



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0045657
(43) 공개일자 2020년05월06일

(51) 국제특허분류(Int. Cl.)
G06Q 50/10 (2012.01) G06N 3/02 (2019.01)
G06Q 50/00 (2018.01)
(52) CPC특허분류
G06Q 50/10 (2013.01)
G06N 3/02 (2019.01)
(21) 출원번호 10-2018-0126459
(22) 출원일자 2018년10월23일
심사청구일자 2018년10월23일

(71) 출원인
글로벌사이버대학교 산학협력단
충청남도 천안시 동남구 목천읍 교천지산길
284-88 (글로벌사이버대학교산학협력단)
(72) 발명자
강장목
서울특별시 강남구 학동로 165, 702호(논현동, 마
일스튜디오빌)
(74) 대리인
부영빈

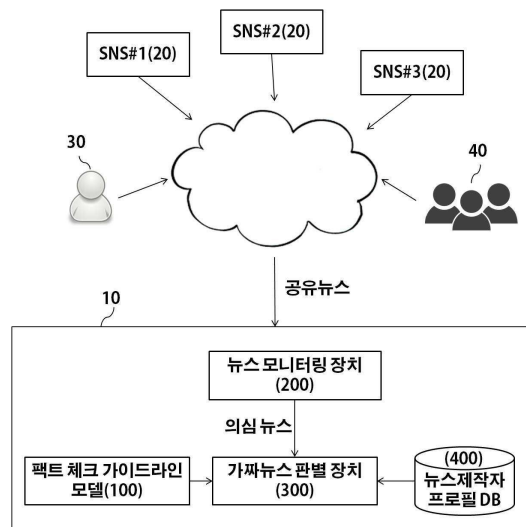
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 인공지능 기반의 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템

(57) 요약

본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템 및 방법이 개시된다. 여기서 팩트 체크는 가짜뉴스를 판단하는 휴리스틱한 방법에 기초한 접근이다. 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템은 가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하도록 모델링한 팩트 체크 가이드라인 모델과, 소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트에 공유되는 뉴스 기사를 모니터링하면서 상기 뉴스 기사의 사용자 반응을 기초로 의심 뉴스 기사를 선정하는 뉴스 모니터링 장치와 상기 선정된 의심 뉴스 기사를 분석하여 상기 팩트 체크 가이드라인 모델에 반영하여 상기 의심 뉴스 기사의 진위를 판별하는 가짜 뉴스 판별 장치를 포함하여 구성된다.

대표도 - 도1



(52) CPC특허분류

G06Q 50/01 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711073494

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 인공지능산업원천기술개발

연구과제명 인공지능 기반 가짜뉴스를 판단하는 알고리즘 설계 및 S/W 모델링

기 여 율 1/1

주관기관 남서울대학교 산학협력단

연구기간 2018.04.01 ~ 2018.12.31

명세서

청구범위

청구항 1

가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별 하도록 모델링한 팩트 체크 가이드라인 모델과,

소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트에 공유되는 뉴스 기사를 모니터링하면서 상기 뉴스 기사의 사용자 반응을 기초로 의심 뉴스 기사를 선정하는 뉴스 모니터링 장치와

상기 선정된 의심 뉴스 기사를 분석하여 상기 팩트 체크 가이드라인 모델을 적용하여 상기 의심 뉴스 기사의 진 위를 판별하는 가짜 뉴스 판별 장치를 포함하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템.

청구항 2

제 1항에 있어서, 상기 가짜 뉴스 판별 장치는

상기 의심 뉴스 기사를 자연어 처리하는 자연어 처리부와,

상기 의심 뉴스 기사의 자연어 처리 결과를 토대로 상기 팩트 체크 가이드라인 모델을 적용하여 상기 의심 뉴스 기사 내의 팩트 체크 대상을 획득하는 팩트 체크 대상 획득부와

상기 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행 하는 팩트 체크 수행부와,

상기 팩트 체크 결과에 기초하여 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 가짜 뉴스 판단부를 포함하 는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템.

청구항 3

제 2항에 있어서, 상기 팩트 체크 수행부는

상기 팩트 체크 대상의 유형이 인용문, 조사 결과, 또는 숫자나 통계치 중 적어도 하나에 해당하는 경우에, 상 기 의심 뉴스 기사 내에 상기 체크 대상의 출처가 표시되어 있는지 여부를 확인하는 출처 표시 확인 모듈과

유형별 가중치에 따라서 상기 모델에서 지정한 방법을 적용하여 상기 출처의 신뢰성 정보를 검증하는 출처 신뢰 성 검증 모듈을 포함하는, 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템.

청구항 4

제 2항에 있어서, 상기 팩트 체크 수행부는

상기 팩트 체크 대상의 유형이 금지어인 경우에, 금지어 어휘를 저장한 금지어 데이터베이스를 검색하여 상기 팩트 체크 대상이 금지어에 해당하는지 확인하는 금지어 확인 모듈과,

상기 팩트 체크 대상이 금지어로 확인되면, 상기 금지어의 예외적 허용 사유에 해당하는지 여부를 확인하는 예 외적 허용 확인 모듈을 포함하는, 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템.

청구항 5

제4항에 있어서,

상기 금지어는 사생활 침해, 비속어, 광고성 문구 및 편향된 표현, 인권 침해적 표현 중 적어도 하나를 포함하 는 것을 특징으로 하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템.

청구항 6

제2항에 있어서,

뉴스 제작자 및 소속 기자에 대한 신뢰성 정보가 저장되는 뉴스 제작자 프로필 데이터베이스를 더 포함하고, 상기 가짜 뉴스 판별 장치는

상기 뉴스 제작자 프로필 데이터베이스로부터 획득한 신뢰성 정보를 기초로 상기 의심 뉴스 기사의 제작자 신뢰성 점수를 부여하는 제작자 신뢰성 평가부를 더 포함하고,

상기 가짜 뉴스 판단부는 상기 의심 뉴스 기사의 제작자 신뢰성 평가 결과와 상기 팩트 체크 결과를 기초로 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템..

청구항 7

SNS 사이트에 공유되는 뉴스 기사를 모니터링 하면서, 상기 뉴스 기사에 대한 사용자 반응이 소정의 임계치를 초과하는 경우에, 상기 뉴스 기사의 최초 공유자를 획득하는 최초 공유자 획득부와,

상기 최초 공유자의 게시글 및 상기 뉴스 기사의 댓글을 분석하여 상기 뉴스 기사의 공유 의도를 분석하여 상기 뉴스 공유의 신뢰도를 평가하는 공유 신뢰도 평가부와,

상기 뉴스 기사의 작성일을 기초로 뉴스 기사의 시의성을 판단하는 시의성 판단부와,

상기 공유 신뢰도 평가 결과와 상기 시의성 판단 결과를 기초로 상기 뉴스 기사의 의심지수를 산출하는 의심 지수 산출부를 포함하는 뉴스 모니터링 장치.

청구항 8

제 7항에 있어서, 상기 공유 신뢰도 평가부는,

상기 최초 공유자가 게시한 공유 의견을 분석하여, 상기 공유 의견 내에 가짜 뉴스를 암시하는 표현, 부추기는 표현, 도발하는 표현 또는 호도하는 표현이 포함되는 빈도에 따라서 제1 공유 신뢰도 평가값을 산출하는 것을 특징으로 하는 뉴스 모니터링 장치.

청구항 9

제 8항에 있어서, 상기 공유 신뢰도 평가부는,

상기 뉴스 기사의 댓글 중 상기 최초 공유자를 향한 의견이 포함되는 댓글을 분석하여 부정 표현 또는 긍정 표현이 포함되는 빈도에 따라서 제2 공유 신뢰도 평가값을 산출하는 것을 특징으로 하는 뉴스 모니터링 장치.

청구항 10

제 7항에 있어서, 상기 시의성 판단부는,

상기 뉴스 기사의 작성일이나 게시일 및 뉴스 기사 공유일 간에 시차를 분석하고, 상기 시차를 기초로 상기 시의성을 판단하는 것을 특징으로 하는 뉴스 모니터링 장치.

청구항 11

소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트에 공유되는 뉴스 기사를 모니터링하면서 가짜 뉴스를 판별하는 장치에서 수행되는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법에 있어서,

뉴스 기사의 최초 공유자의 공유 의도를 분석하여, 상기 분석 결과를 기초로 의심 뉴스 기사를 선정하는 단계와,

상기 의심 뉴스 기사를 자연어 처리하는 자연어 처리 단계와,

가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하도록 모델링한 팩트 체크 가이드라인 모델을 적용하여, 상기 의심 뉴스 기사의 자연어 처리 결과를 토대로 상기 의심 뉴스 기사 내의 팩트 체크 대상을 획득하는 단계와,

상기 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행하는 단계와,

상기 팩트 체크 결과에 기초하여 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 단계를 포함하는 팩트 체크

가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 12

제 11항에 있어서, 상기 팩트 체크를 수행하는 단계는

상기 팩트 체크 대상의 유형이 인용문, 조사 결과, 또는 숫자나 통계치 중 적어도 하나에 해당하는 경우에, 상기 의심 뉴스 기사 내에 상기 체크 대상의 출처가 표시되어 있는지 여부를 확인하는 단계와,

유형별 가중치에 따라서 상기 모델에서 지정한 방법을 적용하여 상기 출처의 신뢰성을 검증하는 단계를 포함하는, 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 13

제 11항에 있어서, 상기 팩트 체크를 수행하는 단계는

상기 팩트 체크 대상의 유형이 금지어인 경우에, 금지어 데이터베이스를 검색하여 금지어에 해당하는지 확인하는 단계와,

상기 팩트 체크 대상이 금지어로 확인되면, 상기 금지어의 예외적 허용 규칙에 따라서 사용되었는지 여부를 확인하는 단계를 포함하는, 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 14

제 13항에 있어서,

상기 금지어는 사생활 침해, 비속어, 광고성 문구 및 편향된 표현, 인권 침해적 표현 중 적어도 하나를 포함하는 것을 특징으로 하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 15

제 11항에 있어서,

뉴스 제작자 및 소속 기자에 대한 신뢰성 정보가 저장되는 뉴스 제작자 프로필 데이터베이스로부터 획득한 신뢰성 정보를 기초로 상기 의심 뉴스 기사의 제작자 신뢰성 점수를 부여하는 제작자 신뢰성 평가단계를 더 포함하고,

상기 가짜 뉴스 여부를 판단하는 단계는 상기 의심 뉴스 기사의 제작자 신뢰성 평가 결과와 상기 팩트 체크 결과를 기초로 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 16

제 11항에 있어서, 의심 뉴스 기사 선정 단계는,

상기 SNS 사이트에 공유되는 뉴스 기사를 모니터링하면서, 상기 뉴스 기사에 대한 사용자 반응이 소정의 임계치를 초과하는 경우에, 상기 뉴스 기사의 최초 공유자를 추적하여 획득하는 단계와,

