



국내 최초 Text 기반 다중 감정 분류기 MR2SentiNet

페르소나 시스템

2019.11.27

진명훈 김호현

목차 Index



Ch01 문제 정의

- 프로젝트 / 개발 개요
- 요구사항 분석
- 감정 범주 정의

Ch02 관련 연구

- 기존 극성 감정 분석
- Multi-Modal 감정 분석
- Differentiation

Ch03 활용 데이터

- NSMC + Crawling
- 감정 단어 사전
- Text 전처리 필요성

Ch04 M2SNet

- 전체 Process 시각화
- Preprocess line
- Labeling line
- Modeling line
- Predict line



Ch05 Version 0.0

- Labeling Detail
- 가설 설정
- 실험 및 검증

Ch06 Version 0.1

- Update 내역

Ch07 결론 및 향후 방향성

- 문제 분석 - Labeling
- 활용 서비스 제안
- 결론 및 향후 방향성

Appendix.

Reference.

Ch01 문제 정의

- 프로젝트 / 개발 개요
- 요구사항 분석
- 감정 범주 정의

01. 문제 정의

프로젝트 / 개발 개요

- ✓ 프로젝트 명

• 희로애락 다중 감정 분류 시스템 구축
- ✓ 프로젝트 기간

• ‘19.09.10 ~ ‘19.11.25
- ✓ 과업배경 및 목표

• IITP, KSA 주관 국비 사업 프로젝트

• 산업 특화형 인공지능 인재 배양을 위한 현장형 프로그램

• 협력 기업 연계 프로젝트 실습을 통해 실제 적용 역량 배양
- ✓ 수행조직 및 일정

• 팀장: 진명훈

• 팀원: 김호현

	2019년		
	9월	10월	11월
관련 연구 탐색	<div></div>		
DB 및 감정 사전 탐색		<div></div>	
데이터 라벨링(JST, 사전)		<div></div>	
분류 모델 구축			<div></div>
성능 향상 및 보고서 작성			<div></div>
모듈화 및 결과 발표			<div></div>

01. 문제 정의

프로젝트 / 개발 개요

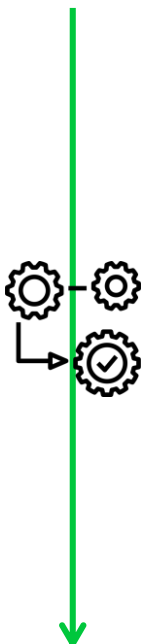
< 히로애락 다중 감정 분류 시스템 구축 >



요구 사항

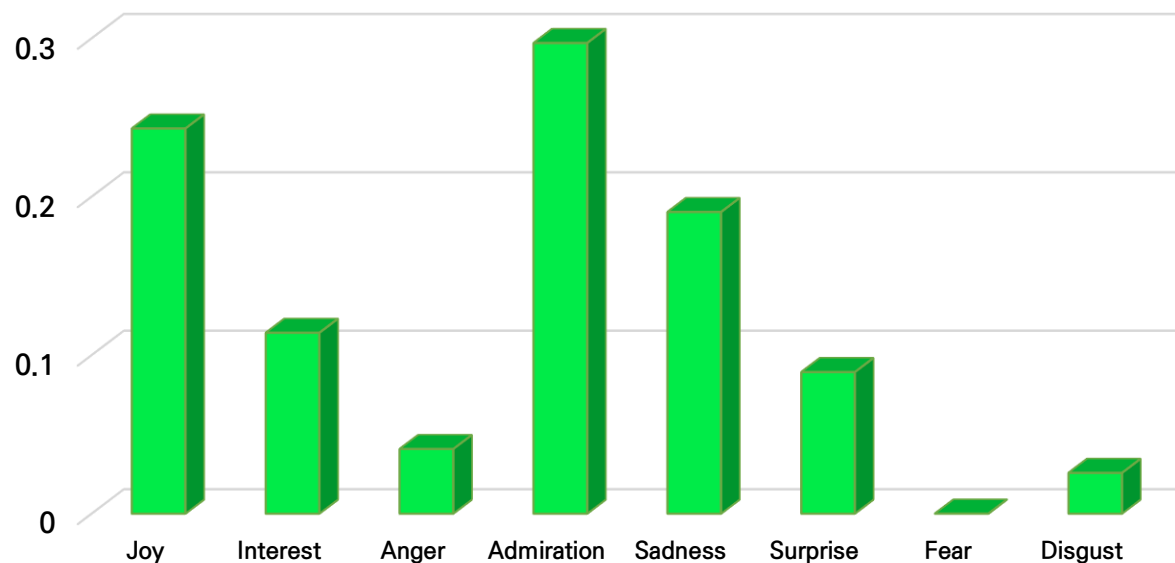
- 텍스트가 Input으로 들어오면 **특정 감정일 확률**을 Output으로 받는 시스템 구축

Input:



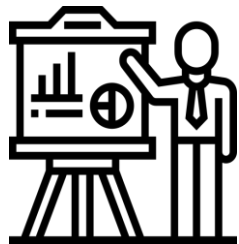
Output:

“IITP, KSA에서 주관한 인공지능 산업형인재 양성 사업 덕분에
더욱 성장할 수 있었습니다. 감사합니다!”



01. 문제 정의

요구사항 분석



- 앞선 요구 사항에 대해 다음의 6가지 해결 과제를 선정
- 3개월 간 다음의 과제를 수행하기 위한 process를 수립 및 실시

“

2D-SLAP

”



Sentiment

- ✓ 감정의 범주에 대한 정의를 어떻게 내릴 것인가?



Differentiation

- ✓ 기존 연구와의 차별성은?



Data

- ✓ 어떤 데이터로 분석할 것인가?



Labeling

- ✓ Label이 없는 데이터로 어떻게 학습할 것인가?



Algorithms

- ✓ 확률 값을 얻기 위해 어떤 방식을 사용할 것인가?

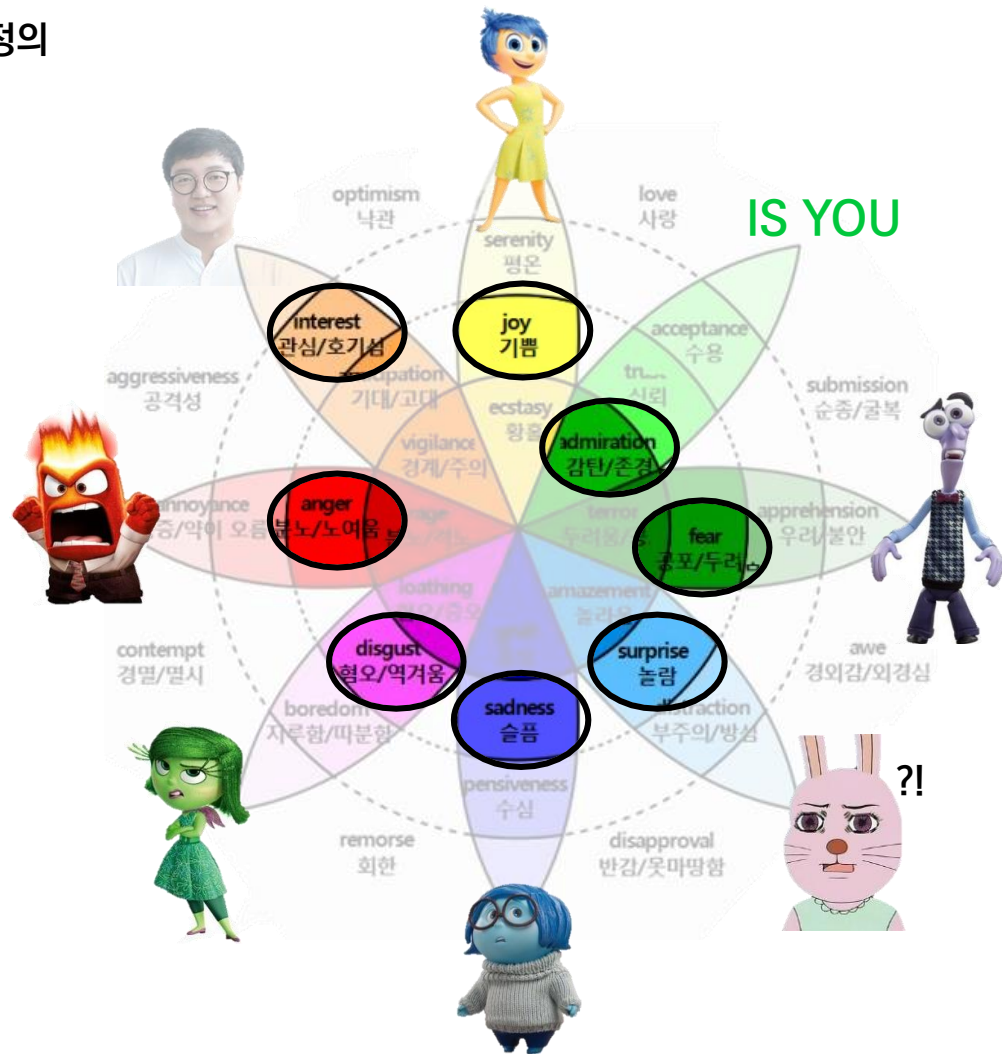


Preprocessing

- ✓ Text 데이터 전처리 및 임베딩을 어떻게 실시할 것인가?

