User-defined Content Detection Framework

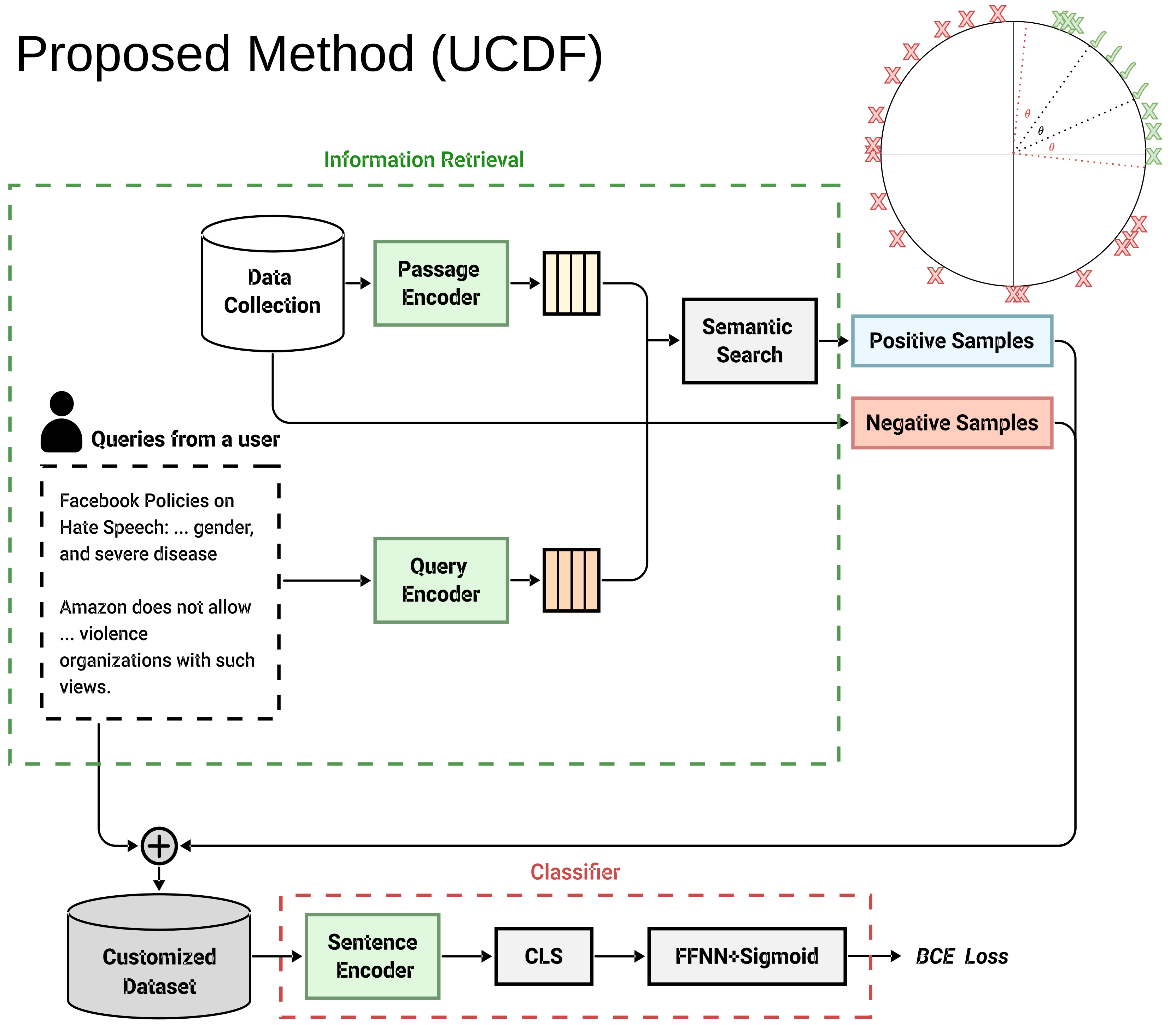
Keonwoo Kim[1,2], Sungzoon Cho[1], Younggun Lee[2]

1 Seoul National University, 2 Neosapience

Introduction

- Various research related to hate speech detection has been conducted
- Introduce the user-defined content detection task as new research topic
- Propose a novel interactive framework called the ***user-defined content detection framework (UCDF)***
- Suggest a novel method to build a customized dataset for the detection task

Proposed Method (UCDF)



The process of UCDF includes two stages

1) Building a customized dataset from the queries using **DPR**

2) Fine-tuning a classifier using the customized dataset

* Additional interactive stage can be placed after the training of the classifier

- Add more queries to refine the customized dataset
- Negative queries of the opponents of the user-defined content can be added

Algorithm 1: Building customized dataset

```
Result: An arracy $C$ of customized dataset
input: An array $S$ of $n$ queries;
$\Theta \leftarrow$ pairwise cosine similarity on an array $S$ ;
for $query = 1, 2, \ldots, N$ do
    retrieve $k$ passages from $DataCollection$ using $NearestNeighbors$;
    for $passage = 1, 2, \ldots, k$ do
        $\alpha \leftarrow CosineSimilarity(query, passage)$ ;
        if $\Theta \leq \alpha$ then
            $C \leftarrow C + passage$
        else
    end
end
```

