(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0141419
(43) 공개일자 2024년09월27일

(51) 국제특허분류(Int. Cl.)
G06F 40/30 (2020.01) *G06F 16/35* (2019.01)
G06F 18/241 (2023.01) *G06F 40/205* (2020.01)
G06N 20/10 (2019.01)
(52) CPC특허분류
G06F 40/30 (2020.01)
G06F 16/35 (2019.01)
(21) 출원번호 10-2023-0035743
(22) 출원일자 2023년03월20일
심사청구일자 2023년03월20일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
김미숙
서울특별시 동작구 동작대로29길 91, 207동 508호
이재훈
대구광역시 수성구 들안로 360, 104동 2504호
(74) 대리인
김준석, 박민욱

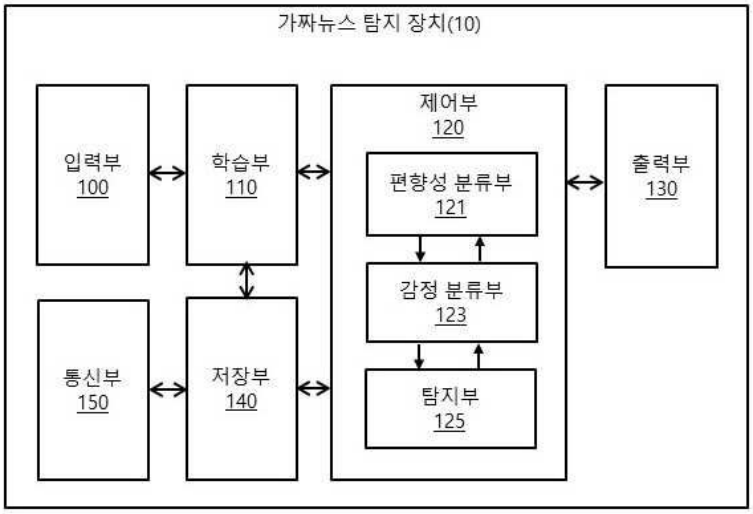
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **가짜뉴스 탐지 장치 및 그 방법**

(57) 요 약

본 발명은 가짜뉴스 탐지 장치 및 가짜뉴스 탐지 방법에 관한 것으로, 가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는 단계; 미리 학습된 편향성 분류 모델을 이용하여 평가 문서의 문서 편향성 점수를 산출하는 단계; 미리 학습된 감정 분류 모델을 이용하여 평가 문서의 문서 감정 점수를 산출하는 단계; 및 문서 편향성 점수 및 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단하는 단계;를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 18/241 (2023.01)
G06F 40/205 (2020.01)
G06N 20/10 (2021.08)

이 발명을 지원한 국가연구개발사업
과제고유번호 1415180329
과제번호 20018758
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 지식서비스산업기술개발
연구과제명 중소 중견 제조산업에서의 KESG 기반 경영성과관리와 생산 프로세스 최적화 서비스 기술개발
기 여 율 50/100
과제수행기관명 (주)비아이매트릭스
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711152617
과제번호 2021-0-00469-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 SW컴퓨팅산업원천기술개발
연구과제명 이종 융합 데이터 탐지 및 추적 기술 개발
기 여 율 50/100
과제수행기관명 (주)피씨엔
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는 단계;

미리 학습된 편향성 분류 모델을 이용하여 상기 평가 문서의 문서 편향성 점수를 산출하는 단계;

미리 학습된 감정 분류 모델을 이용하여 상기 평가 문서의 문서 감정 점수를 산출하는 단계; 및

상기 문서 편향성 점수 및 상기 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단하는 단계;를 포함하는 가짜뉴스 탐지 방법.

청구항 2

제1항에 있어서,

상기 문서 편향성 점수를 산출하는 단계는,

상기 평가 문서를 복수의 문장으로 분할하는 단계;

상기 복수의 문장을 상기 미리 학습된 편향성 분류 모델에 입력하여 개별 문장의 편향성 수치를 산출하는 단계; 및

상기 개별 문장의 편향성 수치에 기초하여 상기 문서 편향성 점수를 산출하는 단계;를 포함하는 가짜뉴스 탐지 방법.

청구항 3

제2항에 있어서,

상기 문서 편향성 점수를 산출하는 단계는,

상기 개별 문장의 편향성 수치에 기초하여 상기 평가 문서의 편향성 분포를 도출하는 단계; 및

상기 평가 문서의 편향성 분포에 기초하여 상기 개별 문장의 편향성 수치를 보정하는 단계;를 더 포함하는 가짜뉴스 탐지 방법.

청구항 4

제3항에 있어서,

상기 개별 문장의 편향성 수치를 보정하는 단계는,

상기 평가 문서의 편향성 분포를 기반으로 편향된 문장이 연속적으로 존재하는 경우에 상기 연속적으로 편향된 문장 각각에 대한 상기 편향성 수치에 가중치를 부여하여 가짜뉴스 탐지 방법.

청구항 5

제2항에 있어서,

상기 문서 감정 점수를 산출하는 단계는,

상기 평가 문서를 복수의 문장으로 분할하는 단계;

상기 복수의 문장을 상기 미리 학습된 감정 분류 모델에 입력하여 문장 감정 점수를 산출하는 단계; 및

상기 문장 감정 점수를 기반으로 상기 문서 감정 점수를 산출하는 단계;를 포함하는 가짜뉴스 탐지 방법.

청구항 6

제5항에 있어서,

상기 문서 감정 점수를 산출하는 단계는,

상기 문장 감정 점수에 기초하여 상기 평가 문서의 감정 분포를 도출하는 단계; 및

상기 평가 문서의 감정 분포에 기초하여 상기 문장 감정 점수를 보정하는 단계;를 더 포함하는 가짜뉴스 탐지 방법.

청구항 7

제5항에 있어서,

상기 문서 편향성 점수를 산출하는 단계는,

상기 문장 감정 점수가 미리 설정된 기준 이상인 경우 상기 개별 문장의 편향성 수치를 보정하는 단계;를 포함하는 가짜뉴스 탐지 방법.

청구항 8

제1항에 있어서,

상기 가짜뉴스 여부를 판단하는 단계는,

Support Vector Machine(SVM)를 기반으로 학습된 상기 가짜뉴스 탐지 모델을 통해 상기 평가 문서가 가짜뉴스인지 진짜뉴스인지 판단하는 가짜뉴스 탐지 방법.

청구항 9

제1항에 있어서,

증강된 학습데이터에 기초하여 상기 편향성 분류 모델을 학습시키는 단계;

상기 증강된 학습데이터에 기초하여 상기 감정 분류 모델을 학습시키는 단계; 및

상기 문서 편향성 점수 및 상기 문서 감정 점수에 기초하여 상기 가짜뉴스 탐지 모델을 학습시키는 단계;을 더 포함하는 가짜뉴스 탐지 방법.

청구항 10

가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는 입력부,

미리 학습된 편향성 분류 모델을 이용하여 상기 평가 문서의 문서 편향성 점수를 산출하는 편향성 분류부,

미리 학습된 감정 분류 모델을 이용하여 상기 평가 문서의 문서 감정 점수를 산출하는 감정 분류부, 및

상기 문서 편향성 점수 및 상기 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단하는 탐지부를 포함하는 가짜뉴스 탐지 장치.

청구항 11

제10항에 있어서,

상기 편향성 분류부는,

상기 평가 문서를 복수의 문장으로 분할하고, 상기 복수의 문장을 상기 미리 학습된 편향성 분류 모델에 입력하여 개별 문장의 편향성 수치를 산출하며, 상기 개별 문장의 편향성 수치에 기초하여 상기 문서 편향성 점수를 산출하는 가짜뉴스 탐지 장치.

청구항 12

제11항에 있어서,

상기 편향성 분류부는,

상기 개별 문장의 편향성 수치에 기초하여 상기 평가 문서의 편향성 분포를 도출하고, 상기 평가 문서의 편향성

분포에 기초하여 상기 개별 문장의 편향성 수치를 보정하는 가짜뉴스 탐지 장치.

