(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년12월02일
(11) 등록번호 10-2334018
(24) 등록일자 2021년11월29일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01) *G06F 16/903* (2019.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06F 40/279 (2020.01)
G06F 16/90344 (2019.01)
(21) 출원번호 10-2019-0151952
(22) 출원일자 2019년11월25일
심사청구일자 2019년11월25일
(65) 공개번호 10-2020-0063067
(43) 공개일자 2020년06월04일
(30) 우선권주장
1020180148087 2018년11월27일 대한민국(KR)
(56) 선행기술조사문헌
KR101913284 B1*
(뒷면에 계속)
전체 청구항 수 : 총 15 항

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
한요섭
서울특별시 은평구 진관1로 77-8, 403동 204호 (진관동,은평뉴타운폭포동아파트)
박준우
경기도 부천시 소향로 11, 20층 B-2002호(상동, 코오롱이데아폴리스)
(74) 대리인
민영준

심사관 : 김경완

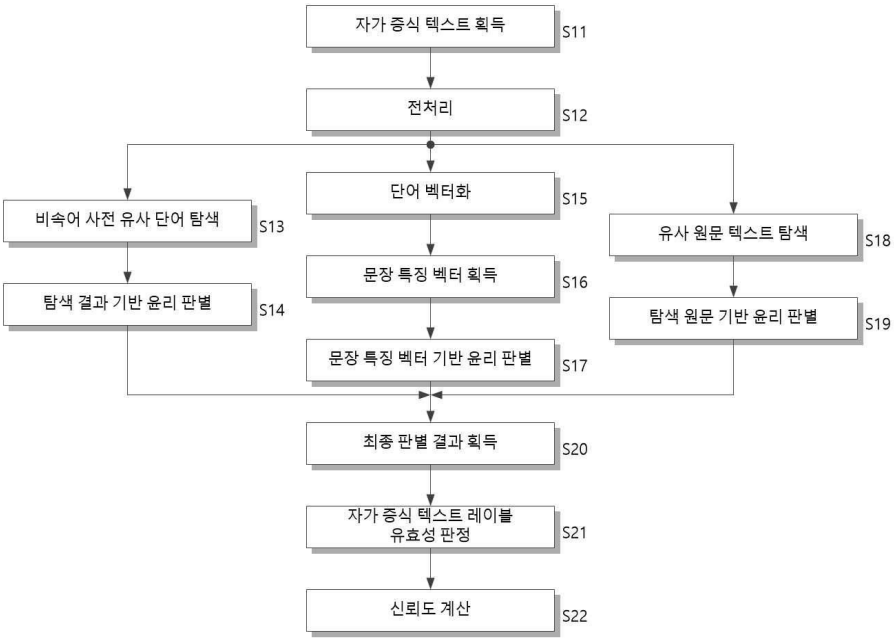
(54) 발명의 명칭 **자가 증식된 비윤리 텍스트의 유효성 검증 장치 및 방법**

(57) 요 약

본 발명은 윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 이용하여 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득하는 텍스트 획득부, 자가 증식 텍스트를 인가받고, 인가된 자가 증식 텍스트에서 미리 획득된 비속어 사전에 등재된 비속어와 기기정된 레벨 이상으로 유사한 단어를 탐색하여 자가 증식 텍스트

(뒷면에 계속)

대 표 도



의 비윤리를 판별하는 사전 기반 판별부, 자가 증식 텍스트를 인가받아 단어 단위로 벡터화하고, 벡터화된 단어로부터 미리 학습된 패턴 추정 방식에 따라 문장 특징 벡터를 추출하여 자가 증식 텍스트의 비윤리를 판별하는 학습 모델 기반 판별부, 자가 증식 텍스트와 가장 유사한 학습용 텍스트를 탐색하고, 탐색된 학습용 텍스트의 레이블에 따라 자가 증식 텍스트의 비윤리를 판별하는 원문 기반 판별부 및 사전 기반 판별부, 학습 모델 기반 판별부 및 원문 기반 판별부 각각에서 판별된 자가 증식 텍스트의 비윤리를 판별 결과를 조합하여, 자가 증식 텍스트에 대한 최종 판별 결과를 획득하는 판별 결과 비교부를 포함하는 텍스트의 유효성 검증 장치 및 방법을 제공할 수 있다.

(52) CPC특허분류

G06N 3/08 (2013.01)

(56) 선행기술조사문헌

KR1020170073354 A*

KR1020180008247 A*

부석준 외 3명, "비윤리적 SNS 댓글 분류를 위한구문기반 CNN 과 의미기반 LSTM 의 앙상블 기법", 2017.05.31., pp.6-19. 1부.*

KR1020130022075 A

*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호	2018-0-00247
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터(NIPA산하)
연구사업명	정보통신방송연구개발사업
연구과제명	[이지바로][주관/창원대학교] GANs를 이용한 딥러닝용 학습데이터 자가 증식 기술 및 유효성 검증 기술 개발 (1/2)
기 여 율	1/1
과제수행기관명	창원대학교 산학협력단
연구기간	2018.04.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 이용하여 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득하는 텍스트 획득부;

자가 증식 텍스트를 인가받고, 인가된 자가 증식 텍스트에서 미리 획득된 비속어 사전에 등재된 비속어와 기기정된 레벨 이상으로 유사한 단어를 탐색하여 상기 자가 증식 텍스트의 비윤리를 판별하는 사전 기반 판별부;

자가 증식 텍스트를 인가받아 단어 단위로 벡터화하고, 벡터화된 단어로부터 미리 학습된 패턴 추정 방식에 따라 문장 특징 벡터를 추출하여 상기 자가 증식 텍스트의 비윤리를 판별하는 학습 모델 기반 판별부;

