

2022 과학기술 · 공공 AI 데이터 분석활용 경진대회

# 여론 분석 AI 모델

- 이상 탐지와 감성 분석 모델 활용 -

---

Nov 16, 2022

AIM HIGH

**POSTECH**

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

# 목차

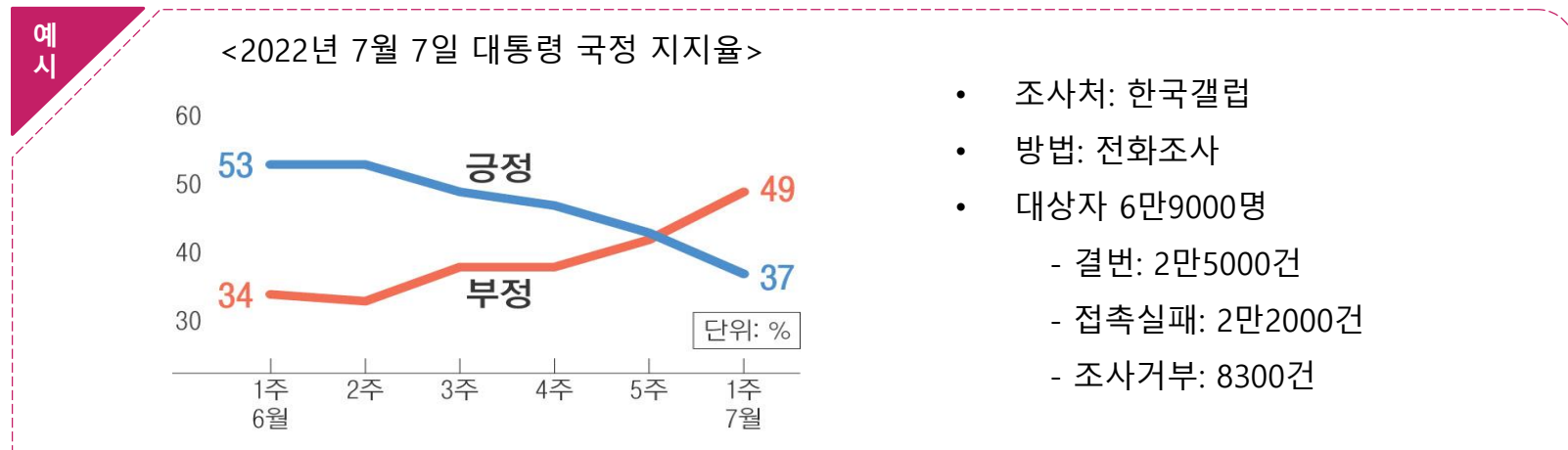
---

1. 프로젝트 개요
2. 활용데이터
3. 모델 개발 방법
  - 이슈 탐지
  - 여론 분석
4. 실험 및 평가
  - 평가 지표
  - 결과
5. 활용 계획 및 기대 효과
6. 시연

# 1. 프로젝트 개요

## • 연구 배경

- 다양한 매체 속 대중의 의견을 반영하는 여론 분석 시스템의 부재
  - 현재 이용되는 많은 여론 지표들이 효율성 및 신뢰성이 떨어짐
- 뉴스 뿐만 아니라 온라인 상에 나타난 국민들의 의견을 분석할 수 있는 모델 필요
  - 뉴스 생산자의 정치적 성향에서 벗어나 왜곡없이 국민들의 의견을 수렴할 필요성 존재
  - 온라인상의 정제되지 않은 표현들까지도 정확하게 분석할 수 있는 모델이 필요



# 1. 프로젝트 개요

- 여론 분석 AI 모델



## 시간과 비용을 최소화한

여론조사에 들어가는 시간과 비용,  
여론조사 결과로 인해 발생하는 사회적 비용을 최소화



## 온라인 상의 방대한 데이터를 활용한

트위터, 블로그, 유튜브, 커뮤니티 등 개인이 자신의 의견을  
자유롭게 표출할 수 있는 웹 상의 다양한 창구를 활용



## 사람의 주관에 최대한 배제한

조사자의 출신, 성향, 선입견에 좌우되지 않는  
객관적이고 신뢰성 있는 여론 분석 모델

## 2. 활용 데이터

### • 데이터 설명

#### ◦ 정책과 채널 기준으로 구분된 크롤링 데이터

- 정책: 임대차3법(54,752), 중대재해처벌법(40,0056), 차별금지법(59,420), 탄소중립 (59,295)
- 채널: 뉴스(75,825), 블로그(53,574), 커뮤니티(27,324), 트위터(56,800)
- 변수: 작성일, 작성자, 제목, 내용, 댓글수

