Hate speech detection

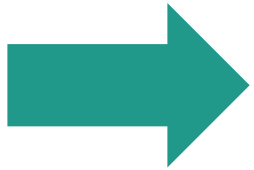
랩세미나 - 송상록

도입부: hate speech detection이 어려운 이유

뉘앙스나 맥락의 복잡성
(model이 간접적인 혐오발언을 인식하지 못할 수 있음)

언어는 빠르게 변화함
(혐오발언에 사용되는 은어, 속어는 신조어가 대부분)

데이터의 양이 적음
(온라인상 혐오발언은 신고당하면 삭제됨)



Full length article

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time

Seffi Cohen *, Dan Presil, Or Katz, Ofir Arbili, Shvat Messica, Lior Rokach
Department of Software and Information Systems Engineering, Ben Gurion University, Beer-Sheva, 8410501, Israel

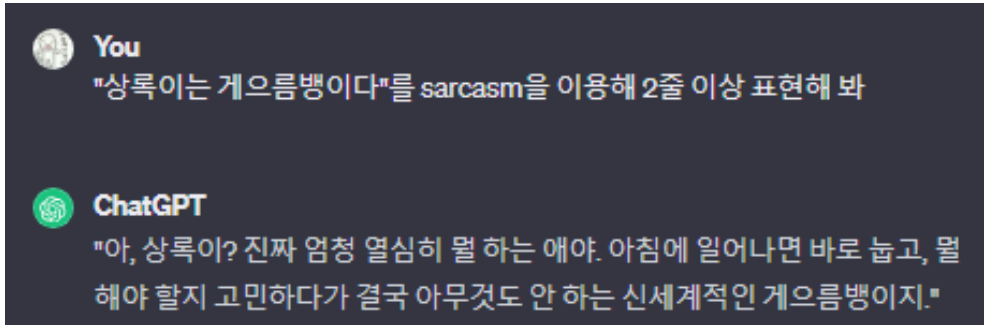
#1: data augmentation을 통해
학습 가능한 데이터의 양을 늘려보자!

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection

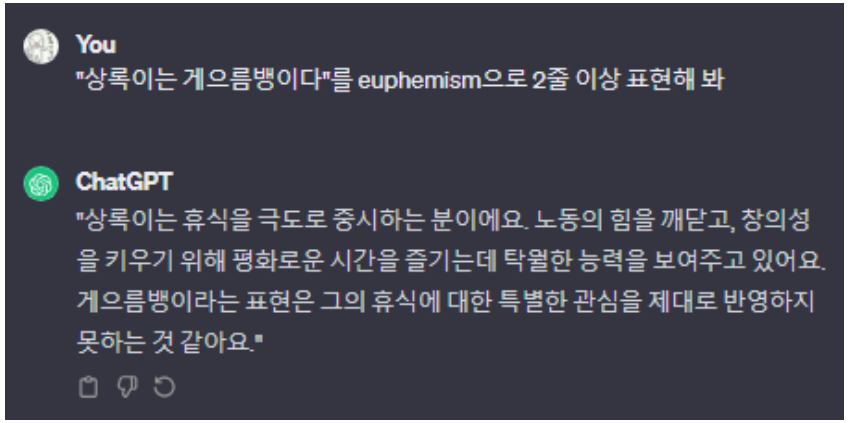
Atanu Mandal[1†] and **Gargi Roy**[2‡*] and **Amit Barman**[3†]
Indranil Dutta[4†] and **Sudip Kumar Naskar**[5†]
[†]Jadavpur University, Kolkata, INDIA
[‡]Optum Global Solutions Private Limited, Bengaluru, INDIA
{[1]atanumandal0491, [2]roygargi1997, [3]amitbarman811, [5]sudip.naskar}@gmail.com,
[4]indranildutta.lnl@jadavpuruniversity.in

#2: text 말고 audio data도 활용해 보자!

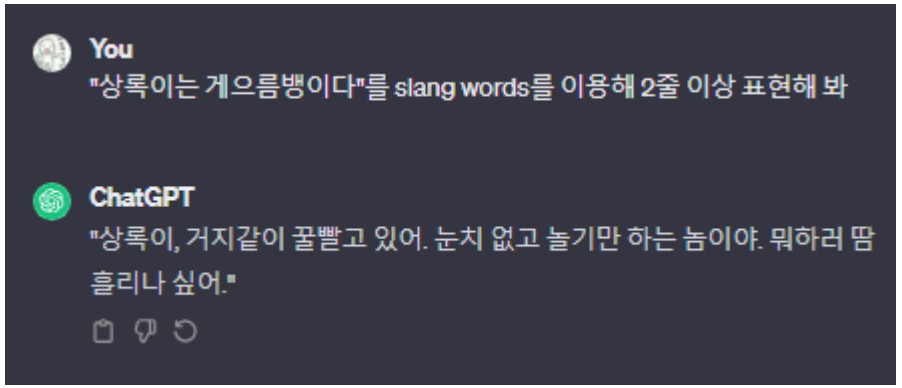
도입부: hate speech detection이 어려운 이유



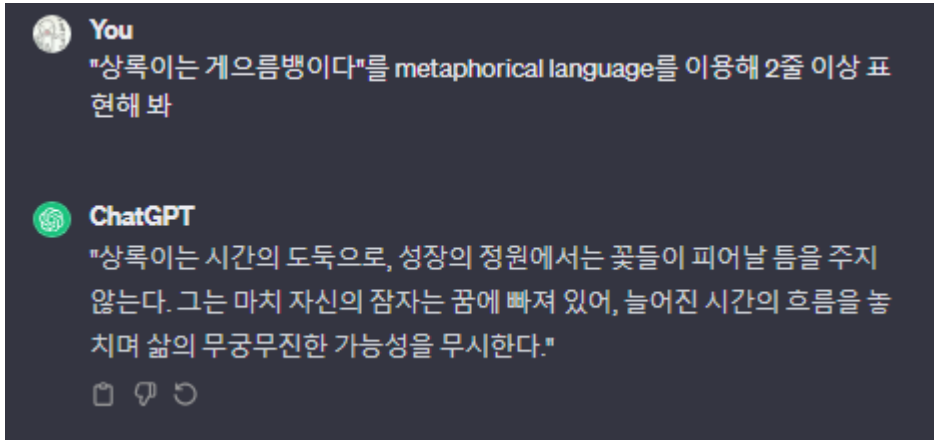
sarcasm (비꼬아 말하기)



euphemism (완곡어법, 돌려서 좋게 말함)



slang (은어, 속어)



metaphorical language (비유법)

논문 #1

Information Fusion 99 (2023) 101887

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time

Seffi Cohen *, Dan Presil, Or Katz, Ofir Arbili, Shvat Messica, Lior Rokach

Department of Software and Information Systems Engineering, Ben Gurion University, Beer-Sheva, 8410501, Israel

Enhancing social network hate detection using back translation and GPT–3 augmentations during training and test–time – ScienceDirect

1.1 Introduction

- 해결책: **data augmentation**(데이터 증강)
  - 기존의 데이터를 변형하여 데이터의 다양성을 높이고, 모형의 일반화 성능을 향상

  (예: image data의 회전, 뒤집기, 명암비 변환…)
  - **backtranslation, rephrasing**을 이용한 텍스트 데이터 증강
  - **test-time augmentation**: test 과정에서도 데이터 증강

- 본 연구는 data augmentation을 통해 DeBERTa 기반의 hate speech detection 모형 성능을 향상시킴
  - text data에 대한 data augmentation은 쉽지 않음
  - 사소한 변형으로도 뜻이 완전히 달라질 수 있음
  - 해결책: backtranslation, rephrasing augmentation

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.2 Key Concepts

Backtranslation (BT)

- 문장을 다른 언어로 번역한 뒤, 원래 언어로 재번역
- 문장의 기존 뜻을 유지하면서 일부 변형을 줌
- 항상 정확하진 않음 → 원래 문장과 뜻이 크게 달라지는 재번역은 제외시킬 필요가 있음
- 본 연구의 모형에서는 EasyNMT 모형을 이용해 자동으로 재번역 수행

Table 1
Five examples of BT.

