Hate Speech Detection based on Sentiment Knowledge Sharing

Nhóm 4

Mạnh Hùng Công Thiện Minh Trí Bảo Lợi Đình Quân

Khoa Khoa học máy tính

Trường đại học Công Nghệ Thông Tin

Ngày 3 tháng 1 năm 2024

Nội dung

Nội dung

Giới thiệu bài báo

Bài báo: https://aclanthology.org/2021.acl-long.556

Nội dung: Hate Speech Detection based on Sentiment Knowledge Sharing

Các tác giả: Xianbing Zhou, Yong Yang, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, Hongfei Lin

Code: https://github.com/1783696285/sks

Trích dẫn: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)

Ngày công bố: Tháng 8 năm 2021

Giới thiệu bài toán

Đặt vấn đề

Ví dụ: Go fucking kill yourself.

- Sự phổ biến của Internet và mạng xã hội
  $\rightarrow$ Điều kiện để hate speech lan truyền một cách rộng rãi
- Hậu quả: Bị phân biệt đối xử, ảnh hưởng tâm lý, bạo lực.. và đây là vấn đề nghiêm trọng của xã hội

$\Rightarrow$ **Vấn đề**: Làm thế nào để có thể phát hiện hate speech tự động, nhanh chóng và chính xác, đồng thời can thiệp sớm để ngăn chặn chúng?

Thách thức

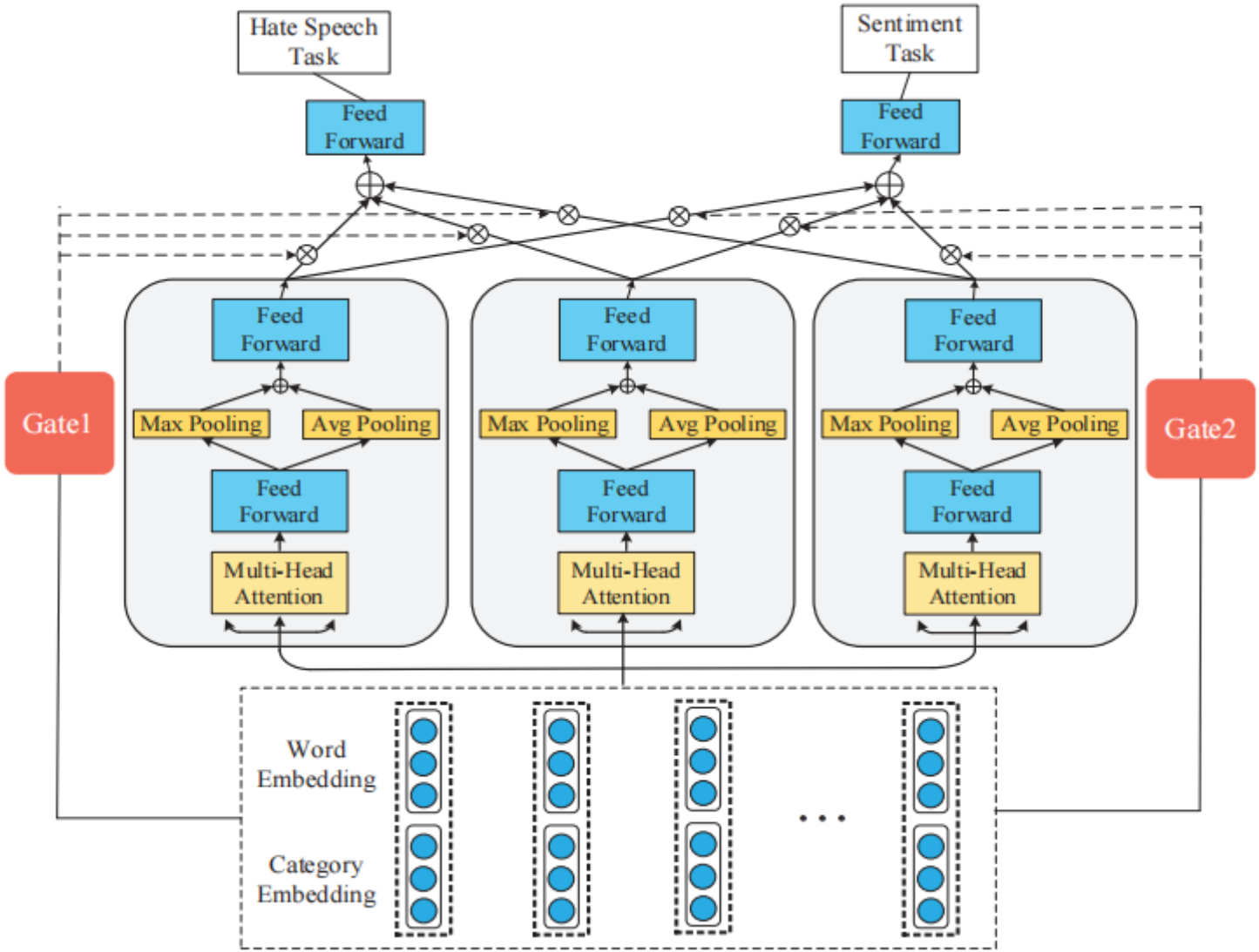
1. Cấu trúc của ngôn ngữ tự nhiên phức tạp
2. Các giải pháp trước chủ yếu dựa vào **quy tắc** hoặc **trích xuất đặc trưng thủ công**
3. **Phương pháp máy học:** sử dụng các đặc trưng nhân tạo ở mức độ nông (shallow features)
4. **Phương pháp học sâu:** chưa khai thác đầy đủ đặc trưng sentiment và nguồn kiến thức sentiment ở bên ngoài

