



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0103427
(43) 공개일자 2022년07월22일

(51) 국제특허분류(Int. Cl.)

G06F 40/30 (2020.01) G06F 16/33 (2019.01)
G06F 16/951 (2019.01) G06F 40/258 (2020.01)
G06N 20/00 (2019.01) G10L 15/26 (2006.01)

(52) CPC특허분류

G06F 40/30 (2020.01)
G06F 16/3335 (2019.01)

(21) 출원번호 10-2021-0006025

(22) 출원일자 2021년01월15일

심사청구일자 2021년01월15일

(71) 출원인

세종대학교산학협력단

서울특별시 광진구 능동로 209 (군자동, 세종대학교)

(72) 발명자

박중현

서울특별시 광진구 능동로 209 세종대학교

오유원

서울특별시 광진구 능동로 209 세종대학교

(74) 대리인

김준석

전체 청구항 수 : 총 18 항

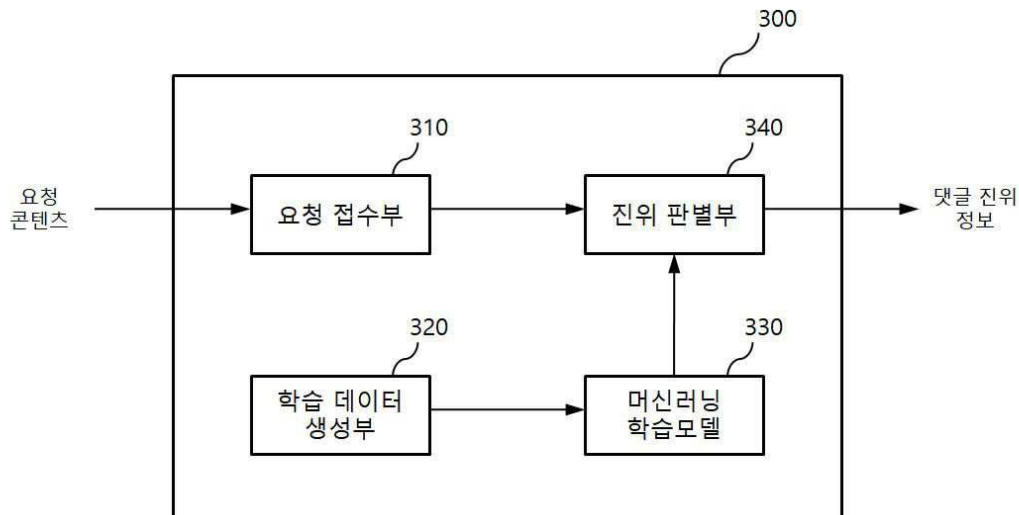
(54) 발명의 명칭 머신 러닝을 이용한 댓글 판별 서버 및 그를 이용한 가짜 댓글 판별 방법

(57) 요약

본 발명의 일 기술적 측면에 따른 머신 러닝을 이용한 가짜 댓글 판별 시스템은, 온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성하는 학습 데이터 생성부, 상기 학습 데이터 생성부에서 제공된 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인

(뒷면에 계속)

대표도 - 도3



콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습하는 머신러닝 학습모델 및 사용자로부터 요청된 요청 콘텐츠에 대하여 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하고, 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 상기 요청 콘텐츠의 댓글의 진위를 판별하는 진위 판별부를 포함 할 수 있다. 본 발명의 일 실시형태에 따르면, 온라인 콘텐츠에 대한 머신러닝 학습을 기반으로 사용자가 요청한 온라인 콘텐츠의 댓글에 대한 진위 확률을 산출하여, 온라인 콘텐츠의 댓글에 대한 진위 여부를 판별할 수 있는 효과가 있다.

(52) CPC특허분류

G06F 16/951 (2019.01)

G06F 40/258 (2020.01)

G06N 20/00 (2021.08)

G10L 15/26 (2013.01)

명세서

청구범위

청구항 1

온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성하는 학습 데이터 생성부;

상기 학습 데이터 생성부에서 제공된 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인 콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습하는 머신러닝 학습모델; 및

사용자로부터 요청된 요청 콘텐츠에 대하여 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하고, 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 상기 요청 콘텐츠의 댓글의 진위를 판별하는 진위 판별부;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 2

제1항에 있어서, 상기 학습 데이터는,

온라인 콘텐츠의 종류를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 제목, 퍼블리셔 정보 및 오너 정보를 포함하는 종류 학습데이터;

온라인 콘텐츠의 주제를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 내용, 퍼블리셔 정보 및 오너 정보를 포함하는 주제 학습데이터; 및

온라인 콘텐츠의 댓글을 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 종류, 주제, 가짜댓글 정보를 포함하는 댓글 학습데이터;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 3

제2항에 있어서, 상기 학습 데이터 생성부는,

온라인 콘텐츠를 수집하는 온라인 콘텐츠 수집모듈;

상기 온라인 콘텐츠 수집모듈에서 수집된 온라인 콘텐츠 중에서, 상기 종류 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집하는 종류 학습데이터 수집모듈;

상기 수집된 온라인 콘텐츠 중에서, 상기 주제 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집하는 주제 학습데이터 수집모듈; 및

상기 수집된 온라인 콘텐츠 중에서, 상기 댓글 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집하는 댓글 학습데이터 수집모듈;

을 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 4

제2항에 있어서, 상기 머신러닝 학습모델은,

온라인 콘텐츠의 텍스트에 대한 불용어를 제거하고 표제어를 추출하고, 상기 학습데이터 별로 다르게 설정된 온라인 콘텐츠의 각 기준 항목을 대상으로 N-gram 언어모델을 적용하고 벡터화를 적용하여 분류 학습을 수행하는

것

을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 5

제2항에 있어서, 상기 머신러닝 학습모델은,

종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 학습하여 온라인 콘텐츠의 종류 확률을 산출하는 종류 학습 모델;

주제 학습데이터를 기초로 온라인 콘텐츠의 내용을 학습하여 온라인 콘텐츠의 주제 확률을 산출하는 주제 학습 모델; 및

댓글 학습데이터를 기초로 온라인 콘텐츠의 종류 및 주제 별 가짜 댓글의 확률을 산출하는 댓글 학습모델;

을 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 6

제5항에 있어서, 상기 머신러닝 학습모델은,

음성 기반 온라인 콘텐츠를 식별하고, 음성 인식을 이용하여 음성 기반 온라인 콘텐츠를 텍스트 콘텐츠로 변환하며, 상기 텍스트 콘텐츠에 대하여 불용어를 제거하고 표제어를 추출하는 전처리를 수행하는 전처리 모듈;

을 더 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 7

제5항에 있어서, 상기 진위 판별부는,

상기 종류 학습모델을 이용하여 상기 요청 콘텐츠의 종류를 판별하는 종류 판별모듈;

상기 주제 학습모델을 이용하여 상기 요청 콘텐츠의 주제를 판별하는 주제 판별모듈; 및

상기 댓글 학습모델을 이용하여, 판별된 상기 요청 콘텐츠의 종류 및 주제에 따른 댓글 진위 확률을 산출하는 댓글 진위 판별모듈;

을 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 8

제7항에 있어서, 상기 진위 판별부는,

음성 기반 온라인 콘텐츠를 식별하고, 음성 인식을 이용하여 음성 기반 온라인 콘텐츠를 텍스트 콘텐츠로 변환하는 전처리 모듈;

을 더 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 9

제7항에 있어서, 상기 진위 판별부는,

상기 종류 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 종류 보정치를 설정하고, 상기 종류 판별모듈의 판별 결과에 상기 종류 보정치를 반영하는 종류 보정모듈; 및

상기 주제 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 주제 보정치를 설정하고, 상기 주제

판별모듈의 판별 결과에 상기 주제 보정치를 반영하는 주제 보정모듈;
을 더 포함하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 10

제7항에 있어서, 상기 진위 판별부는,

$$\tilde{p} = f(\tilde{x} + 3 \times w_{ij}), \text{ where } \tilde{x} = f^{-1}(p)$$

수학식 및

$$f(x) = \frac{1}{1 + e^{-x}}$$

수학식

을 이용하여 상기 댓글 진위 판별모듈의 판별 결과를 보정하는 댓글 진위 보정모듈;
을 더 포함하고,

여기에서, p 는 상기 댓글 진위 판별모듈의 판별 결과를, $\tilde{p}$ 는 상기 댓글 진위 보정모듈에 의한 보정된 확률을 의미하는 것을 특징으로 하는 머신 러닝을 이용한 댓글 판별 서버.

청구항 11

온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성하는 단계;

상기 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인 콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습하여 머신러닝 학습 모델을 생성하는 단계;

사용자로부터 요청된 요청 콘텐츠에 대하여 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하는 단계; 및

온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 상기 요청 콘텐츠의 댓글의 진위를 판별하는 단계;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 12

제11항에 있어서, 상기 학습 데이터는,

온라인 콘텐츠의 종류를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 제목, 퍼블리셔 정보 및 오너 정보를 포함하는 종류 학습데이터;

온라인 콘텐츠의 주제를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 내용, 퍼블리셔 정보 및 오너 정보를 포함하는 주제 학습데이터; 및

온라인 콘텐츠의 댓글을 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 종류, 주제, 가짜댓글 정보를 포함하는 댓글 학습데이터;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 13

제12항에 있어서, 상기 학습 데이터를 생성하는 단계는,

온라인 콘텐츠를 수집하는 단계;

상기 수집된 온라인 콘텐츠 중에서, 상기 종류 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 종류 학습 데이터를 수집하는 단계;

상기 수집된 온라인 콘텐츠 중에서, 상기 주제 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 주제 학습 데이터를 수집하는 단계; 및

상기 수집된 온라인 콘텐츠 중에서, 상기 댓글 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 댓글 학습 데이터를 수집하는 단계;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 14

제12항에 있어서, 상기 머신러닝 학습모델은,

온라인 콘텐츠의 텍스트에 대한 불용어를 제거하고 표제어를 추출하고, 상기 학습데이터 별로 다르게 설정된 온라인 콘텐츠의 각 기준 항목을 대상으로 N-gram 언어모델을 적용하고 벡터화를 적용하여 분류 학습을 수행하는 것

을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 15

제12항에 있어서, 상기 머신러닝 학습 모델을 생성하는 단계는,

종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 학습하여 온라인 콘텐츠의 종류 확률을 산출하는 종류 학습 모델을 생성하는 단계;

주제 학습데이터를 기초로 온라인 콘텐츠의 내용을 학습하여 온라인 콘텐츠의 주제 확률을 산출하는 주제 학습 모델을 생성하는 단계; 및

댓글 학습데이터를 기초로 온라인 콘텐츠의 종류 및 주제 별 가짜 댓글의 확률을 산출하는 댓글 학습모델을 생성하는 단계;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 16

제15항에 있어서, 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하는 단계는,

상기 종류 학습모델을 이용하여 상기 요청 콘텐츠의 종류를 판별하는 단계;

상기 주제 학습모델을 이용하여 상기 요청 콘텐츠의 주제를 판별하는 단계; 및

상기 댓글 학습모델을 이용하여, 판별된 상기 요청 콘텐츠의 종류 및 주제에 따른 댓글 진위 확률을 산출하는 단계;

를 포함하는 것을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 17

제16항에 있어서, 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하는 단계는,

상기 종류 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 종류 보정치를 설정하고, 상기 종류

판별모듈의 판별 결과에 상기 종류 보정치를 반영하는 단계; 및

상기 주제 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 주제 보정치를 설정하고, 상기 주제 판별모듈의 판별 결과에 상기 주제 보정치를 반영하는 단계;

를 더 포함하는 것을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

청구항 18

제16항에 있어서, 상기 요청 콘텐츠의 댓글의 진위를 판별하는 단계는,

$$\tilde{p} = f(\tilde{x} + 3 \times w_{ij}), \text{ where } \tilde{x} = f^{-1}(p)$$

수학식

$$f(x) = \frac{1}{1 + e^{-x}}$$

수학식

을 이용하여 상기 댓글 진위 판별모듈의 판별 결과를 보정하는 단계;

를 더 포함하고,

여기에서, p 는 상기 댓글 진위 판별모듈의 판별 결과, $\tilde{p}$ 는 상기 댓글 진위 보정모듈에 의한 보정된 확률을 의미하는 것

을 특징으로 하는 머신 러닝을 이용한 가짜 댓글 판별 방법.