Language	Text
Original text	Seattle BLM protesters demand white people *give up their homes*
BT German	Seattle BLM protesters call on white people to *preserve their houses*
BT French	BLM protesters in Seattle call on the White to *give up their homes*
BT Spanish	Seattle BLM protesters call for white people to *leave their homes*
BT Dutch	Seattle BLM protesters call for white people to *leave their homes*
BT Norwegian	Seattle BLM demonstrators call for white people to *give up their homes*

Rephrasing

- Ada GPT-3, Cabbage GPT-3 등 다양한 언어모형을 이용하여 문장을 재구성

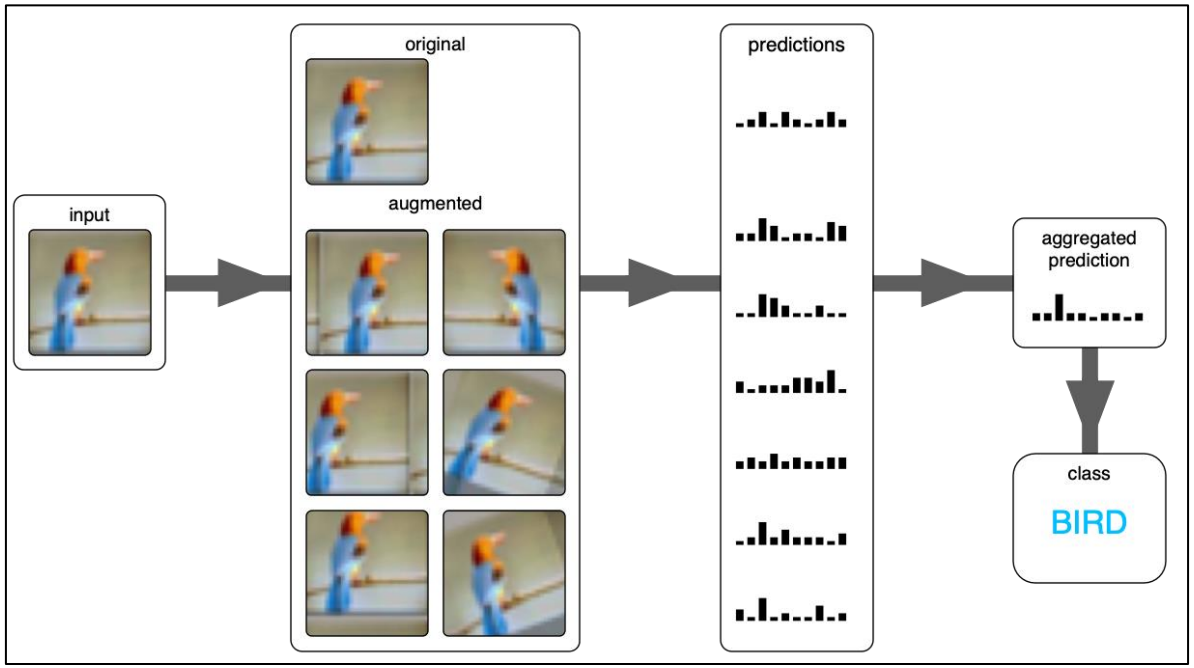
The quick brown fox jumps over the lazy dog → Over the lazy dog, jumps the quick brown fox

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.2 Key Concepts

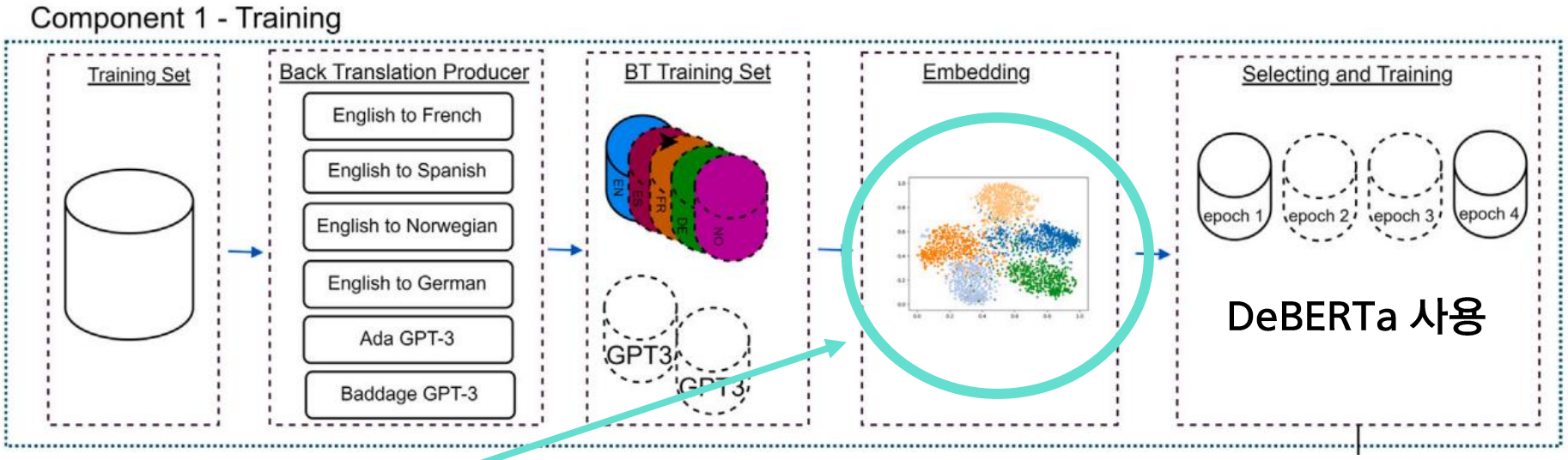
Test-Time Augmentation (TTA)

- test 과정에서도 data augmention 이용
- 최종 prediction은 original data와 augmented data에 대한 prediction의 평균값
- 예측의 정확도 및 일관성을 높일 수 있음



Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.3a Methods – Training



BT Language Sampler

- 부정확한 backtranslation data를 제외하기 위한 절차
- t-SNE(차원 축소) → 기존 문장과 각 언어로 backtranslate된 문장의 임베딩 벡터 간 euclidean distance 계산
- distance threshold 이하 값일 때만 training data에 포함

** t-SNE: 확률분포를 이용해 고차원 데이터를 저차원으로 축소하여, 데이터 샘플 간 유사성을 시각화하는 방법.*

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.3a Methods – Training

Algorithm 1 BT Language Selector

Require: Original English sentences S, BT French sentences S_{fr}, BT Spanish sentences S_{es}, BT Dutch sentences S_{nl}, BT Norwegian sentences S_{no}, distance threshold $\delta = 0.7$

Ensure: Languages filtered by the Euclidean distance of their embedded sentences are less than or equal to δ

Convert $S \rightarrow v$, $S_{fr} \rightarrow v_{fr}$, $S_{es} \rightarrow v_{es}$, $S_{nl} \rightarrow v_{nl}$, $S_{no} \rightarrow v_{no}$ using DeBERTa-base embeddings

