Hate Speech Detection based on Sentiment Knowledge Sharing

Xianbing Zhou, Young Yang, Xiaochao Fan,
Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, Hongfei Lin

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing

성신여자대학교 미래융합기술공학과 이세영 발표

Contents

A. 자연어 처리(Natural Language Processing)

B. 텍스트 분류(Text Classification)

A. 자연어처리, NLP(Natural Language Processing)

NLP란?

Natural Language Processing (자연어처리)
: 텍스트에서 의미 있는 정보를 분석, 추출하고 이해하는 일련의 기술집합

NLP 응용사례 예시

- **텍스트 분류 (Text Classification)**
- 기계 번역 (Machine Translation)
- 텍스트 요약(Summarization)
- 자동 질의응답 (Question Answering, QA)
- etc

출처: Konlpy document (https://konlpy-ko.readthedocs.io/ko/v0.4.3/start/)

B. 텍스트 분류 (Text Classification)

① 데이터셋 준비

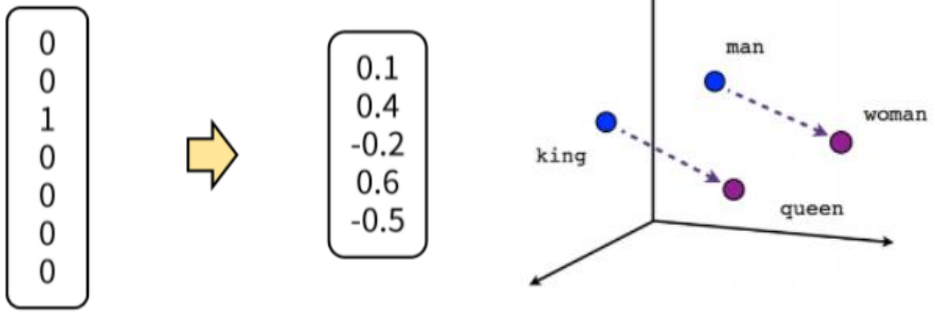
```
$ head ratings_train.txt
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나        1
10265843        너무재밋었다그래서보는것을추천한다      0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정       0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.     0
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

https://github.com/e9t/nsmc

② 데이터 전처리(토크나이징, 특수문자 제거 등, 레이블링 등

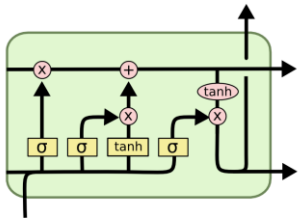
```
안녕하세요. 반갑습니다.

['_안', '녕', '하세요', '.', '_반', '갑', '습니다', '.']