#### ◦ 내용 작성자가 언론사인지 대중인지에 따라 크게 2가지로 재분류

- 언론사: 뉴스

작성일	작성자	제목	내용
2022/04/19	연합뉴스	경총 "중대재해처벌법 제정 후 안전 관련 예산·인력 모두 늘어"	중대재해처벌법 제정 이후 안전에 대한 기업의 관심도와 예산, 인력이 모두 늘었지만 여전히 대다수의 기업이 법 개정의 필요성을 느끼고 있는 것으로 나타났다.. (중략)

- 대중: 블로그, 트위터, 커뮤니티

작성일	작성자	제목	내용
2022/05/04	독취사	형님들 노동정책.. 제가 잘 몰라서 그러는데;;;	중대재해처벌법 없애고 기업에게 자율성을 주어서 중대재해를 예방 한다는데... 이게 가능한 부분인지.....????????...(중략)

## 2. 활용 데이터

### • 데이터 전처리

#### ◦ 키워드 맵핑 사전 생성

- 주요 단어들의 띄어쓰기 방식이 통일되지 않음

✓ Ex. 중대재해처벌법: 중대재해 처벌법, 중대 재해 처벌법, 중대 재해처벌법 등

- 키워드로 추출된 단어가 쓰일 수 있는 경우 리스트 자동 생성

✓ 형태소 단위 띄어쓰기 여부(0, 1)에 따른 경우의 수 리스트 생성

▪ Ex. [0, 0, 0], [0, 0, 1], [0, 1, 0], [0, 1, 1], [1, 0, 0], [1, 0, 1], [1, 1, 0], [1, 1, 1]

