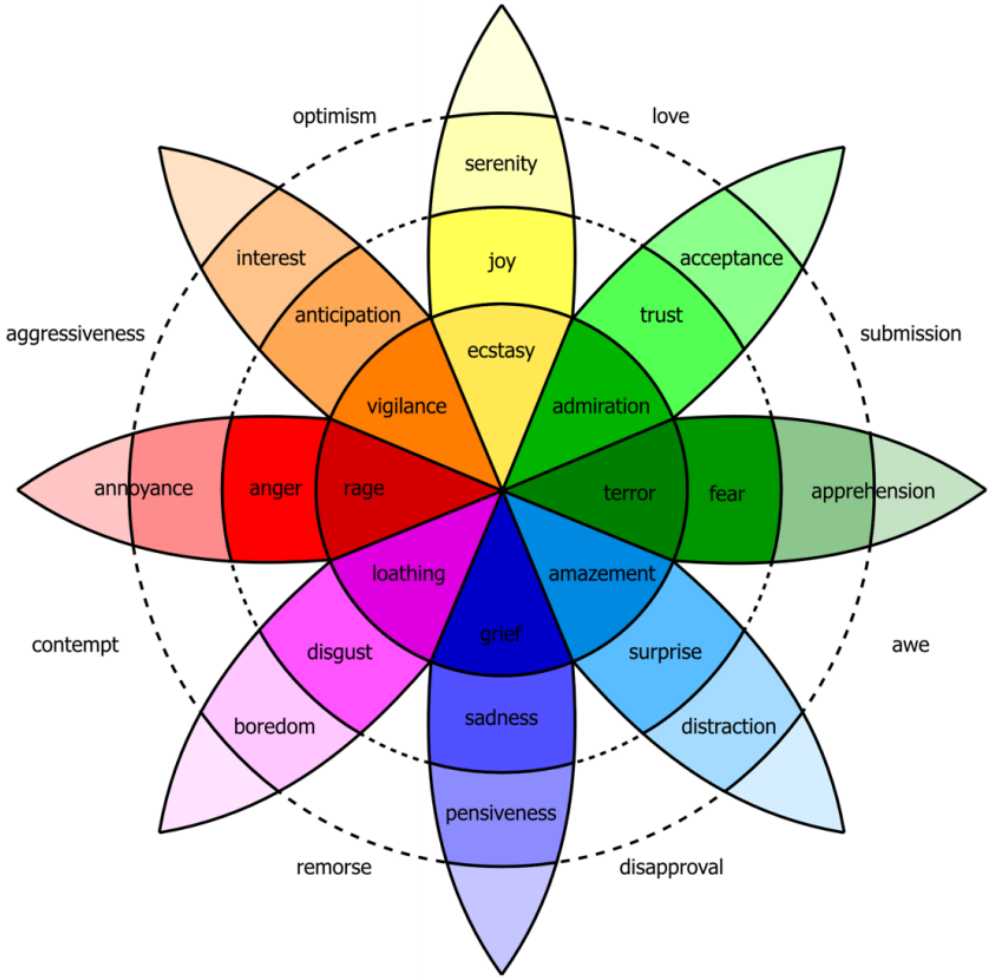


Figure 1: Plutchik's wheel of emotions.

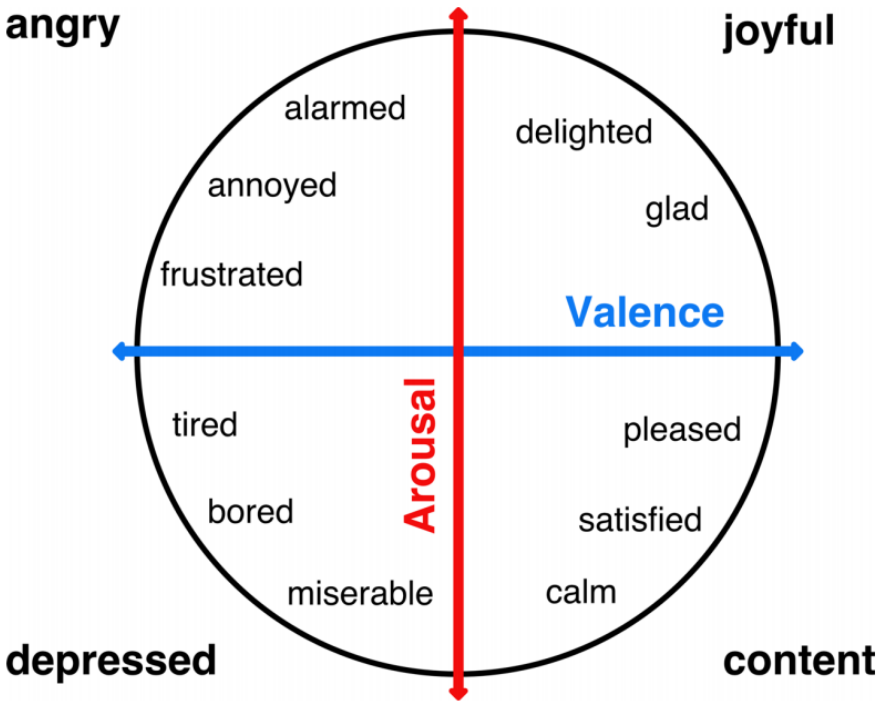


Figure 2: Circumplex model of affect: Horizontal axis represents the valence dimension, the vertical axis represents the arousal dimension. (drawn after Posner, Russell, Peterson 2005)

- [E] Ekman

  anger, disgust, fear, joy, sadness, surprise

- [P] Plutchik

  anger, disgust, fear, joy, sadness, surprise, trust, anticipation

- [CF] CrowdFlower

  enthusiasm, fun, hate, neutral, love, boredom, relief, empty

감정 말뭉치 구축: Annotation 기법

- Standard annotation by experts
- Crowdsourcing platforms
- Social networks with distant supervision (self-labeling)

감정 말뭉치 구축: Annotation by experts

- Standard annotation by experts
  - Aman(2007), Strapparava(2007), Ghazi(2015), Li(2017), Schuff(2017), Liu(2017)

- ISEAR dataset
  - 7가지 감정: joy, sadness, anger, fear, disgust, shame, and guilt
  - 7,666 sentences ~ 1,096x7

  - “Emotion dominant meaning tree”를 이용한 감정분석, M. Razek(2017)

    https://www.researchgate.net/figure/Characteristics-of-the-ISEAR-Dataset_tbl1_313407834

[참고문헌] Annotation by experts

- Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In V´aclav Matouˇsek and Pavel Mautner, editors, Text, Speech and Dialogue, pages 196–205.
- Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 70–74.
- Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, pages 152–165.
- Yanran Li, Hui Su, Xiaoyu Shen,Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957.
- Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Pad´o, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Workshop at Conference on EMNLP.
- Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 477–483

감정 말뭉치 구축: Crowdsourcing platforms

- Crowdsourcing platforms
  - Amazon’s Mechanical Turk: https://www.mturk.com/
  - CrowdFlower: https://www.crowdflower.com/

  - *Often lacks sufficient quality control but some popular datasets have been successfully acquired!*

  - Crowdflower for Cortana dataset
    https://www.crowdflower.com/data/sentiment-analysis-emotion-text/

Social networks with distant supervision

- Social network platforms with *distant supervision* (self-labeling)
- on Twitter
  - add a “#joy” hashtag to a happy tweet
- on Facebook
  - tag personal posts with a “feeling” and people can show an emotional “surprised reaction”

- Two levels of annotation are provided that are relevant
  - both the reader’s and the writer’s emotional state
  - Buechel and Hahn (2017a, 2017b)

감정 분석 방법론

- Rule-based algorithms
  - WordNet, SentiWordNet, LIWC(Linguistic Inquiry and Word Count)
- Tree-based 분류 기법
- 기계학습 기법
  - SVM, Naïve Bayse, kNN, CRF 등
  - 딥러닝 기법: CNN, RNN
- 앙상블 기법: 2가지 이상의 방법론 사용

Tree-based Sentiment Classification

- T. Nakagawa, K. Inui, S. Kurohashi, “Dependency tree-based sentiment classification using CRFs with hidden variables,” In Proceedings of ACL 2010.