상기 최초 공유자가 게시한 공유 의견 및 상기 뉴스 기사의 댓글을 분석하여 상기 뉴스 기사의 공유 의도를 분석함으로써, 상기 뉴스 공유의 신뢰도를 평가하는 단계와,

상기 뉴스 기사의 작성일을 기초로 상기 뉴스 기사의 시의성을 판단하는 단계와,

상기 공유 신뢰도 평가 결과와 상기 시의성 판단 결과를 기초로 상기 뉴스 기사의 의심지수를 산출하는 단계를 포함하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 17

제 16항에 있어서, 상기 뉴스 공유의 신뢰도를 평가하는 단계는,

상기 최초 공유자가 게시한 공유 의견을 분석하여, 상기 공유 의견 내에 가짜 뉴스를 암시하는 표현, 부추기는 표현, 도발하는 표현 또는 호도하는 표현이 포함되는 빈도에 따라서 제1 공유 신뢰도 평가값을 산출하는 것을 특징으로 하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 18

제 17항에 있어서, 상기 공유 신뢰도 평가부는,

상기 뉴스 기사의 댓글 중 상기 최초 공유자를 향한 의견이 포함되는 댓글을 분석하여 부정 표현 또는 긍정 표현이 포함되는 빈도에 따라서 제2 공유 신뢰도 평가값을 산출하는 것을 특징으로 하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 19

제 16항에 있어서, 상기 시의성을 판단하는 단계는,

상기 뉴스 기사의 작성일, 뉴스 기사 게시일 및 뉴스 기사 공유일 간에 시차를 분석하고, 상기 시차를 기초로 상기 시의성을 판단하는 것을 특징으로 하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법.

청구항 20

SNS 사이트에 공유되는 뉴스 기사를 모니터링하면서, 팩트 체크 가이드라인을 기초로 가짜 뉴스를 판별하는 장치에서 수행되는 가짜 뉴스 판별 방법을 실행하는 프로그램이 기록되어 컴퓨터로 읽을 수 있는 기록매체에 있어서, 상기 가짜 뉴스 판별 방법은

뉴스 기사의 최초 공유자의 공유 의도를 분석하여, 상기 분석 결과를 기초로 의심 뉴스 기사를 선정하는 의심 뉴스 기사 선정 단계와,

상기 의심 뉴스를 자연어 처리하는 자연어 처리 단계와,

가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하도록 모델링한 팩트 체크 가이드라인 모델을 적용하여, 상기 의심 뉴스 기사의 자연어 처리 결과를 토대로 상기 의심 뉴스 기사 내의 팩트 체크 대상을 획득하는 단계와,

상기 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행하는 단계와,

상기 팩트 체크 결과에 기초하여 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 단계를 포함하는 것을 특징으로 하는 컴퓨터로 읽을 수 있는 기록매체

발명의 설명

기술 분야

[0001] 본 발명은 팩트 체크 가이드라인 또는 국가 기관 등 검증된 단체를 통해 제공된 뉴스 제작 및 보도 지침 등의 가이드라인을 이용한 가짜 뉴스 판별 시스템에 관한 것으로, 더욱 상세하게는 인공 지능 기반의 팩트체크 가이드라인을 이용하여 가짜 뉴스를 판별하는 시스템에 관한 것이다.

배경 기술

[0002] 일반적으로 가짜 뉴스는 정식 언론 매체나 뉴스 편집 권한을 갖는 포털 사이트를 통해서 유포되기보다는, 페이스북, 트위터, 카카오톡 스토리와 같은 소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트나 인스턴트 메시지의 공개 채팅방, 밴드와 카카오톡 같은 폐쇄형 소셜 망 등을 통해서 유통된다.

[0003] SNS나 메시지를 통한 가짜 뉴스의 유포는 확산속도나 파급력이 크지만 사이버 친구나 연락처를 주고받는 관계에 있는 신뢰할 수 있는 지인을 통해서 유포되는 경우가 많기 때문에, 사용자들은 의심 없이 뉴스의 내용을 사실로서 받아들이게 된다. 또한, 공식 언론 매체에서 제공하는 신뢰성 있는 뉴스 기사의 형식을 갖추고 있기 때문에 진위여부를 판단하기 어렵다.

[0004] 이러한 문제를 해결하기 위하여 최근에는 자연어 처리기술을 통해서 뉴스의 문장을 분석하고, 문맥상 불일치 하는 문장을 찾거나, 신뢰할 수 있는 유사한 뉴스와 비교하여 가짜 뉴스임을 판단하는 방법이 제시되고 있다. 여기서 신뢰할 수 있는 유사한 뉴스란 예를 들어 ‘김정은’에 대한 사생활 등에 대한 블로그 뉴스라면 이와 유사

한 주제나 내용의 조선일보 또는 경향신문 등 대표적인 한국의 미디어 매체에서도 언급된 사실인지를 비교하고 분석하는 방식으로 가짜 유무를 판단할 수 있다. 물론 이 방법으로 가짜 유무를 단정할 수 없지만 초기 단계로 가짜 뉴스의 확률이 높은 뉴스기사로 분류하는데 이용될 수도 있다. 또한, SNS 사이트의 사용자 반응을 분석하여, 반응을 기초로 가짜 뉴스 여부를 판단하는 기술이 제안되고 있다.

- [0005] 그러나 SNS에 공유되는 전체 문서에 대하여 일일이 사용자 반응을 분석하는 것은 많은 연산 시간을 요하기 때문에 현실적으로 적용이 어려운 문제가 있으며, 문맥을 분석하여 오류를 찾아내거나, 신뢰할 수 있는 유사한 뉴스와 비교하여 진위를 판단하는 방법은 의미 분석을 위한 자연어처리에 많은 연산 시간이 소요되는 문제가 있고, 날이 갈수록 교묘하게 진보하는 가짜 뉴스의 유형에 적시에 적용하기 어려운 문제가 있다.

발명의 내용

해결하려는 과제

- [0006] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은, 많은 연산시간이 소요되지 않고 다양하게 변화하는 가짜 뉴스에 적용 가능한 가짜 뉴스를 판별하는 시스템을 제공하는 데 있다.
- [0007] 상기와 같은 문제점을 해결하기 위한 본 발명의 다른 목적은, 많은 연산시간이 소요되지 않고 다양하게 변화하는 가짜 뉴스에 적용 가능한 가짜 뉴스를 모니터링하는 장치를 제공하는 데 있다.
- [0008] 상기와 같은 문제점을 해결하기 위한 본 발명의 또 다른 목적은, 많은 연산시간이 소요되지 않고 다양하게 변화하는 가짜 뉴스에 적용 가능한 가짜 뉴스를 판별하는 방법을 제공하는 데 있다.

과제의 해결 수단

- [0009] 상기 목적을 달성하기 위한 본 발명은, 가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하도록 모델링한 팩트 체크 가이드라인 모델과, 소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트에 공유되는 뉴스 기사를 모니터링하면서 상기 뉴스 기사의 사용자 반응을 기초로 의심 뉴스 기사를 선정하는 뉴스 모니터링 장치와 상기 선정된 의심 뉴스 기사를 분석하여 상기 팩트 체크 가이드라인 모델을 적용하여 상기 의심 뉴스 기사의 진위를 판별하는 가짜 뉴스 판별 장치를 포함하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템을 제공한다.
- [0010] 여기서, 상기 가짜 뉴스 판별 장치는 상기 의심 뉴스 기사를 자연어 처리하는 자연어 처리부와, 상기 의심 뉴스 기사의 자연어 처리 결과를 토대로 상기 팩트 체크 가이드라인 모델을 적용하여 상기 의심 뉴스 기사 내의 팩트 체크 대상을 획득하는 팩트 체크 대상 획득부와 상기 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행하는 팩트 체크 수행부와, 상기 팩트 체크 결과에 기초하여 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 가짜 뉴스 판단부를 포함한다.
- [0011] 또한 여기서, 상기 팩트 체크 수행부는 상기 팩트 체크 대상의 유형이 인용문, 조사 결과, 또는 숫자나 통계치 중 적어도 하나에 해당하는 경우에, 상기 의심 뉴스 기사 내에 상기 팩트 체크 대상의 출처가 표시되어 있는지 여부를 확인하는 출처 표시 확인 모듈과 유형별 가중치에 따라서 상기 모델에서 지정한 방법을 적용하여 상기 출처의 신뢰성 정보를 검증하는 출처 신뢰성 검증 모듈을 포함한다.
- [0012] 또한 여기서, 상기 팩트 체크 수행부는 상기 팩트 체크 대상의 유형이 금지어인 경우에, 금지어 어휘를 저장한 금지어 데이터베이스를 검색하여 상기 팩트 체크 대상이 금지어에 해당하는지 확인하는 금지어 확인 모듈과, 상기 팩트 체크 대상이 금지어로 확인되면, 상기 금지어의 예외적 허용 사유에 해당하는지 여부를 확인하는 예외적 허용 확인 모듈을 포함한다.
- [0013] 또한 여기서, 상기 금지어는 사생활 침해, 비속어, 광고성 문구 및 편향된 표현, 인권 침해적 표현 중 적어도 하나를 포함한다.
- [0014] 또한 여기서, 뉴스 제작자 및 소속 기자에 대한 신뢰성 정보가 저장되는 뉴스 제작자 프로필 데이터베이스를 더 포함하고, 상기 가짜 뉴스 판별 장치는 상기 뉴스 제작자 프로필 데이터베이스로부터 획득한 신뢰성 정보를 기초로 상기 의심 뉴스 기사의 제작자 신뢰성 점수를 부여하는 제작자 신뢰성 평가부를 더 포함하고, 상기 가짜 뉴스 판단부는 상기 의심 뉴스 기사의 제작자 신뢰성 평가 결과와 상기 팩트 체크 결과를 기초로 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단한다. 만약 블로거 등 정식 뉴스 기자 데이터 베이스에 등록되지 않은 경우에는 그 빈도, 팔로워수 등 네트워크 파워 등을 기자의 영향력 지수로 삼을 수 있다.