[10, 4468, 357, 0, 227, 3968, 251, 0]
```

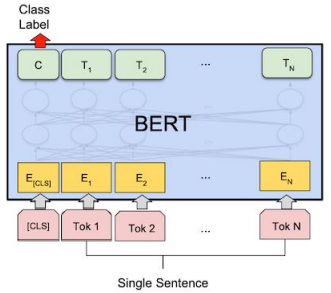
③ 임베딩



Male-Female

④ 모델 학습





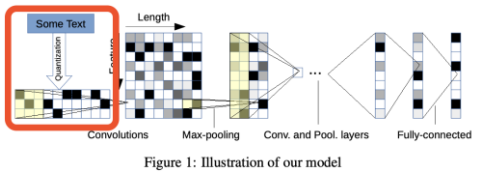


Figure 1: Illustration of our model

LSTM BERT char-CNN

⑤ 성능평가

I. Introduction

연구 배경

- 인터넷, SNS, 모바일의 발달로 혐오표현 증가 및 이로 인한 문제 심화

  => 이를 막기 위한 NLP 기술 필요성이 대두됨.
- 대부분의 혐오 발언은 부정적 감정을 포함하고 있음.
- MoE(Mixture of Expert)에 영감을 받아 sentimental analysis 와 hate speech detection 모델에서 지식을 공유하는 연구 진행함.

Contributions

1) 감정지식을 활용한 Multi-task Learning(MTL)
2) 공유작업을 더 잘하기 위해 multi-head attention mechanism과 Gated Attention을 사용하는 새로운 프레임워크 제안
3) 공개된 2개의 데이터셋에서 다른 baseline 모델들에 비해 최첨단 성능 달성 입증

II. Related Work

Feature engineering 기반 Machine learning

- Zeerak Waseem. 2016. Are you a racist or am i seeing things?
  - n-gram feature, sentimental feature 가 hate speech 탐지에 효과적이라는 것 입증
- Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. lexicon-based approach for hate speech detection.
  - 몇 가지 sentimental feature 구성, 실험을 통해 좋은 성능 입증

=〉 이전의 연구는 감정 특징이 혐오 발언 탐지에 중요한 역할을 한다는 것을 보여줌.

Multi-task Learning 기반 연구

- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis.
  - offensive language detection 을 위한 BERT 기반 MTL 모형 제안
- Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection.
