(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0094050
(43) 공개일자 2024년06월25일

(51) 국제특허분류(Int. Cl.)
G06F 40/226 (2020.01) *G06F 16/901* (2019.01)
G06F 16/9035 (2019.01) *G06F 40/211* (2020.01)
G06F 40/30 (2020.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01) *G06Q 50/10* (2012.01)
(52) CPC특허분류
G06F 40/226 (2020.01)
G06F 16/9024 (2019.01)
(21) 출원번호 10-2022-0154603
(22) 출원일자 2022년11월17일
심사청구일자 2022년11월17일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
강명훈
서울특별시 송파구 올림픽로 435, 224동 3301호
(74) 대리인
김등용

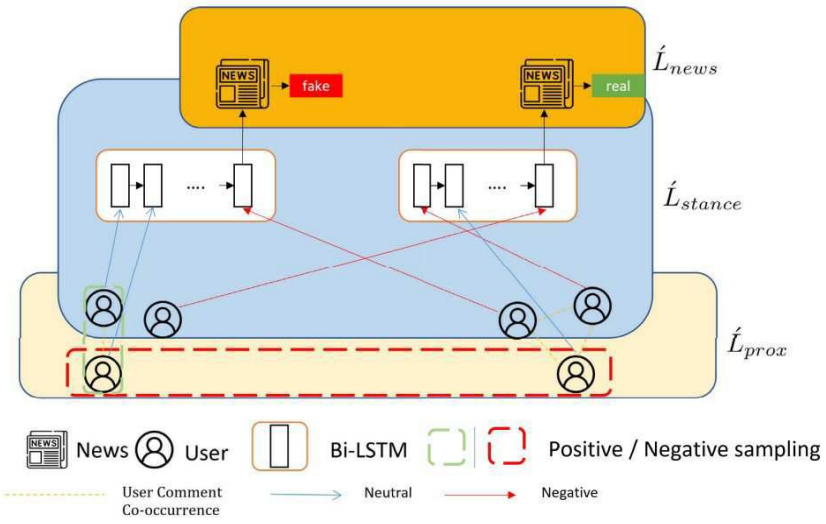
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **독자 정보를 활용하는 가짜뉴스 판별 방법 및 장치**

(57) 요 약

가짜뉴스 판별 모델 생성 장치 및 방법과 이를 이용한 가짜뉴스 판별 장치 및 방법이 개시된다. 상기 가짜뉴스 판별 모델 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 가짜뉴스 판별 모델 생성 방법으로써, 기본 데이터를 수집하는 단계, 상기 기본 데이터를 가공하여 학습 데이터를 구축하는 단계, 및 상기 학습 데이터를 이용하여 모델을 학습함으로써 상기 가짜뉴스 판별 모델을 생성하는 단계를 포함하고, 상기 기본 데이터는 뉴스를 식별할 수 있는 정보와 뉴스의 가짜뉴스 여부에 대한 정보를 포함하는 가짜뉴스 정보와 뉴스, 뉴스에 작성된 댓글, 댓글을 작성한 독자를 식별할 수 있는 정보, 독자들 각각이 작성한 댓글들을 포함하는 독자 반응 정보를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G06F 16/9035 (2019.01)
G06F 40/211 (2020.01)
G06F 40/30 (2020.01)
G06N 3/0454 (2013.01)
G06N 3/08 (2023.01)
G06Q 50/10 (2015.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1345349433
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2022.03.01 ~ 2023.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711159969
과제번호 2018-0-01405-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 가짜뉴스 판별 모델 생성 방법에 있어서,

기본 데이터를 수집하는 단계;

상기 기본 데이터를 가공하여 학습 데이터를 구축하는 단계; 및

상기 학습 데이터를 이용하여 모델을 학습함으로써 상기 가짜뉴스 판별 모델을 생성하는 단계를 포함하고,

상기 기본 데이터는 뉴스를 식별할 수 있는 정보와 뉴스의 가짜뉴스 여부에 대한 정보를 포함하는 가짜뉴스 정보와 뉴스, 뉴스에 작성된 댓글, 댓글을 작성한 독자를 식별할 수 있는 정보, 독자들 각각이 작성한 댓글들을 포함하는 독자 반응 정보를 포함하는, 가짜뉴스 판별 모델 생성 방법.

청구항 2

제1항에 있어서,

상기 학습 데이터는 독자와 뉴스에 대응하는 노드들(nodes)과 독자와 뉴스 사이의 관계 및 독자들 사이의 관계에 대응하는 엣지들(edges)을 포함하는 그래프 구조의 데이터인,

가짜뉴스 판별 모델 생성 방법.

청구항 3

제2항에 있어서,

상기 생성하는 단계는,

2 개 이상의 동일한 뉴스에 댓글을 작성한 독자들 사이의 독자 노드 임베딩 거리가 가까워지도록 학습을 진행하는,

가짜뉴스 판별 모델 생성 방법.

청구항 4

제3항에 있어서,

상기 생성하는 단계는,

동일한 태도를 표현한 독자 노드 임베딩과 뉴스 노드 임베딩 간의 유사도가 최대화되도록 학습을 진행하는,

가짜뉴스 판별 모델 생성 방법.