발명의 설명

기술 분야

[0001] 본 발명은 머신 러닝을 이용한 댓글 판별 서버 및 그를 이용한 가짜 댓글 판별 방법에 관한 것이다.

배경 기술

[0003] IT 기술의 발전에 따라 다양한 정보, 의견, 홍보 등이 온라인에서 수행되고 있다. 이와 같이, 다양한 정보, 의견, 홍보 등은 온라인 콘텐츠로서 온라인에서 게시 및 공유되고 있다.

[0005] 통상적으로, 이러한 온라인 콘텐츠는 상호 소통 가능성을 넓히기 위하여 댓글 기능을 제공한다. 온라인 콘텐츠의 댓글은 해당 온라인 콘텐츠의 또 하나의 서브 콘텐츠가 되며, 사용자들에게 영향을 미치게 된다.

[0007] 이러한 온라인 댓글은 비교적 짧은 내용을 게시하지만 사용자에게는 보다 큰 영향을 미친다. 한편, 이러한 온라인 댓글을 악의적으로 이용하기 위하여 의도적으로 댓글을 생성하는 문제가 있다. 예컨대, 광고나 홍보의 목적으로 특정 의도에 따라 댓글을 생성하는 등의 문제가 발생하고 있다. 이에 온라인 댓글의 진위 판별에 대한 니즈가 발생하고 있다.

[0009] 한편, 종래의 기술들로는, 한국등록특허공보 제10-2094201호와 같이 가짜 뉴스를 식별하는, 즉, 온라인 콘텐츠 자체에 대한 검증을 수행하는 기술을 개시하고는 있으나, 이러한 종래 기술들은, 온라인 콘텐츠에 달린 댓글에 대한 진위를 검증하지 못하여 온라인 댓글에 의한 부정적인 영향을 방지할 수 없는 한계가 있다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 한국등록특허공보 제10-2094201호

발명의 내용

해결하려는 과제

- [0012] 본 발명의 일 기술적 측면은 상기한 종래 기술의 문제점을 해결하기 위한 것으로써, 온라인 콘텐츠에 대한 머신러닝 학습을 기반으로 사용자가 요청한 온라인 콘텐츠의 댓글에 대한 진위 확률을 산출하여, 온라인 콘텐츠의 댓글에 대한 진위 여부를 판별할 수 있는 머신러닝을 이용한 가짜 댓글 판별 방법 및 댓글 판별 서버를 제공하는 것이다.
- [0013] 또한, 본 발명의 일 기술적 측면은, 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 추정하고, 요청된 온라인 콘텐츠의 종류 및 주제를 학습모델에 따라 구분하여 그 댓글의 진위 여부를 판별할 수 있는 머신러닝을 이용한 가짜 댓글 판별 방법 및 댓글 판별 서버를 제공하는 것이다.
- [0014] 또한, 본 발명의 일 기술적 측면은, 온라인 콘텐츠의 퍼블리셔 및 오너 간의 연관성에 따른 가중치를 설정하고, 이를 반영함으로써 온라인 콘텐츠의 분류를 보다 정확하게 수행할 수 있는 머신러닝을 이용한 가짜 댓글 판별 방법 및 댓글 판별 서버를 제공하는 것이다.
- [0016] 본 발명의 상기 목적과 여러 가지 장점은 이 기술분야에 숙련된 사람들에 의해 본 발명의 바람직한 실시예로부터 더욱 명확하게 될 것이다.

과제의 해결 수단

- [0018] 본 발명의 일 기술적 측면은 머신러닝을 이용한 가짜 댓글 판별 시스템을 제안한다. 상기 머신러닝을 이용한 가짜 댓글 판별 시스템은, 온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성하는 학습 데이터 생성부, 상기 학습 데이터 생성부에서 제공된 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인 콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습하는 머신러닝 학습모델 및 사용자로부터 요청된 요청 콘텐츠에 대하여 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하고, 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 상기 요청 콘텐츠의 댓글의 진위를 판별하는 진위 판별부를 포함 할 수 있다.
- [0019] 일 실시예에서, 상기 학습 데이터는, 온라인 콘텐츠의 종류를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 제목, 퍼블리셔 정보 및 오너 정보를 포함하는 종류 학습데이터, 온라인 콘텐츠의 주제를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 내용, 퍼블리셔 정보 및 오너 정보를 포함하는 주제 학습데이터 및 온라인 콘텐츠의 댓글을 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 종류, 주제, 가짜댓글 정보를 포함하는 댓글 학습 데이터를 포함 할 수 있다.
- [0020] 일 실시예에서, 상기 학습 데이터 생성부는, 온라인 콘텐츠를 수집하는 온라인 콘텐츠 수집모듈, 상기 온라인 콘텐츠 수집모듈에서 수집된 온라인 콘텐츠 중에서, 상기 종류 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집하는 종류 학습데이터 수집모듈, 상기 수집된 온라인 콘텐츠 중에서, 상기 주제 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집하는 주제 학습데이터 수집모듈 및 상기 수집된 온라인 콘텐츠 중에서, 상기 댓글 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집하는 댓글 학습데이터 수집모듈을 포함 할 수 있다.
- [0021] 일 실시예에서, 상기 머신러닝 학습모델은, 온라인 콘텐츠의 텍스트에 대한 불용어를 제거하고 표제어를 추출하고, 상기 학습데이터 별로 다르게 설정된 온라인 콘텐츠의 각 기준 항목을 대상으로 N-gram 언어모델을 적용하

고 벡터화를 적용하여 분류 학습을 수행 할 수 있다.

[0022] 일 실시예에서, 상기 머신러닝 학습모델은, 종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 학습하여 온라인 콘텐츠의 종류 확률을 산출하는 종류 학습모델, 주제 학습데이터를 기초로 온라인 콘텐츠의 내용을 학습하여 온라인 콘텐츠의 주제 확률을 산출하는 주제 학습모델 및 댓글 학습데이터를 기초로 온라인 콘텐츠의 종류 및 주제 별 가짜 댓글의 확률을 산출하는 댓글 학습모델을 포함 할 수 있다.

[0023] 일 실시예에서, 상기 머신러닝 학습모델은, 음성 기반 온라인 콘텐츠를 식별하고, 음성 인식을 이용하여 음성 기반 온라인 콘텐츠를 텍스트 콘텐츠로 변환하며, 상기 텍스트 콘텐츠에 대하여 불용어를 제거하고 표제어를 추출하는 전처리를 수행하는 전처리 모듈을 더 포함 할 수 있다.

[0024] 일 실시예에서, 상기 진위 판별부는, 상기 종류 학습모델을 이용하여 상기 요청 콘텐츠의 종류를 판별하는 종류 판별모듈, 상기 주제 학습모델을 이용하여 상기 요청 콘텐츠의 주제를 판별하는 주제 판별모듈 및 상기 댓글 학습모델을 이용하여, 판별된 상기 요청 콘텐츠의 종류 및 주제에 따른 댓글 진위 확률을 산출하는 댓글 진위 판별모듈을 포함 할 수 있다.

[0025] 일 실시예에서, 상기 진위 판별부는, 음성 기반 온라인 콘텐츠를 식별하고, 음성 인식을 이용하여 음성 기반 온라인 콘텐츠를 텍스트 콘텐츠로 변환하는 전처리 모듈을 더 포함 할 수 있다.

[0026] 일 실시예에서, 상기 진위 판별부는, 상기 종류 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 종류 보정치를 설정하고, 상기 종류 판별모듈의 판별 결과에 상기 종류 보정치를 반영하는 종류 보정모듈 및 상기 주제 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 주제 보정치를 설정하고, 상기 주제 판별모듈의 판별 결과에 상기 주제 보정치를 반영하는 주제 보정모듈을 더 포함 할 수 있다.

[0027] 일 실시예에서, 상기 진위 판별부는, 수학적

$$\tilde{p} = f(\tilde{x} + 3 \times w_{ij}), \text{ where } \tilde{x} = f^{-1}(p)$$

및

$$f(x) = \frac{1}{1 + e^{-x}}$$

수학적 을 이용하여 상기 댓글 진위 판별모듈의 판별 결과를 보정하는 댓글 진위 보정 모듈을 더 포함하고, 여기에서, p 는 상기 댓글 진위 판별모듈의 판별 결과를, $\tilde{p}$ 는 상기 댓글 진위 보정모듈에 의한 보정된 확률을 의미 할 수 있다.

[0029] 본 발명의 다른 일 기술적 측면은 머신 러닝을 이용한 가짜 댓글 판별 방법을 제안한다. 상기 머신 러닝을 이용한 가짜 댓글 판별 방법은, 온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성하는 단계, 상기 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인 콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습하여 머신러닝 학습 모델을 생성하는 단계, 사용자로부터 요청된 요청 콘텐츠에 대하여 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하는 단계 및 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 상기 요청 콘텐츠의 댓글의 진위를 판별하는 단계를 포함 할 수 있다.

[0030] 일 실시예에서, 상기 학습 데이터는, 온라인 콘텐츠의 종류를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 제목, 퍼블리셔 정보 및 오너 정보를 포함하는 종류 학습데이터, 온라인 콘텐츠의 주제를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 내용, 퍼블리셔 정보 및 오너 정보를 포함하는 주제 학습데이터 및 온라인 콘텐츠의 댓글을 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 종류, 주제, 가짜댓글 정보를 포함하는 댓글 학습데이터를 포함 할 수 있다.

[0031] 일 실시예에서, 상기 학습 데이터를 생성하는 단계는, 온라인 콘텐츠를 수집하는 단계, 상기 수집된 온라인 콘텐츠 중에서, 상기 종류 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 종류 학습데이터를 수집하는 단계, 상기 수집된 온라인 콘텐츠 중에서, 상기 주제 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 주제 학습데이터를 수집하는 단계 및 상기 수집된 온라인 콘텐츠 중에서, 상기 댓글 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 댓글 학습데이터를 수집하는 단계를 포함 할 수 있다.

