Tampereen yliopisto

Oona Saksman

TEHOKAS VALEUUTISTEN TUNNISTUS

Uutissisällön ja verkostotapahtumien tarkastelun automatisointi ihmisten apuna

Informaatioteknologian ja viestinnän tiedekunta
Kandidaatintutkielma
Maaliskuu 2023

TIIVISTELMÄ

Oona Saksman: Tehokas valeuutisten tunnistus - Uutissisällön ja verkostotapahtumien tarkastelun automatisointi ihmisten apuna
Kandidaatintutkielma
Tampereen yliopisto
Tietojenkäsittelytieteiden tutkinto-ohjelma
Maaliskuu 2023

Digitaalinen media on edistänyt maailmaa monin tavoin. Tiedon jakaminen ja löytäminen on helpompaa kuin koskaan, mutta tälle on varjopuolensa. Valeuutiset nähdään yhtenä suurimmista uhista demokratialle, journalismille ja taloudelle. Ne häiritsevät demokraattisia vaaleja, vahingoittavat ihmisten ja järjestöjen maineita ja heikentävät ihmisten luottamusta hallituksiin. Nämä uhat ovat inspiroineet tutkijoita kehittämään lukuisia systeemejä valeuutisten tunnistamiseen. Tässä tutkielmassa kartoitan kehitettyjä lähestymistapoja valeuutisten tunnistamiseen ja selvitän mitkä seikat ovat keskeisiä, kun prosessista halutaan tehdä mahdollisimman tehokas.

Tämän tutkielman menetelmä on kirjallisuuskatsaus. Käytin lähteinä uusia eli aikaisintaan 2018 julkaistuja vertaisarvioituja tieteellisiä artikkeleita ja tutkimuksia, yhtä päiväämätöntä Euroopan komission julkaisua lukuun ottamatta. Aiheesta oli helppo löytää kirjallisuutta, ja hakuja tehdessäni kävi selväksi, että tutkijat kehittävät uusia tapoja valeuutisten tunnistamiseen hyvin aktiivisesti.

Valeuutiset ovat monimutkainen ongelma ja faktantarkistus on monivaiheinen prosessi, jonka yksi vaihe on valeuutisen tunnistaminen. Automatisoidut tavat tunnistaa valeuutisia vaativat monen eri tekijän huomioon ottamista ja ihmisten suorittama tunnistus paljon tutkimustyötä. Mahdollisimman tehokkaat tunnistusalgoritmit käyttävät sekä uutisen kontekstia, että sen sisältöä tunnistamiseen. Sisällön tarkistelu voi keskittyä moneen eri elementtiin, esimerkiksi uutisen leipätekstiin tai siihen kuuluviin kuviin. Tässäkin useamman eri elementin tarkastelu on hyödyllistä. Kontekstilla tarkoitetaan uutisen valheellisuuden päättelyä hyödyntämällä sen sijaintia sosiaalisessa verkostossa. Esimerkiksi se kuka julkaisi ja jakoi uutista voi auttaa algoritmia identifioimaan valeuutisen.

Algoritmien soveltaminen tehtävissä, jotka vaativat suuren uutismäärän läpikäymistä on tarpeellista, mutta teknologia ei ole kehittynyt tarpeeksi pitkälle pärjätäkseen ilman ihmisten työpanosta. Teknologian tehtävä taistelussa valeuutisia vastaan on ainakin toistaiseksi toimia ihmisten työkaluna. Tämä on seikka joka vaatisi enemmän huomiota tutkijoilta, joiden työ on toistaiseksi keskittynyt pitkälti vain täysin autonomisoituihin tunnistusprosesseihin.

Avainsanat: valeuutiset, koneoppiminen, luonnollisen kielen käsittely

SISÄLLYSLUETTELO

1 Johdanto

Digitaalinen media on edistänyt yhteiskuntaa monella tavalla. Se on tehnyt tiedon jakamisesta helpompaa ja siten yhdistää meitä sosiaalisesti ja kasvattaa tuottavuutta. Ilman ongelmia tämä ei kuitenkaan ole tapahtunut. Informaation helpon jakamisen seurauksena valeuutisia luodaan enemmän kuin koskaan ennen. (Segura-Bedmar ja Alonso-Bartolome, 2022) Valeuutiset nähdään yhtenä isoimmista uhista demokratialle, journalismille ja taloudelle (Zhou ja muut, 2019). Ne aiheuttavat laajamittaista hämmennystä, joka synnyttää kaaosta ja kärsimystä monella saralla (Collins ja muut, 2021). Tämä kaaos häiritsee demokraattisia vaaleja, vahingoittaa ihmisten ja järjestöjen maineita (Choraś ja muut, 2021), heikentää ihmisten luottamusta hallituksiin ja huojuttaa osakemarkkinoita (Zhou ja Zafarani, 2020).

Ilmiö ei ole uusi, mutta aiheesta tulee jatkuvasti tärkeämpi. Sosiaalisen median kasvu antaa käyttäjille aivan uuden tason tilaisuuden kerätä voitot johtamalla ihmisiä harhaan. Tätä mahdollisuutta hyödynnetään poliittisten ja taloudellisten etujen vuoksi (Collins ja muut, 2021). Yksi tunnetuimmista esimerkeistä on joukko makedonialaisia teinejä, jotka rikastuivat USA:n 2016 presidentinvaalien aikaan klikkikalasteluartikkelien (eng. clickbait), eli perättömien ja räikeästi otsikoitujen uutistarinoiden avuin (Zhou ja muut, 2019; Collins ja muut, 2021). Valeuutisia tehdään monista eri aiheista. Poliittisten tarinoiden lisäksi esimerkiksi lääketieteellistä disinformaatiota levitetään paljon (Alaphilippe ja muut 2019).

Vastauksena väärän tiedon aiheuttamaan vaaraan on luotu monenlaisia teknologioita ja systeemejä, joiden avulla valeuutisia voidaan erotella muusta tietovirrasta. Euroopan unioni rahoittaa useampaa hanketta, joiden tarkoituksena on saada aikaan rehellisempää viestintää entistä tehokkaamman valeuutisten tunnistuksen avulla. GoodNews-hankkeessa hyödynnetään syväoppimisteknologiaa tämän tavoitteen saavuttamiseksi, kun taas Fandango-hankkeessa uutisdatan, medialähteiden, sosiaalisen median ja avoimen datan typologioita kootaan ja todennetaan. (Euroopan komissio, ei pvm.)

