SEOUL NATIONAL UNIVERSITY
GRADUATE SCHOOL OF DATA SCIENCE
SNU GSDS 이 윤 섭

자연어처리 기술을 활용한 여론 분석모델

□ 제안 배경

○ **국회 · 정부**, 여론조사 및 언론매체를 통해 사회 주요 이슈에 대한 **국민 여론을 적시에 탐지**하고, **대처**하기 위해 다각적인 노력 中

○ 한편, 전문인력(리서치 기관)의 조사 및 다양한 설문기법을 토대로 **사회문제에 대한 국민 여론 변화**를 분석하고 있으나, **객관적인 예측에 한계**가 있고, **해결책 마련과 입법 · 정책 연결까지 상당한 시간과 비용**이 발생

* 20대 국회, 전체 법안 가결률 13.2% / 평균 처리시간 577.2일 소요(국회입법조사처)

○ 이에, 본 프로젝트에서는 ❶ **효과적인 온라인 上 여론예측**과 ❷ **빠르고 정확한 주요 뉴스의 쟁점 추출이 가능한 두 가지 『자연어처리(NLP) 기반 인공지능 모델』을 제안**하며, 다양한 분야의 정책 결정 도구로 활용되어 國益에 기여할 것으로 기대

TASK ❶ : 『온라인 환경에서 긍 · 부정 여론 예측모델』

키워드 및 핵심쟁점 추출
TASK ❷ : 『클러스터링 기반 사회 주요 뉴스 토픽 모델링』

□ 활용 데이터

【데이터 수집】

- 국회, 주요 입법 관련 온라인 기사 · 트위터 · 커뮤니티 데이터 제공
  * 4대 주요 입법 : 임대차 3법, 중대재해처벌법, 차별금지법, 탄소중립법
- 언어모델 파인튜닝을 위한 **온라인 댓글** 및 **문장 유사성** 데이터 추가 수집
  * **(TASK ①)** 긍 · 부정 여론 분류 모델의 성능 향상을 위한 온라인 댓글 학습 데이터
  * **(TASK ②)** 클러스터링 및 토픽 모델링 성능 향상을 위한 NLI & STS 학습 데이터

□ 데이터 전처리 · 모델평가 전략

【텍스트 데이터 전처리 전략】

- 텍스트 內 한자(漢字)는 한글로 변환하고, 이모지 · 특수문자 · 해시태그 등 **다양한 온라인 용어를 필터링하는 함수 구현**
- 띄어쓰기를 유지하며, 한글 · 영어 · 숫자만으로 텍스트 再구성

【수치 데이터 정규화 전략】

- 수치 데이터 內 서로 다른 지표들을 비교하기 위해 정규화 과정이 필요하며, **선형함수 정규화**(Min-Max Scaling)와 **표준 정규화**(Z-score Norm)가 대표적인 방법

  * Min-Max Scaling : 데이터에 선형변환을 진행하여 결괏값이 [0, 1] 범위에 투영
  * Z-Score Norm : 데이터를 평균이 0이고 표준편차가 1인 정규분포 범위에 투영

- 본 분석의 **TASK ❶** 에서는 **선형함수 정규화**를 통해 시계열 데이터(수치)로 변환된 긍 · 부정 규모를 정규화

【모델평가 전략】

- 모델의 평가 방법으로 홀드아웃(Holdout) 검증 K-fold 교차 검증 등이 있으며, 본 분석에서는 **홀드아웃 검증** 방법론 채택

  * Holdout : 전체 데이터를 9:1 비율로 나누고, 90%는 훈련에 10%는 모델 검증에 사용
  * K-fold : k개의 하위 샘플로 나누고, k개의 샘플을 순차적으로 검정, 나머지는 훈련에 사용

- **긍 · 부정 텍스트 분류 언어모델**의 경우 **정확도**(Accuracy), **재현율**(Recall) **그리고 F1-score**를 모델 성능 평가 도구로 사용
- **트랜스포머 기반 여론 수치 예측모델**의 경우 **실제값과 모델 예측값의 차이**(RSS)를 평가 지표로 사용

TASK ❶ 온라인 여론 예측모델

□ 온라인 여론 예측모델 파이프라인

○ 우리가 디자인한 『**온라인 여론 예측모델**』의 전체 프로세스를 요약하면 다음과 같은 **4단계 과정**으로 진행

❶ **감성분석 말뭉치 준비** : 긍·부정 라벨링 된 트위터 및 댓글 텍스트

BERT-based Text Embedding Classifier

❷ **텍스트 임베딩 분류모델 파인튜닝** : 3가지 방식으로 텍스트 임베딩(또는 벡터)을 생성·분류할 수 있도록 설계한 BERT 기반 언어모델을 학습