[0032] 일 실시예에서, 상기 머신러닝 학습모델은, 온라인 콘텐츠의 텍스트에 대한 불용어를 제거하고 표제어를 추출하

고, 상기 학습데이터 별로 다르게 설정된 온라인 콘텐츠의 각 기준 항목을 대상으로 N-gram 언어모델을 적용하고 벡터화를 적용하여 분류 학습을 수행 할 수 있다.

[0033] 일 실시예에서, 상기 머신러닝 학습 모델을 생성하는 단계는, 종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 학습하여 온라인 콘텐츠의 종류 확률을 산출하는 종류 학습모델을 생성하는 단계, 주제 학습데이터를 기초로 온라인 콘텐츠의 내용을 학습하여 온라인 콘텐츠의 주제 확률을 산출하는 주제 학습모델을 생성하는 단계 및 댓글 학습데이터를 기초로 온라인 콘텐츠의 종류 및 주제 별 가짜 댓글의 확률을 산출하는 댓글 학습모델을 생성하는 단계를 포함 할 수 있다.

[0034] 일 실시예에서, 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하는 단계는, 상기 종류 학습모델을 이용하여 상기 요청 콘텐츠의 종류를 판별하는 단계, 상기 주제 학습모델을 이용하여 상기 요청 콘텐츠의 주제를 판별하는 단계 및 상기 댓글 학습모델을 이용하여, 판별된 상기 요청 콘텐츠의 종류 및 주제에 따른 댓글 진위 확률을 산출하는 단계를 포함 할 수 있다.

[0035] 일 실시예에서, 상기 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하는 단계는, 상기 종류 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 종류 보정치를 설정하고, 상기 종류 판별모델의 판별 결과에 상기 종류 보정치를 반영하는 단계 및 상기 주제 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 주제 보정치를 설정하고, 상기 주제 판별모델의 판별 결과에 상기 주제 보정치를 반영하는 단계를 더 포함 할 수 있다.

[0036] 일 실시예에서, 상기 요청 콘텐츠의 댓글의 진위를 판별하는 단계는, 수학적

$$\tilde{p} = f(\tilde{x} + 3 \times w_{ij}), \text{ where } \tilde{x} = f^{-1}(p) \quad \text{및} \quad f(x) = \frac{1}{1 + e^{-x}} \quad \text{을 이용하여}$$

상기 댓글 진위 판별모델의 판별 결과를 보정하는 단계를 더 포함하고, 여기에서, p 는 상기 댓글 진위 판별모델의 판별 결과를, $\tilde{p}$ 는 상기 댓글 진위 보정모델에 의한 보정된 확률을 의미 할 수 있다.

[0038] 상기한 과제 해결 수단은, 본 발명의 특징을 모두 열거한 것은 아니다. 본 발명의 과제 해결을 위한 다양한 수단들은 이하의 상세한 설명의 구체적인 실시형태를 참조하여 보다 상세하게 이해될 수 있을 것이다.

발명의 효과

[0040] 본 발명의 일 실시형태에 따르면, 온라인 콘텐츠에 대한 머신러닝 학습을 기반으로 사용자가 요청한 온라인 콘텐츠의 댓글에 대한 진위 확률을 산출하여, 온라인 콘텐츠의 댓글에 대한 진위 여부를 판별할 수 있는 효과가 있다.

[0041] 또한, 본 발명의 일 실시형태에 따르면, 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 추정하고, 요청된 온라인 콘텐츠의 종류 및 주제를 학습모델에 따라 구분하여 그 댓글의 진위 여부를 판별할 수 있는 효과가 있다.

[0042] 또한, 본 발명의 일 실시형태에 따르면, 온라인 콘텐츠의 퍼블리셔 및 오너 간의 연관성에 따른 가중치를 설정하고, 이를 반영함으로써 온라인 콘텐츠의 분류를 보다 정확하게 수행할 수 있는 효과가 있다.

도면의 간단한 설명

[0044] 도 1은 본 발명의 일 실시예에 따른 머신 러닝을 이용한 가짜 댓글 판별 시스템의 일 적용예를 설명하는 도면이다.

도 2는 본 발명의 일 실시예에 따른 댓글 판별 서버의 예시적인 컴퓨팅 운영 환경을 설명하는 도면이다.

도 3은 본 발명의 일 실시예에 따른 댓글 판별 서버의 일 실시예를 설명하기 위한 블록 구성도이다.

도 4는 도 3에 도시된 학습 데이터 생성부의 일 실시예를 설명하기 위한 블록 구성도이다.

도 5는 도 3에 도시된 머신러닝 학습모델의 일 실시예를 설명하기 위한 블록 구성도이다.

도 6은 도 3에 도시된 진위 판별부의 일 실시예를 설명하기 위한 블록 구성도이다.

도 7은 도 3에 도시된 진위 판별부의 다른 일 실시예를 설명하기 위한 블록 구성도이다.

도 8은 도 3에 도시된 진위 판별부의 또 다른 일 실시예를 설명하기 위한 블록 구성도이다.

도 9 내지 도 15는 댓글 판별 서버의 각 구성요소를 설명하기 위한 예시적인 데이터들을 도시하는 도면이다.

도 16은 본 발명의 일 실시예에 따른 머신 러닝을 이용한 가짜 댓글 판별 방법을 설명하는 순서도다.

발명을 실시하기 위한 구체적인 내용

- [0045] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시 형태들을 설명한다.
- [0046] 그러나, 본 발명의 실시형태는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 이하 설명하는 실시 형태로 한정되는 것은 아니다. 또한, 본 발명의 실시형태는 당해 기술분야에서 평균적인 지식을 가진 자에게 본 발명을 더욱 완전하게 설명하기 위해서 제공되는 것이다.
- [0047] 즉, 전술한 목적, 특징 및 장점은 첨부된 도면을 참조하여 상세하게 후술되며, 이에 따라 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다. 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 상세한 설명을 생략한다. 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하기로 한다. 도면에서 동일한 참조부호는 동일 또는 유사한 구성요소를 가리키는 것으로 사용된다.
- [0048] 또한, 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서 상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다.
- [0049] 또한, 이하에서 본 발명에 따른 시스템을 설명하기 위하여 다양한 구성요소 및 그의 하부 구성요소에 대하여 설명하고 있다. 이러한 구성요소 및 그의 하부 구성요소들은, 하드웨어, 소프트웨어 또는 이들의 조합 등 다양한 형태로서 구현될 수 있다. 예컨대, 각 요소들은 해당 기능을 수행하기 위한 전자적 구성으로 구현되거나, 또는 전자적 시스템에서 구동 가능한 소프트웨어 자체이거나 그러한 소프트웨어의 일 기능적인 요소로 구현될 수 있다. 또는, 전자적 구성과 그에 대응되는 구동 소프트웨어로 구현될 수 있다.
- [0050] 본 명세서에 설명된 다양한 기법은 하드웨어 또는 소프트웨어와 함께 구현되거나, 적합한 경우에 이들 모두의 조합과 함께 구현될 수 있다. 본 명세서에 사용된 바와 같은 "부(Unit)", "서버(Server)" 및 "시스템(System)" 등의 용어는 마찬가지로 컴퓨터 관련 엔티티(Entity), 즉 하드웨어, 하드웨어 및 소프트웨어의 조합, 소프트웨어 또는 실행 시의 소프트웨어와 등가로 취급할 수 있다. 또한, 본 발명의 시스템에서 실행되는 각 기능은 모듈 단위로 구성될 수 있고, 하나의 물리적 메모리에 기록되거나, 둘 이상의 메모리 및 기록매체 사이에 분산되어 기록될 수 있다.
- [0051] 본 발명의 실시형태를 설명하기 위하여 다양한 순서도가 개시되고 있으나, 이는 각 단계의 설명의 편의를 위한 것으로, 반드시 순서도의 순서에 따라 각 단계가 수행되는 것은 아니다. 즉, 순서도에서의 각 단계는, 서로 동시에 수행되거나, 순서도에 따른 순서대로 수행되거나, 또는 순서도에서의 순서와 반대의 순서로도 수행될 수 있다.
- [0053] 이하에서는, 본 발명에 따른 머신 러닝을 이용한 가짜 댓글 판별 방법 및 댓글 판별 서버의 다양한 실시예들에 대하여 설명한다.
- [0055] 도 1은 본 발명의 일 실시예에 따른 머신 러닝을 이용한 가짜 댓글 판별 시스템의 일 적용예를 설명하는 도면이

다.