청구항 13

제12항에 있어서,

상기 편향성 분류부는,

상기 평가 문서의 편향성 분포를 기반으로 편향된 문장이 연속적으로 존재하는 경우에 상기 연속적으로 편향된 문장 각각에 대한 상기 편향성 수치에 가중치를 부여하여 상기 편향성 수치를 보정하는 가짜뉴스 탐지 장치.

청구항 14

제11항에 있어서,

상기 감정 분류부는,

상기 평가 문서를 복수의 문장으로 분할하고, 상기 복수의 문장을 상기 미리 학습된 감정 분류 모델에 입력하여 문장 감정 점수를 산출하고, 상기 문장 감정 점수를 기반으로 상기 문서 감정 점수를 산출하는 가짜뉴스 탐지 장치.

청구항 15

제14항에 있어서,

상기 감정 분류부는,

상기 문장 감정 점수에 기초하여 상기 평가 문서의 감정 분포를 도출하고, 상기 평가 문서의 감정 분포에 기초하여 상기 문장 감정 점수를 보정하는 가짜뉴스 탐지 장치.

청구항 16

제14항에 있어서,

상기 편향성 분류부는,

상기 문장 감정 점수가 미리 설정된 기준 이상인 경우에 상기 개별 문장의 편향성 수치를 보정하는 가짜뉴스 탐지 장치.

청구항 17

제10항에 있어서,

상기 탐지부는,

Support Vector Machine(SVM)를 기반으로 학습된 상기 가짜뉴스 탐지 모델을 통해 상기 평가 문서가 가짜뉴스인지 진짜뉴스인지 판단하는 가짜뉴스 탐지 장치.

청구항 18

제10항에 있어서,

증강된 학습데이터에 기초하여 상기 편향성 분류 모델을 학습시키고, 상기 증강된 학습데이터에 기초하여 상기 감정 분류 모델을 학습시키며, 상기 문서 편향성 점수 및 상기 문서 감정 점수에 기초하여 상기 가짜뉴스 탐지 모델을 학습시키는 학습부를 더 포함하는 가짜뉴스 탐지 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 가짜뉴스 탐지 장치 및 그 방법에 대한 것이다.

배 경 기 술

[0002] 일반적으로, 뉴스는 시사성이 있다고 판단되어 각종 대중매체를 통해 보도되어지는 내용을 의미하지만, 보다 넓게는 대중에게 아직 알려져 있지 않아 새롭게 전달되는 소식들을 총칭한다.

[0003] 뉴스는 정보의 공유와 확산에 매우 중요한 역할을 하는 매개체 중 하나로, 일반적으로, 매스 미디어라 일컬어지는 TV, 신문, 라디오, 잡지, 영화, 광고 등의 매체들을 통해 정보가 확산되지만, 최근에는 뉴스를 생성하는 주체가 블로그, SNS 등을 활용한 개인이 될 수 있다.

[0004] 대중은 다양한 매체를 통한 여러 분야의 수많은 뉴스들 중에서도 각자 자신이 관심이 있는 것들을 선택하여 접한다. 대중은 뉴스로 새로이 전해지는 소식에서 정보를 얻음으로 전에는 전혀 몰랐던 사실을 알게 되거나 이미 알고 있던 사실이 잘못되었다는 것을 알게 되기도 한다. 이를 통해 개인이 가지고 있던 견해에 변화가 생길 가능성이 있는데 수동적 입장에서는 주체적 입장이 전달하는 뉴스 내용을 대부분 진실한 정보로 수용하게 되기 때문이다.

[0005] 최근, 가짜 정보는 포함하는 뉴스가 대중들에게 전달되는 문제가 발생하였고, 이러한 가짜 정보들은 뉴스 뿐만 아니라, 비디오, 오디오, 이미지, 소셜 네트워크 서비스, 연설, 문서 등과 같이 다양한 형태로 발현되어 대중에게 전달될 수 있다.

[0006] 이러한 가짜뉴스들은 거짓된 내용을 내포하여 마치 진실된 사실인 듯이 대중에게 전달될 수 있다는 문제가 있다.

[0007] 그리고, 이러한 가짜 정보를 통한 새로운 정보의 의심과 함께 기존 얻었던 정보에 대한 신뢰 또한 잃게 하여 정보의 활발한 공유와 확산에 있어 문제를 일으키고 있다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 등록특허공보 제10-2340542 호("자동화 기반의 가짜뉴스 탐지 장치 및 방법", 공개일자: 2021.05.25)

발명의 내용

해결하려는 과제

[0009] 본 발명은 문장의 편향성과 문장의 감정분류 결과를 이용하여 가짜뉴스를 탐지하는 가짜뉴스 탐지 장치 및 가짜뉴스 탐지 방법을 제공하는 것을 해결하고자 하는 과제로 한다.

과제의 해결 수단

[0010] 상술한 과제를 해결하기 위하여 가짜뉴스 탐지 장치 및 가짜뉴스 탐지 방법이 제공된다.

[0011] 본 발명의 일 실시예에 따른 가짜뉴스 탐지 방법은 가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는 단계; 미리 학습된 편향성 분류 모델을 이용하여 상기 평가 문서의 문서 편향성 점수를 산출하는 단계; 미리 학습된 감정 분류 모델을 이용하여 상기 평가 문서의 문서 감정 점수를 산출하는 단계; 및 상기 문서 편향성 점수 및 상기 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단하는 단계;를 포함한다.

[0012] 일 실시예에서, 상기 문서 편향성 점수를 산출하는 단계는, 상기 평가 문서를 복수의 문장으로 분할하는 단계; 상기 복수의 문장을 상기 미리 학습된 편향성 분류 모델에 입력하여 개별 문장의 편향성 수치를 산출하는 단계; 및 상기 개별 문장의 편향성 수치에 기초하여 상기 문서 편향성 점수를 산출하는 단계;를 포함할 수 있다.

[0013] 일 실시예에서, 상기 문서 편향성 점수를 산출하는 단계는, 상기 개별 문장의 편향성 수치에 기초하여 상기 평가 문서의 편향성 분포를 도출하는 단계; 및 상기 평가 문서의 편향성 분포에 기초하여 상기 개별 문장의 편향성 수치를 보정하는 단계;를 더 포함할 수 있다.

[0014] 일 실시예에서, 상기 개별 문장의 편향성 수치를 보정하는 단계는, 상기 평가 문서의 편향성 분포를 기반으로 편향된 문장이 연속적으로 존재하는 경우에 상기 연속적으로 편향된 문장 각각에 대한 상기 편향성 수치에 가중치를 부여할 수 있다.

[0015] 일 실시예에서, 상기 문서 감정 점수를 산출하는 단계는, 상기 평가 문서를 복수의 문장으로 분할하는 단계; 상기 복수의 문장을 상기 미리 학습된 감정 분류 모델에 입력하여 문장 감정 점수를 산출하는 단계; 및 상기 문장 감정 점수를 기반으로 상기 문서 감정 점수를 산출하는 단계;를 포함할 수 있다.

[0016] 일 실시예에서, 상기 문서 감정 점수를 산출하는 단계는, 상기 문장 감정 점수에 기초하여 상기 평가 문서의 감정 분포를 도출하는 단계; 및 상기 평가 문서의 감정 분포에 기초하여 상기 문장 감정 점수를 보정하는 단계;를 더 포함할 수 있다.

[0017] 일 실시예에서, 상기 문서 편향성 점수를 산출하는 단계는, 상기 문장 감정 점수가 미리 설정된 기준 이상인 경우 상기 개별 문장의 편향성 수치를 보정하는 단계;를 포함할 수 있다.

[0018] 일 실시예에서, 상기 가짜뉴스 여부를 판단하는 단계는, Support Vector Machine(SVM)를 기반으로 학습된 상기 가짜뉴스 탐지 모델을 통해 상기 평가 문서가 가짜뉴스인지 진짜뉴스인지 판단할 수 있다.