01. 문제 정의

감정 범주 정의



😊 Sentiment

Joy	Sadness
Interest	Surprise
Anger	Fear
Admiration	Disgust

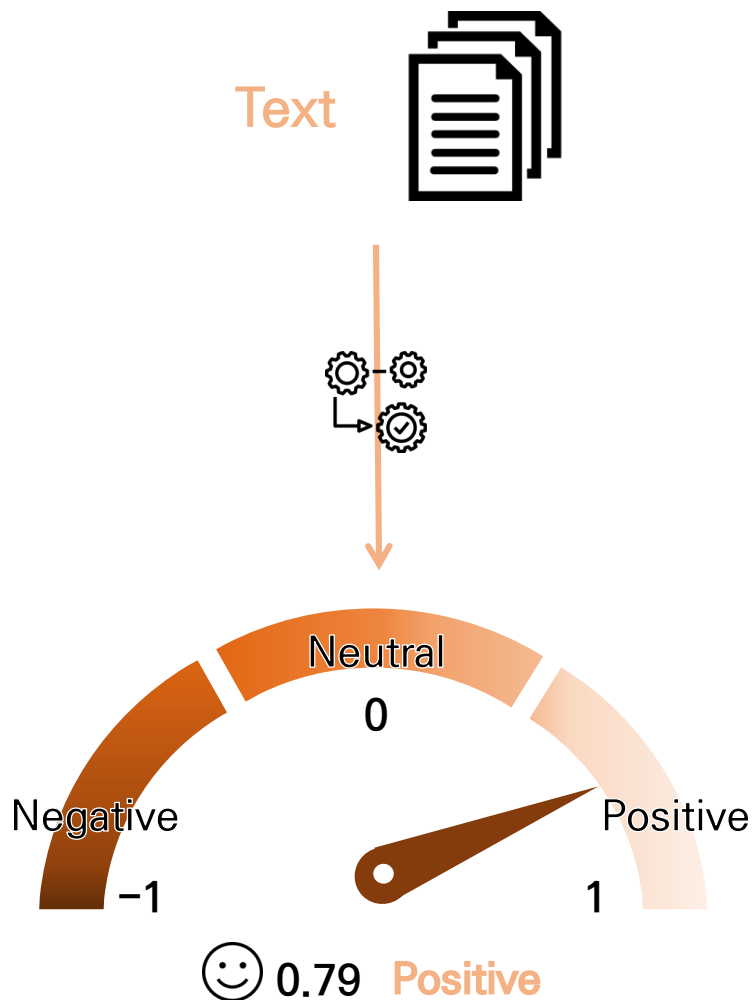
〈 Robert Plutchik's Wheel of Emotions 〉

Ch02 관련 연구

- 기존 극성 감정 분석
- Multi-Modal 감정 분석
- Differentiation

02. 관련 연구

기존 극성 감정 분석



평점 기반 Labeling

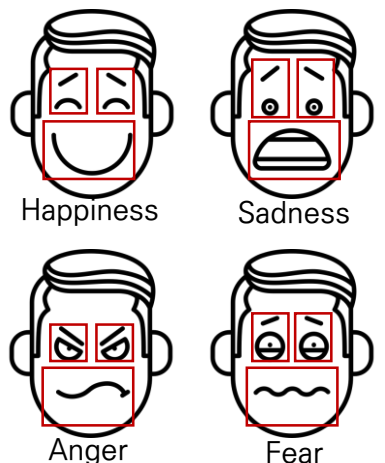
페르소나 시스템: 히로애락 감정 분류 시스템 평점 리스트

감상평	별점	평점
대박이다.. 발표를 듣고 눈물 흘린건 처음...	★★★★★	9.08
감정 사전틀 기반으로 군집화 실시... 크~	★★★★★	10.0
6개월 간의 시간을 보여줬던 발표였습니다	★★★★★	8.76
흠... 이런 점은 조금 아쉬웠던 거 같아요!	★★★★	6.03
개선 사항들 수행하면 대박이겠는걸요?!	★★★★★	7.98

⇒ 단순 극성 분류로는 한계가 존재

02. 관련 연구

Multi-Modal 감정 분석



Happiness

Sadness

Anger

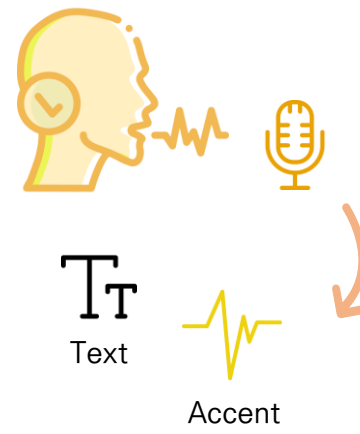
Fear

Emotion Detection



- ✓ 경쾌함
- ✓ 다급한
- ✓ 늘어지는
- ✓ 질질끄는
- ✓ etc

Step Tempo
Taylor Transformation



Text

Accent

Voice Recognition

Multi-Modal
Sentiment
Analysis

Another Methodology

추가적인 데이터 필요

02. 관련 연구

Multi-Modal 감정 분석

Input Text에 대해 Output으로 특정 감정일 확률을 받는 시스템 구축

요구사항 :

- ✓ 경쾌함
- ✓ 다급한
- ✓ 늘어지는
- ✓ 질질끄는

제한사항 :

a-z , 한정적인 자원, 그리고 시간

추가적인 데이터 필요

02. 관련 연구



Differentiation

JST 감정 토픽 모델과 확률적 분류 모델을 활용,

“감정 군집화 및 예측” & “감정에 맞는 대응책 제공”

Ch03 활용 데이터

- NSMC + Crawling
- 감정 단어 사전
- Text 전처리 필요성

03. 활용 데이터



Data NSMC & KSenticNet

NAVER

sentiment movie corpus

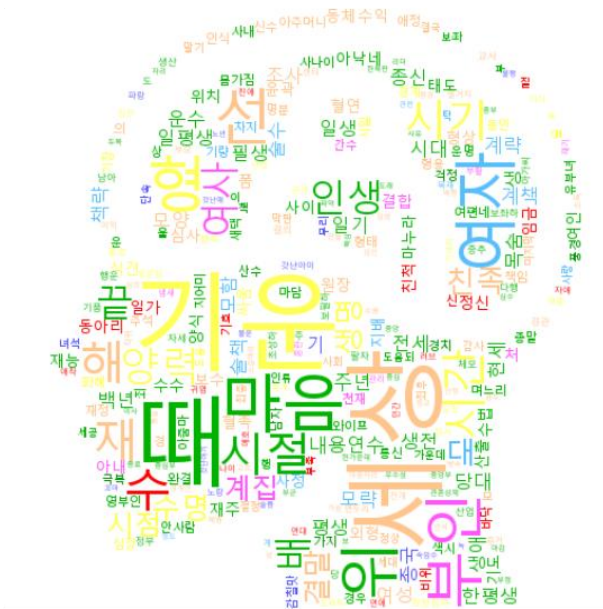
- ✓ NSMC + Crawling
- ✓ 2,981,222 texts

Index	review
0	전체관람가는 아닌것 같아요
1	디렉터스컷으로봐서 거의 3시간짜리인데 참 흡인력있다
2	태어나 처음으로 가슴아리는 영화였다. 20년이상 지났지만.. 생각하면 또 가슴이...
3	어린시절 고딩때 봤던 때랑 또 결혼하고 나서 봤을때의 느낌은 확실히 다르네요. 뭔가...
4	토토에게 넓은 세상을 보여주고픈 알프레도.. 그가 토토를 위해 정을 떼려고 했던 장...
5	인생 최고의 영화. 말이 필요없음. 감독판은 감동이 좀 덜함.
6	아름다운 영화 지금까지 봤던 영화 중 끝까지 감동적이었던 영화
...	...
2981222	이 영화에서 나의 향수를 느꼈다. 알베르토와 토토가 함께 자전거를 타며 배경음악이 ...