상기 자가 증식 텍스트와 가장 유사한 학습용 텍스트를 탐색하고, 탐색된 학습용 텍스트의 레이블에 따라 상기 자가 증식 텍스트의 비윤리를 판별하는 원문 기반 판별부; 및

상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과를 조합하여, 상기 자가 증식 텍스트에 대한 최종 판별 결과를 획득하는 판별 결과 비교부를 포함하되,

상기 생성되는 자가 증식 텍스트 각각은 윤리 또는 비윤리로 레이블링되며,

상기 자가 증식 텍스트의 생성 시에 윤리 또는 비윤리로 레이블링된 레이블과 상기 최종 판별 결과를 비교하여 동일하면 상기 자가 증식 텍스트의 레이블이 유효한 것으로 판정하고, 동일하지 않으면 유효하지 않은 것으로 판정하는 레이블 비교부를 더 포함하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 2

제1 항에 있어서, 상기 사전 기반 판별부는

상기 비속의 사전에 등재된 비속어와 상기 자가 증식 텍스트의 각 단어에 대해 N-그램 유사도 분석을 수행하여, 상기 자가 증식 텍스트에 비속어의 포함 여부를 판정하고, 비속어가 포함된 것으로 판정되면, 상기 자가 증식 텍스트를 비윤리로 판별하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 3

제1 항에 있어서, 상기 학습 모델 기반 판별부는

상기 자가 증식 텍스트의 각 단어를 임베딩하여 벡터화함으로써 다수의 단어 벡터를 획득하는 벡터 변환부;

미리 학습된 패턴 추정 방식에 따라 상기 다수의 단어 벡터의 특징을 누적하여 추출함으로써, 상기 문장 특징 벡터를 획득하는 문장 특징 추출부; 및

미리 학습된 패턴 분류 방식에 따라 상기 문장 특징 벡터를 분류하여, 상기 자가 증식 텍스트의 비윤리를 판별하는 특징 분류부를 포함하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 4

제3 항에 있어서, 상기 문장 특징 추출부는

LSTM(Long Short Term Memory)으로 구현되는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 5

제1 항에 있어서, 상기 판별 결과 비교부는

상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각에서 판별된 상기 자가 증

식 텍스트의 비윤리를 판별 결과에 대해 다수결 원칙을 적용하여 상기 최종 판별 결과를 획득하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 6

제1 항에 있어서, 상기 판별 결과 비교부는

상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과에 각각에 대해 기지정된 서로 다른 가중치를 할당하고, 할당된 가중치에 따라 윤리 또는 비윤리 중 더 높은 가중치가 할당된 결과를 상기 최종 판별 결과로 획득하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 7

삭제

청구항 8

제1 항에 있어서, 상기 레이블 비교부는

다수의 자가 증식 텍스트의 레이블에 대한 유효 판정 결과에 따라 자가 증식 텍스트의 신뢰도를 계산하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 9

제1 항에 있어서, 상기 자가 증식된 텍스트의 유효성 검증 장치는

상기 텍스트 획득부에서 획득된 자가 증식 텍스트에 대해 부가 구성 요소 제거하고, 문장 단위로 구분하여 상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각으로 전달하는 전처리부를 더 포함하는 자가 증식된 텍스트의 유효성 검증 장치.

청구항 10

자가 증식된 텍스트의 유효성 검증 장치에서 수행되는 자가 증식된 텍스트의 유효성 검증 방법으로서,

윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 이용하여 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득하는 자가 증식 텍스트 획득 단계;

자가 증식 텍스트에서 미리 획득된 비속어 사전에 등재된 비속어와 기기정된 레벨 이상으로 유사한 단어를 탐색하여, 상기 자가 증식 텍스트의 비윤리를 판별하는 사전 기반 판별 단계;

자가 증식 텍스트를 인가받아 단어 단위로 벡터화하고, 벡터화된 단어로부터 패턴 추정 방식이 미리 학습된 학습 모델을 이용하여 문장 특징 벡터를 추출하고, 추출된 문장 특징에 기반하여 상기 자가 증식 텍스트의 비윤리를 판별하는 학습 모델 기반 판별 단계;

상기 자가 증식 텍스트와 가장 유사한 학습용 텍스트를 탐색하고, 탐색된 학습용 텍스트의 레이블에 따라 상기 자가 증식 텍스트의 비윤리를 판별하는 원문 기반 판별 단계; 및

상기 사전 기반 판별 단계, 상기 학습 모델 기반 판별 단계 및 상기 원문 기반 판별 단계 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과를 조합하여, 상기 자가 증식 텍스트에 대한 최종 판별 결과를 획득하는 최종 판별 단계를 포함하되,

상기 생성되는 자가 증식 텍스트 각각은 윤리 또는 비윤리로 레이블링되며,

상기 자가 증식 텍스트의 생성 시에 윤리 또는 비윤리로 레이블링된 레이블과 상기 최종 판별 결과를 비교하여 동일하면 상기 자가 증식 텍스트의 레이블이 유효한 것으로 판정하고, 동일하지 않으면 유효하지 않은 것으로 판정하는 레이블 비교 단계를 더 포함하는 자가 증식된 텍스트의 유효성 검증 방법.

청구항 11

제10 항에 있어서, 상기 사전 기반 판별 단계는

상기 비속의 사전에 등재된 비속어와 상기 자가 증식 텍스트의 각 단어에 대해 N-그램 유사도 분석을 수행하여, 상기 자가 증식 텍스트에 비속어의 포함 여부를 판정하는 단계; 및

비속어가 포함된 것으로 판정되면, 상기 자가 증식 텍스트를 비윤리로 판별하는 단계를 포함하는 자가 증식된 텍스트의 유효성 검증 방법.

