

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0128616

(43) 공개일자 2023년09월05일

(51) 국제특허분류(Int. Cl.)

G06F 40/205 (2020.01) G06F 16/31 (2019.01)

G06F 40/117 (2020.01) G06F 40/237 (2020.01)

G06F 40/279 (2020.01) G06Q 50/10 (2012.01)

(52) CPC특허분류

G06F 40/205 (2020.01)

G06F 16/31 (2019.01)

(21) 출원번호 10-2022-0025684

(22) 출원일자 2022년02월28일

심사청구일자 2022년02월28일

(71) 출원인

세종대학교산학협력단

서울특별시 광진구 능동로 209 (군자동, 세종대학교)

(72) 발명자

김미숙

서울특별시 동작구 동작대로29길 91, 207동 508호

장준원

서울특별시 강남구 선릉로 206, 103동 1204호

(74) 대리인

김준석, 박민욱

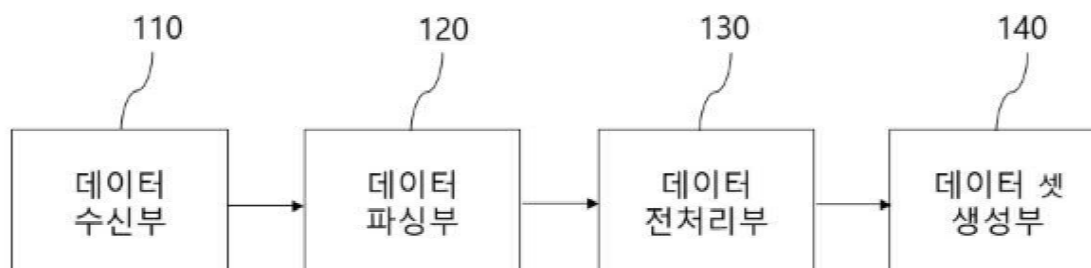
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 가짜 뉴스 탐지용 데이터 셋 생성 장치 및 이의 실행 방법

(57) 요약

본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 장치는 데이터 수신부로부터 뉴스 데이터를 수신하면 뉴스 데이터를 분석하여 특정 키워드 태그에 해당하는 데이터를 과잉하는 데이터 과잉부, 상기 데이터 과잉부에 의해 과잉된 뉴스 데이터에서 특정 태그에 해당하는 데이터를 삭제하여 전처리를 실행하는 데이터 전처리부 및 상기 데이터 전처리부에 의해 전처리된 뉴스 데이터 중 일부 뉴스 데이터 중 특정 태그를 동일하게 가지는 뉴스 데이터를 추출하고, 상기 특정 뉴스 데이터 중 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고 나머지 뉴스 데이터 중 일부 뉴스 데이터를 후보 뉴스 데이터로 결정하고, 상기 기준 뉴스 데이터 및 상기 후보 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하고, 상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 데이터 셋 생성부를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06F 40/117 (2020.01)

G06F 40/237 (2022.01)

G06F 40/279 (2020.01)

G06Q 50/10 (2015.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1375027257

과제번호 R2020040126-0002

부처명 문화체육관광부

과제관리(전문)기관명 한국콘텐츠진흥원

연구사업명 문화기술연구개발(R&D)

연구과제명 「실감형 뉴스」를 위한 빅데이터 분석, 제작, 유통이 가능한 참여형 통합 플랫폼 구축

기 여 율 50/100

과제수행기관명 (주)피씨엔

연구기간 2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711134451

과제번호 2021-0-00469-001

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 SW컴퓨팅산업원천기술개발(R&D, 정보화)

연구과제명 이중 융합 데이터 탐지 및 추적 기술 개발

기 여 율 50/100

과제수행기관명 (주)피씨엔

연구기간 2021.04.01 ~ 2021.12.31

명세서

청구범위

청구항 1

가짜 뉴스 탐지용 데이터 셋 생성 장치에 있어서,

데이터 수신부로부터 뉴스 데이터를 수신하면 뉴스 데이터를 분석하여 특정 키워드 태그에 해당하는 데이터를 파싱하는 데이터 파싱부;

상기 데이터 파싱부에 의해 파싱된 뉴스 데이터에서 특정 태그에 해당하는 데이터를 삭제하여 전처리를 실행하는 데이터 전처리부; 및

상기 데이터 전처리부에 의해 전처리된 뉴스 데이터 중 일부 뉴스 데이터 중 특정 태그를 동일하게 가지는 뉴스 데이터를 추출하고, 상기 특정 뉴스 데이터 중 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고 나머지 뉴스 데이터 중 일부 뉴스 데이터를 후보 뉴스 데이터로 결정하고, 상기 기준 뉴스 데이터 및 상기 후보 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하고, 상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 데이터 셋 생성부를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 장치.

청구항 2

제1항에 있어서,

상기 데이터 셋 생성부는

상기 전처리된 뉴스 데이터를 특정 비율로 나눈 후 그룹핑하여 제1 뉴스 데이터 그룹 및 제2 뉴스 데이터 그룹을 생성하고, 상기 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 각각에 진짜 뉴스 레이블을 할당하는 데이터 그룹 생성 모듈;

상기 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고, 나머지 뉴스 데이터 중 특정 개수의 후보 뉴스 데이터를 추출하는 데이터 추출 모듈;

상기 기준 뉴스 데이터의 본문에서 특정 개수의 문장을 노이즈 적용 문장으로 선택하는 노이즈 적용 문장 선택 모듈;

상기 노이즈 적용 문장을 상기 후보 뉴스 데이터의 문장 중 특정 문장으로 교환하여 가짜 뉴스 데이터 셋을 생성하는 가짜 뉴스 데이터 셋 생성 모듈; 및

상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 가짜 뉴스 탐지용 데이터 셋 생성 모듈을 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 장치.

청구항 3

제2항에 있어서,

상기 데이터 추출 모듈은

상기 기준 뉴스 데이터의 문장 개수에 따라 후보 뉴스 데이터를 추출하는 특정 개수를 결정하고, 상기 기준 뉴스 데이터를 송고 날짜를 기준으로 특정 일 이전 및 이후에 작성된 뉴스 데이터를 상기 특정 개수의 후보 뉴스

데이터의 선택 시 제외하여 후보 뉴스 데이터를 추출하는 것을 특징으로 하는
가짜 뉴스 탐지용 데이터 셋 생성 장치.

청구항 4

제2항에 있어서,

상기 노이즈 적용 문장 선택 모듈은

상기 기준 뉴스 데이터 및 특정 개수(N)의 후보 뉴스 데이터 각각의 본문의 각 문장 사이의 유사도에 따라 상기 기준 뉴스 데이터의 본문에서 특정 개수의 노이즈 적용 문장을 선택하고,

상기 가짜 뉴스 탐지용 데이터 셋 생성 모듈은

상기 노이즈 적용 문장의 단어 개수에 따라 상기 노이즈 적용 문장 및 상기 후보 뉴스 데이터의 문장 중 특정 문장으로 교환하거나 노이즈 적용 문장의 문장 성분의 단어를 상기 후보 뉴스 데이터의 문장 중 특정 문장의 단어로 교환하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 장치.

청구항 5

제4항에 있어서,

상기 가짜 뉴스 탐지용 데이터 셋 생성 모듈은

상기 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수 이상이면, 노이즈 적용 문장을 후보 뉴스 데이터의 모든 문장 중 가장 평균 유사도가 가장 높은 문장으로 교환하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 장치.

청구항 6

제4항에 있어서,

상기 가짜 뉴스 탐지용 데이터 셋 생성 모듈은

상기 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수 이하이면, 후보 뉴스 데이터의 모든 문장 중 노이즈 적용 문장과 유사한 문장을 교환 대상 문장으로 선택하고, 노이즈 적용 문장의 문장 성분 중 특정 문장 성분의 단어 및 교환 대상 문장에서 동일한 문장 성분을 가지지만 일치하지 않은 단어로 교환하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 장치.