[0019] 일 실시예에서, 증강된 학습데이터에 기초하여 상기 편향성 분류 모델을 학습시키는 단계; 상기 증강된 학습데이터에 기초하여 상기 감정 분류 모델을 학습시키는 단계; 및 상기 문서 편향성 점수 및 상기 문서 감정 점수에 기초하여 상기 가짜뉴스 탐지 모델을 학습시키는 단계;을 더 포함할 수 있다.

[0020] 본 발명의 다른 실시예에 따른 가짜뉴스 탐지 장치는, 가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는 입력부, 미리 학습된 편향성 분류 모델을 이용하여 상기 평가 문서의 문서 편향성 점수를 산출하는 편향성 분류부, 미리 학습된 감정 분류 모델을 이용하여 상기 평가 문서의 문서 감정 점수를 산출하는 감정 분류부, 및 상기 문서 편향성 점수 및 상기 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단하는 탐지부를 포함한다.

[0021] 일 실시예에서, 상기 편향성 분류부는, 상기 평가 문서를 복수의 문장으로 분할하고, 상기 복수의 문장을 상기 미리 학습된 편향성 분류 모델에 입력하여 개별 문장의 편향성 수치를 산출하며, 상기 개별 문장의 편향성 수치에 기초하여 상기 문서 편향성 점수를 산출할 수 있다.

[0022] 일 실시예에서, 상기 편향성 분류부는, 상기 개별 문장의 편향성 수치에 기초하여 상기 평가 문서의 편향성 분포를 도출하고, 상기 평가 문서의 편향성 분포에 기초하여 상기 개별 문장의 편향성 수치를 보정할 수 있다.

[0023] 일 실시예에서, 상기 편향성 분류부는, 상기 평가 문서의 편향성 분포를 기반으로 편향된 문장이 연속적으로 존재하는 경우에 상기 연속적으로 편향된 문장 각각에 대한 상기 편향성 수치에 가중치를 부여하여 상기 편향성 수치를 보정할 수 있다.

[0024] 일 실시예에서, 상기 감정 분류부는, 상기 평가 문서를 복수의 문장으로 분할하고, 상기 복수의 문장을 상기 미리 학습된 감정 분류 모델에 입력하여 문장 감정 점수를 산출하고, 상기 문장 감정 점수를 기반으로 상기 문서 감정 점수를 산출할 수 있다.

[0025] 일 실시예에서, 상기 감정 분류부는, 상기 문장 감정 점수에 기초하여 상기 평가 문서의 감정 분포를 도출하고, 상기 평가 문서의 감정 분포에 기초하여 상기 문장 감정 점수를 보정할 수 있다.

[0026] 일 실시예에서, 상기 편향성 분류부는, 상기 문장 감정 점수가 미리 설정된 기준 이상인 경우에 상기 개별 문장의 편향성 수치를 보정할 수 있다.

[0027] 일 실시예에서, 상기 탐지부는, Support Vector Machine(SVM)를 기반으로 학습된 상기 가짜뉴스 탐지 모델을 통해 상기 평가 문서가 가짜뉴스인지 진짜뉴스인지 판단할 수 있다.

[0028] 일 실시예에서, 증강된 학습데이터에 기초하여 상기 편향성 분류 모델을 학습시키고, 상기 증강된 학습데이터에 기초하여 상기 감정 분류 모델을 학습시키며, 상기 문서 편향성 점수 및 상기 문서 감정 점수에 기초하여 상기 가짜뉴스 탐지 모델을 학습시키는 학습부를 더 포함할 수 있다.

발명의 효과

[0029] 상술한 가짜뉴스 탐지 장치 및 가짜뉴스 탐지 방법에 의하면, 문서 내의 문장의 편향성 및 문장의 감정을 기반

으로 가짜뉴스를 탐지할 수 있다.

도면의 간단한 설명

[0030] 도 1은 본 발명의 일 실시예에 따른 가짜뉴스 탐지 장치의 구성을 설명하기 위한 블록도이다.

도 2는 ISOT Fake News Dataset의 구성을 설명하기 위한 도면이다.

도 3은 WIKIBIAS-manual을 설명하기 위한 도면이다.

도 4는 문서 편향성 점수를 산출하는 과정을 설명하기 위한 도면이다.

도 5는 문서 편향성 점수를 설명하기 위한 도면이다.

도 6은 문서 감정 점수를 산출하는 과정을 설명하기 위한 도면이다.

도 7은 가짜뉴스 여부를 판단하는 과정을 설명하기 위한 도면이다.

도 8은 본 발명의 일 실시예에 따른 복수의 분류 모델을 학습하는 과정을 설명하기 위한 순서도이다.

도 9는 본 발명의 일 실시예에 따른 가짜뉴스 탐지 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0032] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 발명에 대해 구체적으로 설명하기로 한다.

[0033] 본 발명에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0034] 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 있음을 의미한다. 또한, 명세서에서 사용되는 "부", "모듈", "유닛" 등의 용어는 적어도 하나의 기능 또는 동작을 처리하는 단위를 의미하며, 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소, 또는 소프트웨어와 하드웨어의 결합으로 구현될 수 있다. 그렇지만 "부", "모듈", "유닛" 등의 용어가 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부", "모듈", "유닛" 등은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 ""부", "모듈", "유닛" 등의 용어는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.

[0035] 아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략한다.

[0036] "제1", "제2" 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 구성 요소들은 용어들에 의해 한정되지는 않는다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는" 이라는 용어는 복수의 관련된 항목들의 조합 또는 복수의 관련된 항목들 중의 어느 하나의 항목을 포함한다.

[0038] 이하, 본 발명의 일 실시예에서는 가짜뉴스 탐지 장치(1)를 설명하도록 한다.

[0039] 도 1은 본 발명의 일 실시예에 따른 가짜뉴스 탐지 장치(1)의 구성을 설명하기 위한 블록도이다.

[0040] 도 1에서 도시한 바와 같이, 가짜뉴스 탐지 장치(10)는 입력부(100), 학습부(110), 제어부(120), 출력부(130), 저장부(140) 및 통신부(150)를 포함한다.

[0041] 입력부(100)는 가짜뉴스 탐지를 학습하기 위한 학습 데이터를 입력받는다.

[0042] 이때, 학습 데이터는 텍스트 형태로 구성될 수 있으며, 예를 들어, 가짜뉴스와 진짜 뉴스로 구성된 데이터 셋인 ISOT Fake News Dataset이 학습 데이터가 될 수 있다.

[0043] 도 2는 ISOT Fake News Dataset의 구성을 설명하기 위한 도면이다.

[0044] 도 2에서 도시한 바와 같이, 입력부(100)는 23,481개의 가짜뉴스와 21,417개의 진짜 뉴스를 포함하여 수집된 2016년부터 2017년까지의 기사를 입력받을 수 있다.

[0045] 또한, 입력부(100)는 편향성 분류 모델을 학습시키기 위해 WIKIBIAS corpus인 학습 데이터를 입력받을 수 있다.

[0046] 입력부(100)는 위키피디아의 문서 편집 목록에서 수집된 NPOV 정책에 관련된 편집 기록으로부터 수정 전의 편향된 문장을 1로 레이블링 된 학습 데이터와 수정 후의 편향되지 않은 문장을 0으로 레이블링 된 학습 데이터를 입력받을 수 있다.

[0047] 입력부(100)는 WIKIBIAS corpus을 기반으로 기 제작된 양식에 따라 일부 데이터에 골드 레이블을 부여하는 데이터 셋인 WIKIBIAS-manual을 학습 데이터로 입력받을 수 있다.

[0048] 도 3은 WIKIBIAS-manual 데이터 셋을 설명하기 위한 도면이다.

[0049] 도 3에서 도시한 바와 같이, 입력부(100)는 'Since 1985, Starostin had been developing STARLING, which is the most advanced linguist's workplace software.' 문장에서 'the most advanced'에 의해 해당 문장을 1로 레이블링 된 학습 데이터를 입력받을 수 있다.

[0050] 그리고, 입력부(100)는 'Since 1985, Starostin had been developing STARLING, which is a linguist's workplace software.'을 0으로 레이블링 된 학습 데이터를 입력받을 수 있다.

[0051] 또한, 입력부(100)는 가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는다.