KSenticNet

Korean Sentiment Dictionary by KAIST

- ✓ 5,465 개의 단어, 어근에 대한 8가지 감정 값 존재
- ✓ JST의 사전 확률 조작에 사용



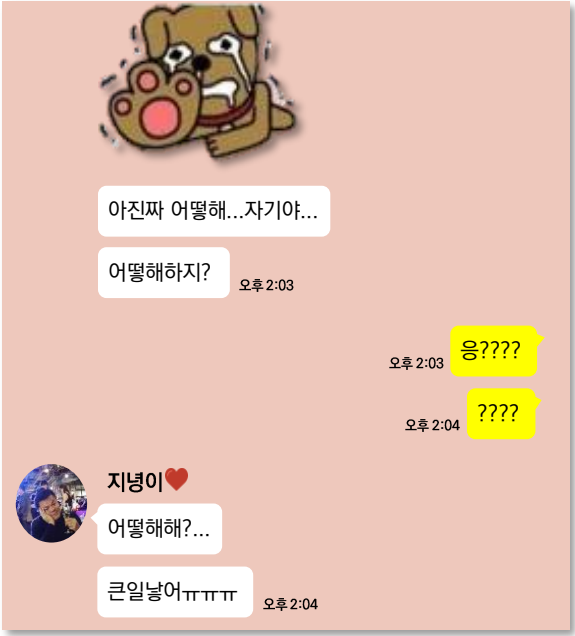
03. 활용 데이터

Text 전처리 필요성

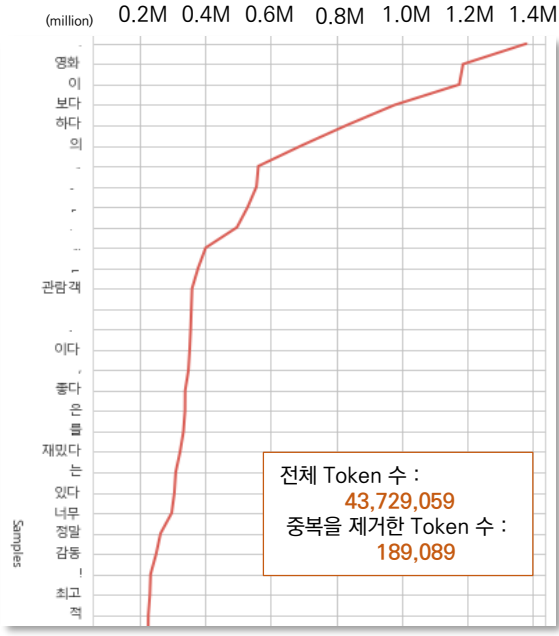
✓ 띄어쓰기



✓ 맞춤법



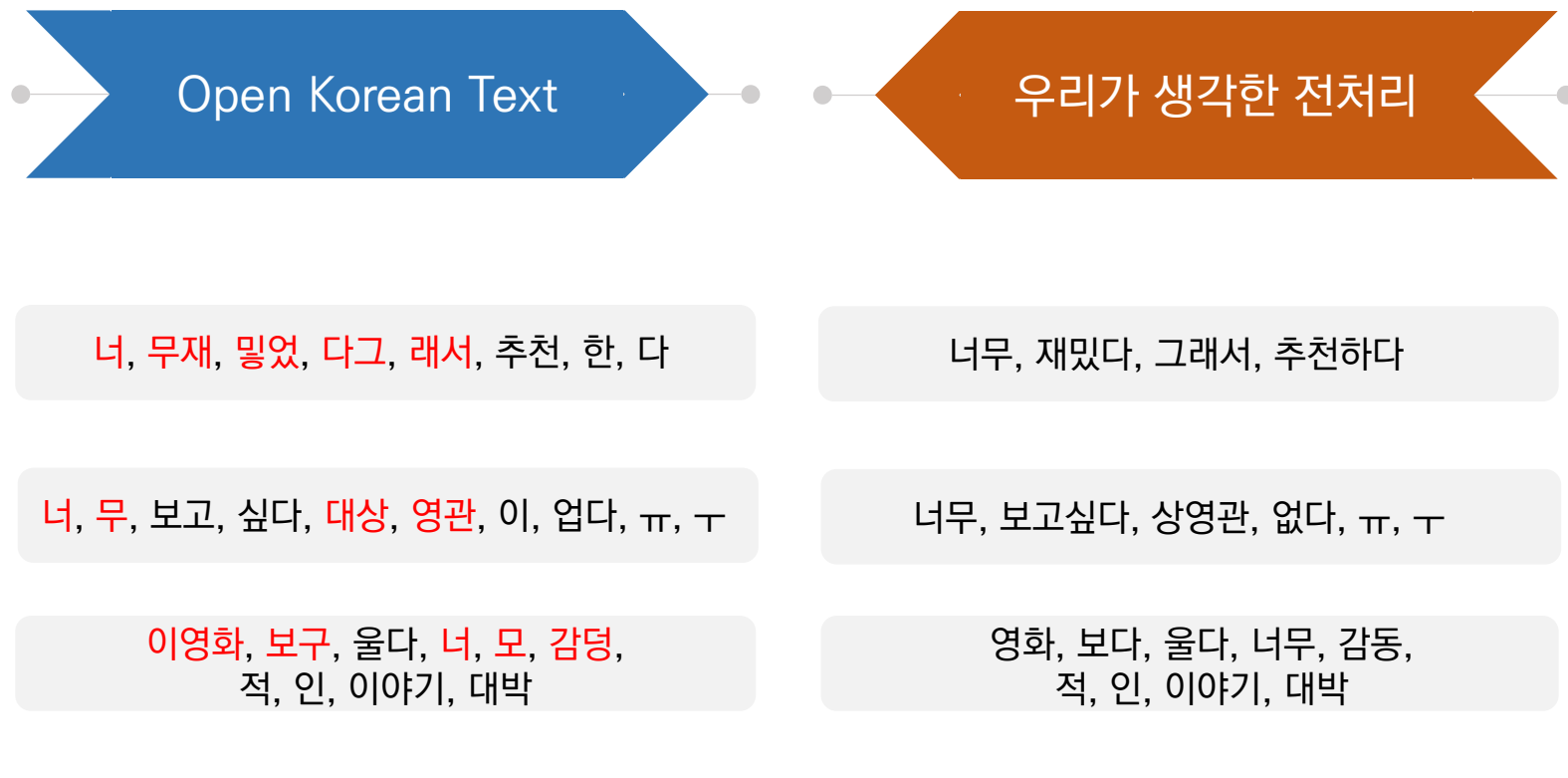
✓ 불용어



군집화 / 분류 성능을 저하시키는 요인들

03. 활용 데이터

Text 전처리 필요성



분석을 위해 오른쪽과 같이 전처리를 해줘야 함

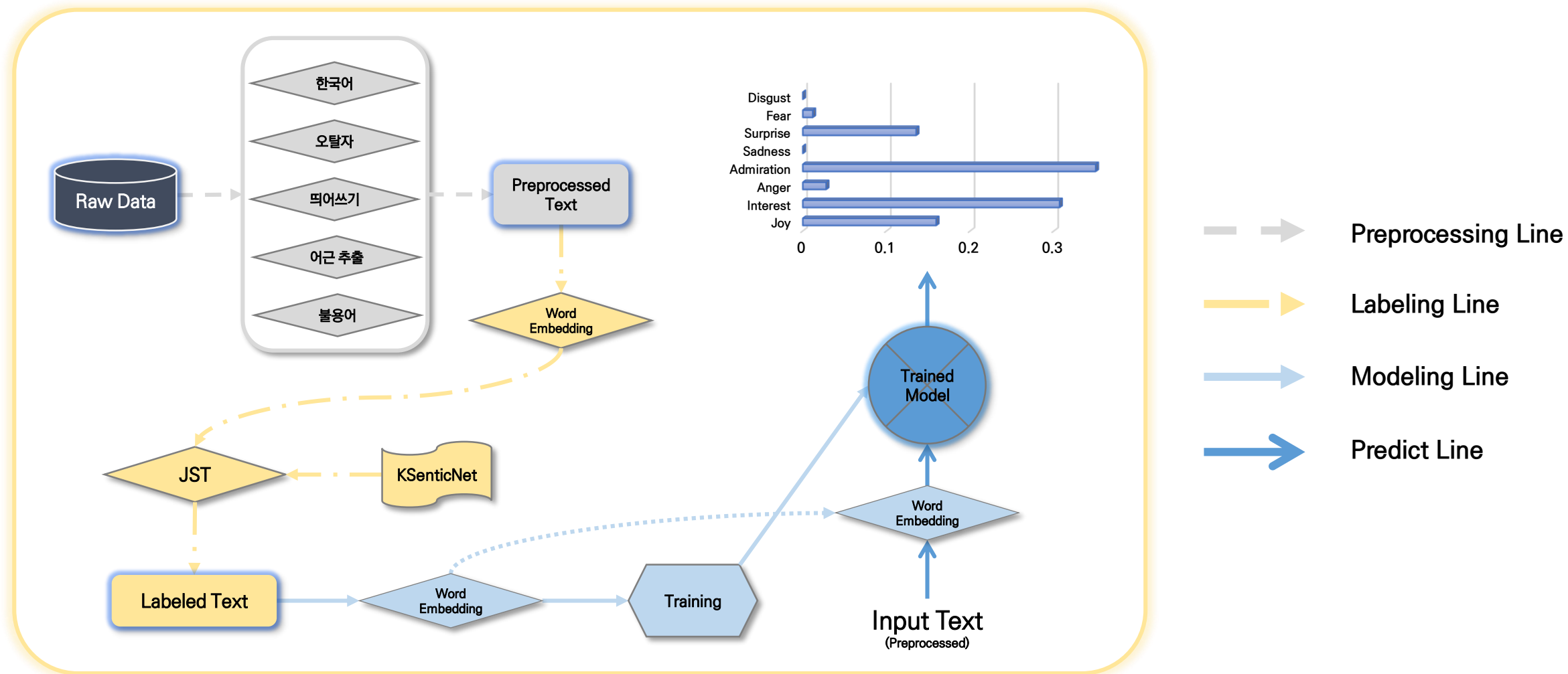
Ch04 M2SNet

- 전체 Process 시각화
- Preprocess line
- Labeling line
- Modeling line
- Predict line