- [0015] 상기 다른 목적을 달성하기 위한 본 발명은 SNS 사이트에 공유되는 뉴스 기사를 모니터링 하면서, 상기 뉴스 기사에 대한 사용자 반응이 소정의 임계치를 초과하는 경우에, 상기 뉴스 기사의 최초 공유자를 획득하는 최초 공유자 획득부와, 상기 최초 공유자의 게시글 및 상기 뉴스 기사의 댓글을 분석하여 상기 뉴스 기사의 공유 의도를 분석하여 상기 뉴스 공유의 신뢰도를 평가하는 공유 신뢰도 평가부와, 상기 뉴스 기사의 작성일을 기초로 뉴스 기사의 시의성을 판단하는 시의성 판단부와, 상기 공유 신뢰도 평가 결과와 상기 시의성 판단 결과를 기초로 상기 뉴스 기사의 의심지수를 산출하는 의심 지수 산출부를 포함하는 뉴스 모니터링 장치를 제공한다.
- [0016] 여기서, 상기 공유 신뢰도 평가부는, 상기 최초 공유자가 게시한 공유 의견을 분석하여, 상기 공유 의견 내에 가짜 뉴스를 암시하는 표현, 부추기는 표현, 도발하는 표현 또는 호도하는 표현이 포함되는 빈도에 따라서 제1 공유 신뢰도 평가값을 산출하는 것을 특징으로 한다.
- [0017] 또한 여기서, 상기 공유 신뢰도 평가부는, 상기 뉴스 기사의 댓글 중 상기 최초 공유자를 향한 의견이 포함되는 댓글을 분석하여 부정 표현 또는 긍정 표현이 포함되는 빈도에 따라서 제2 공유 신뢰도 평가값을 산출하는 것을 특징으로 한다. 댓글은 현재는 공개된 댓글을 크롤링하고 해당 단어, 문장 등이 갖는 차가움과 뜨거움, 공격성과 방어성 등 어휘 사전의 긍정과 부정 등을 중심으로 비교하여 값을 처리한다.
- [0018] 또한 여기서, 상기 시의성 판단부는, 상기 뉴스 기사의 작성일이나 게시일 및 뉴스 기사 공유일 간에 시차를 분석하고, 상기 시차를 기초로 상기 시의성을 판단하는 것을 특징으로 한다.
- [0019] 상기 또 다른 목적을 달성하기 위한 본 발명은 SNS 사이트에 공유되는 뉴스 기사를 모니터링하면서 가짜 뉴스를 판별하는 장치에서 수행되는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법에 있어서, 뉴스 기사의 최초 공유자의 공유 의도를 분석하여, 상기 분석 결과를 기초로 의심 뉴스 기사를 선정하는 단계와, 상기 의심 뉴스 기사를 자연어 처리하는 자연어 처리 단계와, 가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하도록 모델링한 팩트 체크 가이드라인 모델을 적용하여, 상기 의심 뉴스 기사의 자연어 처리 결과를 토대로 상기 의심 뉴스 기사 내의 팩트 체크 대상을 획득하는 단계와, 상기 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행하는 단계와, 상기 팩트 체크 결과에 기초하여 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단하는 단계를 포함하는 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법을 제공한다.
- [0020] 여기서, 상기 팩트 체크를 수행하는 단계는 상기 팩트 체크 대상의 유형이 인용문, 조사 결과, 또는 숫자나 통계치 중 적어도 하나에 해당하는 경우에, 상기 의심 뉴스 기사 내에 상기 팩트 체크 대상의 출처가 표시되어 있는지 여부를 확인하는 단계와, 유형별 가중치에 따라서 상기 모델에서 지정한 방법을 적용하여 상기 출처의 신뢰성을 검증하는 단계를 포함한다. 상기 출처 중 URL인 경우에는 URL 연결이 살아 있는 경우와 끊겨진 경우 즉 늘 변함없이 서비스가 되는 경우를 좀 더 신뢰할 수 있고 진짜 뉴스에 가깝다고 가정하고 가중치를 부여할 수 있다.
- [0021] 또한 여기서, 상기 팩트 체크를 수행하는 단계는 상기 팩트 체크 대상의 유형이 금지어인 경우에, 금지어 데이터베이스를 검색하여 금지어에 해당하는지 확인하는 단계와, 상기 팩트 체크 대상이 금지어로 확인되면, 상기 금지어의 예외적 허용 규칙에 따라서 사용되었는지 여부를 확인하는 단계를 포함한다.
- [0022] 또한 여기서, 상기 금지어는 사생활 침해, 비속어, 광고성 문구 및 편향된 표현, 인권 침해적 표현 중 적어도 하나를 포함하는 것을 특징으로 한다.
- [0023] 또한 여기서, 상기 뉴스 제작자 및 소속 기자에 대한 신뢰성 정보가 저장되는 뉴스 제작자 프로필 데이터베이스로부터 획득한 신뢰성 정보를 기초로 상기 의심 뉴스 기사의 제작자 신뢰성 점수를 부여하는 제작자 신뢰성 평가단계를 더 포함하고, 상기 가짜 뉴스 여부를 판단하는 단계는 상기 의심 뉴스 기사의 제작자 신뢰성 평가 결과와 상기 팩트 체크 결과를 기초로 상기 의심 뉴스 기사의 가짜 뉴스 여부를 판단한다.
- [0024] 또한 여기서, 의심 뉴스 기사 선정 단계는, 상기 SNS 사이트에 공유되는 뉴스 기사를 모니터링하면서, 상기 뉴스 기사에 대한 사용자 반응이 소정의 임계치를 초과하는 경우에, 상기 뉴스 기사의 최초 공유자를 추적하여 획득하는 단계와, 상기 최초 공유자가 게시한 공유 의견 및 상기 뉴스 기사의 댓글을 분석하여 상기 뉴스 기사의 공유 의도를 분석함으로써, 상기 뉴스 공유의 신뢰도를 평가하는 단계와, 상기 뉴스 기사의 작성일을 기초로 상기 뉴스 기사의 시의성을 판단하는 단계와, 상기 공유 신뢰도 평가 결과와 상기 시의성 판단 결과를 기초로 상기 뉴스 기사의 의심지수를 산출하는 단계를 포함한다.
- [0025] 또한 여기서, 상기 뉴스 공유의 신뢰도를 평가하는 단계는, 상기 최초 공유자가 게시한 공유 의견을 분석하여, 상기 공유 의견 내에 가짜 뉴스를 암시하는 표현, 부추기는 표현, 도발하는 표현 또는 호도하는 표현이 포함되

는 빈도에 따라서 제1 공유 신뢰도 평가값을 산출하는 것을 특징으로 한다.

[0026] 또한 여기서, 상기 공유 신뢰도 평가부는, 상기 뉴스 기사의 댓글 중 상기 최초 공유자를 향한 의견이 포함되는 댓글을 분석하여 부정 표현 또는 긍정 표현이 포함되는 빈도에 따라서 제2 공유 신뢰도 평가값을 산출하는 것을 특징으로 한다.

[0027] 또한 여기서, 상기 시의성을 판단하는 단계는, 상기 뉴스 기사의 작성일, 뉴스 기사 게시일 및 뉴스 기사 공유 일 간에 시차를 분석하고, 상기 시차를 기초로 상기 시의성을 판단하는 것을 특징으로 한다.

발명의 효과

[0028] 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템 및 방법을 이용하면, SNS를 모니터링하면서 사용자 반응이 소정의 임계치를 초과하는 뉴스 기사를 선택해서, 상기 뉴스 기사의 최초 공유자의 공유 의도와 사용자 반응을 분석하여 의심 뉴스 기사를 선정함으로써, 분석 대상을 1차적으로 필터링하고, 빅데이터 분석을 통해 생성된 팩트 체크 가이드라인 모델을 적용하여 팩트 체크 대상 문장을 선정하는 2차 필터링 과정을 거치고, 다음으로 팩트 체크 대상에 대하여 팩트 체크 가이드라인 모델에서 제시하는 체크 방법으로 검증을 함으로써, 3단계 필터링 과정을 거침으로써 가짜 뉴스를 신속하고 정확하게 판별할 수 있는 장점이 있다. 상기한 방법의 각 단계는 개별의 가이드라인이면서 이들의 다양한 조합을 통해 가짜뉴스를 보다 효과적이고 저렴하게 발견할 수 있게 돕는다.

도면의 간단한 설명

[0029] 도 1은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템의 개념도이다.

도 2는 본 발명의 일 실시예에 따른 뉴스 모니터링 장치의 구성을 보여주는 블록도이다.

도 3은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치의 구성을 보여주는 블록도이다.

도 4는 본 발명의 일 실시예에 따른 팩트 체크 가이드라인 모델을 적용하여 팩트 체크 대상을 도출하는 예를 보여주는 도면이다.

도 5는 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치의 일 구성인 팩트 체크 수행부의 블록도이다.

도 6은 본 발명의 일 실시예에 따른 팩트 체크 유형별 체크 방법을 도출한 예를 보여주는 도면이다.

도 7은 본 발명의 일 실시예에 따른 팩트 체크 대상으로 선정될 수 있는 금지어 유형의 예를 보여주는 도면이다.

도 8은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법이 수행되는 과정을 보여주는 순서도이다.

도 9는 본 발명의 일 실시예에 따른 의심 뉴스를 선정하는 과정을 보여주는 순서도이다.

도 10은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인에 따라 팩트 체크하는 과정을 보여주는 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0030] 본 발명은 다양한 변형을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변형, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0031] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 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 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0032] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에

직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0033] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0034] 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 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상의 의미와 일치하는 의미로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0036] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0037] 도 1은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템의 개념도이다.

[0038] 도 1을 참조하면, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템(10)은 소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트(20)에서 사용자(30)에 의해 공유되는 뉴스 기사를 모니터링하면서 최초 뉴스 공유자(10)의 뉴스 공유 의도를 분석하여 의심 뉴스 기사를 선정하고, 팩트 체크 가이드라인을 기초로 의심 뉴스 기사를 분석하여 가짜 뉴스인지 여부를 판별할 수 있다. 예를 들면, 평소 활동이 없던 A씨가 갑자기 다수의 특정 주제 뉴스만을 퍼 나르거나, 오늘 급히 만든 ID인 B씨가 갑작스럽게 특정 기사만 퍼 나르는 경우가 있을 수 있다.

[0039] 또한, 도 1을 참조하면, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템(10)은 팩트 체크 가이드라인 모델(100), 뉴스 모니터링 장치(200), 가짜뉴스 판별 장치(300) 및 뉴스 제작자 프로필 데이터베이스(400)를 포함하여 구성될 수 있다.

[0040] 팩트 체크 가이드라인 모델(100)은 가짜 뉴스 및 진짜 뉴스 데이터를 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하도록 모델링한 모델이다. 여기서 팩트 체크는 가짜뉴스를 판단하는 휴리스틱한 방법에 기초한 접근이다

[0041] 뉴스 모니터링 장치(200)는 소셜 네트워킹 서비스(Social Networking Service, SNS) 사이트에 공유되는 뉴스 기사를 모니터링하면서 상기 뉴스 기사의 공유 신뢰도 분석 결과를 기초로 의심 뉴스 기사를 선정하여 가짜뉴스 판별 장치(300)로 전달할 수 있다. 예를 들면, 최초 공유자(30)의 공유 의도를 분석하고, 공유된 기사에 대한 댓글(40)을 분석하여 공유 신뢰도를 평가할 수 있다. 자세한 내용은 후술하기로 한다.