- R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, C. Potts, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Tree-bank,” In Proceedings of EMNLP 2013.

- M. Razek and C. Frasson, “Text-Based Intelligent Learning Emotion System,” Journal of Intelligent Learning Systems and Applications 09(01), pages 17-20, 2017.

Sentence Classification by CNN

- Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, Proceedings of EMNLP, pp.1746-1751, 2014.

- A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, “CNN Features off-the-shelf: an Astounding Baseline,” CoRR, abs/1403.6382.

- N. Kalchbrenner, E. Grefenstette, P. Blunsom, “A Convolutional Neural Network for Modelling Sentences,” In Proceedings of ACL 2014.

- B. Yang, C. Cardie, “Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization,” In Proceedings of ACL 2014.

E. Kim and R. Klinger(2018)

"A Survey on Sentiment and Emotion Analysis for Computational Literary Studies"

https://arxiv.org/ftp/arxiv/papers/1808/1808.03137.pdf

Table 1: Summary of characteristics of methods used in the papers reviewed in this survey.

		Features						Models						
	Citation	TF-IDF	Bag-of-words	Dictionary	Syntax	Embeddings	Annotation	SVM	Naïve Bayes	Decision Trees	Rules	PCA	Deel Learning	Other
Emotion classification	Yu (2008)	*						*	*					
	Barros et al. (2013)			*						*				
	Reed (2018)			*							*			
	Zehe et al. (2016)			*			*	*						
	Reagan et al. (2016)			*								*	*	
	Samothrakis and Fasli (2015)		*	*						*			*	
	Kim et al. (2017a)		*	*						*				
	Henny-Krahmer (2018)			*						*				
Temporal	Heuser et al. (2016)						*							*
	Bruggmann and Fabrikant (2014)			*							*			
	Taboada et al. (2006, 2008)			*			*				*			
	Chen et al. (2012)			*							*			
	Marchetti et al. (2014)			*			*				*			
	Sprugnoli et al. (2016)			*			*				*			
	Buechel et al. (2017)			*		*								
	Buechel et al. (2016)			*		*	*							
	Leemans et al. (2017)			*			*				*			
Network Analysis	Nalisnick and Baird (2013a,b)			*							*			
	Elsner (2012, 2015)			*	*						*			
	Kim and Klinger (2018)					*	*						*	*
	Barth et al. (2018)			*							*			
	Jhavar and Mirza (2018)			*							*			
	Egloff et al. (2018)						*							*
	Rinaldi et al. (2013)													*
	Zhuravlev et al. (2014)													*
	Jarafi et al. (2016)													*
	Kim and Klinger (2019a)						*							*
	Kim and Klinger (2019b)					*	*				*		*	
Other	Anderson and McMaster (1986)			*							*			
	ANderson and McMaster (1993)			*							*			
	Alm and Sproat (2005)						*				*			
	Mohammad (2011, 2012)			*							*			
	Klinger et al. (2016)			*							*			
	Kim et al. (2017b)			*							*			*
	Kakkonen and Galic Kakkonen (2011)			*	*						*			
	Koolen (2018)			*							*			
	Kraicer and Piper (2019)			*										*
	Morin and Acerbi (2017)			*							*			
	Bentley et al. (2014)			*							*			

문장 벡터 생성 기법

- TfidfVectorizer: scikit-learn
- 워드 임베딩 기법: word2vec 등
- 문서 임베딩 기법: BERT, ALBERT 등

감정 분석을 위한 토큰화 기법

- 토큰화 단위
  - Character (unigram)
  - N-gram
  - Morpheme: 형태소 분석기
  - Subword: sentencePiece algorithm

- 토큰화 기법은 위 2가지 이상을 혼합할 수 있음!

- Multi-prototype embedding : 품사 정보, 위치 정보 등

긍정/부정 어휘 가중치 사용 기법

- 긍정, 부정 표현에 사용되는 어휘들을 추출
  - Term clustering 기법
  - Word embedding 기법

- 문장 벡터 생성할 때 긍정/부정 토큰 가중치 부여

감정분석 말뭉치 구축 사례

- Movie review, product review
- Stanford Sentiment Treebank
- 6가지 감정 말뭉치 데이터셋
- 유해성 말뭉치 데이터셋
  - Kaggle.com: Jigsaw Unintended Bias in Toxicity Classification
  - Korean Corpus of Online News Comments for Toxic Speech Detection
- 사용자 의도파악 데이터셋 -- http://aihub.com
- Naver Sentiment Movie Corpus v1.0 및 확장 구축

SST: Stanford Sentiment Treebank
https://nlp.stanford.edu/sentiment/index.html

- 5 sentiment classes for each subtree

- 논문: “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”

