seok830621

5

off

소셜빅데이터 분석을 통한 정신건강정책 제안

머신러닝의 기초 4조

이전 발표 내용

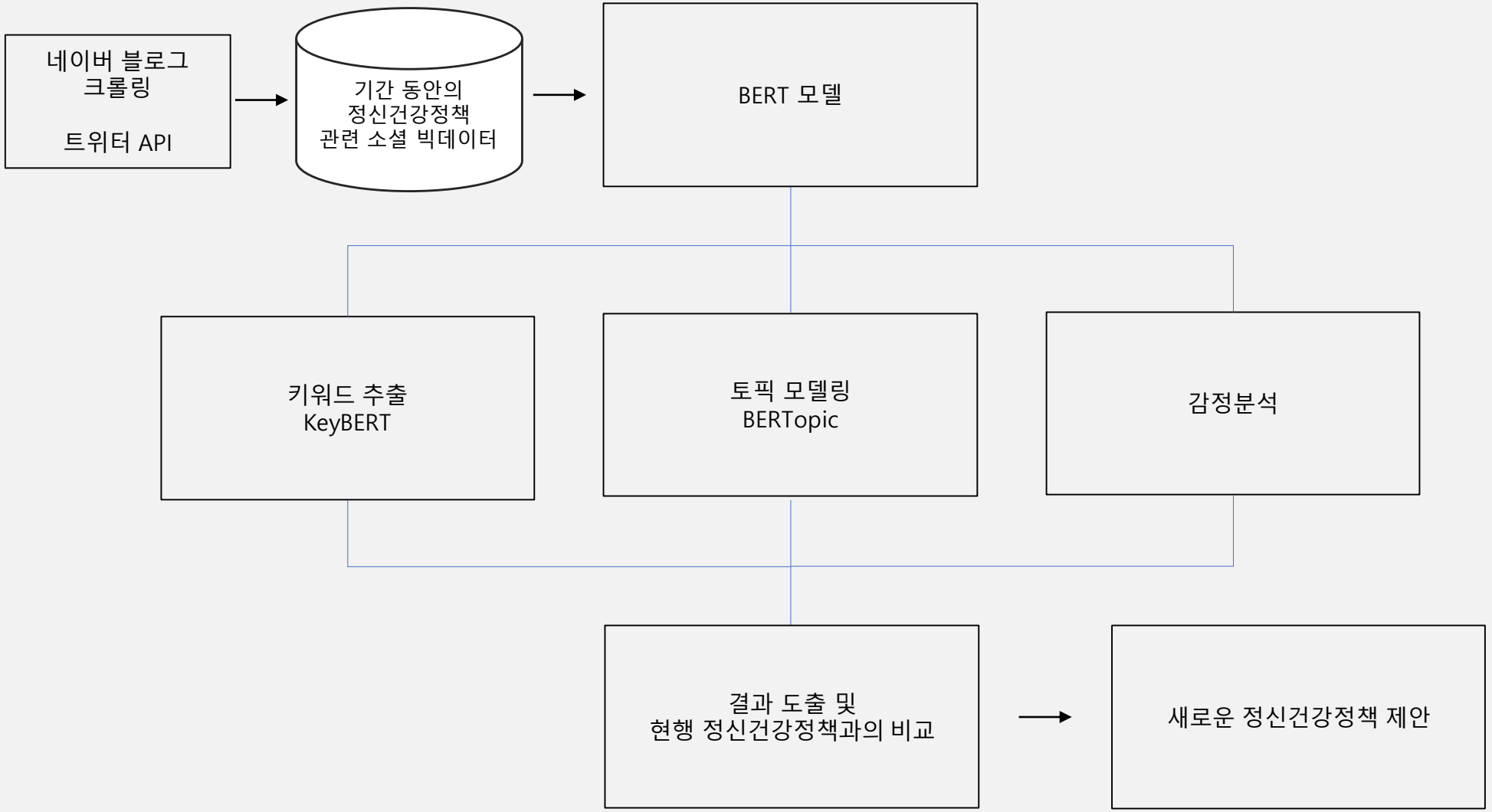
- 자연어처리 프로세스
- 임베딩

  - Word2Vec
  - FastText

- 언어모델

  - Markov 확률 기반 언어모델
  - RNN 기반 언어모델
  - Attention 모델
  - Transformer 모델
  - Bert 모델

연구 프로세스


네이버 블로그
크롤링
트위터 API
기간 동안의
정신건강정책
관련 소셜 빅데이터
BERT 모델
키워드 추출
KeyBERT
토픽 모델링
BERTopic
감정분석
결과 도출 및
현행 정신건강정책과의 비교
새로운 정신건강정책 제안