Tässä tutkielmassa selvitän, mitkä teknologiat ja strategiat ovat keskeisiä tehokkaassa valeuutisten tunnistamisessa. Keskityn erityisesti valeuutisten, eli valheellisten, haitallisten ja usein nopeasti leviävien uutisten, mahdollisimman tehokkaaseen tunnistamiseen, koska sille on tarvetta ison, jatkuvasti kasvavan tietovirran ja ongelman vakavuuden takia. Tämän tutkielman ulkopuolelle jää muun muassa valeuutisten sensurointiin liittyvät eettiset kysymykset, valeuutisten tunnistaminen eri kielillä ja se mitä valeuutisten tunnistussysteemejä tietyillä sosiaalisen median alustoilla tai uutissivustoilla on jo käytössä. Näiden päätösten pohjalta tutkimuskysymykseksi nousi “Mitä tehokkaaseen valeuutisten tunnistukseen kuuluu?”.

Tutkimusmenetelmä tälle tutkielmalle on kirjallisuuskatsaus. Lähteenä olen käyttänyt mahdollisimman tuoreita, eli aikaisintaan 2018 julkaistuja, vertaisarvioituja tieteellisiä julkaisuja tietojenkäsittelyn alalta, sekä yhtä päiväämätöntä Euroopan komission julkaisua. Hain lähteitä muun muassa seuraavilla hakusanoilla: ”fake news detection”, ”fake news detection survey”, ”automated fake news detection” ja ”multimodal fake news detection”. Hakuja on tehty Science Direct, ACM Digital Library, SpringerLink ja Andor -tietokannoista. Tuloksia oli paljon julkaisuvuosien ja julkaisutyypin rajaamisen jälkeenkin. Valitsin syntyneestä joukosta parhaimmat lähteet ensin otsikon ja sitten tiivistelmän perusteella. Osan lähteistä valitsin, koska ne olivat lähteenä toisessa lähteessäni. Osa lähteistä käsittelee laajemmin virheellisen tiedon leviämistä, mutta koska valeuutiset ovat virheellistä tietoa, niistä poimittu tieto pätee myös tämän tutkielman aiheeseen.

Luvussa 2 käyn läpi tutkielman kannalta keskeiset käsitteet. Luvussa 3 pureudun eri tapoihin, joilla valeuutisten tunnistustapoja on luokiteltu. Luku 4 käsittelee avoimia haasteita valeuutisten tunnistamisen alalla. Luku 5 vastaa tutkimuskysymykseen ja siihen miten se suhtautuu tässä kappaleessa esitettyyn tietoon. Pohdin myös millaista jatkotutkimusta aihe vaatisi ja tulokseen yleistettävyyttä sekä siihen liittyviä varauksia. Luku 6 on yhteenveto.

2 Keskeiset käsitteet

Tässä luvussa esittelen lyhyesti tutkielman kannalta keskeisiä käsitteitä.

Valeuutinen. Valeuutisella ei ole yhtä sovittua tai yleisesti hyväksyttyä määritelmää (Collins ja muut, 2021). Guo ja muut (2020) määrittelevät valeuutisen uutisartikkeliksi, jotka ovat tarkoituksella ja vahvistetusti valetta. Jotkut tutkijat määrittelevät termin laajemmin tarkoittaen sillä kaikenlaisia valheellisia julkaisuja, jotka käsittelevät julkisia henkilöitä tai järjestöjä, kun taas toiset rajaavat sen tarkoittamaan vain uutistoimiston julkaisemaa valheellista uutista (Zhou ja Zafarani, 2020). Tässä tutkielmassa käytän laajempaa määritelmää.

Valeuutisia on monentyyppisiä, kuten propaganda tai klikkikalastelu mutta niitä ei tule sekoittaa huhuihin, jotka ovat uutisia, joiden totuudenmukaisuutta ei ole vielä varmistettu (De Souza ja muut, 2020). Valeuutisia ei myöskään tule sekoittaa uutisiin, jotka eivät tue tietyn ryhmän kantaa tai etuja, vaikka julkisessa keskustelussa ja politiikassa termiä käytetäänkin joskus näin (Zhou ja Zafarani, 2020).

Uutisten ja valeuutisten lajitteluun on käytetty esimerkiksi seuraavanlaista skaalaa (Saquete ja muut, 2020):

- *Tosi*: Väite pitää paikkansa ja mitään merkittävää ei puutu.
- *Pääosin tosi*: Väite pitää paikkansa, mutta kaipaa selvennystä tai lisätietoa.

- *Puoliksi tosi*: Väite on osittain paikkaansa pitävä, mutta jättää mainitsematta tärkeitä yksityiskohtia tai asia on otettu pois kontekstista.
- *Pääosin vale*: Väite sisältää totuuden elementtejä, mutta ei ota huomioon kriittisiä faktoja, jotka antaisivat asiasta erilaisen kuvan.
- *Vale*: Väite ei pidä paikkaansa.

Valeuutisten tunnistamisen tehokkuus. Tehokkuudeksi määrittelen kyvyn tunnistaa valeuutisia tavalla, joka minimoi valeuutisten katselukerrat. Esimerkiksi Zhou ja muut (2020) painostavat kuinka tärkeää valeuutisiin on puuttua ennen kuin ne ovat ehtineet leviämään ja aiheuttamaan vahinkoa.

Valeuutisten leviäminen. Leviäminen tarkoittaa tämän tutkielman kontekstissa valeuutisen leviämistä useamman ihmisen näytöille. Tämä voi olla seurausta ihmisten tai bottien jakamisesta (Alaphilippe ja muut 2019) tai suosittelualgoritmien ehdotuksista (Collins ja muut 2021).

Botti. Botit ovat algoritmeja, jotka luovat sisältöä ja ovat vuorovaikutuksessa ihmisten tai toisten bottien kanssa sosiaalisessa mediassa. Niitä voi olla vaikea erottaa oikeista käyttäjistä. (Zhang ja Ghorbani, 2020). Kaikki botit eivät ole haitallisia, mutta tämän tutkielman kontekstissa boteilla tarkoitetaan valeuutisia tuottavia ja levittäviä botteja, joiden tarkoitus on johtaa ihmisiä harhaan. Botteja on käytetty jopa journalistien uhkailuun ja häiritsemiseen (Alaphilippe ja muut 2019).