https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

- https://www.kaggle.com/atulanandjha/stanford-sentiment-treebank-v2-sst2

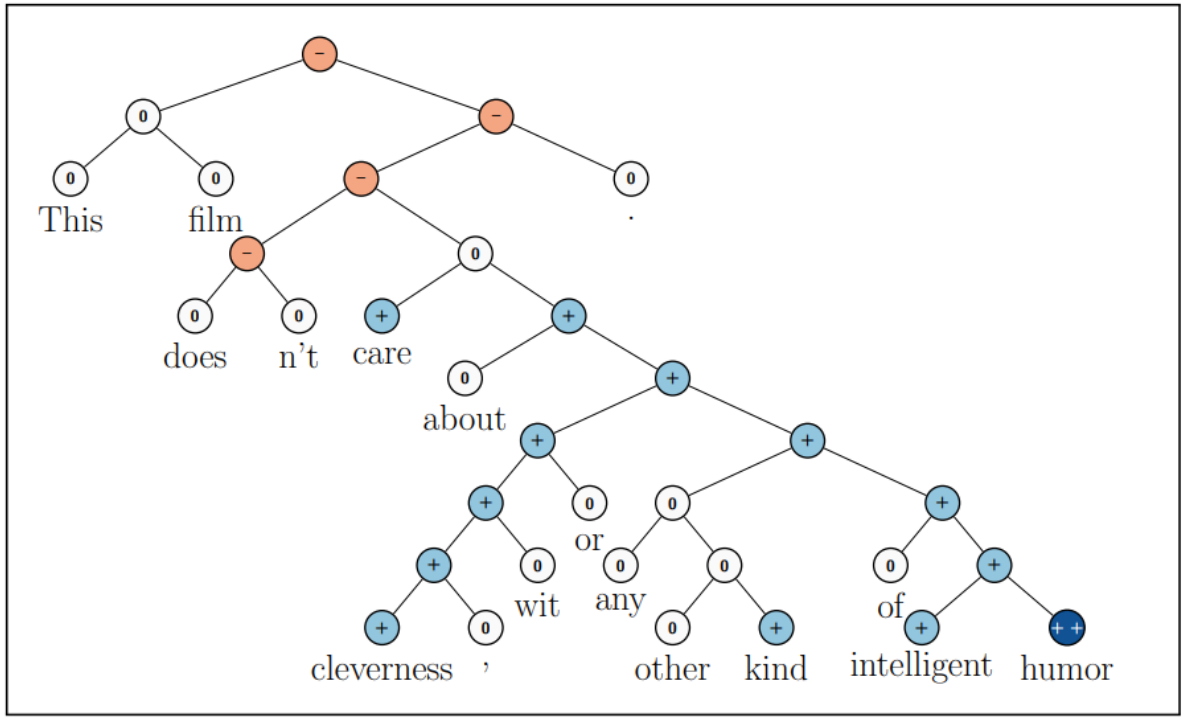


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

Example of contrastive conjunction X but Y

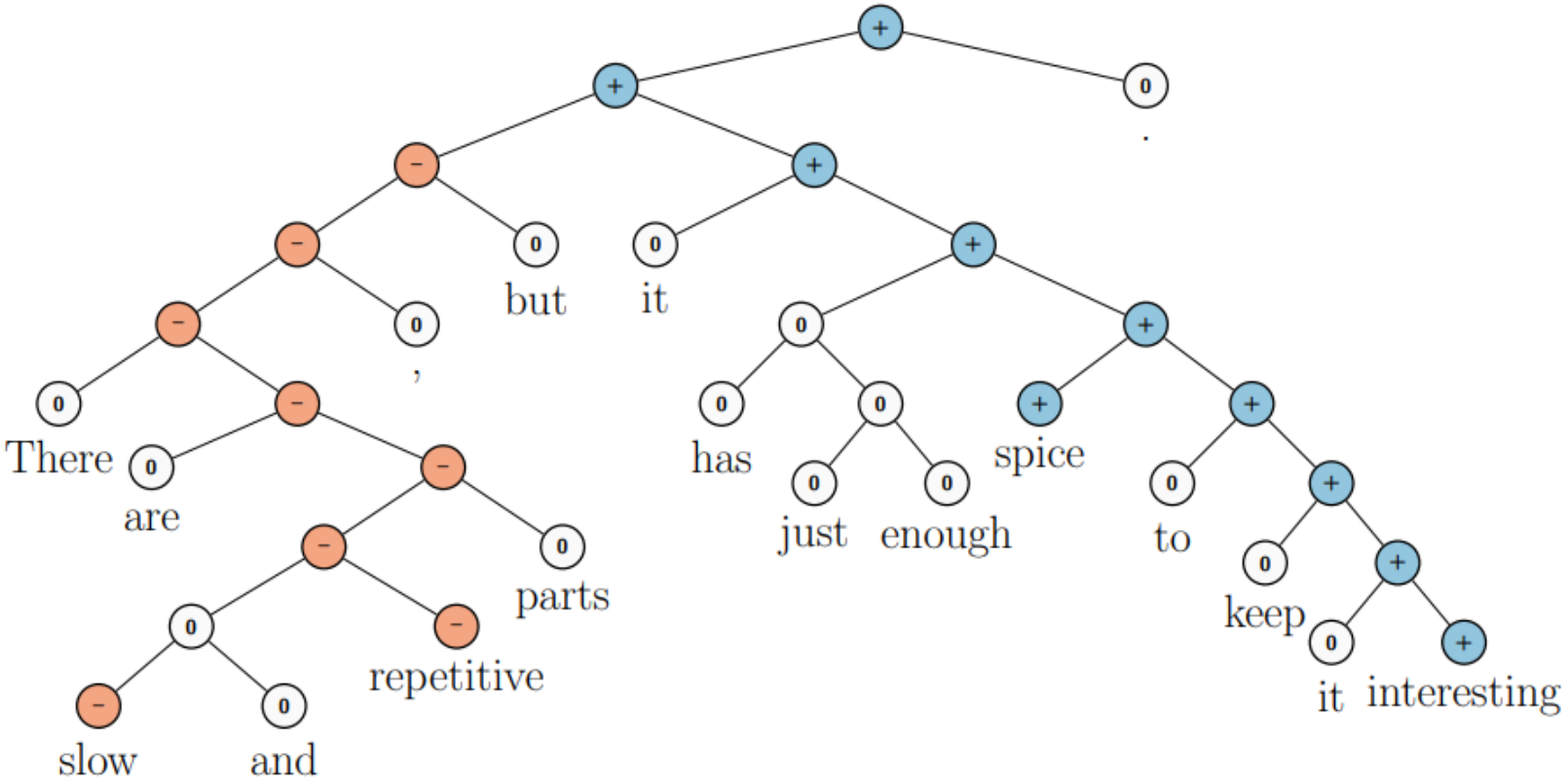
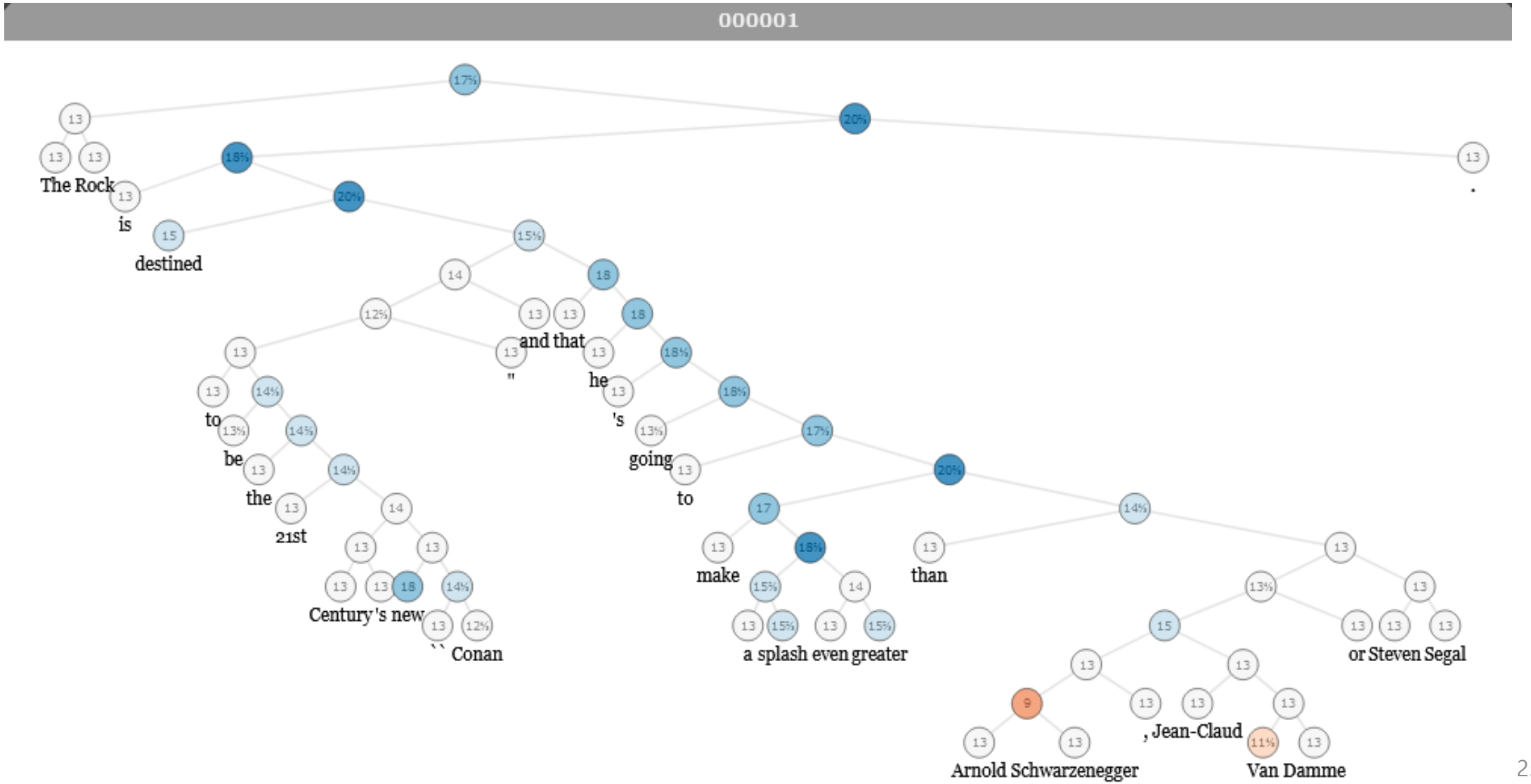


Figure 7: Example of correct prediction for contrastive conjunction *X but Y*.