[0042] 가짜뉴스 판별 장치(300)는 뉴스 모니터링 장치(200)에서 선정된 의심 뉴스 기사를 분석하여 팩트 체크 가이드라인 모델(100)을 적용하여, 의심 뉴스 기사의 진위를 판별할 수 있다.

[0044] 이하, 도면을 참조하여 뉴스 모니터링 장치(200)에 대하여 좀 더 상세하게 설명한다.

[0045] 도 2는 본 발명의 일 실시예에 따른 뉴스 모니터링 장치의 구성을 보여주는 블록도이다.

[0046] 도 2를 참조하면, 본 발명의 일 실시예에 따른 뉴스 모니터링 장치(200)는 최초 공유자 추적부(210), 공유 신뢰도 평가부(220), 시의성 판단부(230) 및 의심 지수 산출부(240)를 포함하여 구성될 수 있다.

[0047] 최초 공유자 추적부(210)는 상기 SNS 사이트(20)에 공유되는 뉴스 기사를 모니터링하면서, 공유되는 뉴스 기사에 대한 사용자 반응(40)이 소정의 임계치를 초과하는 경우에, 상기 뉴스 기사의 최초 공유자(30)를 추적할 수 있다.

[0048] 예를 들면, 페이스북, 트위터, 카카오톡과 같은 SNS 사이트에 공유한 뉴스 기사에 대한 공감표현(예를 들면, 좋아요, 놀라워요, 등), 댓글, 재공유, 트윗/리트윗 등을 분석하여 소정의 임계치를 초과하면 해당 뉴스

기사를 최초로 공유한 사용자(30)를 추적할 수 있다. 또한, 최근에는 스와이프 기술을 활용하여, 좋거나/싫다는 감정의 표시로서, 손가락을 이용하여 오른쪽이나 왼쪽으로 사진이나 글을 밀어서 버리는 방식의 사용자 인터페이스를 사용할 수 있다. 이런 경우 동작기반(예를 들면, 스와이프 기반)의 선호도 분석이 가능할 수 있다.

- [0049] 공유 신뢰도 평가부(220)는 최초 공유자(30)가 공유하면서 게시한 공유 의견과 공유한 뉴스 기사의 댓글을 분석하여 최초 공유자(30)의 공유 의도를 파악하여 뉴스 공유의 신뢰도를 평가할 수 있다.
- [0050] 즉, 공유 신뢰도 평가부(220)는 최초 공유자가 게시한 공유 의견을 분석하여, 공유 의견 내에 가짜 뉴스를 암시하는 표현, 부추기는 표현, 도발하는 표현 또는 호도하는 표현이 포함되는 빈도에 따라서 제1 공유 신뢰도 평가값을 산출할 수 있다.
- [0051] 또한, 공유 신뢰도 평가부(220)는 상기 뉴스 기사의 댓글 중 최초 공유자(30)를 향한 의견이 포함되는 댓글을 분석하여 부정 표현 또는 긍정 표현이 포함되는 빈도에 따라서 제2 공유 신뢰도 평가값을 산출하여, 제1 공유 신뢰도 평가값과 제2 공유 신뢰도 평가값을 기초로 뉴스 공유의 신뢰도를 평가할 수 있다.
- [0053] 예를 들면, SNS 에서 뉴스를 공유하는 일반적인 사용자는 뉴스 기사의 논조나 내용이 자신의 정치적, 경제적, 문화적, 직업적 이해관계나 소신과 일치하는 경우에 해당 뉴스 기사를 공유하여, 많은 사람에게 퍼트리고 공감을 받고자 하는 의도가 있을 수 있다. 따라서 공유 의견에는 기사의 내용을 일부 발췌하거나, 기사의 내용에 대한 자신의 주관적인 의견이 포함될 수 있다. 이러한 기사에 대한 댓글 역시 기사의 내용에 대한 의견 표현이 포함될 수 있다. 이러한 선의의 공유는 뉴스 공유의 신뢰도를 높게 평가할 수 있다.
- [0054] 반면, 자신의 성향과 맞지 않거나 가짜 뉴스로 의심되는 뉴스 기사를 굳이 공유해서 퍼트리려고 하는 사람은 많지 않을 수 있지만, 만일 사용자가 뉴스 기사의 내용이 사실과 다른 가짜 뉴스임을 알면서도 경고의 의미로 공유하는 경우라면, 공유 의견에 가짜 뉴스임을 암시하는 문구나 조심하라는 경고하는 문구가 포함될 수 있고, 이에 대한 댓글의 반응을 통해서 사용자 의도를 좀 더 확실하게 파악할 수 있다. 이 경우는 사용자의 의견에 가짜 뉴스임이 나타나 있을 수 있기 때문에 뉴스 공유의 신뢰도를 상당히 낮게 평가할 수 있다.
- [0055] 위의 경우와 다르게, 오직 사실 관계를 호도하거나 특정 여론을 조성하기 위해서 가짜 뉴스를 공유하는 것이라면, 공유 의견 내에 가짜 뉴스를 암시하는 문구는 없을 수 있다. 그러나 원색적이거나 도발적인 문구를 포함할 수 있고, 해당 뉴스 기사를 많이 퍼뜨리도록 부추기는 표현을 포함할 수 있다. 이에 대한 댓글은 최초 공유자나 뉴스 작성자에 대한 비난이나 격려 등의 글이 포함될 수 있다. 이와 같은 악의적인 공유는 뉴스 공유의 신뢰도가 낮게 평가될 수 있다.
- [0056] 댓글은 공개된 댓글을 크롤링하고 해당 단어, 문장 등이 갖는 차가움과 뜨거움, 공격성과 방어성 등 어휘 사전의 긍정과 부정 등을 중심으로 비교하여 값을 처리할 수 있다.
- [0057] 시의성 판단부(230)는 공유된 뉴스 기사의 작성일을 기초로 뉴스 기사의 시의성을 판단할 수 있다. 즉, 이미 오래 전에(몇년전, 또는 몇달전) 작성되거나 게시된 뉴스여서 뉴스로서의 가치가 없는 기사를 공유하는 경우라면, 여론을 호도하기 위한 가짜 뉴스로 의심할 수 있다. 따라서, 상기 뉴스 기사의 작성일이나 뉴스 기사 게시일과 뉴스 기사 공유일 간에 시차를 분석하고, 시차를 기초로 뉴스 기사의 시의성을 판단하되, 시차가 클수록 뉴스 기사의 시의성이 낮게 판단될 수 있다. 즉, 오래되어서 뉴스로서의 가치가 떨어지는 시의성 없는 기사의 특정 부분만을 인용하고 이를 짜깁기하여 자신의 가짜 뉴스 정보에 설득력을 얻으려는 방법으로 간주할 수 있다.
- [0058] 의심 지수 산출부(240)는 공유 신뢰도 평가 결과와 상기 시의성 판단 결과를 기초로 상기 뉴스 기사의 의심지수를 산출할 수 있다. 공유 신뢰도와 시의성이 낮을수록 의심 지수가 높게 산출될 수 있으며, 산출된 의심 지수가 소정의 임계치를 초과하는 경우에는 상기 뉴스 기사의 가짜 여부 판정을 위하여 가짜뉴스 평가장치(300)로 전달될 수 있다.
- [0060] 도 3은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치의 구성을 보여주는 블록도이다.
- [0061] 도 4는 본 발명의 일 실시예에 따른 팩트 체크 가이드라인 모델을 적용하여 팩트 체크 대상을 도출하는 예를 보여주는 도면이다.
- [0062] 도 3을 참조하면 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치(300)는 자연

어 처리부(310), 팩트 체크 대상 획득부(320), 팩트 체크 수행부(330), 제작자 신뢰성 평가부(340) 및 가짜 뉴스 판단부(350)를 포함하여 구성될 수 있다. 또한, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치(300)는 뉴스 모니터링 장치(200)에서 선정한 의심 뉴스 기사를 전달받아서, 팩트 체크 가이드라인 모델(100)을 적용하여 그 결과를 기초로 의심 뉴스 기사의 진위를 판별할 수 있다.

- [0063] 이하, 도 3을 참조하여 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치의 구성과 각 구성요소간의 관계를 설명한다.
- [0064] 자연어 처리부(310)는 팩트 체크 가이드라인 모델에 적용할 수 있도록 의심 뉴스를 자연어 처리할 수 있다. 자연어 처리 방법은 알려진 기술이므로 상세한 설명을 생략한다.
- [0065] 팩트 체크 대상 획득부(320)는 의심 뉴스 기사의 자연어 처리 결과를 토대로 팩트 체크 가이드라인 모델(100)을 적용하여 의심 뉴스 기사 내의 팩트 체크 대상을 획득할 수 있다.
- [0066] 팩트 체크 가이드라인 모델(100)은 가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 빅데이터 분석하여 가짜 뉴스의 패턴을 도출하고, 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하기 위한 가이드라인을 제시하도록 모델링한 모델이다. 더 나아가 인공지능 기술을 이용하여 딥러닝한 결과로 가이드 라인을 새롭게 작성하여 제시할 수도 있다.
- [0067] 따라서, 자연어 처리된 의심 뉴스 기사에 팩트 체크 가이드라인 모델(100)을 적용하여 의심스럽거나 추가적인 검증이 필요한 문장이나 어휘가 팩트 체크 대상으로 추출될 수 있다.
- [0068] 예를 들면, 도 4의 팩트 체크 가이드라인 모델 적용 결과(100')를 참조하면, 팩트 체크 대상 획득부(320)가 의심 뉴스 기사를 팩트 체크 가이드라인 모델(100)에 적용하면, 모델(100)은 의심 뉴스 기사를 점검하여 적어도 하나의 팩트 체크 대상(101)이 되는 문장이나 어휘를 추출할 수 있다. 팩트 체크 대상(401)은 모델(100)에서 지정한 팩트 체크 유형(102)(예를 들면, 인용문, 연구결과, 통계 등)으로 분류될 수 있으며, 팩트 체크 대상(401)별로 팩트 체크 방법(103)과 가중치(104)가 정의될 수 있다.
- [0069] 팩트 체크 수행부(330)는 팩트 체크 대상 획득부(320)에서 획득한 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행할 수 있다. 좀 더 상세한 내용은 후술한다.
- [0070] 제작자 신뢰성 평가부(340)는 뉴스 제작자 및 소속 기자에 대한 신뢰성 정보가 저장되는 뉴스 제작자 프로필 데이터베이스(400)로부터 획득한 신뢰성 정보를 기초로 의심 뉴스 기사의 제작자 신뢰성 점수를 부여할 수 있다. 만약 블로거 등 정식 뉴스 기자 데이터 베이스에 등록되지 경우에는 그 빈도, 팔로워수 등 네트워크 파워 등을 기자의 영향력 지수로 삼을 수 있다.
- [0071] 가짜 뉴스 판단부(350)는 팩트 체크 수행부(330)의 팩트 체크 결과와 제작자 신뢰성 평가부(340)의 제작자 신뢰성 평가 결과에 기초하여 의심 뉴스 기사의 가짜 뉴스 여부를 판단할 수 있다.
- [0073] 이하, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치의 일 구성인 팩트 체크 수행부(330)의 구성과 기능에 좀 더 상세하게 설명한다.
- [0074] 도 5는 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 장치의 일 구성인 팩트 체크 수행부의 블록도이다.
- [0075] 도 6은 본 발명의 일 실시예에 따른 팩트 체크 유형별 체크 방법을 도출한 예를 보여주는 도면이다.
- [0076] 도 7은 본 발명의 일 실시예에 따른 팩트 체크 대상으로 선정될 수 있는 금지어 유형의 예를 보여주는 도면이다.
- [0077] 도 5를 참조하면, 본 발명의 일 실시예에 따른 팩트 체크 수행부(330)는 출처표시 확인 모듈(331), 출처 신뢰성 검증 모듈(332), 금지어 확인 모듈(333), 예외적 허용 확인 모듈(334) 및 팩트 체크 평가 모듈(335)을 포함하여 구성될 수 있다.
- [0078] 출처표시 확인 모듈(331)은 상기 팩트 체크 대상의 유형이 인용문, 조사 결과, 또는 숫자나 통계치 중 적어도 하나에 해당하는 경우에, 상기 의심 뉴스 기사 내에 상기 체크 대상의 출처가 표시되어 있는지 여부를 확인할 수 있다.
- [0079] 출처 신뢰성 검증 모듈(332)은 팩트 체크 대상의 유형별 가중치에 따라서 팩트체크 가이드라인 모델에서 지정