- [0056] 도 1을 참조하면, 머신 러닝을 이용한 가짜 댓글 판별 시스템은 사용자 단말(100), 온라인 콘텐츠 서버(200) 및 댓글 판별 서버(300)를 포함한다.
- [0057] 사용자 단말(100)은 사용자가 온라인 콘텐츠 서버(200) 또는 댓글 판별 서버(300)를 이용하는데 사용되는 단말로서, 스마트폰, 태블릿 PC, 노트북, PC 등 다양한 컴퓨팅 장치가 적용될 수 있다.
- [0058] 온라인 콘텐츠 서버(200)는 기사, 의견, 홍보 등의 온라인 콘텐츠를 온라인 상에서 생성 또는 유통하는 주체이다. 도시된 예에서, 온라인 콘텐츠 서버(200)는 하나의 서버로 예시되었으나, 이는 설명의 편의를 위하여 예시된 것에 불과하며, 실제로는 다양한 주체에 의하여 운영되는 다양한 온라인 콘텐츠 서버(200)가 존재할 수 있다.
- [0059] 댓글 판별 서버(300)는 온라인 콘텐츠에 대하여 머신러닝 학습을 수행하여 온라인 콘텐츠의 댓글에 대한 진위를 판별할 수 있다.
- [0060] 예컨대, 댓글 판별 서버(300)는 온라인 콘텐츠를 수집하여 머신러닝 학습을 위한 학습 데이터를 생성하고, 이러한 학습 데이터를 학습하여 온라인 콘텐츠에 대한 머신러닝 학습 모델을 생성할 수 있다. 댓글 판별 서버(300)는 머신러닝 학습 모델을 이용하여 사용자 단말(100)에서 제공된 요청 온라인 콘텐츠(이하, '요청 콘텐츠'라 함)의 댓글에 대한 진위 판별을 수행하여 사용자 단말(100)에 그 결과를 제공할 수 있다. 이러한 요청 콘텐츠는 온라인 콘텐츠 및 진위를 파악하고자 하는 댓글을 포함한다.
- [0062] 이하에서, 댓글 판별 서버(300)에 대하여 도 2 내지 도 14를 참조하여 보다 상세히 설명한다.
- [0064] 도 2는 본 발명의 일 실시예에 따른 댓글 판별 서버의 예시적인 컴퓨팅 운영 환경을 설명하는 도면이다.
- [0065] 도 2는 댓글 판별 서버(300)의 실시예들이 구현될 수 있는 적합한 컴퓨팅 환경의 일반적이고 단순화된 설명을 제공하기 위한 것으로, 도 6을 참조하면, 댓글 판별 서버(300)의 일 예로서 컴퓨팅 장치가 도시된다.
- [0066] 컴퓨팅 장치는 적어도 프로세싱 유닛(303)과 시스템 메모리(301)를 포함할 수 있다.
- [0067] 컴퓨팅 장치는 프로그램을 실행할 때 협조하는 복수의 프로세싱 유닛을 포함할 수도 있다. 컴퓨팅 장치의 정확한 구성 및 유형에 의존하여, 시스템 메모리(301)는 휘발성(예컨대, 램(RAM)), 비휘발성(예컨대, 롬(ROM), 플래시 메모리 등) 또는 이들의 조합일 수 있다. 시스템 메모리(301)는 플랫폼의 동작을 제어하기 위한 적합한 운영 체제(302)를 포함하는데, 예컨대 마이크로소프트사로부터의 WINDOWS 운영체제와 같은 것일 수 있다. 시스템 메모리(301)는 프로그램 모듈, 애플리케이션 등의 같은 하나 이상의 소프트웨어 애플리케이션을 포함할 수도 있다.
- [0068] 컴퓨팅 장치는 자기 디스크, 광학적 디스크, 또는 테이프와 같은 추가적인 데이터 저장 장치(304)를 포함할 수 있다. 이러한 추가적 저장소는 이동식 저장소 및/또는 고정식 저장소 일 수 있다. 컴퓨터 판독가능 저장 매체는 컴퓨터 판독가능 인스트럭션, 데이터 구조, 프로그램 모듈, 또는 기타 데이터와 같은 저장정보를 위한 임의의 방법이나 기법으로 구현되는 휘발성 및 비휘발성, 이동식 및 고정식 매체를 포함할 수 있다. 시스템 메모리(301), 저장소(304)는 모두 컴퓨터 판독가능 저장 매체의 예시일 뿐이다. 컴퓨터 판독가능 저장 매체는 램(RAM), 롬(ROM), EEPROM, 플래시 메모리 또는 다른 메모리 기법, CD-ROM, DVD 또는 다른 광학적 저장소, 자기 테이프, 자기 디스크 저장소 또는 다른 자기적 저장 장치, 또는 원하는 정보를 저장하고 컴퓨팅 장치(300)에 의해 액세스될 수 있는 임의의 다른 매체를 포함할 수 있는데, 이에 한정되는 것은 아니다.
- [0069] 컴퓨팅 장치의 입력 장치(305), 예컨대 키보드, 마우스, 펜, 음성 입력 장치, 터치 입력 장치, 및 비교 가능한 입력 장치를 포함할 수 있다. 출력 장치(306)는, 예컨대 디스플레이, 스피커, 프린터, 및 다른 유형의 출력 장치가 포함될 수도 있다. 이들 장치는 본 기술분야에서 널리 알려진 것이므로 자세한 설명은 생략한다.
- [0070] 컴퓨팅 장치는 예컨대 분산 컴퓨팅 환경에서의 네트워크, 예컨대, 유무선 네트워크, 위성 링크, 셀룰러 링크, 근거리 네트워크, 및 비교가능한 메커니즘을 통해 장치가 다른 장치들과 통신하도록 허용하는 통신 장치(307)를 포함할 수도 있다. 통신 장치(307)는 통신 매체의 한가지 예시이며, 통신 매체는 그 안에 컴퓨터 판독 가능 인스트럭션, 데이터 구조, 프로그램 모듈, 또는 다른 데이터를 포함할 수 있다. 예시적으로, 통신 매체는 유선 네

트위크나 직접 유선 접속과 같은 유선 매체, 및 음향, RF, 적외선 및 다른 무선 매체와 같은 무선 매체를 포함하는데, 이에 한정되는 것은 아니다.

- [0071] 댓글 판별 서버(300)는 이러한 컴퓨팅 환경에서 구현되는 기능적 구성으로 설명될 수 있다. 이에 대해서는, 도 3 내지 도 14를 참조하여 댓글 판별 서버의 다양한 실시예에 대하여 보다 상세히 설명한다.
- [0073] 도 3은 본 발명의 일 실시예에 따른 댓글 판별 서버의 일 실시예를 설명하기 위한 블록 구성도이다.
- [0074] 도 3을 참조하면, 댓글 판별 서버(300)는 요청 접수부(310), 학습 데이터 생성부(320), 머신러닝 학습모델(330) 및 진위 판별부(340)을 포함할 수 있다.
- [0075] 요청 접수부(310)는 사용자 단말(100)로부터 요청 콘텐츠를 접수받을 수 있다. 접수받은 요청 콘텐츠를 진위 판별부(340)에 제공할 수 있다.
- [0076] 학습 데이터 생성부(320)는 온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성할 수 있다.
- [0077] 도 9는 학습 데이터 생성부(320)에 의하여 수집된 온라인 콘텐츠의 일 예를 도시하는 도면으로서, 도 9에 도시된 바와 같이, 온라인 콘텐츠는 콘텐츠의 제목, 내용 및 댓글을 포함하고, 온라인 콘텐츠를 제공하는 퍼블리셔 정보 및 온라인 콘텐츠를 생성한 오피니언 정보를 포함할 수 있다.
- [0078] 학습 데이터 생성부(320)는 수집한 온라인 콘텐츠를 3가지 학습 데이터로 분류할 수 있다. 즉, 학습 데이터 생성부(320)에 의하여 생성되는 학습 데이터는, 종류 학습데이터, 주제 학습데이터 및 댓글 학습데이터로 구분될 수 있다.
- [0079] 종류 학습데이터는 온라인 콘텐츠의 종류를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 제목, 퍼블리셔 정보 및 오피니언 정보를 포함할 수 있다. 여기에서, 온라인 콘텐츠의 종류는 뉴스, 연예, 스포츠 및 기타로 구분될 수 있다.
- [0080] 주제 학습데이터는 온라인 콘텐츠의 주제를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 내용, 퍼블리셔 정보 및 오피니언 정보를 포함할 수 있다. 여기에서, 온라인 콘텐츠의 주제는 스트레이트, 오피니언, 홍보 및 기타로 구분될 수 있다. 스트레이트는 온라인 콘텐츠가 사건이나 사고와 관련된 정보를 포함하는 것을 의미하고, 오피니언은 특정 사안에 대해서 가치 또는 평가를 부여하는 정보를 포함하는 경우이다. 홍보는 온라인 콘텐츠가 흥미, 오락 내지 상업적 정보를 제공하고 있을 때 해당된다.
- [0081] 댓글 학습데이터는 가짜 댓글을 식별하기 위한 온라인 콘텐츠의 댓글을 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 종류, 주제, 가짜 댓글 정보를 포함할 수 있다.
- [0082] 일 실시예에서, 댓글 학습데이터는 협력자에 의하여 생성된 온라인 댓글을 기반으로 생성될 수 있다. 즉, 협력자는 온라인 콘텐츠에 대한 댓글 시나리오에 따라 온라인 콘텐츠에 댓글을 달 수 있다. 협력자에 의하여 댓글 시나리오에 의하여 작성된 가짜 댓글을 시나리오 가짜 댓글로, 사용자에게 의하여 자연적으로 생성된 댓글을 자연적 댓글로 구분한다. 학습 데이터 생성부(320)가 수집한 댓글 학습데이터에 대하여, 시나리오 가짜 댓글 및 자연적 댓글에 대하여 진위 라벨링을 부여할 수 있다. 예컨대, 시나리오 가짜 댓글에는 0 라벨을, 자연적 댓글에는 1 라벨을 부여할 수 있다.
- [0083] 이와 같이, 본 명세서에서의 댓글의 진위란, 고의적인 의도로 작성된 가짜 댓글(시나리오 댓글)과 사용자에게 의하여 자연적으로 생성된 진짜 댓글(자연적 댓글)을 구분하는 것을 의미한다. 즉, 댓글의 내용의 진위성이 아닌, 댓글의 생성에 대한 진위성을 의미한다.
- [0084] 일 실시예에서, 학습 데이터 생성부(320)는 초기 댓글이 시나리오 댓글인지 여부에 따라, 자연적으로 발생하는 가짜 댓글의 변동성을 확인하여 진위성 보정 계수를 산출할 수 있다. 이를 위하여, 시나리오 작성 댓글에는 0 라벨이, 협력자가 아닌 타 사용자에게 의하여 생성된 가짜 댓글에 대해서도 0 라벨이, 자연적 진짜 댓글에는 1 라벨이 부여될 수 있다. 즉, 자연적 댓글 중 댓글 시나리오를 만족하는 댓글은 협력자가 아닌 타 사용자에게 의하여 생성된 가짜 댓글로 판정되어 0 라벨이 부여될 수 있다.
- [0085] 이러한 진위 라벨링을 기초로, 온라인 콘텐츠의 초기 댓글이 가짜 댓글인 경우, 그러한 초기 가짜 댓글이 이후에 생성되는 댓글의 진위성에 영향을 미치는 정도를 산출하여 진위성 보정 계수를 산출할 수 있다.