감정 트리: 9,645 문장

https://nlp.stanford.edu/sentiment/treebank.html



Experiments on 5 classes and binary classification

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

Table 1: Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) and for all nodes.

Sentiment Treebank: 한글 번역 데이터셋

http://nlp.kookmin.ac.kr/corpus/Sentiment_Treebank_en_ko.zip

Stanford Sentiment Treebank Korean translation
Total 201,019 words (11,855 sentences), 430,776 subtrees
==========
Sentiment_Treebank train set
1) Sentiment_Treebank_en_ko_train_set.txt
--> 8,544 sentences, 318,583 sub tree

Sentiment_Treebank test set
2) Sentiment_Treebank_en_ko_test_set.txt
--> 2,209 sentences, 71,848 sub tree

Sentiment_Treebank dev set
3) Sentiment_Treebank_en_ko_dev_set.txt
--> 1,102 sentences, 40,345 sub tree

Data Structure -->
###Score&Sentence=<tap>score<tap>eng_sent<tap>kor_sent
score<tap>eng_sub_tree<tap>kor_sub_tree
score<tap>eng_sub_tree<tap>kor_sub_tree

6가지 감정 말뭉치 데이터셋

- E. Saravia, H. Liu, Y. Huang, J. Wu, Y. Chen, "CARER: Contextualized Affect Representations for Emotion Recognition", EMNLP-2018, pp.3687-3697, 2018.

  https://www.aclweb.org/anthology/D18-1404/

- 다운로드 1: 원본 데이터셋

  https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp

- 다운로드 2: 한국어 번역 데이터셋

  http://nlp.kookmin.ac.kr/corpus/emotion_korTran.zip

```
im feeling rather rotten so im not very ambitious right now;sadness
im updating my blog because i feel shitty;sadness
i never make her separate from me because i don t ever want her to feel like i m ashamed with her;sadness
i left with my bouquet of red and yellow tulips under my arm feeling slightly more optimistic than when i
i was feeling a little vain when i did this one;sadness
i cant walk into a shop anywhere where i do not feel uncomfortable;fear
i felt anger when at the end of a telephone call;anger
i explain why i clung to a relationship with a boy who was in many ways immature and uncommitted despite t
i like to have the same breathless feeling as a reader eager to see what will happen next;joy
i jest i feel grumpy tired and pre menstrual which i probably am but then again its only been a week and :
i don t feel particularly agitated;fear
i feel beautifully emotional knowing that these women of whom i knew just a handful were holding me and my
i pay attention it deepens into a feeling of being invaded and helpless;fear
i just feel extremely comfortable with the group of people that i dont even need to hide myself;joy
i find myself in the odd position of feeling supportive of;love
i was feeling as heartbroken as im sure katniss was;sadness
i feel a little mellow today;joy
i feel like my only role now would be to tear your sails with my pessimism and discontent;sadness
i feel just bcoz a fight we get mad to each other n u wanna make a publicity n let the world knows about c
i feel like reds and purples are just so rich and kind of perfect;joy
im not sure the feeling of loss will ever go away but it may dull to a sweet feeling of nostalgia at what
i feel like ive gotten to know many of you through comments and emails and for that im appreciative and g
i survey my own posts over the last few years and only feel pleased with vague snippets of a few of them c
i also tell you in hopes that anyone who is still feeling stigmatized or ashamed of their mental health is
i don t feel guilty like i m not going to be able to cook for him;sadness
i hate it when i feel fearful for absolutely no reason;fear
i am feeling outraged it shows everywhere;anger
i stole a book from one of my all time favorite authors and now i feel like a rotten person;sadness
i do feel insecure sometimes but who doesnt;fear
i highly recommend visiting on a wednesday if youre able because its less crowded so you get to ask the fa
ive been missing him and feeling so restless at home thinking of him;fear
i posted on my facebook page earlier this week ive been feeling a little grumpy and out of sorts the past
i start to feel emotional;sadness
i feel so cold a href http irish;anger
i feel like i m defective or something for not having baby fever;sadness
i feel more virtuous than when i eat veggies dipped in hummus;joy
i feel very honoured to be included in a magzine which prioritises health and clean living so highly im cu
i spent the last two weeks of school feeling miserable;sadness
im feeling very peaceful about our wedding again now after having;joy
```

여러 개의 감정 말뭉치 통합

- L. Bostan and R. Klinger, “An Analysis of Annotated Corpora for Emotion Classification in Text”, Proceedings on Int’l Conferenceon Computational Linguistics, pp.2104-2119, 2018.

- Unified corpus 다운로드
  https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/unifyemotion/

통합 대상 말뭉치 데이터셋

Dataset	Granularity	Annotation	Size	Topic	Source	Avail.
AffectiveText	headlines	E + V	1,250	news	Strapparava (2007)	D-U
Blogs	sentences	E + ne + me	5,025	blogs	Aman (2007)	R
CrowdFlower	tweets	E + CF	40,000	general	Crowdflower (2016)	D-U
DailyDialogs	dialogues	E	13,118	multiple	Li et al. (2017)	D-RO
Electoral-Tweets	tweets	P	4,058	elections	Mohammad (2015)	D-RO
EmoBank	sentences	V+A+D	10,548	multiple	Buechel (2017a)	CC-by4
EmoInt	tweets	E − DS	7,097	general	Mohammad (2017b)	D-RO
Emotion-Stimulus	sentences	E + shame	2,414	general	Ghazi et al. (2015)	D-U
fb-valence-arousal	faceb. posts	V+A	2,895	questionnaire	Preoţiuc (2016)	D-U
Grounded-Emotions	tweets	HS	2,585	weather/events	Liu et al. (2017)	D-U
ISEAR	descriptions	E + SG	7,665	events	Scherer (1994)	GPLv3
Tales	sentences	E	15,302	fairytales	Alm et al. (2005)	GPLv3
SSEC	tweets	P	4,868	general	Schuff et al. (2017)	D-RO
TEC	tweets	E ±S	21,051	general	Mohammad (2012)	D-RO

Table 1: Selection of resources for emotions analysis. Ann. refers to the following annotation schemata: [E] *Ekman: anger, disgust, fear, joy, sadness, surprise*, [P] *Plutchik: anger, disgust, fear, joy, sadness, surprise, trust, anticipation*, [CF] *enthusiasm, fun, hate, neutral, love, boredom, relief, empty*, [DS] *disgust, surprise*, [JS] happy, sad, [V] *valence*, [A] *arousal*, [D] *dominance*, [SG] *shame, guilt*, [±S] *positive surprise, negative surprise*, [ne] *no emotion* [me] *mixed emotion* and Availability refers to the following [D-RO] *Available to download, research only*, [D-U] *Available to download, unknown licensing*, [R] *Available upon request*, [GPLv3] *GNU Public License version 3*, [CC-by 4] *Creative Commons Attribution version 4.0*

Dataset	joy		anger		sadness		disgust		fear		surprise	
	o	m	o	m	o	m	o	m	o	m	o	m
AffectiveText	619	619	374	374	650	650	253	253	537	537	787	787
Blogs	848	536	884	179	883	173	882	172	861	115	847	115
CrowdFlower	5,209	9,220	110	1,421	5,165	5,123	—	179	8,459	8,430	2,187	2,177
DailyDialogs	12,885	12,885	1,022	1,022	1,150	1,150	353	353	74	174	1,823	1,823
Electoral-Tweets	267	349	300	569	34	31	1,937	1,638	80	91	259	251
EmoInt	1,616	1,616	1,701	1,701	1,533	1,533	—	—	2,252	2,252	—	–
Emotion-Stimulus	479	479	483	483	575	575	95	95	423	423	213	213
Grounded-Emotions	1,530	1,530	—	—	1,055	1,055	—	—	—	—	—	—
ISEAR	1,094	1,094	1,096	1,096	3,285	3,285	1,096	1,096	1,095	1,095		—
SSEC	2,067	2,067	2,902	2,902	2,644	2,644	2,183	2,183	1,840	1,840	1,108	1,108
Tales	107	107	195	195	116	116	14	14	111	111	90	90
TEC	8,240	8,237	1,555	1,555	3,830	3,830	761	761	2,816	2,816	3,849	3,849
Total	34,961	38,739	10,622	11,497	21,920	20,165	7,574	6,744	16,708	17,884	11,162	10,413

Table 2: The distribution of categories across the datasets, limited to *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. Non-availability of a class in a set is marked with —. [o]: original distributions, without taking high agreement annotations; [m]: counts after mapping to our labels (see Table 4).

