Tampereen yliopisto

Kiira Rintamäki

MATEMAATTISET MALLIT INFORMAATION LEVIÄMISEN ENNUSTAMISESSA SOSIAALISESSA MEDIASSA

Kandidaatintyö
Tekniikan ja luonnontieteiden tiedekunta
Huhtikuu 2021

TIIVISTELMÄ

Kiira Rintamäki: Matemaattiset mallit informaation leviämisen ennustamisessa sosiaalisessa mediassa
Kandidaatintyö
Tampereen yliopisto
Matematiikka, Tekniikka ja luonnontieteet, TkK
Huhtikuu 2021

Matemaattinen mallintaminen sosiaalisen median informaation leviämisen ennustamisessa on useita eri matematiikan osa-alueita yhdistävä kokonaisuus. Koska sosiaalisen median käyttö on kasvanut huomattavasti viimeisen vuosikymmenen aikana, on informaation leviämisen tarkastelemiseksi kehitetty useita matemaattisia malleja.

Työssä tarkastellaan syvällisesti lineaarista vaikutusmallia, joka tarjoaa ennusteita sosiaalisen median informaatioiden leviämisprosesseista. Lineaarisen vaikutusmallin matemaattisen todistamisen avulla saadaan johdettua malli, jota sovelletaan kirjallisuudesta saatuun dataan informaation leviämisen ennustamiseksi. Tämän mallin toimintaa pyritään lisäksi arvioimaan kriittisesti vertailemalla keskenään mallin tarjoamia ennusteita todellisiin arvoihin aihepiireittäin. Työssä vertaillaan politiikan, talouden, urheilun ja viihteen aihepiirien informaatioiden leviämisen kokonaisvolyymia sekä pohditaan tutkimustuloksiin vaikuttavia tekijöitä.

Työ toteutetaan kirjallisuuskatsauksen ja kirjallisuudesta saadun datan analysoinnin yhdistelmänä. Työ jakaantuu kolmeen osaan. Kirjallisuustutkimusosa esittelee sosiaalisen median käsitteitä ja sovelluksia. Matemaattisessa osassa käydään läpi lineaarisen vaikutusmallin matemaattinen todistaminen. Työn viimeinen osa esittelee kirjallisuudesta saadun datan analysoinnin pohjalta saatuja tutkimustuloksia. Nämä tutkimustulokset mallinnetaan lineaarisen vaikutusmallin tarjoamien ennusteiden avulla aihepiireittäin.

Tutkimustulokset osoittavat, että matemaattinen mallintaminen on tehokas tapa sosiaalisen median informaation leviämisen ennustamisessa. Matemaattinen mallintaminen tuottaa kerätyn datan avulla työlle olennaisia tuloksia, joka tekee sosiaalisen median informaation leviämisen tarkastelusta tieteellisesti laadukasta. Matemaattisen mallintamisen hyödyntäminen sosiaalisen median informaation leviämisen tarkastelussa on tutkimusalueena tuore ja tutkimustulokset vähäisiä. Tutkimustulosten perusteella tutkimusalueessa on tunnistettavissa puutteita, jonka johdosta potentiaali jatkotutkimuksille on merkittävä.

Avainsanat: sosiaalinen media, matemaattinen mallintaminen, lineaarinen vaikutusmalli

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

ALKUSANAT

Haluaisin kiittää Juho Kanniaista työn aiheesta ja kommenteista työn aikana. Työn aihe oli haastava, mutta kirjoittamisprosessi opettavainen ja mielenkiintoinen. Haluaisin lisäksi kiittää perhettäni ja ystäviäni tuesta tämän kevään aikana.

Tampereella, 26. huhtikuuta 2021

Kiira Rintamäki

SISÄLLYSLUETTELO

1. Johdanto 1
1.1 Työn tavoitteet ja työn rajaus 1
1.2 Tutkimusmenetelmä. 2
1.3 Tutkimusaineisto ja sen valinta 3
1.4 Työn rakenne 3
2. Sosiaalinen media 5
3. Matemaattinen mallintaminen informaation leviämisen ennustamisessa 7
3.1 Mallin peruselementit 7
3.2 Mallin parametrinen arviointi 9
4. Tutkimusdata ja ennustevirhe 12
4.1 Tutkimusdata 12
4.2 Lineaarisen vaikutusmallin ennustevirhe. 13
5. Tulokset datan matemaattisesta mallintamisesta 14
5.1 Influessifunktio. 14
5.2 Lineaarinen vaikutusmalli 15
6. Yhteenveto 20
Lähteet 21