청구항 12

제10 항에 있어서, 상기 학습 모델 기반 판별 단계는

상기 자가 증식 텍스트의 각 단어를 임베딩하여 벡터화함으로써 다수의 단어 벡터를 획득하는 단계;

패턴 추정 방식이 미리 학습된 학습 모델을 이용하여 상기 다수의 단어 벡터의 특징을 누적하여 추출함으로써, 상기 문장 특징 벡터를 획득하는 단계; 및

미리 학습된 패턴 분류 방식에 따라 상기 문장 특징 벡터를 분류하여, 상기 자가 증식 텍스트의 비윤리를 판별하는 단계를 포함하는 자가 증식된 텍스트의 유효성 검증 방법.

청구항 13

제12 항에 있어서, 상기 학습 모델은

LSTM(Long Short Term Memory)으로 구현되는 자가 증식된 텍스트의 유효성 검증 방법.

청구항 14

제10 항에 있어서, 상기 최종 판별 단계는

상기 사전 기반 판별 단계, 상기 학습 모델 기반 판별 단계 및 상기 원문 기반 판별 단계 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과에 대해 다수결 원칙을 적용하여 상기 최종 판별 결과를 획득하는 자가 증식된 텍스트의 유효성 검증 방법.

청구항 15

제10 항에 있어서, 상기 최종 판별 단계는

상기 사전 기반 판별 단계, 상기 학습 모델 기반 판별 단계 및 상기 원문 기반 판별 단계 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과에 각각에 대해 기지정된 서로 다른 가중치를 할당하는 단계; 및

할당된 가중치에 따라 윤리 또는 비윤리 중 더 높은 가중치가 할당된 결과를 상기 최종 판별 결과로 획득하는 자가 증식된 텍스트의 유효성 검증 방법.

청구항 16

삭제

청구항 17

제10 항에 있어서, 상기 자가 증식된 텍스트의 유효성 검증 방법은

다수의 자가 증식 텍스트의 레이블에 대한 유효 판정 결과에 따라 자가 증식 텍스트의 신뢰도를 계산하는 신뢰도 계산 단계를 더 포함하는 자가 증식된 텍스트의 유효성 검증 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 텍스트 검증 장치 및 방법에 관한 것으로, 자가 증식된 비윤리 텍스트의 유효성 검증 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 현재 온라인 환경은 많은 사용자들에게 다양한 커뮤니케이션 수단을 제공하였으나, 온라인의 익명성으로 인하여 각종 비속어나 비윤리어가 빈번하게 사용되고 있다. 이에 온라인 서비스 업체들은 비속어나 비윤리어 등을 필터링하여 제거하기 위하여 노력하고 있으나, 비속어나 비윤리어 또한 다양한 형태로 변형되어 이용됨에 따라 필터링이 용이하지 않다는 한계가 있다.

[0003] 이에 최근에는 인공 신경망으로 구성되는 비윤리 텍스트 탐지 장치를 이용하여 비속어나 비윤리어를 검출하고자 하는 시도가 계속되어 왔다. 그러나 인공 신경망을 이용하기 위해서는 학습이 선행되어야 하며, 학습을 위해서는 대량의 학습용 텍스트가 필요하다. 여기서 학습용 데이터는 비속어나 비윤리어가 포함되어 있는지 여부가 사전에 검증되어 레이블된 텍스트이다.

[0004] 기존에는 텍스트를 사람이 직접 검증을 하여 학습용 텍스트로 사용하였으므로, 인공 신경망을 학습시키기에 충분한 양의 학습용 텍스트를 획득하기가 매우 어렵다는 한계가 있었다. 이러한 학습용 텍스트 획득의 어려움을 극복하기 위해 윤리 또는 비윤리가 미리 검증된 적은 양의 학습용 텍스트를 자가 증식 증식시키도록 미리 학습된 인공 신경망을 구현되는 비윤리 텍스트 자가 증식 장치를 이용하여 대량의 학습용 텍스트를 획득하는 방안이 제안되었다. 인공 신경망을 이용하여 학습용 텍스트를 자가 증식시키게 됨으로써, 적은 양의 학습용 텍스트로부터 대량의 학습용 텍스트를 용이하게 획득할 수 있으며, 다양한 변형 형태의 비속어나 비윤리어가 포함된 학습용 텍스트를 획득할 수 있게 되었다.

[0005] 다만, 자가 증식된 학습용 데이터 또한 정확하게 레이블 되었는지 판별될 필요가 있다. 만일 자가 증식된 학습용 데이터가 부정확하게 레이블 되면, 비윤리 텍스트 감지 장치의 학습이 부정확하게 수행되며, 이로 인해 비속어나 비윤리어가 포함된 텍스트를 제대로 필터링하지 못하게 되는 문제가 발생된다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 공개 특허 제10-2019-0108958호 (2019.09.25 공개)

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은 학습용으로 자가 증식된 비윤리 텍스트가 유효한지 여부를 판별할 수 있는 텍스트의 유효성 검증 장치 및 방법을 제공하는데 있다.