[0052] 학습부(110)는 가짜뉴스 여부를 판단하기 위한 편향성 분류 모델, 감정 분류 모델, 가짜뉴스 탐지 모델을 학습시킨다.

[0053] 먼저, 학습부(110)는 입력부(100)로부터 입력받은 학습 데이터를 변경, 삭제 등 다양한 방법을 통해 학습 데이터의 양을 증강시킬 수 있다.

[0054] 구체적으로, 학습부(110)는 Easy Data Augmentation을 기반으로 유의어 교체(Synonym Replacement, SR), 임의의 단어 삽입(Random Insertion, RI), 임의의 단어 삭제(Random Deletion, RD), 단어 위치 변경(Random Swap, RS)을 이용하여 학습 데이터를 증강시킬 수 있다.

[0055] 학습부(110)는 An Easy Data Augmentation을 기반으로 문장 사이에 문장부호를 무작위로 삽입하여 텍스트의 단어 위치를 변형하여 학습 데이터를 증강시킬 수 있다.

[0056] 즉, 본 발명의 일 실시예에서는 학습 데이터의 크기를 증강시켜 복수의 분류 모델을 학습시킬 수 있다.

[0058] 또한, 학습부(110)는 증강된 학습 데이터에 기초하여 편향성 분류 모델을 학습시킬 수 있다.

[0059] 학습부(110)는 증강된 학습 데이터로 문서 내의 각각의 문장을 이용할 수 있다.

[0060] 학습부(110)는 증강된 학습 데이터에 기초하여 개별 문자의 편향성 수치를 산출하도록 편향성 분류 모델을 학습시킬 수 있다.

[0061] 더욱 자세하게는, 학습부(110)는 증강된 학습 데이터를 편향성 분류 모델에 입력하여 문장의 편향성 수치를 산출하도록 편향성 분류 모델을 학습시킬 수 있다.

[0062] 학습부(110)는 Bidirectional Encoder Representations from Transformers(BERT) 언어모델을 기반으로 하는 편

향성 분류 모델을 학습시킬 수 있다.

[0063] 이때, BERT는 대용량의 텍스트 데이터를 이용하여 사전 학습(pre-training)된 언어 모델일 수 있다.

[0064] 학습부(110)는 사전 학습된 BERT를 파인 튜닝(fine-tuning)하여 편향된 문장이 있는지 여부에 대하여 문장의 편향성 수치를 산출하도록 편향성 분류 모델을 학습시킬 수 있다.

[0065] 즉, 학습부(110)는 증강된 학습 데이터를 편향성 분류 모델에 입력하여 문장의 편향성 수치를 출력하도록 학습시킬 수 있다.

[0066] 또한, 학습부(110)는 증강된 학습 데이터에 기초하여 상기 감전 분류 모델을 학습시킨다.

[0067] 학습부(110)는 증강된 학습 데이터에 기초하여 문장 감정 점수를 산출하도록 감정 분류 모델을 학습시킬 수 있다.

[0068] 이때, 학습부(110)는 Sentiment Intensity Analyzer(SIA)를 기반으로 감정 사전을 통해 규칙 기반의 분석 기법을 이용하여 감정 분류 모델을 학습시킬 수 있다.

[0069] 즉, 학습부(110)는 증강된 학습 데이터를 감정 분류 모델에 입력하여 문장 감정 점수를 산출하도록 감정 분류 모델을 학습시킬 수 있다.

[0070] 또한, 학습부(110)는 문서 편향성 점수와 문서 감정 점수에 기초하여 가짜뉴스 탐지 모델을 학습시킨다.

[0071] 학습부(110)는 문서 편향성 점수와 문서 감정 점수에 기초하여 가짜뉴스 여부를 판단하는 가짜뉴스 탐지 모델을 학습시킬 수 있다.

[0072] 가짜뉴스 탐지 모델은 Support Vector Machine(SVM)를 기반으로 학습될 수 있다.

[0073] 학습부(110)는 문서 편향성 점수와 문서 감정 점수를 가짜뉴스 탐지 모델의 입력데이터로 설정하고, 가짜뉴스 또는 진짜뉴스를 출력데이터로 설정하여 가짜뉴스 탐지 모델을 학습시킬 수 있다.

[0074] 학습부(110)는 문서 감정 점수 및 문서 편향성 점수가 기준선보다 높은 경우에 가짜뉴스로 판단하도록 가짜뉴스 탐지 모델을 학습시키고, 문서 감정 점수 및 문서 편향성 점수가 기준선이하인 경우에 진짜뉴스로 판단하도록 가짜뉴스 탐지 모델을 학습시킬 수 있다.

[0075] 이때, 기준선은 기준값으로 설정될 수도 있다.

[0076] 제어부(120)는 평가 문서를 미리 학습된 편향성 분류 모델, 감정 분류 모델에 적용하여 문서 편향성 점수 및 문서 감정 점수를 산출하고, 산출된 문서 편향성 점수 및 문서 감정 점수를 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단한다.

[0077] 제어부(120)는 편향성 분류부(121), 감정 분류부(123) 및 탐지부(125)를 포함한다.

[0078] 편향성 분류부(121)는 미리 학습된 편향성 분류 모델을 이용하여 입력부(100)로부터 입력받은 가짜뉴스 여부를 판단하고자 하는 평가 문서의 문서 편향성 점수를 산출한다.

[0079] 도 4는 문서 편향성 점수를 산출하는 과정을 설명하기 위한 도면이다.

[0080] 더욱 자세하게는, 도 4의 (a) 및 (b)에서 도시한 바와 같이, 편향성 분류부(121)는 입력받은 평가 문서(D)를 복수의 문장(d1 내지 d6)으로 분할할 수 있다.

[0081] 도 4의 (b) 내지 (d)와 같이, 편향성 분류부(121)는 복수의 문장(d1 내지 d6)을 미리 학습된 편향성 분류 모델에 입력하여 개별 문장의 편향성 수치를 산출할 수 있다.

[0082] 이때, 문장의 편향성 수치는 0 과 1에 해당하는 수치로 나타낼 수 있으며, 편향성이 높은 문장일 수로 편향성 수치가 1에 가깝고, 편향성이 낮은 문장일수록 0에 가까운 수치를 가진다.

[0083] 예를 들어, 도 4의 (d)와 같이, 편향성 분류부(121)는 문장 d1의 편향성 수치를 0.7, 문장 d2의 편향성 수치를 0.2, 문장 d3의 편향성 수치를 0.6, 문장 d4의 편향성 수치를 0.7, 문장 d5의 편향성 수치를 0.7, 문장 d6의 편향성 수치를 0.4로 출력할 수 있다.

[0084] 이와 같이, 편향성 분류부(121)는 복수의 문장을 편향성 분류 모델에 입력하여 개별 문장에 대한 편향성 수치를 산출할 수 있다.

[0085] 일 실시예에서, 편향성 분류부(121)는 개별 문장의 편향성 수치에 기초하여 평가 문서의 편향성 분포를 도출할 수 있다.

[0086] 도 4의 (d)와 같이, 편향성 분류부(121)는 문장 d1 내지 문장 d6의 편향성 수치를 통해 평가 문서 전체에 대한 편향성 분포를 도출할 수 있다.

[0087] 이때, 편향성 수치가 0에 가까울수록 편향성이 없는 문장이고, 편향성 수치가 1에 가까울수록 편향된 문장이므로, 문장 d1, 문장 d3 내지 문장 d5에 편향된 문장이 분포에 있고, 문장 d2와 문장 d6은 편향되지 않은 문장이 분포되어 있음을 알 수 있다.

[0088] 도 4의 (d) 및 (e)와 같이, 편향성 분류부(121)는 평가 문서의 편향성 분포에 기초하여 개별 문장의 편향성 수치를 보정할 수 있다.

[0089] 일 실시예에서, 문장 d3 내지 문장 d5와 같이. 편향성 분류부(121)는 평가 문서의 편향성 분포를 기반으로 편향된 문장이 연속적으로 존재하는 경우에 연속적으로 편향된 문장 각각에 대한 편향성 수치에 가중치를 부여하여 개별 문장의 편향성 수치를 보정할 수 있다.