Đóng góp

1. Sử dụng thông tin sentiment
   - Tích hợp các từ mang tính xúc phạm có trong câu vào mạng neural network
   - Sử dụng multi-task learning để mô hình học và chia sẻ kiến thức sentiment
2. Đề xuất một framework mới
   - Sử dụng nhiều đơn vị trích xuất đặc trưng
   - Áp dụng cơ chế Gated Attention để kết hợp các đặc trưng
3. Thực nghiệm trên 2 tập dữ liệu: **SemEval-2019 task-5, Davidson**

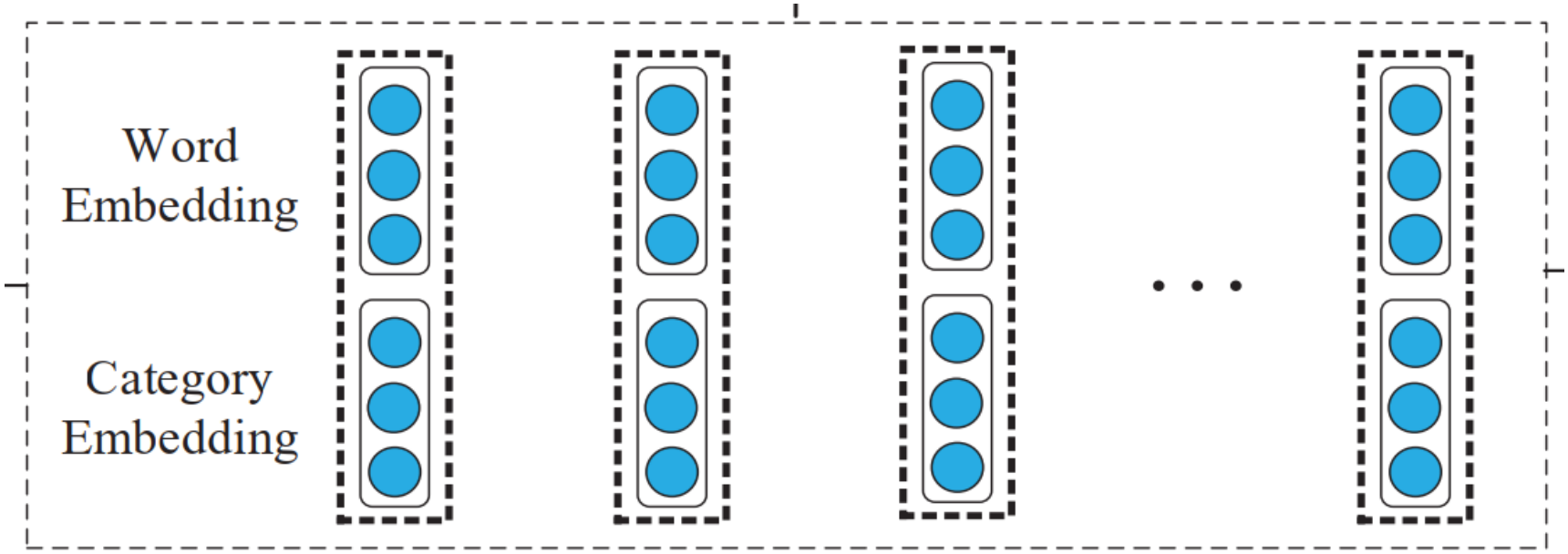
Nội dung

Tổng quan



Hình 1: Kiến trúc tổng quan của phương pháp **Hate Speech detection based on Sentiment Knowledge Sharing (SKS)**