KeyBERT

- SBERT 패키지인 sentence_transformers를 이용해 키워드 추출
- 키워드를 추출하고자 하는 문서와 토큰화된 키워드 2개의 임베딩 필요

```
In [5]: doc
Out[5]: '[메디칼업저버 김나현 기자] 정신질환을 겪는 아동과 청소년이 꾸준히 증가하고 있지만 정부의 정책 지원은 미비한 것으로 나타났다.₩n₩n이에 아동·청소년 전문 기초정신건강복지센터를 지역별로 확충해 질병을 조기발견하고, 정신건강간호사의 인력도 확충할 필요가 있다는 제안이 나왔다.₩n₩n최근 국회입법조사처는 아동·청소년의 정신건강 현황을 분석한 보고서를 통해 지원제도 및 개선방향을 제시했다.₩n₩n2016년부터 2020년까지 아동·청소년의 정신진료 현황을 살펴보면, 최근 5년간 진료를 받은 환자 수는 2016년 22만 5679명에서 2020년 27만 1557명으로 꾸준히 증가했다.₩n₩n한국표준질병·사인분류의 정신 및 행동장애에 해당하는 상병명을 기준으로 최근 5년간 아동·청소년(0세~19세)의 정신질환 상병명을 살펴보면, 2016년을 제외한 2017년부터 운동과다장애가 가장 많았다. 운동과다장애는 ADHD(주의력결핍 과잉행동장애)를 포함한다.₩n₩n아동·청소년(0세~19세)의 상위 5개 정신질환별 환자 수 (단위:명)₩n이어 ▲우울증 ▲기타 불안장애 ▲심한 스트레스에 대한 반응 및 적응장애 ▲전반발달장애 순으로 수진자 수에서 상위권을 차지했다.₩n₩n연령대를 좀 더 세분화해 0세~9세까지의 정신질환별 환자 수를 살펴보면, 2016년부터 매년 ADHD를 포함한 운동과다장애가 가장 많았다. ₩n₩n다음으로는 말하기와 언어의 특정 발달장애, 전반발달장애, 틱장애 순이었다.₩n₩n한편 10세~19세는 2018년 이후부터 우울증 환자가 가장 많았다. 2016년과 2017년에는 운동과다장애가 1위였지만 2018년부터는 2위였던 우울증 환자가 더 많아진 것이다.₩n₩n이 연령에서는 우울증에 이어 운동과다장애, 기타 불안장애, 심한스트레스에 대한 반응 및 적응장애 순으로 환자가 많았다.₩n₩n19세 이하 아동·청소년의 자살률도 꾸준히 늘고 있다.₩n₩n한국생명존중희망재단 자료에 따르면 자살자 수는 2015년 245명에서 2019년 300명으로 늘었다. 인구 10만명 당 자살자 수인 '자살률'도 같은 기간 2.3명에서 3.2명으로 증가했다.₩n₩n또한 같은 기간 전체 자살자 수에서 아동·청소년의 자살이 차지하는 비중도 2015년 이후 지속적으로 2.0%를 상회했다.₩n₩n국민건강증진법에 따라 매년 전국의 중·고등학교 재학생을 대상으로 실시하는 청소년건강행태조사에서도 스트레스와 우울감에 취약한 청소년의 정신건강 상태가 나타났다.₩n₩n이 통계를 만 19세 이상 성인을 대상으로 하는 국민건강영양조사의 정신건강 결과와 비교해보면 성인보다 청소년의 스트레스 인지율이 매년 더 높게 나타났다.₩n₩n2019년 기준으로 성인과 청소년의 스트레스 인지율은 각각 28.6%, 39.9%였다.₩n₩n"정부 추진 사업, 의료비 지원 제외하면 차별성 없다"₩n₩n인구 기준 권역별 센터 설치, 정신질환실태조사 확대 등 제안₩n₩n상황이 이렇자 정부는 교육부와 보건복지부, 여성가족부를 중심으로 아동·청소년의 정신건강 사업을 추진 중이다. 여기서 복지부는 기초정신건강복지센터를 중심으로 아동과 청소년의 정신건강증진사업을 시행하고 있지만 역부족이라는 것이 입법조사처의 판단이다.₩n₩n입법조사처는 "기초정신건강복지센터에서 초기평가, 사례관리, 의료기관 연계 및 의료비 지원, 자살예방 등을 수행하고 있다"면서도 "의료비 지원을 제외하면 해당 사업은 지역주민을 대상으로 동일하게 시행하고 있어 차별성이 없다"고 지적했다.₩n₩n2019년 말 기준으로 241개의 기초정신건강복지센터 중 아동·청소년에 특화해 치료가 가능한 기초정신건강복지센터는 고양시, 부천시, 성남시, 수원시에 설치한 4개소에 불과하다.₩n₩n입법조사처는 "일상적이고 만성적인 우울감, 자살충동에 시달리는 청소년을 조기에 발굴하기 위해 지원서비스 접근 가능성을 제고해야 한다"고 지적했다.₩n₩n정부의 행태조사를 통해 관심군으로 발굴하는 일련의 과정은 상당한 시간이 포함될 뿐 아니라, 연계 의료기관의 부족으로 적절한 치료 시기를 놓치기 쉽다는 것이다.₩n₩n이에 정신건강의 위험요인을 조기발견하고 지속적으로 관리하기 위한 정책 개선방향이 제시됐다.₩n₩n구체적으로 ▲아동·청소년 대상 정신질환실태조사의 주기적 실시 ▲아동·청소년 전문 기초정신건강복지센터의 지역별 확충 ▲학교 내외에서의 정신건강증진사업 강화 등이다.₩n₩n현재 정신건강복지법에 따라 5년 주기로 정신질환실태조사를 실시하고 있으나, 조사대상을 만 18세 이상으로 해 아동·청소년의 정신질환에 대한 통계가 정확하게 파악되지 않는 문제점이 있다.₩n₩n입법조사처는 "아동·청소년의 정신질환률이 정확하게 산출되고, 그에 따른 정책을 시행하고 있는 외국과 다른 상황"이라며 "아동과 청소년을 대상으로 하는 정신질환실태조사의 도입을 검토할 수 있다"고 설명했다.