[0008] 본 발명의 다른 목적은 학습용 비윤리 텍스트를 생성하는 비윤리 텍스트 자가 증식 장치의 유효성을 검증할 수 있는 자가 증식된 텍스트의 유효성 검증 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 자가 증식된 텍스트의 유효성 검증 장치는 윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 이용하여 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득하는 텍스트 획득부; 자가 증식 텍스트를 인가받고, 인가된 자가 증식 텍스트에서 미리 획득된 비속어 사전에 등재된 비속어와 기기정된 레벨 이상으로 유사한 단어를 탐색하여 상기 자가 증식 텍스트의 비윤리를 판별하는 사전 기반 판별부; 자가 증식 텍스트를 인가받아 단어 단위로 벡터화하고, 벡터화된 단어로부터 미리 학습된 패턴 추정 방식에 따라 문장 특징 벡터를 추출하여 상기 자가 증식 텍스트의 비윤리를 판별하는 학습 모델 기반 판별부; 상기 자가 증식 텍스트와 가장 유사한 학습용 텍스트를 탐색하고, 탐색된 학습용 텍스트의 레이블에 따라 상기 자가 증식 텍스트의 비윤리를 판별하는 원문 기반 판별부; 및 상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과를 조합하여, 상기 자가 증식 텍스트에 대한 최종 판별 결과를 획득하는 판별 결과 비교부를 포함한다.

[0010] 상기 사전 기반 판별부는 상기 비속의 사전에 등재된 비속어와 상기 자가 증식 텍스트의 각 단어에 대해 N-그램 유사도 분석을 수행하여, 상기 자가 증식 텍스트에 비속어의 포함 여부를 판정하고, 비속어가 포함된 것으로 판정되면, 상기 자가 증식 텍스트를 비윤리로 판별할 수 있다.

[0011] 상기 학습 모델 기반 판별부는 상기 자가 증식 텍스트의 각 단어를 임베딩하여 벡터화함으로써 다수의 단어 벡

터를 획득하는 벡터 변환부; 미리 학습된 패턴 추정 방식에 따라 상기 다수의 단어 벡터의 특징을 누적하여 추출함으로써, 상기 문장 특징 벡터를 획득하는 문장 특징 추출부; 및 미리 학습된 패턴 분류 방식에 따라 상기 문장 특징 벡터를 분류하여, 상기 자가 증식 텍스트의 비윤리를 판별하는 특징 분류부를 포함할 수 있다.

[0012] 상기 문장 특징 추출부는 LSTM(Long Short Term Memory)으로 구현될 수 있다.

[0013] 상기 판별 결과 비교부는 상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과에 대해 다수결 원칙을 적용하여 상기 최종 판별 결과를 획득할 수 있다.

[0014] 상기 판별 결과 비교부는 상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과에 각각에 대해 기지정된 서로 다른 가중치를 할당하고, 할당된 가중치에 따라 윤리 또는 비윤리 중 더 높은 가중치가 할당된 결과를 상기 최종 판별 결과로 획득할 수 있다.

[0015] 상기 자가 증식된 텍스트의 유효성 검증 장치는 상기 자가 증식 텍스트의 생성 시에 윤리 또는 비윤리로 레이블링된 레이블과 상기 최종 판별 결과를 비교하여 동일하면 상기 자가 증식 텍스트의 레이블이 유효한 것으로 판정하고, 동일하지 않으면 유효하지 않은 것으로 판정하는 레이블 비교부를 더 포함할 수 있다.

[0016] 상기 레이블 비교부는 다수의 자가 증식 텍스트의 레이블에 대한 유효 판정 결과에 따라 자가 증식 텍스트의 신뢰도를 계산할 수 있다.

[0017] 상기 자가 증식된 텍스트의 유효성 검증 장치는 상기 텍스트 획득부에서 획득된 자가 증식 텍스트에 대해 부가 구성 요소 제거하고, 문장 단위로 구분하여 상기 사전 기반 판별부, 상기 학습 모델 기반 판별부 및 상기 원문 기반 판별부 각각으로 전달하는 전처리부를 더 포함할 수 있다.

[0018] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 자가 증식된 텍스트의 유효성 검증 방법은 윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 이용하여 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득하는 자가 증식 텍스트 획득 단계; 자가 증식 텍스트에서 미리 획득된 비속어 사전에 등재된 비속어와 기기정된 레벨 이상으로 유사한 단어를 탐색하여, 상기 자가 증식 텍스트의 비윤리를 판별하는 사전 기반 판별 단계; 자가 증식 텍스트를 인가받아 단어 단위로 벡터화하고, 벡터화된 단어로부터 패턴 추정 방식이 미리 학습된 학습 모델을 이용하여 문장 특징 벡터를 추출하고, 추출된 문장 특징에 기반하여 상기 자가 증식 텍스트의 비윤리를 판별하는 학습 모델 기반 판별 단계; 상기 자가 증식 텍스트와 가장 유사한 학습용 텍스트를 탐색하고, 탐색된 학습용 텍스트의 레이블에 따라 상기 자가 증식 텍스트의 비윤리를 판별하는 원문 기반 판별 단계; 및 상기 사전 기반 판별 단계, 상기 학습 모델 기반 판별 단계 및 상기 원문 기반 판별 단계 각각에서 판별된 상기 자가 증식 텍스트의 비윤리를 판별 결과를 조합하여, 상기 자가 증식 텍스트에 대한 최종 판별 결과를 획득하는 최종 판별 단계를 포함한다.

발명의 효과

[0019] 따라서, 본 발명의 실시예에 따른 자가 증식된 텍스트의 유효성 검증 장치 및 방법은 자가 증식 방법으로 생성되어 레이블링된 대량의 학습용 텍스트의 레이블을 검증함으로써, 자가 증식 방식으로 획득되는 학습용 텍스트의 유효성을 정확하게 검증할 수 있다. 그러므로 인공 신경망으로 구현되어 비속어 또는 비윤리어를 탐지하는 탐지 장치를 학습시키기 위한 학습용 텍스트의 신뢰성을 크게 높일 수 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일 실시예에 따른 자가 증식된 텍스트 유효성 검증 장치의 개략적 구조를 나타낸다.