[0090] 편향성 분류부(121)는 연속적으로 편향된 문장들 중에서 앞 문장보다 뒷 문장에 더 큰 가중치를 부여하여 문장의 편향성 수치를 보정할 수 있다.

[0091] 예를 들어, 편향성 분류부(121)는 문장 d3의 편향성 수치를 0.7로 보정할 수 있고, 문장 d4의 편향성 수치를 0.9로 보정할 수 있고, 문장 d5의 편향성 수치를 1.0으로 보정할 수 있다.

[0092] 본 발명의 일 실시예에서, 편향성 분류부(121)는 연속적으로 나타난 편향된 문장 중에서 뒷 문장으로 갈수록 더 큰 가중치를 부여하여 개별 문장의 편향성 수치를 보정하는 것으로 기술하였으나, 한정하지 않는다.

[0093] 다른 실시예에서, 편향성 분류부(121)는 문서 중 미리 설정된 기준 위치에 편향된 문장이 포함되는 경우에 편향된 문장 각각에 대해 가중치를 부여하여 문장의 편향성 수치를 보정할 수 있다.

[0094] 예를 들어, 평가 문서 중 복수의 문장(d1 내지 d6)이 존재하고, 미리 설정된 기준 위치가 문장 d4 내지 문장 d6라고 가정한다.

[0095] 기준 위치는 변경될 수 있고, 한정하지 않는다.

[0096] 편향성 분류부(121)는 문장 d4 내지 문장 d6 중 적어도 하나의 문장이 편향된 문장인 경우에 해당 문장의 편향성 수치에 가중치를 부여하여 문장의 편향성 수치를 보정할 수 있다.

[0097] 그리고, 편향성 분류부(121)는 문장 d1 내지 문장 d3 중에 편향된 문장이 있더라도 가중치를 부여하지 않을 수 있다.

[0098] 또 다른 실시예에서, 편향성 분류부(121)는 문장 감정 점수가 미리 설정된 기준 이상인 경우에 개별 문장의 편향성 수치를 보정할 수 있다.

[0099] 이때, 편향성 분류부(121)는 문장 감정 점수의 절대값이 미리 설정된 기준이상인지 판단할 수 있으며, 기준은 변경될 수 있고, 한정하지 않는다.

[0100] 예를 들어, 도 4 및 도 6에서의 문장 d1 내지 문장 d6과 같이, 편향성 분류부(121)는 문장 감정 점수의 절대값이 0.5이상인 문장 d5 및 d6에 대하여 편향성 수치를 보정할 수 있다.

[0101] 즉, 편향성 분류부(121)는 동일한 평가 문서에서 감정이 포함된 문장에 대하여 해당 문장의 편향성 수치에 가중치를 부여하여 해당 문장의 편향성 수치를 보정할 수 있다.

[0102] 도 4의 (e) 및 (f)와 같이, 편향성 분류부(121)는 개별 문장의 편향성 수치에 기초하여 문서 편향성 점수를 산출할 수 있다.

[0103] 편향성 분류부(121)는 평가 문서 내의 복수의 문장들끼리의 평균값을 문서 편향성 점수로 산출할 수 있다.

[0104] 편향성 분류부(121)는 문서 편향성 점수를 0이상 1이하의 값으로 산출할 수 있고, 0에 가까울수록 편향되지 않은 문서이고, 1에 가까울수록 편향된 문서일 수 있다.

[0105] 도 5는 문서 편향성 점수를 설명하기 위한 도면이다.

[0106] 도 5에서 도시한 바와 같이, 편향성 분류부(121)는 문서의 편향성 분포를 고려하여 보정된 문장의 편향성 수치

를 이용하여 문서 편향성 점수를 산출할 수 있다.

[0107] 도 4의 (d)와 같이, 문서의 편향성 분포를 고려하지 않은 문장의 편향성 수치를 기반으로 한 경우, 편향성 분류부(121)는 문서 편향성 점수를 0.55로 산출할 수 있고, 도 4의 (e)와 같이, 문서의 편향성 분포를 고려하여 보정된 문장의 편향성 수치를 기반으로 한 경우, 편향성 분류부(121)는 문서 편향성 점수를 0.65로 산출할 수 있다.

[0108] 이와 같이, 편향성 분류부(121)는 문서의 편향성 분포를 기반으로 보정된 문장의 편향성 수치를 이용하여 문서 편향성 점수를 산출할 수 있다.

[0109] 감정 분류부(123)는 미리 학습된 감정 분류 모델을 이용하여 평가 문서의 문서 감정 점수를 산출할 수 있다.

[0110] 도 6은 문서 감정 점수를 산출하는 과정을 설명하기 위한 도면이다.

[0111] 도 6의 (a) 및 (b)에서 도시한 바와 같이, 감정 분류부(123)는 입력받은 평가 문서를 복수의 문장(d1 내지 d6)으로 분할할 수 있다.

[0112] 일 실시예에서, 도 6의 (b) 내지 (d)와 같이, 감정 분류부(123)는 복수의 문장(d1 내지 d6)을 미리 학습된 감정 분류 모델에 입력하여 각각의 문장에 대한 문장 감정 점수를 산출할 수 있다.

[0113] 감정 분류부(123)는 -1부터 1까지의 문장 감정 점수를 산출할 수 있다.

[0114] 여기서, 문장 감정 점수가 -1이상 0미만, 0초과 1이하인 경우에 감정이 포함된 문장이고, 문장 감정 점수가 0인 경우에 감정이 포함되지 않은 중립 문장일 수 있다.

[0115] 이때, 문장 감정 점수의 범위는 한정하지 않으며, 관리자에 의해 변경될 수 있다.

[0116] 예를 들어, 도 6의 (d)와 같이, 감정 분류부(123)는 문장 d1을 0.2로, 문장 d2 및 문장 d3를 0으로, 문장 d4를 -0.3으로, 문장 d5를 -0.5로, 문장 d6을 0.6으로 문장 감정 점수를 출력할 수 있다.

[0117] 일 실시예에서, 감정 분류부(123)는 문장 감정 점수에 기초하여 평가 문서의 감정 분포를 도출할 수 있다.

[0118] 도 6의 (d)와 같이, 감정 분류부(123)는 문장 d1 내지 문장 d6에 대한 문장 감정 점수를 통해 평가 문서에서의 감정을 가지는 문장과 감정을 가지지 않는 문서의 분포를 도출할 수 있다.

[0119] 감정 분류부(123)는 평가 문서의 문서의 감정 분포를 기초하여 문장 감정 점수를 보정할 수 있다.

[0120] 일 실시예에서, 복수의 문장(d1 내지 d6) 중에 연속적으로 -1이상 0미만, 0초과 1이하인 값을 가지는 문장이 존재하는 경우, 감정 분류부(123)는 해당 문장의 문장 감정 점수를 보정할 수 있다.

[0121] 도 6의 (d)에서 문장 d4 내지 문장 d6와 같이 연속적으로 문장의 문장 감정 점수가 -1이상 0미만, 0초과 1이하인 경우, 감정 분류부(123)는 문장 d4 내지 문장 d6 각각에 대하여 가중치를 부여하여 문장 감정 점수를 보정할 수 있다.

[0122] 예를 들어, 감정 분류부(123)는 문장 d4의 문장 감정 점수를 -0.4, 문장 d5의 문장 감정 점수를 -0.7, 문장 d6의 문장 감정 점수를 0.9로 보정할 수 있다.

[0123] 본 발명의 일 실시예에서, 감정 분류부(123)는 연속적으로 나타난 -1이상 0미만, 0초과 1이하의 문장 감정 점수 중에서 뒷 문장으로 갈수록 더 높은 가중치를 부여하여 문장 감정 점수를 보정하는 것으로 기술하였으나, 한정하지 않는다.

[0124] 다른 실시예에서, 감정 분류부(123)는 문서 중 미리 설정된 기준 위치에 감정 점수가 -1이상 0미만, 0초과 1이하인 문장이 존재하는 경우에 해당 문장에 대해 가중치를 부여하여 문장 감정 점수를 보정할 수 있다.