청구항 7

가짜 뉴스 탐지용 데이터 셋 생성 장치에서 실행되는 가짜 뉴스 탐지용 데이터 셋 생성 방법에 있어서,

뉴스 데이터를 수신하면 뉴스 데이터를 분석하여 특정 키워드 태그에 해당하는 데이터를 파싱하는 단계;

상기파싱된 뉴스 데이터에서 특정 태그에 해당하는 데이터를 삭제하여 전처리를 실행하는 단계;

상기 데이터 전처리부에 의해 전처리된 뉴스 데이터 중 일부 뉴스 데이터 중 특정 태그를 동일하게 가지는 뉴스 데이터를 추출하고, 상기 특정 뉴스 데이터 중 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고 나머지 뉴스 데이터 중 일부 뉴스 데이터를 후보 뉴스 데이터로 결정하는 단계;

상기 기준 뉴스 데이터 및 상기 후보 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하는 단계;

상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 단계를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 방법.

청구항 8

제7항에 있어서,

상기 전처리된 뉴스 데이터를 특정 비율로 나눈 후 그룹핑하여 제1 뉴스 데이터 그룹 및 제2 뉴스 데이터 그룹을 생성하고, 상기 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 각각에 진짜 뉴스 레이블을 할당하는 단계;

상기 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고, 나머지 뉴스 데이터 중 특정 개수의 후보 뉴스 데이터를 추출하는 단계;

상기 기준 뉴스 데이터의 본문에서 특정 개수의 문장을 노이즈 적용 문장으로 선택하는 단계;

상기 노이즈 적용 문장을 상기 후보 뉴스 데이터의 문장 중 특정 문장으로 교환하여 가짜 뉴스 데이터 셋을 생성하는 단계; 및

상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 단계를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 방법.

청구항 9

제8항에 있어서,

상기 특정 개수의 후보 뉴스 데이터를 추출하는 단계는

상기 기준 뉴스 데이터의 문장 개수에 따라 후보 뉴스 데이터를 추출하는 특정 개수를 결정하는 단계; 및

상기 기준 뉴스 데이터를 송고 날짜를 기준으로 특정 일 이전 및 이후에 작성된 뉴스 데이터를 상기 특정 개수의 후보 뉴스 데이터의 선택 시 제외하여 후보 뉴스 데이터를 추출하는 단계를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 방법.

청구항 10

제8항에 있어서,

상기 노이즈 적용 문장으로 선택하는 단계는

상기 기준 뉴스 데이터 및 특정 개수(N)의 후보 뉴스 데이터 각각의 본문의 각 문장 사이의 유사도에 따라 상기 기준 뉴스 데이터의 본문에서 특정 개수의 노이즈 문장을 선택하는 단계를 포함하고,

상기 가짜 뉴스 탐지용 데이터 셋을 생성하는 단계는

상기 노이즈 적용 문장의 단어 개수에 따라 상기 노이즈 적용 문장 및 상기 후보 뉴스 데이터의 문장 중 특정 문장으로 교환하거나 노이즈 적용 문장의 문장 성분의 단어를 상기 후보 뉴스 데이터의 문장 중 특정 문장의 단어로 교환하는 단계를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 방법.

청구항 11

제10항에 있어서,

상기 가짜 뉴스 탐지용 데이터 셋을 생성하는 단계는

상기 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수 이상이면, 노이즈 적용 문장을 후보 뉴스 데이터의 모든 문장 중 가장 평균 유사도가 가장 높은 문장으로 교환하는 단계를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 방법.

청구항 12

제10항에 있어서,

상기 가짜 뉴스 탐지용 데이터 셋을 생성하는 단계는

상기 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수 이하이면, 후보 뉴스 데이터의 모든 문장 중 노이즈 적용 문장과 유사한 문장을 교환 대상 문장으로 선택하고, 노이즈 적용 문장의 문장 성분 중 특정 문장 성분의 단어 및 교환 대상 문장에서 동일한 문장 성분을 가지지만 일치하지 않은 단어로 교환하는 단계를 포함하는 것을 특징으로 하는

가짜 뉴스 탐지용 데이터 셋 생성 방법.

발명의 설명

기술 분야

[0001] 본 발명은 가짜 뉴스 탐지용 데이터 셋 생성 장치 및 이의 실행 방법에 관한 것으로, 보다 구체적으로 제목 및 본문이 상이한 가짜뉴스를 탐지할 수 있는 가짜 뉴스 탐지용 데이터 셋 생성 장치 및 이의 실행 방법에 관한 것이다.

배경 기술

[0003] 오늘날 매스 미디어의 발전으로 인해 대중들은 시공간의 제약 없이 다양한 매개체를 통해 아직 알려지지 않은 새로운 소식과 관련 정보들을 접할 수 있게 되었다. 대중들은 많은 정보들 가운데 관심 있는 것들만 주목하여 취사 선택하는데, 수동적 입장에 있는 대중으로서는 주체적 입장에 있는 매스미디어의 보도 내용을 대부분 진실된 정보로 수용할 수 밖에 없게 된다.

[0004] 그러나 사실이 아닌 정보를 포함하는 가짜 뉴스가 유포되어 대중들에게 혼란을 주며 여론과 의사결정 과정을 왜곡하고 선거나 정치 과정의 결과를 바꾸는 등 사회적 문제를 야기하고 있다.

[0005] 이러한 가짜 뉴스는 정교해서 뉴스 사용자들이 진실 여부를 확인하기가 어렵고, 가짜 뉴스가 뉴스 사용자들에게 그대로 전달될 경우 뉴스 사용자들에게 혼란을 줄 뿐만 아니라, 뉴스 사용자들이 이러한 가짜 뉴스를 여과 없이 믿게 되면 사회적인 피해가 발생할 수도 있다.

[0006] 예를 들어, 대중들에게 혼란을 주어 여론과 의사결정 과정을 왜곡하거나 조회수를 올리기 위해 제작된 제목과 본문의 내용이 다른 가짜뉴스는 불필요한 사회적 비용을 야기할 수 있다.

[0007] 이를 해결하기 위해 데이터셋을 활용해 인공지능 모델을 훈련해 가짜뉴스를 자동 탐지하고자 했으나 기존의 제목과 본문이 다른 가짜뉴스 데이터셋은 하나의 뉴스 본문의 특정 문단에 다른 뉴스 본문의 문단을 단순 삽입하는 방식으로 제작되어 2가지 한계점이 존재한다.

[0008] 첫번째, 삽입되는 뉴스의 주제를 고려하지 않고 가짜뉴스를 제작하는 경우 모델이 풀기 쉬운 데이터셋이 제작된다. 예를 들어, 정치 분야 뉴스 본문에 스포츠 분야 뉴스 본문이 삽입될 경우 모델은 뉴스 제목과 본문의 불일치성을 내용적인 측면보다 단순 키워드 기반으로 학습될 확률이 높으며, 이후 모델은 정치 분야 뉴스 본문끼리 제작된 가짜뉴스는 제대로 판별하기 어렵다.

[0009] 두번째, 기존 뉴스 본문 대비 삽입되는 뉴스 본문의 비율이 일정하지 않아 모델의 안정적인 학습이 어렵다. 제작된 가짜뉴스 본문마다 제목과의 불일치 비율이 다를 경우 모델의 분산이 높아져 강건한 판별 성능을 갖춘 모델 제작이 어렵다는 문제가 있다.

[0010] 한국공개특허 10-2228737는 AI를 이용한 가짜뉴스 판독 방법 및 장치에 관한 것으로, 가짜뉴스를 판독하여 증권 정보들 중에서 가짜뉴스와 신뢰할 수 있는 뉴스를 가려낼 수 있도록 한다는 내용이 개시되어 있다. 하지만, 상기의 문제점을 해결하기 위한 방안이 개시되어 있지 않다.