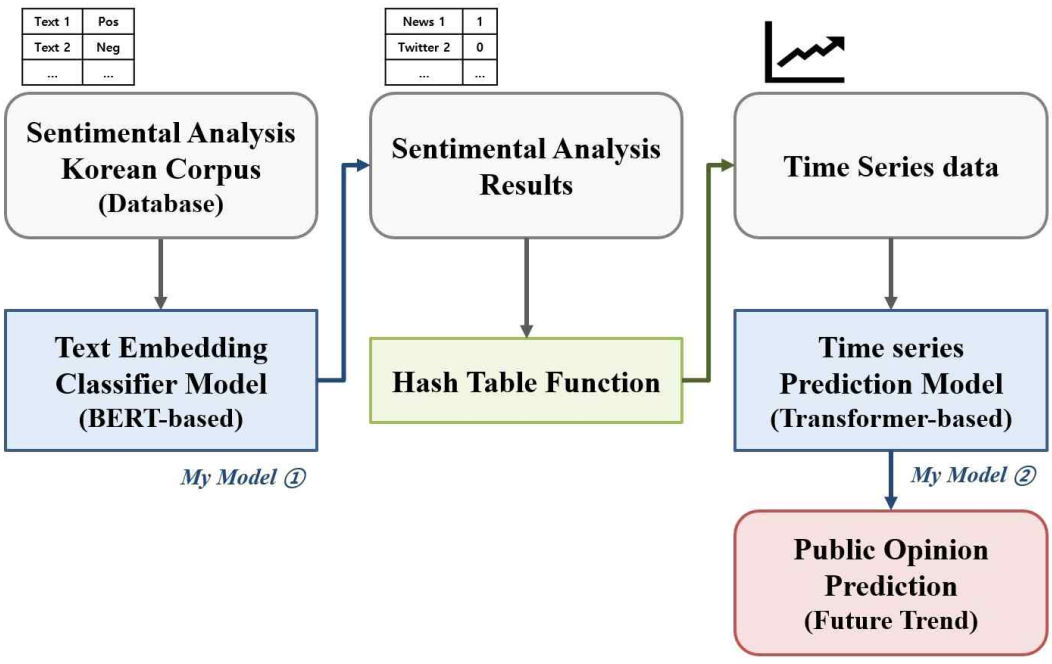
Convert predictions to time series

❸ **시계열 데이터 변환** : 언어모델의 긍·부정 예측값을 시계열 테이블로 변환

Transformer-based Time series Prediction Model

❹ **트랜스포머 시계열 예측모델 적용** : Transformers 기반으로 설계한 시계열 데이터 예측모델로 학습 後 미래의 긍·부정 여론 추이 예측

《 Online Public Opinion Prediction Model Procedure 》

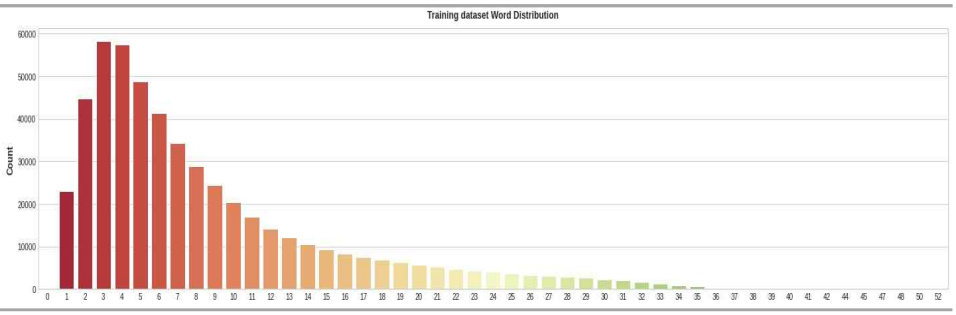


□ 모델 관련 세부 내용

【STEP 1 : 감성분석 말뭉치 준비】

- 국회에서 제공한 뉴스 제목, 트위터 그리고 감성분석을 위해 자체 수집한 온라인 댓글 등 **총 53만개의 텍스트 데이터**를 종합 및 전처리
  * 최소 1개 ~ 최대 60개의 단어로 구성된 말뭉치이며, 2 ~ 9개 단어 빈도가 높음.
- 해당 텍스트 데이터를 긍 · 부정 여론에 따라 라벨링(부정 : 0, 긍정 : 1)

Text length distribution



Labeling

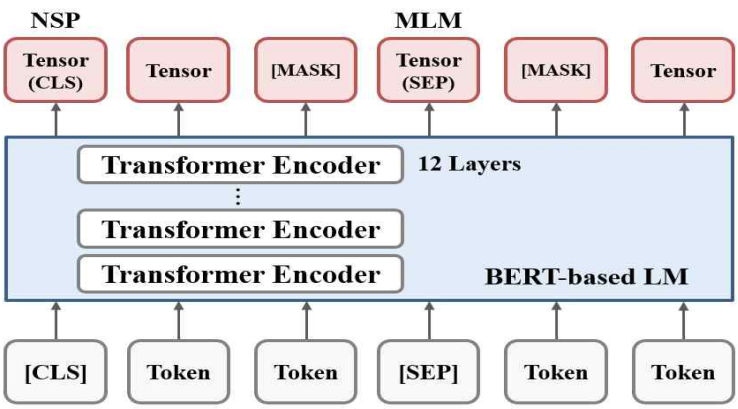
	Text	Sentiment
522514	임대차 3법은 오히려 지금 필요한데, 인수위에서 거론되는 것은 임대차3법을 폐지하겠...	1
522515	윤호중, 인수위 임대차3법 폐지 축소에 대단한 흔란 올 것.	1
522516	임대차 3법 아마 지금 한바퀴 돌아 효과나오기 직전일건데 그걸 또 없앤다고.	1
522517	윤석렬정부의 임대차3법 폐지는 민생파기, 안정적인 전월세 또 급등 우려. 민간임대부...	1
522518	집이나 상가 임대하여 살거나 영업하시는 분들 열 받게 생겼네. 임대차 3법을 윤석열...	1

【STEP 2 : 텍스트 임베딩 분류모델】

- BERT 모델은 **트랜스포머 인코더**로 구성되어 있으며, MLM · NSP 두 가지 태스크를 통해 텍스트에서 단어(하위 단어) 간의 **문맥 관계**(Contextual relations)를 **사전학습한 언어모델**(PLMs)

  Masked Language Modeling
  * MLM : 문맥을 활용하여 랜덤으로 마스킹된 단어(15%)를 예측하며 학습하는 방법

  Next Sentence Prediction
  * NSP : 2개의 문장이 연속적인 문장인지 아닌지를 구분하는 분류 학습 방법



- 사전학습된 BERT 기반 언어모델(PLMs)을 활용해서 **고정된 크기의 텍스트 벡터값**(Fixed-size contextual vector)을 얻을 수 있으며, 이는 **단어(토큰) 수준의 임베딩**(Token-level Embedding)을 의미

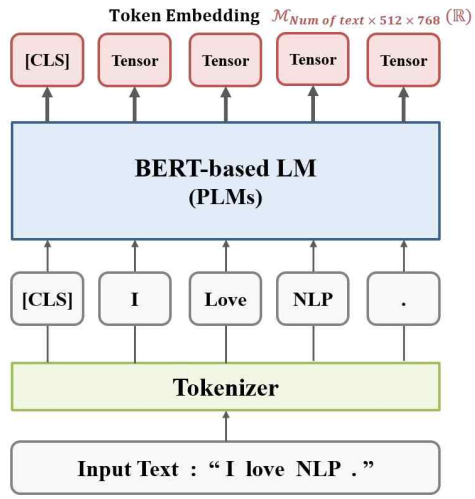
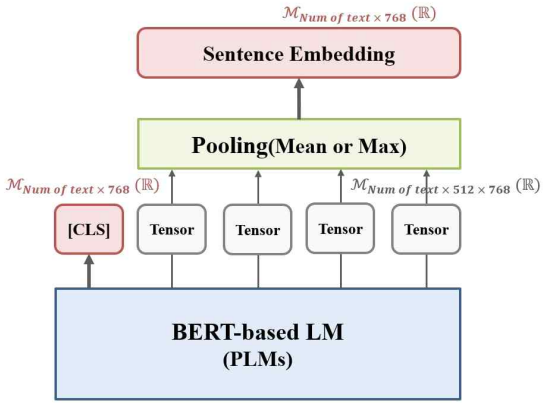
  * 여기서, 토큰 단위 임베딩은 각 어휘 요소를 벡터로 표현한 방법을 학습한 결과값