[0125] 예를 들어, 문서 중 복수의 문장(d1 내지 d6)이 존재하고, 미리 설정된 기준 위치가 문장 d2 내지 문장 d4라고 가정한다.

[0126] 기준 위치는 변경될 수 있고, 한정하지 않는다.

[0127] 감정 분류부(123)는 문장 d2 내지 문장 d4 중 적어도 하나의 문장의 문장 감정 점수가 -1이상 0미만, 0초과 1이하인 문장에 대하여 가중치를 부여하여 문장 감정 점수를 보정할 수 있다.

[0128] 만일, 문장 d1, 문장 d5 및 문장 d6 중에 문장 감정 점수가 -1이상 0미만, 0초과 1이하인 값을 가지더라도, 감

정 분류부(123)는 해당 문장에 대하여 가중치를 부여하지 않을 수 있다.

[0129] 또 다른 실시예에서, 감정 분류부(123)는 개별 문장의 편향성 수치가 미리 설정된 기준 이상인 경우에 문장 감정 점수를 보정할 수 있다.

[0130] 이때, 기준은 변경될 수 있으며, 한정하지 않는다.

[0131] 예를 들어, 도 4 및 도 6의 (d)에서 문장 d1 내지 문장 d6과 같이, 감정 분류부(123)는 개별 문장의 편향성 수치가 0.7이상인 문장 d1, 문장 d4 및 문장 d5에 대하여 문장 감정 점수를 보정할 수 있다.

[0132] 본 발명에서는 문장 감정 점수를 기초로 개별 문장의 편향성 수치를 보정하거나, 개별 문장의 편향성 수치를 기초로 문장 감정 점수를 보정하는 것을 기술하였으나, 문장 감정 점수 및 개별 문장의 편향성 수치가 모두 기준 이상인 경우에 문장 감정 점수 및 개별 문장의 편향성 수치를 각각 보정할 수도 있다.

[0133] 감정 분류부(123)는 문장 감정 점수를 기반으로 문서 감정 점수를 산출할 수 있다.

[0134] 감정 분류부(123)는 문장 감정 점수의 절대값을 이용하여 문서 감정 점수를 산출할 수 있다.

[0135] 감정 분류부(123)는 평가 문서에 포함된 복수의 문장들끼리의 평균 값을 문서 감정 점수로 산출할 수 있다.

[0136] 예를 들어, 도 6의 (e)와 같이, 감정 분류부(123)는 문서 감정 점수를 0.26으로 산출할 수 있다.

[0137] 감정 분류부(123)는 보정된 문장 감정 점수도 동일한 방식으로 문서 감정 점수를 산출할 수 있다.

[0138] 즉, 감정 분류부(123)는 평가 문서를 기반으로 미리 학습된 감정 분류 모델을 통해 문서 감정 점수를 산출할 수 있다.

[0139] 탐지부(125)는 문서 편향성 점수 및 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단할 수 있다

[0140] 탐지부(125)는 Support Vector Machine(SVM)를 기반으로 학습된 가짜뉴스 탐지 모델을 통해 평가 문서가 가짜뉴스인지 진짜뉴스인지 판단할 수 있다.

[0141] 일 실시예에서, 탐지부(125)는 문서 편향성 점수 및 문서 감정 점수가 미리 설정된 기준선보다 큰 경우에 가짜뉴스로 판단하고, 기준선이하인 경우에 진짜뉴스로 판단할 수 있다.

[0142] 도 7은 가짜뉴스 여부를 판단하는 과정을 설명하기 위한 도면이다.

[0143] 도 7에서 도시한 바와 같이, 탐지부(125)는 미리 설정된 기준선에 대응하여 문서 편향성 점수 및 문서 감정 점수를 기반으로 가짜뉴스 여부를 판단할 수 있다.

[0144] 탐지부(125)는 미리 설정된 기준선보다 상단에 있는 문서에 대하여 가짜뉴스로 판단하고, 미리 설정된 기준선보다 하단에 있는 문서에 대하여 진짜뉴스로 판단할 수 있다.

[0145] 이때, x는 문서 감정 점수이고, y는 문서 편향성 점수일 수 있고, 기준선은 기준값이 될 수도 있고, 관리자에 의해 변경될 수 있으므로 한정하지 않는다.

[0146] 즉, 제어부(120)는 평가 문서를 기반으로 편향성 분류부(121), 감정 분류부(123)에 적용하여 문서 편향성 점수 및 문서 감정 점수를 산출하고, 산출된 문서 편향성 점수 및 문서 감정 점수를 기반으로 탐지부(125)를 통해 평가 문서가 가짜뉴스인지, 진짜 뉴스인지 분류할 수 있다.

[0147] 제어부(120)는 중앙 처리 장치(CPU: Central Processing Unit), 그래픽 처리 장치(GPU: Graphic Processing Unit), 마이크로 컨트롤러 유닛(MCU: Micro Controller Unit), 애플리케이션 프로세서(AP: Application Processor), 전자 제어 유닛(ECU: Electronic Controlling Unit) 및/또는 이외 각종 연산 및 제어 처리를 수행할 수 있는 적어도 하나의 전자 장치 등을 포함할 수 있다. 이들 장치는, 예를 들어, 하나 또는 둘 이상의 반도체 칩, 회로 또는 관련 부품 등을 단독으로 이용하거나 조합하여 구현된 것일 수도 있다.

[0149] 출력부(130)는 탐지부(125)를 통해 평가 문서가 가짜뉴스인지, 진짜뉴스인지 판단된 결과값을 출력할 수 있다.

[0150] 출력부(130)는, 예를 들어, 디스플레이, 프린터 장치, 스피커 장치, 영상 출력 단자, 데이터 입출력 단자 또는 통신 모듈 등을 포함할 수 있으나, 이에 한정되는 것은 아니다.

[0152] 저장부(140)는 가짜뉴스 탐지 장치의 운영 프로그램, 가짜뉴스 탐지 프로그램, 영상처리 프로그램 등과 같이 장치의 운영에 필요한 정보를 저장할 수 있으며, 제어부(120)로부터 분류된 해당 문서가 가짜뉴스인지 진짜 뉴스인지 여부를 저장할 수 있다.

[0153] 저장부(140)는 입력부(100)로부터 입력받은 학습 데이터 및 평가 문서를 저장할 수 있다.

[0154] 저장부(140)는 학습부(110)를 통해 학습된 편향성 분류 모델, 감정 분류 모델, 가짜뉴스 탐지 모델의 결과 값을 저장할 수 있다.

[0155] 저장부(140)는 주기억장치 및 보조기억장치 중 적어도 하나를 포함할 수 있다. 주기억장치는 예를 들어, 롬(ROM) 및/또는 램(RAM)과 같은 반도체 저장 매체를 이용하여 구현된 것일 수 있고, 보조기억장치는, 플래시 메모리 장치(솔리드 스테이트 드라이브(SSD, Solid State Drive) 등), SD(Secure Digital) 카드, 하드 디스크 드라이브(HDD, Hard Disc Drive), 콤팩트 디스크, 디브이디(DVD) 또는 레이저 디스크 등과 같이 데이터를 영구적 또는 반영구적으로 저장 가능한 장치를 기반으로 구현될 수 있다.

[0157] 통신부(150)는 입력받은 학습 데이터를 학습부(110)로 전송하고, 입력받은 평가 문서를 제어부(120)로 전송할 수 있다.

[0158] 통신부(150)는 학습부(110)에서 미리 학습된 편향성 분류 모델, 감정 분류 모델, 가짜뉴스 탐지 모델을 제어부(120)로 전송할 수 있다.

[0159] 통신부(150)는 제어부(120)의 편향성 분류부(121), 감정 분류부(123) 및 탐지부(125)가 서로 연결되어 출력되어 나온 결과값을 송수신할 수 있도록 하고, 탐지부(125)로부터 판단된 출력값인 가짜뉴스 또는 진짜뉴스를 출력부(130)로 전송할 수 있다.