Faktantarkistus. Faktantarkistushankkeiden tarkoitus on taistella esimerkiksi väärää tietoa ja propagandaa vastaan (Alaphilippe ja muut 2019). Saqueten ja muiden (2020) mukaan väärän tiedon, ja siten valeuutisten tunnistamiseen kuuluu neljä vaihetta, jotka ovat samat sekä automatisoidussa että journalistien tai tutkijoiden suorittamassa manuaalisessa faktantarkastuksessa. Ensimmäinen vaihe on **monitorointi**. Väitteen konteksti selvitetään. Vastataan siihen kuka sanoi tai julkaisi, ja missä, milloin ja kenelle. Seuraavaksi on **havaitseminen**, johon kuuluu useampi tehtävä. Tässä vaiheessa tunnistetaan mitkä väitteet on jo faktatarkistettu ja mitä ei ole. Väitteiden tärkeydestä tehdään päätöksiä ja käsitellään väitteet, jotka on ilmaistu eri tavoin, mutta tarkoittavat samaa asiaa. **Tunnistus**vaiheessa luokitellaan väitteet niiden totuudenmukaisuuden perusteella. Lopuksi, **Luominen ja julkaisu** -vaiheessa tulokset muotoillaan ja esitetään. Tämä tutkielma keskittyy pääasiassa tunnistusvaiheeseen, mutta sitä ympäröivät askeleet ovat tärkeitä, jotta tunnistus saadaan tehtyä tehokkaasti, ja jotta tuloksista on hyötyä. Näihin vaiheisiin ja siihen kuka suorittaa ne ja mitä teknologiaa hyödyntäen faktantarkastusorganisaatioissa pureudutaan syvemmin kohdassa 4.1.

3 Valeuutisten tunnistamistapojen luokittelu

Tässä luvussa käyn läpi eri lähestymistapoja valeuutisten tunnistamiseen. Aluksi (kohta 3.1), vertailen automatisoituja ja manuaalisia keinoja ja syvennyn faktantarkistusprosessiin. Seuraavaksi (kohta 3.2) käsittelen kolmessa eri tutkimusartikkelissa käytettyjä ryhmittelyitä. Tarkastelen miten ne eroavat toisistaan, miten niitä on rajattu ja miten ne täydentävät toisiaan. Valitsin kyseiset tutkimuspaperit niiden erilaisten lähestymistapojen vuoksi. Kohta sisältää myös diagrammeja visualisoimaan mitkä käsitteet liittyvät toisiinsa ja miten. Diagrammit on otsikoitu artikkelin otsikoinnin perusteella tai kuvaus luokittelusta on poimittu tekstistä. Tarkoituksena on luoda kokonaiskuva erilaisista lähestymistavoista valeuutisten tunnistamiseen.

3.1 Automatisoidut, osittain automatisoidut ja manuaaliset lähestymistavat

Valeuutisia identifioidaan automatisoiduin, osittain automatisoiduin tai manuaalisin keinoin (Collins ja muut, 2021). Tämä tarkoittaa, että tehtävän voi suorittaa siihen tehdyllä teknologialla, osittain sen avulla tai täysin ihmisten voimin. Pelkällä manuaalisella faktojen tarkistamisella ei pystytä kunnolla käsittelemään valtavaa määrää sosiaalisessa mediassa syntyvää uutta tietoa (Zhou ja muut, 2020). Ekspertti faktantarkastajat eivät aina pysty vastaamaan valeuutisten kasvavaan määrään ja käyttäjien valjastaminen valeuutisten tunnistukseen tuo mukanaan asiantuntemuksen puutteesta johtuvat ongelmat. Esimerkiksi huonosti tunnetut uutissivustot saatetaan virheellisesti luokitella eiluotetuiksi. (Collins ja muut 2021). Faktantarkastamisella on myös hintansa tarkastajien mielenterveydelle. He ovat usein käyttäjien ja salaliittoteoriitikkojen uhkausten kohteina ja sisältö jota heidän pitää tutkia voi olla raskasta. (Juneja ja Mitra, 2022)

Puhtaasti automatisoiduilla tunnistamisella on myös ongelmansa. Ensinnäkin, vaikka algoritmi pystyisi tekemään tunnistamista 98 % tarkkuudella, lajiteltavien uutisten määrä on niin valtava, että iso määrä niistä päätyisi silti väärään kategoriaan. Algoritmeja käytetään usein myös yhdistelmänä, jolloin toinen algoritmi voi kasvattaa virhettä ensimmäisen tuloksien pohjalta. Toiseksi, virheet tunnistamisessa voivat tulla kalliiksi. Valeuutinen joka pääsee systeemistä läpi voi kylvää paljon haittaa, mutta oikean uutisen poistaminen on mahdollisesti vielä vahingollisempaa. Näiden seikkojen perusteella paras vaihtoehto on toteuttaa sosio-teknologisia systeemejä, joissa tekoälyalgoritmit avustavat ihmisten faktantarkistusta. (Alaphilippe ja muut, 2019) Tämä onkin tapa, jota monet johtavat faktantarkistusjärjestöt käyttävät (Juneja ja Mitra, 2022).

Juneja ja Mitra (2022) haastattelivat 26 osallistujaa, jotka kuuluvat 16 eri faktantarkistustiimiin ja -organisaatioon. He tunnistivat kuusi sidosryhmää, joilla on omat roolinsa faktantarkistusprosessissa. Tutkimuksen tarkoituksena oli muun muassa saada kuva prosessin infrastruktuurista eli kartoittaa ihmisten ja teknologian tehtäviä ja yhteistyötä. Infrastruktuureja on kaksi, yksi suorittamaan lyhyen aikavälin faktantarkistusta ja toinen pitkän aikavälin.