- [0086] 예컨대, 학습 데이터 생성부(320)는 초기 댓글이 자연적 진짜 댓글인 경우 이후에 발생하는 댓글의 진위성을 산출하고, 초기 댓글이 시나리오 가짜 댓글인 경우 이후에 발생하는 댓글의 진위성을 산출한 후, 이들을 비교하여 초기 댓글이 시나리오 가짜 댓글인 경우에 의한 댓글의 진위성에 영향을 미치는 정도를 산출할 수 있다.
- [0087] 머신러닝 학습모델(330)은 학습 데이터 생성부(320)에서 제공된 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인 콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습할 수 있다.
- [0088] 일 실시예에서, 머신러닝 학습모델(330)은 온라인 콘텐츠의 종류에 대한 종류 학습모델, 온라인 콘텐츠의 주제에 대한 주제 학습모델 및 온라인 콘텐츠의 댓글에 대한 댓글 학습모델을 포함할 수 있다.
- [0089] 일 예로, 머신러닝 학습모델(330)은 온라인 콘텐츠의 텍스트에 대한 불용어를 제거하고 표제어를 추출하고, 학습데이터 별로 다르게 설정된 온라인 콘텐츠의 각 기준 항목을 대상으로 N-gram 언어모델을 적용하고 벡터화를 적용하여 분류 학습을 수행함으로써 학습 모델을 생성할 수 있다. 다만, 이는 학습 모델의 예시적인 일 예를 설명한 것으로, 본 발명에서의 머신러닝 학습모델(330)이 이러한 알고리즘 기법으로만 한정되는 것은 아니며 다양한 다른 머신러닝 학습기법으로 구현 가능함은 자명하다.
- [0090] 진위 판별부(340)는 사용자로부터 요청된 요청 콘텐츠에 대하여 머신러닝 학습모델(330)을 적용하여 요청 콘텐츠의 종류 및 주제를 식별하고, 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 요청 콘텐츠의 댓글의 진위를 판별할 수 있다. 진위 판별부(340)는 판별된 요청 콘텐츠의 댓글의 진위에 대한 정보를 사용자 단말(100)에 제공할 수 있다.
- [0091] 이러한 진위 판별부(340)에 대해서는 도 6 내지 도 8을 참조하여 이하에서 보다 상세히 설명한다.
- [0093] 도 4는 도 3에 도시된 학습 데이터 생성부의 일 실시예를 설명하기 위한 블록 구성도이다.
- [0094] 도 4를 참조하면, 학습 데이터 생성부(320)는 온라인 콘텐츠 수집모듈(321), 종류 학습데이터 수집모듈(322), 주제 학습데이터 수집모듈(323) 및 댓글 학습데이터 수집모듈(324)을 포함할 수 있다.
- [0095] 온라인 콘텐츠 수집모듈(321)은 온라인 콘텐츠를 수집하여 종류 학습데이터 수집모듈(322) 내지 댓글 학습데이터 수집모듈(324)에 제공한다.
- [0096] 종류 학습데이터 수집모듈(322)은 수집된 온라인 콘텐츠 중에서, 종류 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집할 수 있다. 종류 학습데이터 수집모듈(322)은 생성된 종류 학습데이터에 대하여, 각 종류 별로 퍼블리셔-오너 연관 정보를 생성할 수 있다. 도 10은 종류 학습데이터 수집모듈(322)에 의하여 생성되는 종류 학습 데이터의 종류별 퍼블리셔-오너 연관 정보의 일 예를 도시한다. 도 10에 도시된 예와 같이, 종류 학습데이터 수집모듈(322)은 뉴스 종류, 연예 종류, 스포츠 종류 각각에 대하여 퍼블리셔와 오너 간의 연관성에 대한 정보를 생성할 수 있다.
- [0097] 주제 학습데이터 수집모듈(323)은 수집된 온라인 콘텐츠 중에서, 주제 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집할 수 있다. 생성된 주제 학습데이터에 대하여, 각 주제 별로 퍼블리셔-오너 연관 정보를 생성할 수 있다. 도 11은 주제 학습데이터 수집모듈(323)에 의하여 생성되는 주제 학습 데이터의 주제별 퍼블리셔-오너 연관 정보의 일 예를 도시한다. 도 11에 도시된 예와 같이, 주제 학습데이터 수집모듈(323)은 스트레이트 주제, 오피니온 주제 및 홍보 주제 각각에 대하여 퍼블리셔와 오너 간의 연관성에 대한 정보를 생성할 수 있다.
- [0098] 이와 같이, 특정 종류 또는 특정 주제에 대하여 퍼블리셔-오너 연관성이 높다는 것은, 그러한 퍼블리셔-오너 관계에서 해당 종류 해당 주제의 온라인 콘텐츠를 다수 생산하는 것을 반영하는 것이므로, 이러한 퍼블리셔-오너 연관성은 진위 판별부(340)에 의하여 진위 판별 시 적용되는 보정을 위하여 사용된다.
- [0099] 댓글 학습데이터 수집모듈(324)은 수집된 온라인 콘텐츠 중에서, 댓글 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 수집할 수 있다.
- [0100] 댓글 학습데이터 수집모듈(324)은 생성된 댓글 학습 데이터에 대하여, 각 종류-주제 별 가짜 댓글 정보를 생성할 수 있다. 도 12는 댓글 학습데이터 수집모듈(324)에 의하여 생성되는 댓글 학습 데이터의 종류-주제별 가짜 댓글 정보의 일 예를 도시한다. 여기에서, N_{ij} 는 온라인 콘텐츠의 종류 i 및 주제 j 에 해당하는 가짜 댓글의 수를 의미한다.
- [0101] 이와 같이 생성된 댓글 학습 데이터의 종류-주제별 가짜 댓글 정보는 요청 콘텐츠의 종류-주제에 따른 댓글 확

를을 산출하는데 사용된다.

- [0103] 도 5는 도 3에 도시된 머신러닝 학습모델의 일 실시예를 설명하기 위한 블록 구성도이다.
- [0104] 도 5를 참조하면, 머신러닝 학습모델(330)은 종류 학습모델(332), 주제 학습모델(333) 및 댓글 학습모델(334)을 포함할 수 있다. 실시예에 따라, 머신러닝 학습모델(330)은 전처리 모듈(331)을 더 포함할 수 있다.
- [0105] 전처리 모듈(331)은 온라인 콘텐츠를 텍스트화 하고, 학습에 필요한 소정의 전처리를 수행할 수 있다. 예컨대, 전처리 모듈(331)은 음성 기반 온라인 콘텐츠를 식별하고, 음성 인식을 이용하여 음성 기반 온라인 콘텐츠를 텍스트 콘텐츠로 변환하며, 텍스트 콘텐츠에 대하여 불용어를 제거하고 표제어를 추출하는 전처리를 수행할 수 있다.
- [0106] 종류 학습모델(332)은 종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 학습하여 온라인 콘텐츠의 종류 확률을 산출할 수 있다. 예컨대, 종류 학습모델(332)은 종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 벡터화 하고, 그러한 벡터에 대하여 SVM 또는 DNN 을 적용하여 종류 분류 학습을 수행할 수 있다.
- [0107] 따라서, 종류 학습모델(332)은 온라인 콘텐츠의 제목을 학습하여 해당 온라인 콘텐츠의 종류에 대한 확률, 즉, 종류 확률을 산출할 수 있다. 일 예로, 온라인 콘텐츠의 종류가 '뉴스, 연예, 스포츠, 기타'로 구성되는 경우, 종류 학습모델(332)은 " $P(\text{뉴스}) = 0.3$, $P(\text{연예}) = 0.4$, $P(\text{스포츠}) = 0.2$, $P(\text{기타}) = 0.1$ " 등과 같이, 온라인 콘텐츠의 제목으로부터 어느 종류에 해당될 가능성(종류 확률)을 산출할 수 있다.
- [0108] 주제 학습모델(333)은 주제 학습데이터를 기초로 온라인 콘텐츠의 내용을 학습하여 온라인 콘텐츠의 주제 확률을 산출할 수 있다. 마찬가지로, 주제 학습모델(333)은 온라인 콘텐츠의 내용을 학습하여 해당 온라인 콘텐츠의 주제에 대한 확률, 즉, 주제 확률을 산출할 수 있다. 일 예로, 온라인 콘텐츠의 주제가 '스트레이트, 오피니언, 홍보, 기타'로 구성되는 경우, 종류 학습모델(332)은 " $P(\text{스트레이트}) = 0.2$, $P(\text{오피니언}) = 0.6$, $P(\text{홍보}) = 0.2$, $P(\text{기타}) = 0.1$ " 등과 같이, 온라인 콘텐츠의 내용으로부터 어느 주제에 해당될 가능성(주제 확률)을 산출할 수 있다.
- [0109] 댓글 학습모델(334)은 댓글 학습데이터를 기초로 온라인 콘텐츠의 종류 및 주제 별 가짜 댓글의 확률을 산출할 수 있다.
- [0110] 도 13은 댓글 학습모델(334)에 의하여 생성된 온라인 콘텐츠의 종류-주제 별 가짜 댓글의 비율의 일 예를 예시하는 도면으로서, 여기에서, w_{ij} 는 온라인 콘텐츠의 종류 i 및 주제 j 일 때의 가짜 댓글의 수를 전체 가짜 댓글의 수로 나눈 값일 수 있다.
- [0112] 도 6은 도 3에 도시된 진위 판별부의 일 실시예를 설명하기 위한 블록 구성도이다.
- [0113] 진위 판별부(340)는 사용자 단말(100)로부터 댓글 진위 확인을 요청하는 요청 콘텐츠를 제공받고, 요청 콘텐츠의 댓글 진위를 판별 할 수 있다.
- [0114] 도 6을 참조하면, 진위 판별부(340)는 종류 판별모듈(342), 주제 판별모듈(343) 및 댓글 진위 판별모듈(344)를 포함할 수 있다. 실시예에 따라, 진위 판별부(340)는 전처리 모듈(341)을 더 포함할 수 있다.
- [0115] 전처리 모듈(341)은 온라인 콘텐츠를 텍스트화 할 수 있다. 예컨대, 전처리 모듈(341)은 음성 기반 온라인 콘텐츠를 식별하고, 음성 인식을 이용하여 음성 기반 온라인 콘텐츠를 텍스트 콘텐츠로 변환할 수 있다.
- [0116] 종류 판별모듈(342)은 종류 학습모델(332)을 이용하여 요청 콘텐츠의 종류를 판별할 수 있다. 주제 판별모듈(343)은 주제 학습모델(333)을 이용하여 요청 콘텐츠의 주제를 판별할 수 있다.
- [0117] 댓글 진위 판별모듈(344)은 댓글 학습모델(334)을 이용하여, 판별된 요청 콘텐츠의 종류 및 주제에 따른 댓글 진위 확률을 산출할 수 있다.
- [0118] 도 14는 요청 콘텐츠의 일 예를 도시하고 있다.
- [0119] 도 14의 예를 들면, 종류 판별모듈(342)은 종류 학습모델(332)을 이용하여 요청 콘텐츠의 종류를 판단할 수 있다. 즉, 종류 학습모델(332)은 요청 콘텐츠의 제목에 대하여 학습을 수행하여 요청 콘텐츠의 종류 확률을 산출할 수 있다. 종류 학습모델(332)에서 산출된 종류 확률이 " $P(\text{뉴스}) = 0.4$, $P(\text{연예}) = 0.3$, $P(\text{스포츠}) = 0.2$, $P(\text{기타}) = 0.1$ " 등과 같이, 온라인 콘텐츠의 제목으로부터 어느 종류에 해당될 가능성(종류 확률)을 산출할 수 있다.

타)=0.1"라고 하면, 종류 판별모듈(342)은 가장 종류 확률이 높은 '뉴스' 종류를 요청 콘텐츠의 종류로서 판단할 수 있다.

[0120] 또한, 도 14의 예에서, 주제 판별모듈(343)은 주제 학습모델(333)을 이용하여 요청 콘텐츠의 주제를 판별할 수 있다. 즉, 주제 학습모델(333)은 요청 콘텐츠의 내용에 대하여 학습을 수행하여 요청 콘텐츠의 주제 확률을 산출할 수 있다. 주제 학습모델(333)에서 산출된 주제 확률이 "P(스트레이트)= 0.5, P(오피니언)=0.3, P(홍보) = 0.1, P(기타)=0.1"라고 하면, 주제 판별모듈(343)은 가장 주제 확률이 높은 '스트레이트' 주제를 요청 콘텐츠의 주제로서 판단할 수 있다.

[0121] 즉, 도 14의 예에서, 댓글 학습모델(334)은 '뉴스 종류-스트레이트 주제'의 댓글 진위 확률을 산출할 수 있으며, 도 13의 예를 참조하면 댓글 학습모델(334)은 '뉴스 종류-스트레이트 주제'에 해당하는 댓글 진위 비율 ω_{11} 을 확인하고, 댓글 진위 판별모듈(344)은 ω_{11} 을 댓글 진위 확률로서 산출할 수 있다.

[0123] 도 7은 도 3에 도시된 진위 판별부의 다른 일 실시예를 설명하기 위한 블록 구성도이다. 도 7에 도시된 예는, 진위 판별부에 종류 보정 및 주제 보정이 적용된 실시예에 관한 것이다.

[0124] 도 7을 참조하면, 진위 판별부(340)는 전처리 모듈(341), 종류 판별모듈(342), 주제 판별모듈(343), 댓글 진위 판별모듈(344), 종류 보정모듈(345) 및 주제 보정모듈(346)을 포함할 수 있다.

[0125] 전처리 모듈(341) 내지 댓글 진위 판별모듈(344)에 대해서는 도 6을 참조하여 상술한 바로부터 쉽게 이해할 수 있으므로, 여기에서는 그 설명을 생략한다.