₩n₩n관련 시설을 지역별로 확충해야 한다는 지적도 나왔다. 2019년 기준 아동·청소년에 특화해 정신질환의 예방과 조기발견, 치료가 가능한 정신건강복지센터는 전국 241개소 중 4개소에 불과하다.₩n₩n또한 퇴원 후 사회복귀를 지원하는 정신재활시설은 전국 349개소 중 아동·청소년 정신건강지원시설은 13개이며 서울에 11개소, 제주시에 2개소가 설치됐다.₩n₩n입법조사처는 "급격한 설치는 불가능하겠지만, 아동·청소년 인구를 기준으로 권역별로 설치해 시설의 수도권 편중을 해결해야 한다"며 "전국적으로 균등한 정신건강서비스를 제공할 수 있는 방안이 필요하다"고 제안했다.₩n₩n이어 "정신건강에 대한 사업 수행인력의 전문성이 부족해보이므로 Wee클래스에 정신건강간호사 등 정신건강전문요원 인력을 확충해야 한다"고 강조했다.₩n'
```

```
In [7]: okt = Okt()

In [8]: tokenized_doc = okt.pos(doc)

In [9]: tokenized_nouns = ' '.join([word[0] for word in tokenized_doc if word[1] == 'Noun'])

In [10]: tokenized_nouns
Out[10]: '칼 업저버 김 나현 기자 정신질환 아동 청소년 증가 정부 정책 지원 것 이 아동 청소년 전문 기초 정신건강 복지 센터 지역별 확충 질병 발견 정신건강 간호사 인력 확충 필요 제안 최근 국회입법조사처 아동 청소년 정신건강 현황 분석 보고서 통해 지원 제도 및 개선 방향 제시 아동 청소년 정신 진료 현황 최근 진료 환자 수 명 명 증가 한국 표준 질병 사인 분류 정신 및 행동 장애 해당 상병 기준 최근 아동 청소년 세 세 의 정신질환 병명 제외 운동 과 장애 가장 운동 과 장애 주의력 결핍 과잉행동 장애 를 포함 아동 청소년 세 세 의 상위 개 정신질환 별 환자 수 딘위 명 우울증 기타 불안장애 스트레스 대한 반응 및 적응장애 전반 발달장애 순 수진 수 상위 차지 연령 좀 더 세분 화해 세 세 정신질환 별 환자 수 기간 매년 를 포함 운동 과 장애 가장 다음 말 언어 특정 발달장애 전반 발달장애 틱장애 순 반면 세 이후 우울증 환자 가장 운동 과 장애 위 위 우울증 환자 더 것 이 연령 우울증 운동 과 장애 기타 불안장애 스트레스 대한 반응 및 적응장애 순 환자 세 이하 아동 청소년 자살률 늘 한국 생명 존중 희망 재단 자료 자살 수 명 명 인구 당 자살 수인 자살률 도 기간 명 명 증가 또한 기간 전체 자살 수 아동 청소년 자살 차지 비중 이후 지속 를 상회 국민 건강 증 진법 매년 전국 중 고등학교 재학생 대상 실시 청소년 건강 행태 조사 스트레스 우울 감 취약 청소년 정신건강 상태 이 통계 만 세 이상 성인 대상 국민 건강 조사 정신건강 결과 비교 성인 청소년 스트레스 인지 율 매년 더 기준 성인 청소년 스트레스 인지 율 각각 정부 추진 사업 의료 비 지원 제외 차별성 인구 기준 권역별 센터 설치 정신질환 실태 조사 확대 등 제안 상황 정부 교육부 보건복지부 여성가족부 중심 아동 청소년 정신건강 사업 추진 중이 여기 복지부 기초 정신건강 복지 센터 중심 아동 청소년 정신건강 증진 사업 시행 역 부족 것 입법조사처 판단 입법조사처 기초 정신건강 복지 센터 초기 평가 사례 관리 의료 기관 계 및 의료 비 지원 자살 예방 등 수행 의료 비 지원 제외 해당 사업 지역 주민 대상 시행 차별성 고 지적 말 기준 개 기초 정신건강 복지 센터 중 아동 청소년 특 화해 치료 기초 정신건강 복지 센터 고양시 부천시 성남시 수원시 설치 개소 입법조사처 일상 성적 우울 감 자살 충 에 청소년 조기 발굴 위해 지원 서비스 접근 가능성 고해 고 지적 정부 행 태조 통해 관심 발굴 일련 과정 포함 뿐 계 의료 기관 부족 치료 시기 것 이 정신건강 위험 요인 발견 지속 관리 위 정책 개선 방향 제시 구체 아동 청소년 대상 정신질환 실태 조사 기적 실시 아동 청소년 전문 기초 정신건강 복지 센터 지역별 확충 학교 내외 정신건강 증진 사업 강화 등 현재 정신건강 복지 법 주기 정신질환 실태 조사 실시 대상 만 세 이상 해 아동 청소년 정신질환 대한 통계 파악 문제점 입법조사처 아동 청소년 정신질환 율 산출 그 정책 시행 외국 다른 상황 라며 아동 청소년 대상 정신질환 실태 조사 도입 검토 수 고 설명 관련 시설 지역별 확충 지적도 기준 아동 청소년 특 화해 정신질환 예방 조기 발견 치료 정신건강 복지 센터 전국 중 개소 또한 퇴원 후 사회 복귀 지원 정신 재활 시설 전국 중 아동 청소년 정신건강 시설 개 서울 제주시 개 소가 설치 입법조사처 아동 청소년 인구 기준 권역 별로 설치 시설 수도권 편중 해결 며 전국 정신건강 서비스 제공 수 방안 고 제안 정신건강 대한 사업 행인 전문성 부족 클래스 정신건강 간호사 등 정신건강 전문요원 인력 확충 고 강조'
```