Lyhyen aikavälin faktantarkistus tarkoittaa yksittäisten valheellisten väitösten kumoamista. Tätä käsiteltiin jo luvussa 2. Perehdyn nyt tarkemmin siihen kuka tai mikä teknologia on vastuussa mistäkin tehtävästä. Ensin faktantarkistajat tarkkailevat nettisivuja virheellisen tiedon varalta. He identifioivat kumottavat väitteet ja tuovat editorille kumoamissuunnitelman. Kun suunnitelma saa editorin hyväksynnän faktantarkistajat arkistoivat sisällön ja etsivät väitteen lähteen. Väitettä tutkitaan nettityökalujen, eksperttien neuvojen ja julkisten virallisten tietojen (esimerkiksi tilastot) avulla. Tämän työn tuloksena syntyy raportti, joka sisältää käytetyt lähteet ja väitös luokitellaan sen totuudenmukaisuuden perusteella. Raportin oikeellisuus tarkastetaan perusteellisesti. Valheellisen väitteen julkaisijaan otetaan yhteyttä korjauksesta. Lopuksi sosiaalisen median vuorovaikutusvastaavat julkaisevat faktankorjauksen järjestön nettisivuilla ja käyttävät monenlaisia strategioita kasvattaakseen yleisön vuorovaikutusta julkaisun kanssa. Suurin osa järjestöistä tekee myös pitkän aikavälin faktantarkistusta, joka tarkoittaa pitkäaikaista kampanjaa, johon kuuluu tutkimus-, vaikutus- ja tiedotustyötä. (Juneja ja Mitra, 2022)

Tutkimuksessa selvisi myös, että faktantarkistajat ovat skeptisiä puhtaasti automatisoiduista ratkaisuista ja kokevat, että tutkijoiden työ harvoin vaikuttaa heidän tekemiseensä. Seuraavassa kohdassa (3.2) on hyvä pitää mielessä kuilu tutkijoiden suorittaman työn ja faktantarkastajien arjen välillä.

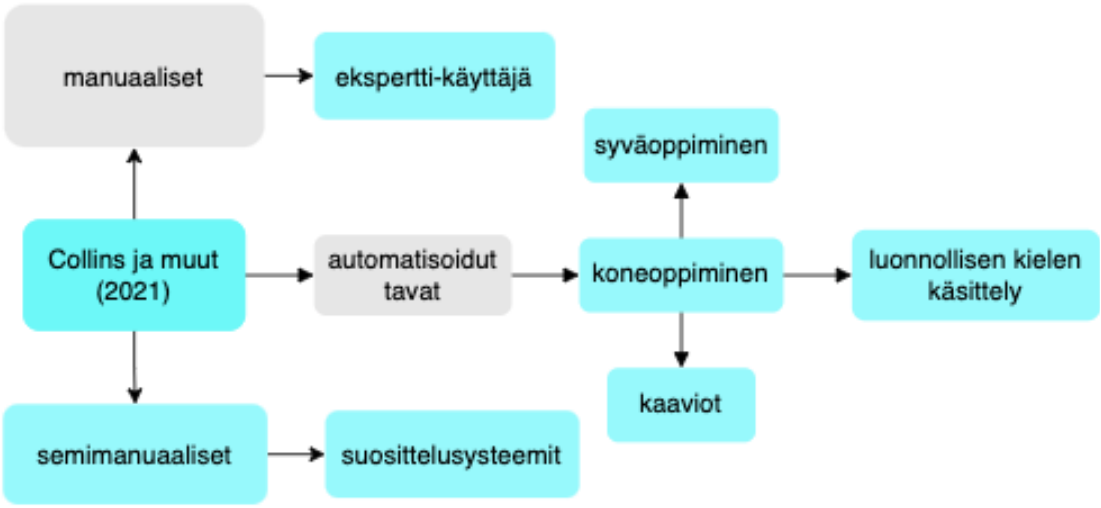
3.2 Lähestymistapojen luokittelua

Collins ja muut (2021) jaottelevat automatisoidut tavat tunnistaa valeuutisia koneoppimiseen, syväoppimiseen, kaavioihin ja luonnollisen kielen käsittelyyn pohjautuviin malleihin. Suosittelusysteemit on semimanuaalinen lähestymistapa, kun taas eksperttikäyttäjä on esimerkki täysin manuaalisesta tunnistusmallista (kuva 1). Collinsin ja muiden (2021) tutkimusartikkelin luokittelussa automatisoidut kategoriat ovat tunnistamiseen käytettävien teknologioiden yläkäsitteitä. Lisäsin automatisoidut ja manuaaliset tavat selventämään kaavion rakennetta.

Semimanuaaliset tavat ovat hybriditekniikoita, koska niissä on käytössä joku tietokoneen suorittama tehtävä ja joku ihmisen suorittama tehtävä. Suosittelusysteemit ovat keskeisessä roolissa uutisten toimittamisessa ihmisille ja niitä pyörittävät algoritmit voivat aiheuttaa ongelmia työntämällä ihmisten näytöille valeuutisia. Artikkelin mukaan suosittelusysteemitaktiikka yrittää varmistaa joitain uutisia tosiksi ja suosittelee niitä. Tätä ei voi sanoa valeuutisten tunnistamiseksi, mutta artikkelissa on kuitenkin annettu seuraavan kappaleen esimerkki, jossa tehdään tunnistusta. (Collins ja muut, 2021)

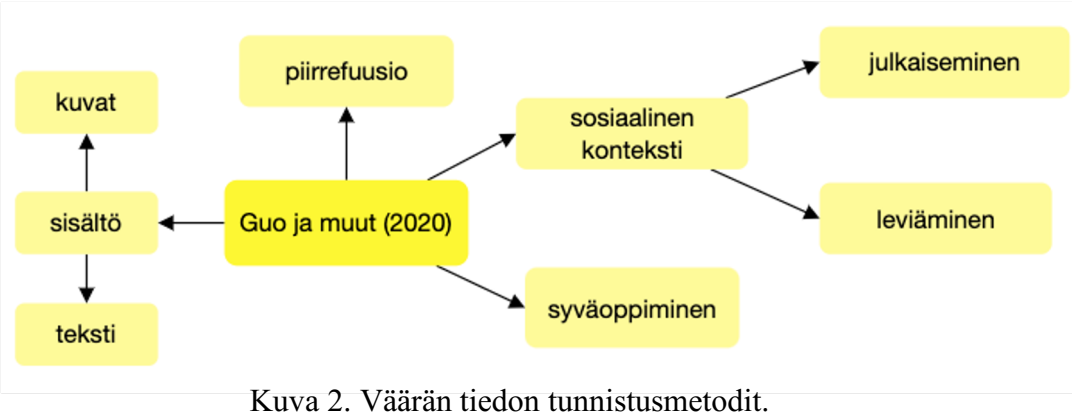
Vo ja Lee (2018) ottivat yhteyttä internetissä aktiivisiin käyttäjiin, joilla oli tapana lisätä korjaavaa tietoa sisältäviä linkkejä valheellisiin julkaisuihin. He kehittivät näille korjaajille personalisoidun suosittelusysteemin. Suosittelusysteemi ehdottaa faktantar-

kistuslinkkejä kannustaakseen käyttäjiä korjaamaan väärää tietoa ja tekemään sitä tehokkaammin.