청구항 5

제4항에 있어서,

상기 생성하는 단계는,

뉴스 표현형을 완전 연결 레이어에 투과하여 뉴스의 허위정보 포함 여부를 판별하는,

가짜뉴스 판별 모델 생성 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인터넷 뉴스의 독자 정보를 이용한 가짜뉴스 판별 모델 생성 장치 및 방법과 이를 이용한 가짜뉴스

판별 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 가짜뉴스(fake news) 판별 시스템은 어떤 뉴스에 허위정보가 담겨있는지 자동으로 판별하는 인공지능(Artificial Intelligence, AI) 시스템을 말한다. 기존 인공지능 기반의 가짜뉴스 탐지 시스템은 기사 제목이나 본문과 같은 텍스트 정보만을 활용하여 허위정보 여부를 판별하는 방식으로 개발되었다. 그러나 기사 내부 정보만을 이용하는 방식은 이전에 학습하지 못한 새로운 내용과 주제가 담긴 기사가 입력으로 주어질 때 정확도가 떨어진다는 문제점이 있다. 이는 특정 시점의 내용과 주제가 담긴 기사 텍스트 정보만을 가지고 학습되기 때문이다. 따라서 기존 시스템이 추론 과정에서 생성하는 기사 표현 벡터(news representation vector)는 분류 강건성 측면에서 품질이 떨어진다.

[0003] 이와 같이 기사 텍스트 단일 정보만을 활용하는 시스템에서 벗어나 외부 정보를 결합하여 가짜뉴스 탐지를 시도하는 연구들이 최근 진행되고 있다. 대표적인 외부 정보는 위키백과 정보와 같은 지식정보와 독자 반응정보가 존재한다. 이 중 독자 반응정보는 어떤 기사가 독자에게 주는 뉘앙스(또는 태도)로 허위정보 여부를 간접적으로 파악하는데 유용한 자료로 사용될 수 있다.

[0004] 본 발명에서는 (한국) 독자의 기사 소비행태가 주로 인터넷 포털에서 이루어지는 점과 인터넷 포털에서 댓글을 통해 자유롭게 기사와 관련된 의견의 공유가 이뤄지는 점을 고려하여 독자 정보를 활용하는 (한국어) 가짜뉴스 판별 시스템을 만들고자 한다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 등록특허 제1869815호 (2018.06.22. 공고)

(특허문헌 0002) 대한민국 공개특허 제2021-0059544호 (2021.05.25. 공개)

(특허문헌 0003) 대한민국 공개특허 제2020-0064943호 (2020.06.08. 공개)

(특허문헌 0004) 대한민국 공개특허 제2021-0092979호 (2021.07.27. 공개)

비특허문헌

[0006] (비특허문헌 0001) M. Kang, J. Seo, and H. Lim, "Korean fake news detection with user graph," in Proc. Annu. Conf. Hum. Lang. Technol., 2021, pp. 97-102.

(비특허문헌 0002) M Kang, J Seo, C Park, and H Lim, "Utilization Strategy of User Engagements in Korean Fake News Detection", IEEE Access, 2022, 10: 79516-79525.

발명의 내용

해결하려는 과제

[0007] 본 발명이 이루고자 하는 기술적인 과제는 독자 정보를 이용하여 뉴스의 허위 여부, 즉 가짜뉴스를 판별하는 방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0008] 본 발명의 일 실시예에 따른 가짜뉴스 판별 모델 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되고, 기본 데이터를 수집하는 단계, 상기 기본 데이터를 가공하여 학습 데이터를 구축하는 단계, 및 상기 학습 데이터를 이용하여 모델을 학습함으로써 상기 가짜뉴스 판별 모델을 생성하는 단계를 포함하고, 상기 기본 데이터는 뉴스를 식별할 수 있는 정보와 뉴스의 가짜뉴스 여부에 대한 정보를 포함하는 가짜뉴스 정보와 뉴스, 뉴스에 작성된 댓글, 댓글을 작성한 독자를 식별할 수 있는 정보, 독자들 각각이 작성한 댓글들을 포함하는 독자 반응 정보를 포함한다.

발명의 효과

[0009] 본 발명은 기존 텍스트 단일 정보만을 활용하는 가짜뉴스 탐지 시스템의 문제점, 즉 새로운 입력이 주어질 때에도 강건성을 갖는 가짜뉴스 탐지 기법을 제공할 수 있다.

[0010] 또한, 가짜뉴스 탐지 시스템의 학습을 위한 뉴스-독자 그래프 구축 알고리즘을 통해서 보다 풍부한 정보를 가진 한국어 가짜뉴스 탐지 데이터셋 구축을 할 수 있다.

도면의 간단한 설명

[0011] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명에서 제안하는 전체적인 데이터셋 구축 과정의 흐름도이다.

도 2는 본 발명에서 제안하는 데이터셋의 그래프 구조를 도시한다.

도 3은 본 발명에서 제안하는 학습 알고리즘의 학습 아키텍쳐를 도시한다.

도 4는 본 발명의 일 실시예에 따른 가짜뉴스 판별 모델 생성 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0013] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0014] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0015] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0016] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0017] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0018] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0019] 우선, 본 발명에서는 그래프 기반 (한국어) 가짜뉴스(fake news) 데이터셋 구축 방법을 제안한다. 종래의 연구들에서는, 뉴스 콘텐츠(news content)만을 주된 특징으로 고려하는 데이터셋 구축 방법이 제안되었다. 그러나, 뉴스 콘텐츠만을 단일의 특징으로 고려하는 것은 모델의 성능 저하를 야기한다.