04. M2SNet

전체 Process 시각화

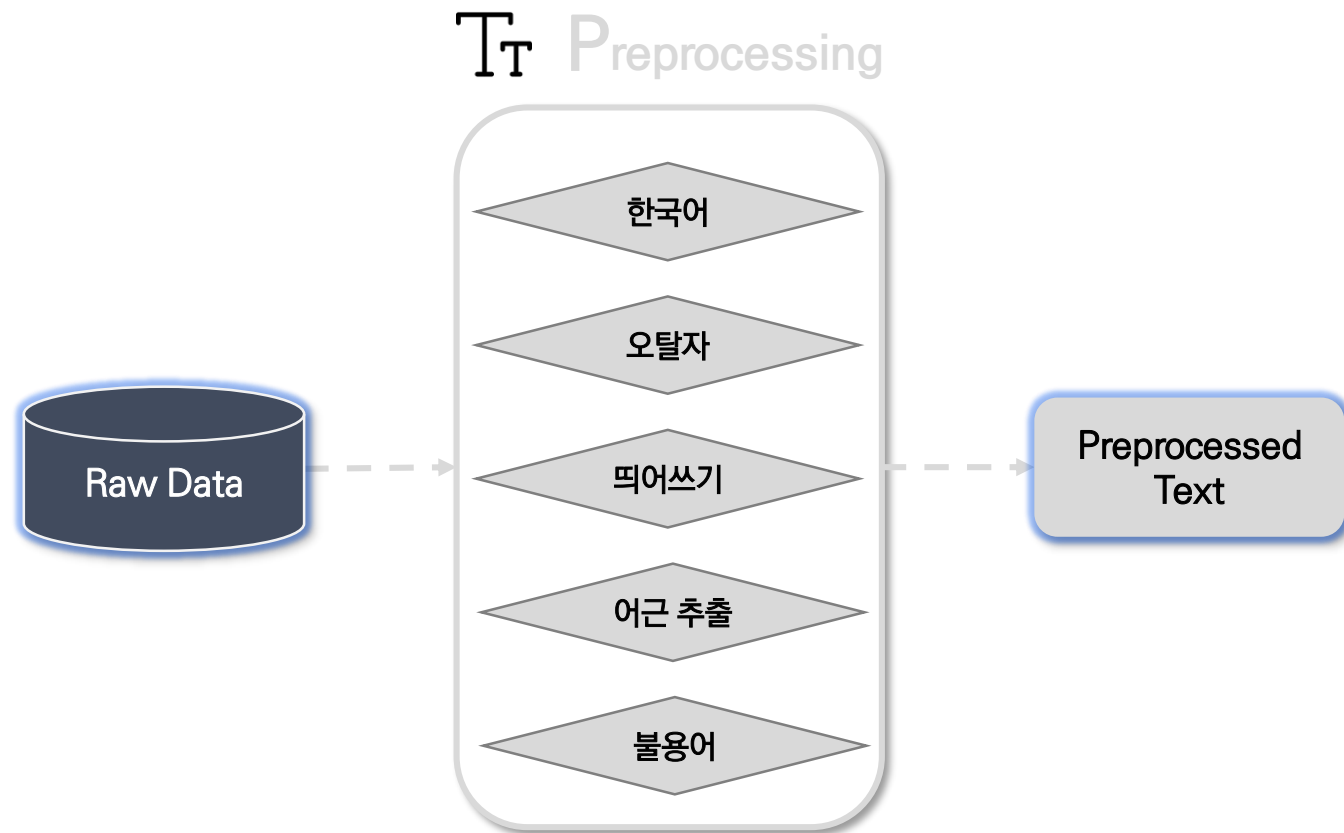
〈 MR2SentiNet 구조도 〉



04. M2SNet

Preprocess line

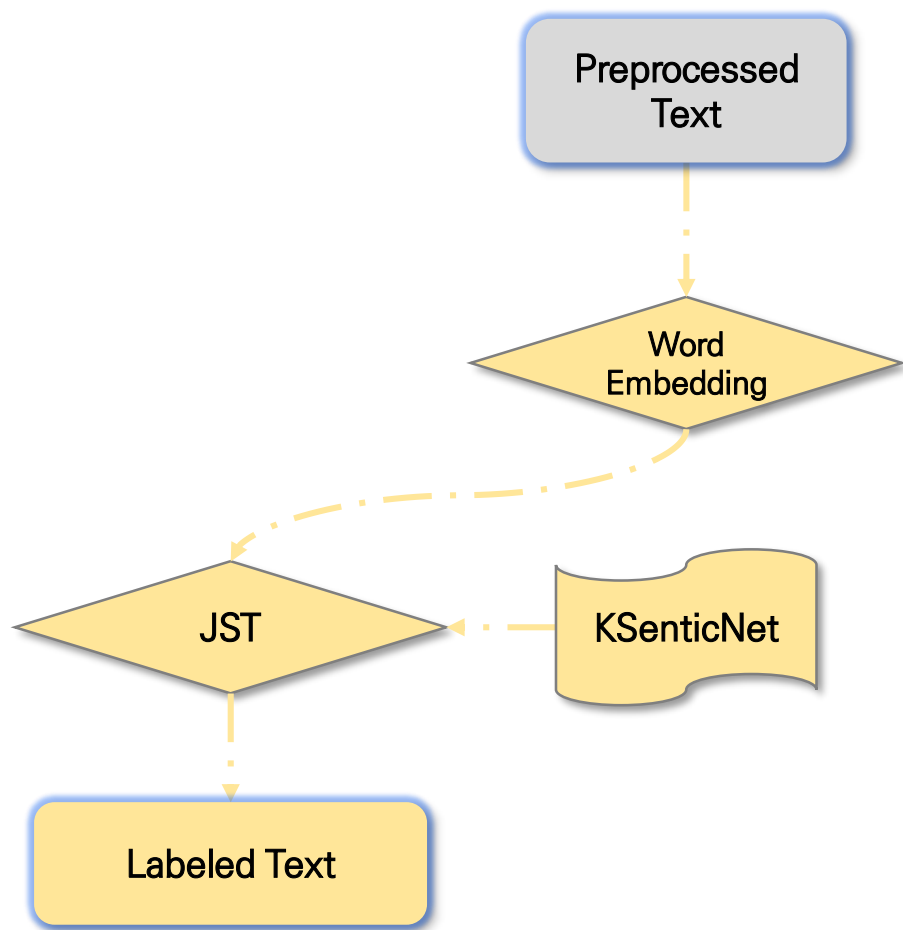
〈 Preprocessing Line 〉



04. M2SNet

Labeling line

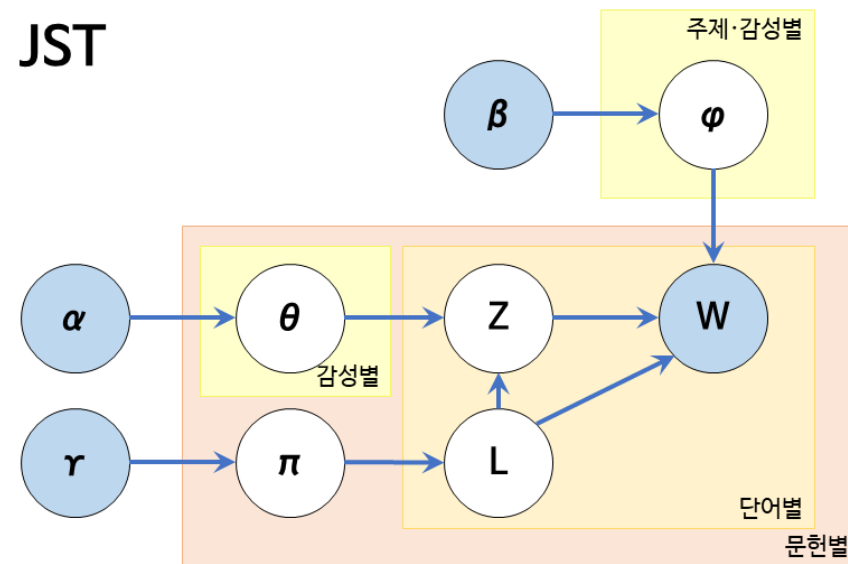
< Labeling Line >



Labeling

- ✓ Label이 없는 데이터로 어떻게 학습할 것인가?