KeyBERT

- n_gram을 2와 3으로 설정하여 2~3단어로 이루어진 키워드를 추출하고자 함
- BERT 모델은 다국어SBERT 모델을 로드
- 문서와 n_gram 키워드 후보를 임베딩

```
n_gram_range = (2,3)
count = CountVectorizer(ngram_range = n_gram_range).fit([tokenized_nouns])
candidates = count.get_feature_names_out()
print('bigram과 trigram 개수:', len(candidates))
```

```
bigram과 trigram 개수: 790
```

다국어 sbert load

```
model = SentenceTransformer('sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens')
doc_embedding = model.encode([doc])
candidate_embeddings = model.encode(candidates)
```

Cosine Similarity 기준 키워드 추출

```
top_n = 5
distances = cosine_similarity(doc_embedding, candidate_embeddings)
keywords = [candidates[index] for index in distances.argsort()[0][-top_n:]]
```

```
keywords
```

```
['최근 국회입법조사처 아동', '청소년 화해 정신질환', '문제점 입법조사처 아동', '제도 개선 방향',
'청소년 정신건강 증진']
```

KeyBERT

- 문서와 키워드임베딩 사이의 코사인 유사도 이외에도 결과 키워드의 다양성을 위해
- MSS(Max Sum Similarity), MMR(Maximal Marginal Relevance) 알고리즘 사용가능

nr_candidates 10

```
max_sum_sim(doc_embedding, candidate_embeddings, candidates, top_n=5, nr_candidates=10)
```

```
['환자 증가 한국', '해결 전국 정신건강', '청소년 정신건강 사업', '지원 제도 개선', '최근 국회입법조사처 아동']
```

nr_candidates 30

nr_candidates 를 높이면 더 다양한 키워드 추출 가능

```
max_sum_sim(doc_embedding, candidate_embeddings, candidates, top_n=5, nr_candidates=30)
```

```
['복지 주기 정신질환', '간호사 인력 확충', '정책 개선', '국회입법조사처 아동 청소년', '환자 증가 한국']
```

MSS는 문서에서 코사인 유사도를 기준으로 상위 n개의 단어를 선택하고 각 키워드들 중 가장 덜 유사한 키워드들간의 조합을 계산

diversity 0.2

```
mmr(doc_embedding, candidate_embeddings, candidates, top_n=5, diversity=0.2)
```

```
['청소년 정신건강 증진', '환자 증가 한국', '최근 국회입법조사처 아동', '제도 개선 방향', '해결 전국 정신건강']
```

diversity 0.7

diversity 를 높이면 다양한 키워드 추출 가능

```
mmr(doc_embedding, candidate_embeddings, candidates, top_n=5, diversity=0.7)
```

```
['청소년 정신건강 증진', '최근 국회입법조사처', '의료 기관 부족', '고양시 부천시 성남시', '차별성 인구 기준']
```

MMR은 문서와 가장 유사한 키워드를 선택하고 이후 문서와 비슷하면서 이미 선택한 키워드와 비슷하지 않은 새 후보를 반복적으로 선택

KoBERTopic

- BERTopic 모델은 현재 github에서 3000개 이상의 star를 받은 토픽모델링 방식
- https://github.com/MaartenGr?tab=repositories

- 이를 한국어 데이터에 적용할 수 있도록 tokenizer와 사용 BERT을 수정한 KoBERTopic

- Tokenizer: CountVectorize에서는 단순 띄어쓰기 > Mecab
- 단순 띄어쓰기 토큰화 > 형태소 분석을 통한 토큰화

- BERT model: 다국어 SBERT 'sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens'

- Input: 하나의 line에 하나의 doc로 구성

```
mecab.morphs("아버지가방에들어가신다")
```

```
['아버지', '가', '방', '에', '들어가', '신다']
```

	토크나이저	토큰화 소요시간
0	komoran	1.146485
1	okt	0.108238
2	mecab	0.008010
3	hannanum	0.481190
4	kkma	0.552523

KoBERTopic

1. 텍스트 데이터 BERT로 임베딩

2. 임베딩의 차원축소(UMAP), 임베딩 클러스터링 (HDBSCAN) 이후 의미적으로 유사한 문서 클러스터 생성

3. c-TF-IDF(클래스 기반 TF-IDF)를 통해 토픽 추출

UMAP: high dimension space에서 데이터를 그래프로 만들어 low dimension으로 graph projection하는 차원축소 알고리즘 – 빠르고 global structure를 잘 보존하는 차원축소

HDBSCAN: DBSCAN을 기반으로 local density에 대한 정보를 반영하지 못하고 데이터의 계층적 구조를 반영한 클러스터링이 불가능하다는 한계점을 개선한 알고리즘