[0126] 종류 보정모듈(345)은 종류 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 종류 보정치를 설정하고, 종류 판별모듈(342)의 판별 결과에 종류 보정치를 반영할 수 있다.

[0127] 주제 보정모듈(346)은 주제 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 주제 보정치를 설정하고, 주제 판별모듈(343)의 판별 결과에 주제 보정치를 반영할 수 있다.

[0128] 도 14에 도시된 요청 콘텐츠를 기초로 종류 보정 및 주제 보정에 대하여 보다 상세히 설명한다.

[0129] 도 14에 도시된 요청 콘텐츠에 대한 종류 판별모듈(342)의 판별 결과가 "P(뉴스)= 0.3, P(연예)=0.4, P(스포츠) = 0.2, P(기타)=0.1"이고, 종류 보정모듈(345)에서 퍼블리셔-오너 간 연관성, 즉, '네이버-News1'간의 연관성이 뉴스 종류에 대하여 1.5 보정치를 가진다고 하자. 따라서, 종류 판별모듈(342)의 판별 결과에 종류 보정모듈(345)의 보정치를 반영하면, 'P(뉴스)= 0.3*1.5=0.45'가 되므로 'P(연예)=0.4'보다 높게 설정되어, 요청 콘텐츠는 '뉴스 종류'로 판정된다.

[0130] 또한, 도 14에 도시된 요청 콘텐츠에 대한 주제 판별모듈(343)의 판별 결과가 "P(스트레이트)= 0.2, P(오피니언)=0.6, P(홍보) = 0.2, P(기타)=0.1"이고, 주제 보정모듈(346)에서 퍼블리셔-오너 간 연관성, 즉, '네이버-News1'간의 연관성이 스트레이트 주제에 대하여 1.5 보정치를 가진다고 하자. 따라서, 주제 판별모듈(343)의 판별 결과에 주제 보정모듈(346)의 보정치를 반영하면, P(스트레이트)= 0.2*1.5=0.3'가 되나 여전히 'P(오피니언)=0.6'이 가장 높게 설정되므로, 요청 콘텐츠는 '오피니언 주제'로 판정된다.

[0132] 도 8은 도 3에 도시된 진위 판별부의 또 다른 일 실시예를 설명하기 위한 블록 구성도이다. 도 8에 도시된 예는, 진위 판별부에 댓글 보정이 적용된 실시예에 관한 것이다.

[0133] 도 8을 참조하면, 진위 판별부(340)는 전처리 모듈(341), 종류 판별모듈(342), 주제 판별모듈(343), 댓글 진위 판별모듈(344) 및 댓글 진위 보정모듈(347)을 포함할 수 있다. 도시되지는 않았으나, 실시예에 따라 진위 판별부(340)는 종류 보정모듈(345) 및 주제 보정모듈(346)를 더 포함할 수 있다.

[0134] 전처리 모듈(341) 내지 주제 보정모듈(346)에 대해서는 도 6 내지 도 7을 참조하여 상술한 바로부터 쉽게 이해할 수 있으므로, 여기에서는 그 설명을 생략한다.

[0135] 댓글 진위 보정모듈(347)은 도 15에 도시된 예와 같이, 아래의 수학식 1 및 2를 이용하여 댓글 진위 판별모듈(344)의 판별 결과를 보정할 수 있다.

[0136] [수학식 1]

$$\tilde{p} = f(\tilde{x} + 3 \times w_{ij}), \text{ where } \tilde{x} = f^{-1}(p)$$

[0137]

[0138] [수학식 2]

$$f(x) = \frac{1}{1 + e^{-x}}$$

[0139]

[0140] 여기에서, p 는 상기 댓글 진위 판별모듈의 판별 결과를, $\tilde{p}$ 는 상기 댓글 진위 보정모듈에 의한 보정된 확률을 의미한다.

[0141] 즉, 댓글 진위 보정모듈(347)은 sigmoid 함수의 역함수를 활용하여 댓글 진위를 보정할 수 있으며, 상술한 수학식 1의 예 외에도, 함수에 적용되는 수식의 변형이 가능하다. 예컨대, w_{ij} 에 승산되는 상수의 값이 3이 아닌 2나 1로도 적용 가능하다.

[0142] 일 실시예에서, 댓글 진위 보정모듈(347)은 요청 콘텐츠의 주제 중 홍보 주제에 대한 부가 가중치를 반영하여 댓글 진위 확률을 보정할 수 있다. 이는, 홍보 종류의 온라인 콘텐츠는 다수의 인위적인 댓글들이 달리는 것이 통상적이므로, 주제 판별모듈(243) 내지 주제 보정모듈(346)에 의하여 판정된 요청 콘텐츠의 주제가 '홍보 주제'로 판별되는 경우, 댓글 진위 보정모듈(347)은 거짓 확률에 부가 가중치를 부여할 수 있다.

[0143] 예컨대, 댓글 진위 판별모듈(344)에서 '연예 뉴스'의 '홍보 종류'에 대한 요청 콘텐츠를 판별한 결과, "P(진짜)=0.6, P(가짜)=0.4)"와 같이 산출되었고, 댓글 진위 보정모듈(347)이 '연예 뉴스'의 '홍보 종류'에 대한 가중치를 도 13으로부터 $w_{23}=0.3$ 으로 설정하였다고 하면, 댓글 진위 보정모듈(347)에 의하여 보정되면 "P(가짜)= $f(x(1+w_{23}))=f(-0.4+3*0.3)=f(0.5)=0.62$ "이므로 'P(진짜)'보다 높아 가짜 댓글로 판단할 수 있다.

[0145] 일 실시예에서, 요청 콘텐츠는 온라인 콘텐츠와 그의 첫 댓글 및 진위 판단하고자 하는 댓글을 포함할 수 있고, 댓글 진위 보정모듈(347)은 진위성 보정 계수를 이용하여 진위 판단하고자 하는 댓글의 진위 판단을 보정할 수 있다.

[0146] 즉, 전술한 바와 같이, 학습 데이터 생성부(320)는 초기 댓글이 시나리오 댓글인지 여부에 따라, 자연적으로 발생하는 가짜 댓글의 변동성을 확인하여 진위성 보정 계수를 산출할 수 있으며, 댓글 진위 보정모듈(347)은 이러한 진위성 보정 계수를 반영하여 댓글 진위를 보정할 수 있다.

[0147] 즉, 댓글 진위 보정모듈(347)은 요청 콘텐츠의 제1 댓글의 진위성을 판단하고, 제1 댓글이 거짓 댓글인 경우, 그 이후에 달린 진위를 판단하고자 하는 제2 댓글에 진위성 보정 계수-제1 댓글의 진위에 따라 결정되는 계수를 반영할 수 있다.

[0148] 예컨대, 댓글 진위 판별모듈(344)이 진위 판단하고자 하는 제2 댓글에 대한 진위성을 "P(진짜) = 0.6, P(가짜) = 0.4"로 산출하였고, 진위성 보정 계수가 2이라고 가정하자. 댓글 진위 보정모듈(347)은 댓글 진위 판별모듈(344)을 이용하여 요청 콘텐츠의 제1 댓글의 진위성을 판단할 수 있고, 만약 요청 콘텐츠의 제1 댓글이 진짜인 경우, 제2 댓글에 대하여 댓글 진위 판별모듈(344)에서 판별한 진위성, 즉, "P(진짜) = 0.6, P(가짜) = 0.4"을 그대로 이용하여, 진위 판단하고자 하는 제2 댓글은 진짜로 판별될 수 있다.

[0149] 한편, 요청 콘텐츠의 제1 댓글이 가짜인 경우, 댓글 진위 판별모듈(344)에서 판별한 진위성에 진위성 보정 계수 2를 적용하여 제2 댓글에 대한 댓글 진위를 판별할 수 있다. 즉, 제2 댓글에 대한 댓글 진위를 판별 시, 제1 댓글이 가짜임에 의하여 진위성 보정 계수 2를 적용할 수 있으며, 그에 따라 'P(가짜) = 0.4 * 2 = 0.8'이 되어 'P(진짜) = 0.6'보다 높으므로, 진위 판단하고자 하는 댓글은 가짜로 판별될 수 있다.

[0151] 일 실시예에서, 댓글 진위 보정모듈(347)은 가짜 댓글의 빈도수를 일정 주기로 업데이트하여 보정계수를 관리할 수 있다. 이는 그동안의 사용자의 요청에 의한 온라인 콘텐츠의 분류를 기반으로 업데이트 하는 경우를 들어 설명할 수 있다.