Experimental Setup

- Data Collection: Wikipedia dump 2018
- Bi-encoder in DPR (Passage encoder, Query encoder)
  : SimCSE (sup-bert-base-uncased) trained on NaturalQuestion dataset
- Classifier (Sentence encoder)
  : SimCSE (sup-bert-base-uncased)

(1) Hate speech detection task

Examples of queries used in hate speech detection task

Company name	Query
Facebook	Facebook Policies on Hate Speech: We define hate speech as a direct attack against people — rather than concepts or institutions— based on protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and severe disease.
Youtube	We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: race or ethnic origin, religion, disability, gender, age, veteran status, or sexual orientation/gender identity.
Twitter	You may not promote violence against or directly attack or threaten other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

- Test data: Tweets hate speech detection / Davidson et al.
- Baseline models
  - RoBERTa [a]: trained on Jigsaw dataset (SkolkovoInstitute)
  - RoBERTa [b]: trained on machine-generated toxic texts (tomh)

(2) User-defined content detection experiment

Examples of queries used in user-defined content detection (Religion)

Class	Query
Positive	Judaism is the world's oldest monotheistic religion, dating back nearly 4,000 years.
Positive	Christianity is the most widely practiced religion in the world, with more than 2 billion followers.
Positive	Islam is the second largest religion in the world after Christianity, with about 1.8 billion Muslims worldwide.
Negative	Buddhism is a faith that was founded by Siddhartha Gautama ("the Buddha") more than 2,500 years ago in India.
Negative	Hinduism is the world's oldest religion, according to many scholars, with roots and customs dating back more than 4,000 years.

Examples of queries used in user-defined content detection (South Korea)

Class	Query
Positive	South Korea, K-pop, Seoul, Kimchi
Negative	North Korea, Japan, China

- Test data: Customized Religion / South Korea dataset
  - Include hard negative / easy samples on test data
  - Build customized negative samples with three methods
    1. without negative queries (0%) + only with random sampling (100%)
    2. only with negative queries (100%) + without random sampling (0%)
    3. with negative queries (50%) + with random sampling (50%)

Experimental Results

Qualitative results on hate speech detection task

	Tweets Hate Speech Detection			Davidson et al.		
Model	F1-score (%)	Accuracy (%)	AUC (%)	F1-score (%)	Accuracy (%)	AUC (%)
RoBERTa [a][2]	19.94	90.96	67.13	**37.61**	46.41	**63.52**
RoBERTa [b][3]	19.95	90.21	23.35	34.84	45.80	59.31
UCDF - SimCSE (avg)	**32.06**	**91.65**	78.47	13.02	**77.72**	22.21
UCDF - SimCSE (min)	30.54	79.51	**81.13**	23.11	74.84	58.38

Qualitative results on User-defined content detection experiment

	Customized Religion			Customized South Korea		
Sampling method	F1-score (%)	Accuracy (%)	AUC (%)	F1-score (%)	Accuracy (%)	AUC (%)
w/o nq, w/ rand	78.95	79.49	84.69	78.10	79.46	84.18
w/ nq, w/o rand	68.97	76.92	82.70	72.58	69.64	76.79
w/ nq, w/ rand	**95.24**	**96.15**	**97.53**	**85.44**	**86.61**	**93.94**

Examples of customized test data (Religion)

Class	Example
Positive	Jewish people believe there's only one God who has established a covenant—or special agreement—with them.
Positive	Muslims are monotheistic and worship one, all-knowing God, who in Arabic is known as Allah.
Easy Negative	Enter your destination & your Tesla will automatically include Supercharging stops in your route
Easy Negative	Comments section of Yahoo controlled by alt-reality biased moderators supporting lies harmful to the Nation.
Hard Negative	The religion's founder, Buddha, is considered an extraordinary being, but not a god. The word Buddha means "enlightened."
Hard Negative	Hinduism is unique in that it's not a single religion but a compilation of many traditions and philosophies.

Examples of customized test data (South Korea)

Class	Example
Positive	Bong Joon Ho's Parasite made history for bagging 3 awards at the 2020 Oscars, which was the most of any film nominated.
Positive	Hangul classifies as one of the Altaic languages, is affiliated to Japanese, and contains some Chinese loanwords.
Easy Negative	Recently after more than 20 years as a Google account holder my YouTube channel was suspended without warning and without any reason given.
Easy Negative	Jordi Cruyff has signed his contract as FC Barcelona's new sporting director of football. He has already been serving in the role since July 1.
Hard Negative	The greatest health threat in North Korea is hunger.
Hard Negative	The Huizhou Ancient Town is a famous historical and cultural city in southern Anhui Province with over 2000 years of history.

References

- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-tau. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

- Gao, T., Yao, X., & Chen, D. (2021). SIMCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

- Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1).