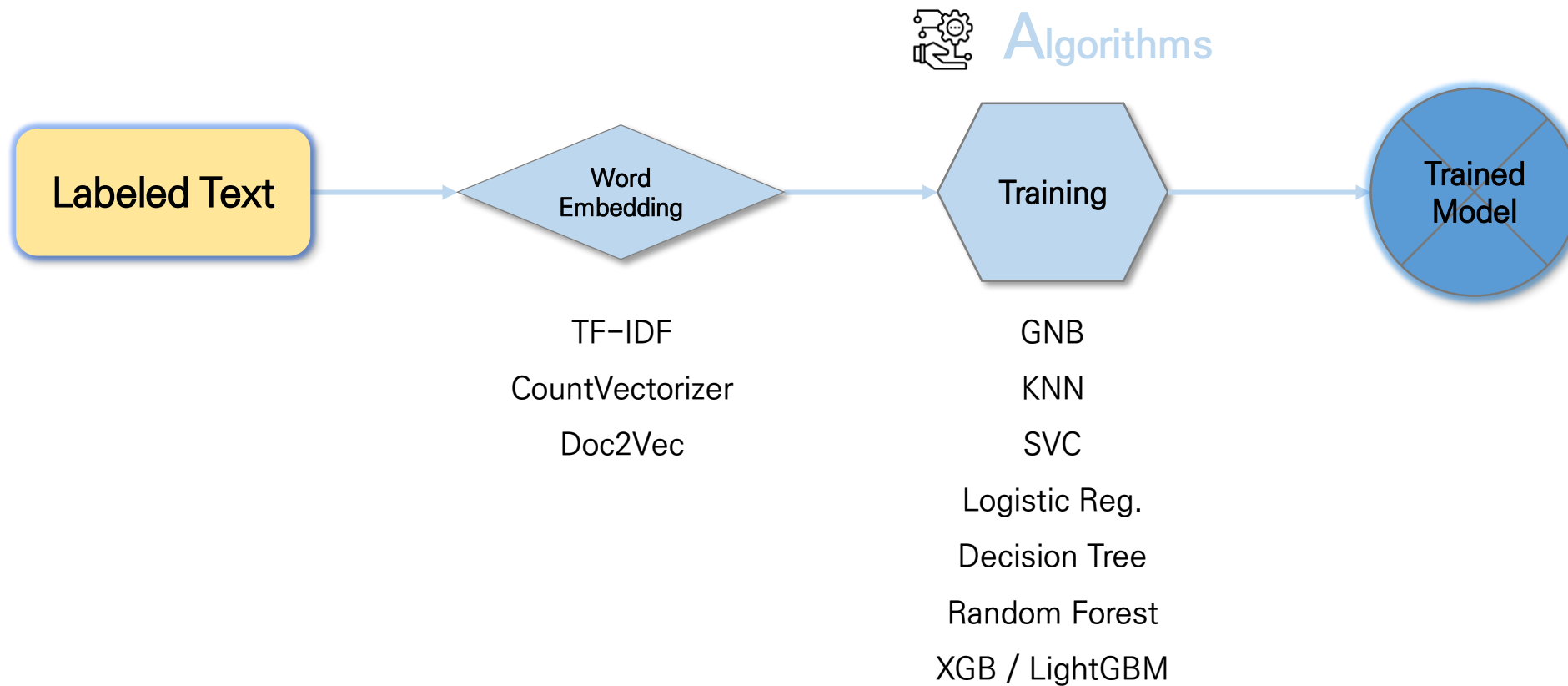
JST



04. M2SNet

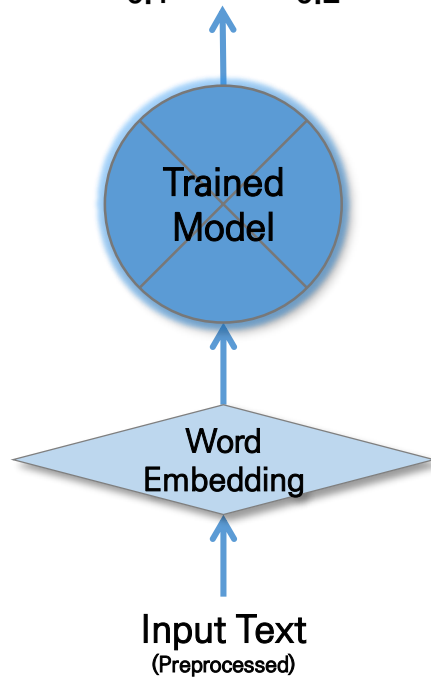
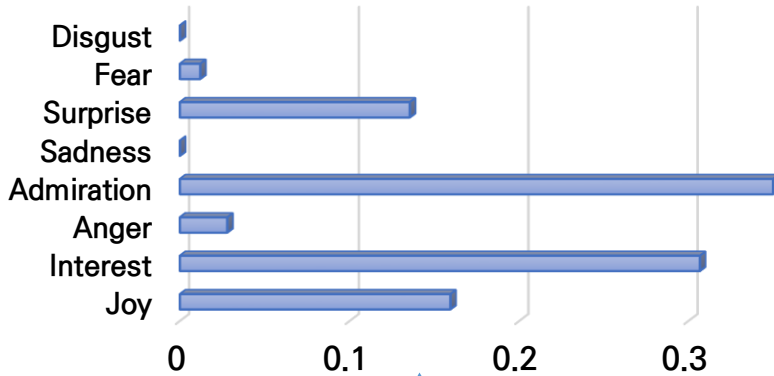
Modeling line

〈 Modeling Line 〉

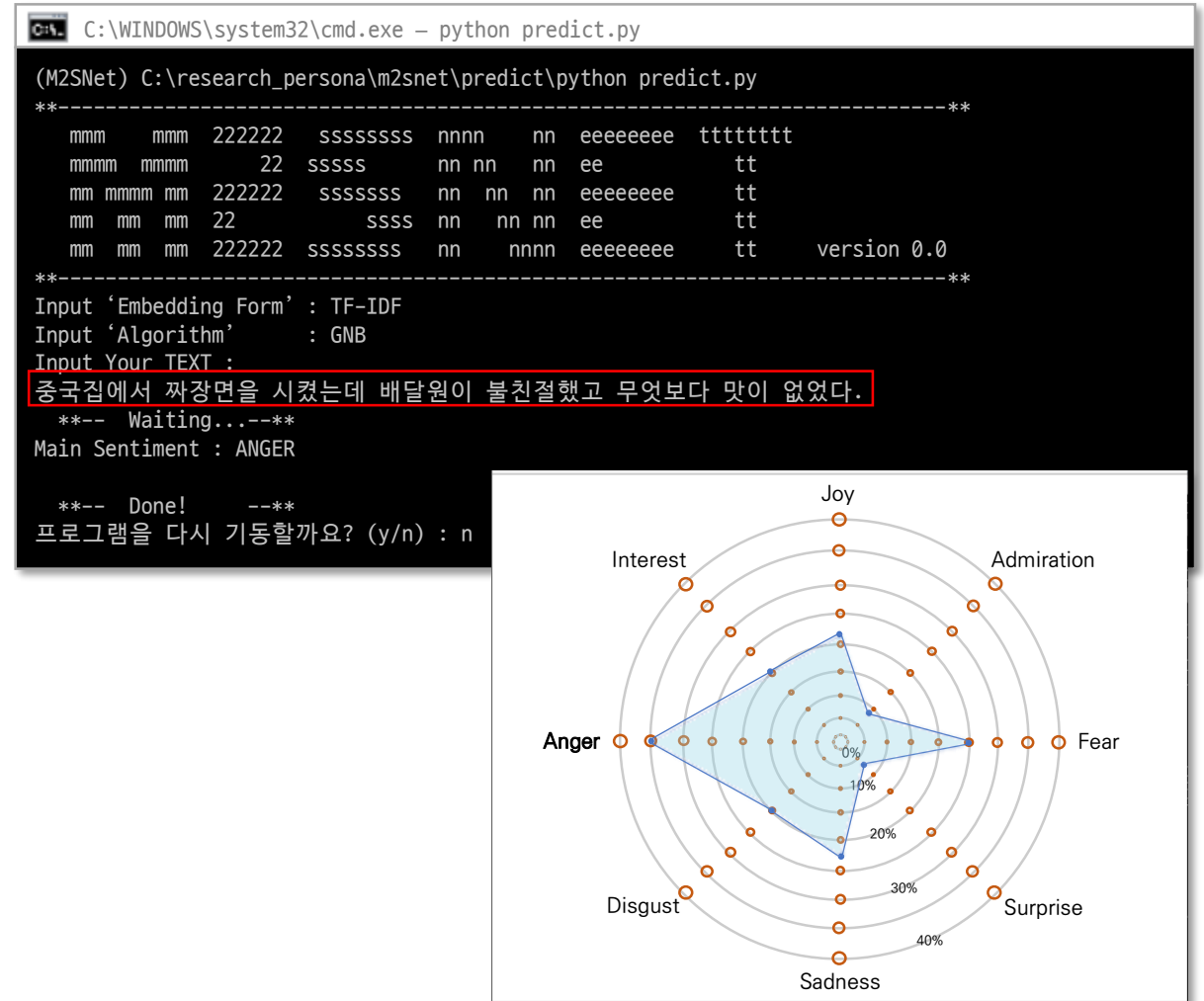


04. M2SNet

Predict line



〈 Predict Line 〉



Ch05 Version 0.0

- Labeling Detail
- 가설 설정
- 성능 평가

Ch06 Version 0.1

- Bagging Tree Ensemble

05. Version 0.0 (baseline)

23

Labeling Detail

〈Labeling line〉

```
class SentimentLDAGibbsSampler:
    ...
    def run(self, reviews, st, maxIters=30, do_preprocess=True, senti_dict=None):
        self._initialize(reviews, st, do_preprocess, senti_dict)
        numDocs, vocabSize = self.wordOccurenceMatrix.shape
        for iteration in range(maxIters):
            gc.collect()
            print('Starting iteration {} of {}'.format(iteration + 1, maxIters))
            for d in range(numDocs):
                for i, v in enumerate(word_indices(self.wordOccurenceMatrix[d, :].toarray()[0])):
                    t = self.topics[(d, i)]
                    s = self.sentiments[(d, i)]
                    self.n_dt[d, t] -= 1
                    self.n_d[d] -= 1
                    self.n_dts[d, t, s] -= 1
                    self.n_vts[v, t, s] -= 1
                    self.n_ts[t, s] -= 1

                    probabilites_ts = self.conditionalDistribution(d, v)
                    if v in self.priorSentiment:
                        s = self.priorSentiment[v]
                        t = sampleFromCategorical(probabilites_ts[:, s])
                    else:
                        ind = sampleFromCategorical(probabilites_ts.flatten())
                        t, s = np.unravel_index(ind, probabilites_ts.shape)

                    self.topics[(d, i)] = t
                    self.sentiments[(d, i)] = s
                    self.n_dt[d, t] += 1
                    self.n_d[d] += 1
                    self.n_dts[d, t, s] += 1
                    self.n_vts[v, t, s] += 1
                    self.n_ts[t, s] += 1
```