  - 혐오 발언 탐지 성능을 개선하기 위해 여러 관련 분류 작업에서 유용한 정보를 활용하는 심층 다중 작업 학습(MTL) 프레임워크를 제안

=〉 multi-task learning 모델에서 감정 분석 작업과 혐오 발언 감지 작업 사이의 상관관계를 사용함으로써

혐오 발언 감지 모델의 성능과 일반화 능력을 향상시킬 수 있다는 것을 보여줌

III. Methodology

Main Idea

- 문장에서 감정 지식(sentiment knowledge)을 고려해서 Hate speech detection 모델 개선

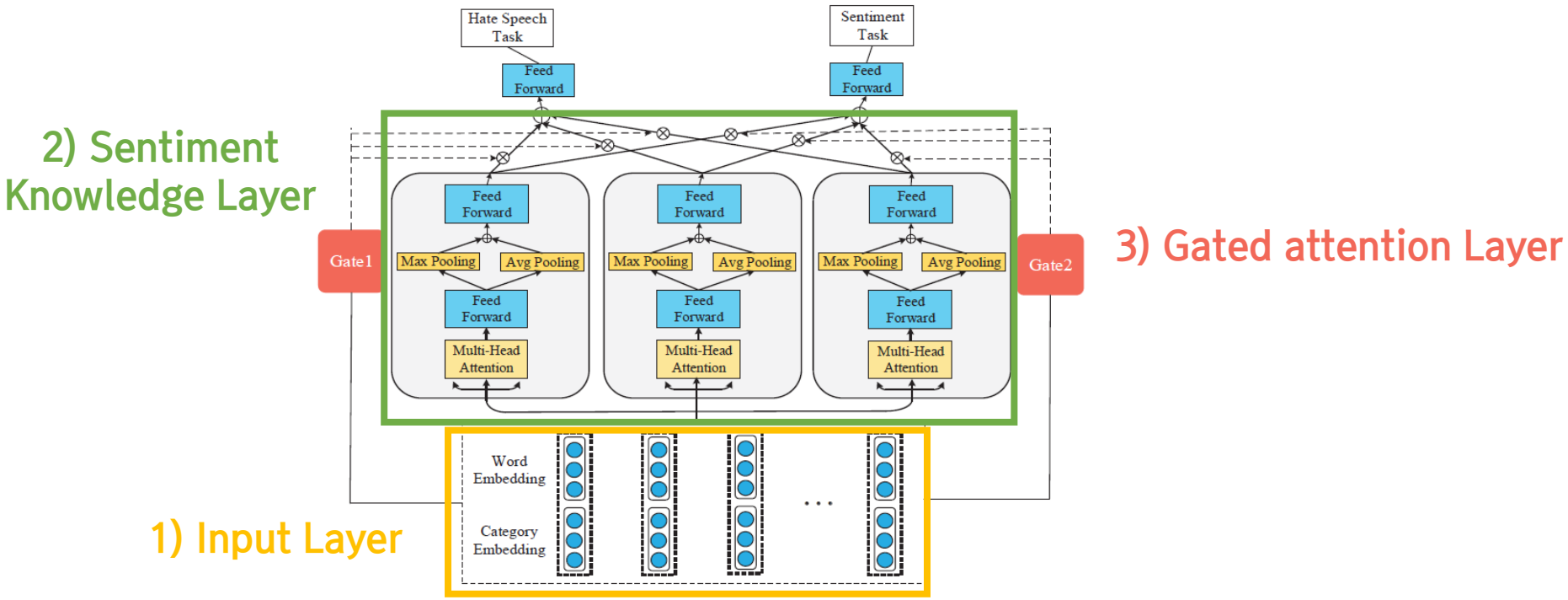


Figure 1: The overall framework of our proposed Hate Speech Detection based on Sentiment Knowledge Sharing(SKS).

III. Methodology

1) Input Layer

- Hate speech 는 부정적인 정서를 담고 있는 경우가 많다.

  ex. 가족들이 다 인물이 없네 ㅉㅉ, 개돼지집단 개한민족의 실상.

  - 문장의 단어가 경멸적인(derogatory) 단어인지 포착하는 것에 주의를 기울이면 모델 성능 개선에 도움이 될 수 있음.

Word embedding

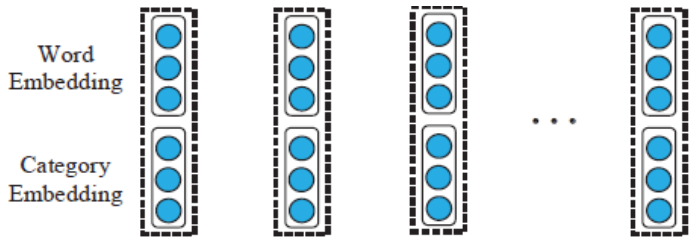
- $S = \{w_1, w_2, \dots, w_N\}$ 에 대해 word embedding 사용 -〉 w_i를 vector x_i 로 변환 ($x_i \in \mathbb{R}^d$, d는 dimension)

  I, am, so, happy

Category embedding

- 장애, LGBT, 민족, 종교, 혐오 단어로 이루어진 경멸어 사전 제작
- 문장 S 가 들어오면, w_i 경멸어 포함 or 포함하지 않음 두 카테고리로 나눈 다음, 각 단어(w_i) 를 카테고리에 할당
- 카테고리의 각 단어는 vector C로 랜덤하게 초기화 $C = (c_1, c_2, \dots, c_n), c_i \in \mathbb{R}^{d'}$
- 경멸어 정보를 활용하기 위해 category embedding c_i 를 word embedding x_i에 합침

$$x_i' = x_i \oplus c_i \quad (\oplus\text{는 vector concatenation 연산자})$$