선행기술문헌

특허문헌

[0012] (특허문헌 0001) (선행특허 1)한국공개특허 10-2228737호

발명의 내용

해결하려는 과제

[0013] 본 발명은 제목 및 본문이 상이한 가짜 뉴스를 탐지할 수 있는 가짜 뉴스 탐지용 데이터 셋 생성 장치 및 이의 실행 방법을 제공하는 것을 목적으로 한다.

[0014] 또한, 본 발명은 기존의 하나의 뉴스 본문의 특정 부분에 다른 뉴스 본문의 내용의 일부분을 단순 삽입하는 방식이 아닌 추가적인 제약조건을 활용해 가짜뉴스의 제목과 본문의 불일치성을 일정수준으로 유지할 수 있도록 하는 가짜 뉴스 탐지용 데이터 셋 생성 장치 및 이의 실행 방법을 제공하는 것을 목적으로 한다.

[0015] 또한, 본 발명은 가짜 뉴스 데이터 셋은 모델이 자연어에 대한 이해를 기반으로 제목과 본문의 불일치성을 판별 하도록 훈련시킬 수 있으며, 모델의 분산을 낮게 유지시킬 수 있도록 하는 가짜 뉴스 탐지용 데이터 셋 생성 장치 및 이의 실행 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0017] 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 장치는 데이터 수신부로부터 뉴스 데이터를 수신 하면 뉴스 데이터를 분석하여 특정 키워드 태그에 해당하는 데이터를 과싱하는 데이터 과싱부, 상기 데이터 과싱부에 의해 과싱된 뉴스 데이터에서 특정 태그에 해당하는 데이터를 삭제하여 전처리를 실행하는 데이터 전처리부 및 상기 데이터 전처리부에 의해 전처리된 뉴스 데이터 중 일부 뉴스 데이터 중 특정 태그를 동일하게 가지는 뉴스 데이터를 추출하고, 상기 특정 뉴스 데이터 중 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고 나머지 뉴스 데이터 중 일부 뉴스 데이터를 후보 뉴스 데이터로 결정하고, 상기 기준 뉴스 데이터 및 상기 후보 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하고, 상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 데이터 셋 생성부를 포함한다.

[0018] 또한, 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 장치에서 실행되는 가짜 뉴스 탐지용 데이터 셋 생성 방법은 뉴스 데이터를 수신하면 뉴스 데이터를 분석하여 특정 키워드 태그에 해당하는 데이터를 과싱하는 단계, 상기과싱된 뉴스 데이터에서 특정 태그에 해당하는 데이터를 삭제하여 전처리를 실행하는 단계, 상기 데이터 전처리부에 의해 전처리된 뉴스 데이터 중 일부 뉴스 데이터 중 특정 태그를 동일하게 가지는 뉴스 데이터를 추출하고, 상기 특정 뉴스 데이터 중 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정하고 나머지 뉴스 데이터 중 일부 뉴스 데이터를 후보 뉴스 데이터로 결정하는 단계, 상기 기준 뉴스 데이터 및 상기 후보 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하는 단계 및 상기 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성하는 단계를 포함한다.

발명의 효과

- [0020] 상기와 같은 본 발명에 의하면, 제목 및 본문이 상이한 가짜뉴스를 탐지할 수 있다는 장점이 있다.
- [0021] 또한 본 발명에 의하면, 기존의 하나의 뉴스 본문의 특정 부분에 다른 뉴스 본문의 내용의 일부분을 단순 삽입하는 방식이 아닌 추가적인 제약조건을 활용해 가짜뉴스의 제목과 본문의 불일치성을 일정수준으로 유지할 수 있다는 장점이 있다.
- [0022] 또한 본 발명에 의하면, 가짜 뉴스 데이터 셋은 모델이 자연어에 대한 이해를 기반으로 제목과 본문의 불일치성을 판별하도록 훈련시킬 수 있으며, 모델의 분산을 낮게 유지시킬 수 있다는 장점이 있다.

도면의 간단한 설명

- [0024] 도 1은 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 장치의 내부 구조를 설명하기 위한 블록도이다.
- 도 2는 본 발명의 일 실시예에 따른 데이터 셋 생성부의 내부 구조를 설명하기 위한 블록도이다.
- 도 3은 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 과정을 설명하기 위한 예시도이다.
- 도 4는 본 발명의 일 실시예에 따른 가짜 뉴스 데이터의 생성 과정을 설명하기 위한 예시도이다.
- 도 5는 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 과정을 통해 생성된 가짜 뉴스 데이터를 설명하기 위한 예시도이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 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 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.
- [0027] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 발명에 대해 구체적으로 설명하기로 한다.
- [0028] 본 발명에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 관례, 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.
- [0029] 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.
- [0030] 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.

- [0031] "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.
- [0033] 도 1은 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 장치의 내부 구조를 설명하기 위한 블록도이다.
- [0034] 도 1을 참조하면, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 데이터 수신부(110), 데이터 파싱부(120), 데이터 전처리부(130) 및 데이터 셋 생성부(140)를 포함한다.
- [0035] 데이터 수신부(110)는 외부 서버로부터 뉴스 데이터를 수신하여 데이터 파싱부(120)에 제공한다. 이때, 뉴스 데이터는 텍스트 데이터 및 메타 데이터가 xml 형식으로 저장되어 있는 데이터이다. 텍스트 데이터는 제목, 부제목, 본문, 캡션을 포함하고, 메타 데이터는 송고날짜(send date) 및 시간(send time), 대분류(category), 소분류(classcode), 기자 부서(deskcode), 기자명(writer) 등을 포함한다.
- [0036] 데이터 파싱부(120)는 데이터 수신부(110)로부터 뉴스 데이터를 수신하면, 뉴스 데이터를 분석하여 특정 키워드 태그에 해당하는 데이터를 파싱한다.
- [0037] 일 실시예에서, 데이터 파싱부(120)는 뉴스 데이터의 xml을 분석하여 제목(<Title>) 태그, 본문(<Textbody>) 태그, 카테고리(<Category>) 태그 및 소분류(<Classcode>) 태그 각각에 해당하는 데이터를 파싱한다. 이때, 제목 태그에 해당하는 데이터는 뉴스의 제목을 의미하고, 본문 태그에 해당하는 데이터는 뉴스의 본문을 의미하고, 카테고리 태그에 해당하는 데이터는 카테고리를 의미하고, 소분류 태그에 해당하는 데이터는 세부 카테고리를 의미한다. 예를 들어, 카테고리 태그 “정치”에 해당하는 뉴스 데이터라도 소분류 태그에 따라 “청와대일반”, “국회일반”, “청문회” 등으로 세부화가 가능하다.
- [0038] 상기의 실시예에서, 데이터 파싱부(120)는 카테고리 태그 및 소분류 태그 각각에 해당하는 데이터를 파싱하는 이유는 소분류 태그에 해당하는 데이터가 같은 기사를 이용하여 제목 및 본문이 상이한 가짜 뉴스 탐지용 데이터 셋을 생성하기 위해서이다. 가짜 뉴스 탐지용 데이터 셋을 생성하는 과정은 이하의 데이터 셋 생성부(140)에서 보다 구체적으로 설명하기로 한다.
- [0039] 데이터 전처리부(130)는 데이터 파싱부(120)에 의해 파싱된 뉴스 데이터에서 특정 태그에 해당하는 데이터를 삭제하여 전처리를 실행한다. 이와 같은 이유는, 데이터 파싱부(120)에 의해 파싱된 뉴스 데이터는 텍스트 데이터 내의 제어 문자, 특수 문자, 이메일 주소 등과 같은 노이즈 데이터를 포함하며, 이러한 노이즈 데이터는 최종 데이터 셋의 품질에 영향을 줄 수 있기 때문이다.
- [0040] 데이터 셋 생성부(140)는 데이터 전처리부(130)에 의해 전처리된 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성한다.
- [0041] 먼저, 데이터 셋 생성부(140)는 데이터 전처리부(130)에 의해 전처리된 뉴스 데이터를 특정 비율로 나눈 후 그룹핑하여 제1 뉴스 데이터 그룹 및 제2 뉴스 데이터 그룹을 생성한다. 이때, 제1 뉴스 데이터 그룹은 진짜 뉴스 데이터로 활용되는 뉴스 데이터가 포함되어 있는 그룹이라고, 제2 뉴스 데이터 그룹은 제목 및 본문이 상이한 가짜 뉴스 데이터 셋의 제작에 활용되는 뉴스 데이터가 포함되어 있는 그룹이다.
- [0042] 일 실시예에서, 데이터 셋 생성부(140)는 더 많은 비율의 뉴스 데이터를 제2 뉴스 데이터 그룹으로 그룹핑한다. 이와 같은 이유는 제목 및 본문이 상이한 가짜 뉴스 데이터 셋의 제작 시 데이터 손실이 일어날 수 있기 때문에 더 많은 비율의 뉴스 데이터를 제2 뉴스 데이터 그룹으로 그룹핑한다.
- [0043] 예를 들어, 데이터 셋 생성부(140)는 데이터 전처리부(130)에 의해 전처리된 뉴스 데이터의 전체를 35:65의 비율로 나눈 후 35%에 해당하는 뉴스 데이터를 제1 뉴스 데이터 그룹으로 그룹핑하고, 65%에 해당하는 뉴스 데이터를 제2 뉴스 데이터 그룹으로 그룹핑할 수 있다.
- [0044] 그런 다음, 데이터 셋 생성부(140)는 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터에 대한 별도의 처리를 진행하지 않고 뉴스 데이터 각각에 진짜 뉴스 레이블을 할당하며, 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터를 이용하여 제목 및 본문이 상이한 가짜 뉴스 데이터 셋을 생성한다.