- 단어 단위가 아닌 **문장 단위 임베딩**(Sentence-level Embedding)을 얻기 위해 **[CLS] 토큰**을 사용하거나, 토큰 임베딩에 **풀링**(Pooling) **기법**을 적용

  ❶ [CLS] Token : 문장 內 전체 토큰의 의미가 포함된 단어 단위 벡터

  ❷ Mean Pooling : 전체 토큰의 의미표현이 종합된 문장 단위 벡터

  ❸ Max Pooling : 중요한 토큰의 의미표현이 종합된 문장 단위 벡터

Token Embedding (Fixed-size contextual vector)	**Sentence Embedding**
	

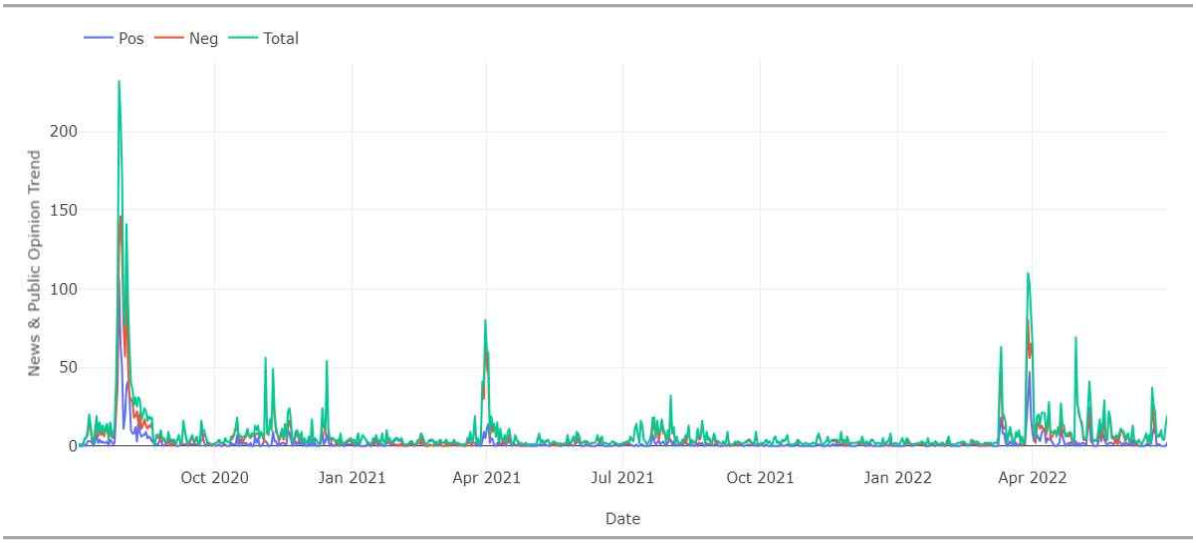
- **『텍스트 임베딩 분류모델』**은 단순한 분류모델이 아닌, 앞서 소개한 **3가지 방식의 임베딩에 대한 이해**를 바탕으로 긍·부정 문장(감성분석 말뭉치)을 학습하고 추론할 수 있는 언어모델

  * **모델 성능평가 결과** : 3가지 방식 모두 91~92% 이상의 텍스트 분류 정확도 ☞

Model type	Precision		Recall		F1-score	
	Neg	Pos	Neg	Pos	Neg	Pos
[CLS] Token	0.91	0.91	0.90	0.92	0.91	0.91
Mean Pooling	0.91	0.91	0.90	0.92	0.91	0.92
Max Pooling	0.92	0.91	0.91	0.91	0.91	0.91

【STEP 3 : 시계열 데이터 변환】

- **『텍스트 임베딩 분류모델』**을 통해 '임대차 3법' 관련 기사 및 트위터('20.1.~'22.6월)를 긍·부정 결과값으로 분류(예측) 後 **『해시 테이블 함수』**를 통해 **시계열 데이터(시간에 따른 수치 변화)**로 변환



【STEP 4 : 트랜스포머 기반 시계열 예측모델】

- **트랜스포머**(Transformer), 어텐션 메커니즘을 적용하여 기존 RNN 기반 모델들이 직면한 문제들을 해결하고, 계산속도를 대폭 향상
  * 타임 스텝이 길어질수록 초기 입력 정보를 소실하는 ❶ **장기의존성 문제**와 역전파 과정(Backpropagation)에서 반복적인 가중치 행렬 곱에 따른 ❷ **기울기 소실 문제**

Long-term Dependency
Vanishing Gradient

- 특히, **어텐션**(Attention)은 트랜스포머의 핵심 개념이며, 이를 통해 **모델의 신경망이 문맥 정보**(Contextual Information) **이해가 가능**하며, 현재 위치의 단어(토큰)와 유사한 단어들에 집중하면서 학습 및 추론
  * 주어진 데이터 內 유사도를 계산(Dot-Product)하여 **Attention Score**를 구하고, **해당 점수가 높은 순으로 가중치를 부여**하여 현재 타임스텝과 밀접한 데이터에 집중