Kuva 1. Valeuutisten tunnistusmalleja.

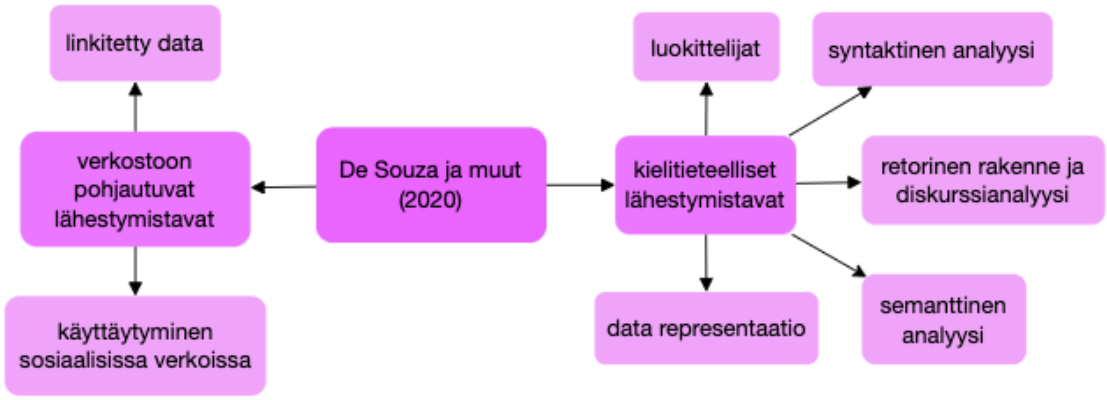
Guo ja muut (2020) tekevät jaon sisältöön, sosiaaliseen kontekstiin ja syväoppimiseen pohjautuviin metodeihin sekä piirrefuusioon (kuva 2). Viimeisellä tarkoitetaan sisällön ja sosiaalisen verkoston yhdistelmää. Sisältö, sosiaalinen konteksti ja piirrefuusio kuvaavat sitä, mitä piirrettä hyödyntämällä tunnistamista tehdään. Syväoppiminen tarkoittaa käytettyä teknologiaa. Tämä luokittelu koskee laajemmin väärän tiedon tunnistamista, mutta pätee myös valeuutisiin.



Kuva 2. Väärän tiedon tunnistusmetodit.

De Souza, ja muut (2020) tekivät systemaattisen kartoituksen olemassa olevista tavoista tunnistaa valeuutisia automatisoidusti. He analysoivat 85 tutkimusartikkelia ja selvittivät, että lähestymistavat jakautuvat seuraaviin: luokittelija, retorinen rakenne ja diskurssianalyysi, semanttinen analyysi, syntaktinen analyysi, datarepresentaatio, käyttäytyminen sosiaalisissa verkostoissa, ja linkitetty data (kuva 3). Kaksi viimeistä kategoriaa

kuuluvat verkostoon pohjautuviin lähestymistapoihin ja muut kielitieteellisiin lähestymistapoihin. Tämä jako on tarkin, mutta samalla suppein. Se sulkee ulkopuolelleen esimerkiksi kuvien avulla tehdyn tunnistuksen tai hybriditekniikat. Tutkimuksessa (De Souza, ja muut, 2020) ilmeni, että suosituimmat metodit ovat luokittelijat (käytetty 59 artikkelissa) ja käytös sosiaalisessa mediassa (käytetty 19 artikkelissa).



Kuva 3. Lähestymistavat valeuutisten tunnistamiseen.

Jaottelua tehdään siis hyvin eri näkökulmista ja vain yhdenkin jaottelusysteemin puitteissa teknologiat kuuluvat monesti useampaan kategoriaan. Osa kategorioista kuvaa sitä, minkä piirteen perusteella uutisia lajitellaan, esimerkiksi miten valeuutista jaetaan ja levitetään, kun taas osa kuvaa teknologiaa joka on käytetty, esimerkiksi luonnollisen kielen käsittelyä.

Kuva 4 havainnollistaa kuinka luokittelut menevät osittain päällekkäin ja kuinka ne täydentävät toisiaan. Lähteet on värikoodattu, jotta on helpompi käsittää mistä lähteestä tai lähteistä kategoria on peräisin. Useammasta lähteestä peräisin olevat kategoriat seuraavat logiikkaa, jonka voi tarkistaa oheisesta kuviosta. Jos kategoria mainitaan useammassa lähteessä, sen taustavärinä on lähteiden värien yhdistelmä. Esimerkiksi syväoppiminen on vihreällä taustalla, koska se on mainittu keltaisessa kuvassa (Guo ja muut, 2020) ja syaaninsinisessä kuvassa (Collins ja muut, 2021). Värejä ei ole kuitenkaan pakko hyödyntää, sillä kaikki erillisissä kuvissa mainitut nimet ovat myös tässä. Esimerkiksi luonnollisen kielen käsittelyä voi tehdä myös syväoppimista hyödyntäen ja monia kategorioita pystyy varmasti jakamaan pienempiin osiin, joten diagrammi ei selitä kategorioiden välisiä yhteyksiä kokonaisvaltaisesti. Seuraavissa alakohdissa perehdyn kuvan 4 keskeisimpiin kategorioihin.

3.2.1 Sisältö

Sisältökategoria (kuva 2 ja kuva 4) kattaa lähestymistavat, jotka tekevät tunnistuksen uutisen sisällön, eli esimerkiksi tekstin, kuvien, äänten tai videoiden perusteella (Guo ja muut, 2020). Sisältöön perustuvan tunnistamisen hyödyntäminen vaatii syvää tiedonlouhintaa, eli suurien uutismäärien monitorointia ja lajittelua (Zhou ja muut, 2020).

Multimodaalinen lähestymistapa tarkoittaa kahden tai useamman eri sisältötyypin perusteella tehtyä tunnistusta. Multimodaalisella valeuutisten tunnistamisella tekstin ja kuvien avulla on saatu parempia tuloksia kuin vain yhteen sisältötyyppiin pohjautuvista lähestymistavoista. (Segura-Bedmar ja Alonso-Bartolome, 2022)

Kuva 4. Lähestymistavat valeuutisten tunnistamiseen.