LYHENTEET JA MERKINNÄT

$A(t)$	Joukko aktivoituneita solmukohtia ennen tarkastelua ajanhetkeä t
$E_k(t)$	Todellisen ja ennustetun kokonaisvolyymin välinen erotus
$I_u(l)$	Solmukohdan u influenssifunktio aikayksikön l funktiona
K	Joukko erilaisia informaatioita
L	Influenssivektorin $\mathbf{I_u}$ pituus
$M_{u,k}(t)$	Indikaattorifunktio
N	Joukko solmukohtia eli verkkosivustoja
T	Vektorin $\mathbf{V_k}$ pituus
$V(t)$	Infektoituneiden solmukohtien kokonaisvolyymi ajanhetkellä t
$V_k(t)$	Informaation k ennen ajanhetkeä t saavuttaneiden solmukohtien kokonaisvolyymi
$\hat{I}$	Influenssin ennuste
$\hat{V}_k(t)$	Ennustettu kokonaisvolyymi
$\mathbf{I_u}$	Yksittäisen solmukohdan u influenssivektori
$\mathbf{I}$	Influenssivektori
$\mathbf{M_{u,k}}$	Influenssin indikaattori-matriisi yksittäisestä informaatio-solmukohta parista
$\mathbf{M}$	Influenssin indikaattori-matriisi
$\mathbf{V_k}$	Yksittäisen informaation k volyymivektori
$\mathbf{V}$	Volyymivektori
t	Aika
t_u	Ajanhetki, jolloin solmukohta u saavuttaa tietyn informaation
t_v	Ajanhetki, jolloin solmukohta v saavuttaa tietyn informaation
t_w	Ajanhetki, jolloin solmukohta w saavuttaa tietyn informaation
u	Yksittäinen tarkasteltava solmukohta
v	Yksittäinen tarkasteltava solmukohta
w	Yksittäinen tarkasteltava solmukohta
RE	Suhteellinen virhe

UGC	Sosiaalisen median käyttäjien luoma sisältö
Web 2.0	Vuonna 2004 kehitetty termi kuvaamaan uudenlaista tapaa käyttää www-pohjaista teknologiaa

1. JOHDANTO

Sosiaalinen media on maailmanlaajuinen ilmiö, jonka käyttö niin yksityisten käyttäjien kuin myös yritysten parissa on kasvanut räjähdysmäisesti viimeisen vuosikymmenen aikana. Tämän jatkuvasti kasvavan ja kehittyvän verkkoviestintäympäristön informaation leviämisen ennustaminen on kompleksisen luonteensa vuoksi usein hankalaa.

Tämä kandidaatintyö pyrkii tarkastelemaan ja ennustamaan sosiaalisen median informaation leviämistä matemaattisen mallintaminen avulla. Työssä pyritään myös selvittämään, miten eri aihepiirit vaikuttavat informaation kokonaisvolyymeihin sosiaalisen median verkostoissa. Työ on toteutettu kirjallisuuskatsauksen ja kirjallisuudesta saadun datan analysoinnin yhdistelmänä.

Tässä johdantoluvussa esitellään työn aihe sekä työlle asetetut tavoitteet. Johdantoluvussa rajataan myös työn päätutkimusongelma ja esitellään työssä käytetty tutkimusmenetelmä. Viimeisessä alaluvussa esitellään työn rakenne.

1.1 Työn tavoitteet ja työn rajaus

Tämän kandidaatintyön tavoitteena on löytää vastaus päätutkimusongelmaan: "Miten sosiaalisen median informaation leviämistä voidaan ennustaa matemaattisen mallinnuksen avulla?". Jotta päätutkimusongelmaan löydettäisiin kattava vastaus, luodaan päätutkimusongelman tueksi seuraavia apututkimuskysymyksiä:

1. Tarjoaako matemaattiset mallit täsmällisiä ennusteita informaation leviämisestä?
2. Mitkä tekijät vaikuttavat informaation leviämiseen?
3. Miten informaation aihepiiri vaikuttaa informaation leviämiseen?