- [0045] 보다 구체적으로, 데이터 셋 생성부(140)는 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 뉴스 데이터를 추출한 후, 추출된 동일한 소분류 태그를 가지는 뉴스 데이터 중 특정 조건에 해당하는 뉴스 데이터를 추출한다.
- [0046] 먼저, 데이터 셋 생성부(140)는 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터를 추출한다. 이와 같은 이유는 대분류 태그를 기준으로 분류되는 뉴스 데이터를 소분류 태그를 기준으로 나누는 경우 동일한 소분류 태그를 가지는 뉴스 데이터의 본문 간의 내용 교환이 이루어질 수 있어 보다 어려운 데이터 셋을 생성할 수 있기 때문이다.
- [0047] 예를 들어, 데이터 셋 생성부(140)는 대분류 태그 “생활” 및 “건강”으로 분류되는 뉴스 데이터를 소분류 태그를 기준으로 나누는 경우, “놀이시설공원”, “레이싱”, “바둑” 등으로 분류되기 때문에 동일한 소분류 태그를 가진 뉴스 데이터의 본문 간의 내용 교환이 이루어지는 경우 보다 어려운 데이터 셋을 생성할 수 있다.
- [0048] 그 후, 데이터 셋 생성부(140)는 동일한 소분류 태그를 가지는 어느 하나의 기준 뉴스 데이터를 제외한 나머지 뉴스 데이터 중 특정 조건에 해당하는 특정 개수(N)의 뉴스 데이터를 추출한다. 이와 같은 이유는, 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터 및 특정 조건에 해당하는 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하기 위해서이다.
- [0049] 일 실시예에서, 데이터 셋 생성부(140)는 기준 뉴스 데이터의 문장 개수에 따라 후보 뉴스 데이터 개수(N)를 선형적으로 증가시킨다. 즉, 데이터 셋 생성부(140)는 [수학식 1]을 기초로 산출된 기준 뉴스 데이터의 문장 개수에 따라 후보 뉴스 데이터 개수(N)를 선형적으로 증가시킨다. 예를 들어, 기준 뉴스 데이터의 문장 개수가 10개인 경우 후보 뉴스 데이터 개수(N)는 5개로 결정하고, 기준 뉴스 데이터의 문장 개수가 15개인 경우 후보 뉴스 데이터 개수(N)는 8개로 결정할 수 있다.
- [0051] [수학식 1]
- [0052]
$$N = \text{round}(M \times 0.5)$$
- [0054] N: 후보 뉴스 데이터의 선택 개수
- [0055] M: 문장 개수,
- [0056] $\text{round}()$: $M \times 0.5$ 의 값을 소수점 첫째자리에서 반올림하여 정수를 반환하는 함수
- [0058] 다른 일 실시예에서, 데이터 셋 생성부(140)는 동일한 소분류 태그를 가지는 뉴스 데이터 중 어느 하나의 기준 뉴스 데이터의 송고 날짜를 기준으로 특정 일 이전 및 이후에 작성된 뉴스 데이터를 특정 개수의 후보 뉴스 데이터의 선택 시 제외한다. 이와 같은 이유는 동일한 소분류 태그를 가지는 뉴스 데이터 중 동일한 송고 날짜를 가진 뉴스 데이터들은 초안 기사와 정정 기사 사이의 본문 내용 교환이 일어나 위음성 데이터를 만들 가능성이 존재하기 때문이다.
- [0059] 예를 들어, 카테고리 태그 “정치”, “사회” 및 “문화”에 해당하는 뉴스 데이터의 경우 1주일 이내에 동일한 사건에 대해 상반된 견해를 작성한 뉴스 데이터가 작성될 수 있기 때문에 기준 뉴스 데이터의 송고 날짜를 기준으로 7일 이전 및 7일 이후에 작성된 뉴스 데이터를 후보 뉴스 데이터에서 제외한다.
- [0060] 상기와 같이, 데이터 셋 생성부(140)는 기준 뉴스 데이터와 동일한 소분류 태그를 가지는 뉴스 데이터 중 1-gram, 2-gram, 3-gram keyword overlap의 가중 평균이 가장 높은 특정 개수(N)의 후보 뉴스 데이터를 선택한다. 이때, 가중 평균은 $0.6 \times 1\text{-gram overlap} + 0.4 \times 2\text{-gram overlap} + 0.3 \times 3\text{-gram overlap}$ 등이될 수 있다.
- [0061] 그런 다음, 데이터 셋 생성부(140)는 기준 뉴스 데이터를 가짜 뉴스 데이터로 변경하기 위해서 기준 뉴스 데이터 및 특정 개수(N)의 후보 뉴스 데이터 각각의 본문의 각 문장에 대해서 1-gram, 2-gram, 3-gram, 5-gram 집합과 KoBERT [CLS] Embedding 벡터를 계산한다. 이때, N-gram은 기준 뉴스 데이터 및 후보 뉴스 데이터 각각의 두 문장 간의 통사적 유사도 측정을 위해, KoBERT [CLS] Embedding은 두 문장 간의 맥락적 유사도 측정을 위해 활용된다.