도 2는 도 1의 학습 모델 기반 판별부의 상세 구성을 나타낸다.

도 3은 본 발명의 일 실시예에 따른 자가 증식된 텍스트 유효성 검증 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0022] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러

나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0023] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0024] 도 1은 본 발명의 일 실시예에 따른 자가 증식된 텍스트 유효성 검증 장치의 개략적 구조를 나타내고, 도 2는 도 1의 학습 모델 기반 판별부의 상세 구성을 나타낸다.

[0025] 도 1을 참조하면, 본 실시예에 따른 자가 증식된 텍스트 유효성 검증 장치는 텍스트 획득부(100), 전처리부(200), 사전 기반 판별부(300), 학습 모델 기반 판별부(400), 원문 기반 판별부(500), 판별 결과 비교부(600) 및 레이블 비교부(700)를 포함할 수 있다.

[0026] 텍스트 획득부(100)는 윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 기반으로 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득한다. 여기서 자가 증식 텍스트는 적은 수의 학습용 텍스트를 이용하여 대량의 학습용 데이터를 생성하기 위해 미리 학습된 자가 증식 장치가 학습용 텍스트를 인가받아 생성한 다수의 텍스트로서, 입력된 학습용 텍스트와 마찬가지로 윤리 또는 비윤리가 레이블된 텍스트이다. 이때 텍스트 내에 다수의 문장이 포함된 경우, 각 문장 단위로 윤리 또는 비윤리가 레이블될 수 있으며, 자가 증식 장치는 학습된 방식에 따라 비윤리로 레이블된 학습용 텍스트로부터 윤리로 레이블된 자가 증식 텍스트를 생성하거나 윤리로 레이블된 학습용 텍스트로부터 비윤리로 레이블된 자가 증식 텍스트를 생성할 수도 있다.

[0027] 즉 텍스트 획득부(100)는 자가 증식 장치가 생성한 다수의 자가 증식 텍스트를 획득하며, 자가 증식 장치가 생성한 다수의 자가 증식 텍스트를 저장하는 저장 장치 또는 데이터 베이스 등으로 구현될 수 있다.

[0028] 또한 텍스트 획득부(100)는 자가 증식 텍스트와 함께 자가 증식 텍스트를 생성하기 위해 이용된 원문 학습용 텍스트를 함께 저장할 수 있다.

[0029] 전처리부(200)는 텍스트 획득부(100)에서 획득된 자가 증식 텍스트를 인가받아 기지정된 전처리 작업을 수행한다. 이때 전처리부(200)는 자가 증식 텍스트에서 레이블을 함께 인가받도록 구성될 수도 있으나, 레이블을 제외한 텍스트만을 인가받도록 구성될 수도 있다.

[0030] 전처리부(200)는 자가 증식 텍스트 내에서 문자, 공백, 구두점 등과 같이 문장을 구성하는 문장 구성 요소 이외에 나머지 구성 요소인 특수 문자, URL, SNS 지정 특성 문자(# 해쉬태그, @ 언급)등의 부가 구성 요소 모두 제거한다. 이는 문자와 문장 기호 및 공백과 같이 문장을 구성하는 문장 구성 요소 이외의 부가 구성 요소들은 비속어나 비윤리어로 이용될 가능성이 거의 없으므로 탐지 대상에서 배제하기 위해서이다.

[0031] 다만 자가 증식 장치에 의해 생성된 자가 증식 텍스트에서는 부가 구성 요소가 포함되지 않도록 생성될 수 있으며, 이 경우 부가 구성 요소 모두 제거하는 과정은 생략될 수 있다.

[0032] 그리고 전처리부(200)는 부가 구성 요소가 제거된 텍스트 내에 포함된 각 문장을 구분하여, 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500) 각각으로 전달한다.

[0033] 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500)는 각각 서로 다른 지정된 방식으로 자가 증식 텍스트의 윤리 또는 비윤리를 판별한다.

[0034] 우선 사전 기반 판별부(300)는 비속어를 포함하는 문장은 혐오 문장 또는 비윤리적 문장일 가능성이 크다는 점을 고려하여, 전처리부(200)에서 인가되는 문장에서 비속어의 포함 여부를 분석하여 윤리 또는 비윤리 여부를 판별한다.

[0035] 사전 기반 판별부(300)는 일예로 비속어 사전을 이용하여 문장 내의 비속어 포함 여부를 분석할 수 있다. 여기서 비속어 사전은 이미 공개된 비속어 사전을 이용하거나, 미리 작성되어 획득될 수 있으며, 경우에 따라서는 원본 학습용 텍스트 또는 자가 증식 텍스트로부터 미리 학습된 방식에 따라 비속어를 분류하여 비속어 사전을 생성하여 이용할 수 있다. 비속어 사전은 이미 작성되어 공개되어 있으며, 비속어 사전을 생성하는 방식 또한 공지된 기술이므로 여기서는 상세하게 설명하지 않는다.

[0036] 이때 사전 기반 판별부(300)는 다양하게 변형되는 비속어에 대응할 수 있도록 완전히 일치하는 비속어만을 탐색

하는 것이 아니라 비속의 사전에 등재된 비속어와 인가된 문장의 각 단어에 대해 N-그램(N-gram) 유사도 분석을 수행하여, 각 문장에 비속어의 포함 여부를 판정할 수 있다. 일예로 사전 기반 판별부(300)는 비속어 사전에 등재된 각 단어와 인가된 문장에 포함된 단어들을 비교하여 매칭 문자의 수를 기반으로 대응 여부를 판정하고, 판정 결과에 따라 해당 문장이 비윤리 문장인지 여부를 판별할 수 있다.