KoBERTopic

```
! docs
```

```
['메디칼업저버 김나환 기자 정신질환을 겪는 아동과 청소년이 꾸준히 증가하고 있지만 정부의 정책 지원은 미비한 것으로 나타났다.  이에 아동청소년 전문 기초정신건강복지센터를 지역별로 확충해 질병을 조기발견하고 정신건강간호사의 인력도 확충할 필요가 있다는 제안이 나왔다.  최근 국회입법조사처는 아동청소년의 정신건강 현황을 분석한 보고서를 통해 지원체도 및 개선방향을 제시했다.  2016년부터 2020년까지 아동청소년의 정신진료 현황을 살펴보면 최근 5년간 진료를 받은 환자 수는 2016년 22만 587명에서 2020년 27만 1557명으로 꾸준히 증가했다.  한국표준질병사인분류의 정신 및 행동장애에 해당하는 상병명을 기준으로 최근 5년간 아동청소년(0세-19세)의 정신질환 병명을 살펴보면 2016년을 제외한 2017년부터 문동과다장애가 가장 많았다.  운동과다장애는 ADHD주의력결핍 과잉행동장애를 포함한다.  아동청소년(0세-19세)의 상위 5개 정신질환별 환자 수 단위:명 이어 우울증 기타 불안장애 심한 스트레스에 대한 반응 및 적응장애 전반발달장애 순으로 수진자 수에서 상위권을 차지했다.  연령대를 좀 더 세분화해 0세-9세까지의 정신질환별 환자 수를 같은 기간 살펴보면 2016년부터 매년 ADHD를 포함한 문동과다장애가 가장 많았다.  다음으로는 말하기와 언어의 특정 발달장애 전반발달장애 탁장애 순이었다.  반면 10세-19세는 2018년 이후부터 우울증 환자가 가장 많았다.  2016년과 2017년에는 문동과다장애가 1위였지만 2018년부터는 2위였던 우울증 환자가 더 많아진 것이다.  이 연령에서는 우울증에 이어 문동과다장애 기타 불안장애 심한스트레스에 대한 반응 및 적응장애 순으로 환자가 많았다.  19세 이하 아동청소년의 자살률도 꾸준히 늘고 있다.  한국생명존중희망재단 자료에 따르면 자살자 수는 2015년 246명에서 2019년 300명으로 늘었다.  인구 10만명 당 자살자 수인 자살률도 같은 기간 2.3명에서 3.2명으로 증가했다.  또한 같은 기간 전체 자살자 수에서 아동청소년의 자살이 차지하는 비중도 2015년 이후 지속적으로 2.0%를 상회했다.  국민건강증진법에 따라 매년 전국의 중고등학교 재학생을 대상으로 실시하는 청소년건강행태조사에서도 스트레스와 우울감에 취약한 청소년의 정신건강 상태가 나타났다.  이 통계를 만 19세 이상 성인을 대상으로 하는 국민건강영양조사의 정신건강 결과와 비교해보면 성인보다 청소년의 스트레스 인지율이 매년 더 높게 나타났다.  2019년 기준으로 성인과 청소년의 스트레스 인지율은 각각 28.6% 39.9%였다.  정부 추진 사업 의료비 지원 제외하면 차별성 없다  인구 기준 권역별 센터 설치 정신질환실태조사 확대 등 제안  상황이 이렇자 정부는 교육부와 보건복지부 여성가족부를 중심으로 아동청소년의 정신건강 사업을 추진 중이다.  여기서 복지부는 기초정신건강복지센터를 중심으로 아동과 청소년의 정신건강증진사업을 시행하고 있지만 역부족이라는 것이 입법조사처의 판단이다.  입법조사처는 기초정신건강복지센터에서 초기평가 사례관리 의료기관 연계 및 의료비 지원 자살예방 등을 수행하고 있다면서도 의료비 지원을 제외하면 해당 사업은 지역주민을 대상으로 동일하게 시행하고 있어 차별성이 없다고 지적했다.  2019년 말 기준으로 241개의 기초정신건강복지센터 중 아동청소년에 특화된 치료가 가능한 기초정신건강복지센터는 고양시 부천시 성남시 수원시에 설치한 4개소에 불과하다.  입법조사처는 일상적이고 만성적인 우울감 자살충동에 시달리는 청소년을 조기에 발굴하기 위해 지원서비스 접근 가능성을 제고해야 한다고 지적했다.  정부의 협태조사를 통해 관심군으로 발굴하는 원칙의 과정은 상당한 시간이 소합될 뿐 아니라 연계 의료기관의 부족으로 적절한 치료 시기를 놓치기 쉽다는 것이다.  이에 정신건강의 위험요인을 조기발견하고 지속적으로 관리하기 위한 정책 개선방향이 제시됐다.  구체적으로 아동청소년 대상 정신질환실태조사의 주기적 실시 아동청소년 전문 기초정신건강복지센터의 지역별 확충 학교 내외에서의 정신건강증진사업 강화 등이다.  현재 정신건강복지법에 따라 5년 주기로 정신질환실태조사를 실시하고 있으나 조사대상을 만 18세 이상으로 해 아동청소년의 정신질환에 대한 통계가 정확하게 파악되지 않는 문제점이 있다.  입법조사처는 아동청소년의 정신질환율이 정확하게 산출되고 그에 따른 정책을 시행하고 있는 외국과 다른 상황이라며 아동과 청소년을 대상으로 하는 정신질환실태조사의 도입을 검토할 수 있다고 설명했다.  관련 시설을 지역별로 확충해야 한다는 지적도 나왔다.  2019년 기준 아동청소년에 특화해 정신질환의 예방과 조기발견 치료가 가능한 정신건강복지센터는 전국 241개소 중 4개소에 불과하다.  또한 퇴원 후 사회복귀를 지원하는 정신재활시설은 전국 349개소 중 아동청소년 정신건강지원시설은 13개이며 서울에 11개소 제주시에 2개소가 설치했다.  입법조사처는 급격한 설치는 불가능하겠지만 아동청소년 인구를 기준으로 권역별로 설치해 시분의 수도권 편중을 해결해야 한다며 전국적으로 균등한 정신건강서비스를 제공할 수 있는 방안이 필요하다고 제안했다.  이어 정신건강에 대한 사업 수행인력의 전문성이 부족해보이므로 #####레스에 정신건강간호사 등 정신건강전문요원 인력을 확충해야 한다고 강조했다.  ',