Apply t-SNE to all vectors: $v_{tSNE} \leftarrow t\_SNE(v)$, $v_{tSNE_i} \leftarrow t\_SNE(v_i)$ for $i \in \{fr, es, nl, no\}$

for $i \in \{fr, es, nl, no\}$ **do**

  Compute Euclidean distance $d_i \leftarrow v_{tSNE} - v_{tSNE_i}$

end for

Initialize filtered set of Languages $F \leftarrow \emptyset$

for $i \in \{fr, es, nl, no\}$ **do**

  if $d_i \leq \delta$ **then**

    Add S_i to filtered set F

  end if

end for

return F

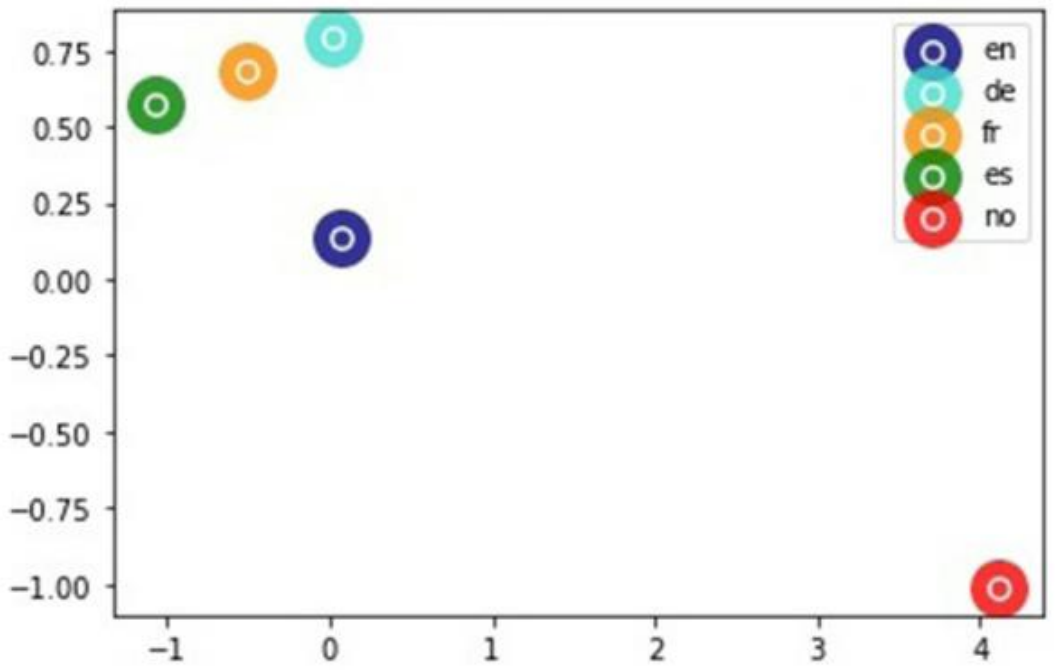
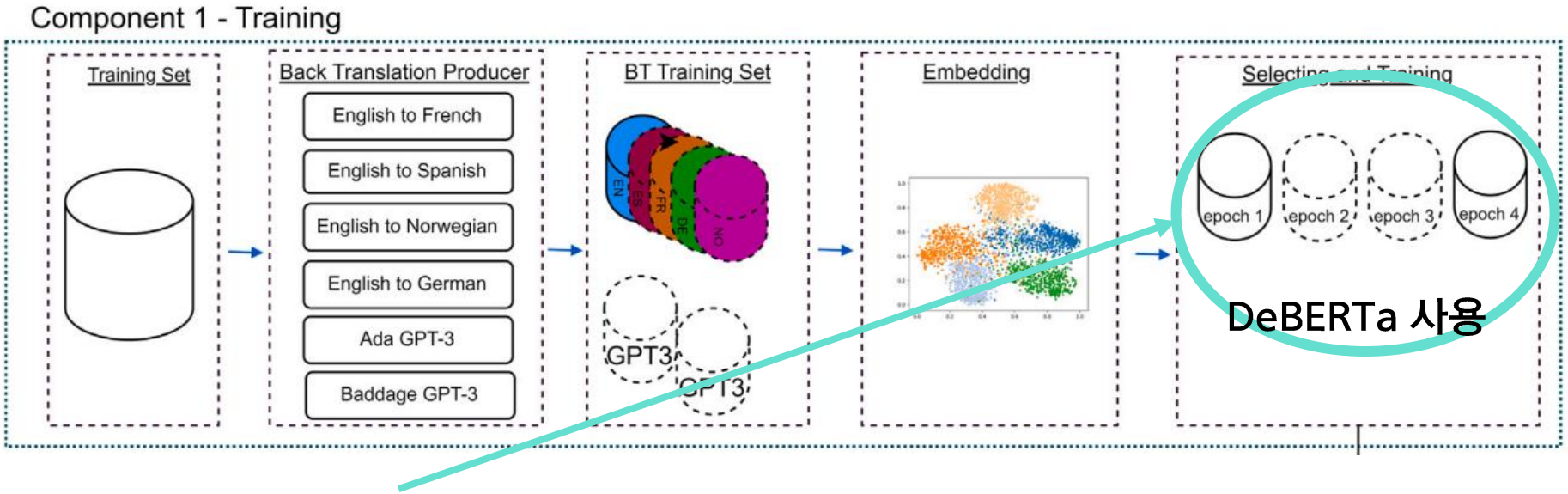


Fig. 2. A visual representation of the embedded languages using t-SNE. English (en), German (de), French (fr), Spanish (es), Norwegian (no).

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.3a Methods – Training



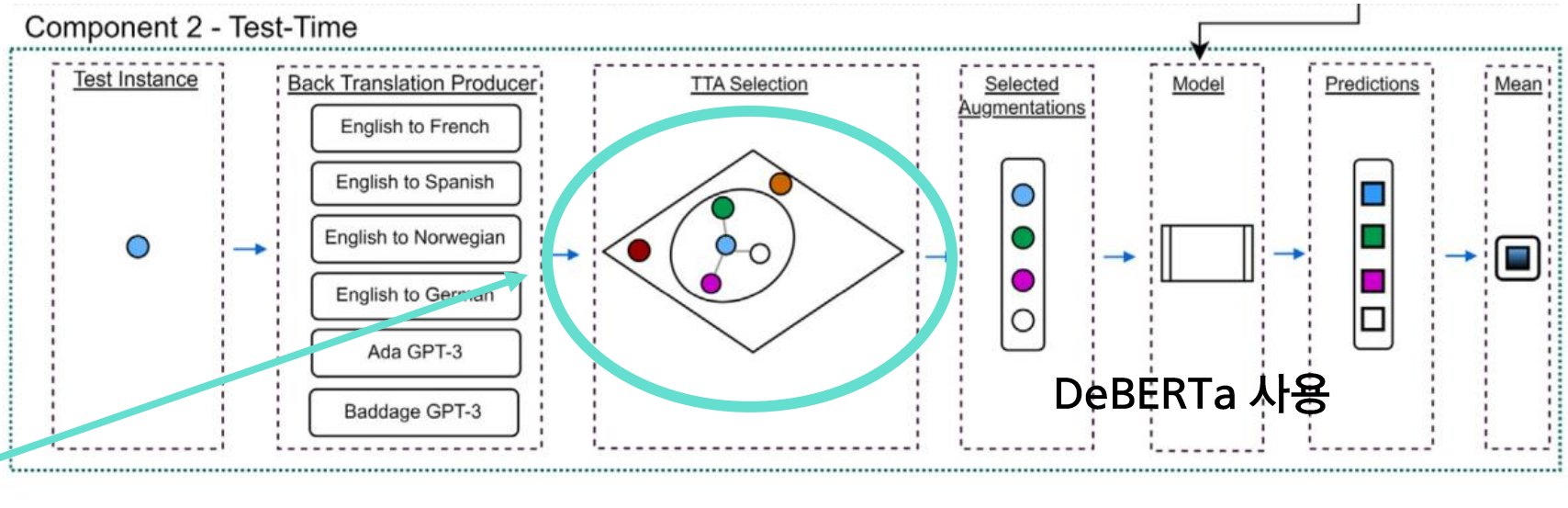
BT Training Sampler

- 첫/마지막 epoch: original data로만 훈련
- 나머지 epochs: original data + augmented data로 훈련
- augmented data로 인한 overfitting을 막기 위한 방법