[0020] 본 발명에서 제안하는 데이터셋은, 엔티티들(entities)로써 뉴스(news)와 독자(user)를, 그리고 각 노드 사이 또는 노드들 내의 엣지들을 포함하는 이종 그래프(heterogeneous graph)의 형태를 갖는다. 이러한 그래프 구조의 데이터셋은 모델이 그래프 기반 학습 알고리즘을 통해 학습하도록 한다. 그래프 기반 학습 알고리즘은 1) 노드 특징 정보(node feature information)를 이용하여 각 노드의 표현을 학습하고, 2) 각 노드 사이 또는 노드들 내의 관계를 이용하여 구조적 정보를 학습한다. 데이터셋은 한국어 뉴스 콘텐츠와 독자 반응 정보(user engagement information)을 수집함으로써 구축될 수 있다.

[0021] A. OVERALL PROCEDURE

[0022] 전체적인 데이터셋 구축 과정은 도 1의 흐름도에 도시되어 있다. 라벨링된 뉴스 데이터를 고려하여, 제안하는 데이터셋을 구축하기 위하여 뉴스 (텍스트) 콘텐츠와 독자 반응이 검색된다. 독자 반응의 경우, 각 독자의 최근 댓글 작성 행동에 기초하여 독자 댓글들(comments)과 정보가 수집될 수 있다. 결과적으로, 제안하는 데이터셋의 최종 형태는 독자-뉴스 그래프(user-news graph)이다.

[0023] B. DATA COLLECTION

[0024] (한국어) 가짜뉴스 탐지 모델을 위한 학습 데이터셋(training dataset)을 구축하기 위하여, 팩트체크 플랫폼(fact-checnking platform, 팩트체크 서비스를 제공하는 서버를 의미할 수 있음)으로부터 팩트체크된 뉴스(fact-checked news)가 수집될 수 있다. 예시적으로, 라벨링된 뉴스를 획득하기 위하여 SNU FactCHeck 플랫폼(https://factcheck.snu.ac.kr)이 이용될 수 있다. SNU FactCheck는 한국어 팩트체크 플랫폼이다. 즉, 크롤링 동작을 통해, 뉴스에 대한 정보(뉴스 콘텐츠 및/또는 뉴스를 식별하기 위한 식별 정보 등을 의미할 수 있음) 및/또는 해당 뉴스의 가짜뉴스 여부에 대한 정보(즉, 라벨)가 수집될 수 있다.

[0025] 다만, 본 발명이 라벨링된 뉴스를 획득하기 위한 소스(source)에 제한되는 것은 아니다. 실시예에 따라, 상이한 팩트체크 플랫폼 또는 가짜뉴스 여부에 대한 라벨을 제공하는 임의의 서버 등으로부터 획득할 수도 있다. 또한, 라벨링된 뉴스가 미리 확보된 경우, 라벨링된 뉴스에 대한 별도의 크롤링 과정은 생략될 수도 있다.

[0026] 한국어 가짜뉴스 감지에서 부가적인 특징들을 온전히 이용하는 모델을 구축하기 위하여, 독자 반응과 정보가 수집될 수 있다. 한국어 독자들은 주로 뉴스 포털의 인터넷 뉴스를 구독한다. 따라서, 네이버 뉴스(Naver News)와 같은 포털(포털 서비스를 제공하는 서버를 의미할 수 있음)로부터 데이터가 수집될 수 있다. 팩트체크된 뉴스에 관하여, 각 뉴스에 대한 댓글(comments)이 수집될 수 있다. 이때, 팩트체크된 뉴스를 식별할 수 있는 정보에 기초하여 해당 뉴스가 수집될 수 있다. 또한, 독자 반응은 독자가 작성한 댓글 자체를 의미하고, 정보는 댓글을 작성한 독자의 아이디, 닉네임, 식별자, 이름 등과 같이 댓글을을 작성한 독자를 식별할 수 있는 정보를 의미할 수 있다.

[0027] 그러나, 본 발명이 독자 반응 및/또는 정보가 수집되는 포털에 제한되는 것은 아니다. 실시예에 따라, 다양한 포털로부터 수집이 가능하고, 독자 반응 및/또는 정보가 획득되어 있는 경우, 이에 대한 별도의 크롤링 과정은 생략될 수 있다.

[0028] C. PREPROCESS

[0029] 가짜뉴스 탐지 문제는 이진 분류 문제(binary classification problem)로 정의 된다. 실시예에 따라, 팩트체크 플랫폼으로부터 크롤링된 라벨링된 뉴스는 복수의 카테고리(예컨대, 6개의 카테고리)로 분류될 수 있다. 또한, "False"와 "True"로 라벨링된 뉴스는 각각 가짜뉴스(fake news)와 실제뉴스(real news)로 취급된다. 또한, 로 데이터(raw data)에는 소정의 필터링 과정이 수행될 수도 있다.

[0030] 그런 다음, 독자 정보(user information)가 전처리될 수 있다. 실시예에 따라, 독자반응(user engagements)은 독자 자체에 대한 정보(식별자 등)와 독자가 작성한 댓글을 포함하는 개념으로 이해될 수 있다. 따라서, 독자반응으로부터 1) 각 뉴스에 대한 독자의 태도(stance)와 2) 독자 정보(user information)가 도출될 수 있다.