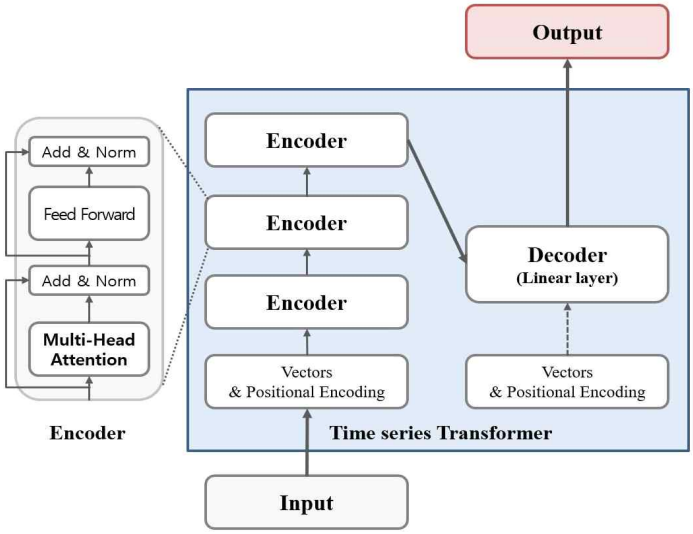
- 우리의『시계열 예측모델』은 **3개의 트랜스포머 인코더**를 쌓고, **1개의 선형회귀 모델을 디코더**로 구성한 Seq2Seq 모델
  * **(d_model = 512)** : 인코더와 디코더에 **동일한 입·출력 512 차원** 유지
  * **(multi_head = 8)** : 입력값을 **8개로 분할**하여 **병렬로 어텐션을 수행**하여 성능 향상

Attention Algorithm

Input: $e \in \mathbb{R}^{d_{\text{in}}}$, vector representation of the current token
Input: $e_t \in \mathbb{R}^{d_{\text{in}}}$, vector representations of context tokens $t \in [T]$.
Output: $\tilde{v} \in \mathbb{R}^{d_{\text{out}}}$, vector representation of the token and context combined.
Parameters: $W_q, W_k \in \mathbb{R}^{d_{\text{attn}} \times d_{\text{in}}}$, $b_q, b_k \in \mathbb{R}^{d_{\text{attn}}}$, the query and key linear projections.
Parameters: $W_v \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $b_v \in \mathbb{R}^{d_{\text{out}}}$, the value linear projection.

$q \leftarrow W_q e + b_q$
$\forall t: \; k_t \leftarrow W_k e_t + b_k$
$\forall t: \; v_t \leftarrow W_v e_t + b_v$
$\forall t: \; \alpha_t = \frac{\exp(q^\top k_t / \sqrt{d_{\text{attn}}})}{\sum_u \exp(q^\top k_u / \sqrt{d_{\text{attn}}})}$
return $\tilde{v} = \sum_{t=1}^{T} \alpha_t v_t$

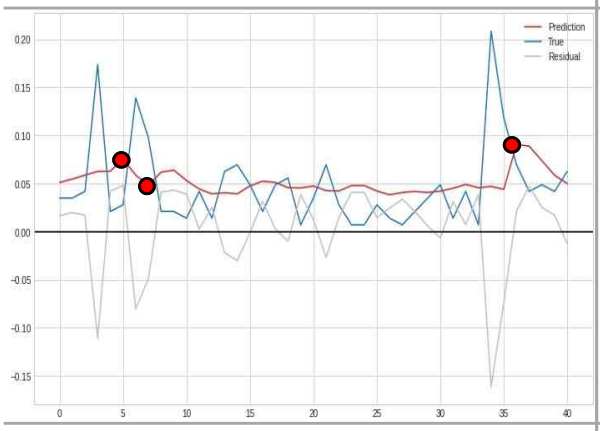
Time Series Prediction Model



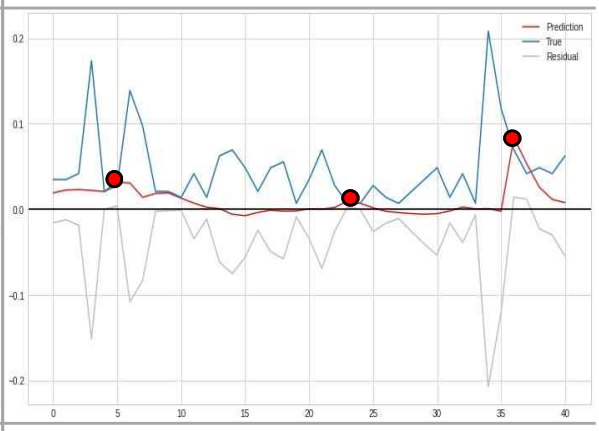
- 『시계열 예측모델』 훈련 결과, 입력 데이터의 **주요 반등·반락 변곡점에 집중(큰 가중치 부여)**하며 **세밀한 시계열 패턴을 학습**

  * (파란색 라인) 실제값, (붉은색 라인) 모델 예측값, (회색 라인) 잔차(실제값 - 예측값)