 '보건복지부는 21일 오후 서울 영등포구 여의도 글래드 호텔에서 마음투자 공개토론회로의 제3차 마음투자 정책콘서트를 개최했다.  이번 제3차 마음투자정책콘서트는 1부 발제와 2부 대담회로 진행되며 정신건강에 더 많은 관심과 지원이 필요하다는 메시지를 전달한다.  1부 발제에서는 코로나19 대유행 1인 가구 증가 고령화 등 환경변화를 고려한 정신건강 발전 방향과 정신건강 분야에 신기술을 접목하는 방법과 정신건강에 대한 사회적 관심과 마음 투자의 필요성 방법을 공유했다.  발제자 홍익철 교수서울대 경제학부 최연우 대표연합팀 스코리아 김도원 위원장대한민국 청소년 정신건강 위원회 유현재 교수서강대 신문방송학과가 각 마음 투자의 가치와 재증진Reboot 전략 우리들이 원하는 정신건강 투자 ITInformation Tecchnology가 마음투자를 만나면을 주제로 60분 동안 발제했다.  2부 대담회에서는 사전 질문 현장 참석자 유튜브 실시간 연결 등의 방법으로 국민의 다양한 목소리를 듣고 함께 소통하는 시간을 가졌다.  전명숙 과장보건복지부 정신건강정책과 이주현 기자한겨레신문 배배언니는 조울의 사막을 건넌이 저자 정재훈 공동회장국회 보건의료발전연구회 아주편한병원장이 패널로 참석했다.  ',
 '한덕수 국무총리는 1일 이태원사고와 관련 정부는 유가족과 부상자는 물론일반 시민들도 심리 상담과 치료를 받을 수 있도록 국가트라우마센터와 서울시 정신건강복지센터를 통해 적극 지원하겠다고 밝혔다.  한 총리는 이날 서울 세종로 정부서울청사에서열린 이태원 사고 중앙재난안전대책본부중대본 회의를 주재한 자리에서부의의 사고로 슬픔에 빠진 유가족 뿐만 아니라 현장에 계셨거나 뉴스를 통해 소식을 전한 많은 시민들께서 정신적으로 큰 충격을 받으셨다며 이 같이 말했다.  한덕수 국무총리가 1일 서울 세종로 정부서울청사에서 열린 이태원 사고 중대본 회의에서 발언하고 있다 사진=국무조정실 한 총리는 재발 방지를 위한 제도개선도 본격적으로 추진해 나가겠다며 현재 경찰청에서 명확한 사고원인을 규명하기 위해 조사와 분석이 진행 중이라고 설명했다.  그러면서 이를 토대로 주최자가 없는 자발적 집단행사에서도 시민들의 안전이 철저히 담보될 수 있는 방안을 마련하겠다며 이 과정에서 해외사례 등을 참고해 전문가들과 함께 과학적 관리기법도 모색하겠다고 덧붙였다.  이어 이번 사고로 어린 학생들의 피해도 컸다며 다중 밀집장소에서의 안전 수칙 등을 포함한 안전교육 강화방안을 마련해 안전교육이 내실있게 이뤄지도록 하겠다고 밝혔다.  한 총리는 전날 서울광장 합동분향소에서 조문한 것을 언급하며 유가족과 함께 슬픔을 나누고 위로해주고 계신 국민 여러분께 깊은 감사의 말씀을 드린다고 말했다.  또 일부 언론에서 자극적인 장면의 보도를 자제하는 조치를 취했고 이러한 움직임이 점차 확산하고 있다며 감사의 뜻을 전했다.  한 총리는 정부는 사고 수습과 돌아가신 분 다치신 분에 대한 지원에 총력을 기울이면서 미래세대에게 다시는 이런 사고가 없도록 제도개선에 최선을 다하겠다고 강조했다.  11일 오전 서울광장에 마련된 이태원 사고 사망자 합동 분향소에서 시민들이 조문하고 있다.  사진=저작권자c 연합뉴스 무단 전재-재배포 금지 이날 이태원 사고 중대본 브리핑에서는 유사한 사고를 예방하기 위해 오는 3일부터 지역축제에 대한 점부합동점검도 실시한겠다고 밝혔다.  브리핑에 나선 김성호 행정안전부 재난안전관리본부장은 사고 원인 조사를 위한 수사 진행과 재발 방지를 위한 대책도 추진 중에 있다면서 이같이 말했다.  이어 경찰과 국과수 합동 현장감식을 실시했고 이번 사례와 같이 주최자 없는 행사를 위한 안전관리방안도 마련해 나가도록 하겠다고 덧붙였다.  한편 정부는 정부는 오는 5일까지 국가애도기간을 지정해서 전국 지자체에서 총 59개소의 합동분향소를 운영하고 있고 유가족을 위한 세심한 지원을 계속하고 있다.  이에 유가족 전담 공무원을 1:1로 매칭해서 지원하고 있고 장례비는 유가족 주소지에 있는 지자체를 통해서 최대한 신속하게 지급될 수 있도록 하고 있다.  또한 화장시설도 부족함이 없도록 조치하고 있다.  아울러 학생 피해자가 많은 점을 감안해서 사상자가 있는 학교를 대상으로 심리치료와 정서상담도 실시하고 학생 안전을 위한 안전교육도 강화할 계획이다.  김 본부장은 거듭 이번 사고로 유명을 달리하신 분들의 명복을 빌며 유가족께 깊은 위로의 말씀을 드리고 부상자분들의 빠른 쾌유도 기원한다며 브리핑을 마쳤다.  출처 대한민국 정책브리핑### korea.kr']
```