[0037] 학습 모델 기반 판별부(400)는 인가되는 문장의 각 단어를 임베딩하여 벡터화하고, 벡터화된 단어를 인가받아 문장 특징을 추출하고, 추출된 문장 특징 벡터를 기반으로 해당 문장이 비윤리 문장인지 여부를 판별한다.

[0038] 도 2를 참조하면, 학습 모델 기반 판별부(400)는 벡터 변환부(410)와 문장 특징 추출부(420) 및 특징 분류부(430)를 포함할 수 있다.

[0039] 벡터 변환부(410)는 인가되는 문장에 포함된 단어 각각을 임베딩하여 벡터화함으로써 다수의 단어 벡터를 획득한다. 벡터 변환부(410)는 미리 학습된 임베딩 모델을 이용하여 문장 내의 단어 각각을 단어 벡터로 변환할 수 있다. 벡터 변환부(410)는 Word2Vec, fastText 등과 같이 단어를 단어 벡터로 변환하도록 공개된 임베딩 모델을 이용하여 단어를 단어 벡터로 변환할 수 있다.

[0040] 문장 특징 추출부(420)는 벡터 변환부(410)로부터 단어 벡터를 인가받고 인가되는 단어 벡터의 특징을 누적하여 추출함으로써, 문장 특징 벡터를 획득한다.

[0041] 문장 특징 추출부(420)는 패턴 추정 방식이 미리 학습된 인공 신경망으로 구현될 수 있으며, 특히 LSTM(Long Short Term Memory)으로 구현될 수 있다. LSTM은 순환 신경망(Recurrent Neural Network: RNN)이 장기간(Long Term) 특징을 반영할 수 있도록 개선된 구조를 갖는 신경망으로서, 이전 추출된 단어 벡터의 특징이 이후 입력되는 단어 벡터에 누적 반영됨으로써 문장 특징을 획득하기 용이하다는 장점이 있다.

[0042] 그리고 특징 분류부(430)는 문장 특징 추출부(420)에서 획득된 문장 특징 벡터를 인가받고, 미리 학습된 패턴 분류 방식에 따라 문장 특징 벡터를 분류하여, 윤리 또는 비윤리를 판별한다. 특징 분류부(430)는 인공 신경망의 완전 연결 레이어(Fully Connected layer)로 구현되어 문장 특징 벡터를 이진 분류함으로써, 윤리 또는 비윤리를 판별할 수 있다.

[0043] 한편, 원문 기반 판별부(500)는 전처리부(200)로부터 문장을 인가받고, 인가된 문장을 자가 증식 장치에서 자가 증식 텍스트를 생성하기 위해 이용된 원문 학습 텍스트와 비교하여 가장 유사한 원문 학습 텍스트를 탐색한다. 그리고 탐색된 원문 학습 텍스트의 레이블에 따라 문장을 윤리 또는 비윤리로 판별한다. 여기서 원문 기반 판별부(500) 또한 N-그램 유사도 분석을 수행하여, 가장 유사한 원문 학습 텍스트를 판별할 수 있다.

[0044] 판별 결과 비교부(600)는 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500) 각각이 자가 증식 텍스트에 대해 비윤리 여부를 판별한 결과에 기초하여 인가된 자가 증식 텍스트의 윤리 또는 비윤리를 최종 판별한다.

[0045] 여기서 판별 결과 비교부(600)는 단순히 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500) 각각의 판별 결과를 기초로 다수결 원칙에 따라 자가 증식 텍스트의 윤리 또는 비윤리를 판별할 수 있다.

[0046] 그러나 경우에 따라서 판별 결과 비교부(600)는 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500) 각각의 판별 결과에 서로 다른 가중치를 가중하여 자가 증식 텍스트의 윤리 또는 비윤리를 판별할 수도 있다. 즉 미리 설정되는 판별 결과의 중요도에 따라 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500) 각각의 판별 결과에 서로 다른 가중치를 가중할 수 있다.

[0047] 일예로 사전 기반 판별부(300)의 경우, 이미 검증된 비속어 사전을 기반으로 하여 자가 증식 텍스트의 윤리 또는 비윤리를 판별하므로, 학습 모델 기반 판별부(400)나 원문 기반 판별부(500)에 비해 더 높은 가중치를 가중한 후, 가중치가 가중된 사전 기반 판별부(300), 학습 모델 기반 판별부(400) 및 원문 기반 판별부(500)의 판별 결과에서 윤리 또는 비윤리 중 더 높은 가중치가 부여된 결과를 선택할 수 있다.

[0048] 한편, 레이블 비교부(700)는 판별 결과 비교부(600)에서 최종 판별된 결과와 텍스트 획득부(100)에 저장된 대응하는 자가 증식 텍스트의 레이블을 비교하여 동일하면, 자가 증식 텍스트의 레이블이 유효한 것으로 판정하고, 동일하지 않으면 유효하지 않은 것으로 판정한다.

[0049] 레이블 비교부(700)는 다수의 자가 증식 텍스트의 레이블에 대한 판정 결과를 누적하여 자가 증식 장치에서 생성된 자가 증식 텍스트의 신뢰도를 계산할 수 있다. 일예로 전체 자가 증식 텍스트에서 유효한 것으로 판정된

자가 증식 텍스트의 비율로 자가 증식 텍스트의 신뢰도를 계산할 수 있다.

[0050] 도 3은 본 발명의 일 실시예에 따른 자가 증식된 텍스트 유효성 검증 방법을 나타낸다.