✓ 위의 경우의 수 리스트에 따른 단어 리스트 추출



### 3. 모델 개발 방법 - 개괄

#### STEP 1. 이슈 탐지

- 정책 여론 데이터 기반 이슈 탐지
- 입법 과정에 대한 Talking Point 탐지
- Talking Point별 주요 이슈 내용 추출

Anomaly Detection

Extractive  
Summarization

#### STEP 2. 여론 분석

- 정책 여론 데이터 기반 감성분석
- 해당 법안에 대한 대중의 반응을 5가지로 구분하여 분석함

Few-Shot Learning

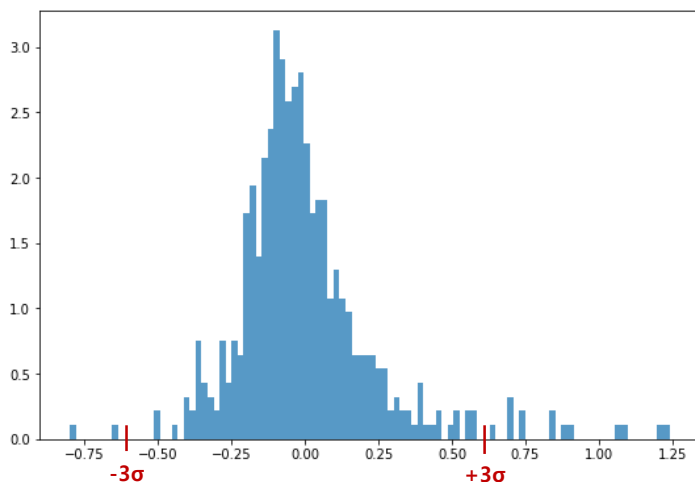
Sentiment Analysis

### 3. 모델 개발 방법 - 이슈 탐지

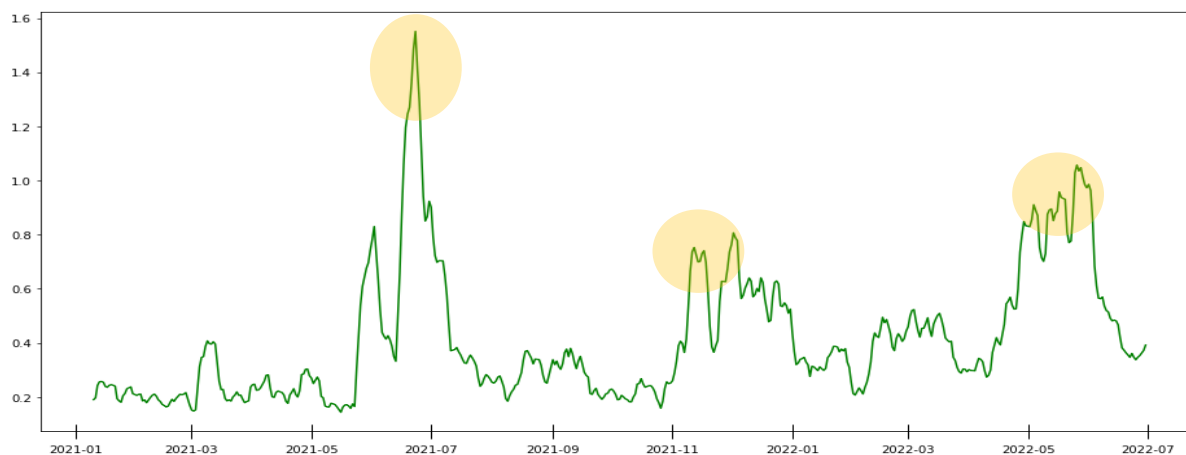
- 이상 탐지(Anomaly Detection)

- 이상 탐지란? 정상 범주에서 벗어난 데이터를 탐지하는 방법론
- **목적:** 각 사회 이슈별로 가장 화제가 된 날짜를 탐지하고자 함
- 화제가 된 날(Talking Point) 탐지를 위해 시계열 데이터 Anomaly Detection 진행
- 신뢰구간 이상치 탐색 방법으로 Talking Point 추출
- 차별금지법의 경우 신뢰구간 이상 탐지로 크게 3가지 Talking Point를 확인할 수 있었음

Talking Point 추출을 위한 신뢰구간



차별금지법의 게시글 빈도 분포와 Talking Point



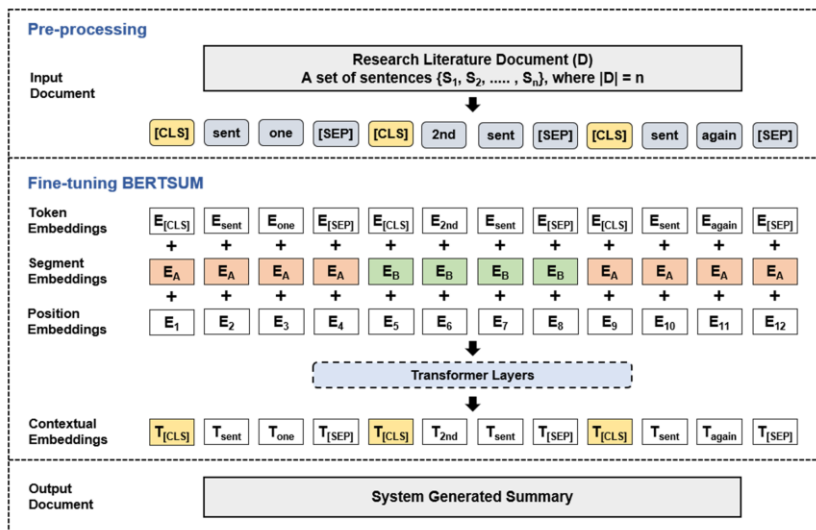


### 3. 모델 개발 방법 - 이슈 탐지

#### • 텍스트 추출 요약 (Extractive Summarization)

- 추출 요약이란? 원문에서 중요한 핵심 문장을 뽑아 이들로 구성된 요약문을 만드는 방법
- **목적:** 이슈별 Talking Point 를 추출하여 해당 날짜에 어떤 이슈가 발생하였는지 확인
- BERTSum을 통해 Talking Point 날짜에 출고된 뉴스 중 가장 댓글이 많은 뉴스의 본문을 요약 (AI 허브에서 제공하는 학습 데이터 2만개를 이용하여 모델을 Fine-tuning함)

#### Fine-tuning BERTSum 구조도



#### 텍스트 추출 요약 예시

##### 문서

'슬기로운 의사생활'이 또다시 최고 시청률을 경신하며 고공행진을 이어갔다. 26일 방송된 tvN 2020 목요스페셜 '슬기로운 의사생활'3회는 케이블, IPTV, 위성을 통합한 유료플랫폼에서 가구 평균 8.6%, 최고 10%의 시청률을 기록했다. 3주 연속 시청률 상승세다.