Ensimmäiseen apututkimukseen pyritään löytämään vastaus kirjallisuudesta. Tarkoituksena on löytää matemaattinen malli, jota voidaan tehokkaasti soveltaa sosiaalisen median informaation leviämisen matemaattiseen mallintamiseen. Tämän mallin löydyttyä pyritään vastaamaan toiseen apututkimuskysymykseen. Tarkoituksena on siis tarkastella eri tekijöitä, jotka vaikuttavat mallin toimintaan ja siten myös informaation leviämiseen. Viimeiseen apututkimuskysymykseen pyritään löytämään vastaus soveltamalla valittua mallia kerättyyn dataan.

Sosiaalisen median informaation leviämiseen voitaisiin soveltaa useitakin matemaattisia malleja, mutta työn laajuuden rajoituksien takia pyritään tarkastelemaan vain yhtä kirjallisuuden pohjalta löydettyä mallia.

1.2 Tutkimusmenetelmä

Työ toteutetaan osittain kirjallisuuskatsauksena, jonka avulla tietyn aihepiirin kartoittaminen on läpinäkyvää ja tieteellisesti aiempien tutkimusten pohjalta toistettavaa [1]. Salmisen mukaan [2] kirjallisuuskatsaus on tehokas tapa kartoittaa tieteellisten tulosten kannalta tärkeitä tutkimuksia. Tämän kartoituksen avulla voidaan tehokkaasti testata hypoteeseja, esittää tutkimustuloksia tiivistetyssä muodossa sekä arvioida näitä tuloksia johdonmukaisesti. Kirjallisuuskatsaus voi myös paljastaa aikaisempien tutkimusten puutteet ja muodostaa täten uusia tutkimustarpeita. [2]

Työ sisältää lisäksi osuuden, jossa analysoidaan kirjallisuudesta saatua dataa. Dataa on kerätty informaation leviämisen tarkastelemiseksi useilta eri verkkosivustoilta aihepiirettäin.

Työn eteneminen noudattaa mukaillen Snyderin [3] esittämää kirjallisuuskatsauksen metodologiaa. Snyderin mukaan kirjallisuustutkimusprosessin voidaan neljään merkittävään vaiheeseen. Snyderin kirjallisuustutkimusprosessin vaiheet ovat:

1. suunnittelu *(design)*
2. toteutus *(conduct)*
3. analysointi *(analysis)*
4. järjestäminen ja kirjoittaminen *(structuring and writing)*

Ensimmäisessä vaiheessa määritellään tutkimusongelma, jonka avulla laaditaan tutkimuskysymykset. Näiden tutkimuskysymysten pohjalta muodostetaan hakutermejä, joiden avulla pyritään löytämään työlle olennaista ja tieteellisesti laadukasta kirjallisuutta. Tämän työn tutkimusongelma ja tutkimuskysymykset ovat esitettynä luvussa 1.1. Toinen vaihe koostuu tiedonhausta, jossa haetaan tietoa erilaisista tietokannoista ja kirjallisuudesta hakutermien avulla. Näistä tuloksista valitaan erilaisten laadullisten kriteerien avulla työhön mukaan otettava kirjallisuus. Viimeisissä vaiheissa pyritään löytämään vastaus aikaisemmin määriteltyyn tutkimusongelmaan ja esittelemään nämä tulokset tieteellisestä näkökulmasta mahdollisimman selkeästi sekä ymmärrettävästi. [3]

1.3 Tutkimusaineisto ja sen valinta

Tämän työn tutkimusaineisto kerättiin seuraavista tietokannoista:

- Elsevier Science Direct
- Web of Science
- Elsevier Scopus
- ACM Digital Library
- Andor

Erilaisten tietokantojen käytöllä pyrittiin löytämään mahdollisimman laaja-alainen otos eri tieteenalojen julkaisuja. Elsevier Science Direct, Web of Science ja Elsevier Scopus ovat tietokantoja, joista löytyy laajasti eri tieteenalojen julkaisuja. ACM Digital Library tuo työhön sosiaalisen median tietoteknillistä näkökulmaa. Andor yhdistää eri tieteenalojen julkaisuja sekä esimerkiksi opiskelijoiden tuottamia diplomitöitä tietokannassaan, jolloin otos eri julkaisuista saadaan mahdollisimman laajaksi.