유해성 말뭉치 데이터셋: kaggle.com
(Jigsaw Unintended Bias in Toxicity Classification)

- 위키피디아 댓글 159,571개를 6개의 레이블로 분류

- 총 4개의 csv 파일로 구성
  - train.csv (159,571 data) (68.8MB)
  - test.csv (153,164 data) (60.4 MB)
  - test_labels.csv (5MB)
  - sample_submission.csv

- 다운로드: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
- 챌린지: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/notebooks

- 8 Columns (total 159,571 data):
  - id
  - comment_text (6,963,526 tokens, 533,015 unique tokens)
  - toxic (악성)
  - severe_toxic (심한 악성)
  - obscene (외설)
  - threat (위협)
  - insult (모욕)
  - identity_hate (정체성 혐오)

Toxic:1
Severe_toxic: 0
Obscene:1
Threat: 0
Insult: 1
Identity_hate: 1

"You are gay or antisemmitian? \n\nArchangel WHite Tiger\n\nMeow! Greetingshhh!\n\nUh, there are two ways, why you do erased my comment about WW2, that holocaust was brutally slaying of Jews and not gays/Gypsys/Slavs/anyone...\n\n1 - I f you are anti-semitian, than shave your head bald and go to the skinhead meetings!\n\n2 - If you doubt words of the Bible, that homosexuality is a deadly sin, make a pentagram tatoo on your forehead go to the satanistic masses with y our gay pals!\n\n3 - First and last warning, you fucking gay - I won't appreciate if any more nazi shwain would write in my page! I don't wish to talk to you anymore!\n\nBeware of the Dark Side!"

Toxic: 1
Severe_toxic: 0
Obscene: 0
Threat: 0
Insult: 0
Identity_hate: 0

"Bye! \n\nDon't look, come or think of comming back! Tosser."

Korean Corpus of Online News Comments for Toxic Speech Detection

- 9,381 manually labeled entertainment news comments for identifying Korean toxic speech
- Collected from a widely used online news platform in Korea.
- Dataset description:
    - labeled: train.tsv (914KB), dev.tsv (55KB)
    - test.no_label.tsv (96KB)
    - unlabeled (237.1MB): 레이블이 없는 댓글 2,033,893개
    - news_title: 댓글을 가져온 기사의 제목

https://github.com/kocohub/korean-hate-speech

- Labeled data

  - train.tsv: 7,896 data / 66,321 tokens / 37,723 unique tokens

	comments	contain_gender_bias	bias	hate
0	(현재 호텔주인 심정) 아18 난 마른하늘에 날벼락맞고 호텔망하게생겼는데 누군 계속...	False	others	hate
1	....한국적인 미인의 대표적인 분...너무나 곱고아름다운모습...그모습뒤의 슬픔을...	False	none	none
2	...못된 넘들...남의 고통을 즐겼던 넘들..이젠 마땅한 처벌을 받아야지..,그래...	False	none	hate
3	1,2화 어설펐는데 3,4화 지나서부터는 갈수록 너무 재밌던데	False	none	none
4	1. 사람 얼굴 손톱으로 긁은것은 인격살해이고2. 동영상이 몰카냐? 메걸리안들 생각...	True	gender	hate

→ contain_gender_bias (False: 6,664, True: 1,232)
bias (none: 5,148, others: 1,516, gender: 1,232)
hate (none:3,486, offensive: 2,499, hate: 1,911)

  - dev.tsv: 471 data / 4,012 tokens / 3,276 unique tokens

→ contain_gender_bias (False: 404, True: 67)
bias (none: 342, others: 62, gender: 67)
hate (none: 160, offensive: 189, hate: 122)

- test.no_label.tsv (96KB)
    - 974 data
    - 8,332 tokens / 6,227 unique tokens

- Unlabeled (237.1MB)
    - 2,033,893 comments
    - 20,013,949 tokens / 4,051,695 unique tokens

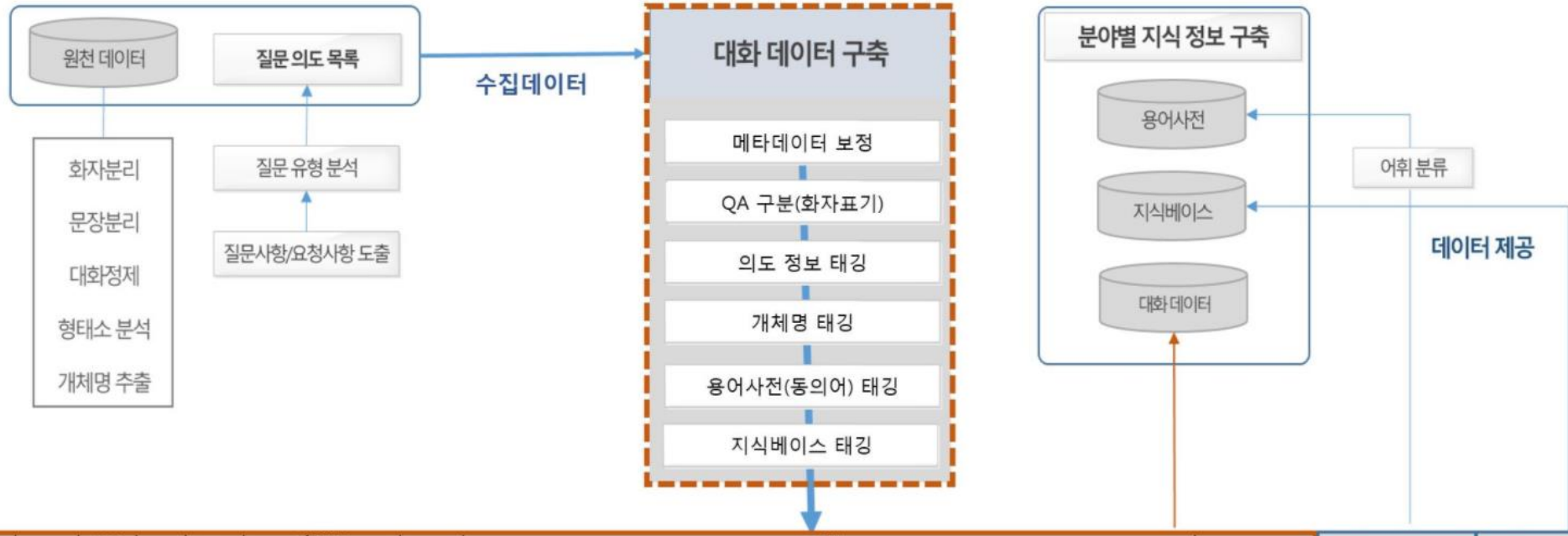
사용자 의도파악 데이터셋 (한국어)
http://www.aihub.or.kr/aidata/85