KoBERTopic

```
1 !wget https://raw.githubusercontent.com/lovit/soynlp/master/tutorials/2016-10-20.txt
```

```
--2022-11-03 15:14:14--  https://raw.githubusercontent.com/lovit/soynlp/master/tutorials/2016-10-20.txt
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133,
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 43694449 (42M) [text/plain]
Saving to: '2016-10-20.txt'

2016-10-20.txt      100%[===================>]  41.67M   268MB/s    in 0.2s

2022-11-03 15:14:14 (268 MB/s) - '2016-10-20.txt' saved [43694449/43694449]
```

```
1 text_file = "2016-10-20.txt"
```

```
1 documents = [line.strip() for line in open(text_file, encoding="utf-8").readlines()]
```

```
1 preprocessed_documents = []
2
3 for line in tqdm(documents):
4   # 빈 문자열이거나 숫자로만 이루어진 줄은 제외
5   if line and not line.replace(' ', '').isdecimal():
6     preprocessed_documents.append(line)
```

```
100%|██████████| 30091/30091 [00:00<00:00, 140163.83it/s]
```

```
1 len(preprocessed_documents)
```

```
27540
```

27540개의 한국어 뉴스기사로 이루어진 데이터셋을 토픽모델링

Line 하나에 뉴스기사 하나가 담겨있는 데이터셋

preprocessed_documents

붙잡힌 직후 나 자살하려고 한 거다 맞아 죽어도 괜찮다 고 말한 것으로 전해졌다 성씨 자신도 경찰이 발사한 공포탄 1발 실탄 3발 중 실탄 1발을 배에 맞았으나 방탄조끼를 입은 상태여서 부상하지는 않았다 경찰은 인근을 수색해 성씨가 만든 사제총 16정과 칼 7개를 압수했다 실제 폭발할지는 알 수 없는 요구르트병에 무언가를 채워두고 심지를 꽂은 사제 폭탄도 발견됐다 일부는 숲에서 발견됐고 일부는 성씨가 소지한 가방 안에 있었다',
 '테헤란 연합뉴스 강훈상 특파원 이용 승객수 기준 세계 최대 공항인 아랍에미리트 두바이국제공항은 19일 현지시간 이 공항을 이륙하는 모든 항공기의 탑승객은 삼성전자의 갤럭시노트7을 휴대하면 안 된다고 밝혔다 두바이국제공항은 여러 항공 관련 기구의 권고에 따라 안전성에 우려가 있는 스마트폰 갤럭시노트7을 휴대하고 비행기를 타면 안 된다 며 탑승 전 검색 중 발견되면 압수할 계획 이라고 발표했다 공항 측은 갤럭시노트7의 배터리가 폭발 우려가 제기된 만큼 이 제품을 갖고 공항 안으로 들어오지 말라고 이용객에 당부했다 이런 조치는 두바이국제공항 뿐 아니라 신공항인 두바이월드센터에도 적용된다 배터리 폭발문제로 회수된 갤럭시노트7 연합뉴스자료사진',
 '브뤼셀 연합뉴스 김병수 특파원 독일 정부는 19일 원자력발전소를 폐쇄하기로 함에 따라 원자력 발전소 운영자들에게 핵폐기물 처리를 지원하는 펀드에 235억 유로 260억 달러 29조 원 를 지불하도록 하는 계획을 승인했다고 언론들이 보도했다 앞서 독일은 5년 전 일본 후쿠시마 원전사태 이후 오는 2022년까지 원전 17기를 모두 폐쇄하기로 하고 오는 2050년까지 전기생산량의 80 를 재생에너지로 충당하는 것을 목표로 세웠다 이날 내각을 통과한 법안은 원전 운영자들이 원전 해체와 핵폐기물 처리를 위한 포장을 책임지고 정부는 핵폐기물 보관을 책임지도록 했다 독일 경제부는 전력회사들과 공식적인 접촉은 아직 합의되지 않았다고 밝혔다 독일 원자력 발전소 연합뉴스 자료사진',