Hakulausekkeet muodostetaan yhdistämällä sosiaalisen median aihepiirin hakusanoja matemaattisen mallinnuksen aihepiirin sanoihin. Hakutuloksia siis haetaan julkaisuista, jotka sisältävät ainakin yhden sosiaalisen median ja matemaattisen mallinnuksen hakusanan. Käyttämällä laajasti erilaisia hakutermejä voidaan varmistua mahdollisimman laajan ja työlle olennaisen otoksen muodostumisesta. Sosiaalisen median aihepiiristä käytetään seuraavia hakusanoja: *social media, online social network, social platform, online media, real-time information, social networking, social computing*. Matemaattisen mallintamisen aihepiirin hakusanat ovat seuraavat: *mathematical modeling, mathematical pattern, information diffusion, statistical methods, diffusion dynamics, analysis, computational intelligence, complex system*.

Kandidaatintyön aihe on hyvin ajankohtainen, jolloin osa löydetyistä lähteistä on hyvinkin tuoreita. Työn lähteet pyritään kuitenkin pitämään tieteellisesti luotettavia tarkastelemalla niiden julkaisufoorumeja, viittauksia, tekijöitä ja julkaisuajankohtaa. Tässä työssä käytettyjä lähteitä on pyritty arvioimaan kriittisesti mahdollisimman laadukkaan työn tuottamiseksi.

1.4 Työn rakenne

Kandidaatintyön toinen luku käsittelee laajemmin sosiaalista mediaa. Luvussa määritellään työn kannalta oleellisia käsitteitä sosiaalisen median aihepiiristä. Luvussa luokitellaan myös erilaisia sosiaalisen median lähteitä ja esitellään erilaisia sosiaalisen median esimerkkisovelluksia.

Kolmas luku esittelee työn matemaattista osuutta, jossa kuvataan kirjallisuuden pohjalta

valitun lineaarisen vaikutusmallin peruselementtejä. Kolmannessa luvussa syvennytään lisäksi lineaarisen vaikutusmallin kannalta oleelliseen parametriseen arviointiin.

Datan soveltamista lineaarisen vaikutusmallin avulla on esitetty luvussa neljä, jossa sovelletaan Stanfordin yliopiston professori Jure Leskovecin keräämää dataa informaation leviämisestä [4]. Luvussa pyritään myös tarkastelemaan, miten eri aihepiirit vaikuttavat informaation diffuusionopeuksiin sosiaalisen median verkostoissa.

Viimeisessä luvussa kerätään yhteen työn tulokset ja pohditaan, onko asetettuun päätutkimusongelmaan löydetty vastaus. Lisäksi yhteenvedossa pyritään miettimään mallin käytön haasteita sekä tarjoamia mahdollisuuksia jatkotutkimuksien näkökulmasta.

2. SOSIAALINEN MEDIA

Tässä luvussa määritellään käsitteet sosiaalinen media, web 2.0 ja UGC, jotka ovat olennaisia tutkimusongelman ymmärtämiseksi. Lisäksi luvussa annetaan esimerkkejä sosiaalisen median sovelluksista.

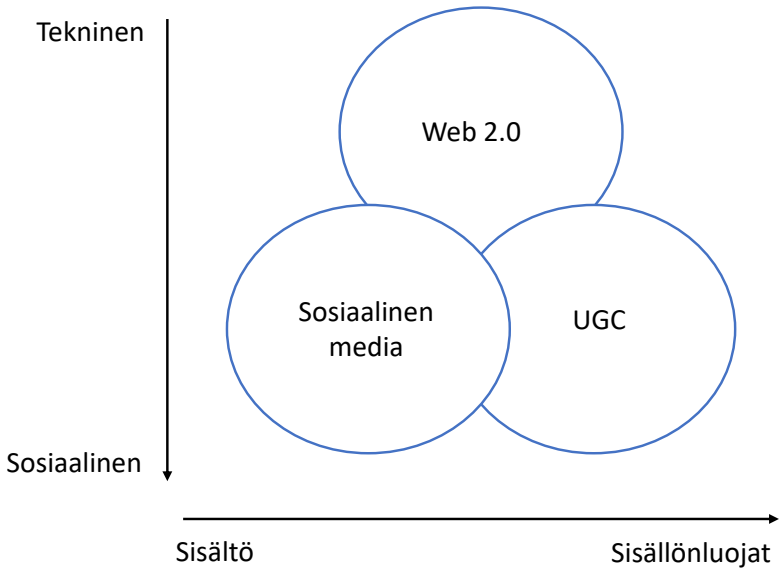
Sosiaalisen median käsitettä käytetään kuvaamaan verkkoviestintäpalveluita, jotka saavat valtaosan sisällöstään käyttäjiltä itseltään tai käyttäjien yhdistäessä muiden sivustojen sisältöä omaan sosiaalisen median syötteeseensä. Sosiaalisen median toiminta perustuu web 2.0 -pohjaisen teknologian sekä UGC:n yhdistämiseen. [5] Jotta siis sosiaalisen median käsitteen voisi täysimittaisesti ymmärtää, on tarpeellista määritellä myös käsitteet *web 2.0* ja *UGC*.