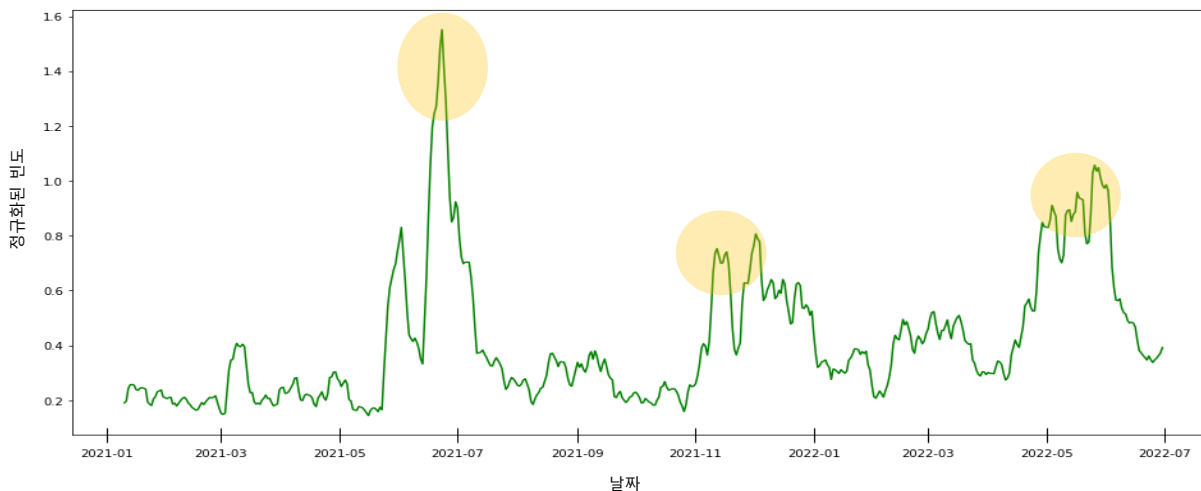
##### 모델이 추출한 문장

'슬기로운 의사생활'이 또다시 최고 시청률을 경신하며 고공행진을 이어갔다.

### 3. 모델 개발 방법 - 이슈 탐지

- 이슈 탐지 및 요약 결과

<차별금지법 Talking Point 뉴스 텍스트 추출요약>



**[2021-06-14]**

**차별금지법 제정**을 요구하는 국회 국민동의청원이 10만명 동의를 얻어 국회 **심사**를 받게 됐다. 2007년 첫 발의 뒤 14년간 국회 문턱을 넘지 못한 채 표류해 왔던 차별금지법 논의가 국회청원을 계기로 동력을 얻을 수 있을지 주목된다.

**[2021-11-10]**

**문재인 대통령**은 25일 차별금지법과 관련해 "우리가 인권선진국이 되기 위해 반드시 넘어서야 할 과제"라고 밝혔다. 보수 기독교계가 강하게 반대하는 **차별금지법 제정 필요성을 언급**한 것이다.

**[2022-05-11]**

**트랜스젠더** 방송인 **하리수씨**가 **더불어민주당 지도부를 만나** 15년째 국회 문턱을 넘지 못하고 있는 차별금지법 제정을 촉구했다. 민주당 지도부는 하루빨리 공청회를 개최해 차별금지법 공론화에 속도를 내겠다고 약속했다.

### 3. 모델 개발 방법 - 여론분석

- 감성 분석 (Sentiment Analysis)

- 감성 분석이란? 텍스트에 포함된 견해, 감정, 태도 등 주관적인 정보를 분석하는 방법
- **목적:** 각 사회 이슈 별 Talking Point를 기점으로 여론 전후변화 비교

감성 분류 기준 5가지 → 완전 찬성, 약간 찬성, 중립, 약간 반대, 완전 반대

- ✓ 완전 찬성: 법안에 적극적으로 동의하며 주변에 찬성을 권유하는 상태
- ✓ 약간 찬성: 법안에 기본적으로 동의하지만 일부 수정, 보완을 원하는 상태
- ✓ 중립: 법안에 찬성, 반대 중 어떤 의견도 없는 상태 (질문, 단순 정보 전달 등 포함)
- ✓ 약간 반대: 법안에 기본적으로 반대하지만 수정, 보완된다면 받아들일 수 있는 상태
- ✓ 완전 반대: 법안에 적극적으로 반대하며 주변에 반대를 권유하는 상태