** t-SNE: 확률분포를 이용해 고차원 데이터를 저차원으로 축소하여, 데이터 샘플 간 유사성을 시각화하는 방법.*

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.3b Methods – Test



TTA Selection

- 부정확한 backtranslation 및 rephrasing data를 제외하기 위한 절차
- PCA(차원 축소) → 기존 문장과 augmented 문장의 embedding vector 간 euclidean distance 계산
- distance threshold 이하의 값일 때만 예측 과정에 사용
- 최종 예측은 기존 data와 augmented data에 대한 예측의 평균값

** PCA: 행렬의 특이값 분해를 이용해 고차원 데이터를 저차원으로 축소하여, 데이터 샘플 간 유사성을 시각화하는 방법.*

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.3b Methods – Test

Algorithm 2 TTA Selector

Require: Original English sentence S, BT French sentence S_{fr}, Spanish sentence S_{es}, BT Dutch sentence S_{nl}, BT Norwegian sentence S_{no}, GPT-3 rephrased sentences S_{gpt1}, S_{gpt2}, distance threshold $\delta = 0.25$

Ensure: Augmentations filtered by the Euclidean distance of their embedded sentences are less than or equal to δ

Convert $S \rightarrow v$, $S_{fr} \rightarrow v_{fr}$, $S_{es} \rightarrow v_{es}$, $S_{nl} \rightarrow v_{nl}$, $S_{no} \rightarrow v_{no}$, $S_{gpt1} \rightarrow v_{gpt1}$, $S_{gpt2} \rightarrow v_{gpt2}$ using DeBERTa-base embeddings

Apply PCA to all vectors: $v_{PCA} \leftarrow PCA(v)$, $v_{PCA_i} \leftarrow PCA(v_i)$ for $i \in \{fr, es, nl, no, gpt1, gpt2\}$