한 방법을 적용하여, 출처의 신뢰성 정보를 검증할 수 있다.

- [0080] 예를 들면, 도 6에 도시된 것처럼 팩트 체크 대상의 유형(102)이 수치나 통계인 경우에 체크 방법(103)으로 자료 출처 검증이 제시될 수 있고, 팩트 체크 대상의 유형(102)연구결과나 조사결과인 경우 연구자 검증이 체크 방법(103)으로 지정될 수 있다. 만일 법률적 사안이나 소송 중인 사안이 대상이라면 관련 재판 번호나 법률 조항이 체크 방법(103)으로 지정될 수도 있다. 체크 방법(103)은 모델(100)이 빅데이터 분석을 통해서 도출한 자료를 토대로 제시되는 것일 수 있다. 또한, 기사 내에 출처나 인용자가 표기되었다고 해도 허위 표기 등을 고려해, 가중치(104)가 높은 팩트 체크 대상은 인용자나 출처의 신뢰성 검증을 위해서 추가적으로 인터넷 검색을 통한 검증이 요구될 수 있다.
- [0081] 한편, 출처가 URL인 경우에는 URL 연결이 살아 있는 경우와 끊겨진 경우 즉 늘 변함없이 서비스가 되는 경우를 좀 더 신뢰할 수 있고 진짜 뉴스에 가깝다고 가정하고 가중치를 부여할 수 있다.
- [0082] 금지어 확인 모듈(333)은 팩트 체크 대상의 유형이 금지어인 경우에, 금지어 데이터베이스를 검색하여 금지어인지 확인할 수 있다.
- [0083] 예외적 허용 확인 모듈(334) 팩트 체크 대상이 금지어로 확인되면, 소정의 금지어의 예외적 허용 규칙이 적용될 수 있는지 여부를 확인할 수 있다.
- [0084] 예를 들면, 도 7에 도시된 것처럼 금지어 유형(105)은 사생활 침해적 표현, 비속어, 광고성 문구 및 편향된 표현, 인권 침해적 표현 등을 포함할 수 있다. 이러한 금지어가 발견되면 가짜 뉴스 판단에 부정적 평가 요소가 될 수 있다. 다만, 모델(100)에서 금지어 유형(105)에 따라 예외적 허용 사유(106)를 지정하였다면, 예외적 허용 사유에 해당하는지 판단할 수 있다. 예를 들면, 뉴스 기사의 객관성을 해치는 주관적 표현인 "나는"이나 "우리는"이라는 1인칭 주어가 사용되었지만, 인용부호와 함께 사용되어 인터뷰나 타인의 주장을 인용한 것이라면 예외적으로 허용될 수 있다. 통상적인 뉴스가 가지는 육하원칙에 어긋나는 어순이나 글쓰기 방식 역시 정통적인 방식으로 작성된 뉴스로 보기 어렵고 동시에 가짜뉴스일 가능성이 높은 것으로 볼 수 있다.
- [0085] 또한, 기사 내에 사생활 침해적 표현이 포함되어 있지만, 그 대상이 공인이나 연예인이거나 합법적으로 신원이 공개된 범죄자라면 예외적 허용 사유에 들어갈 수 있다.
- [0086] 팩트 체크 평가 모듈(335)은 팩트 체크 결과에 기초하여 의심 뉴스 기사의 가짜 뉴스 여부를 판단할 수 있다.
- [0088] 이하, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법에 대해서 설명한다.
- [0089] 도 8은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법이 수행되는 과정을 보여주는 순서도이다.
- [0090] 도 8을 참조하면, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법은, 의심 뉴스 선정 단계(S100), 자연어 처리단계(S200), 팩트 체크 대상 획득단계(S300), 팩트 체크 수행단계(S400), 제작자 신뢰성 평가 단계(S500) 및 가짜 뉴스 판단 단계(S5100)를 포함하여 구성될 수 있다.
- [0091] 이하, 도 8 참조하여 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 방법의 각 단계를 설명한다.
- [0092] 의심 뉴스 선정 단계(S100)는 뉴스 기사의 최초 공유자의 공유 의도를 분석하여, 상기 분석 결과를 기초로 의심 뉴스 기사를 선정할 수 있다. 예를 들면, SNS 사이트에 공유되는 뉴스 기사를 모니터링하면서 상기 뉴스 기사의 공유 신뢰도 분석 결과를 기초로 의심 뉴스 기사를 선정할 수 있다. 즉 최초 공유자의 공유 의도를 분석하고, 공유된 기사에 대한 댓글을 분석하여 공유 신뢰도를 평가할 수 있다. 의심 뉴스 기사를 선정하는 방법은 추후 좀 더 상세하게 설명하기로 한다.
- [0093] 자연어 처리단계(S200)는 팩트 체크 가이드라인 모델에 적용할 수 있도록 의심 뉴스를 자연어 처리할 수 있다. 자연어 처리 방법은 알려진 기술이므로 상세한 설명을 생략한다.
- [0094] 팩트 체크 대상 획득단계(S300)는 의심 뉴스 기사의 자연어 처리 결과를 토대로 팩트 체크 가이드라인 모델을 적용하여 의심 뉴스 기사 내의 팩트 체크 대상을 획득할 수 있다.
- [0095] 여기서 팩트 체크 가이드라인 모델은 가짜 뉴스 및 진짜 뉴스 데이터를 대비하여 빅데이터 분석하여 가짜 뉴스의 패턴을 도출하고, 패턴 분석한 결과를 기초로 뉴스 기사 내의 팩트 체크 대상을 판별하기 위한 가이드라인을

제시하도록 모델링한 모델일 수 있다. 따라서, 자연어 처리된 의심 뉴스 기사에 팩트 체크 가이드라인 모델을 적용하여 의심스럽거나 추가적인 검증이 필요한 문장이나 어휘가 팩트 체크 대상으로 추출될 수 있다.

- [0096] 팩트 체크 수행단계(S400)는 팩트 체크 대상 획득단계(S300)에서 획득한 팩트 체크 대상을 유형에 따라서 분류하여 가중치를 부여하고 미리 정해진 방법에 따른 팩트 체크를 수행할 수 있다.
- [0097] 제작자 신뢰성 평가 단계(S500)는 뉴스 제작자 및 소속 기자에 대한 신뢰성 정보가 저장되는 뉴스 제작자 프로필 데이터베이스로부터 획득한 신뢰성 정보를 기초로 의심 뉴스 기사의 제작자 신뢰성 점수를 부여할 수 있다.
- [0098] 가짜 뉴스 판단 단계(S5100)는 팩트 체크 결과와 제작자 신뢰성 평가 결과에 기초하여 의심 뉴스 기사의 가짜 뉴스 여부를 판단할 수 있다.
- [0100] 이하, 도면을 참조하여 의심 뉴스 선정 단계(S100)에 대하여 좀 더 상세하게 설명한다.
- [0101] 도 9는 본 발명의 일 실시예에 따른 의심 뉴스를 선정하는 과정을 보여주는 순서도이다.
- [0102] 도 9를 참조하면, 본 발명의 일 실시예에 따른 의심 뉴스 선정 단계(S100)는 최초 공유자 추적단계(S110), 공유 신뢰도 평가단계(S120), 시의성 판단단계(S130) 및 의심 지수 산출단계(S140)를 포함하여 구성될 수 있다.
- [0103] 최초 공유자 추적단계(S110)는 SNS 사이트에 공유되는 뉴스 기사를 모니터링 하면서, 공유되는 뉴스 기사에 대한 사용자 반응이 소정의 임계치를 초과하는 경우에, 뉴스 기사의 최초 공유자를 추적할 수 있다.
- [0104] 예를 들면, 페이스북, 트위터, 카카오톡과 같은 SNS 사이트에 공유한 뉴스 기사에 대한 공감표현, 댓글, 재공유, 트윗/리트윗 등을 분석하여 소정의 임계치를 초과하면 해당 뉴스 기사를 최초로 공유한 사용자를 추적할 수 있다.
- [0105] 공유 신뢰도 평가단계(S120)는 최초 공유자가 공유하면서 게시한 공유 의견과 공유된 뉴스 기사의 댓글을 분석하여 최초 공유자의 공유 의도를 파악하여 뉴스 공유의 신뢰도를 평가할 수 있다.
- [0106] 이때, 최초 공유자가 게시한 공유 의견을 분석하여, 공유 의견 내에 가짜 뉴스를 암시하는 표현, 부추기는 표현, 도발하는 표현 또는 호도하는 표현이 포함되는 빈도에 따라서 제1 공유 신뢰도 평가값을 산출할 수 있다.
- [0107] 또한, 상기 뉴스 기사의 댓글 중 최초 공유자를 향한 의견이 포함되는 댓글을 분석하여 부정 표현 또는 긍정 표현이 포함되는 빈도에 따라서 제2 공유 신뢰도 평가값을 산출하여, 제1 공유 신뢰도 평가값과 제2 공유 신뢰도 평가값을 기초로 뉴스 공유의 신뢰도를 평가할 수 있다.
- [0108] 예를 들면, SNS 에서 뉴스를 공유하는 일반적인 사용자는 뉴스 기사의 논조나 내용이 자신의 정치적, 경제적, 문화적, 직업적 이해관계나 소신과 일치하는 경우에 공유하여, 여러 사람에게 퍼트리고 공감을 받고자 하는 의도가 있을 수 있다. 따라서 공유 의견에는 기사의 내용을 일부 발췌하거나, 기사의 내용에 대한 자신의 주관적인 의견이 포함될 수 있다. 이러한 기사에 대한 댓글 역시 기사의 내용에 대한 의견 표현이 포함될 수 있다. 역시 이러한 공유는 뉴스 공유의 신뢰도를 높게 평가할 수 있다.
- [0109] 반면, 자신의 성향과 맞지 않거나 가짜 뉴스로 의심되는 뉴스 기사를 굳이 공유해서 퍼트리려고 하는 사람은 많지 않을 수 있지만, 만일 사용자가 뉴스 기사의 내용이 사실과 다른 가짜 뉴스임을 알면서도 경고의 의미로 공유하는 경우라면, 공유 의견에 가짜 뉴스임을 암시하는 문구나 조심하라는 경고하는 문구가 포함될 수 있고, 이에 대한 댓글의 반응을 통해서 사용자 의도를 좀 더 확실하게 파악할 수 있다. 이 경우는 사용자의 의견에 가짜 뉴스임이 나타나 있을 수 있기 때문에 뉴스 공유의 신뢰도를 상당히 낮게 평가할 수 있다.
- [0110] 위의 경우와 다르게, 오직 사실 관계를 호도하거나 특정 여론을 조성하기 위해서 가짜 뉴스를 공유하는 것이라면, 공유 의견 내에 가짜 뉴스를 암시하는 문구는 없을 수 있다. 그러나 원색적이거나 도발적인 문구를 포함할 수 있고, 해당 뉴스 기사를 많이 퍼뜨리도록 부추기는 표현을 포함할 수 있다. 이에 대한 댓글은 최초 공유자나 뉴스 작성자에 대한 비난이나 격려 등의 글이 포함될 수 있다. 이 경우는 사용자의 뉴스 공유의 신뢰도가 낮게 평가될 수 있다.
- [0111] 시의성 판단단계(S130)는 공유된 뉴스 기사의 작성일을 기초로 뉴스 기사의 시의성을 판단할 수 있다. 즉, 이미 오래 전(몇년전, 또는 몇달전)에 작성되거나 게시된 뉴스로서의 가치가 없는 예전 기사를 공유하는 경우라면, 여론을 호도하기 위한 가짜 뉴스로 의심할 수 있다. 따라서, 상기 뉴스 기사의 작성일이나 뉴스 기사 게시일과 뉴스 기사 공유일 간에 시차를 분석하고, 시차를 기초로 뉴스 기사의 시의성을 판단하되, 시차가 클수록 뉴스