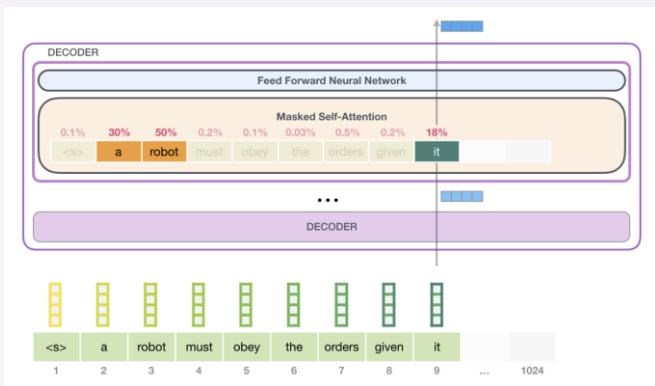


# 3. 모델 개발 방법 - 여론분석

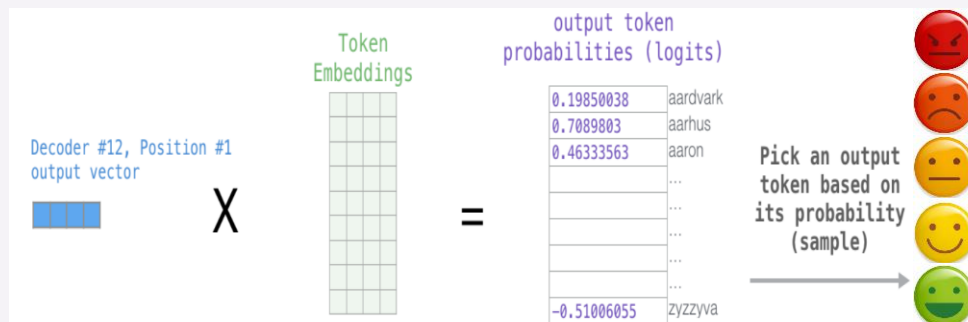
## • Few-Shot Learning & GPT-2

- 매우 적은 양의 데이터로 패턴을 학습하여 모델 파라미터를 효율적으로 학습하는 방법
- 감성을 분류하기 위해 GPT-2 Medium 모델을 Few-Shot Learning시킴
  - GPT-2: Transformer-based Modeling을 통해 여러가지 task에서 fine-tuning시 SOTA 달성
- 모델 파라미터가 클수록 좋은 성능을 발휘하여 자연어처리 모델링에 잘 부합함
- 여론 분류에 객관성을 유지하기 위해 채택
  - 5개 클래스에 대해 각각 20개 정도의 텍스트 정답 레이블 힌트만 줌

### GPT-2 Medium



### Few-shot Learning and Text Classification



# 4. 실험 및 평가

- 결과 비교

- 감성 분석 결과

- 모델이 종결 어미에 따라 바뀌는 문장의 미묘한 감정까지 정확하게 구분
- 예시: 임대차3법

```
prompt = "다주택자분들 어떻게 하실 예정이신가요?"  
pipe(prompt, top_k=1, **tokenizer_kwargs)
```

```
[{'label': '중립', 'score': 1.0}]
```

```
prompt = "다주택자분들 어떻게 하실 예정이신가요"  
pipe(prompt, top_k=1, **tokenizer_kwargs)
```

```
[{'label': '중립', 'score': 0.789466142654419}]
```

```
prompt = "다주택자분들 어떻게 하실 예정이신가요... 휴"  
pipe(prompt, top_k=1, **tokenizer_kwargs)
```