[0160] 통신부(150)는 유무선 통신망을 포함할 수 있고, 유선 통신망은 케이블망이나 공중 전화망(PSTN)과 같은 인터넷망을 포함할 수 있으며, 무선 통신망은 CDMA, WCDMA, GSM, EPC(Evolved Packet Core), LTE(Long Term Evolution), 와이브로(Wibro) 망, 5G 통신망 등을 포함할 수 있다.

[0161] 통신부(150)는 이에 한정되는 것이 아니며, 향후 구현될 차세대 이동통신 시스템의 접속망으로서 가령 클라우드 컴퓨팅 환경하의 클라우드 컴퓨팅망, 5G망 등에 사용될 수 있고, 통신부(150)가 유선 통신망인 경우 통신부(150) 내의 액세스포인트는 전화국의 교환국 등에 접속할 수 있지만, 무선 통신망인 경우에는 통신사에서 운용하는 SGSN 또는 GGSN(Gateway GPRS Support Node)에 접속하여 데이터를 처리하거나, BTS(Base Station Transmission), NodeB, e-NodeB 등의 다양한 중계기에 접속하여 데이터를 처리할 수 있다.

[0163] 이하에서는, 도면을 참조하여 본 발명의 일 실시예에 따른 복수의 분류 모델을 학습하는 방법을 설명한다.

[0164] 도 8은 본 발명의 일 실시예에 따른 복수의 분류 모델을 학습하는 과정을 설명하기 위한 순서도이다.

[0165] 먼저, 입력부(100)는 가짜뉴스 탐지를 학습하기 위한 학습 데이터를 입력받는다(S110).

[0166] S110 단계는 도 1의 입력부(100)에서 이미 설명하였는 바, 자세한 설명은 생략한다.

[0167] 학습부(110)는 학습 데이터의 양을 증강시킨다(S120).

[0168] 학습부(110)는 증강된 학습 데이터에 기초하여 편향성 분류 모델을 학습한다(S130).

[0169] 학습부(110)는 증강된 학습 데이터를 편향성 분류 모델에 입력하여 개별 문장의 편향성 수치를 출력하도록 편향성 분류 모델을 학습시킬 수 있다.

[0170] 학습부(110)는 증강된 학습 데이터에 기초하여 감정 분류 모델을 학습한다(S140).

[0171] 학습부(110)는 증강된 학습 데이터를 감정 분류 모델에 입력하여 문장 감정 점수를 출력하도록 감정 분류 모델을 학습시킬 수 있다.

[0172] 이때, S140 단계는 생략되거나, S130 단계와 병렬적으로 수행될 수 있다.

[0173] 학습부(110)는 문서 편향성 점수 및 문서 감정 점수를 기초하여 가짜뉴스 탐지 모델을 학습한다(S150).

[0174] 학습부(110)는 문서 편향성 점수 및 문서 감정 점수를 가짜뉴스 탐지 모델에 입력하여 입력된 학습 데이터가 가짜뉴스인지, 진짜뉴스인지 판단하도록 가짜뉴스 탐지 모델을 학습시킬 수 있다.

[0175] S140 단계가 생략되는 경우, 학습부(110)는 S150 단계에서 문서 편향성 점수만을 이용하여 가짜뉴스 탐지 모델을 통해 학습 데이터가 가짜뉴스인지, 진짜뉴스인지 판단하도록 학습시킬 수 있다.

[0176] S120 내지 S150 단계는 도 1의 학습부(110)에서 이미 설명하였는 바, 자세한 설명은 생략한다.

[0178] 이하에서는, 도면을 참조하여 본 발명의 일 실시예에 따른 가짜뉴스 탐지 방법을 설명한다.

[0179] 도 9는 본 발명의 일 실시예에 따른 가짜뉴스 탐지 방법을 설명하기 위한 순서도이다.

[0180] 먼저, 입력부(100)는 가짜뉴스 여부를 판단하고자 하는 평가 문서를 입력받는다(S210).

[0181] S210 단계는 도 1의 입력부(100)에서 이미 설명하였는 바, 자세한 설명은 생략한다.

[0182] 편향성 분류부(121)는 미리 학습된 편향성 분류 모델을 이용하여 평가 문서의 문서 편향성 점수를 산출한다(S220).

[0183] S220 단계는 도 1의 편향성 분류부(121)에서 이미 설명하였는 바, 자세한 설명은 생략한다.

[0184] 감정 분류부(123)는 미리 학습된 감정 분류 모델을 이용하여 평가 문서의 문서 감정 점수를 산출한다(S230).

[0185] S230 단계는 도 1의 감정 분류부(123)에서 이미 설명하였는 바, 자세한 설명은 생략한다.

[0186] S230 단계는 생략되거나, S220 단계와 병렬적으로 수행될 수 있다.

[0187] 탐지부(125)는 문서 편향성 점수 및 문서 감정 점수를 미리 학습된 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단한다(S240).

[0188] S230 단계가 생략되는 경우, 탐지부(125)는 S240 단계에서 문서 편향성 점수만을 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단할 수 있다.

[0189] S240 단계는 도 1의 탐지부(125)에서 이미 설명하였는 바, 자세한 설명은 생략한다.

[0191] 이상에서 설명한 바와 같이, 본 발명에 따르면, 문서 내의 문장의 편향성 및 문장의 감정을 기반으로 가짜뉴스를 탐지할 수 있다.

[0193] 본원 발명의 실시예들과 관련된 기술 분야에서 통상의 지식을 가진 자는 상기 기재의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로, 개시된 방법들은 한정적인 관점이 아닌 설명적 관점에서 고려되어야 한다. 본 발명의 범위는 발명의 상세한 설명이 아닌 특허청구 범위에 나타나며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0194] 1: 가짜뉴스 탐지 장치

100: 입력부

110: 학습부

120: 제어부

121: 편향성 분류부

123: 감정 분류부

125: 탐지부

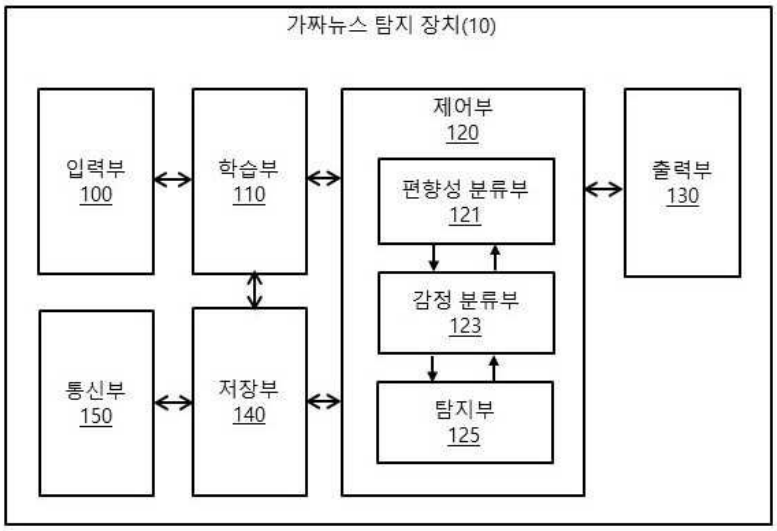
130: 출력부

140: 저장부

150: 통신부

도면

도면1



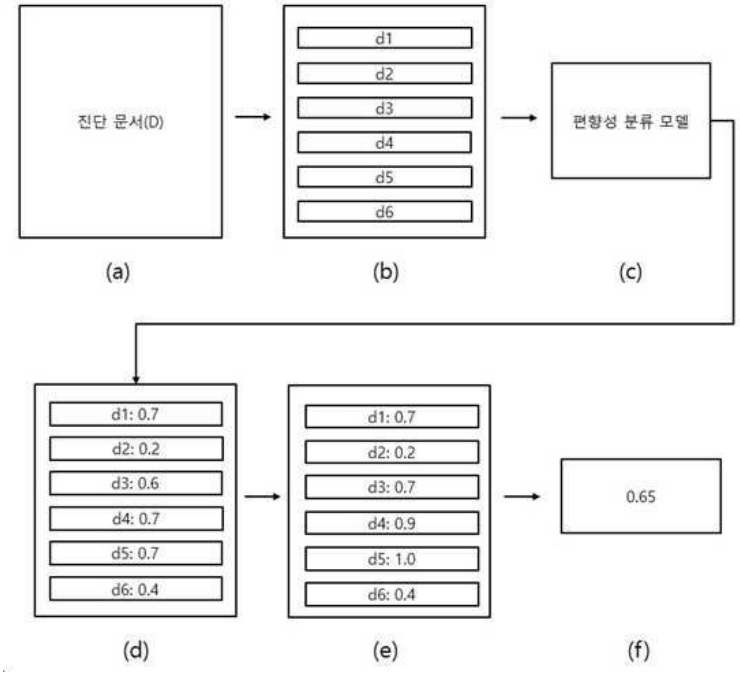
도면2

News	Size	Subjects	
		Type	Size
Real-News	21417	World-news	10145
		Politics-news	11272
Fake-News	23481	Government-News	1570
		Middle-east	778
		US News	783
		Left-news	4459
		Politics	6841
		News	9050
Total			44898