III. Methodology

2) Sentiment Knowledge Sharing Layer (이 레이어를 사용하는 이유)

- 문화적으로 모욕적인 의미를 가진 단어가 감성어에는 반영되지 않을 수 있음

  ex. 유대인들은 다 저급한 돼지들이야.

  =〉 '돼지'는 중립적인 단어지만, 이 문장에서는 유대인(사람)을 돼지로 비유했으므로 모욕적인 발언이 됨.

  ex. I'm so fucking good

  =〉 경멸어이지만, 부정의 의미보다는 good 을 강조하는 부사로 쓰임.

혐오표현을 (사전을 이용해) 감지하는 것만으로는 만족스러운 성과를 얻을 수 없음

III. Methodology

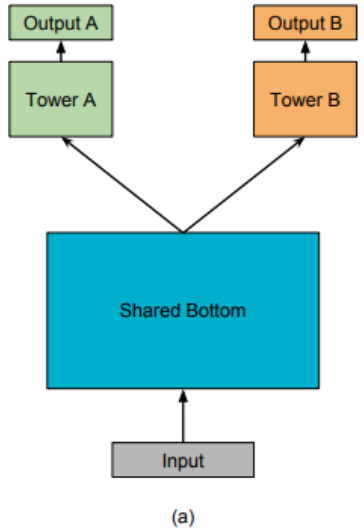
2) Sentiment Knowledge Sharing Layer (이 레이어를 사용하는 이유)

- Hate speech 빅데이터는 부족한 반면, sentimental analysis 는 더 오랫동안 연구되어왔으므로 고품질 레이블링 데이터셋 풍부

⇒ 따라서 Hate speech detection 과 sentimental analysis 의 Multi-task learning 방법 사용

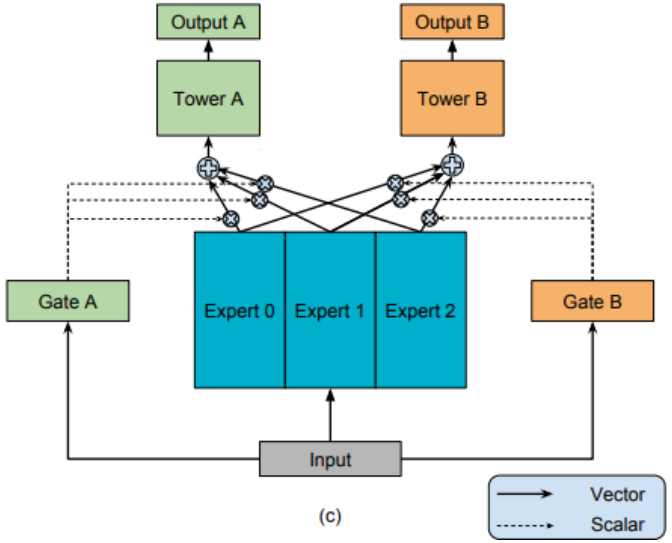
Sentiment Knowledge Sharing Layer 의 구조

- 보통의 Multi-task learning 에서는 shared-bottom 구조를 주로 사용, 하지만 우리는 Mixture of Expert 구조 채택



Shared bottom model

shared bottom 구조를 사용하면
하위 은닉 구조 계층 공유



Multi-gate MoE model

Task 간 차이를 알게 하고,
선택적으로 Expert 사용

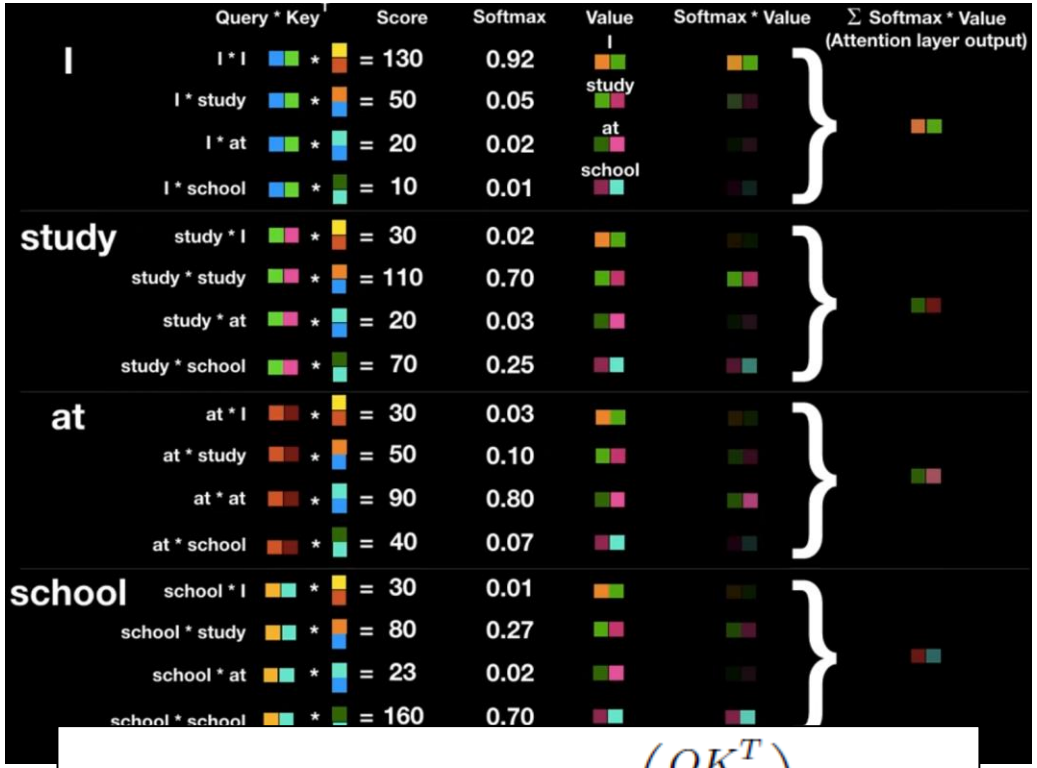
Jiaqi Ma et al. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts

III. Methodology

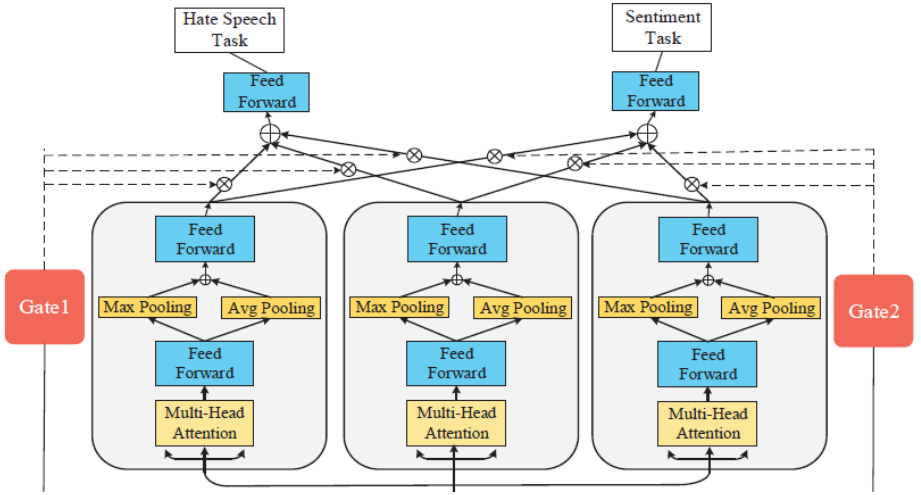
2) Sentiment Knowledge Sharing Layer

Multi-head Attention Layer

- self-attention 메커니즘은 문장 내 각 단어와 다른 단어의 의미 유사성과 의미 특징 계산