Input layer



Hình 2: Input layer

Ý tưởng chung

- Pretrained word embedding
- Category embedding
- Kết hợp (word embedding, category embedding)

Input layer

Word Embedding

- Dựa trên giả định về phân phối
- Ánh xạ các từ vào không gian đặc trưng nhiều chiều
- Duy trì thông tin về ngữ nghĩa

Input layer

Quan sát của tác giả

- Hate speech chứa những từ ngữ xúc phạm
  Ví dụ: *Go fucking kill yourself and die already ugly pile of shit scumbag*
- Việc xác định những từ ngữ xúc phạm góp phần lớn vào việc xác định hate speech
  $\longrightarrow$ Tạo một từ điển chứa các từ ngữ xúc phạm

Input layer

Category Embedding

Bộ từ vựng của từ điển

- Nguồn: Wikipedia, trang Noswearing
- 5 loại chính: Hate speech, Disability, LGBT, Ethnic, Religious

Chức năng

- Dùng để đánh giá mỗi từ trong câu có phải từ căm thù hay không

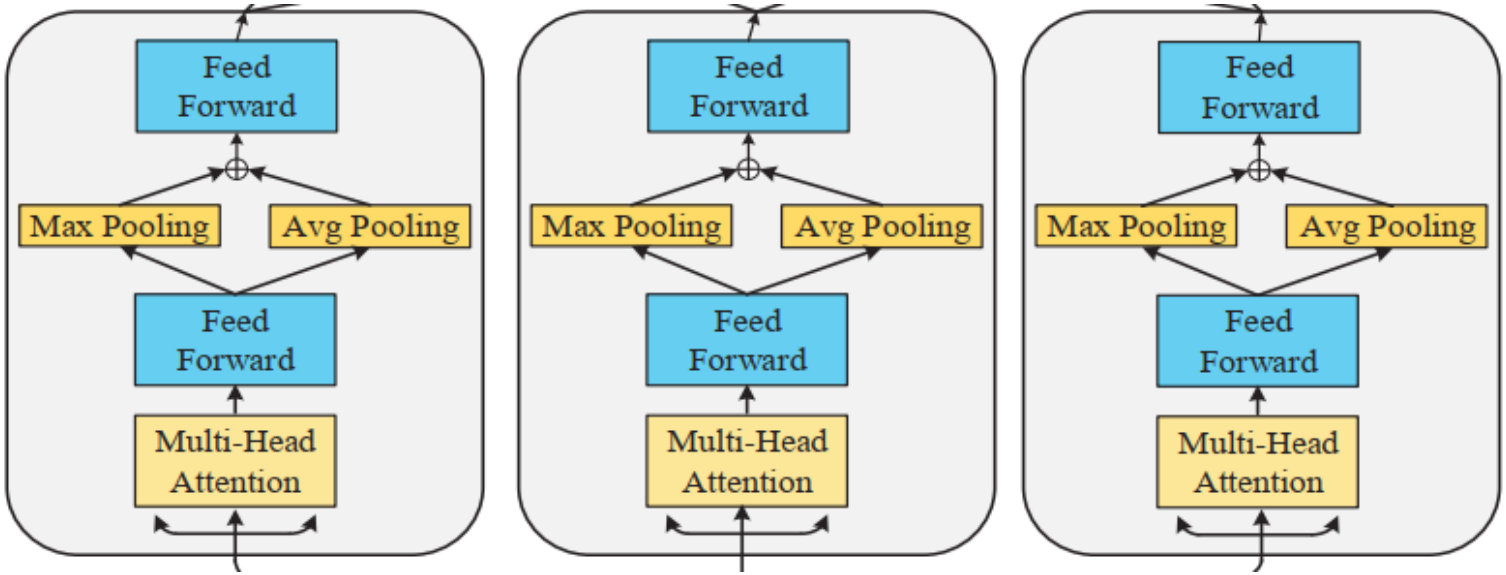
Sentiment Knowledge Sharing layer

Bối cảnh

- Ý nghĩa xúc phạm ẩn trong ngữ cảnh văn hóa, không còn thể hiện ở mỗi từ ngữ
  - **non-hate:** i'm so fucking ready!
- **Hate speech detection:** Thiếu dữ liệu huấn luyện chất lượng cao
- **Sentiment analysis:** Có nhiều dữ liệu chất lượng được gán nhãn