[0051] 도 1 및 도 2를 참조하여, 도 3의 자가 증식된 텍스트 유효성 검증 방법을 설명하면, 우선 윤리 또는 비윤리가 미리 검증되어 레이블된 학습용 텍스트를 기반으로 자가 증식 방식으로 생성된 다수의 자가 증식 텍스트를 획득한다(S11). 여기서 레이블은 자가 증식 텍스트 내의 문장 단위로 레이블링 될 수 있다.

[0052] 그리고 획득된 자가 증식 텍스트에 대해 부가 구성 요소 제거하고, 문장 단위로 구분하는 등의 기지정된 전처리 작업을 수행한다(S12).

[0053] 자가 증식 텍스트가 전처리되면, 전처리된 자가 증식 텍스트에 대해 서로 다른 지정된 방식으로 자가 증식 텍스트의 윤리 또는 비윤리를 판별한다.

[0054] 우선 비속어 사전을 이용하여 윤리 또는 비윤리를 판별한다. 비속어 사전을 이용하는 경우, 먼저 비속어 사전에 기재된 비속어와 기지정된 레벨 이상으로 유사 단어를 자가 증식 텍스트의 각 문장에서 탐색한다(S13). 그리고 탐색 결과에 기반하여 문장의 윤리 또는 비윤리를 판별한다(S14). 즉 비속어와 유사한 것으로 판별되는 단어가 탐색되면 비윤리로 판별하고, 탐색되지 않으면 윤리로 판별할 수 있다.

[0055] 한편, 미리 학습된 학습 모델에 기반하여 자가 증식 텍스트의 윤리 또는 비윤리를 판별한다.

[0056] 이를 위해 우선 자가 증식 텍스트의 각 문장에 포함된 단어를 미리 학습된 임베딩 모델을 이용하여 벡터화함으로써 다수의 단어 벡터를 획득한다(S15). 그리고 획득된 다수의 단어 벡터를 패턴 추정 방식이 미리 학습된 인공 신경망으로 입력하여 문장에 대한 특징을 나타내는 문장 특징 벡터를 획득한다(S16). 여기서 인공 신경망은 LSTM으로 구현될 수 있다.

[0057] 문장 특징 벡터가 획득되면, 미리 학습된 패턴 분류 방식에 따라 문장 특징 벡터를 분류하여, 윤리 또는 비윤리를 판별한다(S17).

[0058] 또한 자가 증식 텍스트를 생성하기 위해 이용된 원문 학습 텍스트를 이용하여 자가 증식 텍스트의 윤리 또는 비윤리를 판별한다.

[0059] 즉 원문 학습 텍스트 중 인가된 자가 증식 텍스트와 가장 유사한 원문 학습 텍스트를 탐색한다(S18). 가장 유사한 원문 학습 텍스트가 탐색되면, 탐색된 원문 학습 텍스트의 레이블에 따라 문장을 윤리 또는 비윤리로 판별한다(S19).

[0060] 이후, 비속어 사전을 이용한 판별 결과와 학습 모델에 기반한 판별 결과 및 원문 학습 텍스트를 이용한 판별 결과를 기지정된 방식으로 조합하여 자가 증식 텍스트에 대한 윤리 또는 비윤리의 최종 판별 결과를 획득한다(S20). 여기서 최종 판별 결과는 다수결의 원칙에 따라 판별하거나, 각 판별 결과에 대해 기지정된 가중치를 할당하여 윤리 또는 비윤리 중 높은 가중치가 부가된 쪽을 최종 판별 결과로 획득할 수 있다.

[0061] 최종 판별 결과가 획득되면, 획득된 최종 판별 결과와 자가 증식 텍스트의 레이블을 비교하여, 자가 증식 텍스트의 유효성을 판정한다(S21). 그리고 다수의 자가 증식 텍스트에 대한 유효성 판정 결과를 누적하여, 자가 증식 방식으로 생성된 자가 증식 텍스트의 신뢰도를 계산한다(S22).

[0062] 본 발명에 따른 방법은 컴퓨터에서 실행시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0063] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0064] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0065] 100: 텍스트 획득부 200: 전처리부