$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{d_1}\right)\text{V} \qquad (1)$$



$$M_i = \text{Attention}(\text{QW}_i^Q, \text{KW}_i^K, \text{VW}_i^V) \qquad (2)$$

$$H^s = \text{concat}\,(M_1, M_2, \ldots, M_l)\, W_o \qquad (3)$$

→ final representation

III. Methodology

2) Sentiment Knowledge Sharing Layer

Pooling Layer

- Shen et al. 2018. On the use of word embeddings alone to represent natural language sequences.

  => max pooling 과 average pooling 을 동시에 사용하여 단일 pooling 보다 성능향상을 보여줌.

- 따라서 우리도 max pooling 과 average pooling 동시에 사용

Max Pooling

1	0	0	4
5	2	4	8
2	1	2	6
0	1	2	2

→

5	8
2	6

Average Pooling

1	0	0	4
5	2	4	8
2	1	2	6
0	1	2	2

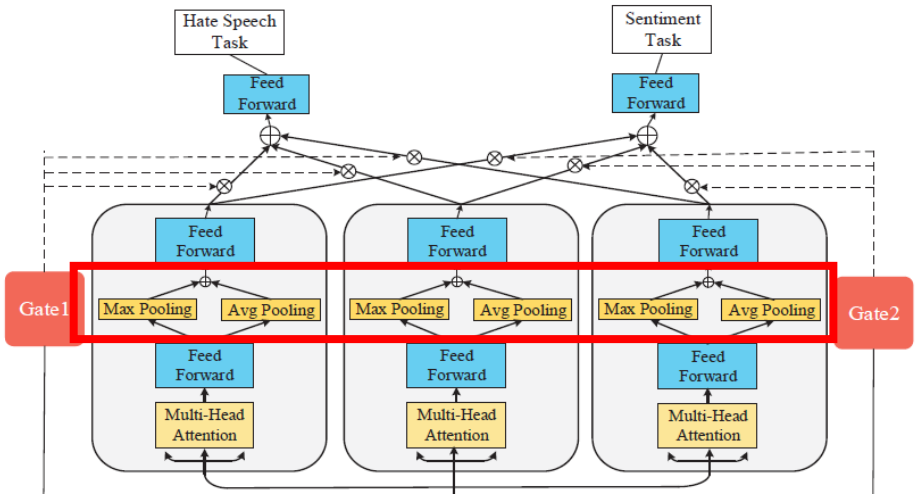
→

2	4
1	3

$$P_m = \text{Pooling\_max}\left(\mathrm{H}^s\right) \qquad (4)$$

$$P_a = \text{Pooling\_average}\left(\mathrm{H}^s\right) \qquad (5)$$

$$P_s = \text{concat}\left(\mathrm{P_m}, \mathrm{P_a}\right) \qquad (6)$$



III. Methodology

3) Gated Attention

- Gated Attention은 Gate 가 입력에 따라 사용할 Expert 를 선택하는 방법을 배울 수 있음.
- Task(hate speech, sentiment) 마다 가중치 선택이 다르므로 task 마다 gate 가 있음.
- 특정 Gate k 의 출력은 각 Expert 가 선택될 확률을 나타냄.

$$g^k(x) = \text{softmax}\left(W_{gn} * gate\left(x\right)\right) \qquad (7)$$

$$f^k(x) = \sum_{i=1}^{n} g^k(x)_i f_i(x) \qquad (8)$$

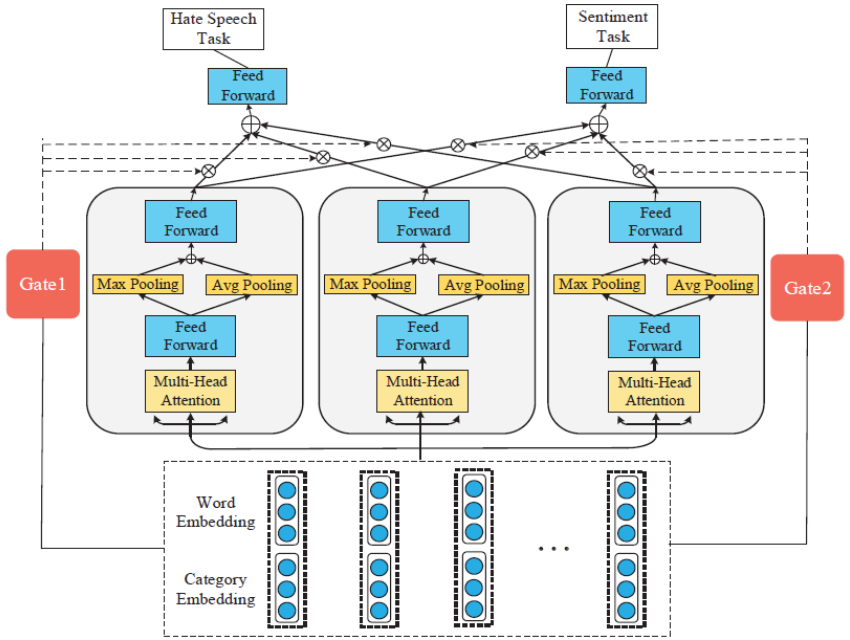


Figure 1: The overall framework of our proposed Hate Speech Detection based on Sentiment Knowledge Sharing(SKS).

III. Methodology

4) Model Training

- loss function 으로 categorical-cross-entropy 에 L2 regularization 을 더한 것을 사용

$$loss = -\sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2 \qquad (10)$$

i : 문장의 인덱스
j : class
λ : L_2 regularization 변수
θ : parameter set

IV. Experiments

Dataset

- DV: Davidson dataset (Davidson et al. 2017. Automated hate speech detection and the problem of offensive language)
  - hate speech 데이터셋, hate 가 적은 불균형 데이터셋임.
- SE: SemEval2019 task5 (Basile et al. 2019. Semeval-2019 task 5: Multilinugual detection of hate speech against immigrants and women in tweeter)
  - train 9000, validation 1000, test 2971 개의 문장으로 구성되어있음.
- SA: Sentiment Analysis (dataset from Kaggle 2018 https://www.kaggle.com/dv1453/twitter-sentiment-analysis-analytics-vidya)

Evaluation Metrics

- Accuracy, F1-score 사용

IV. Experiments

Model	DV		SE	
	Acc	F1(wei)	Acc	F1(macro)
SVM*	-	87.0	49.2	45.1
LSTM*	94.5	93.7	55.0	53.0
GRU*	94.5	93.9	54.0	52.0
CNN-GRU*	-	94.0	62.0	61.5
BiLSTM*	94.4	93.7	53.5	51.9
BiGRU_Stacked*	-	-	56.0	54.6
USE_SVM*	-	-	65.3	65.1
BERT*	94.8	95.8	-	48.8
GPT*	-	-	-	51.5
SKS	**95.1**	**96.3**	**65.9**	**65.2**

Table 2: Comparison with existing methods. The results with superscript * are imported from the literature. The best results in each type are highlighted.

- DV의 경우, 5-fold cross validation 을 사용해서 평균 accuracy 와 weighted F1 사용
- SE의 경우, test set의 성능으로 accuracy와 macro f1 사용

IV. Experiments

Model	DV		SE	
	Acc	F1(wei)	Acc	F1(macro)
SVM*	-	87.0	49.2	45.1
LSTM*	94.5	93.7	55.0	53.0
GRU*	94.5	93.9	54.0	52.0
CNN-GRU*	-	94.0	62.0	61.5
BiLSTM*	94.4	93.7	53.5	51.9
BiGRU_Stacked*	-	-	56.0	54.6
USE_SVM*	-	-	65.3	65.1
BERT*	94.8	95.8	-	48.8
GPT*	-	-	-	51.5
SKS	**95.1**	**96.3**	**65.9**	**65.2**

- **features 를 기반으로 한 SVM 성능은 NN 보다 훨씬 떨어짐. 특히, SE 데이터셋에서는 Acc, F1 이 50% 도 안되는 것을 보여줌.**
  - ✓ 신경망모델이 hate speech detection 을 위한 단어의 의미 관계를 더 잘 포착할 수 있음을 나타냄.
- **Hybrid-NN(CNN-GRU, BiGRU-capsule) 의 성능은 기존 RNN(LSTM 등)과 비교해보면 더 우수함.**
  - ✓ 신경망 모델에 다른 층을 쌓음으로써, 딥러닝 모델은 높은 수준의 특징을 학습할 수 있게 됨.

V. Ablation Study

Model	DV		SE	
	Acc	F1(wei)	Acc	F1(macro)
-sc	94.0	94.0	59.6	59.3
-s	94.5	94.3	61.3	61.3
SKS	**95.1**	**96.3**	**65.9**	**65.2**

Table 3: the results of ablation experiments The best results in each type are highlighted.