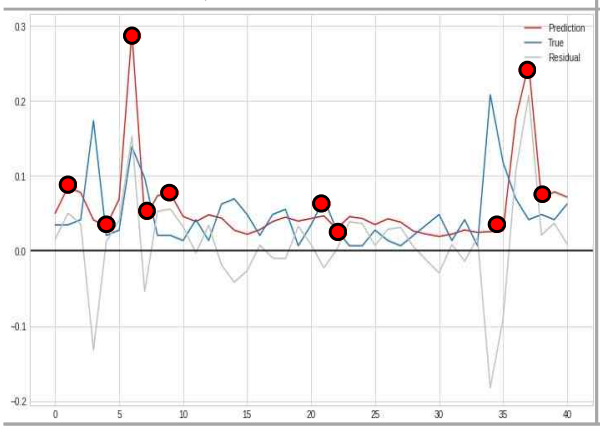
100 Iteration



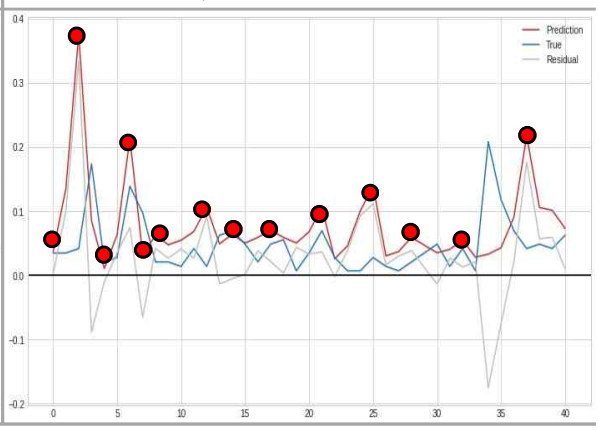
500 Iteration



1,000 Iteration



1,500 Iteration



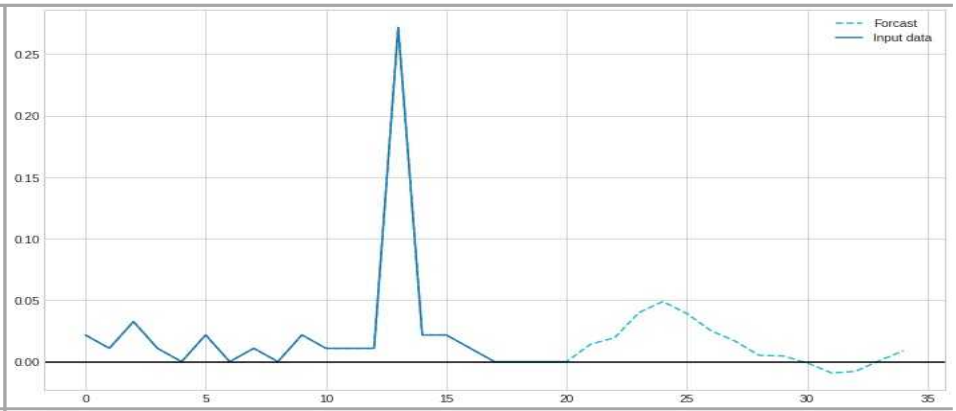
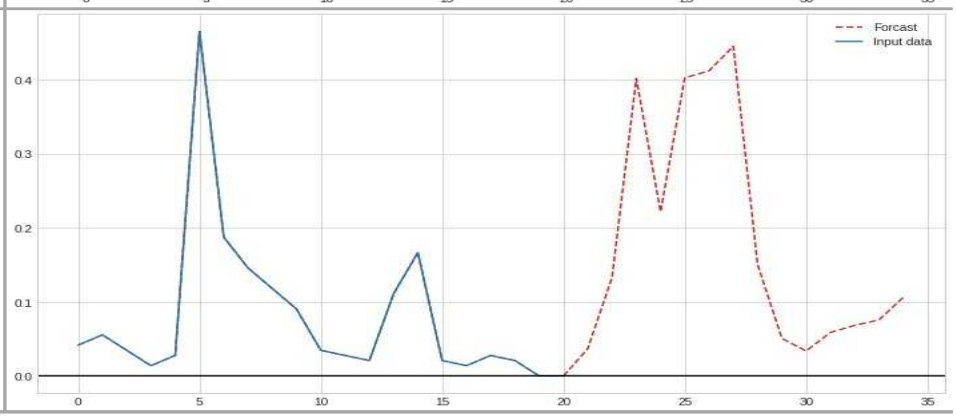
□ 온라인 여론 예측모델의 예측 결과

○ 훈련된 모델로 **향후 2주간 '임대차 3법'** 관련 트렌드 예측 결과,

* 사용자가 모델 훈련시 사용할 타임스텝(input & output_window)과 모델 예측시 미래 타임스텝(prediction step)을 자유롭게 설정할 수 있게 디자인

- **긍정 트렌드**는 1차례 소폭 증가 · 감소 後 **낮은 수준 유지 전망**
- **부정 트렌드**는 2차례 급격한 증감을 반복 後 **완만한 증가세 전망**

《 My Model Prediction Graphs 》

Positive Opinion Trend	
Negative Opinion Trend	

□ Github

○ 깃허브를 통해 우리가 디자인한 『**자연어 처리 기술을 활용한 여론 예측모델**』 소개 및 개발 코드 공개

* https://github.com/Navy10021/Public_Opinion_Prediction

TASK ❷ 클러스터링 기반 뉴스 토픽 모델링

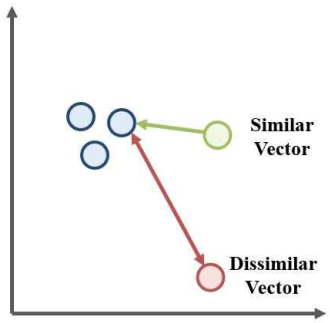
□ 배경 지식

○ **토픽 모델링**(Topic Modeling) 이란? 텍스트 데이터를 구성하는 단어들의 패턴을 분석하여 **문서 집합에서 주제 또는 쟁점 등을 자동으로 추출하는 비지도 학습 프로세스**

○ **클러스터링 기반 토픽 모델링**은 클러스터링 아키텍처에 언어 모델의 임베딩(Embedding)을 접목한 방법론

○ 본 태스크에서는 **병렬 클러스터링**과 **의미론적 임베딩**(Semantic Embedding)을 활용하여 토픽 모델링 모델을 구현

* 벡터 공간 上 **의미적으로 유사·관련된 문장들끼리 가깝게**(Euclidean-dist) 또는 **벡터 유사도**(Cosine-sim)**가 높도록 훈련된 언어모델**(PLMs)이 생성한 **임베딩**



□ 병렬 클러스터링 기반 뉴스 토픽 모델링 파이프라인

○ 우리가 디자인한 『**병렬 클러스터링 기반 뉴스 토픽모델링**』의 전체 프로세스를 요약하면 다음과 같은 **4단계 과정**으로 진행

Unsupervised-training Language model

❶ **언어모델 비지도 학습** : 언어모델이 주어진 뉴스 기사의 문맥을 이해하고, 도메인에 최적화하기 위해 TSDAE 방식의 비지도 학습을 수행

* TSDAE : Transformer-based and Sequential Denosing Auto-Encoder

Fine-tuning on NLI & STS dataset

❷ **언어모델 지도 학습** : 언어모델이 문장·문서 간의 유사성을 구분하고, 의미론적 임베딩을 생성할 수 있도록 한국어 NLI·STS 데이터셋을 학습