- [0062] 일 실시예에서, 데이터 셋 생성부(140)는 기준 뉴스 데이터의 본문에서 특정 개수(k)의 노이즈 적용 문장을 규칙적으로 선택한다. 이때, 노이즈 적용 문장 개수(k)는 기준 뉴스 데이터 및 후보 뉴스 데이터 각각의 본문의 문장을 교환하는 교환 비율에 따라 달라지나 편향되지 않은 가짜 뉴스 탐지용 데이터 셋을 생성하기 위해 [수학식 1]에 의해 산출된 후보 뉴스 데이터 개수(N)과 동일하게 설정된다.
- [0063] 상기의 실시예에서, 데이터 셋 생성부(140)는 기준 뉴스 데이터의 본문의 문장에 대해서 후보 뉴스 데이터 각각의 본문의 문장 사이의 1-gram, 2-gram, 3-gram, KoBERT [CLS] Embedding의 가중평균 유사도를 계산한다. 이때, N-gram Overlap은 matching 비율, KoBERT [CLS] Embedding의 경우에는 코사인 유사도를 활용한다.
- [0064] 상기와 같이, 데이터 셋 생성부(140)는 기준 뉴스 데이터의 본문의 문장에 대해서 후보 뉴스 데이터 각각의 본문의 문장 사이의 유사도를 산출한 후, 유사도가 높은 문장 중 특정 개수(k)의 문장을 기준 뉴스 데이터의 본문에 노이즈로 적용될 노이즈 적용 문장으로 선택한다.
- [0065] 이를 통해, 기준 뉴스 데이터의 본문의 문장은 후보 뉴스 데이터 개수(N)의 후보 뉴스 데이터와 가장 유사한 문장과의 교환이 일어나 보다 어려운 가짜 뉴스 데이터의 제작이 가능하다.
- [0066] 그 후, 데이터 셋 생성부(140)는 기준 뉴스 데이터에서 선택된 특정 개수(k)의 노이즈 적용 문장에 대해서 노이즈를 추가한다.
- [0067] 일 실시예에서, 데이터 셋 생성부(140)는 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수(예를 들어, 7개) 이상이면, 노이즈 적용 문장을 후보 뉴스 데이터의 모든 문장 중 가장 평균 유사도가 가장 높은 문장으로 교환한다. 이때, 노이즈 적용 문장이 아래의 조건을 만족하는 경우 그 다음으로 가장 평균 유사도가 높은 문장으로 교환한다.
- [0068] 다른 일 실시예에서, 데이터 셋 생성부(140)는 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수(예를 들어, 7개) 이하이면, 파서를 통해 노이즈 적용 문장을 파싱하여 문장 성분으로 분해하고, 후보 뉴스 데이터의 모든 문장 중 노이즈 적용 문장과 유사한 문장을 교환 대상 문장으로 선택하고, 선택된 교환 대상 문장을 파싱하여 문장 성분으로 분해한다. 그런 다음, 데이터 셋 생성부(140)는 노이즈 적용 문장의 문장 성분 중 특정 문장 성분(예를 들어, 랜덤하게 선택된 문장 성분)의 단어 및 교환 대상 문장에서 동일한 문장 성분을 가지지만 일치하지 않은 단어를 교환한다.
- [0069] 그런 다음, 데이터 셋 생성부(140)는 기준 뉴스 데이터에서 선택된 특정 개수(k)의 노이즈 적용 문장에 노이즈가 적용된 후 기준 뉴스 데이터에 가짜 뉴스 레이블을 할당한다.
- [0070] 최종적으로, 데이터 셋 생성부(140)는 가짜 뉴스 레이블이 할당된 기준 뉴스 데이터 및 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 셋을 동일한 비율이 되도록 처리해 가짜 뉴스 탐지용 데이터 셋을 생성한다.
- [0071] 상기의 실시예와는 달리, 데이터 셋 생성부(140)는 동일한 소분류 태그를 가지는 뉴스 데이터 중 제1 뉴스 데이터 및 제2 뉴스 데이터 각각의 본문의 문장 개수의 차이가 특정 개수 이하에 해당하는 뉴스 데이터를 추출한다. 이와 같은 이유는 제목 및 본문이 상이한 가짜 뉴스의 제목 및 본문의 불일치성을 40%~60%로 유지하기 위해서이다.
- [0072] 그런 다음, 데이터 셋 생성부(140)는 동일한 소분류 태그를 가지는 뉴스 데이터 중 특정 조건에 해당하는 뉴스 데이터가 추출되면, 추출된 뉴스 데이터 중 두 개의 뉴스 데이터를 추출하여 본문의 문장 개수가 더 많은 뉴스 데이터를 제1 뉴스 데이터로 결정하고 더 적은 뉴스 데이터를 제2 뉴스 데이터로 결정하며, 제1 뉴스 데이터의 본문 문장을 한 문장씩 반복적으로 순환하면서 특정 확률(40%~60%)로 제2 뉴스 데이터의 본문의 문장으로 교체하여 가짜 뉴스 데이터 셋을 생성한다.
- [0073] 상기와 같이, 데이터 셋 생성부(140)는 제1 뉴스 데이터의 본문이 기존의 뉴스 데이터의 본문 대비 특정 확률(40%~60%) 만큼 내용이 변동되었으면, 제1 뉴스 데이터를 가짜 뉴스 탐지용 데이터 셋에 추가하며, 제1 뉴스 데이터의 본문이 기존의 뉴스 데이터의 본문 대비 특정 확률 이하로 변경되었으면 기존의 뉴스 데이터의 본문 대비 특정 확률(40%~60%) 이상만큼 내용이 변동될 때까지 제1 뉴스 데이터의 본문 문장을 한 문장씩 반복적으로 순환하면서 제2 뉴스 데이터의 본문의 문장으로 교체하여 가짜 뉴스 데이터 셋을 생성한다.
- [0074] 최종적으로, 데이터 셋 생성부(140)는 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 셋 및 가짜 뉴스 데이터 셋을 동일한 비율이 되도록 처리해 가짜 뉴스 탐지용 데이터 셋을 생성한다.

- [0076] 도 2는 본 발명의 일 실시예에 따른 데이터 셋 생성부의 내부 구조를 설명하기 위한 블록도이다.
- [0077] 도 2를 참조하면, 데이터 셋 생성부(140)는 데이터 그룹 생성 모듈(141), 데이터 추출 모듈(142), 노이즈 적용 문장 선택 모듈(143), 가짜 뉴스 데이터 셋 생성 모듈(144) 및 가짜 뉴스 탐지용 데이터 셋 생성 모듈(145)을 포함한다.
- [0078] 데이터 그룹 생성 모듈(141)은 전처리된 뉴스 데이터를 특정 비율로 나눈 후 그룹핑하여 제1 뉴스 데이터 그룹 및 제2 뉴스 데이터 그룹을 생성하고, 상기 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 각각에 진짜 뉴스 레이블을 할당한다.
- [0079] 먼저, 데이터 그룹 생성 모듈(141)은 데이터 전처리부(130)에 의해 전처리된 뉴스 데이터를 특정 비율로 나눈 후 그룹핑하여 제1 뉴스 데이터 그룹 및 제2 뉴스 데이터 그룹을 생성한다. 이때, 제1 뉴스 데이터 그룹은 진짜 뉴스 데이터로 활용되는 뉴스 데이터가 포함되어 있는 그룹이라고, 제2 뉴스 데이터 그룹은 제목 및 본문이 상이한 가짜 뉴스 데이터 셋의 제작에 활용되는 뉴스 데이터가 포함되어 있는 그룹이다.
- [0080] 일 실시예에서, 데이터 그룹 생성 모듈(141)은 전처리된 뉴스 데이터를 특정 비율로 나눈 후 더 많은 비율의 뉴스 데이터를 제2 뉴스 데이터 그룹으로 그룹핑한다. 이와 같은 이유는 제목 및 본문이 상이한 가짜 뉴스 데이터 셋의 제작 시 데이터 손실이 일어날 수 있기 때문에 더 많은 비율의 뉴스 데이터를 제2 뉴스 데이터 그룹으로 그룹핑한다.
- [0081] 그런 다음, 데이터 그룹 생성 모듈(141)은 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터에 대한 별도의 처리를 진행하지 않고 뉴스 데이터 각각에 진짜 뉴스 레이블을 할당하며, 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터를 이용하여 제목 및 본문이 상이한 가짜 뉴스 데이터 셋을 생성한다.
- [0082] 데이터 추출 모듈(142)은 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 뉴스 데이터를 추출한 후, 추출된 동일한 소분류 태그를 가지는 뉴스 데이터 중 특정 조건에 해당하는 뉴스 데이터를 추출한다.
- [0083] 먼저, 데이터 추출 모듈(142)은 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터를 추출한다. 이와 같은 이유는 대분류 태그를 기준으로 분류되는 뉴스 데이터를 소분류 태그를 기준으로 나누는 경우 동일한 소분류 태그를 가지는 뉴스 데이터의 본문 간의 내용 교환이 이루어질 수 있어 보다 어려운 데이터 셋을 생성할 수 있기 때문이다.
- [0084] 그 후, 데이터 추출 모듈(142)은 동일한 소분류 태그를 가지는 어느 하나의 기준 뉴스 데이터를 제외한 나머지 뉴스 데이터 중 특정 조건에 해당하는 특정 개수(N)의 뉴스 데이터를 추출한다. 이와 같은 이유는, 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터 및 특정 조건에 해당하는 뉴스 데이터를 이용하여 가짜 뉴스 데이터 셋을 생성하기 위해서이다.
- [0085] 일 실시예에서, 데이터 추출 모듈(142)은 기준 뉴스 데이터의 문장 개수에 따라 후보 뉴스 데이터 개수(N)를 선형적으로 증가시킨다.
- [0086] 다른 일 실시예에서, 데이터 추출 모듈(142)은 동일한 소분류 태그를 가지는 뉴스 데이터 중 어느 하나의 기준 뉴스 데이터의 송고 날짜를 기준으로 특정 일 이전 및 이후에 작성된 뉴스 데이터를 특정 개수의 후보 뉴스 데이터의 선택 시 제외한다. 이와 같은 이유는 동일한 소분류 태그를 가지는 뉴스 데이터 중 동일한 송고 날짜를 가진 뉴스 데이터들은 초안 기사와 정정 기사 사이의 본문 내용 교환이 일어나 위음성 데이터를 만들 가능성이 존재하기 때문이다.
- [0087] 상기과 같이, 데이터 추출 모듈(142)은 기준 뉴스 데이터와 동일한 소분류 태그를 가지는 뉴스 데이터 중 1-gram, 2-gram, 3-gram keyword overlap의 가중 평균이 가장 높은 특정 개수(N)의 후보 뉴스 데이터를 선택한다. 이때, 가중 평균은 $0.6 \times 1\text{-gram overlap} + 0.4 \times 2\text{-gram overlap} + 0.3 \times 3\text{-gram overlap}$ 등이 될 수 있다.
- [0088] 그런 다음, 데이터 추출 모듈(142)은 기준 뉴스 데이터를 가짜 뉴스 데이터로 변경하기 위해서 기준 뉴스 데이터 및 특정 개수(N)의 후보 뉴스 데이터 각각의 본문의 각 문장에 대해서 1-gram, 2-gram, 3-gram, 5-gram 집합과 KoBERT [CLS] Embedding 벡터를 계산한다. 이때, N-gram은 기준 뉴스 데이터 및 후보 뉴스 데이터 각각의 두 문장 간의 통사적 유사도 측정을 위해, KoBERT [CLS] Embedding은 두 문장 간의 맥락적 유사도 측정을 위해 활용된다.