[0031] 뉴스에 대한 독자의 태도는 독자의 댓글을 입력으로 하는 임의의 감성 분석 모델의 출력을 통해 획득(또는 도출)될 수 있다. 예시적으로, 트랜스포머(Transformer) 기반의 KoBERT 감성 분석 모델에 독자의 댓글을 입력함으로써 이진의 독자 태도(binary stance)를 산출할 수 있다. 감성 분석 모델의 출력은, 예시적으로, "긍정"과 "부정" 중 어느 하나(또는 "긍정", "부정", 및 "중립" 중 어느 하나이거나 "긍정"과 "중립" 중 어느 하나일 수

도 있음)일 수 있다.

[0032] 독자 정보의 경우, 네이버 뉴스 플랫폼은 독자에 대한 소개 정보(user text description)를 제공하지 않는다. 따라서, 각 독자의 최근 댓글 행위에 태도가 반영되어 있다고 가정할 수 있다. 독자가 최근에 작성한 복수의 댓글들(예컨대, 가장 최근에 작성한 복수개의 댓글들(예시적으로 5개의 댓글들))이 독자 정보로 수집될 수 있다.

[0033] 데이터 수집 이후에는, 독자-뉴스 이종 그래프가 구축될 수 있다.

[0034] D. GRAPH CONSTRUCTION

[0035] 일 실시예에 따르면, 제안하는 데이터셋은 인터넷 뉴스에 기초하여 구축될 수 있다. 페이퍼 뉴스(paper news)에 비해, 댓글을 작성하거나 기사를 공유하는 등과 같이 인터넷 뉴스에 독자들이 활발하게 참여하는 특성이 있기 때문이다. 게다가, 모든 독자반응이 공개되어 있다. 본 발명에서는, 독자 엔티티 및 뉴스 엔티티와 엣지(edge)로 표현되는 독자와 뉴스 간의 상호 작용을 포함하는 이종 그래프 구조의 데이터셋(heterogeneous graph-structured dataset)을 구축한다. 제안하는 데이터셋의 그래프 구조는 도 2에 도시되어 있다. 노드들은, 독자 정보의 텍스트 임베딩과 뉴스 콘텐츠의 텍스트 임베딩으로 표현되는 특징을 갖는 독자 노드와 뉴스 노드를 포함한다.

[0036] 다음으로, 엣지는 독자-독자 관계와 독자-뉴스 관계로 구성된다. 트위터(Twitter) 등과는 대조적으로, 네이버 뉴스 플랫폼에는 독자들 간의 명확한 관계(explicit relationship)가 존재하지 않는다. 따라서, 독자-독자 엣지는 암시적인 관계(implicit relationship)에 기초하여 구축될 수 있다. 독자-독자 관계는, 도 2의 노란색 선으로 표시된 것과 같이, 독자 댓글 동시 등장 관계로 독자 간의 관계를 구성할 수 있다. 즉, 임의의 두 독자가 동일한 2개 이상의 기사에 댓글을 작성한 경우에 두 독자 사이에 관계가 있는 것으로 취급하여 독자 관계를 구축할 수 있다. 예컨대, 독자 u_j와 u_{j+1}이 공통으로 뉴스 a_i와 a_{i+1}에 반응하였다면(또는 댓글을 작성하였다면), 독자 u_j와 u_{j+1} 사이에는 독자 댓글 동시 등장 관계가 성립할 수 있다.

[0037] 또한, 독자-뉴스 관계는, 도 2의 빨강 애로 라인(arrow line)과 파랑 애로 라인으로 표시된 것과 같이, 뉴스에 대한 독자의 태도(stance)로 표현될 수 있다. 빨강 애로 라인과 파랑 애로 라인 각각은, 뉴스에 대한 독자의 태도로써 부정(negative)과 중립(neutral, 또는 긍정(positive))을 나타낸다.

[0038] 마지막으로, 독자-뉴스 그래프 G가 정의된다.

[0039] . $A = \{a_1, a_2, \ldots, a_i\}$는 뉴스 노드들의 리스트이고, $U = \{u_1, u_2, \ldots, u_j\}$는 독자 노드들의 리스트이다.

[0040] . $E = \{e_1, e_2, \ldots, e_k\}$는 엣지들의 리스트이다. 각 엣지 $e_k = \{v_i, v_{i+1 or j}, t, x_e\}$는 타임 스탬프(time stamp) t에서 두 엔티티들 $v_i, v_{i+1 or j}$ 사이의 관계 라벨(relation label) x_e로 모델링된다(또는 포함한다).

[0041] · 독자-독자 관계의 경우, 라벨(label) x_e는 $\{\text{True: 0, False: 1}\}$로, 독자-뉴스 관계는 $\{\text{Negative: 0, Neutral: 1}\}$로 정의된다. 즉, 독자 댓글 동시 등장 관계에 있는 독자 사이의 관계의 라벨 값은 0(또는 1)을, 그렇지 않은 경우 1(또는 0)의 값을 가질 수 있다. 또한, 해당 뉴스에 대해 부정적인 태도를 갖는 독자와 해당 뉴스의 라벨은 0의 값을, 그렇지 않은 경우(예컨대, 중립적인 태도를 갖는 경우)의 라벨은 1의 값을 가질 수 있다.