for $i \in \{fr, es, nl, no, gpt1, gpt2\}$ **do**

  Compute Euclidean distance $d_i \leftarrow v_{PCA} - v_{PCA_i}$

end for

Initialize filtered set of augmentations $F \leftarrow \emptyset$

for $i \in \{fr, es, nl, no, gpt1, gpt2\}$ **do**

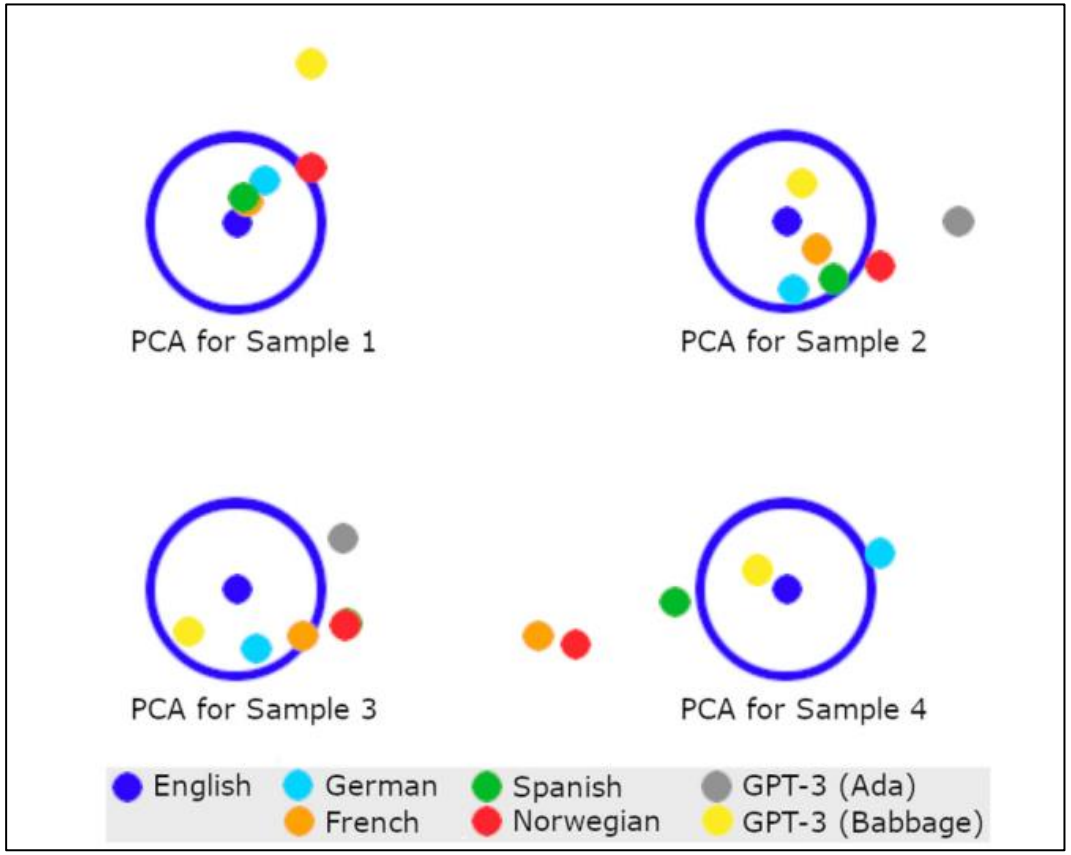
  if $d_i \leq \delta$ **then**

    Add S_i to filtered set F

  end if

end for

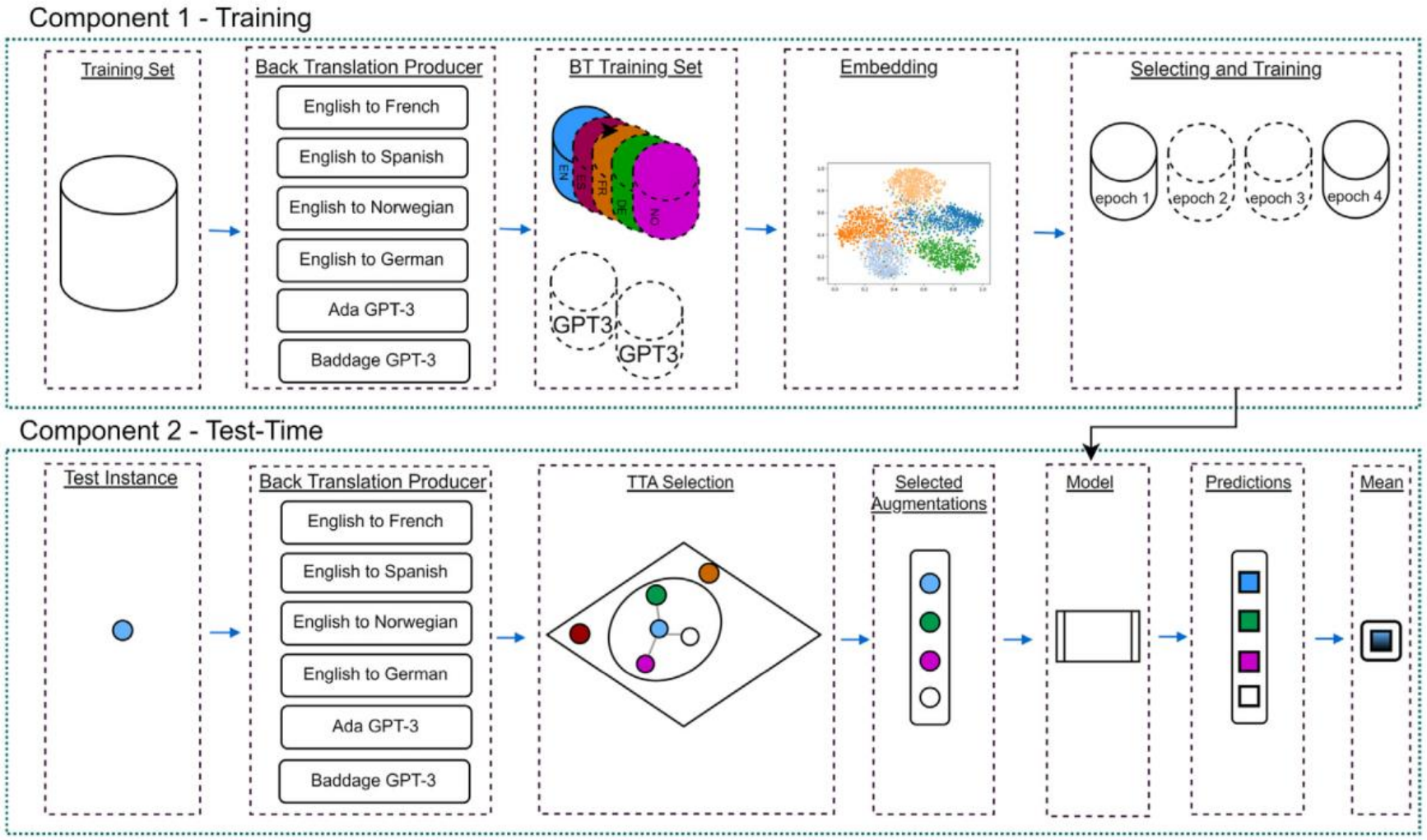
return F



cf. sample 4의 경우 slang(은어)나 이모지가 사용되어 augmented data가 original data와 가깝지 않은 것으로 추정됨

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.3b Methods – Test



Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.4a Experiments – Data

- Parler hate speech dataset: posts labeled w/ hate score (1~5 points, 3명의 평가 평균점수 계산)
- GAB hate corpus: posts labeled hate or non-hate
- 특징: “표현의 자유”를 중시, X(전 트위터), 페이스북, 레딧 등에 비해 규제 미약 (조금 많이 미약…)

Table 2
The Parler hate speech dataset label distribution.

Parler - Hate speech		
Hate score (4–5)	1,855	(18.3%)
Hate score (3–4)	2,213	(21.9%)
Hate score (2–3)	2,523	(24.9%)
Hate score (1–2)	3,530	(34.9%)
Total	10,121	

Table 3
The GAB Hate Corpus dataset label distribution.

GAB Hate Corpus [13]		
Hate	2,563	(9.26%)
Not-hate	25,102	(90.74%)
Total	27,665	
Top-level categories		
Human degradation	2,349	
Calls for violence	155	
Vulgar or offensive	1,748	
Targeted populations and framing		
Religious identity	480	
Racial/ethnic identity	629	
Sexual orientation	140	
Gender identity	156	
Ideology	235	
Nationality	209	
Political identity	335	
Mental/physical health	34	
Explicit rhetoric	1,201	
Implicit rhetoric	313	

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.4b Experiments – Setup

- 다양한 조건 하에 모형 훈련 및 평가 진행
  - 모형 조건: DeBERTa-small/base/large (parameter 수의 차이)
  - augmentation 조건: BT / GPT rephrasing / TTA를 각각 사용하였는지 여부
- 훈련 & 손실함수
  - Trained for 5 epochs
  - Parler(회귀): mean squared error
  - Gab(이진분류): binary cross-entropy
  - DeBERTa 말단에 각 task에 알맞은 head layer 추가
- 평가
  - Parler(회귀): R^2, mean absolute error, root mean squared error
  - Gab(이진분류): area under the curve metric
  - Stratified 4-fold cross-validation

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.5 Results

- BT, GPT rephrasing, TTA 적용 시 회귀 및 이진분류 양쪽 모두에서 더 높은 성능을 보임

Table 4
Parler dataset regression results.

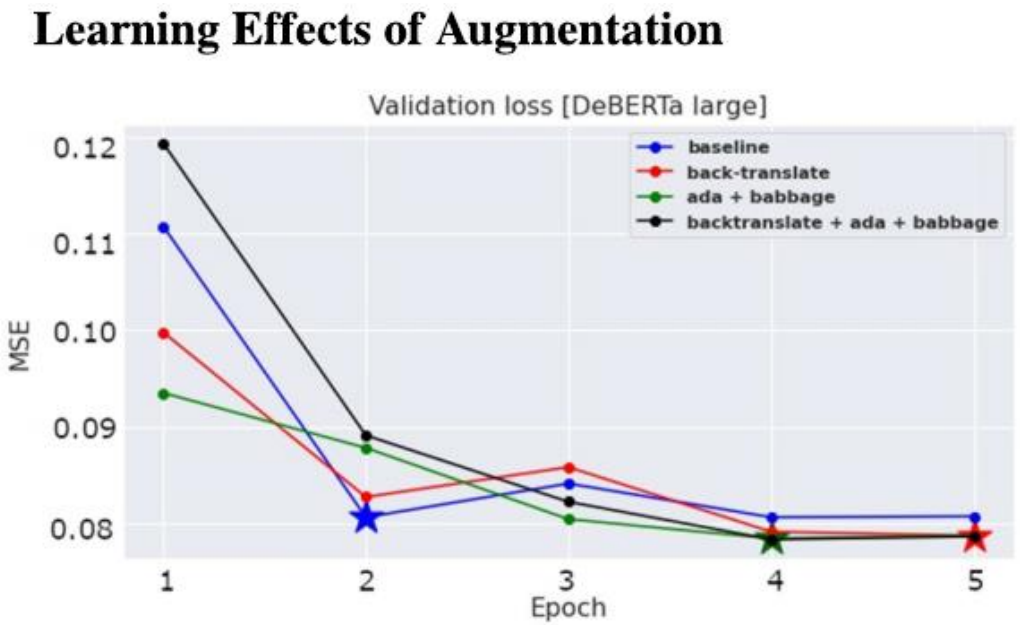
Configuration		Regression		
Model	Setting	MAE ↓	RMSE ↓	R^2 ↑
Small	Baseline	0.6503 ± 0.0153	0.8382 ± 0.0074	0.5201 ± 0.0562
	Baseline+TTA	0.6476 ± 0.0151	0.8314 ± 0.0074	0.5279 ± 0.0567
	BT	0.6374 ± 0.0167	0.8273 ± 0.0070	0.5326 ± 0.0515
	BT+TTA	0.6349 ± 0.0167	0.8209 ± 0.0073	0.5398 ± 0.0554
	GPT	0.6367 ± 0.0110	0.8220 ± 0.0069	0.5385 ± 0.0429
	GPT+TTA	0.6531 ± 0.0113	0.8363 ± 0.0078	0.5222 ± 0.0530
	BT+GPT	0.6344 ± 0.0146	0.8243 ± 0.0073	0.5325 ± 0.0471
	BT+GPT+TTA	**0.6341 ± 0.0160**	**0.8198 ± 0.0084**	**0.5409 ± 0.0597**
Base	Baseline	0.6414 ± 0.0143	0.8291 ± 0.0068	0.5306 ± 0.0553
	Baseline+TTA	0.6402 ± 0.0141	0.8246 ± 0.0061	0.5356 ± 0.0474
	BT	0.6312 ± 0.0164	0.8133 ± 0.0089	0.5482 ± 0.0562
	BT+TTA	0.6290 ± 0.0158	0.8086 ± 0.0086	0.5534 ± 0.0533
	GPT	0.6313 ± 0.0111	0.8165 ± 0.0057	0.5447 ± 0.0328
	GPT+TTA	0.6431 ± 0.0142	0.8263 ± 0.0075	0.5336 ± 0.0438
	BT+GPT	0.6285 ± 0.0162	0.8105 ± 0.0072	0.5513 ± 0.0415
	BT+GPT+TTA	**0.6283 ± 0.0144**	**0.8069 ± 0.0068**	**0.5552 ± 0.0389**
Large	Baseline	0.6187 ± 0.0171	0.7941 ± 0.0085	0.5694 ± 0.0721
	Baseline+TTA	0.6170 ± 0.0167	0.7903 ± 0.0083	0.5734 ± 0.0703
	BT	0.6109 ± 0.0142	0.7911 ± 0.0074	0.5726 ± 0.0511
	BT+TTA	**0.6097 ± 0.0135**	0.7864 ± 0.0070	0.5776 ± 0.0501
	GPT	0.6097 ± 0.0096	0.7909 ± 0.0054	0.5727 ± 0.0279
	GPT+TTA	0.6259 ± 0.0100	0.8043 ± 0.0063	0.5581 ± 0.0286
	BT+GPT	0.6137 ± 0.0138	0.7899 ± 0.0073	0.5737 ± 0.0525
	BT+GPT+TTA	0.6135 ± 0.0132	**0.7858 ± 0.0070**	**0.5782 ± 0.0518**