- [0089] 노이즈 적용 문장 선택 모듈(143)은 기준 뉴스 데이터의 본문에서 특정 개수(k)의 노이즈 적용 문장을 규칙적으로 선택한다. 이때, 노이즈 적용 문장 개수(k)는 기준 뉴스 데이터 및 후보 뉴스 데이터 각각의 본문의 문장을 교환하는 교환 비율에 따라 달라지나 편향되지 않은 가짜 뉴스 탐지용 데이터 셋을 생성하기 위해 [수학식 1]에 의해 산출된 후보 뉴스 데이터 개수(N)과 동일하게 설정된다.
- [0090] 상기의 실시예에서, 노이즈 적용 문장 선택 모듈(143)은 기준 뉴스 데이터의 본문의 문장에 대해서 후보 뉴스 데이터 각각의 본문의 문장 사이의 1-gram, 2-gram, 3-gram, KoBERT [CLS] Embedding의 가중평균 유사도를 계산한다. 이때, N-gram Overlap은 matching 비율, KoBERT [CLS] Embedding의 경우에는 코사인 유사도를 활용한다.
- [0091] 상기와 같이, 노이즈 적용 문장 선택 모듈(143)은 기준 뉴스 데이터의 본문의 문장에 대해서 후보 뉴스 데이터 각각의 본문의 문장 사이의 유사도를 산출한 후, 유사도가 높은 문장 중 특정 개수(k)의 문장을 기준 뉴스 데이터의 본문에 노이즈로 적용될 노이즈 적용 문장으로 선택한다.
- [0092] 이를 통해, 기준 뉴스 데이터의 본문의 문장은 후보 뉴스 데이터 개수(N)의 후보 뉴스 데이터와 가장 유사한 문장과의 교환이 일어나 보다 어려운 가짜 뉴스 데이터의 제작이 가능하다.
- [0093] 그 후, 노이즈 적용 문장 선택 모듈(143)은 기준 뉴스 데이터에서 선택된 특정 개수(k)의 노이즈 적용 문장에 대해서 노이즈를 추가한다.
- [0094] 일 실시예에서, 노이즈 적용 문장 선택 모듈(143)은 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수(예를 들어, 7개) 이상이면, 노이즈 적용 문장을 후보 뉴스 데이터의 모든 문장 중 가중 평균 유사도가 가장 높은 문장으로 교환한다. 이때, 노이즈 적용 문장이 아래의 조건을 만족하는 경우 그 다음으로 가중 평균 유사도가 높은 문장으로 교환한다.
- [0095] 다른 일 실시예에서, 노이즈 적용 문장 선택 모듈(143)은 기준 뉴스 데이터에서 선택된 노이즈 적용 문장의 단어 개수가 특정 개수(예를 들어, 7개) 이하이면, 파서를 통해 노이즈 적용 문장을 파싱하여 문장 성분으로 분해하고, 후보 뉴스 데이터의 모든 문장 중 노이즈 적용 문장과 유사한 문장을 교환 대상 문장으로 선택하고, 선택된 교환 대상 문장을 파싱하여 문장 성분으로 분해한다. 그런 다음, 노이즈 적용 문장 선택 모듈(143)은 노이즈 적용 문장의 문장 성분 중 특정 문장 성분(예를 들어, 랜덤하게 선택된 문장 성분)의 단어 및 교환 대상 문장에서 동일한 문장 성분을 가지지만 일치하지 않은 단어를 교환한다.
- [0096] 가짜 뉴스 데이터 셋 생성 모듈(144)은 노이즈 적용 문장을 상기 후보 뉴스 데이터의 문장 중 특정 문장으로 교환하여 가짜 뉴스 데이터 셋을 생성한다.
- [0097] 먼저, 가짜 뉴스 데이터 셋 생성 모듈(144)은 기준 뉴스 데이터에서 선택된 특정 개수(k)의 노이즈 적용 문장에 노이즈가 적용된 후 기준 뉴스 데이터에 가짜 뉴스 레이블을 할당한다. 최종적으로, 가짜 뉴스 데이터 셋 생성 모듈(144)은 가짜 뉴스 레이블이 할당된 기준 뉴스 데이터 및 제1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 셋을 동일한 비율이 되도록 처리해 가짜 뉴스 탐지용 데이터 셋을 생성한다.
- [0098] 상기의 실시예와는 달리, 가짜 뉴스 데이터 셋 생성 모듈(144)은 동일한 소분류 태그를 가지는 뉴스 데이터 중 제1 뉴스 데이터 및 제2 뉴스 데이터 각각의 본문의 문장 개수의 차이가 특정 개수 이하에 해당하는 뉴스 데이터를 추출한다. 이와 같은 이유는 제목 및 본문이 상이한 가짜 뉴스의 제목 및 본문의 불일치성을 40%~60%로 유지하기 위해서이다.
- [0099] 그런 다음, 가짜 뉴스 데이터 셋 생성 모듈(144)은 동일한 소분류 태그를 가지는 뉴스 데이터 중 특정 조건에 해당하는 뉴스 데이터가 추출되면, 추출된 뉴스 데이터 중 두 개의 뉴스 데이터를 추출하여 본문의 문장 개수가 더 많은 뉴스 데이터를 제1 뉴스 데이터로 결정하고 더 적은 뉴스 데이터를 제2 뉴스 데이터로 결정하며, 제1 뉴스 데이터의 본문 문장을 한 문장씩 반복적으로 순환하면서 특정 확률(40%~60%)로 제2 뉴스 데이터의 본문의 문장으로 교체하여 가짜 뉴스 데이터 셋을 생성한다.
- [0100] 상기와 같이, 가짜 뉴스 데이터 셋 생성 모듈(144)은 제1 뉴스 데이터의 본문이 기존의 뉴스 데이터의 본문 대비 특정 확률(40%~60%) 만큼 내용이 변동되었으면, 제1 뉴스 데이터를 가짜 뉴스 탐지용 데이터 셋에 추가하며, 제1 뉴스 데이터의 본문이 기존의 뉴스 데이터의 본문 대비 특정 확률 이하로 변경되었으면 기존의 뉴스 데이터의 본문 대비 특정 확률(40%~60%) 이상만큼 내용이 변동될 때까지 제1 뉴스 데이터의 본문 문장을 한 문장씩 반복적으로 순환하면서 제2 뉴스 데이터의 본문의 문장으로 교체하여 가짜 뉴스 데이터 셋을 생성한다.
- [0101] 가짜 뉴스 탐지용 데이터 셋 생성 모듈(145)은 가짜 뉴스 데이터 셋 및 상기 전처리된 뉴스 데이터 중 나머지

뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성한다.

- [0103] 도 3은 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 과정을 설명하기 위한 예시도이다.
- [0104] 도 3을 참조하면, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 데이터베이스()에 저장되어 있는 뉴스 데이터를 이용하여 가짜 뉴스 탐지용 데이터 셋을 생성한다.
- [0105] 먼저, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 데이터베이스()에 저장되어 있는 뉴스 데이터를 35:65의 비율로 나눈 후 35%에 해당하는 뉴스 데이터를 제1 뉴스 데이터 그룹으로 그룹핑하고, 65%에 해당하는 뉴스 데이터를 제2 뉴스 데이터 그룹으로 그룹핑할 수 있다. 이때, 제1 뉴스 데이터 그룹은 진짜 뉴스 데이터로 활용되는 뉴스 데이터가 포함되어 있는 그룹이라고, 제2 뉴스 데이터 그룹은 제목 및 본문이 상이한 가짜 뉴스 데이터 셋의 제작에 활용되는 뉴스 데이터가 포함되어 있는 그룹이다.
- [0106] 그 후, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 뉴스 데이터를 추출한 후, 추출된 동일한 소분류 태그를 가지는 뉴스 데이터 중 특정 조건에 해당하는 뉴스 데이터를 추출한다.
- [0107] 먼저, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 동일한 소분류 태그를 가지는 어느 하나의 뉴스 데이터를 기준 뉴스 데이터로 결정한다. 즉, 제2 뉴스 데이터 그룹에 해당하는 뉴스 데이터 중 A 뉴스 데이터를 기준 뉴스 데이터로 결정한다.
- [0108] 그 후, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)의 문장 개수에 따라 후보 뉴스 데이터 개수(N)를 선형적으로 증가시킨다. 즉, 이때, 후보 뉴스 데이터는 B 뉴스, C 뉴스, ..., F 뉴스, G 뉴스 등을 포함할 수 있다.
- [0109] 그 후, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)의 본문에서 특정 개수(k)의 노이즈 적용 문장을 규칙적으로 선택한다.
- [0110] 일 실시예에서, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)의 본문의 문장에 대해서 후보 뉴스 데이터(B 뉴스, C 뉴스, ..., F 뉴스, G 뉴스) 각각의 본문의 문장 사이의 유사도를 산출한 후, 유사도가 높은 문장 중 특정 개수(k)의 문장을 기준 뉴스 데이터의 본문에 노이즈로 적용될 노이즈 적용 문장으로 선택한다.
- [0111] 이를 통해, 기준 뉴스 데이터(A 뉴스)의 본문의 문장은 후보 뉴스 데이터 개수(N)의 후보 뉴스 데이터와 가장 유사한 문장과의 교환이 일어나 보다 어려운 가짜 뉴스 데이터의 제작이 가능하다.
- [0112] 그 후, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)에서 선택된 특정 개수(k)의 노이즈 적용 문장(문장 1, 문장 3, 문장 6, 문장 8, 문장 12)에 대해서 노이즈를 추가한다.
- [0113] 일 실시예에서, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)에서 선택된 노이즈 적용 문장(문장 1, 문장 3, 문장 6, 문장 8, 문장 12)의 단어 개수가 특정 개수(예를 들어, 7개) 이상이면, 노이즈 적용 문장(문장 1, 문장 3, 문장 6, 문장 8, 문장 12)을 후보 뉴스 데이터(B 뉴스, C 뉴스, ..., F 뉴스, G 뉴스)의 모든 문장 중 가중 평균 유사도가 가장 높은 문장으로 교환한다. 이때, 노이즈 적용 문장이 아래의 조건을 만족하는 경우 그 다음으로 가중 평균 유사도가 높은 문장으로 교환한다.
- [0114] 다른 일 실시예에서, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)에서 선택된 노이즈 적용 문장(문장 1, 문장 3, 문장 6, 문장 8, 문장 12)의 단어 개수가 특정 개수(예를 들어, 7개) 이하이면, 파서를 통해 노이즈 적용 문장을 파싱하여 문장 성분으로 분해하고, 후보 뉴스 데이터(B 뉴스, C 뉴스, ..., F 뉴스, G 뉴스)의 모든 문장 중 노이즈 적용 문장(문장 1, 문장 3, 문장 6, 문장 8, 문장 12)과 유사한 문장을 교환 대상 문장으로 선택하고, 선택된 교환 대상 문장을 파싱하여 문장 성분으로 분해한다. 그런 다음, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 노이즈 적용 문장(문장 1, 문장 3, 문장 6, 문장 8, 문장 12)의 문장 성분 중 특정 문장 성분(예를 들어, 랜덤하게 선택된 문장 성분)의 단어 및 교환 대상 문장에서 동일한 문장 성분을 가지지만 일치하지 않은 단어를 교환한다.
- [0115] 그런 다음, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터에서 선택된 특정 개수(k)의 노이즈 적용 문장에 노이즈가 적용된 후 기준 뉴스 데이터에 가짜 뉴스 레이블을 할당한다.