Capuano ja muut (2023) tekivät systemaattisen katsauksen kone- ja syväoppimismalleihin, jotka tunnistavat valeuutisia uutissisällön perusteella. Tutkijat arvioivat kuinka monessa tutkimuksessa kyseistä mallia oli käytetty, kuinka monella tietokannalla mallia oli testattu ja tunnistustarkkuutta. Gradienttivahvistus (eng. gradient boosting), eXtreme gradienttivahvistus, monikerroksinen perseptroniverkko (engl. multilayer perceptron) ja naiivi Bayesin luokitin pärjäsivät tutkimuksessa parhaiten.

Gradienttivahvistus tekee tunnistusta yhdistämällä ennustuksia heikoilta ennustusmalleilta, yleensä päätöspuilta. Ennustuksia tehdään peräkkäin siten, että seuraava yrit-

tää korjata edeltävän tekemiä virheitä. EXtreme gradienttivahvistus on yksi nopeimmista gradienttivahvistusalgoritmeista. Monikerroksinen perseptroniverkko koostuu syötekerroksesta, piilotetusta kerroksesta ja tulostekerroksesta. Piilotettu kerros ja tulostekerros koostuvat keinotekoisista neuroneista, jotka suorittavat epälineaarisia aktivointifunktioita. Naiivi Bayesin luokitin soveltaa Bayesin lauseketta. Algoritmin ”naiivius” tulee oletuksesta, että havainnot ovat toisistaan ehdollisesti riippumattomia. (Capuano ja muut, 2023)

3.2.2 Luonnollisen kielen käsittely

De Souza ja muut (2020) kuvailevat kielitieteellistä lähestymistapaa tavaksi tunnistaa valeuutisia hyödyntämällä tekstistä löytyviä pieniä vihjeitä, jotka erottavat valeuutisen muista uutisista. Collins ja muut (2021) taas kuvaavat luonnollisen kielen käsittelyä tekniikaksi, joka hyödyntää monenlaisia koneoppimistaktiikoita tunnistaakseen sanallisia tai sanastollisia vihjeitä, jotka tuovat esiin kielellisiä eroja valehtelun ja totuuden kertomisen välillä. Guo ja muut (2020) viittaavat aikaisempaan tutkimukseen, jonka mukaan sisältöön pohjautuvat metodit perustuvat pääasiassa sanastollisiin piirteisiin, syntaktisiin piirteisiin ja aihepiirteisiin. Sisältöön pohjautuvat metodit kattavat kuvat ja tekstit, mutta tämä kuvaus ei päde visuaaliseen mediaan. Kaikissa luokitteluissa on siis esitetty sama kategoria eri nimillä. Jatkossa käytän luonnollisen kielen käsittely -termiä.

Kuten edellisessä kappaleessa selvisi, luonnollisen kielen käsittely hyödyntää monia erilaisia uutisten kielessä esiintyviä piirteitä selvittääkseen ovatko ne valheellisia. De Oliveira ja muut (2021) esittävät, että nämä piirteet voidaan kategorisoida seuraavasti:

- *Määrä*: Esimerkiksi kuinka monta merkkiä, sanaa, lausetta tai verbiä uutinen sisältää.
- *Muodollisuus:* Kuinka monta kirjoitusvirhettä tekstissä on suhteessa sen pituuteen. Zhou ja muut (2020) lisäävät tähän kategoriaan myös kirosanat, internetkielen (”lol”, ”btw”), suostumussanat (”OK”), epäsujuvuussanat (”öö”, ”hmm”) ja täyttösanat (”niinku”).
- *Monimutkaisuus*: Kuinka monta merkkiä sanassa, sanaa lauseessa, lauseketta lauseessa tai välimerkkiä lauseessa.
- *Epävarmuus*: Piirteet, jotka ilmaisevat epävarmuutta. Näihin kuuluu muun muassa modaaliverbit, termit jotka viittaavat varmuuteen, termit jotka viittaavat yleistämiseen ja kysymysmerkkien määrä.
- *Välittömyys*: Passiivinen ääni ja pronominit yksikön ensimmäisessä tai monikon ensimmäisessä, toisessa ja kolmannessa persoonassa.
- *Moninaisuus*: Esimerkiksi kuinka monta prosenttia sanoista on uniikkeja.
- *Tunteet*: Positiivisten ja negatiivisten sanojen osamäärä, huutomerkkien määrä ja humoristinen tai sarkastinen sisältö.

Valeuutiset ovat useammin kognitiivisesti yksinkertaisempia ja negatiivisempia. Ne sisältävät vähemmän asiaankuuluvia sanoja ja enemmän sosiaalisia sanoja, superlatiiveja, epävarmuutta ilmaisevia sanoja, itseviittauksia, kieltäviä toteamuksia, valitusta ja yleistystä. (De Oliveira ja muut, 2021)

3.2.3 Verkostoon pohjautuvat lähestymistavat

Sosiaalisessa mediassa samoja piirteitä jakavat käyttäjät linkittyvät toisiinsa. Kaavioita hyödyntämällä väärää tietoa levittävät yksilöt ja siten heidän julkaisemat valeuutiset voidaan löytää toisista valeuutisista tai valeuutisten levittäjistä johtamalla. (Collins ja muut, 2021). De Souza ja muut (2020) eivät kerro verkostoon pohjautuvista lähestymistavoista paljon, mutta nimen ja alaluokkien perusteella voidaan päätellä niiden tarkoittavan myös toisiinsa linkitettyjen entiteettien soveltamista valeuutisten tunnistamiseksi. He sanovat tämän tavan olevan käytössä kasvavissa määrin, koska valeuutiset leviävät sosiaalisissa verkostoissa. Käytän tässä tutkielmassa termiä verkosto puhuttaessa tästä tekniikasta.

Twitterin verkoston tutkimisen avulla on huomattu, että valeuutisen aiheuttaman keskustelun on tapana nousta ja laskea useamman kerran säännöllisin väliajoin, kun taas totuudenmukainen uutinen aiheuttaa vain yhden selvän huipun keskustelussa. Verkostotekniikkaa hyödyntämällä valeuutisten on myös huomattu leviävän tosia uutisia nopeammin, pidemmälle ja laajemmin, erityisesti kun kyseessä on poliittinen uutinen. (Oliveira ja muut, 2021) Sitä, miksi pelkän verkoston perusteella tunnistaminen ei ole suositeltavaa käsitellään seuraavassa luvussa.