- 소상공인 및 공공민원 10개 분야에 대한 1만건 이상의 대화 (Dialog) 데이터 구축

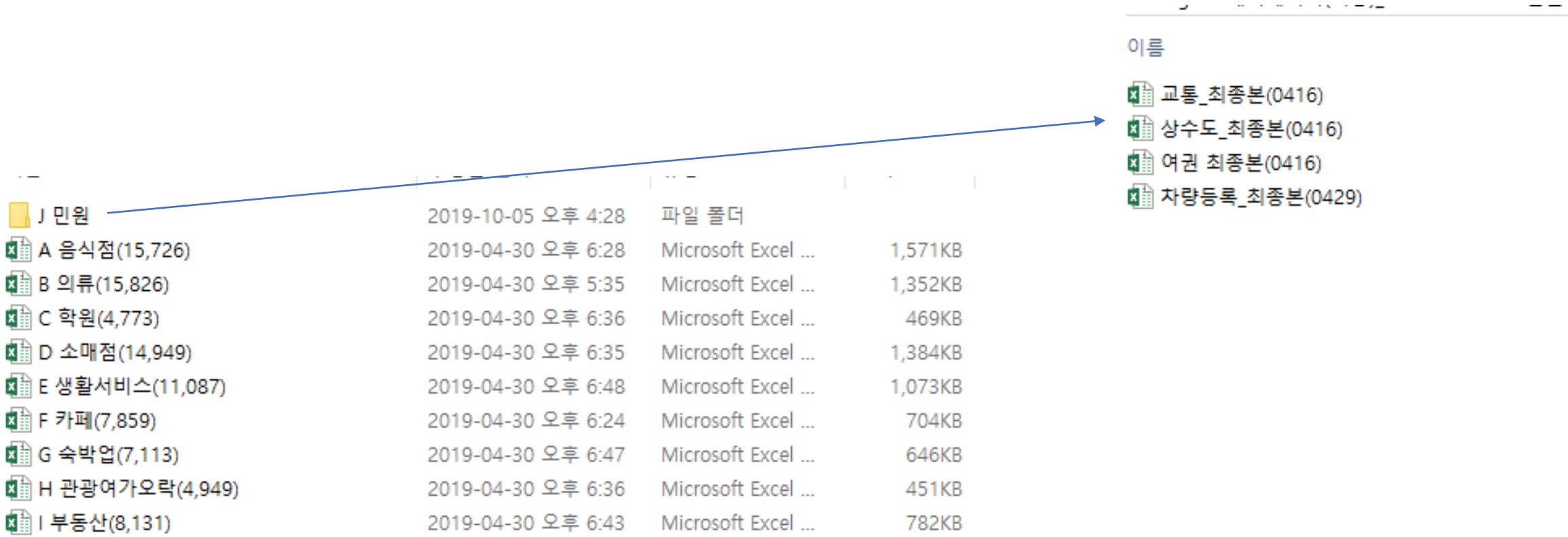
데이터 구조

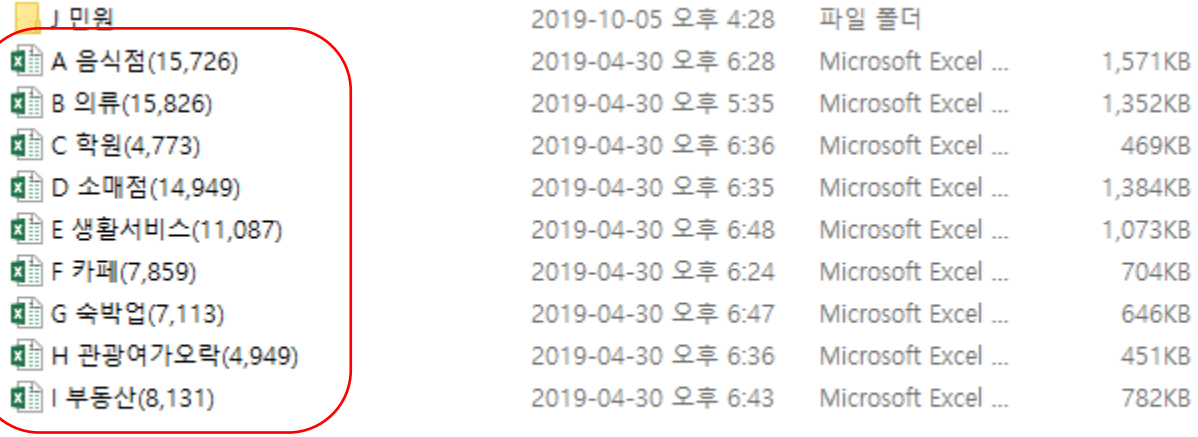
- 사용자의 입력의 의도를 파악하여 그에 따라 시스템이 응답 또는 행동을 하는 대화 생성/관리로 구현
- 대화 데이터의 기본 구조는 Q&A(질의/응답)로 구성되며, 화자가 분리된 각각의 질문(Q)에 대하여 메인 의도(Main Intent)와 서브 의도(Sub Intent)로 구분하여 의도 정보 태깅
- 대화 데이터의 각 문장은 사용자(손님) 질문(Main Question), 메인 질문에 추가적으로 필요한 시스템(점원)의 서브 질문(Sub Question), 서브 질문에 대한 사용자(손님) 응답(User Answer), 시스템(점원) 최종 응답(System Answer)로 구분
- 각 문장에서 고유명사와 복합명사, 수식표현 등 사용자 의도가 반영된 개체(Entity)를 추출하여 시소러스 및 소상공인, 공공민원 분야에 대한 지식정보 구축

한국어 대화데이터 분야 구조 이미지 예시



도메인	카테고리	대화셋 일련번호	화자	문장 번호	의도정보		QA	원본				개체명	용어사전	지식베이스
					main	sub1		main Q	sub Q	User Answer	System Answer(Response)			
카페	카페	1	1	1	일반주문		Q	아메 한 잔 주세요				아메, 한, 잔	아메/아메리카노, 한/하나	아메리카노/종류, 하나/수량
			0	2	일반주문	온도	Q		따뜻한 걸로 드릴까요?			따뜻	따뜻/핫	핫/온도
			1	3			A			아이스로 주세요		아이스		아이스/온도
			0	4	일반주문	사이즈	Q		사이즈는 어떻게 하시겠어요?			사이즈		사이즈/사이즈
			1	5			A			레귤러로 주세요		레귤러		레귤러/사이즈
			1	6	가격		Q	얼마예요?				얼마		
			0	7			A				4500원입니다.	4500		4500/가격
			0	8	결제		Q	결제는 어떻게 하시겠어요?				결제		
			1	9			A			현금이요		현금		
			0	10	결제	현금영수증	Q		현금영수증 해드릴까요			현금영수증		
			1	11			A			네		네		
			0	12	결제	현금영수증	Q		현금영수증 번호 어떻게 되나요?			현금영수증, 번호		
			1	13			A			010-1234-1234입니다		010-1234-1234		


이름
교통_최종본(0416)
상수도_최종본(0416)
여권 최종본(0416)
차량등록_최종본(0429)
J 민원 2019-10-05 오후 4:28 파일 폴더
A 음식점(15,726) 2019-04-30 오후 6:28 Microsoft Excel ... 1,571KB
B 의류(15,826) 2019-04-30 오후 5:35 Microsoft Excel ... 1,352KB
C 학원(4,773) 2019-04-30 오후 6:36 Microsoft Excel ... 469KB
D 소매점(14,949) 2019-04-30 오후 6:35 Microsoft Excel ... 1,384KB
E 생활서비스(11,087) 2019-04-30 오후 6:48 Microsoft Excel ... 1,073KB
F 카페(7,859) 2019-04-30 오후 6:24 Microsoft Excel ... 704KB
G 숙박업(7,113) 2019-04-30 오후 6:47 Microsoft Excel ... 646KB
H 관광여가오락(4,949) 2019-04-30 오후 6:36 Microsoft Excel ... 451KB
I 부동산(8,131) 2019-04-30 오후 6:43 Microsoft Excel ... 782KB