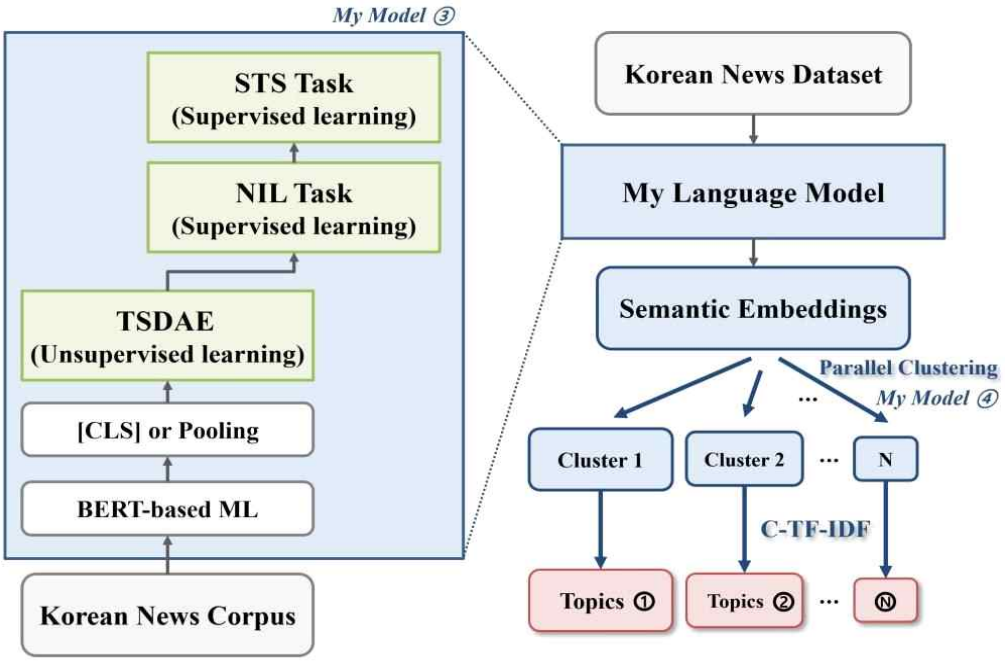
Parallel Clustering

❸ **병렬 클러스터링** : 속도와 안정성에 초점을 두고 직접 설계한 클러스터링 방법

Keyword extraction

❹ **키워드(주제) 추출** : C-TF-IDF(Class-based Term Freq-Inverse Doc Freq) 계산법을 활용하여 클러스터링된 그룹에서 중요한 단어들 추출

《 Parallel Clustering-based Topic Modeling Procedure 》



□ 모델 관련 세부 내용

【STEP 1 : 언어모델 비지도 학습】

- **언어모델이 도메인(뉴스)에 최적화된 임베딩을 생성**하기 위해 ① **레이블이 지정된 대용량의 데이터를 마련**하거나, ② 트랜스포머 기반 **사전학습 모델(PLMs)을 직접 디자인하고 대규모 학습을 진행**하는 등 상당한 시간과 노력이 필요

- 이에, 레이블이 지정되지 않은 채 수집된 **뉴스기사 데이터에 최적화**하고, 동시에 **의미론적 임베딩을 생성**할 수 있도록 언어모델을 미세조정하는 **『TSDAE 비지도 학습 방법론』**을 적용

  Transformer-based and Sequential Denoising Auto-Encoder

- TSDAE 비지도 학습 방법은 다음 3가지 단계로 진행

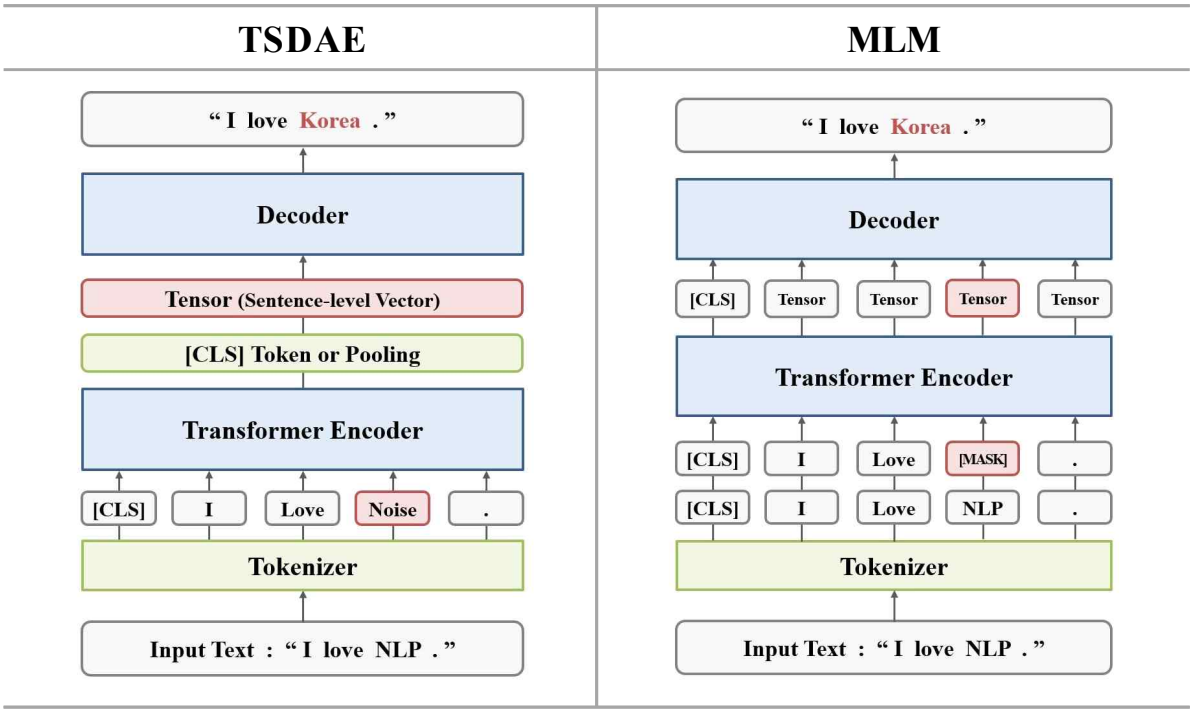
  ❶ 입력 시퀀스(뉴스 문장) 內 토큰을 **삭제 또는 교체**하여 **노이즈(Noise)**를 생성