3.2.4 Hybriditekniikat

Collins ja muut (2021) mainitsevat semimanuaalisen tekniikan, kun taas Guo ja muut (2020) puhuvat piirrefuusiosta. Nämä tarkoittavat tunnistamistapojen yhdistämistä. Guon ja muiden (2020) piirrefuusio viittaa sekä sisältöön, että sosiaaliseen kontekstiin pohjautuvan tekniikan avulla tunnistamista. Collinsin ja muiden (2021) semimanuaalinen lähestymistapa tarkoittaa nimensä mukaisesti sekä tietokoneen että ihmisten käyttämistä. Useamman eri tekniikan hyödyntäminen on keskeistä valeuutisten epämääräisen ja monimutkaisen luonteen vuoksi (Collins ja muut, 2021).

4 Haasteet

Tämä luku käsittelee valeuutisten tunnistamiseen liittyviä avoimia haasteita, joiden huomiointi parantaisi tunnistuksen tehokkuutta. Valeuutisen elinkaari koostuu sen luomisesta, julkaisusta ja leviämisestä. Yksi keskeisimmistä haasteista liittyen valeuutisten tunnistamiseen on **varhainen tunnistus**, joka tapahtuu ennen kuin uutinen on levinnyt ja saanut vahinkoa aikaan. Jotta tunnistus voi tapahtua ennen leviämistä on käytettävä sisällön perusteella tunnistavaa lähestymistapaa pelkän verkoston avulla tunnistamisen

sijasta. (Zhou ja muut, 2020) Vaikka jotkin rajattua datasettiä tarkkailevat algoritmit toimivat lähes reaaliajassa, kaiken julkaistun tiedon tarkkailu ja prosessointi esimerkiksi Twitterissä olisi valtava haaste (Alaphilippe ja muut, 2019).

Selittävä tunnistus (eng. explanatory false information detection) on yksi uudemmista haasteista valeuutisten tunnistuksen alalla (Guo ja muut, 2020). Ymmärrettävän selityksen luominen tekoälyn päätelmistä tulee yleensä sitä vaikeammaksi, mitä monimutkaisempi tekoälyalgoritmi on käytössä. Suurin osa valeuutisten tunnistusteknologioista ei tuota mitään selitystä tekemästään jaottelusta. Syyn antaminen voisi olla hyödyllistä, koska käyttäjä ei yleensä luota malliin, jonka toimintaa hän ei ymmärrä liiallisen monimutkaisuuden tai piilotetun prosessin takia (Mishima ja Yamana, 2022). Myös ammatilliset faktantarkistajat kaipaavat selittäviä työkaluja (Juneja ja Mitra, 2022). Selityksen voi antaa siitä, miten tunnistusmetodi toimii tai prosessin tuloksista, eli tulokset esitettäisiin visualisoimalla päätöksentekoprosessia tai analysoimalla faktoja. (Zhou ja muut, 2020).

Kun tavat tunnistaa valeuutisia kehittyvät, kehittyvät myös valeuutiset. Valeuutisista halutaan tehdä uskottavampia ja algoritmien kehittyessä levitystä tekevät botitkin tulevat mahdollisesti käyttäytymään tavalla jota on vaikeampi erottaa ihmisistä. (Alaphilippe ja muut, 2019) **Itseään kehittävä tunnistusprosessi** (eng. adaptive continuous learning model) on siis kaivattu. Sekä Choraś ja muut (2021), että Capuano ja muut (2023) tuovat esille, että valeuutisten muuttuva luonne huomioidaan vain harvoissa tutkimuspapereissa.

Koska valeuutisten levittämisen takana on usein poliittinen motiivi, valeuutisia ei ole hyödyllistä katsoa vain erillisinä sattumina. Valheellisen sisällön kampanjoita testataan usein vähemmän näkyvillä alustoilla, kuten Discordissa tai 4chanissä. Tällaisilla alustoilla syntyvien **väärän tiedon verkostojen ymmärtäminen** voi auttaa estämään valheiden leviämisen valtavirtaisemmille sivustoille. (Alaphilippe ja muut, 2019)

5 Keskustelu

Tutkimuskysymykseen ei ole yksinkertaista vastausta. Siitä, että automatisoidut tavat eivät pärjää vielä itsekseen löytyi konsensus, mutta eri lähteet painottavat eri lähestymistapoja ja mikään teknologia ei ole selvästi suosituin. Monissa lähteissä kuitenkin puollettiin lähestymistapojen yhdistämistä parhaan tuloksen saamiseksi. Seuraavat seikat nousivat tärkeiksi laadukkaassa valeuutisten tunnistamisessa: tunnistus hyvällä tarkkuudella, isojen datamäärien nopea prosessointi ja valeuutisten tunnistus ja niihin puuttuminen ennen kuin ne ovat ehtineet leviämään ja aiheuttamaan vahinkoa.

Jo johdannossa selvisi, että tutkijoiden aktiivinen työ on tuottanut monenlaisia automatisoituja lähestymistapoja valeuutisten tunnistamiseksi. Tätä havainnollistin näiden lähestymistapojen luokittelulla ja termien avaamisella. Tutkielma nosti kuitenkin myös

esille, kuinka aiheesta tehtävä tutkimus vaatisi uudelleensuuntaamista. Junejan ja Mitran (2022) tutkimuksessa ilmi tullut kuilu ammatillisten faktantarkistajien ja akateemikkojen välillä näkyi selkeästi kirjallisuuskatsausta suorittaessa. Keskittyminen on puhtaasti automatisoiduissa keinoissa eikä niinkään faktantarkastajien työn tukemisessa.

Uskon, että kohdan 3.2 kuvissa esiintyneiden termien standardisoinnista alalla olisi hyötyä. Vaikka termit ovat kuvaavia, vaikuttaa siltä, että tutkijat valitsevat ne mielivaltaisesti, joka tekee aiheen tutkimisesta työläämpää kuin sen tarvitsee olla. Diagrammini (kuva 4) kattoi vain kolmen eri tutkimuspaperin tulokset. Näistä tutkimuspapereista vain yksi sisällytti täysin tai osittain manuaaliset keinot. Mielestäni laajempi selvitys valeuutisten tunnistamiseen kehitettyjen lähestymistapojen luokittelusta ja nimeämiskäytännöistä olisi resurssina hyödyllinen ihmisille, jotka haluavat perehtyä alan tutkimukseen.