J 민원	2019-10-05 오후 4:28	파일 폴더	
A 음식점(15,726)	2019-04-30 오후 6:28	Microsoft Excel ...	1,571KB
B 의류(15,826)	2019-04-30 오후 5:35	Microsoft Excel ...	1,352KB
C 학원(4,773)	2019-04-30 오후 6:36	Microsoft Excel ...	469KB
D 소매점(14,949)	2019-04-30 오후 6:35	Microsoft Excel ...	1,384KB
E 생활서비스(11,087)	2019-04-30 오후 6:48	Microsoft Excel ...	1,073KB
F 카페(7,859)	2019-04-30 오후 6:24	Microsoft Excel ...	704KB
G 숙박업(7,113)	2019-04-30 오후 6:47	Microsoft Excel ...	646KB
H 관광여가오락(4,949)	2019-04-30 오후 6:36	Microsoft Excel ...	451KB
I 부동산(8,131)	2019-04-30 오후 6:43	Microsoft Excel ...	782KB

	A	B	C	D	E	F	G	H	I	J
1	SPEAKER	SENTENCE	DATAID	DOMAINID	DOMAIN	CATEGORY	SPEAKE	SENTEN	MAIN	SUB
2	고객	애가 고등학교 1학년인데 태권도 하려면 수강료가 얼마쯤?	장미연_태권도_1	C	학원	태권도	1	1	교육비문의	
3	점원	12만 원입니다	장미연_태권도_1	C	학원	태권도	0	2	교육비문의	
4	점원	뭐 때문에 하시려는데요?	장미연_태권도_1	C	학원	태권도	0	3	상담문의	수강목적
5	고객	애가 겁도 많고 그래서 태권도 한 번 시켜보려고요	장미연_태권도_1	C	학원	태권도	1	4	상담문의	수강목적
6	고객	고1이 몇 명쯤 있나요?	장미연_태권도_1	C	학원	태권도	1	5	수강생구성문의	
7	점원	대학교 올라가는 친구도 있습니다	장미연_태권도_1	C	학원	태권도	0	6	수강생구성문의	
8	고객	저녁 수업 시간이?	장미연_태권도_1	C	학원	태권도	1	7	강의시간문의	
9	점원	8시에서 9시에요	장미연_태권도_1	C	학원	태권도	0	8	강의시간문의	
10	점원	학생이 8시에 시간이 돼요?	장미연_태권도_1	C	학원	태권도	0	9	상담문의	수강가능시간
11	고객	내면 돼요	장미연_태권도_1	C	학원	태권도	1	10	상담문의	수강가능시간
12	고객	토 일은요?	장미연_태권도_1	C	학원	태권도	1	11	주말프로그램문의	
13	점원	토 일은 특강으로 수업이 진행되서 따로 돈을 받고 하고 있어요	장미연_태권도_1	C	학원	태권도	0	12	주말프로그램문의	
14	고객	특강료 얼마쯤해요?	장미연_태권도_1	C	학원	태권도	1	13	교육비문의	
15	점원	1월 2월 방학때는 안 하고요 초등학교 저학년 대상이에요	장미연_태권도_1	C	학원	태권도	0	14	교육비문의	
16	고객	매일 와야 돼요?	장미연_태권도_1	C	학원	태권도	1	15	수업일수문의	
17	점원	운동을 시키실 생각 있으시면 매일 오는 게 적응하기가 편해요	장미연_태권도_1	C	학원	태권도	0	16	수업일수문의	
18	고객	태권도복은 별도로 사야돼요?	장미연_태권도_1	C	학원	태권도	1	17	교육비에도복비포함문의	
19	점원	네 그건 별도에요	장미연_태권도_1	C	학원	태권도	0	18	교육비에도복비포함문의	
20	점원	운동을 한 적은 없나요?	장미연_태권도_1	C	학원	태권도	0	19	상담문의	운동경험문의
21	고객	없어요 그래서 애가 너무 비만이에요	장미연_태권도_1	C	학원	태권도	1	20	상담문의	운동경험문의
22	점원	사시는 곳은 어디신가요?	장미연_태권도_1	C	학원	태권도	0	21	상담문의	학생위치
23	고객	상인동 여기 살아요	장미연_태권도_1	C	학원	태권도	1	22	상담문의	학생위치
24	점원	차량은 안 해도 되고요?	장미연_태권도_1	C	학원	태권도	0	23	학원차량가능문의	차량이용문의
25	고객	네 당연하죠. 걸어서 오면	장미연_태권도_1	C	학원	태권도	1	24	학원차량가능문의	차량이용문의
26	점원	학생이 한 번 해보고 결정하시는 건 어때요?	장미연_태권도_1	C	학원	태권도	0	25	테스트프로그램가능여부문의	체험권유
27	고객	그게 좋겠네요	장미연_태권도_1	C	학원	태권도	1	26	테스트프로그램가능여부문의	체험권유

이름

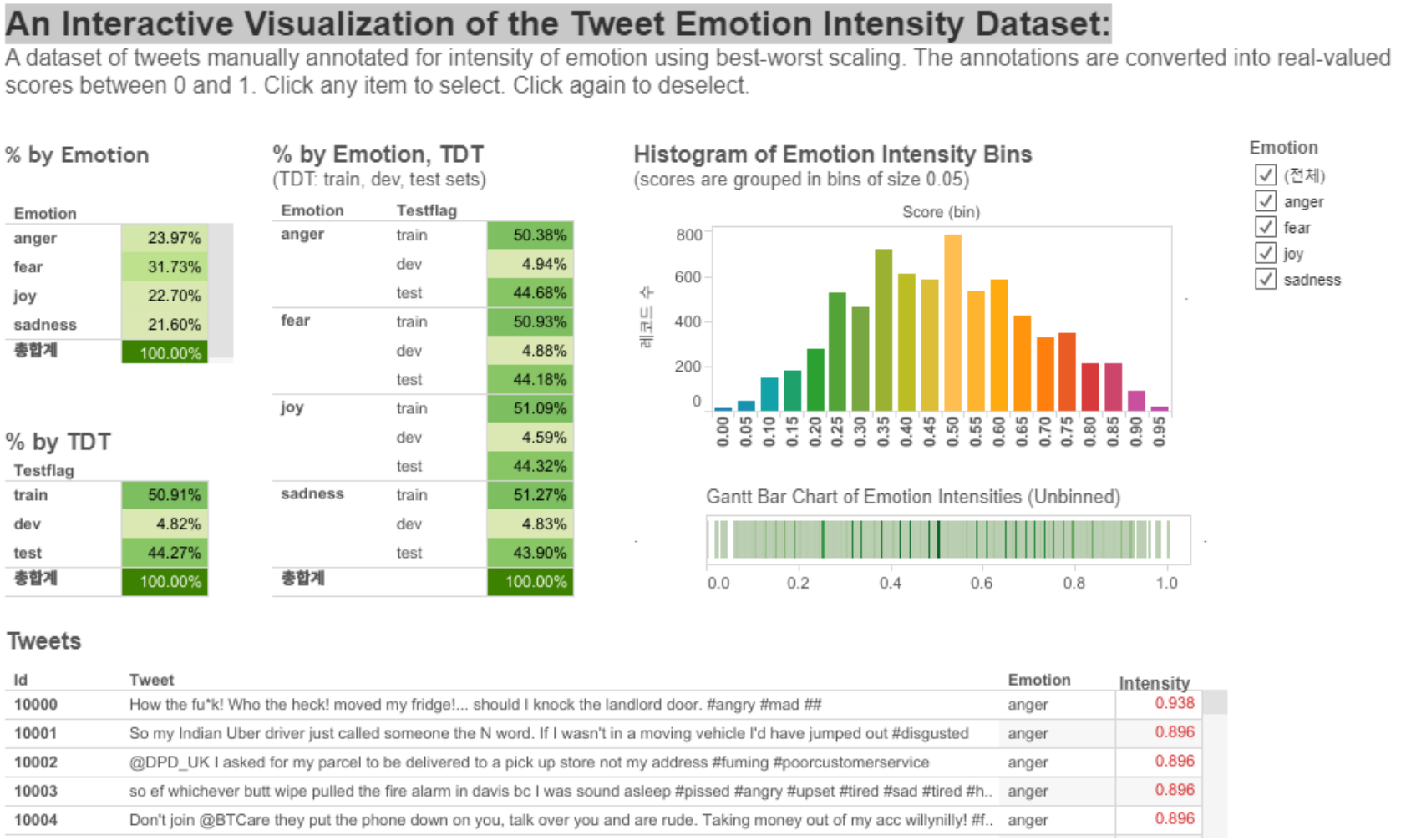
- 교통_최종본(0416)
- 상수도_최종본(0416)
- 여권 최종본(0416)
- 차량등록_최종본(0429)