- [0152] 예컨대, 100개의 사용자 요청에 의해 분류된 10개의 가짜 댓글 중에서 각 종류-주제별 빈도를 도 12에 도시된 표의 각 경우에 맞춰 가중 합산 할 수 있다. 예를 들어, 10개의 가짜 댓글이 모두 뉴스-스트레이트인 경우 기존의 N_{11} 이 ' $N_{11} + 10$ '으로 보정되게 되고, 그에 따라, 도 13에 도시된 비율도 이에 따라 자동으로 변경되어 반영될 수 있다.
- [0154] 이상에서는 도 1 내지 도 15를 참조하여, 머신 러닝을 이용한 가짜 댓글 판별 시스템의 다양한 실시예들에 대하여 설명하였다.
- [0155] 이하에서는, 본 발명의 일 실시예에 따른 머신 러닝을 이용한 가짜 댓글 판별 방법에 대하여 설명한다.
- [0156] 이하에서 설명할 머신 러닝을 이용한 가짜 댓글 판별 방법은, 도 1 내지 도 15를 참조하여 기 설명한 머신 러닝을 이용한 가짜 댓글 판별 시스템을 기초로 수행되므로, 도 1 내지 도 15에서 기 설명한 내용을 참조하여 보다 쉽게 이해할 수 있다.
- [0158] 도 16은 본 발명의 일 실시예에 따른 머신 러닝을 이용한 가짜 댓글 판별 방법을 설명하는 순서도다.
- [0159] 도 16을 참조하면, 댓글 판별 서버(300)는 온라인 콘텐츠를 수집하고, 수집된 온라인 콘텐츠를 온라인 콘텐츠의 종류, 주제 및 댓글을 기준으로 구분하여 학습 데이터를 생성 할 수 있다(S1610).
- [0160] 댓글 판별 서버(300)는 학습 데이터를 기반으로 머신러닝 학습을 수행하여, 온라인 콘텐츠의 종류, 주제 및 댓글 각각에 대한 개별적인 확률을 학습하여 머신러닝 학습 모델을 생성 할 수 있다(S1620).
- [0161] 댓글 판별 서버(300)는 사용자로부터 요청된 요청 콘텐츠에 대하여 머신러닝 학습 모델을 적용하여 상기 요청 콘텐츠의 종류 및 주제를 식별하고(S1630), 온라인 콘텐츠의 종류 및 주제에 따른 댓글의 진위 성향을 이용하여 상기 요청 콘텐츠의 댓글의 진위를 판별 할 수 있다(S1640).
- [0162] 일 실시예에서, 학습 데이터는, 온라인 콘텐츠의 종류를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 제목, 퍼블리셔 정보 및 오너 정보를 포함하는 종류 학습데이터, 온라인 콘텐츠의 주제를 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 내용, 퍼블리셔 정보 및 오너 정보를 포함하는 주제 학습데이터 및 온라인 콘텐츠의 댓글을 학습하기 위한 학습 데이터로서, 온라인 콘텐츠의 종류, 주제, 가짜댓글 정보를 포함하는 댓글 학습 데이터를 포함할 수 있다.
- [0163] 단계 S1610에 대한 일 실시예에서, 댓글 판별 서버(300)는, 온라인 콘텐츠를 수집하는 단계, 수집된 온라인 콘텐츠 중에서 종류 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 종류 학습데이터를 수집하는 단계, 수집된 온라인 콘텐츠 중에서 주제 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 주제 학습데이터를 수집하는 단계 및 수집된 온라인 콘텐츠 중에서 댓글 학습데이터로 적용 가능한 온라인 콘텐츠를 선별하여 댓글 학습 데이터를 수집하는 단계를 수행할 수 있다.
- [0164] 일 실시예에서, 머신러닝 학습모델은, 온라인 콘텐츠의 텍스트에 대한 불용어를 제거하고 표제어를 추출하고, 학습데이터 별로 다르게 설정된 온라인 콘텐츠의 각 기준 항목을 대상으로 N-gram 언어모델을 적용하고 벡터화를 적용하여 분류 학습을 수행할 수 있다.
- [0165] 단계 S1620에 대한 일 실시예에서, 댓글 판별 서버(300)는, 종류 학습데이터를 기초로 온라인 콘텐츠의 제목을 학습하여 온라인 콘텐츠의 종류 확률을 산출하는 종류 학습모델을 생성하는 단계, 주제 학습데이터를 기초로 온라인 콘텐츠의 내용을 학습하여 온라인 콘텐츠의 주제 확률을 산출하는 주제 학습모델을 생성하는 단계 및 댓글 학습데이터를 기초로 온라인 콘텐츠의 종류 및 주제 별 가짜 댓글의 확률을 산출하는 댓글 학습모델을 생성하는 단계를 수행할 수 있다.
- [0166] 단계 S1630에 대한 일 실시예에서, 댓글 판별 서버(300)는, 종류 학습모델을 이용하여 요청 콘텐츠의 종류를 판별하는 단계, 주제 학습모델을 이용하여 요청 콘텐츠의 주제를 판별하는 단계 및 댓글 학습모델을 이용하여 요청 콘텐츠의 종류 및 주제에 따른 댓글 진위 확률을 산출하는 단계를 수행할 수 있다.
- [0167] 단계 S1630에 대한 일 실시예에서, 댓글 판별 서버(300)는, 종류 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 종류 보정치를 설정하고, 종류 판별모델의 판별 결과에 상기 종류 보정치를 반영하는 단계 및 주제 학습데이터를 대상으로 퍼블리셔와 오너 간의 연관성을 기초로 주제 보정치를 설정하고, 주제 판별모델의

판별 결과에 상기 주제 보정치를 반영하는 단계를 수행할 수 있다.

[0168] 단계 S1640에 대한 일 실시예에서, 댓글 판별 서버(300)는 상술한 수학적 1 및 2를 이용하여 상기 댓글 진위 판별모듈의 판별 결과를 보정하는 단계를 수행할 수 있다.

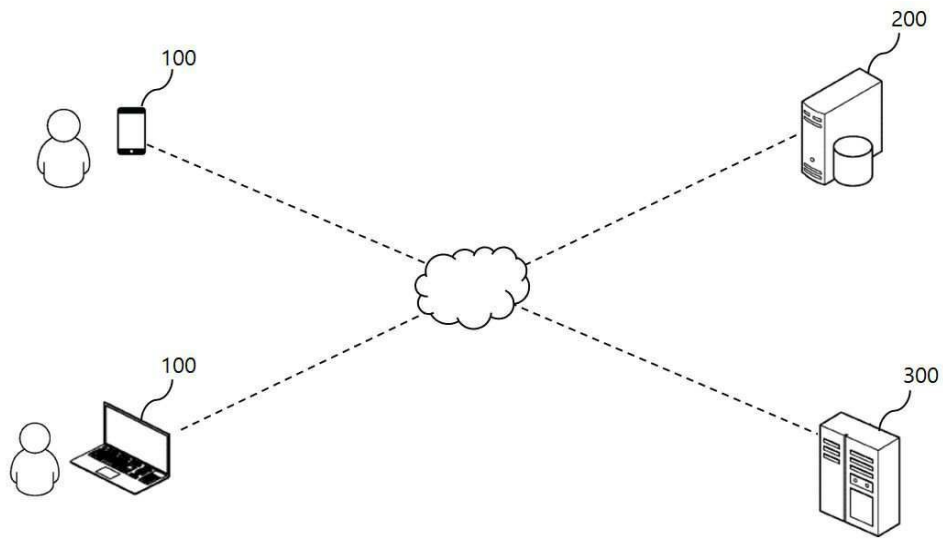
[0170] 이상에서 설명한 본 발명은 전술한 실시예 및 첨부된 도면에 의해 한정되는 것이 아니고 후술하는 특허청구범위에 의해 한정되며, 본 발명의 구성은 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 그 구성을 다양하게 변경 및 개조할 수 있다는 것을 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 쉽게 알 수 있다.

부호의 설명

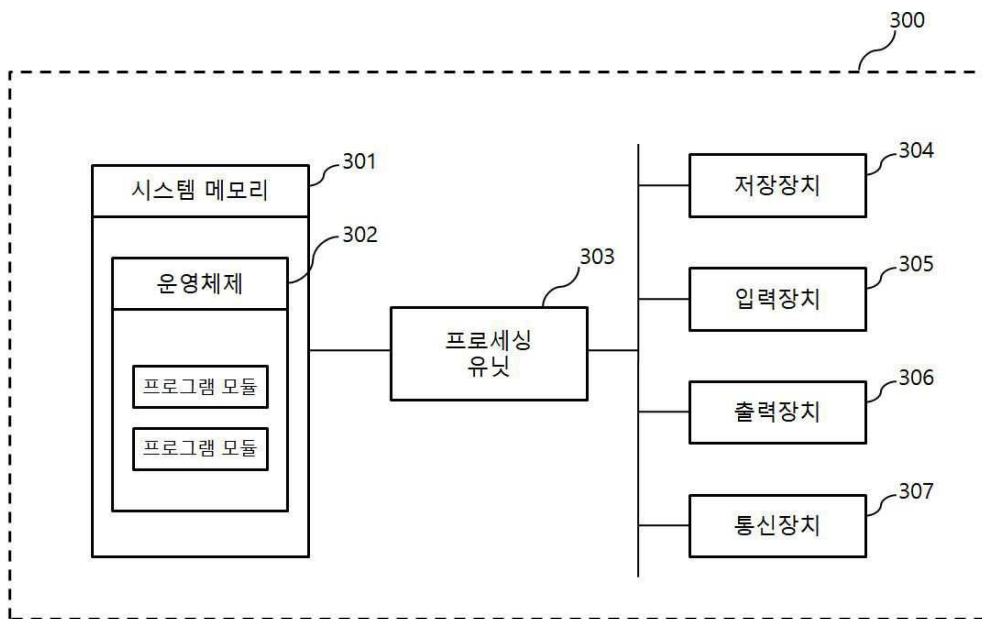
[0172] 100 : 사용자 단말
200 : 온라인 콘텐츠 서버
300 : 댓글 판별 서버
301 : 시스템 메모리 302 : 운영체제
303 : 프로세싱 유닛 304 : 저장소
305 : 입력장치 306 : 출력장치
307 : 통신장치
310 : 요청 접수부 320 : 학습 데이터 생성부
330 : 머신러닝 학습모델 340 : 진위 판별부
321 : 온라인 콘텐츠 수집모듈 322 : 종류 학습데이터 수집모듈
323 : 주제 학습데이터 수집모듈 324 : 댓글 학습데이터 수집모듈
331 : 전처리 모듈 332 : 종류 학습모델
333 : 주제 학습모델 334 : 댓글 학습모델
341 : 전처리 모듈 342 : 종류 판별모듈
343 : 주제 판별모듈 344 : 댓글 진위 판별모듈
345 : 종류 보정모듈 346 : 주제 보정모듈
347 : 댓글진위 보정모듈