Koska valeuutisia koskevia lähteitä on niin valtava määrä ja kanditutkielma laajuudeltaan suppea, on mahdollista, että tässä tutkielmassa on aukkoja. Käsittelin tarkoituksella enemmän automatisoituja tapoja pysyäkseni tietojenkäsittelyn aihealueella.

Tekniset lähestymistavat tai ihmisten suorittama faktantarkistus netissä eivät välttämättä ole tehokkaimpia tapoja vastata valeuutisongelmaan, koska ne eivät korjaa ongelman ydintä. Valeuutiset ovat poliittisten ja taloudellisten motiivien synnyttämiä ja olemukseltaan muuttuvia, eikä niiden tunnistaminen ja faktoilla korjaaminen ole kestävä ratkaisu. Tutkielma vastasi kysymykseen ”Mitä tehokkaaseen valeuutisten tunnistamiseen kuuluu?”, mutta kysymys tehokkaimmasta tavasta estää valeuutisia tai niiden aiheuttamaa vahinkoa on auki. Tämä tutkielma, kuten valtaosa alan tutkimuksesta, on keskittynyt pääasiassa lyhyen aikavälin faktantarkistukseen ja sen vaiheisiin, mutta pitkän aikavälin faktantarkistustyötä ei pidä vähätellä.

6 Yhteenveto

Tutkielmassa on selvinnyt, että valeuutiset ovat monimutkainen ongelma, jolle ei ole yksinkertaista ratkaisua. Valeuutisten tunnistaminen on vain yksi vaihe lyhyen aikavälin faktantarkistusprosessissa, mutta mahdollisesti myös keskeisin. Vaikka tutkimusta automatisoiduista tavoista tunnistaa valeuutisia tehdään paljon, tehtävä on silti pääosin ihmisten vastuulla. Tällä hetkellä teknologian tärkein tehtävä on avustaa ihmisiä faktantarkistuksessa. Toistaiseksi uutisten valheellisuuden päättely automatisoiduin keinoin perustuu tekijöihin, jotka eivät itsessään tee uutisesta valheellista, kuten sen lauseiden monimutkaisuuteen, kuvamateriaaliin tai uutisen julkaisijaan.

Vaikka valeuutisia ei pystytä tunnistamaan algoritmein käytännössä hyödyllisellä tavalla ja niiden torjumis- ja oikaisupyrkimykset ovat vielä kovin puutteellisia, ei työ ole ollut turhaa. Valeuutisten luonteesta on saatu paljon selville ja tutkijoiden tähän asti ottamat askeleet ovat mahdollisesti tarpeellisia päästäksemme joku päivä tilanteeseen, jossa tiedosta internetissä ei tarvitse olla ihan yhtä skeptinen.

Lähdeluettelo

Capuano, N., Fenza G., Loia V., &, Nota F. D *(2023). Content-based fake news netection with machine and deep learning: A systematic review.* Neurocomputing, 530, 91-103, 0925-2312. https://doi.org/10.1016/j.neucom.2023.02.005

Chorąś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., Urda, D., & Woźniak, M. *(2021). Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study.* Applied Soft Computing, 101, 107050–. https://doi.org/10.1016/j.asoc.2020.107050

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. *(2021). Trends in combating fake news on social media - A survey.* Journal of Information and Telecommunication (Print), 5(2), 247–266. https://doi.org/10.1080/24751839.2020.1847379

de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., & Mattos, D. M. F. *(2021). Identifying fake news on social networks based on natural language processing: Trends and challenges.* Information, 12(1), 38–. https://doi.org/10.3390/info12010038

de Souza, J. V., Gomes Jr, J., Souza Filho, F. M. de, Oliveira Julio, A. M. & de Souza, J. F. *(2020). A systematic mapping on automatic classification of fake news in social media.* Social Network Analysis and Mining, 10. https://doi.org/10.1007/s13278-020-00659-2

Euroopan komissio. EU-rahoitteisia hankkeita disinformaation torjunnan alalla. https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/funded-projects-fight-against-disinformation_fi.

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. *(2020). The future of false information detection on social media: New perspectives and trends.* ACM Computing Surveys, 53(4), 1–36. https://doi.org/10.1145/3393880

Hanot, C., Bontcheva, K., Alaphilippe, A., & Gizikis, A. *(2019)*. Automated tackling of di*sinformation: major challenges ahead.* Euroopan Parlamentti. https://data.europa.eu/doi/10.2861/368879

Juneja, P., & Mitra, T. *(2022). Human and technological infrastructures of fact-checking. Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–36. https://doi.org/10.1145/3555143

Mishima, K., & Yamana, H. *(2022). A survey on explainable fake news detection.* IEICE Transactions on Information and Systems, E105.D(7), 1249–1257. https://doi.org/10.1587/transinf.2021EDR0003

Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., & Palomar, M. *(2020). Fighting post-truth using natural language processing: A review and open challenges.* Expert Systems with Applications, 141, 112943. https://doi.org/10.1016/j.eswa.2019.112943

Segura-Bedmar, I., & Alonso-Bartolome, S. *(2022). Multimodal fake news detection. information* (Basel), 13(6), 284–. https://doi.org/10.3390/info13060284

Vo, N., & Lee, K. *(2018). The rise of guardians: Fact-checking URL recommendation to combat fake news.* The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 275–284. https://doi.org/10.1145/3209978.3210037

Zhang, X., & Ghorbani, A. A. *(2020). An overview of online fake news: Characterization, detection, and discussion.* Information Processing & Management, 57(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhou, X., Jain, A., Phoha, V., & Zafarani, R. *(2020). Fake News Early Detection: A Theory-driven Model.* Digital Threats (Print), 1(2), 1–25. https://doi.org/10.1145/3377478

Zhou, X., & Zafarani, R. *(2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities.* ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046

Zhou, X., Zafarani, R., Shu, K., & Liu, H. *(2019). Fake News: Fundamental Theories, Detection Strategies and Challenges.* Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 836–837. https://doi.org/10.1145/3289600.3291382