-sc: sentiment knowledge sharing, categorical embedding 제거

-s : sentiment knowledge sharing 제거

- 두 데이터셋의 성능은 -sc 로 크게 감소
- 하지만 감소해도 우리 모델(SKS)은 기존 Hybrid-NN 보다 우수

=> SKS 는 문장의 잠재의미 특징을 더 잘 학습함.

- -s 일 때 -sc 보다 성능이 약간 향상되는 이유

=> 경멸어 사전 정보는 Hate speech 와 관련이 높지만, 모델을 너무 민감하게 만듦. 이것은 성능에 제한적 영향을 줌.

Model	DV		SE	
	Acc	F1(wei)	Acc	F1(macro)
no-gate	94.8	95.9	64.7	64.3
SKS	**95.1**	**96.3**	**65.9**	**65.2**

Table 4: the influence of gated attention.

모델에서 Gated Attention 의 역할을 분석

- no-gate 보다 SKS 가 성능이 향상됨.

=> 서로 다른 게이트로 인한 분리가 서로 어떻게 겹치는지를 결정함으로써 작업 관계를 정교한 방식으로 모델링 함.

❖ **Task 들이 연관성이 높다면, 지식 공유가 더 나은 성과를 보임.**

V. Ablation Study

Sentiment dataset 크기에 따른 성능 분석

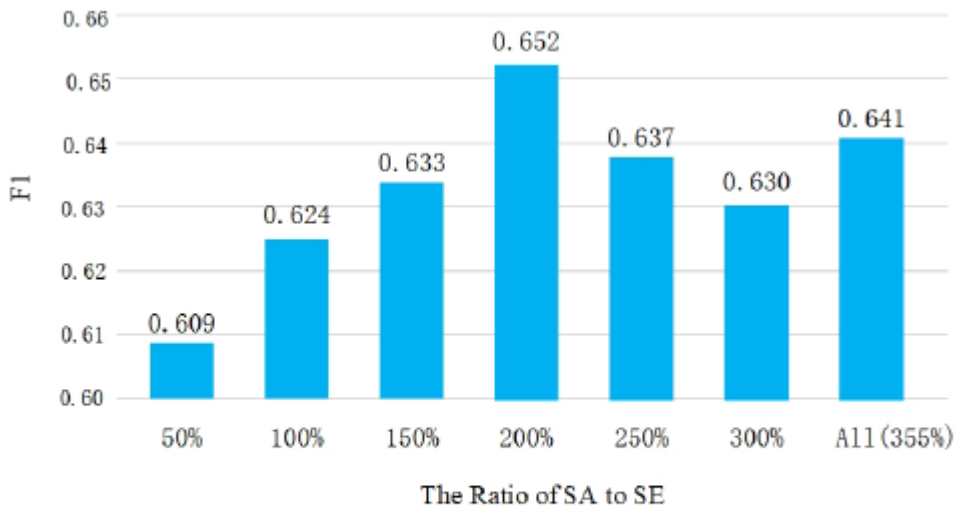


Figure 2: the influence of the scale of sentiment data set.

hate sentiment hate

- DV 와 SA 의 규모는 비슷하므로 SE 에 대한 분석에 집중함.

sentiment hate

- SA : SE 가 1 : 2 일때 가장 낮은 성능,
- SA : SE 가 2 : 1 일 때 가장 높은 성능, 그 이후 성능 하락

=〉 Multi-task learning 에서 데이터의 비율도 성능에 직접적인 영향을 미치고 있음을 보여주고 있음.

VI. Conclusion

Contributions

1) 감정지식을 활용한 Multi-task Learning(MTL) 모델 소개
2) 공유작업을 더 잘하기 위해 multi-head attention mechanism과 Gated Attention을 사용하는 새로운 프레임워크 제안
3) 공개된 2개의 데이터셋에서 다른 baseline 모델들에 비해 최첨단 성능 달성 입증