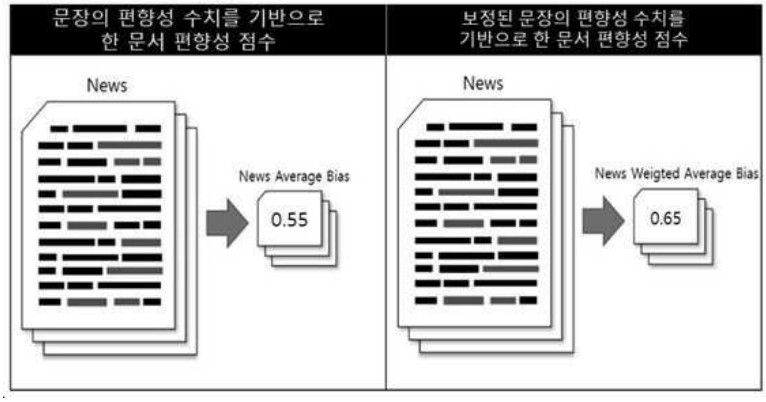
도면3

Sentence	Label
Since 1985, Starostin had been developing STARLING, which is **the most advanced** linguist's workplace software.	1
Since 1985, Starostin had been developing STARLING, which is **a** linguist's workplace software.	0

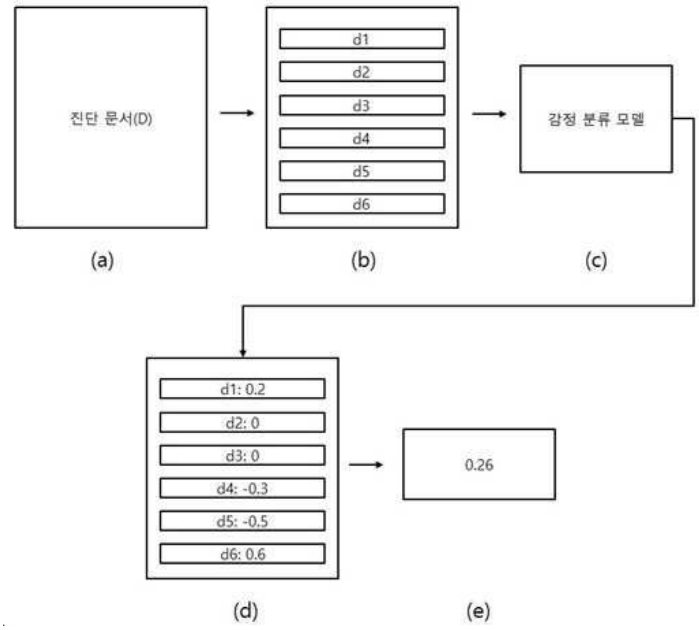
도면4


진단 문서(D)
d1
d2
d3
d4
d5
d6
편향성 분류 모델
(a)
(b)
(c)
d1: 0.7
d2: 0.2
d3: 0.6
d4: 0.7
d5: 0.7
d6: 0.4
d1: 0.7
d2: 0.2
d3: 0.7
d4: 0.9
d5: 1.0
d6: 0.4
0.65
(d)
(e)
(f)

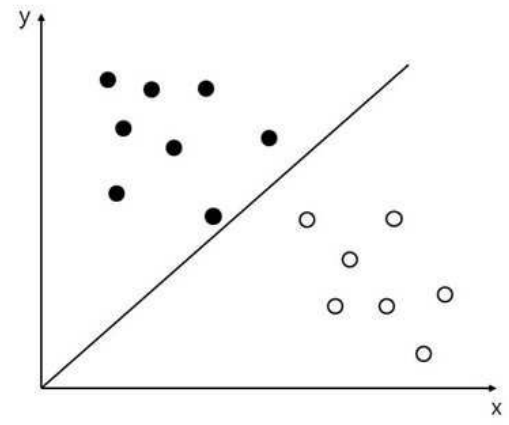
도면5


문장의 편향성 수치를 기반으로 한 문서 편향성 점수
News
News Average Bias
0.55
보정된 문장의 편향성 수치를 기반으로 한 문서 편향성 점수
News
News Weigted Average Bias
0.65

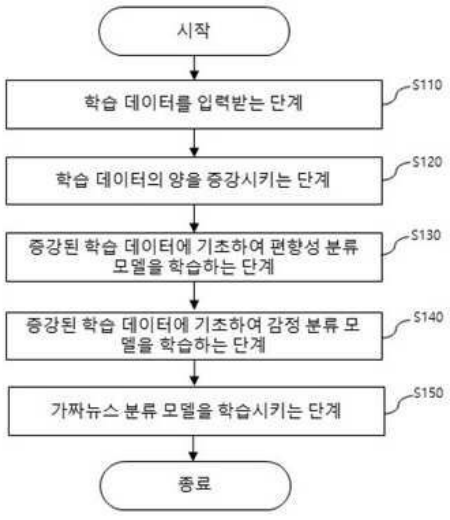
도면6


진단 문서(D)
d1
d2
d3
d4
d5
d6
감정 분류 모델
(a)
(b)
(c)
d1: 0.2
d2: 0
d3: 0
d4: -0.3
d5: -0.5
d6: 0.6
0.26
(d)
(e)

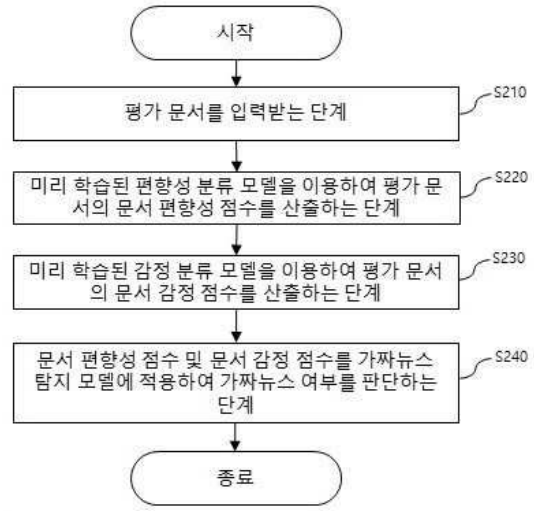
도면7


y
x

도면8


시작
학습 데이터를 입력받는 단계
S110
학습 데이터의 양을 증강시키는 단계
S120
증강된 학습 데이터에 기초하여 편향성 분류 모델을 학습하는 단계
S130
증강된 학습 데이터에 기초하여 감정 분류 모델을 학습하는 단계
S140
가짜뉴스 분류 모델을 학습시키는 단계
S150
종료

도면9


시작
평가 문서를 입력받는 단계
S210
미리 학습된 편향성 분류 모델을 이용하여 평가 문서의 문서 편향성 점수를 산출하는 단계
S220
미리 학습된 감정 분류 모델을 이용하여 평가 문서의 문서 감정 점수를 산출하는 단계
S230
문서 편향성 점수 및 문서 감정 점수를 가짜뉴스 탐지 모델에 적용하여 가짜뉴스 여부를 판단하는 단계
S240
종료