- [0116] 최종적으로, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 가짜 뉴스 레이블이 할당된 기준 뉴스 데이터 및 제 1 뉴스 데이터 그룹에 해당하는 뉴스 데이터 셋을 동일한 비율이 되도록 처리해 가짜 뉴스 탐지용 데이터 셋을 생성한다.
- [0118] 도 4는 본 발명의 일 실시예에 따른 가짜 뉴스 데이터의 생성 과정을 설명하기 위한 예시도이다.
- [0119] 도 4를 참조하면, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스) 및 후보 뉴스 데이터(B 뉴스)를 이용하여 가짜 뉴스 데이터를 생성한다. 이때, 기준 뉴스 데이터(A 뉴스) 및 후보 뉴스 데이터(B 뉴스)는 소분류가 “레이싱”으로 동일하며 송고 날짜가 상이하다.
- [0120] 기준 뉴스 데이터(A 뉴스)에서 선택된 특정 개수(k)의 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)에 대해서 노이즈를 추가한다.
- [0121] 일 실시예에서, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)에서 선택된 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)의 단어 개수가 특정 개수(예를 들어, 7개) 이상이면, 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)을 후보 뉴스 데이터(B 뉴스)의 모든 문장 중 가장 평균 유사도가 가장 높은 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)로 교환한다. 이때, 노이즈 적용 문장이 아래의 조건을 만족하는 경우 그 다음으로 가장 평균 유사도가 높은 문장으로 교환한다.
- [0122] 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 기준 뉴스 데이터(A 뉴스)에서 선택된 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)의 단어 개수가 특정 개수(예를 들어, 7개) 이하이면, 파서를 통해 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)을 파싱하여 문장 성분으로 분해하고, 후보 뉴스 데이터(B 뉴스)의 모든 문장 중 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)과 유사한 문장을 교환 대상 문장으로 선택하고, 선택된 교환 대상 문장을 파싱하여 문장 성분으로 분해한다. 그런 다음, 가짜 뉴스 탐지용 데이터 셋 생성 장치(100)는 노이즈 적용 문장(문장 1, 문장 2, 문장 3, 문장 4, 문장 5)의 문장 성분 중 특정 문장 성분(예를 들어, 랜덤하게 선택된 문장 성분)의 단어 및 교환 대상 문장에서 동일한 문장 성분을 가지지만 일치하지 않은 단어를 교환한다.
- [0124] 도 5는 본 발명의 일 실시예에 따른 가짜 뉴스 탐지용 데이터 셋 생성 과정을 통해 생성된 가짜 뉴스 데이터를 설명하기 위한 예시도이다.
- [0125] 도 5를 참조하면, 본 발명에 따른 가짜 뉴스 데이터의 제목은 열차 충돌 사고가 발생한 장소가 “인도 남부”이지만, 본문의 내용은 열차 충돌 사고가 발생한 장소가 “인도 북부”로 되어 있다. 또한, 본 발명에 따른 가짜 뉴스 데이터의 제목은 “열차 충돌” 사고가 발생하였지만, 본문의 내용은 “열차 탈선” 사고로 되어 있다.
- [0126] 상기와 같이, 가짜 뉴스 탐지용 데이터 셋 생성 과정을 통해 생성된 가짜 뉴스 데이터는 제목 및 본문이 상이하다. 이와 같은 이유는 제목 및 본문이 상이한 가짜 뉴스를 탐지하기 위해서이다.
- [0128] 본원 발명의 실시예 등과 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.

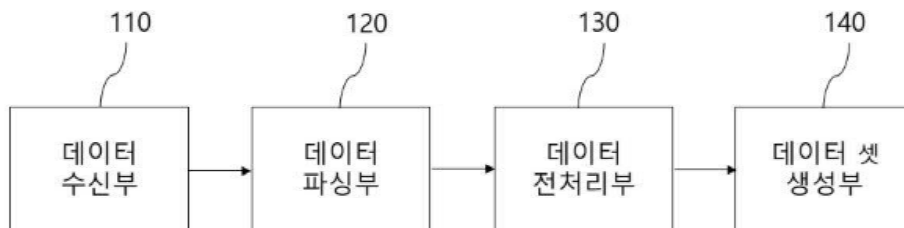
부호의 설명

- [0129] 100: 가짜 뉴스 탐지용 데이터 셋 생성 장치,
110: 데이터 수신부,
120: 데이터 파싱부,
130: 데이터 전처리부,

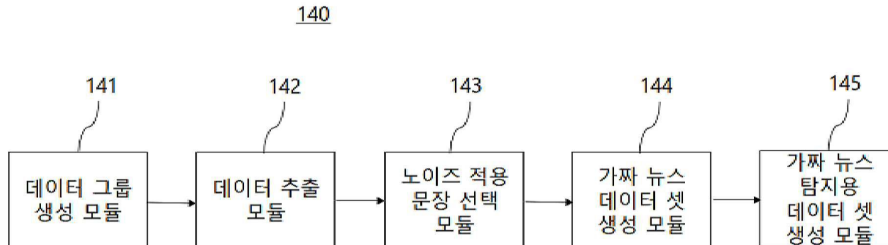
- 140: 데이터 셋 생성부,
- 141: 데이터 그룹 생성 모듈,
- 142: 데이터 추출 모듈,
- 143: 노이즈 적용 문장 선택 모듈,
- 144: 가짜 뉴스 데이터 셋 생성 모듈,
- 145: 가짜 뉴스 탐지용 데이터 셋 생성 모듈

도면

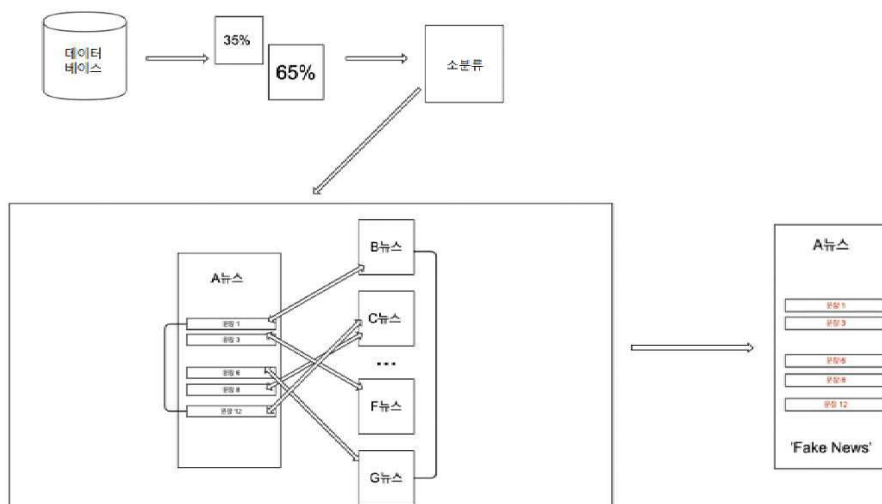
도면1



도면2

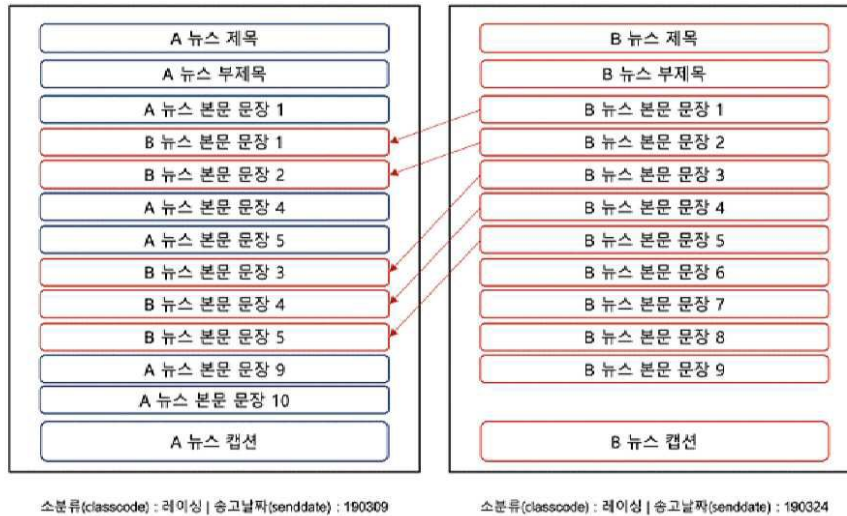


도면3



도면4

[제목과 본문이 다른 가짜뉴스 데이터]



도면5

제목	'인도 남부서 열차 충돌로 12명 부상...사망자는 없어'
본문	인도 북부에서 20일 오전 열차 탈선 사고가 발생해 14명 이상이 다쳤다고 지뉴스 등 현지언론이 보도했다. 이날 사고는 우타르프라데시주 칸푸르 지역 루마역 근처에서 발생했다. 우타르프라데시 프라야그라지에서 수도 뉴델리로 향하던 열차가 갑자기 궤도를 이탈하면서 전체 12량 가운데 10량이 탈선했고 4량이 완전히 넘어졌다. 이 사고로 승객 14명 이상이 병원으로 실려 갔으며 40여명의 구조대원이 현장으로 급파돼 현장 수습에 나섰다. ... 우타르프라데시주에서는 2016년에도 기차가 탈선해 150여명이 사망했으며, 지난해 10월에도 탈선 사고가 발생, 승객 7명이 목숨을 잃었다. 지난해 10월 편자브주에서는 달리던 열차가 철로 위에서 축제를 즐기던 인파를 덮치면서 61명 이상이 사망하는 참사가 벌어지기도 했다.