기사의 시의성이 낮게 판단될 수 있다.

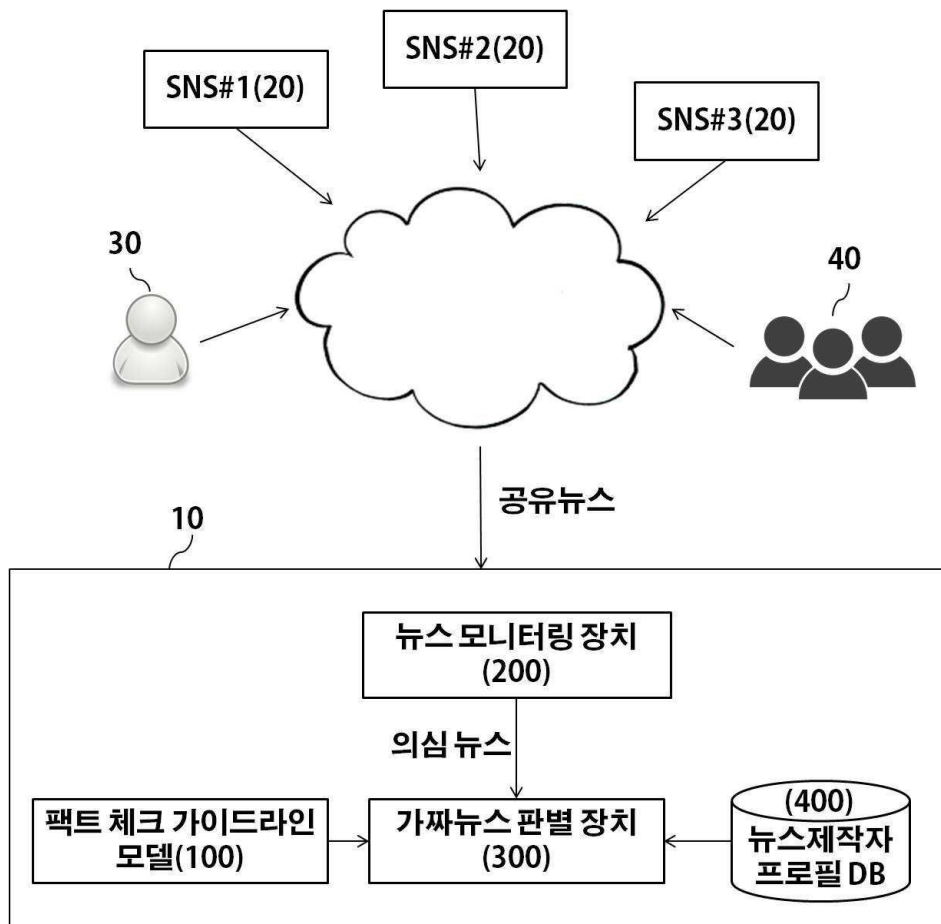
- [0112] 의심 지수 산출단계(S140)는 공유 신뢰도 평가 결과와 상기 시의성 판단 결과를 기초로 상기 뉴스 기사의 의심 지수를 산출할 수 있다. 공유 신뢰도와 시의성이 낮을수록 의심 지수가 높게 산출될 수 있으며, 산출된 의심 지수가 소정의 임계치를 초과하는 경우에는 의심 뉴스로 선정될 수 있다.
- [0114] 이하, 이하, 도면을 참조하여 팩트 체크 수행단계(S400)에 대하여 좀 더 상세하게 설명한다.
- [0115] 도 10은 본 발명의 일 실시예에 따른 팩트 체크 가이드라인에 따라 팩트 체크하는 과정을 보여주는 순서도이다.
- [0116] 도 10를 참조하면, 본 발명의 일 실시예에 따른 팩트 체크 수행단계(S400)는 출처표시 확인 단계(S410), 출처 신뢰성 검증 단계(S420), 금지어 확인 단계(S430), 예외적 허용 확인 단계(S440) 및 팩트 체크 평가 단계(S450)를 포함하여 구성될 수 있다.
- [0117] 출처표시 확인 단계(S410)는 의심 뉴스 내의 팩트 체크 대상의 유형이 인용문, 조사 결과, 또는 숫자나 통계치 중 적어도 하나에 해당하는 경우에, 상기 의심 뉴스 기사 내에 상기 체크 대상의 출처가 표시되어있는지 여부를 확인할 수 있다.
- [0118] 출처 신뢰성 검증 단계(S420)는 팩트 체크 대상의 유형별 가중치에 따라서 팩트체크 가이드라인 모델에서 지정한 방법을 적용하여, 출처의 신뢰성 정보를 검증할 수 있다.
- [0119] 예를 들면, 팩트 체크 대상의 유형이 수치나 통계인 경우에 체크 방법으로 자료 출처 검증으로 제시될 수 있고 연구결과나 조사결과인 경우 연구자 검증이 체크 방법으로 지정될 수 있다. 만일 법률적 사안이나 소송중인 사안이 대상이라면 관련 재판 번호나 법률 조항이 체크 방법으로 지정될 수도 있다. 또한, 기사 내에 출처나 인용자가 표기되었다고 해도 허위 표기 등을 고려해, 가중치가 높은 팩트 대상은 인용자나 출처의 신뢰성 검증을 위해서 추가적으로 인터넷 검색을 통한 검증이 필요할 수 있다.
- [0120] 금지어 확인 단계(S430)는 팩트 체크 대상의 유형이 금지어인 경우에, 금지어 데이터베이스를 검색하여 팩트 체크 대상이 금지어에 해당하는지 확인할 수 있다.
- [0121] 예외적 허용 확인 단계(S440)는 팩트 체크 대상이 금지어로 확인되면, 소정의 금지어의 예외적 허용 사유에 해당하는지 여부를 확인할 수 있다.
- [0122] 예를 들면, 금지어 유형은 사생활 침해적 표현, 비속어, 광고성 문구 및 편향된 표현, 인권 침해적 표현 등을 포함할 수 있다. 이러한 금지어가 발견되면 가짜 뉴스 판단에 부정적 평가 요소가 될 수 있다. 다만, 팩트 체크 가이드라인 모델에서 금지어 유형에 따라 예외적 허용 사유를 지정하였다면, 예외적 허용사유에 해당되는지 여부를 판단할 수 있다.
- [0123] 예를 들면, 뉴스 기사의 객관성을 해치는 주관적 표현인 "나는"이나 "우리는"이라는 1인칭 주어가 사용되었지만, 인용부호와 함께 사용되어 인터뷰나 타인의 주장을 인용한 것이라면 예외적으로 허용될 수 있다. 또한, 기사 내에 사생활 침해적 표현이 포함되어 있지만, 그 대상이 공인이나 연예인이거나 합법적으로 신원이 공개된 범죄자라면 예외적 허용 사유에 들어갈 수 있다.
- [0124] 팩트 체크 평가 단계(S450)는 팩트 체크 결과에 기초하여 의심 뉴스 기사의 가짜 뉴스 여부를 판단할 수 있다.
- [0125] 가짜 뉴스로 판단된 뉴스 기사는 가짜 뉴스 데이터베이스에 저장되거나 뉴스 모니터링 장치로 회신하여 사용자에게 가짜 뉴스임을 알리도록 할수 있다. 가짜뉴스 판단 결과를 처리하는 방법은 다양하게 활용될 수 있으며 본 실시예에 한정되지 않는다.
- [0127] 이상에서 설명한 바와 같이, 본 발명의 일 실시예에 따른 팩트 체크 가이드라인을 이용한 가짜 뉴스 판별 시스템 및 방법은, SNS를 모니터링하면서 사용자 반응이 소정의 임계치를 초과하는 뉴스 기사를 선택해서, 뉴스 기사의 최초 공유자의 공유 의도와 사용자 반응을 분석하여 의심 뉴스 기사를 선정하여, 분석 대상을 1차적으로 필터링하고, 빅데이터 분석을 통해 생성된 팩트 체크 가이드라인 모델을 적용하여 팩트 체크 대상 문장을 선정하는 2차 필터링 과정을 거치고, 다음으로 팩트 체크 대상에 대하여 팩트 체크 가이드라인 모델에서 제시하는 체크 방법으로 검증을 함으로써, 가짜 뉴스를 신속하고 정확하게 판별할 수 있는 장점이 있다.