KoBERTopic

```python
class CustomTokenizer:
    def __init__(self, tagger):
        self.tagger = tagger
    def __call__(self, sent):
        sent = sent[:1000000]
        word_tokens = self.tagger.morphs(sent)
        result = [word for word in word_tokens if len(word) > 1]
        return result
```

```python
custom_tokenizer = CustomTokenizer(Mecab())
```

```python
vectorizer = CountVectorizer(tokenizer=custom_tokenizer, max_features=3000)
```

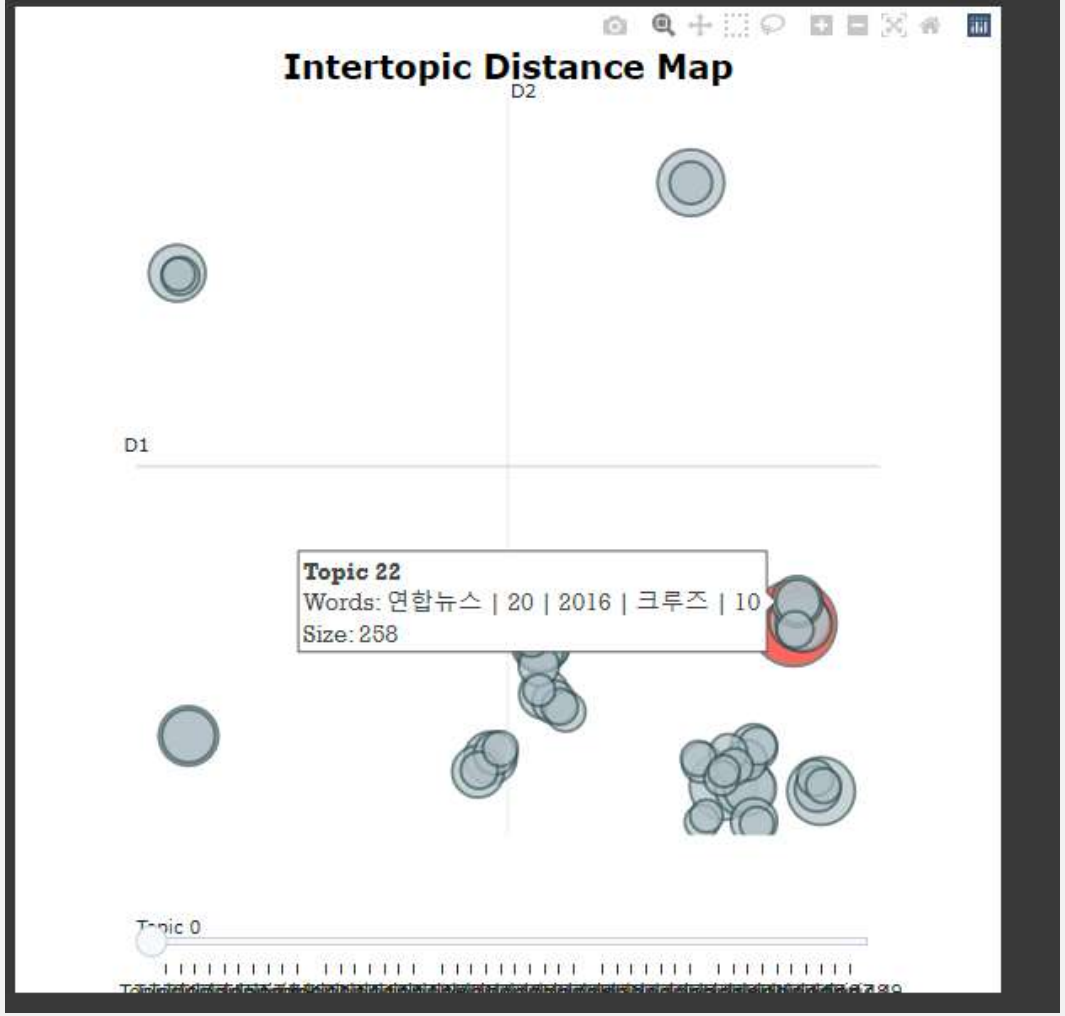
```python
model = BERTopic(embedding_model="sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens", ₩
                 vectorizer_model=vectorizer,
                 nr_topics=50,
                 top_n_words=10,
                 calculate_probabilities=True)
```

BERTopic 모델에 tokenizer를 Mecab으로 설정하고 BERT모델을 다국어 SBERT로 설정

KoBERTopic

```
topics, probs = model.fit_transform(preprocessed_documents)
```

```
model.visualize_topics()
```



Topic과 Topic Probability 도출

이후 계획

- 소셜 빅데이터 수집. 이후 불용어처리 등의 전처리

- BERT모델 기반 감정분석 구현
  - 블로그, 트위터 글에 대한 긍부정 감정 분류 라벨링 작업 수행 필요
  - 모델학습에 파라미터 튜닝

- 다른 모델과의 비교

- 다양한 한국어 BERT 모델 사용
  - BERT_multi(Google), KorBERT_Morphology/WordPiece(ETRI), KoBERT(SKT), HanBERT(TwoBlock AI), LMKor, KalBERT, DistilKoBERT ...

- 평가지표에 대한 연구
  - 토픽모델링: topic coherence, topic diversity...
  - 감정분석: accuracy...

- 꾸준히 선행연구논문 찾아 읽기