도메인	화자	intent	subintent	question	answer	q_entity	a_entity	동의어	지식베이스	
46	민원인	요금 문의		수도 요금 이거 낼려고 그러는데요	그러면 청구된 금액 확인해 보니까요 선생님 지금 구월달 십일월달 지금 두 달치 지금 두 번의 청구금액이 납부가 되지 않은 상태라 총 납부하실 금액 팔만구천백십원인데 제가 금액하고 입금하실 계좌번호 문자로 하나 넣어 드릴까요	수도 요금,	청구, 금액, 확인, 구월달, 십일월달, 두 달, 두 번, 청	수도 요금/수도 요금,		13
46	상담사	요금 문의	관리번호	아 수도 요금이요 제가 요금 확인 도와드릴텐데요 선생님 혹시 고지서 상에 관리번호 확인 가능하십니까?	이공공사 영영삼하나 육육이요	수도 요금, 요금 확인, 고지서, 관리번호, 확인, 가능,	이공공사 영영삼하나 육육,	수도 요금/	이공공사 영	13
46	상담사	요금 문의	이름	선생님 고지서 상에 성함이 어떻게 되어 있습니까?	아 백일호로 되어 있는데요	고지서, 성함,	백일호,	고지서/상	백일호/이	13
46	민원인	요금 문의		상수도 요금 못 냈는데 이게 가상계좌로 지금 못 넣잖아요 날짜가 지나면	아 가상계좌로 입금은 가능하세요	상수도 요금, 가상계좌, 날짜,	가상계좌, 입금, 가능,	상수도 요금/수도 요금		13
46	상담사	요금 문의	관리번호	혹시 관리번호 확인 가능할까요	이영영 팔영영영 일칠칠	관리번호, 확인, 가능,	이영영 팔영영영 일칠칠,	관리번호/	이영영 팔	13
46	상담사	요금 문의	이름	명의자가 혹시 어떻게 되어 있으십니까?	이성훈입니다.	명의자,	이성훈,	명의자/명	이성훈/이	13

WASSA-2017 Shared Task on Emotion Intensity

- http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html

- An Interactive Visualization of the Tweet Emotion Intensity Dataset:



An Interactive Visualization of the Tweet Emotion Intensity Dataset:

A dataset of tweets manually annotated for intensity of emotion using best-worst scaling. The annotations are converted into real-valued scores between 0 and 1. Click any item to select. Click again to deselect.

% by Emotion

Emotion	
anger	23.97%
fear	31.73%
joy	22.70%
sadness	21.60%
총합계	100.00%

% by TDT

Testflag	
train	50.91%
dev	4.82%
test	44.27%
총합계	100.00%

% by Emotion, TDT

(TDT: train, dev, test sets)

Emotion	Testflag	
anger	train	50.38%
	dev	4.94%
	test	44.68%
fear	train	50.93%
	dev	4.88%
	test	44.18%
joy	train	51.09%
	dev	4.59%
	test	44.32%
sadness	train	51.27%
	dev	4.83%
	test	43.90%
총합계		100.00%

Histogram of Emotion Intensity Bins

(scores are grouped in bins of size 0.05)

Gantt Bar Chart of Emotion Intensities (Unbinned)

Emotion
- (전체)
- anger
- fear
- joy
- sadness

Tweets

Id	Tweet	Emotion	Intensity
10000	How the fu*k! Who the heck! moved my fridge!... should I knock the landlord door. #angry #mad ##	anger	0.938
10001	So my Indian Uber driver just called someone the N word. If I wasn't in a moving vehicle I'd have jumped out #disgusted	anger	0.896
10002	@DPD_UK I asked for my parcel to be delivered to a pick up store not my address #fuming #poorcustomerservice	anger	0.896
10003	so ef whichever butt wipe pulled the fire alarm in davis bc I was sound asleep #pissed #angry #upset #tired #sad #tired #h..	anger	0.896
10004	Don't join @BTCare they put the phone down on you, talk over you and are rude. Taking money out of my acc willynilly! #f..	anger	0.896

http://saifmohammad.com/WebPages/TweetEmotionIntensity-dataviz.html

관련 도서

Emotion detection from text and speech: a survey

- https://link.springer.com/article/10.1007/s13278-018-0505-2

Table 13 Datasets

From: Emotion detection from text and speech: a survey

Dataset	Datasize	Description
EmoBank[a]	10K sentences	Double annotation with valence, arousal and dominance were used from the perspectives of both writer and reader
The valence and arousal Facebook posts (Preotiuc-Pietro et al. 2016)	2895 Facebook posts	Double annotation with valence and arousal values
The emotion in text data set[b]	40,000 Tweets	Annotated with Anger, Boredom, Empty, Enthusiasm, Fun, Happiness, Hate, Love, Relief, Sadness, Surprise, Worry, Neutral
EmoLex (Mohammad and Bravo-Marquez 2017)	14,182 Unigrams	Annotated with sentiments—Negative, Positive and Emotions—Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust
ISEAR[c]	7666 sentences	Contains responses of questionnaires on seven emotions (joy, fear, anger, sadness, disgust, shame, and guilt) from 37 countries from 5 continents
Affective Text (Strapparava and Mihalcea 2008)	1200 News Headlines	Annotated with 6 basic emotions from Ekman's model and polarity

[a] JULIELab. Emobank

[b] CrowdFlower

[c] AAAC Emotion Research

맺음말: 감정분석 말뭉치 다운로드

- Stanford Sentiment Treebank
  - 영어 원본 -- https://nlp.stanford.edu/sentiment/index.html
  - 한글 번역 -- http://nlp.kookmin.ac.kr/corpus/Sentiment_Treebank_en_ko.zip

- 6가지 감정 말뭉치 데이터셋
  - 원본 데이터 -- https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp
  - 한국어 번역 -- http://nlp.kookmin.ac.kr/corpus/emotion_korTran.zip

- 여러 개의 감정 말뭉치 통합(영어)
  https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/unifyemotion/

- WASSA-2017 Shared Task on Emotion Intensity (4가지 감정: anger, fear, joy, sadness)
  http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html#test

- 유해성 말뭉치 데이터셋: kaggle.com
  https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
- 한국어 Toxic Speech Detection 말뭉치
  https://github.com/kocohub/korean-hate-speech
- 사용자 의도파악 데이터셋 (한국어) -- http://www.aihub.or.kr/aidata/85
- 네이버 영화평 -- https://github.com/e9t/nsmc
  - "자동 띄어쓰기" 및 "토큰화(형태소 분석)" 버전 다운로드
    http://nlp.kookmin.ac.kr/corpus/NaverSentiMovieCorpus_ASP.zip
  - 20만+8만 확장
    http://nlp.kookmin.ac.kr/corpus/NSMC_extended_282K.zip

[참고] NLP 관련 자료

- https://cafe.naver.com/nlpk -- NLP, 딥러닝 강의자료 등
- https://cafe.naver.com/nlpkang -- KLT2000 형태소 분석기 등