[0042] **제안 모델(K-FANG)**

[0043] 본 발명에서는, 독자-뉴스 인터랙션을 이용함으로서 가짜뉴스을 탐지할 수 있는 학습 알고리즘을 제안한다. 제안 모델(K-FANG)의 학습 아키텍처는 도 3에 도시되어 있다. K-FANG은 독자 엔티티들과 뉴스 엔티티들 사이의 관계 정보를 활용함으로써 표현 벡터(representation vector)를 생성할 수 있다. 또한, 표현 벡터의 생성과 업데이트는 제1 로스(태도 로스(stance loss)로 명명될 수도 있음) $\acute{L}_{stance}$, 제2 로스(근접 로스(proximity loss)로 명명될 수도 있음) $\acute{L}_{prox}$, 및 제3 로스(뉴스 로스(news loss)로 명명될 수도 있음) $\acute{L}_{news}$를 이용함하여 수행될 수 있다.

[0044] 제안 모델(K-FANG)은 그래프 기반 학습 모델이다. 그래프 기반 모델은 노드 간의 관계 정보와 그래프 구조 정보(graph structural information)를 임베딩하는 표현 벡터를 생성할 수 있어야 한다. 또한, 생성된 그래프 표현 벡터는 리얼 월드의 개념 정보(real-world conceptual information)를 포함할 수 있다. 이를 구현하기 위해,

본 발명에서는 그래프 표현 학습 모델 *GraphSAGE*를 이용하였다. *GraphSAGE*는 k(k는 임의의 자연수로써, 사용자나 관리자에 의해 미리 결정될 수 있음) 개의 로컬 이웃 노드들(k local neighbor nodes)의 임베딩 벡터를 샘플링(sampling)하고 집속(aggregating)함으로써 표현을 학습한다(W. L. Hamilton, R. Ying, and J. Leskovec, ‘‘Inductive representation learning on large graphs,’’ in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 1025-1035.). 즉, 노드별로 자기 자신의 정보뿐만이 아니라 연결된 노드(이웃 노드)의 정보까지 포함하여 임베딩 벡터를 업데이트(학습)하므로 각 노드는 이웃 노드의 정보와 그래프의 구조적 정보를 포함하게 된다. 로컬 노드 임베딩 집속으로 인하여, 표현 벡터는 그래프 구조 정보 또한 포함할 수 있다. 표현 학습 과정은 수학식 1에 나타나 있다.

[0045] [수학식 1]

[0046] $$h_v^k = \sigma(W^{(k)} \cdot concat(h_v^{(k-1)}, \mathrm{AGG}(h_u^{(k-1)}, \quad \forall u \in N(v))))$$

[0047] 수학식 1에서, v는 표현 벡터의 업데이트를 위한 타겟 노드(target node), u는 v의 이웃, σ는 활성화 함수(activation function), 그리고 AGG는 집속 함수(aggregation function)이다. 본 발명에서, 표현 학습을 위해 평균 집속 기법(mean aggregation method)을 이용할 수 있다.

[0048] 다음으로, 리얼 월드 개념 정보를 나타내는 표현 벡터들을 생성하기 위한 가설이 설정된다.

[0049] · H_1: 독자 표현과 뉴스 표현 사이의 거리(distance)는, 독자가 뉴스에 대하여 특정 태도를 보일 때, 가까워질 것이다.

[0050] · H_2: 각 독자 표현은, 각 독자가 댓글 동시 등장 관계에 있을 때, 가까워질 것이다.

[0051] 이러한 가설 하에, 제안 모델(K-FANG)은 뉴스 표현 벡터와 독자 표현 벡터를 학습한다. 뉴스 표현을 위해, 제안 모델(K-FANG)은 뉴스 콘텐츠와 뉴스에 대한 독자의 태도를 동시에 고려한다. 이에 따라, 독자-뉴스 반응 어레이(user-news engagements array)를 입력으로 이용하여, v_a^{temp}를 생성하기 위해 Bi-LSTM 모델이 이용된다. 독자-뉴스 반응 어레이는 $\{e_1, e_2, \ldots, e_{|U|}\}$, 어레이의 각 값은 $e_{|U|} = (z_u, meta(x_{e_i}))$이다. 여기서, z_u는 *GraphSAGE*(u)에 의해 생성된 독자 표현 벡터이고, $meta(x_{e_i}) = (t, stance(e_i))$이다. 이는, 타임 스탬프 t에서 독자의 태도 $stance(e_i)$를 나타낸다. 독자-뉴스 반응 어레이가 Bi-LSTM 모델을 투과하면, 포워드 히든 벡터(forward hidden vector)와 백워드 히든 벡터(backward hidden vector)의 접합(concatenation)을 통해 v_a^{temp}이 생성된다. Bi-LSTM은 입력을 왼쪽에서 오른쪽으로 받아 학습을 진행하는 포워드 모델과 입력을 오른쪽에서 왼쪽으로 받아 학습을 진행하는 백워드 모델이 결합된 텍스트 표현 벡터 생성 모델이다. 뉴스 표현의 최종 형태는 수학식 2에 나타나 있고, v_a^{temp}와 뉴스 노드 표현 벡터의 접합을 통해 획득될 수 있다.

[0052] [수학식 2]

[0053] $$z_a = v_a^{temp} + GraphSAGE(a)$$

[0054] 독자 표현 벡터와 뉴스 표현 벡터가 설정되면, 제안 모델(K-FANG)은 표현을 업데이트한다. 이는 도 3에 도시되어 있다. 첫째로, 독자 표현이 업데이트된다. 추가적으로, H_2 하에서 비지도 방식으로 근접 로스(proximity loss)가 계산된다. H_2를 만족시키기 위해, 모델이 네거티브 샘플들과 넌네거티브 샘플들(non-negative samples) 중에서 구별되는 표현을 학습하도록 하는 네거티브 샘플링(negative sampling)이 이용될 수 있다. 즉, GraphSAGE로 생성된 노드 임베딩을 이용하여 관계가 있는 독자 노드 임베딩 간의 거리는 가깝게 하고 그렇지 않은 경우 임베딩 간의 거리를 멀게 하는 비지도 학습 기반의 근접 로스를 산출할 수 있다. 본 발명에서는 네거티브 샘플을 다른 독자와의 컨넥션(관계)을 갖지 않은 독자들의 그룹으로 정의한다. 근접 로스의 최종 형태는 수학식 3에 나타난다.

[0055] [수학식 3]

[0056] $$\acute{L}_{prox} = -\sum_{r_p \in P_r} \log(\sigma(z_r^\top z_{r_p})) + \sum_{r_n \in N_r} \log(\sigma(-z_r^\top z_{r_n}))$$

[0057] 수학식 3에서, Pr과 Nr은 각각 넌네거티브 샘플링과 네거티브 샘플링을, z_{r_p}은 넌네거티브 샘플링 표현 벡터, z_{r_n}은 네거티브 샘플링 표현 벡터를 나타낸다.

[0058] 다음으로, 뉴스 표현과 독자 표현이 (동시에) 업데이트된다. 태도 로스(stance loss)는 $H1$ 하에서 자기지도(self-supervised) 방식으로 계산될 수 있다. 즉, 모델은, 각 태도에 대하여 뉴스 표현과 독자 표현 사이의 코릴레이션이 최대화하도록 학습된다. 따라서, 부정 태도(negative stance)와 연관된 뉴스 표현은 코릴레이션을 최대화하도록 학습되고, 중립 태도(neutral stance)와 연관된 뉴스 표현은 코릴레이션을 최대화하도록 학습된다. 또한, 독자 표현과 뉴스 표현 사이의 유사도(similarity)를 계산하기 위해, 모델은 최적의 투영 가중치(optimal projection weight)를 결정하도록 학습된다. $\alpha_c(u) = A_c z_u$은 독자 투영 객체(user projection object)이고, $\beta_c(a) = B_c z_a$는 뉴스 투영 객체(news projection object)이다. 여기서 c는 독자-뉴스 태도 라벨(label) 인덱스를 나타내고, Ac와 Bc는 각각 독자 투영 객체와 뉴스 투영 객체를 위한 학습가능한 투영 가중치들을 나타낸다. 따라서, 태도 로스는 투영 객체 산물(projection object product)을 소프트맥스 로스 레이어(softmax loss layer)에 투과시킴으로써 도출될 수 있다. 태도 로스의 최종 형태는 수학식 4와 같다.

[0059] [수학식 4]

[0060] $$\acute{L}_{stance} = -\sum_{u,a}\sum_{c} \text{softmax}(\alpha_c(u)^{\top}\beta_c(a))$$

[0061] 다시 말하면, 동일한 태도를 표현한 독자 노드 임베딩과 뉴스 노드 임베딩 간의 유사도가 최대화되도록 학습을 진행한다. 학습을 진행하기에 앞서, 뉴스 표현형을 Bi-LSTM 모델을 이용하여 생성한다. 독자 댓글 텍스트 임베딩이 입력으로 주어져 Bi-LSTM 모델에 투과되어 뉴스-독자 반응 임베딩을 생성한다. 뉴스 노드 GraphSAGE 임베딩과 뉴스-독자 반응 임베딩이 결합되어 뉴스 표현형을 구성하고 태도 클래스별로 뉴스 표현형과 독자 표현형의 유사도가 최대화되도록 태도 로스를 산출한다.

[0062] 다음으로, 뉴스 분류(news classification)를 위한 뉴스 표현이 업데이트된다. 여기서, 지도 방식(supervised manner)으로 계산되는 뉴스 로스가 이용된다. 뉴스 로스는 완전 연결 레이어(fully connected layer)의 출력 $o_a = Wv_a + b$을 교차-엔트로피 로스 레이어(cross-entropy loss layer)에 투과시킴으로써 계산된다. 뉴스 로스는 수학식 5와 같다.

[0063] [수학식 5]

[0064] $$\acute{L}_{news} = \frac{1}{T}\sum_{a} y_a log(\sigma(o_a)) + (1 - y_a) log(1 - \sigma(o_a))$$

[0065] 수학식 5에서, T는 배치 크기(batch size)를, y_a는 실측된 뉴스 라벨(ground truth news label)을, 그리고 σ는 활성화 함수를 나타낸다.

[0066] 다시 말하면, 뉴스 표현형을 완전 연결 레이어에 투과하여 생성된 최종 표현형으로 뉴스의 허위정보 포함 여부를 판별하여 뉴스 로스를 산출한다.

[0067] 수학식 6과 같이, 전체 로스(total loss)는 상술한 3가지의 로스의 합으로 정의된다. 사용된 파라미터들은 역전파(backpropagation)을 통해 업데이트된다.

[0068] [수학식 6]

[0069] $$\acute{L}_{total} = SUM(\acute{L}_{news} + \acute{L}_{stance} + \acute{L}_{prox})$$

[0070] 도 4는 본 발명의 일 실시예에 따른 가짜뉴스 판별 모델 생성 방법 또는 가짜뉴스 판별 방법을 설명하기 위한 흐름도이다. 상기 방법들을 이루는 단계들 중 적어도 일부는 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 컴퓨팅 장치는 PC(Personal Computer), 태블릿 PC, 노트북, 및 서버(server)를 포함한다. 또한, 컴퓨팅 장치는 하나의 물리적 장치를 의미하거나 복수개로 분리된 물리적 장치들을 의미할 수도 있다. 이하에서 설명되는 방법을 설명함에 있어, 앞선 기재와 중복되는 내용에 관하여는 그 구체적인 기재를 생략하기로 한다.

[0071] 우선, 학습 데이터를 구축하기 위한 기본 데이터(로 데이터)가 수집될 수 있다(S100). 기본 데이터는 소정의 크

롤링을 통해 수집될 수 있다. 이때, 수집되는 기본 데이터는 가짜뉴스 여부 정보를 포함할 수 있다. 가짜뉴스 여부 정보는 뉴스의 제목, 뉴스의 본문, 뉴스를 식별할 수 있는 정보, 및 뉴스의 가짜뉴스 여부에 대한 정보(라벨을 의미할 수도 있음) 적어도 하나를 포함할 수 있다. 또한, 기본 데이터는 독자 반응 정보를 포함할 수 있다. 독자 반응 정보는 뉴스의 독자를 식별할 수 있는 정보, 뉴스에 대하여 작성된 댓글, 독자가 (가장) 최근에 작성한 복수의 댓글들을 포함할 수 있다. 실시예에 따라, 독자 반응 정보는 뉴스의 제목 및/또는 뉴스의 본문을 포함할 수도 있다. 가짜뉴스 여부 정보에 뉴스의 제목 및/또는 뉴스의 본문이 포함되지 않을 수도 있기 때문이다.

[0072] 다음으로, 학습 데이터가 구축될 수 있다(S200). 학습 데이터는 복수의 노드들과 노드들 사이의 엣지를 포함하는 그래프 구조의 데이터를 의미할 수 있다. 이를 위해, 독자 댓글 동시 등장 관계, 뉴스에 대한 독자들 각각의 태도 등이 도출될 수 있다. 구축된 학습 데이터는 뉴스-독자 그래프로 명명될 수도 있다.

[0073] 그런 다음, 구축된 학습 데이터를 통해 모델을 학습시킴으로써 가짜뉴스 탐지 모델이 생성될 수 있다(S300). 가짜뉴스 탐지 모델(또는 가짜뉴스 판별 모델)을 생성하기 위해서는 제1 로스, 제2 로스, 및 제3 로스가 이용된다. 또한, 역전파를 통한 파라미터(예컨대, 가중치) 업데이트를 통해 가짜뉴스 탐지 모델이 생성될 수 있다.

[0074] 타겟 뉴스에 대한 가짜뉴스 여부가 판단될 수 있다(S400). 이를 위해 타겟 뉴스에 대한 정보가 입력되어야 하며, 타겟 뉴스에 대한 정보는 뉴스 제목, 뉴스 본문, 뉴스에 대한 댓글들, 각 댓글들의 작성자인 독자에 대한 정보(독자를 식별할 수 있는 정보, 독자가 (가장) 최근에 작성한 복수의 댓글들)이 포함될 수 있다.

[0075] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

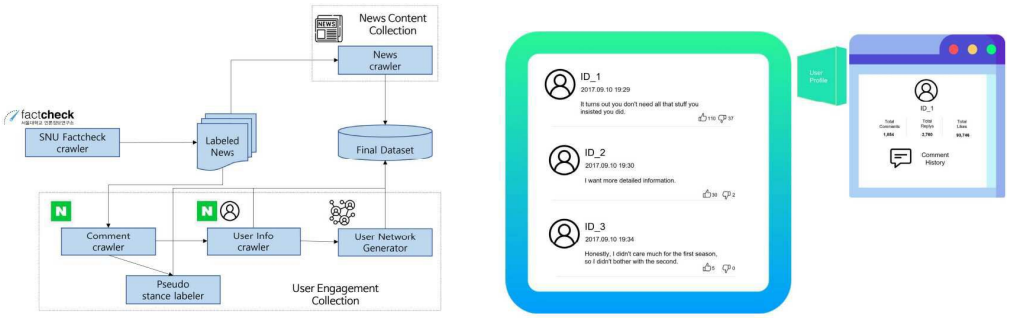
[0076] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0077] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0078] 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. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

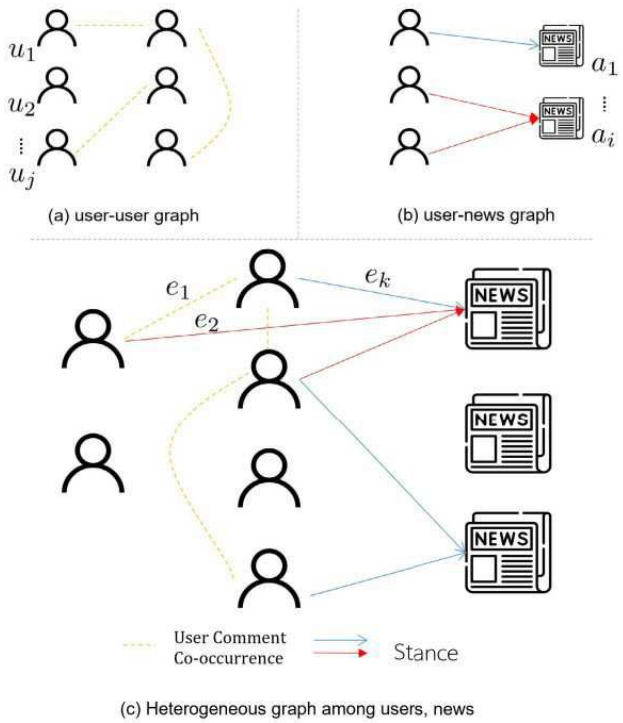
도면

도면1



(a) Flowchart of dataset construction (b) Example of User Engagement

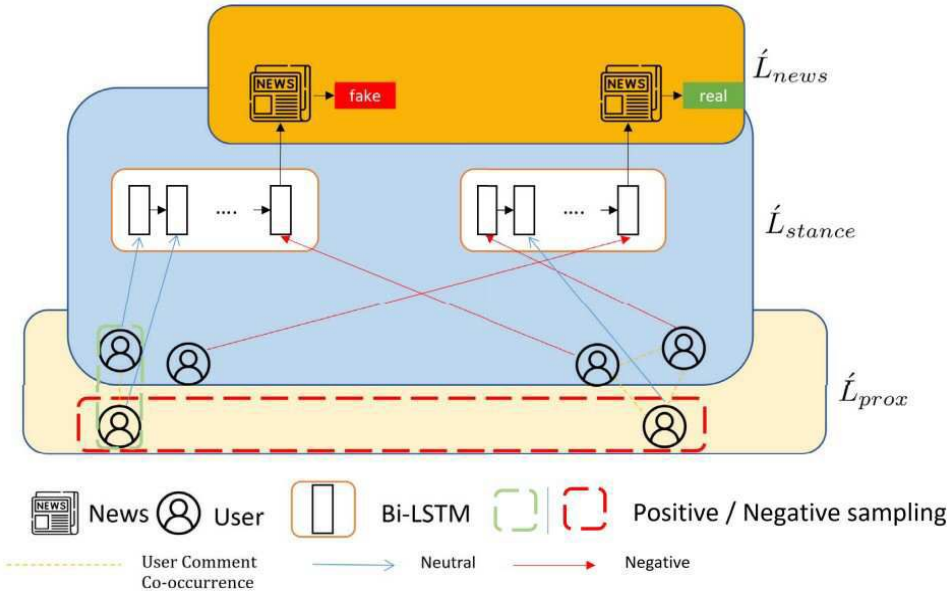
도면2



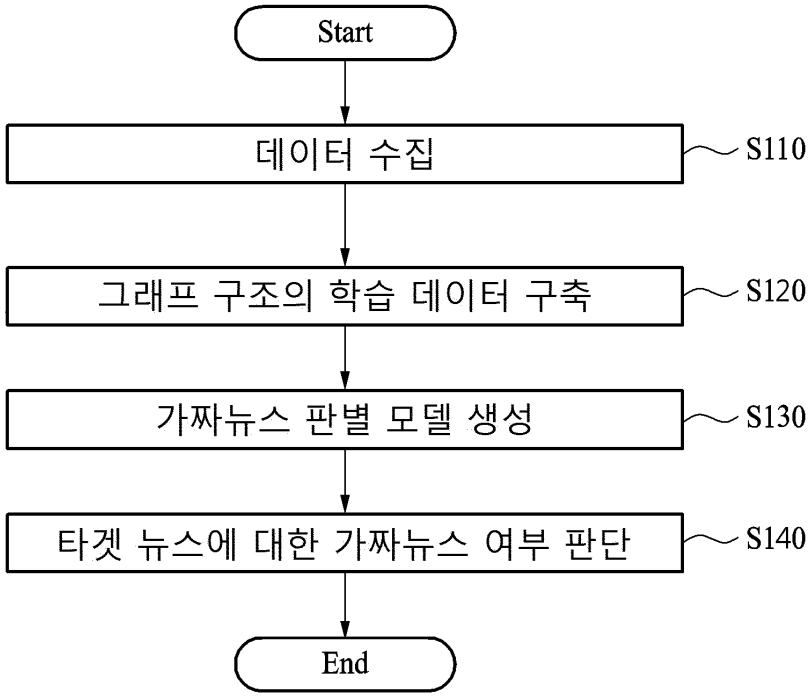
(a) user-user graph (b) user-news graph

(c) Heterogeneous graph among users, news

도면3


fake
real
$\acute{L}_{news}$
$\acute{L}_{stance}$
$\acute{L}_{prox}$
News
User
Bi-LSTM
Positive / Negative sampling
User Comment
Co-occurrence
Neutral
Negative

도면4


Start
데이터 수집
S110
그래프 구조의 학습 데이터 구축
S120
가짜뉴스 판별 모델 생성
S130
타겟 뉴스에 대한 가짜뉴스 여부 판단
S140
End