〈Compressed Sparse Row format〉

1. 초기화 작업 수행
2. Gibbs Sampling
3. 조건부확률분포 계산 후 할당
4. 2~3 과정을 Iteration만큼 반복

```
print('--* KSenticNet으로 사전 확률 조작 중... *--')
# 감정 사전 (KSenticNet)을 사용하여 사전 확률을 조작 중.
for i, word in enumerate(self.vectorizer.get_feature_names()):
    w = senti_dict.keys.get(word)
    if not w: continue
    synsets = senti_dict.scores[w, :]
    self.priorSentiment[i] = np.random.choice(
        self.numSentiments, p=synsets)
```

```
def conditionalDistribution(self, d, v):
    probabilites_ts = np.ones((self.numTopics, self.numSentiments))
    firstFactor = (self.n_dt[d] + self.alpha) / \
        (self.n_d[d] + self.numTopics * self.alpha)
    secondFactor = (self.n_dts[d, :, :] + self.gamma) / \
        (self.n_dt[d, :] + self.numSentiments * self.gamma)[:, np.newaxis]
    thirdFactor = (self.n_vts[v, :, :] + self.beta) / \
        (self.n_ts + self.n_vts.shape[0] * self.beta)
    probabilites_ts *= firstFactor[:, np.newaxis]
    probabilites_ts *= secondFactor * thirdFactor
    probabilites_ts /= np.sum(probabilites_ts)
    return probabilites_ts
```

05. Version 0.0 (baseline)

24

가설 설정

Hypothesis

1. Preprocessing

- 전처리 X
- 전처리 O

2. Embedding

- TF-IDF
- CountVectorizer
- Doc2Vec

3. Modeling

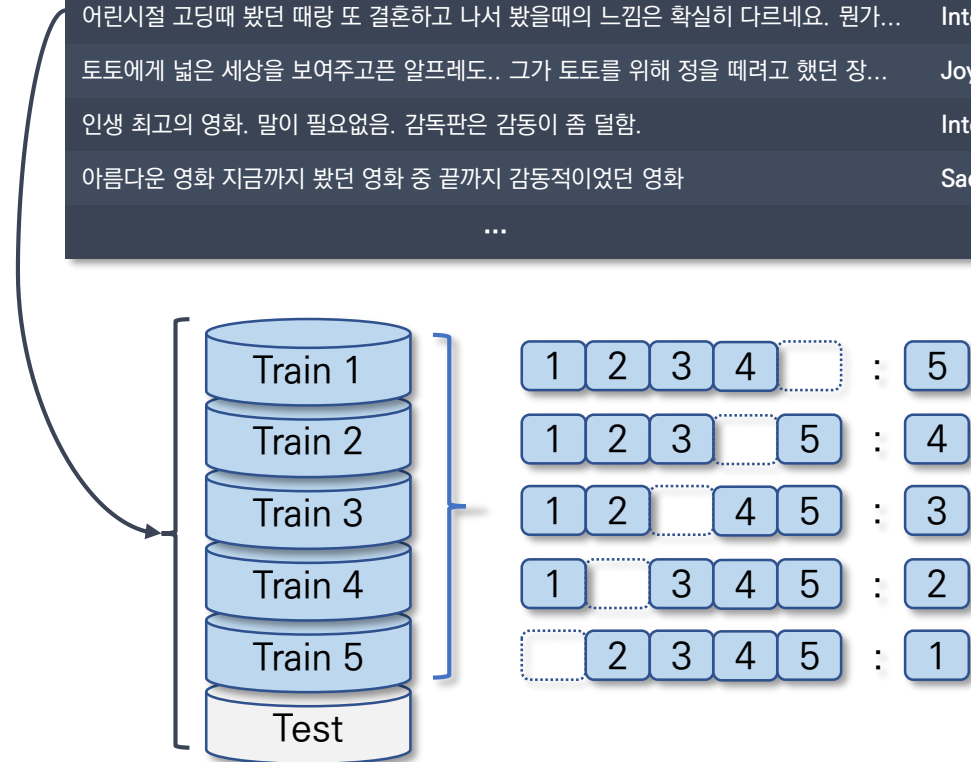
- GNB
- SVM
- KNN
- Logistic Regression

+ Random Search

- Decision Tree
- Random Forest
- XGBoost / LightGBM

- ✓ Train : Test = 5 : 1
- ✓ 5-folds CV
- ✓ MAX_VOCAB_SIZE = 10,000

review	sentiment
전체관람가는 아닌것 같아요	Fear
디렉터스컷으로봐서 거의 3시간짜리인데 참 흡인력있다	Surprise
태어나 처음으로 가슴아리는 영화였다. 20년이상 지났지만.. 생각하면 또 가슴이...	Admiration
어린시절 고딩때 봤던 때랑 또 결혼하고 나서 봤을때의 느낌은 확실히 다르네요. 원가...	Interest
토토에게 넓은 세상을 보여주고픈 알프레도.. 그가 토토를 위해 정을 때려고 했던 장...	Joy
인생 최고의 영화. 말이 필요없음. 감독판은 감동이 좀 덜함.	Interest
아름다운 영화 지금까지 봤던 영화 중 끝까지 감동적이었던 영화	Sadness
...	...



05. Version 0.0 (baseline)

25

시간효율 Bad : 예측에 1시간 이상 소요

실험 및 검증

		시간효율 Good		시간효율 Bad		시간효율 Bad		시간효율 Good		시간효율 Bad		시간효율 Bad		시간효율 Bad		시간효율 Bad	
		GNB		KNN		SVM		Logistic Regression		Decision Tree		Random Forest		XGBClassifier		LGBMClassifier	
Embedding	전처리	Acc	Recall	Acc	Recall	Acc	Recall	Acc	Recall	Acc	Recall	Acc	Recall	Acc	Recall	Acc	Recall
TF-IDF	O	52.15%	51.79%					50.36%	46.78%								
	X	47.05%	47.36%					50.22%	45.44%								
Count Vectorizer	O	52.79%	52.56%					51.80%	49.80%								
	X	48.14%	47.89%					51.36%	49.66%								
Doc2Vec	O	12.70%	13.67%					13.72%	12.22%								
	X	12.92%	12.19%					12.24%	12.71%								

⇒ 전처리 수행 여부에 따라 성능 차이 존재

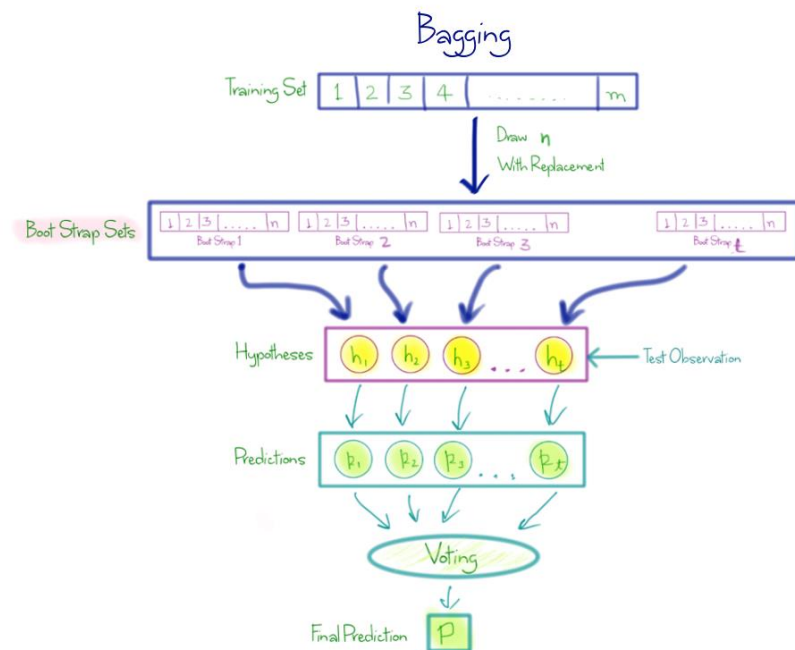
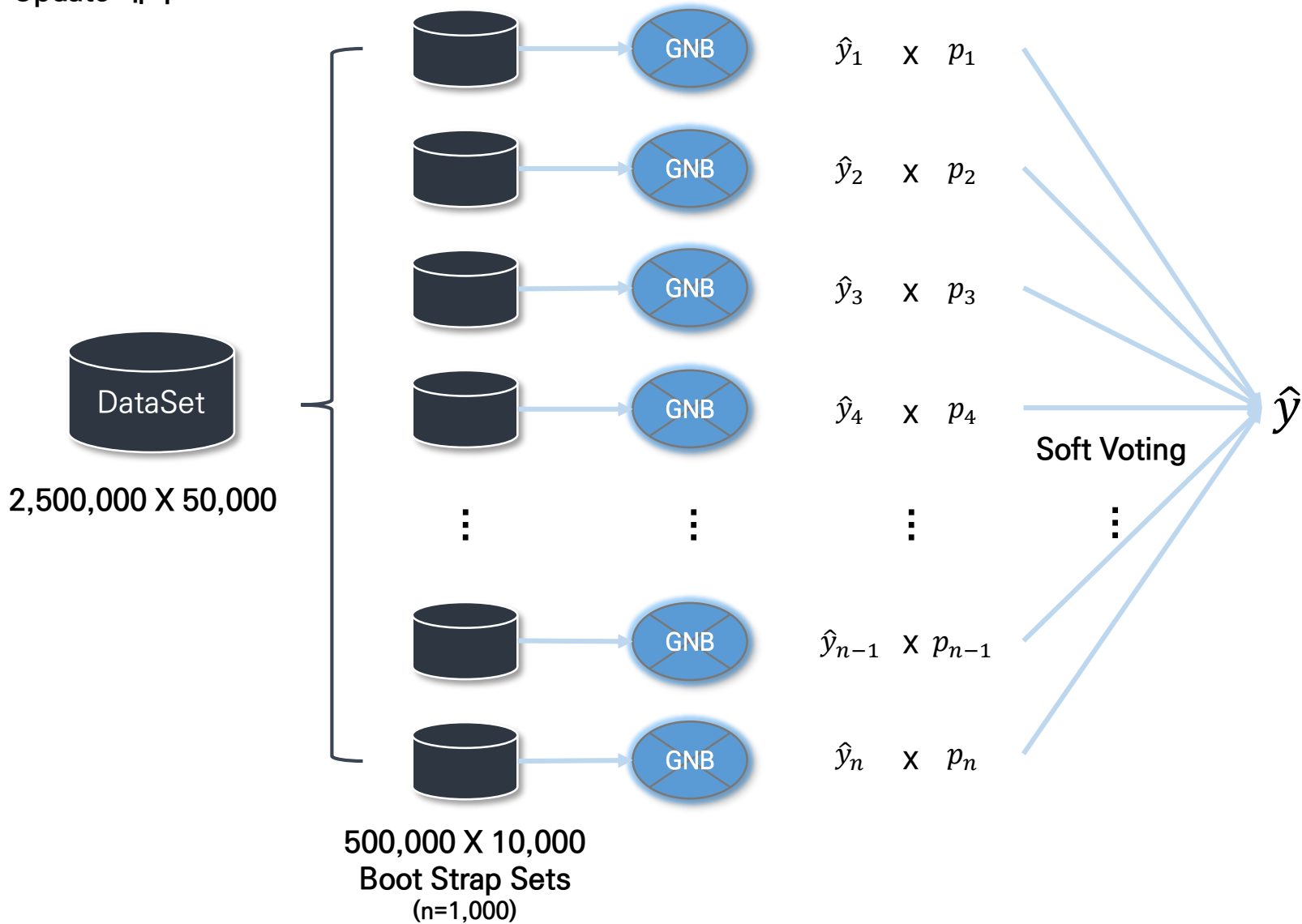
⇒ JST와 유사한 Embedding 방식을 사용해야 재현율 상승