[0129] 한편, 지금까지 설명한 본 발명의 방법 및 장치는 실제로 컴퓨터 프로그램에 의해 구현될 수 있고, 컴퓨터에서 실행될 때 컴퓨터 판독 가능한 기록 매체에 저장될 수 있다. 컴퓨터 판독 가능한 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있도록 프로그램 및 데이터가 저장되는 모든 종류의 기록매체를 포함하며, ROM, RAM, CD, DVD-ROM, 자기테이프, 플로피 디스크, 광데이터 저장장치 등이 있으며, 또한 인터넷을 통한 전송되는 형태로 구현되는 것도 포함될 수 있다. 즉, 이와 같은 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

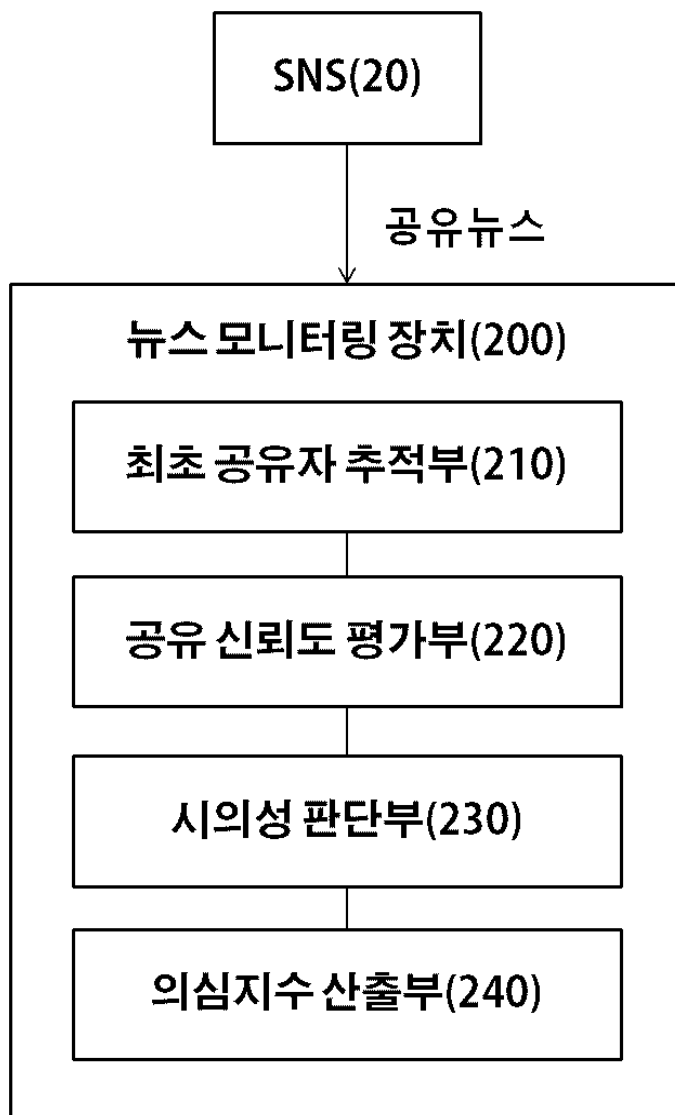
[0131] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