  ❷ 손상된 입력 시퀀스는 **트랜스포머 인코더(Encoder) 네트워크**를 통해 문장 수준의 벡터(Sentence Embedding)로 인코딩

  ❸ **디코더(Decoder) 네트워크**는 손상된 문장 임베딩을 원래 **손상 이전 입력 시퀀스로 복원 및 예측**하며 학습

- 해당 학습 방법은 BERT 기반 언어모델의 MLM(Masked-language modeling) 사전학습 방법론과 유사하나,
- **MLM 태스크**의 경우 **디코더 네트워크**가 인코더로부터 생성된 **모든 단일 토큰 벡터(Token Embedding)에 접근**할 수 있지만, **TSDAE 디코더**는 인코더에 의해 생성된 **문장 수준의 벡터(Sentence Embedding)에만 접근**할 수 있다는 차이점 존재

《 TSDAE Task VS. MLM Task 》

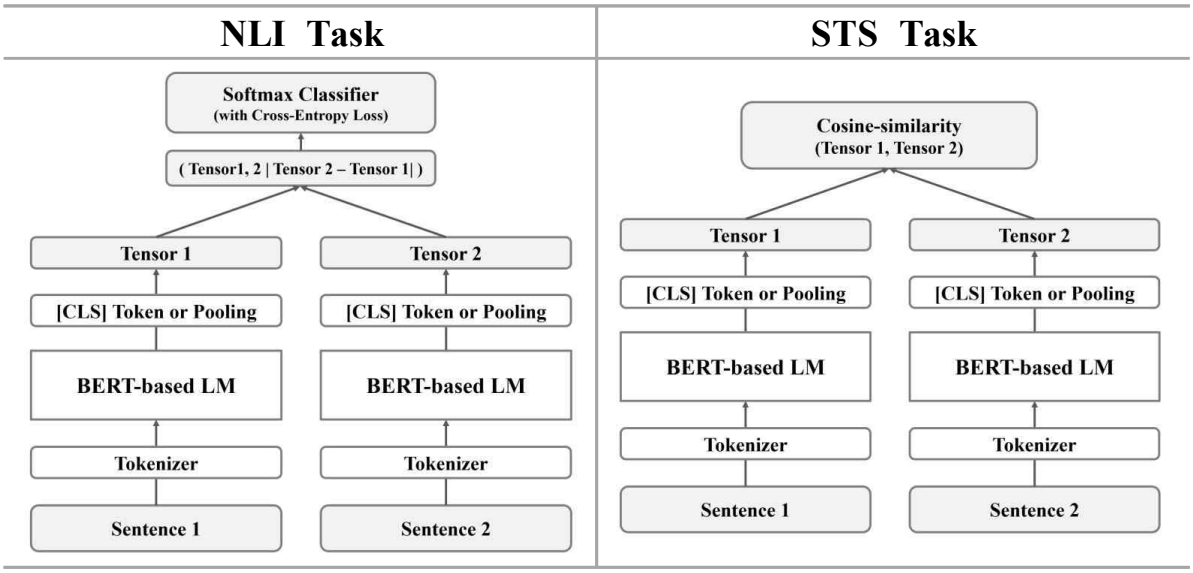


* **MLM 학습 과정**은 **토큰 수준의 세부 정보까지 디코더에 제공**하여 토큰 벡터를 출력하는 반면, **TSDAE**은 원본 텍스트를 예측하는데 **문장 수준의 벡터 정보만을 제공하고 문장 벡터 출력**

【STEP 2 : 언어모델 지도 학습】

- 높은 성능의 텍스트 클러스터링과 토픽 모델링을 구현하기 위해서는 **좋은 의미론적 임베딩(Semantic Embedding)을 생성**하는 언어모델이 중요
- 이에, 의미론적 임베딩을 위해 유사한 문장간 벡터 공간에서 가까워지도록 미세조정하는 NLI(Natural Language Inference)과 STS(Semantic Textual Similarity) **2가지 지도학습**을 추가로 수행
  - * Kor NLI : 94만개의 문장쌍으로 구성, 두 문장의 관계를 유사、모순、중립으로 라벨링된 데이터
  - * Kor STS : 5,750개의 문장쌍으로 구성, 두 문장의 관계를 0~5점 유사도로 라벨링된 데이터

《 NLI & STS Task 》



【STEP 3 : 병렬 클러스터링】

- 『**병렬 클러스터링(Parallel Clustering)**』은 **계산 속도**와 **안정적인 임베딩 그룹화에 초점**을 맞춰 우리가 디자인한 클러스터링 알고리즘

Parallel Clustering Algorithm

1. Randomly split up the entire text embedding into N group size. These serve as initial N cluster assignments for the observations.

2. Iteration until the cluster assignments stop changing.

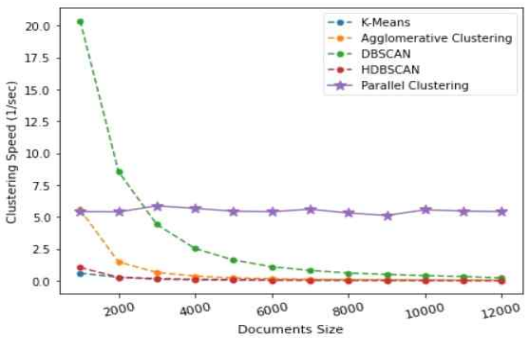
2-1. Parallelly for each of the N groups, compute the group centroid(or group head) and then filter embedding with low similarity to the centroid. Here the n-th cluster centroid is the embedding of the highest cosine similarity score in the n-th cluster.

2-2. Calculate the cosine similarity between group centroids, then merge groups with high similarity scores.

2-3. For all ungrouped embedding, perform a nearest-neighbor search with all centroids, then assign them to the nearest group if they are over the threshold.

3. Stack the clustered results in order of cluster size. As a result of the parallel clustering contextual embedding, news documents are grouped into semantically similar documents and rearranged by cluster size.