$\Rightarrow$ Mối tương quan cao giữa 2 task, áp dụng Multi-task learning để chia sẻ kiến thức sentiment

Sentiment Knowledge Sharing layer



Hình 3: Sentiment Knowledge Sharing layer

Ý tưởng chung

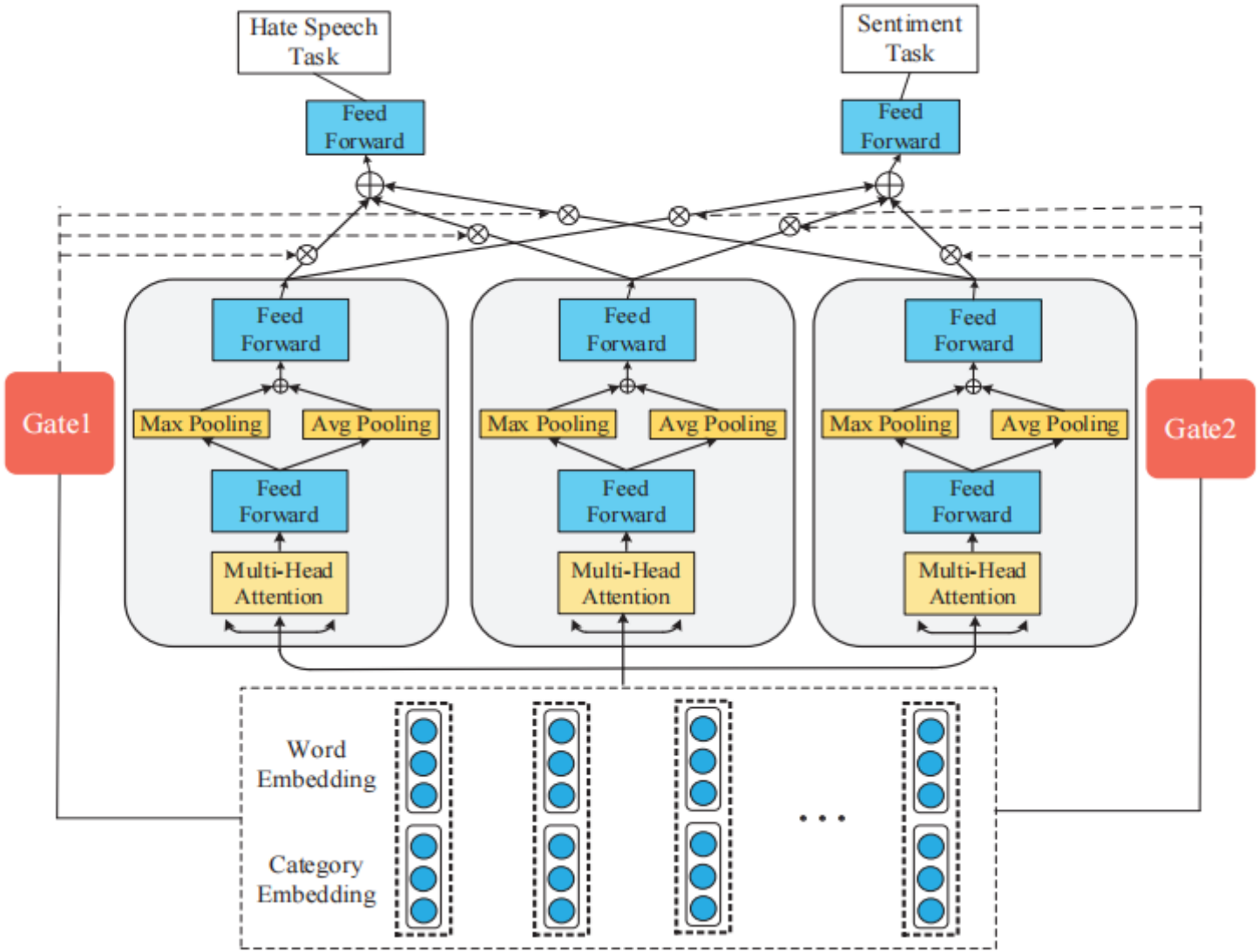
- Sử dụng cấu trúc framework Mix-of-Expert (MoE): gồm nhiều đơn vị trích xuất đặc trưng giống nhau
- Các đơn vị trích xuất đặc trưng: bao gồm một lớp multi-head attention, hai lớp pooling và hai lớp feed forward