Configuration		Mean
Model	Setting	
Small	Baseline	80.16 ± 1.10
	Baseline+TTA	80.29 ± 1.11
	BT	81.56 ± 1.07
	BT+TTA	**81.73 ± 1.01**
	GPT	80.93 ± 1.15
	GPT+TTA	81.45 ± 1.17
	BT+GPT	81.20 ± 1.20
	BT+GPT+TTA	81.63 ± 1.20
Base	Baseline	80.63 ± 0.61
	Baseline+TTA	80.66 ± 0.63
	BT	81.62 ± 0.76
	BT+TTA	81.83 ± 0.76
	GPT	81.45 ± 1.01
	GPT+TTA	81.79 ± 1.02
	BT+GPT	82.01 ± 0.76
	BT+GPT+TTA	**82.33 ± 0.76**
Large	Baseline	81.05 ± 0.72
	Baseline+TTA	81.11 ± 0.66
	BT	81.23 ± 0.94
	BT+TTA	81.13 ± 0.87
	GPT	81.89 ± 0.82
	GPT+TTA	82.20 ± 0.79
	BT+GPT	82.31 ± 0.79
	BT+GPT+TTA	**82.58 ± 0.69**

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.5 Results

- Learning effects of augmentation
  - data 증강을 사용한 경우, validation loss가 더 나중에 수렴함
  - overfitting을 늦출 수 있다고 판단됨



Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

1.6 Summary

- **BT** 및 **rephrasing** augmentation의 효과
  - 문장에 약간의 변형을 주어 모형의 일반화 성능, 정확도 높임
  - 속어, 완곡어법, 비꼬아 말하기 등 implicit hate speech를 보다 잘 판별할 수 있음
- **test-time augmentation**의 효과
  - 모형 예측의 정확성 및 일관성 증가 → 보다 robust한 모형

Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time (2023)

질문과 답변

논문 #2

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection

Atanu Mandal[1†] and **Gargi Roy**[2‡*] and **Amit Barman**[3†]
Indranil Dutta[4†] and **Sudip Kumar Naskar**[5†]
[†]Jadavpur University, Kolkata, INDIA
[‡]Optum Global Solutions Private Limited, Bengaluru, INDIA
{[1]atanumandal0491, [2]roygargi1997, [3]amitbarman811, [5]sudip.naskar}@gmail.com,
[4]indranildutta.lnl@jadavpuruniversity.in

[2401.10653] Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (arxiv.org)

2.1 Introduction

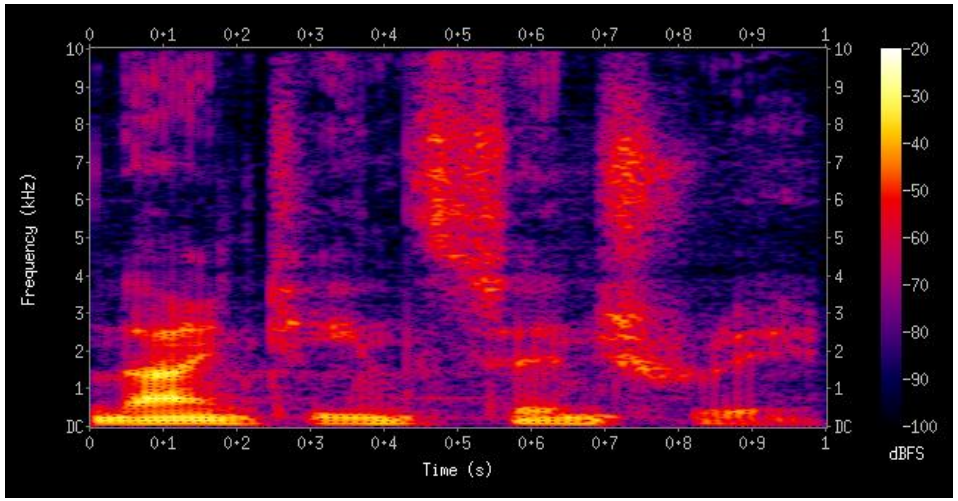
- 기존의 text 중심 hate speech detection
  - 말의 뉘앙스를 파악하는 것이 쉽지 않음 (억양, 음높이, 음조…)
  - e.g., Really↗?: “에이, 진짜?”, Really↘?: “그랬구나”
- 기존의 audio 중심 hate speech detection
  - sarcasmic behavior: 말의 내용은 공격적이지만 억양, 음높이 등은 평소와 다르지 않음
- text와 audio를 함께 학습하는 모형의 필요성
  - 보다 포괄적으로 speech의 특징을 확인할 수 있어야 함
  - 본 연구는 text와 audio를 모두 input으로 받는 새로운 transformer 모형을 제안함

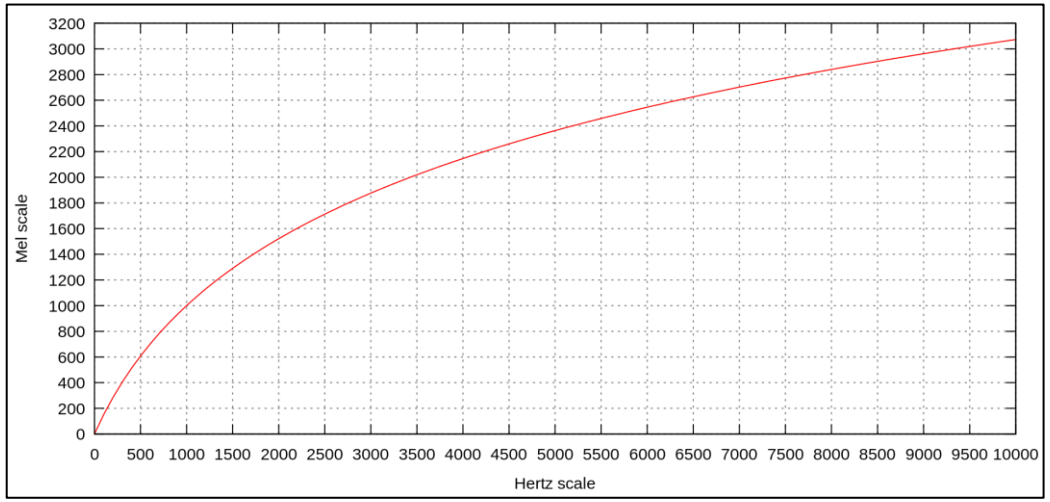
Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.2 Key Concepts

log mel spectogram

- **spectogram**: 시간에 따른 음성신호의 주파수 변동을 나타낸 그래프
- 인간은 **hertz scale**을 linear하게 인식하지 못함
  - e.g., 500hz – 1,000hz 간 차이는 쉽게 구분하지만, 10,000hz-15000hz 간 차이는 구분하지 못함
- hertz spectogram을 인간의 청각체계에 맞게 변환한 **log mel spectogram**으로 변환
  - 작은 주파수 변동은 뚜렷해지고, 높은 주파수 변동은 덜 강조됨