도면

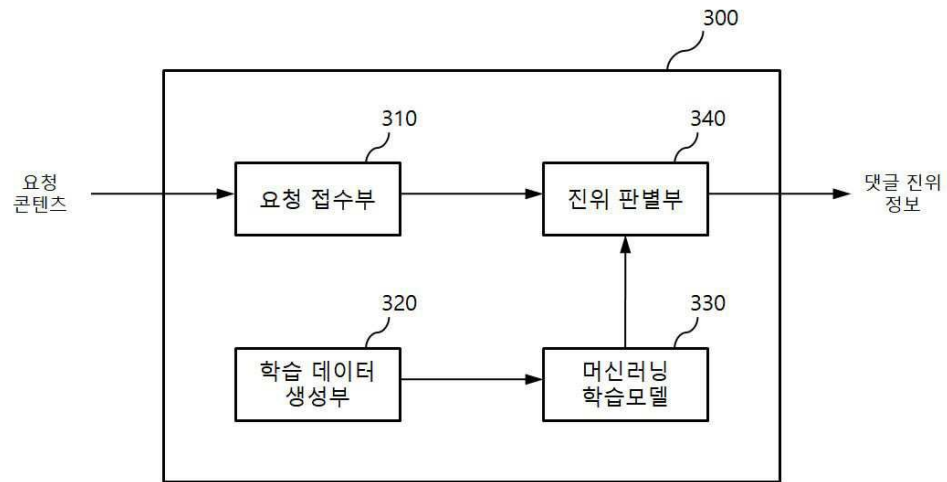
도면1



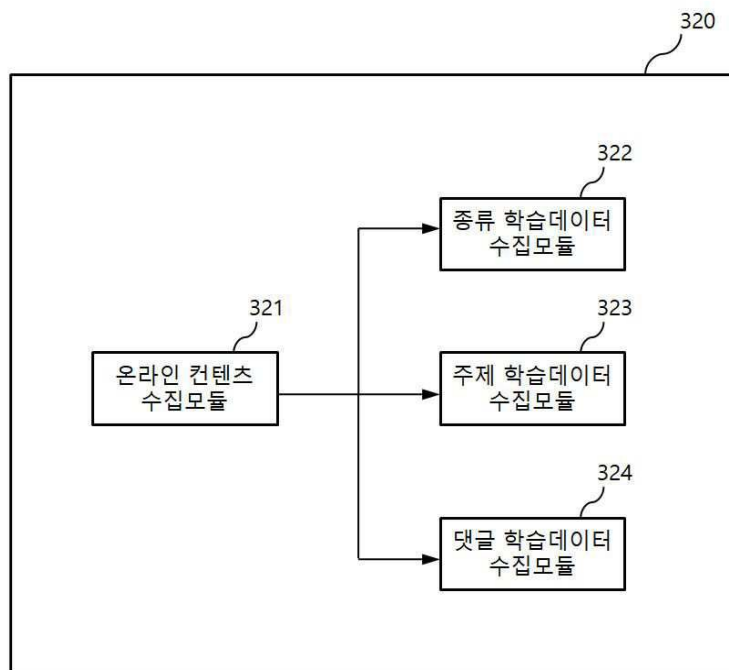
도면2



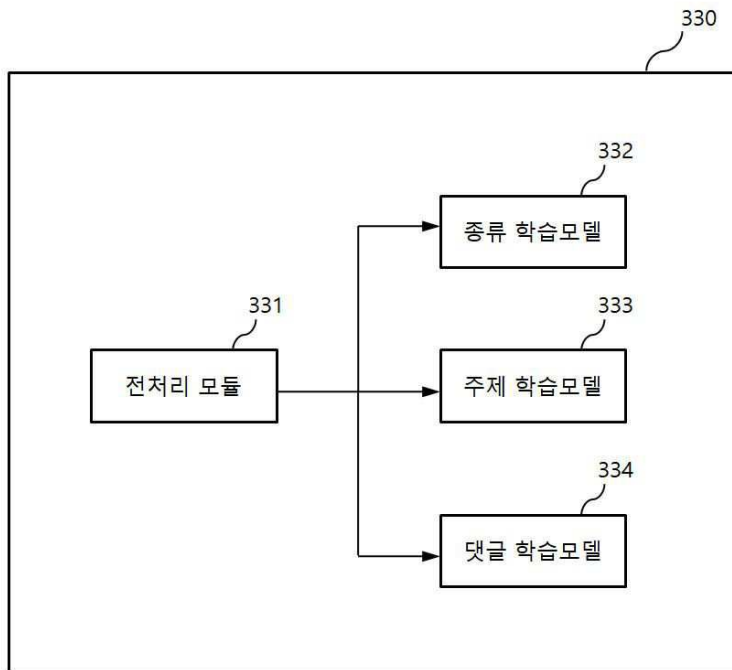
도면3



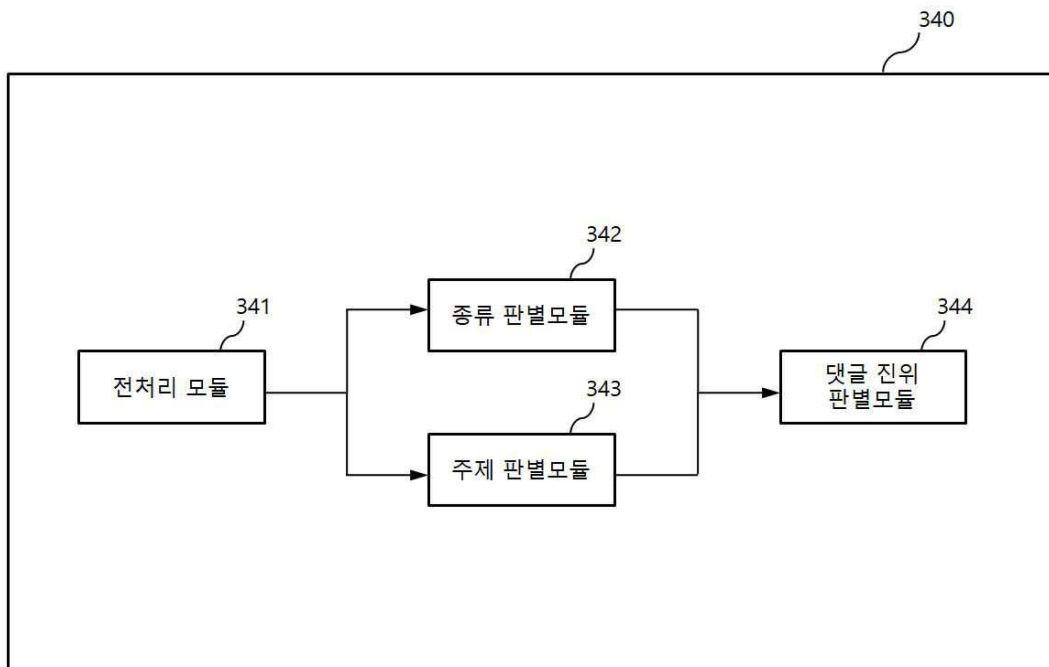
도면4



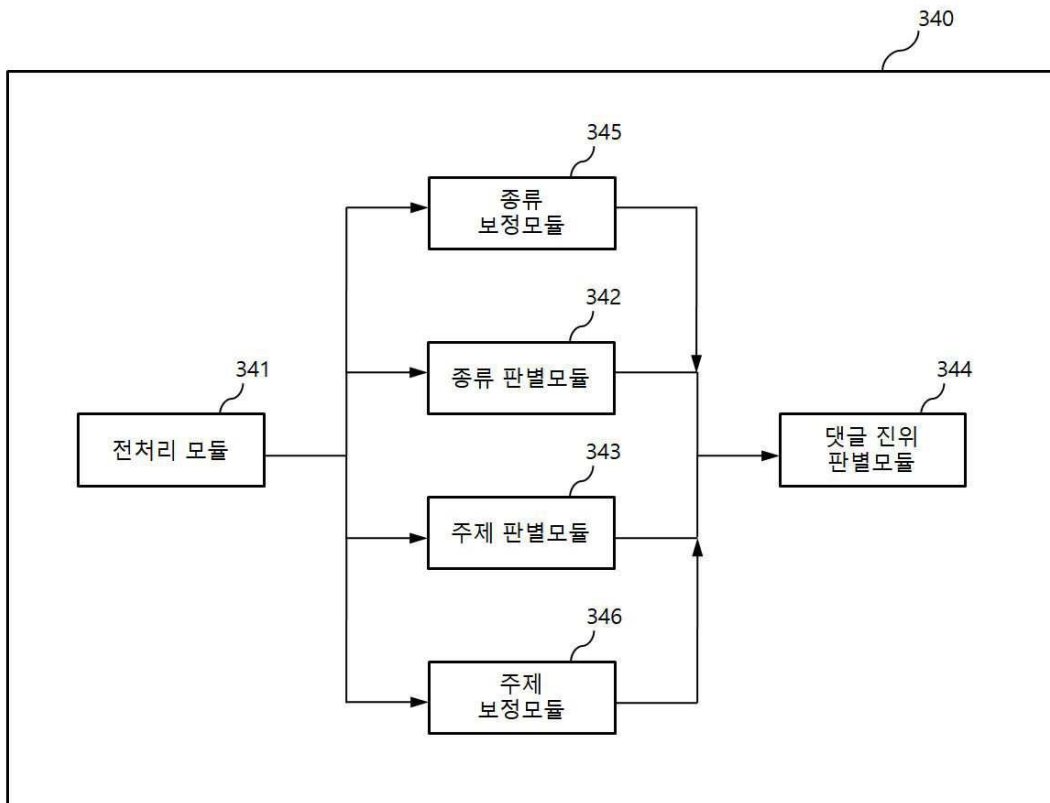
도면5



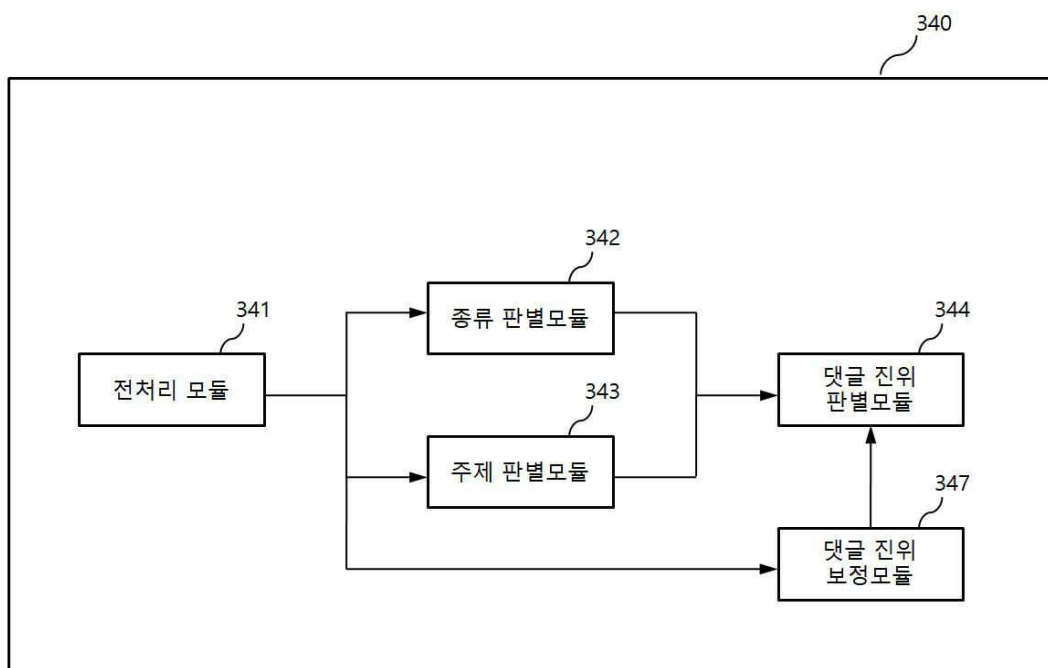
도면6



도면7



도면8



도면9

온라인 콘텐츠	
• 퍼블리셔 (publisher): 네이버	
• 제목 (title): 내일도 600명대 넘을 듯. 9일 오후 6시 현재 483명(종합)	
• 오너 (owner): News1	
• 내용 (contents): 신종 코로나바이러스 감염증(코로나19) 전국 확진자가 발생하고 있으며...	
• 댓글 (comments): (id, comment)	

도면10

종류	퍼블리셔	오너
뉴스	네이버	YTN, MBC, 중앙일보, News1
	유튜브	YTN, MBC, 중앙일보
연예	네이버	디스패치, OSEN, 스포츠동아
	유튜브	Mnet, SM entertainment, BTS
스포츠	네이버	인터풋볼, 스포츠동아
	유튜브	SPOTV, ESPN, LA Dodgers

도면11

주제	퍼블리셔	오너
스트레이트	네이버	YTN, MBC, 중앙일보
	유튜브	YTN, MBC, 중앙일보
오피니언	네이버	미디어오늘, 시사in, 일요신문
	유튜브	가로세로연구소, 다스퇴이다
홍보	네이버	디스패치, 엠스플뉴스
	유튜브	Mnet, SM entertainment

도면12

구분	스트레이트	오피니언	홍보
뉴스	N_{11}	N_{12}	N_{13}
연예	N_{21}	N_{22}	N_{23}
스포츠	N_{31}	N_{32}	N_{33}

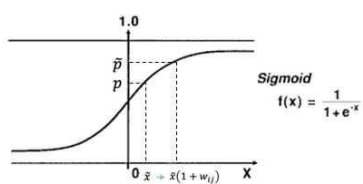
도면13

구분	스트레이트	오피니언	홍보
뉴스	w_{11}	w_{12}	w_{13}
연예	w_{21}	w_{22}	w_{23}
스포츠	w_{31}	w_{32}	w_{33}

도면14

요청 콘텐츠
• 퍼블리셔 (publisher): 네이버 • 제목 (title): 내일도 600명대 넘을 듯...9일 오후 6시 현재 483명(종합) • 오너 (owner): News1 • 내용 (contents): 신종 코로나바이러스 감염증(코로나19) 전국 확진자가... • 댓글 (comments): (id, comment)

도면15



도면16