⇒ CSR type의 Sparse Matrix를 Handling할 때 GNB가 압도적으로 우수

∴ Embedding : CountVect, Model : GNB

06. Version 0.1

Update 내역



Bagging Tree Ensemble

기존 GNB 대비

Recall 3.7% 상승
(56.26%)

Ch07 결론 및 향후 방향성

- 문제 분석 – Labeling
- 활용 서비스 제안
- 결론 및 향후 방향성

07. 결론 및 향후 방향성

문제 분석 – Labeling line

Latent Dirichlet Allocation

David M. Blei, **Andrew Y. Ng** and Michael I. Jordan
University of California, Berkeley
Berkeley, CA 94720

Abstract

We propose a generative model for text and other collections of discrete data that generalizes or improves on several previous models including naive Bayes/unigram, mixture of unigrams [6], and Hofmann's asymmetric LDA. We show that the model can be learned by variational Bayes, and that it can be used for document indexing, topic modeling, and other tasks. We also show that the model can be used as a latent space for document classification, where the latent space is learned as a by-product of the model training process. The model is carried out on a large collection of documents, and the empirical results show that it outperforms other models in document modeling.

Joint Sentiment/Topic Model for Sentiment Analysis

Chenghua Lin
School of Engineering, Computing and Mathematics
University of Exeter
North Park Road, Exeter EX4 4QF, UK
cl322@exeter.ac.uk

Yulan He
Knowledge Media Institute
The Open University
Milton Keynes MK7 6AA, UK
y.l.he.01@cantab.net

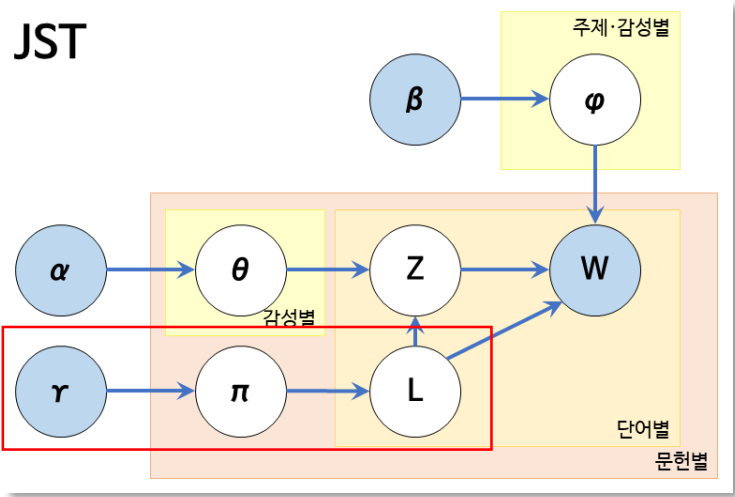
ABSTRACT

Sentiment analysis or opinion mining aims to use automated tools to detect subjective information such as opinions, attitudes, and feelings expressed in text. This paper proposes a novel probabilistic modeling framework based on Latent Dirichlet Allocation (LDA) and Sentiment Dirichlet Distribution (SDD). The model has been much interests in the natural language processing community to develop novel text mining techniques with the capability of accurately extracting customers' opinions from large volumes of unstructured text data. Among various opinion mining tasks, one of them is sentiment classification, i.e., whether the semantic orientation of

- ## Joint Sentiment Topic model

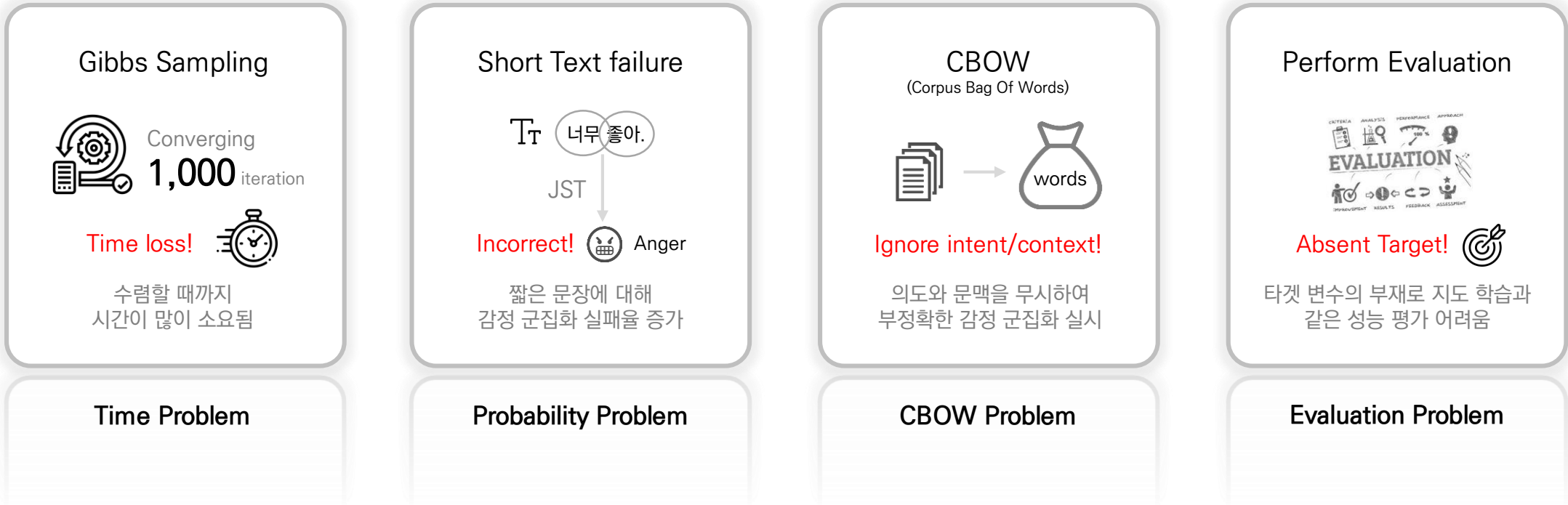
(Sentiment LDA)

- ✓ LDA에 Sentiment Dirichlet Distribution을 추가한 모델
 - ✓ LDA는 Gibbs Sampling을 통해 문서가 생성되는 과정을 확률모형으로 모델링