- 영화 텍스트 데이터(MovieLens / 영문)를 활용하여 **속도와 성능 비교 실험 결과, 他 유명 클러스터링 방법론보다 빠르고**, 데이터 크기가 커져도 **안정적인 텍스트 클러스터링** 가능



【STEP 4 : 병렬 클러스터링 기반 토픽 모델링】

- TF-IDF는 단어의 빈도(TF)와 역문서 빈도(IDF)를 사용하여 **문서 內 주요 단어들에 가중치를 부여**하여 키워드를 추출하는 통계적 방법

  * 전체 문서 內 빈도수가 높은 단어는 낮은 가중치, 특정 문서 內 빈도수가 높은 단어는 높은 가중치

- Class-based TF-IDF는 **클러스터링된 그룹(Class)을 하나의 주제로 간주**하고, 해당 그룹의 모든 문서를 결합 後 토픽별 TF-IDF 가중치를 부여하며 **뉴스 주제별 핵심 키워드 추출**

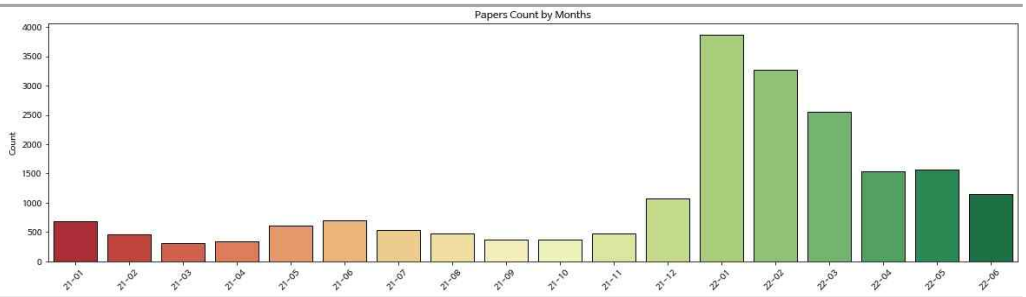
TF-IDF	C-TF-IDF
$W_{t,d} = tf_{t,d} \times \log(\frac{N}{1+df_t})$	$W_{t,C} = tf_{t,C} \times \log(1+\frac{A}{f_t})$
$tf_{t,d}$: Frequency of term ***t*** in document ***d***	$tf_{t,C}$: Frequency of term ***t*** in class ***C***
df_t : Number of documents containing ***t***	f_t : Frequency of term ***t*** across all classes
N : Total number of documents	A : Average number of words per class

□ 뉴스 토픽 모델링 시각화 결과

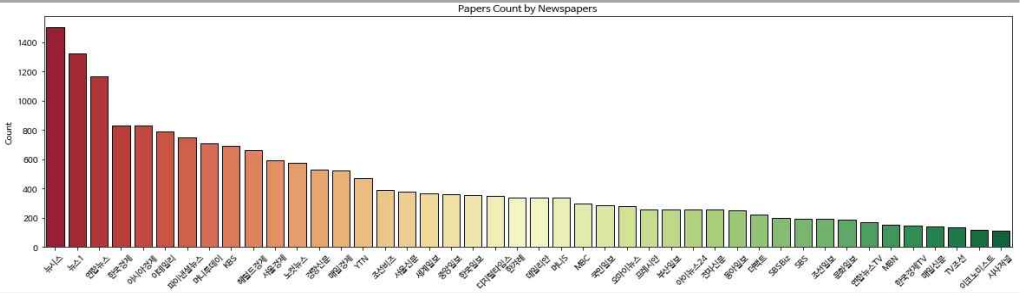
○ '**중대재해법**' 관련 뉴스 기사 데이터를 본 모델에 적용

- 시계열에 따라 기사수를 시각화한 결과, **'22.1.~3월간 대중과 언론의 관심이 높았으며, 상위 5개 언론매체**에서 전체의 50% 이상 기사를 제공

Count by Months



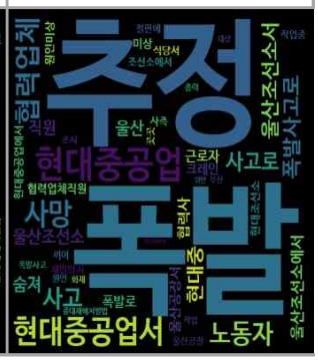
Count by Newspaper



• 우리가 구현한 토픽 모델링 모델로 '중대재해법' 관련 상위 Top 8개 토픽별 키워드를 추출한 결과, **주제별 핵심정보를 자동으로 제공**

* 클러스터링 크기순으로 자동 정렬되며, 사용자가 토픽 및 키워드 개수를 설정할 수 있게 디자인

《 My Topic Modeling Results 》

법 정보	노동자 사망 정보	두성산업 사고 정보	삼표산업 사고 정보
			
정부 대응 정보	**대검 수사 정보**	**판교 승강기 사고 정보**	**현대重 사고 정보**
			

□ Github

○ 깃허브를 통해 우리가 디자인한 『**병렬클러스터링 기반 뉴스 토픽 모델링**』 모델 소개 및 개발 코드 공개

* https://github.com/Navy10021/Parallel_Clustering_based_Topic_Modeling

□ 결　론 / 기대 효과

○ 국회의 입법 활동과 국정 철학에 기반한 정부의 정책 과제 수립 및 집행은 기본적인 책무이자 국민의 편안한 삶과 직결된다.

○ 특히, **입법과 정책은 빠르고, 정확한 여론예측 기술이 동반**돼야 적시에 국민의 삶에 기회요인을 제공하고, 위험요인에 유연하게 대처할 수 있으며, **체계적인 데이터 분석과 시뮬레이션은 주요 입법 · 정책 현안에 과학적 근거를 제공**한다.

○ 본 분석에서는 주요 입법 관련 데이터에 **자연어처리(NLP) 기술**을 접목하여 효율적인 **『국민 여론 분석모델』을 구현**하였다.

○ 세부적으로, 국회에서 제공한 트위터 · 댓글을 활용한 ❶ **『온라인 여론 예측모델』**과 뉴스 기사를 활용한 ❷ **『병렬클러스터 기반 뉴스 토픽 모델링』을 디자인**하였으며, 두 모델 모두 우수한 성능(여론예측 및 핵심쟁점 추출)을 보였다.

○ 향후, 입법 · 행정 · 언론 기관간 협업을 통해 사회 주요 이슈 관련 데이터를 종합하고, **다양한 프로젝트에 우리의 분석모델을 적용**한다면 **주요 법안과 정책 효과에 대한 불확실성 해소 및 국민 삶의 질 향상에 이바지**할 것이다.

※ 국내 데이터 과학자들의 연대와 노력으로 오는 2023년이 『AI 강국 대한민국』으로 전환점이 되길 희망합니다.