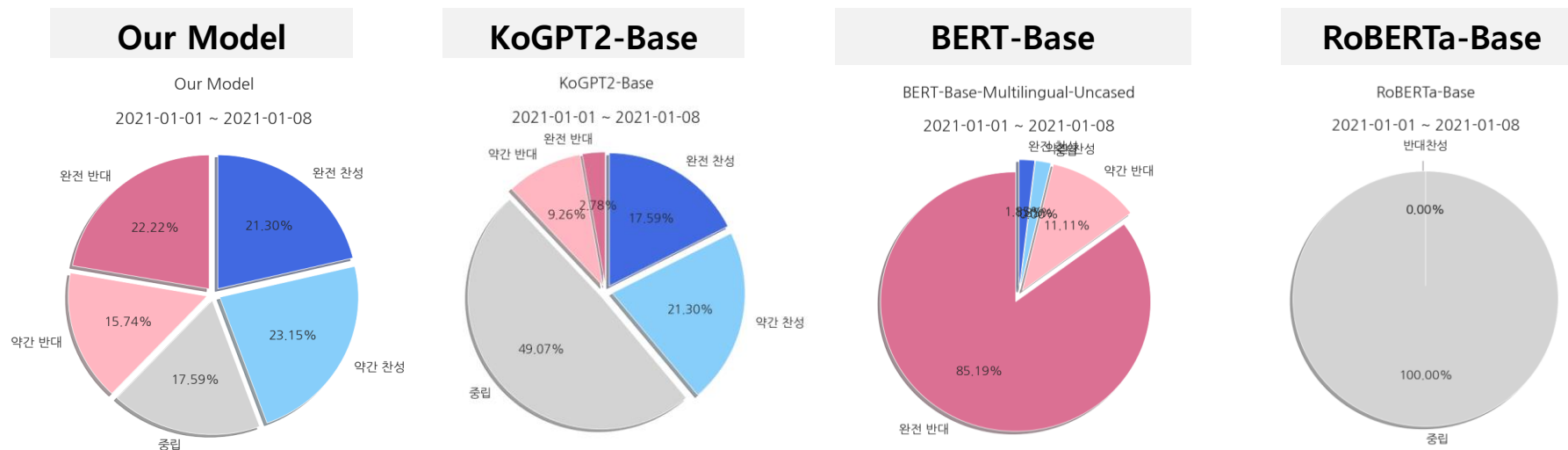
```
[{'label': '약간 반대', 'score': 0.9997629523277283}]
```

# 4. 실험 및 평가

## • 결과 비교

### ◦ 감성 분석 결과

- 예시: 중대재해처벌법 (2021.01.08 – 법안 본회의 통과일)
- 기사 제목: "[2보] '산재사망에 경영자 처벌' 중대재해법 본회의 통과"
- 기사 본문 요약: 국회는 8일 본회의를 열어 산업재해에 기업과 경영자 처벌을 강화하는 중대재해 처벌법 제정안을 의결했다. 산재나 사고로 노동자가 숨지면 해당 사업주나 경영 책임자는 1년 이상 징역이나 10억원 이하 벌금으로 처벌받게 된다. 법인이나 기관도 50억원 이하의 벌금형에 처해진다.



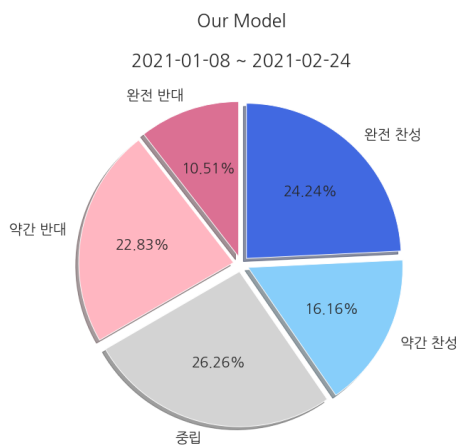
# 4. 실험 및 평가

## • 결과 비교

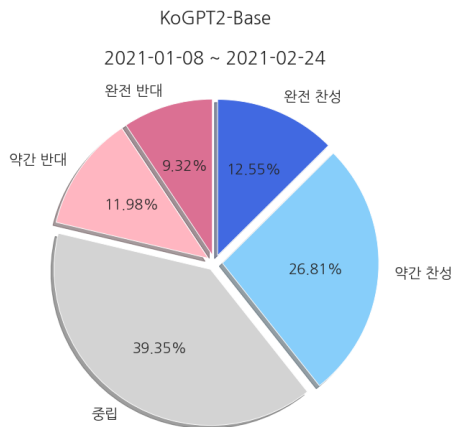
### ◦ 감성 분석 결과

- 예시: 중대재해처벌법 (2021.01.08 – 법안 본회의 통과일)
- 기사 제목: "[2보] '산재사망에 경영자 처벌' 중대재해법 본회의 통과"
- 기사 본문 요약: 국회는 8일 본회의를 열어 산업재해에 기업과 경영자 처벌을 강화하는 중대재해 처벌법 제정안을 의결했다. 산재나 사고로 노동자가 숨지면 해당 사업주나 경영 책임자는 1년 이상 징역이나 10억원 이하 벌금으로 처벌받게 된다. 법인이나 기관도 50억원 이하의 벌금형에 처해진다.