Sentiment Knowledge Sharing layer

Multi-head Attention Layer

- Cơ chế self-attention: tính toán mức độ tương quan về ngữ nghĩa giữa các từ trong câu để xác định mối liên hệ giữa chúng
- Cơ chế multi-head attention:
  - Sử dụng nhiều head self-attention chồng lên nhau để phân tích mối quan hệ giữa các phần trong câu
  - Mỗi head tập trung vào một khía cạnh khác của câu input
  - Kết hợp kết quả của các head lại để tạo ra vector biểu diễn feature cuối cùng

Gated Attention



Hình 4: Kiến trúc tổng quan của phương pháp **Hate Speech detection based on Sentiment Knowledge Sharing (SKS)**

Gated Attention

Ý tưởng

- Gated attention chọn một tập các feature extraction unit để sử dụng, dựa trên nội dung và ngữ cảnh của input đó
- Tác vụ khác nhau $\rightarrow$ có Gate khác nhau
- Cấu trúc của Gate unit $\sim$ feature extraction unit

Kết quả

- Output Gate thứ k: là một vector thể hiện xác suất của các feature extraction unit (FEU) được chọn
- Đối với mỗi Gate:
  $\sum(\textbf{output mỗi gate} * \textbf{FEU}) \rightarrow$ vector biểu diễn cuối cùng của input

Nội dung

Dataset

Dataset	total	Classes
SemEval2019 task5(SE)	11.971	hate (5.035)
		non-hate (6.936)
Davidson(DV)	24.783	hate (1.430)
		non-hate (23.353)
Sentiment Analysis(SA)	31.962	negative(2.242)
		positive(29,720)

Bảng 1: Số liệu thống kê về các dataset sử dụng trong thực nghiệm

So sánh với các phương pháp baseline

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**

Bảng 2: So sánh với các phương pháp hiện có

Thực nghiệm của bài báo

	DV		SE	
Model \ Metric	Acc	F1(wei)	Acc	F1(macro)
-sc	94.0	94.0	59.6	59.3
-s	94.5	94.3	61.3	61.3
no-gate	94.8	95.9	64.7	64.3
SKS	**95.1**	**96.3**	**65.9**	**65.2**

Bảng 3: Bảng kết quả các thực nghiệm của các model

Trong đó:

- "–sc": không sử dụng sentiment feature và category embedding
- "-s": không sử dụng sentiment feature
- "no-gate": không sử dụng Gated attention layer

Thực nghiệm của nhóm

	Paper	**Ours**
embedding data	glove.840B.300d	glove.6B.300d
learning rate	0.001	0.002

Bảng 4: Các tham số thay đổi khi training

	DV		**SE**	
metrics	Acc	F1(wei)	Acc	F1(macro)
SKS	95.1	96.3	65.9	65.2
Ours	93.0	93.0	61.0	61.0

Bảng 5: Kết quả thực nghiệm trên 2 tập datasets

Nội dung

Bàn luận

Ưu điểm

- Đạt hiệu quả vượt trội so với các phương pháp baseline

Nhược điểm

- Khó khăn trong việc thu thập và lựa chọn các nguồn dữ liệu phù hợp
- Tốn kém tài nguyên tính toán do phức tạp của mô hình

Kết luận

1. Sử dụng nhiều **feature extraction unit** để các task chia sẻ tham số, tận dụng được kiến thức liên quan
2. Áp dụng cơ chế gated attention để kết hợp các đặc trưng một cách linh hoạt
3. Mô hình đề xuất có thể sử dụng đầy đủ thông tin sentiment của câu input và nguồn dữ liệu sentiment bên ngoài

Hướng nghiên cứu tiếp theo

- Áp dụng cho nhiều ngôn ngữ khác ngoài tiếng Anh
- Tập trung vào lựa chọn và mở rộng các nguồn dữ liệu huấn luyện, bao gồm các dạng hate speech khác nhau, các loại dữ liệu sentiment phù hợp với quy mô lớn hơn

Nội dung

Tài liệu tham khảo chính

 https://aclanthology.org/2021.acl-long.556

Cảm ơn các bạn và cô đã lắng nghe!