07. 결론 및 향후 방향성

문제 분석 – Labeling line



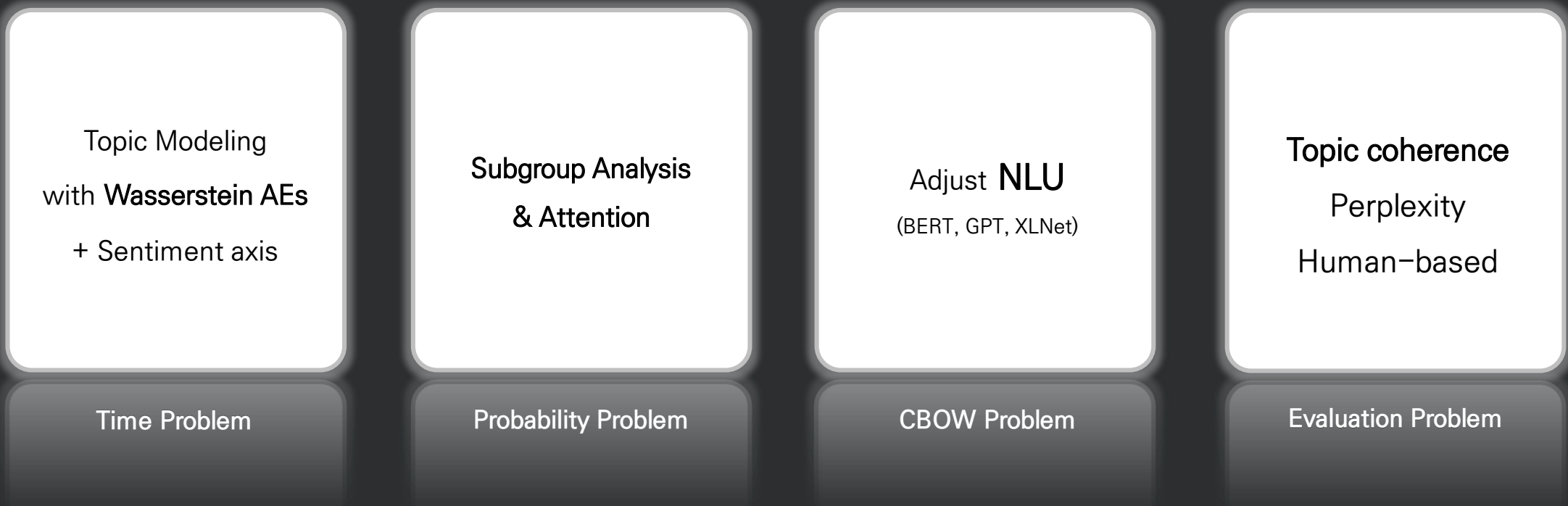
때문에 위와 같은 문제를 갖습니다.

07. 결론 및 향후 방향성

문제 분석 - Labeling line



저희 페르소나 시스템 조는,

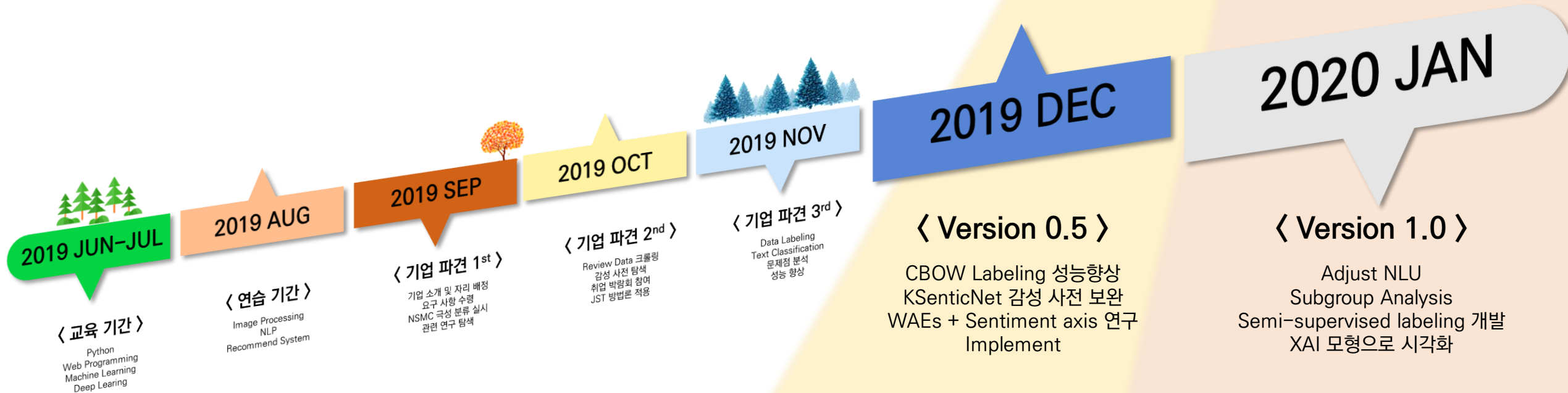


각 문제를 위의 방법으로 해결하겠습니다.

07. 결론 및 향후 방향성

결론 및 향후 방향성

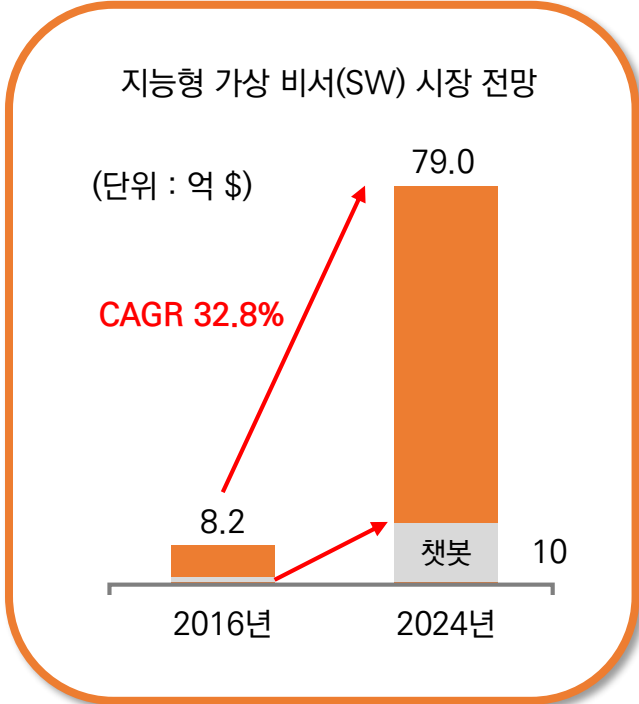
30



저희 프로젝트는 아직 끝나지 않았습니다.

07. 결론 및 향후 방향성

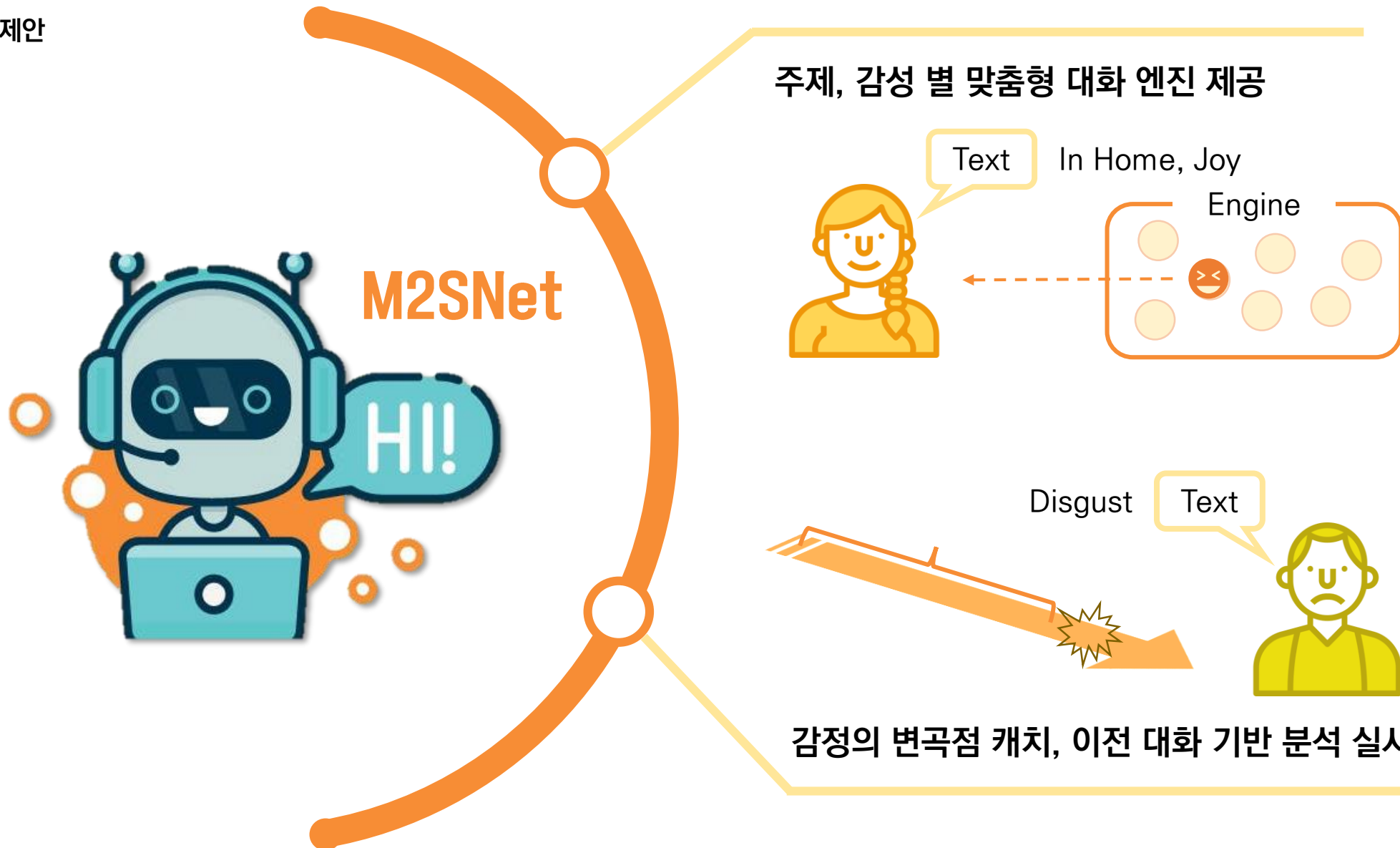
활용 서비스 제안



챗봇은 2024년 약 10억 달러의 시장규모로
예상되며 고속성장중

07. 결론 및 향후 방향성

활용 서비스 제안



감사합니다.