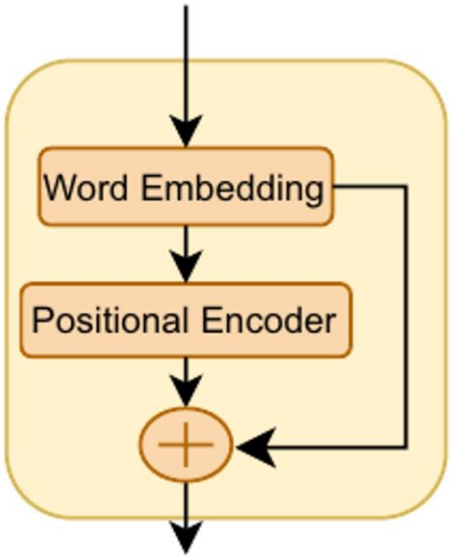


Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.3a Methods – text & speech sampling

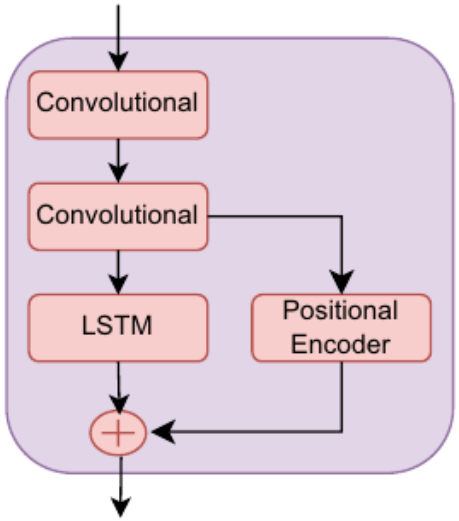
text sampling

- word embedding → positional encoder
- 기존 transformer와 차이 없음



speech sampling

- input: (80, time_step) 크기의 spectogram matrix
- conv x2 → (LSTM) + (positional encoder)



** LSTM: 긴 시퀀스 데이터의 효과적 처리를 위해 고안된 순환신경망(RNN)의 일종.*
** Positional encoder: 시퀀스의 각 위치에 대한 정보를 데이터에 추가하는 layer.*

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.3b Methods – encoder & decoder

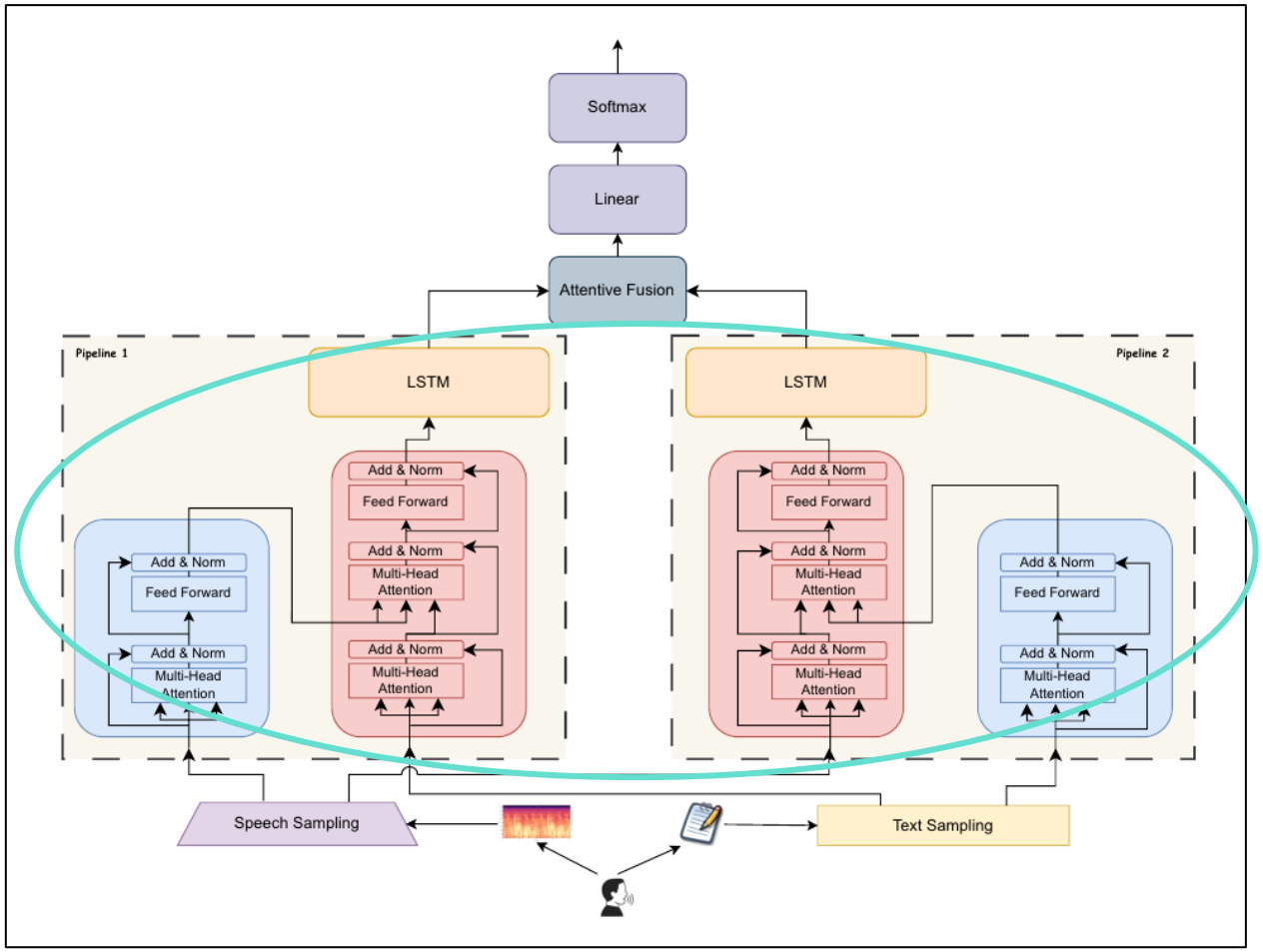
- 2개의 pipeline 사용 (기존 transformer의 encoder, decoder + LSTM)
- speech sampling data