300: 사전 기반 판별부

400: 학습 모델 기반 판별부

410: 벡터 변환부

420: 문장 특징 추출부

430: 특징 분류부

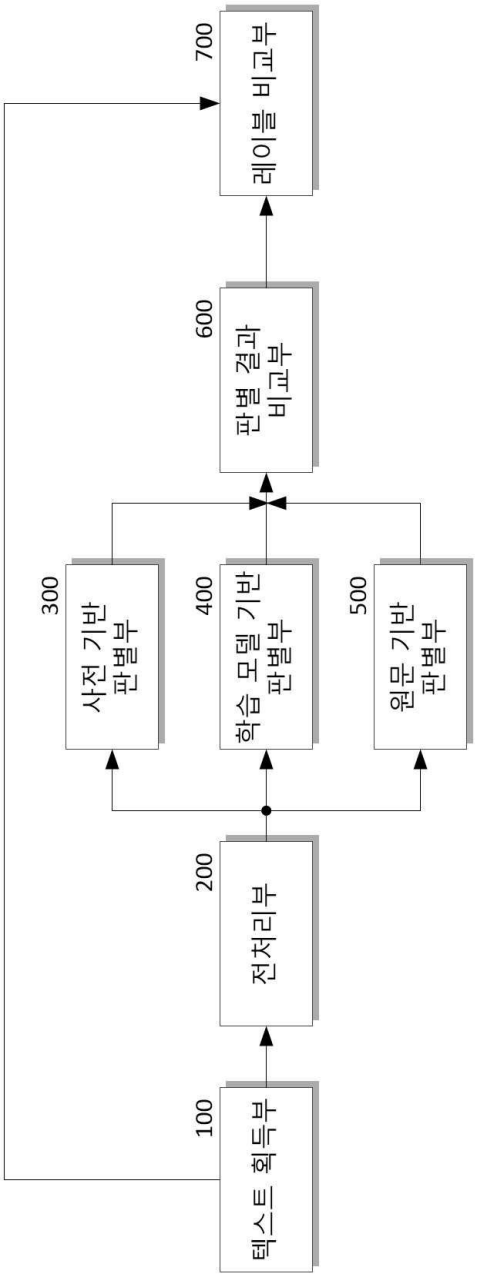
500: 원문 기반 판별부

600: 판별 결과 비교부

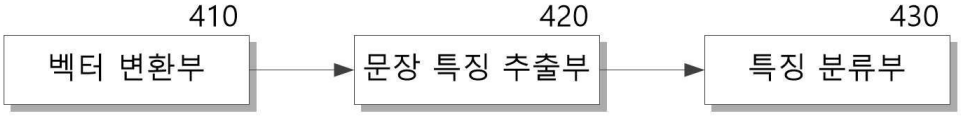
700: 레이블 비교부

도면

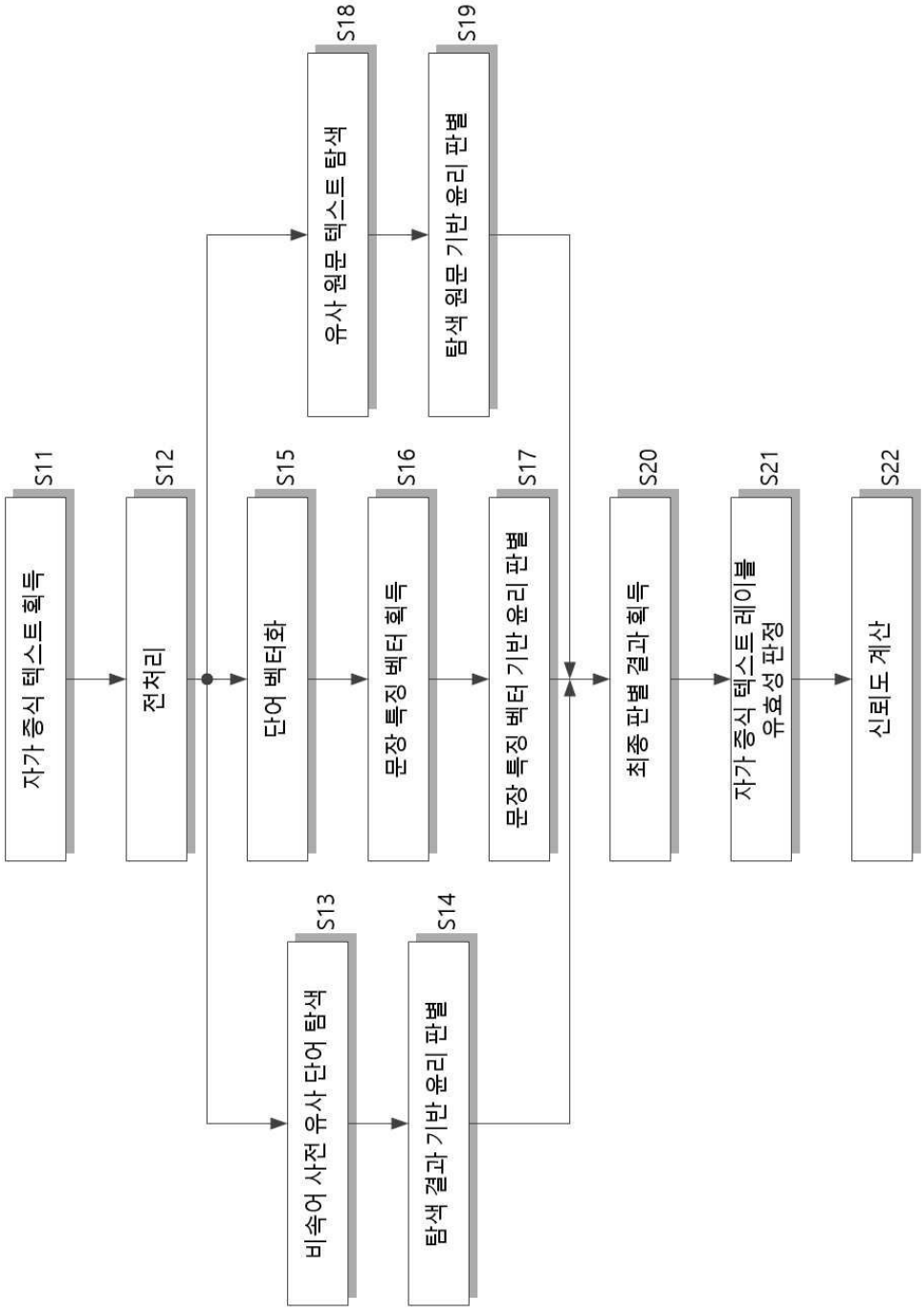
도면1



도면2



도면3


S11
자가 증식 텍스트 획득
S12
전처리
S15
단어 벡터화
S16
문장 특징 벡터 획득
S17
문장 특징 벡터 기반 윤리 판별
S18
유사 원문 텍스트 탐색
S19
탐색 원문 기반 윤리 판별
S13
비속어 사전 유사 단어 탐색
S14
탐색 결과 기반 윤리 판별
S20
최종 판별 결과 획득
S21
자가 증식 텍스트 레이블
유효성 판정
S22
신뢰도 계산