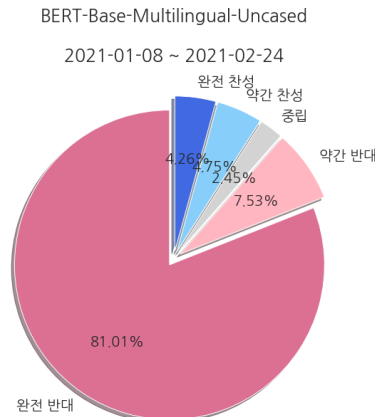
### Our Model



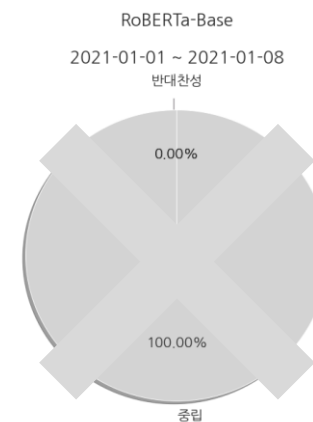
### KoGPT2-Base



### BERT-Base



### RoBERTa-Base



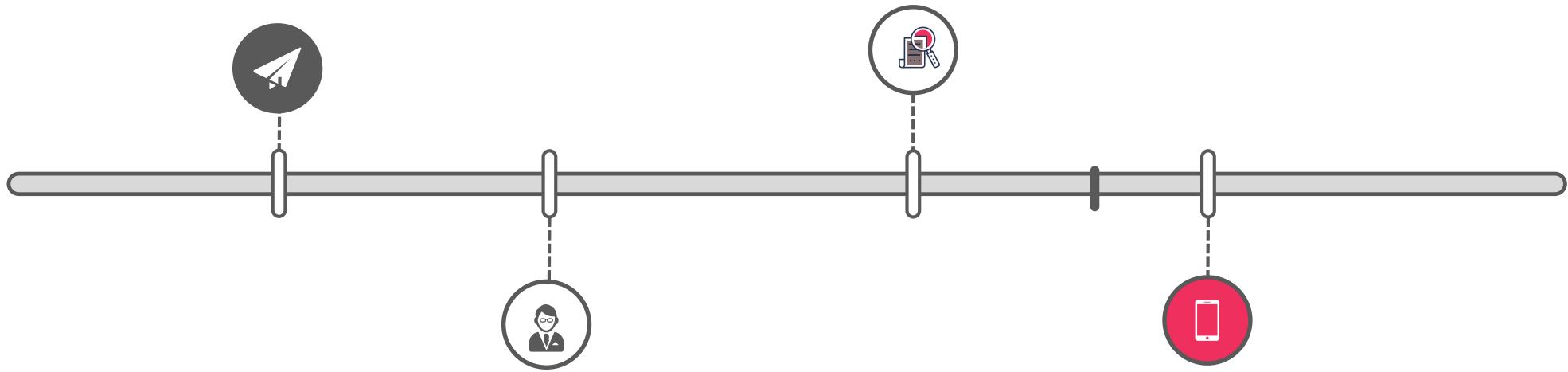
# 5. 활용 및 기대 효과

## 여론을 반영한 법안 구성

특정 사안에 대한 대중들의 여론을  
취합하여 정책 아젠다 셋팅에 반영

## 정책 효과 분석

정책이 시행된 후 실제 국민들의 생활에서  
얼마나 효과적으로 의도에 맞게 적용되는지 확인



## 발의된 법안의 여론 분석

발의된 법안에 대해 찬/반 등  
대중의 감정 온도 파악

## 각 정부 부처의 국정 홍보에 활용

특정 사안에 대한 대중의 관심 정도나 이해  
수준을 파악하여 국정 홍보 전략에 활용



# 동영상

1. Talking Points Extraction
2. Summarization
3. Sentiment Analysis

# Thank You

---

AIM HIGH

***POSTECH***  
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY