(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0098030
(43) 공개일자 2025년07월01일

(51) 국제특허분류(Int. Cl.)
G06F 16/335 (2019.01) *G06F 16/33* (2025.01)
G06F 16/332 (2025.01) *G06F 16/338* (2019.01)
G06F 16/35 (2025.01) *G06F 18/241* (2023.01)
G06N 3/0455 (2023.01) *G06N 3/0475* (2023.01)
G06N 3/096 (2023.01) *G06N 5/04* (2023.01)
(52) CPC특허분류
G06F 16/335 (2019.01)
G06F 16/3322 (2019.01)
(21) 출원번호 10-2023-0188138
(22) 출원일자 2023년12월21일
심사청구일자 2023년12월21일

(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

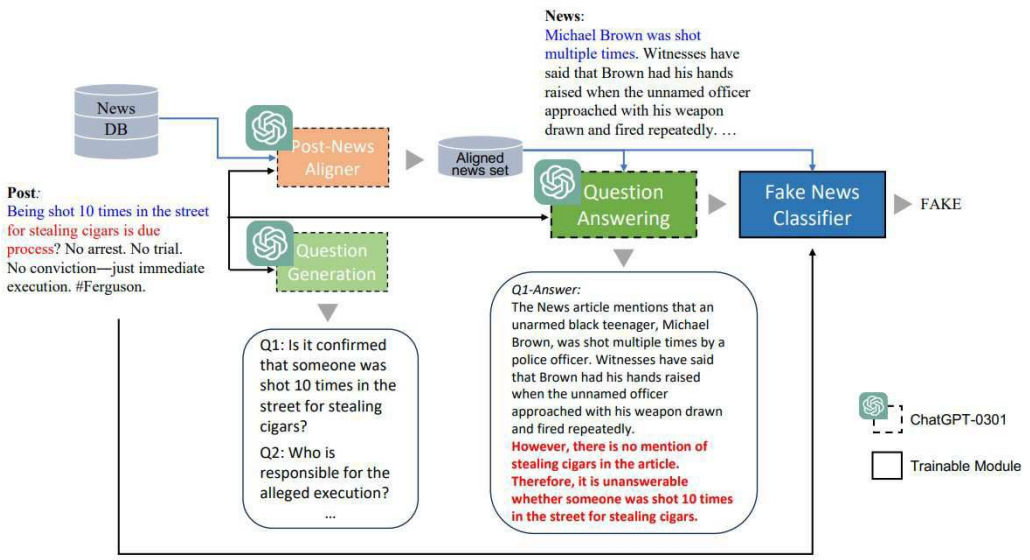
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **질의응답 쌍 생성을 통한 설명 가능한 가짜 뉴스 탐지 장치 및 방법**

(57) 요 약

질의응답 쌍 생성을 통한 설명 가능한 가짜 뉴스 탐지 장치 및 방법이 개시된다. 상기 가짜 뉴스 탐지 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 뉴스 집합에 대한 제1 필터링 동작을 수행하여 관련 뉴스들을 추출하는 단계, 상기 관련 뉴스들에 대한 제2 필터링 동작을 수행하여 참조 뉴스들을 추출하는 단계, 게시글(post)의 사실 검증에 유용한 질의를 생성하는 단계, 상기 질의에 대한 응답들을 생성하는 단계, 질의응답 쌍들 중에서 추론에 사용한 어느 하나의 질의응답 쌍을 선택하는 단계, 및 상기 어느 하나의 질의응답 쌍을 이용하여 상기 게시글의 진위 여부를 검증하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/3347 (2019.01)
G06F 16/338 (2019.01)
G06F 16/35 (2019.01)
G06F 18/241 (2023.01)
G06N 3/0455 (2023.01)
G06N 3/0475 (2023.01)
G06N 3/096 (2023.01)
G06N 5/04 (2023.01)
Y10S 707/943 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193711
과제번호 2018-0-01405-006
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345365402
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 가짜 뉴스 탐지 방법에 있어서,

뉴스 집합에 대한 제1 필터링 동작을 수행하여 관련 뉴스들을 추출하는 단계;

상기 관련 뉴스들에 대한 제2 필터링 동작을 수행하여 참조 뉴스들을 추출하는 단계;

게시글(post)의 사실 검증에 유용한 질의를 생성하는 단계;

상기 질의에 대한 응답들을 생성하는 단계;

질의응답 쌍들 중에서 추론에 사용한 어느 하나의 질의응답 쌍을 선택하는 단계; 및

상기 어느 하나의 질의응답 쌍을 이용하여 상기 게시글의 진위 여부를 검증하는 단계를 포함하는 가짜 뉴스 탐지 방법.

청구항 2

제1항에 있어서,

상기 관련 뉴스들을 추출하는 단계는,

상기 뉴스 집합으로부터, 상기 게시글의 발행일로부터 T(T는 임의의 자연수)일 이내에 발행되고 상기 게시글과 임계치 이상의 유사도를 갖는 뉴스들을 상기 관련 뉴스들로 추출하는,

가짜 뉴스 탐지 방법.

청구항 3

제2항에 있어서,

상기 유사도는 상기 게시글에 대한 벡터 표현과 뉴스에 대한 벡터 표현 사이의 코사인 유사도를 의미하는,

가짜 뉴스 탐지 방법.

청구항 4

제3항에 있어서,

상기 참조 뉴스들을 추출하는 단계는,

상기 게시글과 뉴스 사이의 관련성에 대한 이진 분류를 지시하는 프롬프트를 상기 게시글 및 뉴스와 함께 거대 언어 모델(Large Language Model, LLM)에 입력함으로써 상기 참조 뉴스들을 추출하는,

가짜 뉴스 탐지 방법.

청구항 5

제4항에 있어서,

상기 질의를 생성하는 단계는,

상기 게시글의 진실성을 검증하기 위해 유용한 질의를 생성하도록 지시하는 프롬프트를 상기 게시글과 함께 상기 거대 언어 모델(LLM)에 입력함으로써 상기 질의를 생성하는,

가짜 뉴스 탐지 방법.

청구항 6

제5항에 있어서,

상기 응답들을 생성하는 단계는,

상기 참조 뉴스들 각각에 대하여, 상기 게시글과 상기 질의에 대하여 참조 뉴스 내에서 응답을 찾도록 지시하는 프롬프트를 상기 게시글, 상기 참조 뉴스, 및 상기 질의화 함께 상기 거대 언어 모델(LLM)에 입력함으로써 상기 응답들을 생성하는,

가짜 뉴스 탐지 방법.

청구항 7

제6항에 있어서,

상기 어느 하나의 질의응답 쌍을 선택하는 단계는,

가장 높은 ROUGE 점수를 나타내는 질의응답 쌍을 상기 어느 하나의 질의응답 쌍으로 선택하는,

가짜 뉴스 탐지 방법.

청구항 8

제7항에 있어서,

상기 게시글의 진위 여부를 검증하는 단계는,

상기 게시글과 상기 어느 하나의 질의응답 쌍을 사전학습된 가짜 뉴스 분류 모델에 입력함으로써 상기 게시글이 진위 여부를 검증하는,

가짜 뉴스 탐지 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 질의응답 쌍(Question Answer Pair) 생성을 통하여 인간이 이해 가능한 자연어 형태로 추론 중간 과정을 제공하는 가짜 뉴스 탐지 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 가짜 뉴스(Fake news)는 그 특성상 심각한 사회적 위협을 야기할 수 있다. 최근의 사회적 사건이나 이슈를 활용하는 가짜 정보는, 미디어, 민주주의, 및 의사 결정에 대한 불신을 퍼트릴 수 있다. 이러한 위협을 완화하기 위하여, 기존의 가짜 뉴스 탐지 기법은 게시물 수준의 텍스트 패턴을 포착하는 콘텐츠 기반 방법이나 소셜 미디어 상호작용 신호 및 외부 지식 베이스와 같은 컨텍스트 기반 방법을 활용하였다. 그러나, 이러한 연구들은 "가짜 뉴스가 뉴스 환경에서 다루는 문제의 적시성" 문제를 등한시 한다. 다시 말하면, 가짜 뉴스 탐지을 위한 시스템은 가짜 뉴스가 생성되고 전파되는 맥락을 효과적으로 포착하기 위해 뉴스 환경을 고려하여야 한다.

[0003] 가짜 뉴스의 문맥을 포착하기 위하여, Sheng 등(Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4543-4556, Dublin, Ireland. Association for Computational Linguistics.)은, 뉴스 환경과 관련하여 가짜 뉴스를 탐지하기 위해 인기(popularity)와 새로움(novelty) 개념을 활용하는 NEP(News Environment Perception for fake news detection)를 제안하였다. 이 아이디어는 간단하다. 게시물이 매크로 환경에서 최근 이슈를 사용하고(popularity), 미크로 환경에서 독특한 콘텐츠를 더 많이 제시할수록(novelty), 가짜 뉴스일 가능성이 높다는 것이다. 이를 통해, NEP는, 타겟 게시물과 관련 뉴스 간의 의미적 유사성을 활용하여 가짜 뉴스를 탐지하기 위해 커널 풀링(kernel pooling) 및 문장 표현(sentence representation)을 사용한다.

[0004] 뉴스 환경을 활용함에도 불구하고, NEP는 게시물의 사실성(factuality)을 무시하여 탐지에서 심각한 오류를 야기한다. 표 1에 나타난 바와 같이, 게시물(post)은 경찰에 의한 총격의 무고한 피해자로서의 Michael Brown에 관한 가짜 뉴스를 다루고 있다. NEP가 의미적 유사도 접근법을 사용하기 때문에, "Hundreds of peaceful

protesters march in Ferguson" 또는 "St. Louis protesters dig in: 'I'm sick of this police brutality'" 와 같은 유사한 어조의 뉴스를 얻는다. 그러나, 이러한 뉴스 환경의 어느 것도 (1) 충격이 실제인지 아닌지, (2) 담배 절도가 충격의 실제 원인이었는지를 증명하지 못한다. 따라서, 게시물의 사실적 요소를 포함하고 뉴스 환경 내에서 이를 증명할 수 있는 진보된 접근 방식이 필요하다.

[0005] [표 1]

Post
Being shot 10 times in the street for stealing cigars is due process? No arrest. No trial. No conviction—just immediate execution. #Ferguson.
Top-5 similar news headlines
1. Calls for justice, and calm, in Ferguson: #tellusatoday 2. Alderman, 2 reporters arrested as Ferguson erupts for 4th night 3. Hundreds of peaceful protesters march in Ferguson 4. Wonkbook: What you need to know about the chaos last night in Ferguson 5. St. Louis protesters dig in: 'I'm sick of this police brutality'
Ground Truth Label
FAKE
Predicted Answers
REAL

[0006]

[0007] 상기 표 1은 NEP의 오류 케이스를 나타낸다. 유사한 뉴스 헤드라인은, 가짜 뉴스 탐지를 위한 증거로 사용될 때, 제한된 사실 정보를 가지고 있으며 명시적인 설명이 부족하다.

[0008] 게시물의 가짜 여부를 효과적으로 탐지하기 위하여는, 표면 수준을 넘어서 기저에 깔려 있는 가정을 고려하여야 한다. 이를 위해, 간단히 "Is it confirmed that someone was shot 10 times in the street for stealing cigars?"라는 질문을 하고, 그에 대한 응답을 뉴스 환경에서 찾을 수 있다. 이러한 접근법은 자동 가짜 뉴스 탐지에 관한 여러 이점을 제공한다. 즉, (1) 질문 생성을 통해 암시적인 사실적인 포인트를 포착하고, (2) 질문 응답을 통해 뉴스로부터 명시적인 답변을 얻고, (3) 가독성 있는 텍스트를 통해 설명 가능성을 보장할 수 있다.

[0009] 이를 위해, 본 발명에서는, 질의응답 생성을 통한 설명 가능한 가짜 뉴스 탐지 프레임워크인 XFD-QAG(Explainable Fake News Detection Framework with Question Answer Generation)를 제안한다. 제안하는 프레임워크는, 게시물(post, 검증 발언으로 명명될 수 있음)에서 사실적인 포인트(factual points)를 얻고, 뉴스 환경을 활용하기 위해 QAG(Question-Answering pair Generation, 질의응답 쌍 생성)을 이용한다. 거대 언어 모델(Large Language Models, LLMs)을 활용하여, XFD-QAG는 사실 검증(fact-checking)을 위한 질의응답 쌍을 생성하여 가짜 뉴스 탐지를 지원한다. 질의응답 쌍은 가독성 있는 텍스트(readable text)로 생성되기 때문에, XFD-QAG는 설명 가능한 자동 가짜 뉴스 탐지 방법을 제공할 수 있다.

[0010] 가짜 뉴스 탐지 데이터셋을 사용하여 XFD-QAG와 NEP를 비교하였다. 실험 결과는 QAG가 가짜 뉴스 탐지를 지원할 수 있음을 보여주며, XFD-QAG가 NEP 보다 더 우수한 성능을 달성하는 것으로 나타났다. 또한, XFD-QAG는 인간 평가에서 우수한 성능을 보이며, 가짜 뉴스 탐지에서의 설명 가능성을 보장한다. 비록, LLM이 종종 잘못된 정보를 생성한다는 것에 대한 비판이 있지만, 본 발명은, LLM을 사실 검증을 위한 보조 도구로 활용할 수 있는 가능성을 보여 준다.

선행기술문헌

비특허문헌

[0011] (비특허문헌 0001) Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4543-4556, Dublin, Ireland. Association for Computational Linguistics.

(비특허문헌 0002) Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, Songeun Lee,

Changwoo Chun, Sungsoo Park, and Heuiseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks.

(비특허문헌 0003) Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 731-744, Dublin, Ireland. Association for Computational Linguistics.

(비특허문헌 0004) Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5073-5085, Dublin, Ireland. Association for Computational Linguistics.

(비특허문헌 0005) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

발명의 내용

해결하려는 과제

[0012] 본 발명이 이루고자 하는 기술적인 과제는 질의응답 쌍 생성을 통한 설명 가능한 가짜 뉴스 탐지 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0013] 본 발명의 일 실시예에 따른, 질의응답 쌍 생성을 통한 설명 가능한 가짜 뉴스 탐지 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 뉴스 집합에 대한 제1 필터링 동작을 수행하여 관련 뉴스들을 추출하는 단계, 상기 관련 뉴스들에 대한 제2 필터링 동작을 수행하여 참조 뉴스들을 추출하는 단계, 게시글(post)의 사실 검증에 유용한 질의를 생성하는 단계, 상기 질의에 대한 응답들을 생성하는 단계, 질의응답 쌍들 중에서 추론에 사용한 어느 하나의 질의응답 쌍을 선택하는 단계, 및 상기 어느 하나의 질의응답 쌍을 이용하여 상기 게시글의 진위 여부를 검증하는 단계를 포함한다.

발명의 효과

[0014] 본 발명의 실시예에 따른, 질의응답 쌍 생성을 통한 설명 가능한 가짜 뉴스 탐지 장치 및 방법에 의할 경우, 우수한 성능으로 가짜 뉴스를 탐지할 뿐만 아니라 추론 과정을 제공할 수 있는 효과가 있다.

도면의 간단한 설명

[0015] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른 가까 뉴스 탐지 방법을 설명하기 위한 개요도이다.

도 2는 생성된 질의응답 쌍에 대한 예시를 도시한다.

도 3은 본 발명의 일 실시예에 따른 가짜 뉴스 탐지 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0016] 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.

[0017] 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.

[0018] 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.

[0019] 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 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0020] 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.

[0021] 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.

[0022] 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.

[0023] 도 1은 본 발명의 일 실시예에 따른 가짜 뉴스 탐지 방법을 설명하기 위한 개요도이다. 도 1에서, 파란색(blue)은 뉴스 환경으로부터 도출된 명시적인 정보(explicit information)를, 빨간색(red)은 질의응답에 의해 전파된 정보(debunked information)를 나타낸다. 뉴스 환경에서 게시물의 사실적인 포인트를 포착하기 위해, 우선 게시물에서 사실적인 포인트를 찾기 위한 질문을 생성한다. 그런 다음, 각 게시물에 대한 관련 뉴스 집합을 획득한다. 생성된 질의와 정렬된 뉴스를 사용하여, 게시물 내용에 관한 뉴스 집합에서 응답을 찾는다. 마지막으로, 게시물이 가짜인지 여부를 탐지한다.

[0024] 본 발명세서는, 설명 가능한 가짜 뉴스 탐지를 위해, LLM과 QAG를 활용한다. 제안하는 프레임워크는 4 개의 파트, 즉 (1) 게시물(post) 및 뉴스 정렬(alignment), (2) 게시물 중심 질의 생성, (3) 뉴스 환경으로부터의 질의응답, 및 (4) 가짜 뉴스 탐지로 구성된다. (1), (2), 및 (3)을 위해, 본 발명에서는, 게시물에 대한 정렬된 뉴스 환경을 획득하고 질의응답 쌍 생성을 위해, ChatGPT와 같은 거대 언어 모델(LLM)을 이용한다. 이를 위한, 예시적인 LLM 프롬프트는 표 2와 같다.

[0025] [표 2]

Prompt Objective	Text
System	You are a fact checking assistant for verifying SNS post or News.
Post-News Alignment	Given the post and news, tell me whether the news is relevant to the post or not. Just say "yes" or "no". {post} {news}
Question Generation	Given this SNS post, What are the questions that could be useful for verifying the veracity?. Here is the post: "{post}". Generate questions in short and clear manner.
Question Answering	Given the SNS post and question, find the answer in the News. If you can't answer, you can say "unanswerable". {post} {news} {question}

[0026]

[0027] 예컨대, LLM에게, SNS 게시물이나 뉴스를 검증하기 위한, 사실 검증 보조원의 임무를 수행하도록 지시하고, 뉴스와 게시물의 관련성의 유무를 파악하기 위해서 주어진 게시물({post})과 뉴스({news})의 관련성 여부에 대한 이진 분류를 지시하는 프롬프트를 제공하고, 사실 검증을 위한 질의를 생성하기 위해서 주어진 게시글의 질실성을 검증하기 위해 유용한 질의를 생성하도록 하는 프롬프트를 제공하고, 질의에 대한 응답을 생성하기 위해서 주어진 게시글과 질의에 대한 응답을 뉴스에서 찾도록 하는 프롬프트를 제공할 수 있다.

[0028] 이하에서는, 제안 기법을 상세히 설명한다.

[0029] **1. 게시물-뉴스 정렬기(Post-News Aligner)**

[0030] 게시물(post, 또는 검증 발언) p와 뉴스 환경 집합 e_{news}를 입력으로 이용하여, 각 게시물을 관련된 뉴스 환경에 정렬한다. 여기서, 뉴스 환경 집합은 뉴스에 대한 집합을 의미할 수 있으며, 정렬이라 함은 뉴스 집합에서 게시물과 관련성이 있는 뉴스를 추출(또는 선택)하는 동작으로 이해될 수 있다. 예컨대, 게시물 p가 발행되기 전 T일 내에 발행되고(또는 생성되고) 게시물 p와 높은 의미적 유사도를 갖는 게시물로 뉴스 환경을 초기화할 수 있다. 그런 다음, 각 게시물-뉴스 쌍이 정렬되었는지 확인하기 위해 LLM에 프롬프트를 줌으로서, 정렬된 뉴스 환경 집합 e^{*}_{news}를 얻을 수 있다.

[0031] **2. 질의 생성(Question Generation)**

[0032] 게시물을 검증하기 위해 필요한 사실적인 포인트를 찾기 위한 질의를 생성한다. 복수의 질의, 예컨대 5 내지 10개의 사실 확인 스타일의 질의 q를 생성하기 위해 ChatGPT와 같은 거대 언어 모델(LLM)을 사용한다. 질의 생성을 통한 접근법의 장점은, 게시물에서 무엇을 검증하여야 하며 뉴스 환경에서 사실적인 포인트를 찾을 때 어디에 더 주의를 기울여야 하는지를 프레임워크에 안내할 수 있다는 것이다.

[0033] **3. 질의 응답(Question Answering)**

[0034] 전문 사실 검증자가 게시물의 진위를 체크할 때, 주요 전략은 게시물의 기저 가정을 따르거나 반박하여 증거를 찾는 것이다. 또한, 게시물의 진위를 확인하기 위한 증거는 게시물 발행의 해당 기간 내에 있어야 한다. 이를 위해, 본 발명에서는, 생성적 질의-응답 방법(generative question-answering method)을 활용한다. 게시물 p, 정렬된 환경 집합 e^{*}_{news}, 및 생성된 질의 q에 대하여, ChatGPT와 같은 거대 언어 모델(LLM)에게 게시물 내용에 관한 뉴스 환경에서 응답을 찾도록 지시한다. 세세한 질의-응답 쌍을 얻기 위하여, 두 단계로 구현한다. 첫번째 단계에서, 모든 정렬된 뉴스 환경에서 질의에 대한 각 응답을 얻는다. 질의-응답 쌍에 대한 모든 후보를 얻은 후에, 두번째 단계에서, 각 질의-응답 쌍에 대한 최상의 답변을 찾는다. 본 발명에서는, 최종 분류를 위한 최상의 질의-응답 쌍을 선택하기 위하여, 어휘 중복 점수(ROGUE) 기반 접근 방식(Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, Songeun Lee, Changwoo Chun, Sungsoo Park, and Heuiseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks., Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 731-744, Dublin, Ireland. Association for Computational Linguistics., Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5073-5085, Dublin, Ireland. Association for Computational Linguistics.)을 활용한다. 즉, 추출된 응답 후보들과 질의 쌍 집합 중에서 뉴스와의 어휘 중복 점수(ROUGE)가 높은 질의-응답 쌍을 선택할 수 있다. 최종 출력을 위해, 각 질의-응답 쌍 싸이에 참조된 뉴스 단락을 연결한 다음 각 p에 대한 $QA^* = \{(q_1, e_j^{a_1}, a_1)^*; (q_2, e_j^{a_2}, a_2)^*; ...; (q_i, e_j^{a_i}, a_i)^*\}$를 얻을 수 있다.

[0035] 생성된 질의-응답 쌍에 대한 예시는 도 2에 도시되어 있다. 도 2에서, 사실적인 포인트는 파란색으로, 제안 모델에 의해 제공되는 정보는 빨간색으로 나타내었다.

[0036] **4. 가짜 뉴스 분류기(Fake News Classifier)**

[0037] 본 발명에서는, 멀티-헤드 어텐션 레이어(Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.)와 하나의 선형 레이어로 이루어진 분류기를 미세조정(fine-tune)한다. 우선, 사전학습된 인코더의 마지막 히든 표현(hidden representation)을 평균화(averaging)하여, 게시물 표현(post representation) P = avg(Enc(p)) 및 질의응답 쌍 표현 E = avg(Enc(QA*))을 얻는다. 그런 다음, P를 쿼리(query)로 그리고 E를 키값(key·value)로 입력함으로써, qa로 보강된(enriched) 게시물 표현 P_{QA}, 즉, 멀티-헤드 어텐션 레이어를 위한 값을 획득할 수 있다. qa로 보강된 게시물 표현으로, 게시물이 가짜인지 여부를 탐지할 수 있다.

[0038] **실험 결과**

[0039] **Datasets** 실험에서는, 게시물(post)과 레이블(labels)을 위해, 영문 NEP 데이터셋(Sheng et al., 2022)을 사용한다. 뉴스 환경으로는, BigNews 데이터셋(Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1354-1374, Seattle, United States. Association for Computational Linguistics.)을 이용한다. 제안하는 뉴스 환경은 뉴스의 헤드라인 및 내용을 포함하며, 여섯 개의 언론사를 포함한다. 이는 표 3에 나타내었다. 즉, 표 3에는, 제안하는 뉴스 환경에서의 언론사 및 뉴스의 개수가 나타나 있다.

[0040] [표 3]

[0041]

Press	# news
The Washington Post	198,551
The Newyork Times	173,755
USA TODAY	170,740
AP news	279,349
FOX NEWS	330,235
The Washington Times	243,229

[0042] **Models** 본 발명에서는, BERT(Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.)-버전 NEP를 베이스라인(baseline)으로 이용한다. 제안하는 기법에서는, 게시물-뉴스 정렬기(post-news aligner), 질의 생성(question generation), 및 질의에 대한 응답(question answering)에서는 ChatGPT(gpt-3.5-turbo-0301)을 이용한다. 또한, 가짜 뉴스 분류기에서는 SimCSE(Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894-6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.)를 인코더로 사용한다.

[0043] **Fake News Detection** 제안 모델과 베이스라인 모델에 대하여, 테스트셋으로 Acc(accuracy), macF1(macro F1), PR(precision), 및 RC(recall)을 평가하였다. 표 4는 제안 모델과 베이스라인 모델의 가짜 뉴스 탐지 결과를 나타낸다. 실험 결과, 제안 모델은 Acc에서 0.71%에서 macF1에서 2.34% 까지 베이스라인을 앞서는 것으로 나타났다. 또한, 제안 모델은, 가짜 뉴스를 실제 뉴스로 잘못 분류하는 것이 가짜 뉴스 탐지에서 치명적인 오류이기 때문에, 타입 1 오류에 관하여는 우수한 성능을 보이는 것으로 판단할 수 있다. 이는, 제안 모델의 뉴스 환경에서 사실적인 포인트를 잡아낼 수 있는 능력을 증명한다.

[0044] [표 4]

[0045]

Model	Acc	macF1	RC	PR
NEP	60.93	58.62	58.91	60.04
XFD-QAG	**61.64**	**60.96**	**60.99**	**61.04**
ㄴ w/o QAG	56.5	36.10	50.00	28.25

[0046] 도 3는 본 발명의 일 실시예에 따른, 질의응답 쌍 생성을 통한 설명 가능한 가짜 뉴스 탐지 방법을 설명하기 위한 흐름도이다. 가짜 뉴스 탐지 방법은 적어도 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말하면, 가짜 뉴스 탐지 방법을 이루는 단계들 중 적어도 일부는 가짜 뉴스 탐지 장치로 명명될 수 있는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 컴퓨팅 장치는 PC(Personal), 서버(Server), 랩탑 컴퓨터, 태블릿 PC 등을 포함할 수 있다. 이하에서, 가짜 뉴스 탐지 방법을 설명함에 있어, 앞선 기재와 중복되는 내용에 관하여는, 그 구체적인 기재를 생략하기로 한다.

[0047] 우선, 뉴스에 대한 제1 필터링 동작이 수행될 수 있다(S110). 제1 필터링은, 뉴스 DB(또는 뉴스 집합)로부터 주어진 게시물과 연관성이 인정되는 뉴스를 추출하거나 필터링하는 동작을 의미할 수 있다. 이를 위해, 게시물 및 뉴스 집합의 뉴스를 벡터 형식으로 변환할 수 있다. 또한, 게시물의 발행 시점으로부터 T(T는 임의의 자연수로써, 예시적인 값은 3일 수 있음)일 이내에 발행된 뉴스로써, 임계값 이상의 유사도를 보이는 뉴스가 선택될 수 있다. 예시적인 유사도는 코사인 유사도일 수 있다. 따라서, 게시물 및/또는 각 뉴스는 발행 시점 또는 게시 시점(발행일이나 게시일을 의미할 수 있음)에 대한 정보를 포함할 수 있다. 또한, 뉴스 집합은 크롤링 동작 등을 통하여 사전에 수집된 후 컴퓨팅 장치에 저장되어 있거나, 크롤링 동작을 통하여 수집될 수 있다. 제1 필터링의 결과로써, 게시물과 추출된 뉴스 쌍들이 생성될 수 있다. 또한, 추출된 뉴스는 관련 뉴스로 명명될 수 있다.

[0048] 제1 필터링 결과에 대한 제2 필터링 동작이 수행될 수 있다(S120). 제2 필터링 동작은 거대 언어 모델(LLM)을 이용하여 수행될 수 있다. 예컨대, 주어진 게시글과 뉴스의 연관성에 관하여 이진 분류를 지시하는 프롬프트와 함께 게시글 및 뉴스를 거대 언어 모델에 입력함으로서, 제2 필터링 동작이 수행될 수 있다. 이를 위해, 컴퓨팅 장치는 거대 언어 모델(LLM) 서비스를 제공하는 서버로 상기 프롬프트(게시글과 뉴스를 포함할 수 있음)를 전송하고, 이에 대한 응답으로 관련성 유무에 대한 응답을 수신할 수 있다. 다른 예로, 거대 언어 모델은 컴퓨팅 장치에 저장되어 있을 수도 있다. S120 단계의 수행 결과로, 거대 언어 모델(LLM)이 관련성 있다고 판단한 뉴스만이 추출(또는 선택)될 수 있다.

[0049] 주어진 게시글의 사실 검증을 위한 질의가 생성된다(S130). 질의는 거대 언어 모델(LLM)을 이용하여 생성될 수 있다. 예컨대, 주어진 게시글의 진실성을 검증하기 위해 유용한 질의를 생성하도록 지시하는 프롬프트와 함께 게시글을 거대 언어 모델(LLM)에 입력함으로써, 질의가 생성될 수 있다. 이를 위해, 컴퓨팅 장치는 거대 언어 모델(LLM) 서비스를 제공하는 서버로 상기 프롬프트(게시글을 포함할 수 있음)를 전송하고, 이에 대한 응답으로 질의를 수신할 수 있다. 다른 예로, 거대 언어 모델은 컴퓨팅 장치에 저장되어 있을 수도 있다. S130 단계의 수행 결과로, 게시글의 사실 검증을 위한 적어도 하나의 질의가 생성될 수 있다.

[0050] 생성된 질의에 대한 응답이 생성된다(S140). 응답은 거대 언어 모델(LLM)을 이용하여 생성될 수 있다. 예컨대, 게시글, 뉴스(제2 필터링된 결과물을 의미할 수 있음), 및 질의에 대하여, 질의에 대한 응답을 뉴스에서 찾도록 지시하는 프롬프트와 함께 게시글, 뉴스, 및 질의를 입력함으로써, 응답이 생성될 수 있다. 이를 위해, 컴퓨팅 장치는 거대 언어 모델(LLM) 서비스를 제공하는 서버로 상기 프롬프트(게시글, 뉴스, 질의를 포함할 수 있음)를 전송하고, 이에 대한 응답으로 응답을 수신할 수 있다. 다른 예로, 거대 언어 모델은 컴퓨팅 장치에 저장되어 있을 수도 있다. S140 단계의 수행 결과로, 질의응답 쌍이 생성될 수 있다. 또한, 질의응답 쌍은 참조 뉴스를 포함하는 개념으로 이해될 수 있다.

[0051] 복수의 질의응답 쌍들 중에서 최적의 질의응답 쌍이 선택된다(S150). 최적의 질의응답 쌍은 가장 높은 ROUGE 점수를 갖는 질의응답 쌍을 의미할 수 있다.

[0052] 게시물의 진위 여부가 검증된다(S160). 진위 여부는 사전학습된 분류기를 통해 검증될 수 있다. 예컨대, 가짜 뉴스의 분류는 크로스-어텐션(Cross-Attention) 방법을 사용하여 쿼리로는 게시글을, 키밸류(key·value)로는 생성된 질의응답 쌍을 이용하여 게시글과 질의응답 쌍 간의 관련성을 담은 표현형을 학습하여 해당 표현형을 완전연결계층에 투과하여 크로스-엔트로피 손실(cross-entropy loss)을 이용하여 가짜 뉴스 탐지를 진행할 수 있다. 또한, 분류 모델은 임의의 신경망 모델로써 학습 데이터(게시글, 질의응답 쌍, 게시글의 진위 여부에 대한 레이블)를 이용하여, 게시글의 진위 여부를 추론하도록 사전에 학습된 모델을 의미할 수 있다.

[0053] 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.

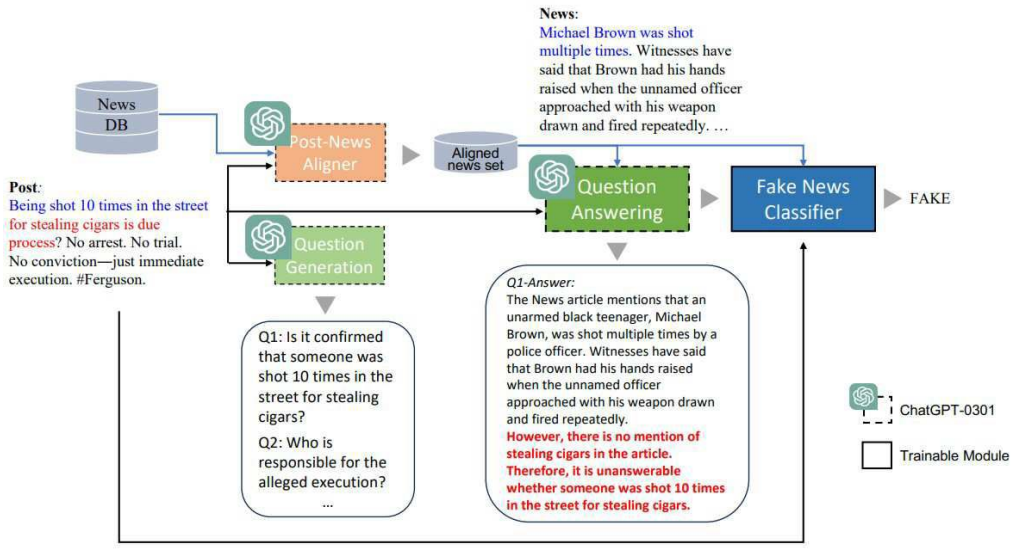
[0054] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 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.

[0055] 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.

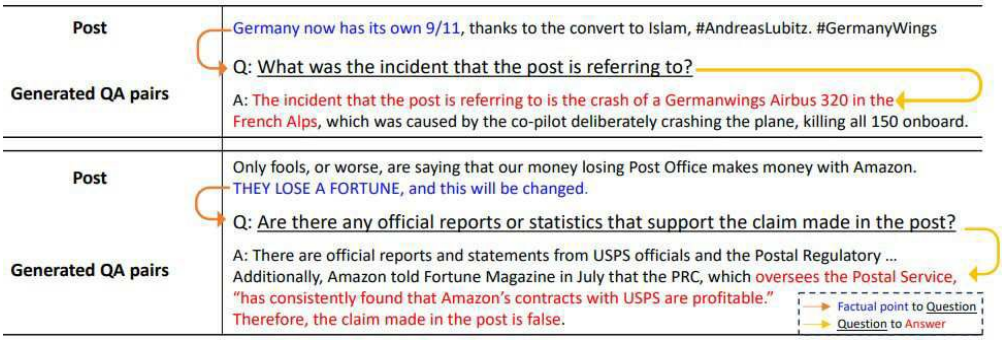
[0056] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

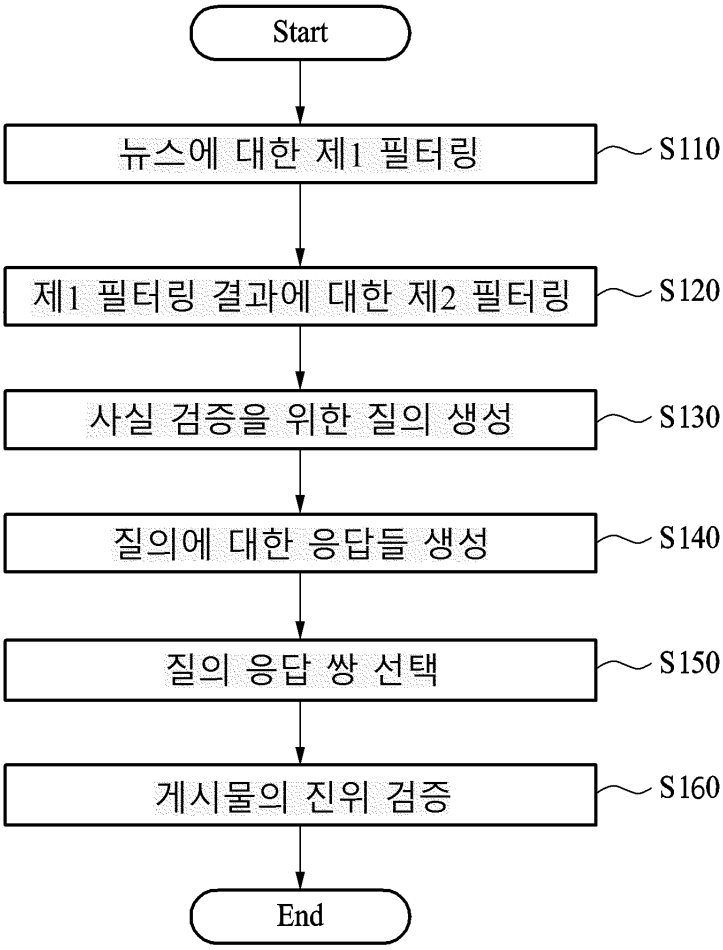
도면1


News:
Michael Brown was shot multiple times. Witnesses have said that Brown had his hands raised when the unnamed officer approached with his weapon drawn and fired repeatedly. …
News DB
Post-News Aligner
Aligned news set
Question Answering
Fake News Classifier
FAKE
Post:
Being shot 10 times in the street for stealing cigars is due process? No arrest. No trial. No conviction—just immediate execution. #Ferguson.
Question Generation
Q1: Is it confirmed that someone was shot 10 times in the street for stealing cigars?
Q2: Who is responsible for the alleged execution?
…
Q1-Answer:
The News article mentions that an unarmed black teenager, Michael Brown, was shot multiple times by a police officer. Witnesses have said that Brown had his hands raised when the unnamed officer approached with his weapon drawn and fired repeatedly.
However, there is no mention of stealing cigars in the article. Therefore, it is unanswerable whether someone was shot 10 times in the street for stealing cigars.
ChatGPT-0301
Trainable Module

도면2


Post
Germany now has its own 9/11, thanks to the convert to Islam, #AndreasLubitz. #GermanyWings
Generated QA pairs
Q: What was the incident that the post is referring to?
A: The incident that the post is referring to is the crash of a Germanwings Airbus 320 in the French Alps, which was caused by the co-pilot deliberately crashing the plane, killing all 150 onboard.
Post
Only fools, or worse, are saying that our money losing Post Office makes money with Amazon. THEY LOSE A FORTUNE, and this will be changed.
Generated QA pairs
Q: Are there any official reports or statistics that support the claim made in the post?
A: There are official reports and statements from USPS officials and the Postal Regulatory … Additionally, Amazon told Fortune Magazine in July that the PRC, which oversees the Postal Service, "has consistently found that Amazon's contracts with USPS are profitable." Therefore, the claim made in the post is false.
Factual point to Question
Question to Answer

도면3


Start
뉴스에 대한 제1 필터링
S110
제1 필터링 결과에 대한 제2 필터링
S120
사실 검증을 위한 질의 생성
S130
질의에 대한 응답들 생성
S140
질의 응답 쌍 선택
S150
게시물의 진위 검증
S160
End