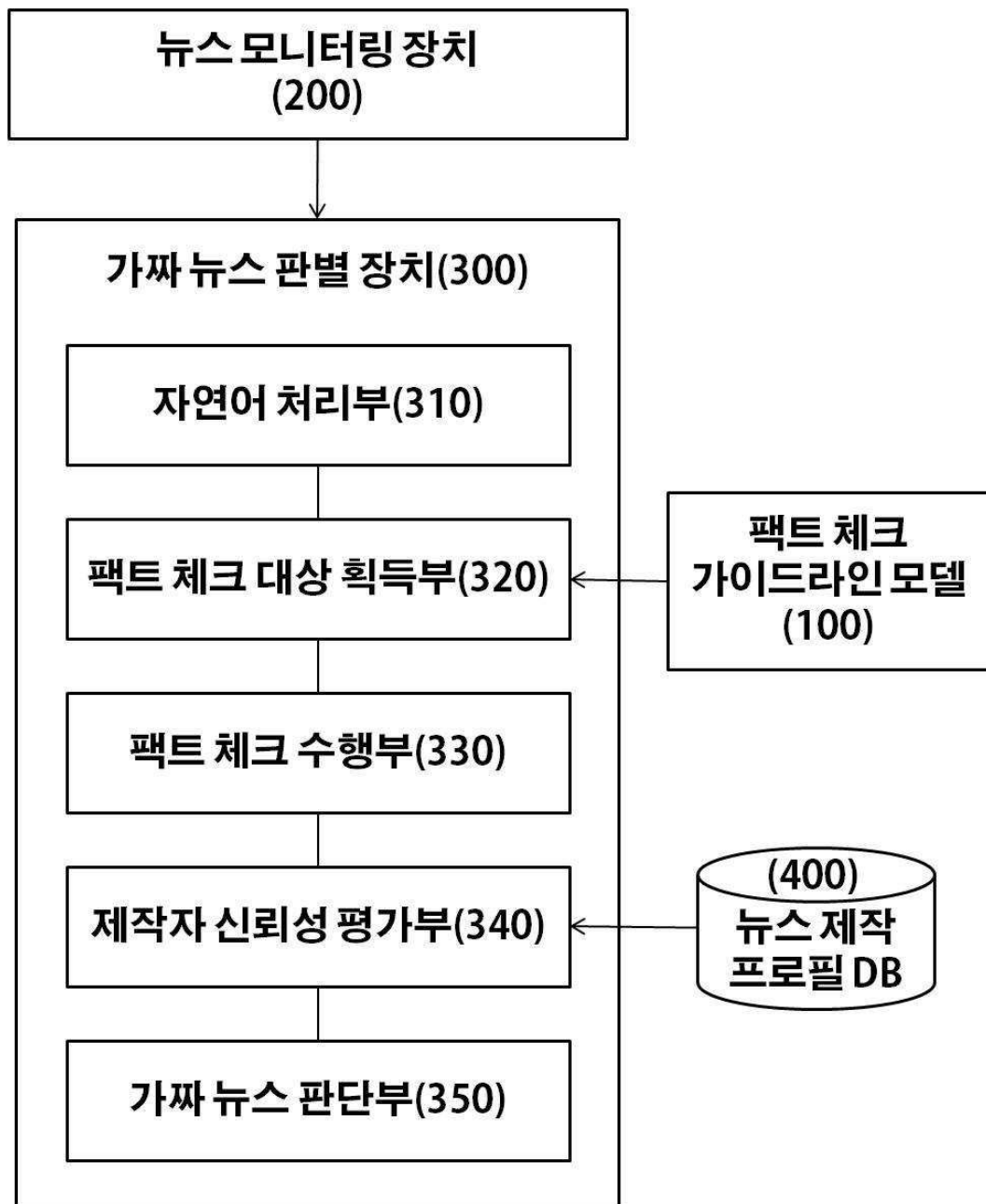
도면1



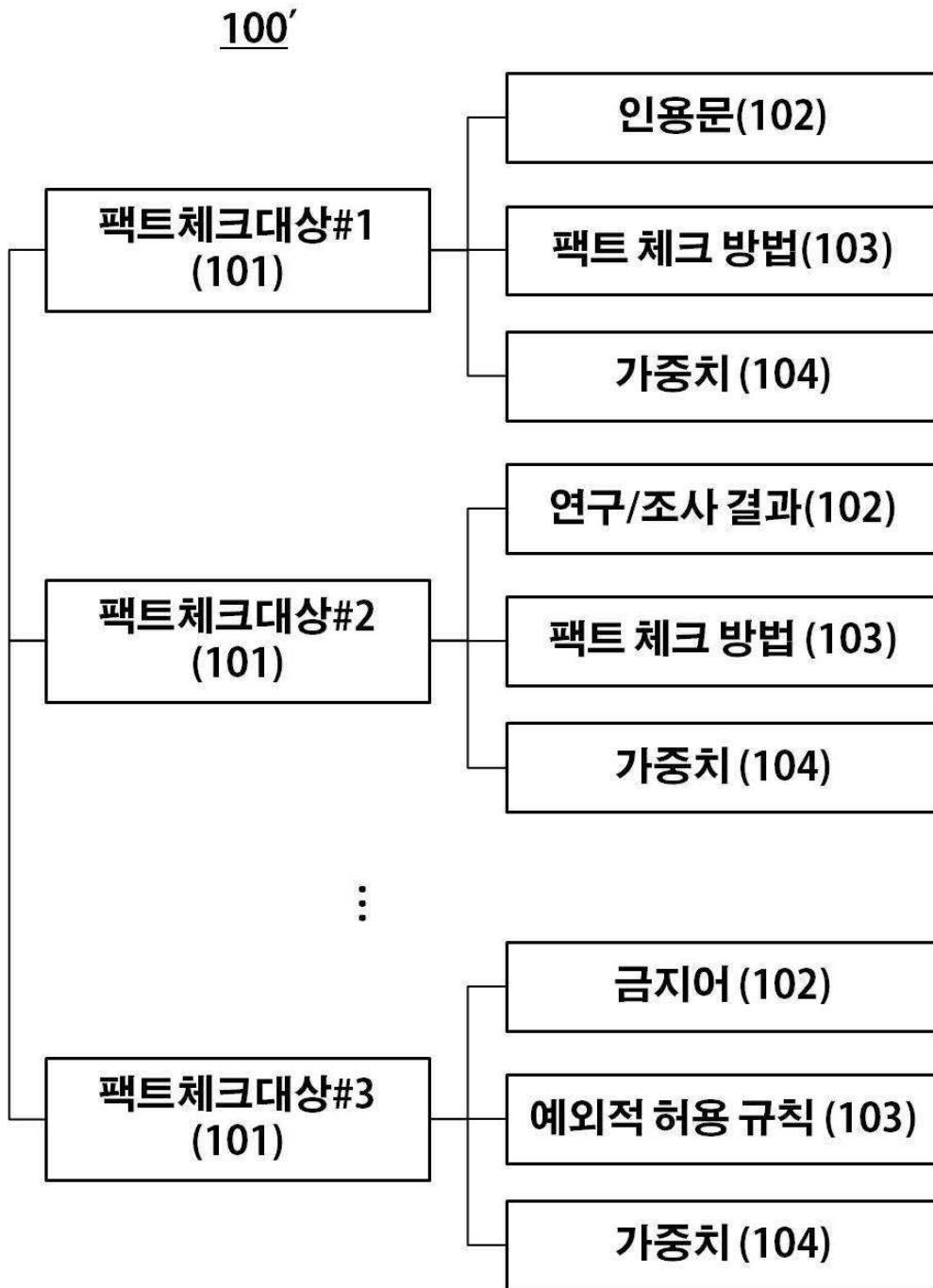
도면2



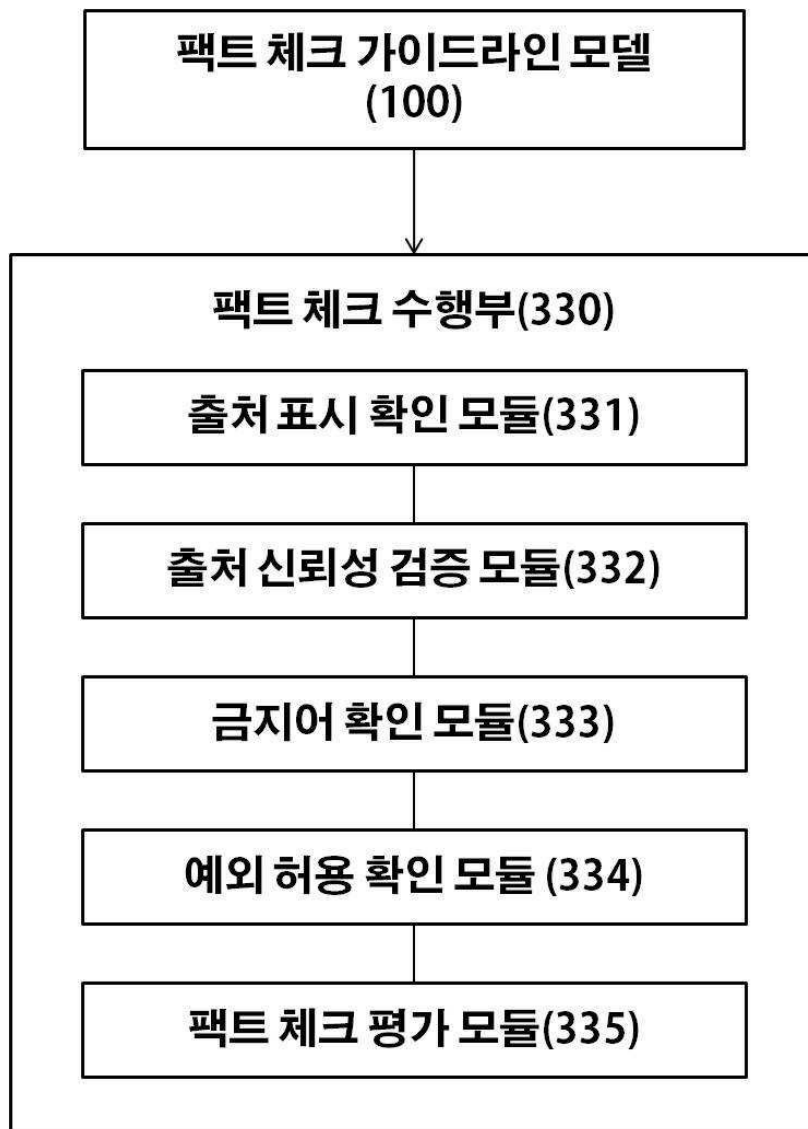
도면3



도면4



도면5



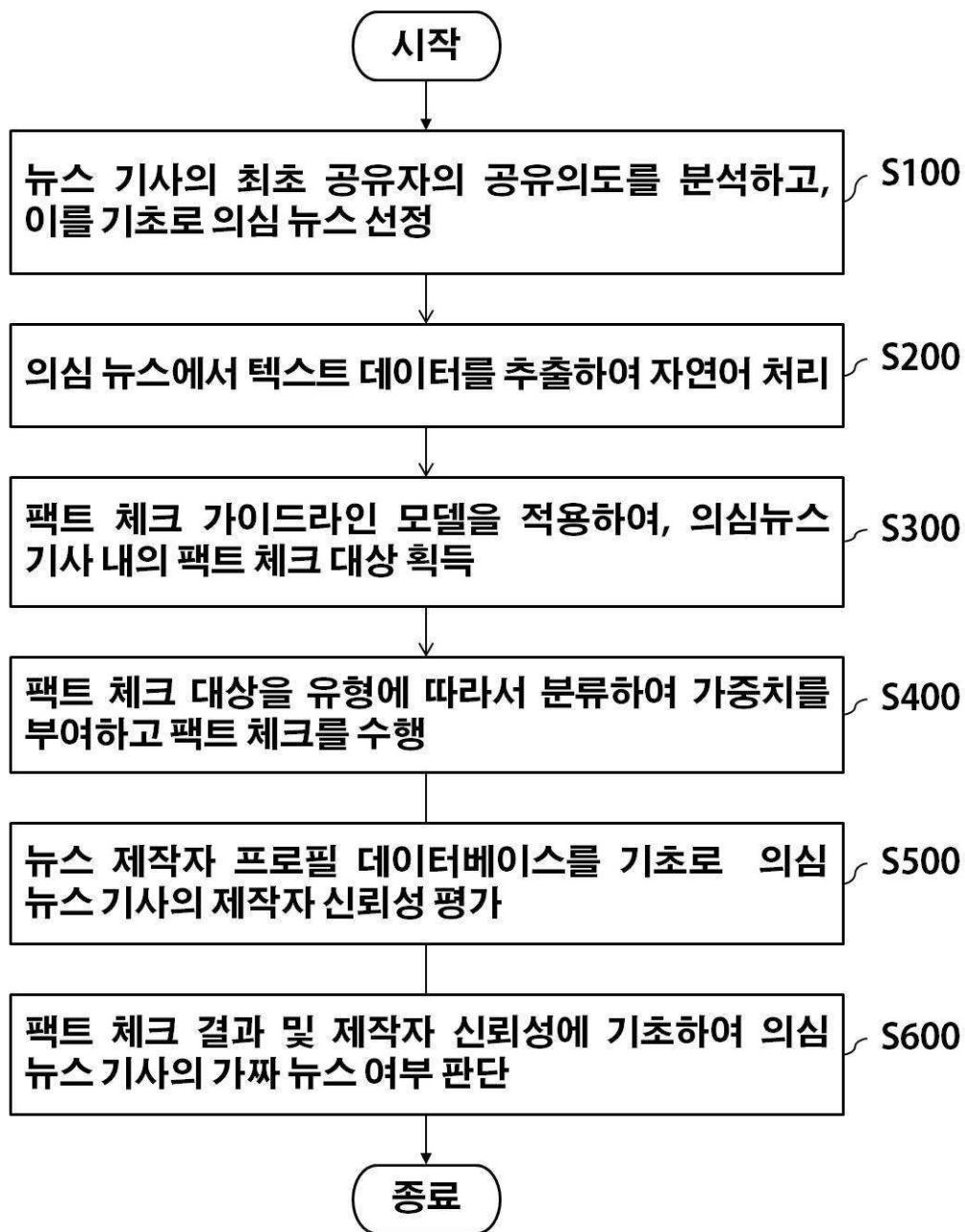
도면6

팩트 유형 (102)	체크 방법 (103)	가중치 (104)
수치표현/통계	자료 출처	1
연구/조사결과	연구자	2
일방의 의견	반대의견/관련이슈	3
법률적 사안	재판문서	2
...	...	
...	...	
...	...	
...	...	

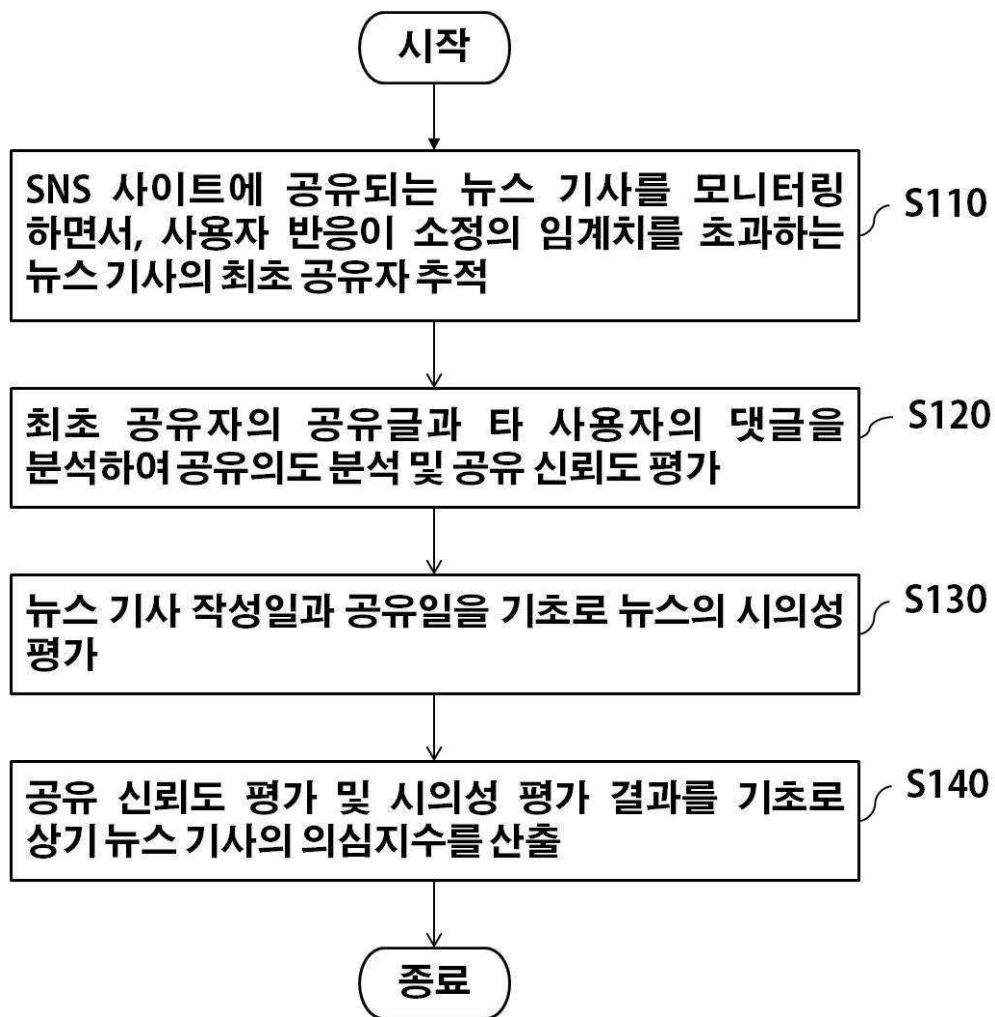
도면7

금지어 유형(105)	예외적 허용(106)
주관적 표현	인용부호사용
사생활 침해표현	공인/연예인/범죄자
비속어	...
광고성 문구	...
편향된 표현	...
인권침해	...
...	...
...	...

도면8



도면9



도면10