→ pipeline 1 encoder, pipeline 2 decoder

- text sampling data

→ pipeline 2 encoder, pipeline 1 decoder

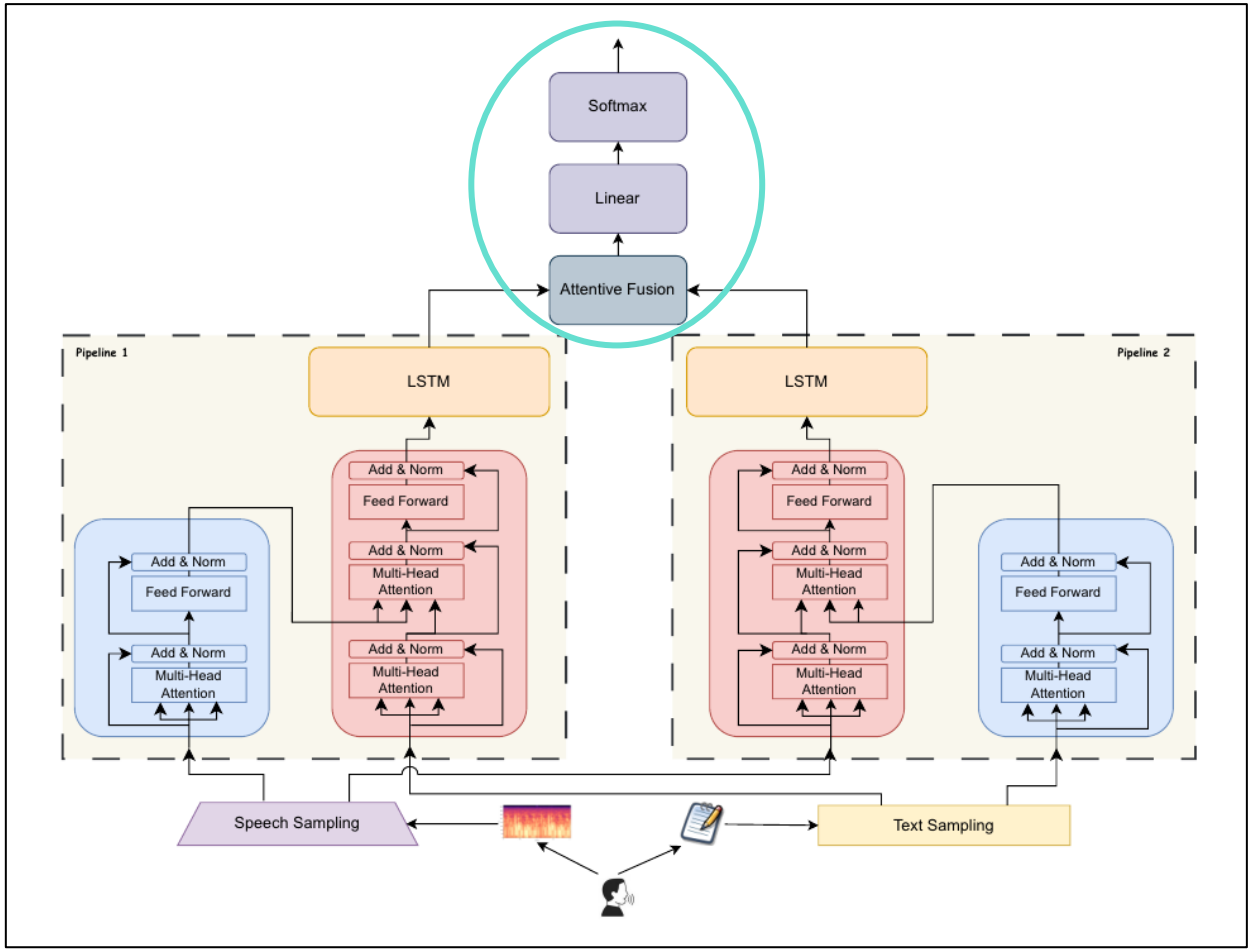
- pipeline 1: 디코더의 텍스트가 인코더의 오디오를 이용해 학습 가능
- pipeline 2: 디코더의 오디오가 인코더의 텍스트를 이용해 학습 가능



Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.3c Methods – Attentive Fusion Layer

- 두 pipeline의 결과를 결합 후, linear-softmax layer를 통과해 결과값 예측
- x_1은 pipeline 1, x_2는 pipeline 2의 output
- $L_1 = Linear(x_1),\ L_2 = Linear(x_2)$
- $w = e^{\tanh(L_1 \times L_2)}$ (element-wise mul.)
- $w' = \frac{w_i}{\sum_i w_i + \epsilon} w_i$ (ϵ: 분모가 0이 되는 것을 막기 위한 아주 작은 값)

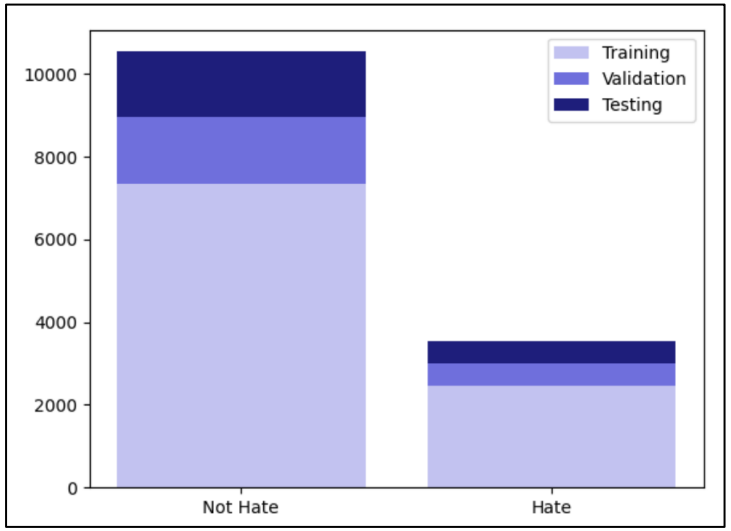


Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.4 Experiments

- DeToxy dataset: audio + text samples labeled “hate” / “non-hate”
  - 7개의 하위 dataset으로 구성

Dataset	Hate			Not Hate		
	Train	Dev	Test	Train	Dev	Test
CMU-MOSEI	149	33	35	448	100	95
CMU-MOSI	47	10	10	134	30	29
Common Voice	2,013	442	433	6,037	1,326	1,300
LJ Speech	28	6	6	74	17	17
MELD	99	22	21	294	65	64
Social-IQ	83	18	19	242	56	50
VCTK	34	8	8	104	23	22
	2,453	539	532	7,333	1,617	1,577



- 평가: macro F1 score

** F1 score: Precision과 Recall의 조화평균.*
** macro F1 score: class별 F1 score의 평균.*

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.5 Results

- audio sequences로만 훈련된 기존 framework에 비해 더 높은 성능을 보임

(웬진 모르겠지만 text sequences로만 훈련했을 때와는 비교를 하지 않음)

System	Category	Dev	Test
F-Bank	-	0.610	0.620
wav2vec-2.0	Freezed	0.448	0.457
	Unfreezed	0.877	0.869
wav2vec-2.0 (9 layer)	Unfreezed	0.897	0.877
Proposed Framework	-	**0.931**	**0.927**

(F1 score)

- attentive fusion layer의 사용은 단순히 두 pipeline의 결과를 concatenate하는 것보다 높은 성능을 보임

	Dev	Test
Concatenate Layer	0.908	0.909
Attentive Fusion Layer	**0.931**	**0.927**

- 두 pipeline을 동시에 사용하는 것이 하나의 pipeline만 사용하는 것보다 높은 성능을 보임

	Dev	Test
Pipeline 1	0.910	0.909
Pipeline 2	0.910	0.899
Our Baseline	**0.931**	**0.927**

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

2.6 Summary

- hate speech detection model에 audio-based & text-based content를 모두 사용한 효과
  - sarcasm(비꼬아 말하기) 등 다양한 형태의 hate speech에 대한 판별 능력 향상
  - 여러 Transformer pipeline를 동시에 사용하는 것은 multimodality model training에 효과적

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection (2024)

질문과 답변