Web 2.0 on Tim O'Reillyn vuonna 2004 kehittämä käsite, jolla kuvataan sovelluskehittäjien ja loppukäyttäjien siirtymistä toiminnallisempiin www-teknologia pohjaisiin sovelluksiin [6]. Web 2.0 siis edustaa merkittävää muutosta tapaan, jolla digitaalista sisältöä luodaan, jaetaan, tallennetaan ja käsitellään [7].

UGC tulee englannin kielen sanoista User Generated Content, joka tarkoittaa käyttäjien luomaa sisältöä. Käyttäjien julkaisemalla sisällöllä on tiettyjä yhteisiä ominaisuuksia: Sisältö on julkisesti muiden käyttäjien saatavilla, sisällön luominen vaatii tietyn määrän luovaa työskentelyä ja sisältö on tyypillisesti luotu ammatillisen työskentelyn ulkopuolella vapaa-ajalla. [8] Tällainen sisältö voi olla esimerkiksi tekstiä, kuvia, kommentteja, käyttäjänimiä, tykkäyksiä tai muuta mediaa.

Käsitteitä sosiaalinen media, web 2.0 ja UGC käytetään usein kuluttajien keskuudessa epätarkasti. Syy liittyy osaksi siihen, että nämä käsitteet liittyvät vahvasti toisiinsa sekä ovat toisistaan riippuvaisia. Nämä ovat käsitteellisesti kuitenkin erillisiä. Kuten kuvasta 2.1. voidaan nähdä, web 2.0:n voidaan ajatella olevan tekninen infrastruktuuri, joka mahdollistaa sosiaalisen median ilmiön ja helpottaa kuluttajien sisällön luomista. Yksinkertaistetusti ilmaistuna, web 2.0 mahdollistaa käyttäjien sisällön luomisen ja jakamisen, jota kutsutaan sosiaaliseksi mediaksi.[9]

Sosiaalisen median kanavia voidaan jaotella niiden vaatiman sosiaalisen kanssakäymisen ja sosiaalisten tavoitteiden mukaan. Näitä luokkia ovat: blogit, mikroblogit, sosiaaliset verkostot, sisältökeskeiset yhteisöt, sosiaaliset virtuaalimaailmat, virtuaaliset peliympäristöt ja yhteistoiminnalliset projektit [10][11].



Kuva 2.1. *Sosiaalisen median käsitteen muodostuminen. Perustuu lähteeseen [9].*

Aikaisinta sosiaalisen median muotoa edustavat blogit, jotka ovat tyypillisesti erityyppisiä yksityisten käyttäjien ylläpitämiä verkkosivustoja kirjoittajan omasta yksityiselämästään tai tietystä aihealueesta [10]. Nykyään suosiotaan ovat kasvattaneet myös mikroblogipalvelut, kuten Twitter, joka sallii mitaltaan vain 140 merkkiin rajattujen päivistysten tekemisen [11].

Sosiaaliset verkostot ovat ihmisten tai organisaatioiden muodostamia sosiaalisia rakenteita, joissa verkoston jäseniä yhdistää toisiinsa esimerkiksi ystävyys, sukulaisuus, työsuhde tai harrastus. Tunnetuimpia sosiaalisia verkostoja ovat Facebook ja ammatillisempi LinkedIn. Sisältökeskeisissä yhteisöissä taas painoarvo on puhtaasti sisällön jakamisessa eikä niinkään verkostoitumisessa muiden verkoston jäsenien kanssa. Tunnetuin tällainen sisältökeskeinen yhteisö on Youtube. [10][11]

Sosiaalisilla virtuaalimaailmoilla tarkoitetaan tietokoneavusteisesti simuloituja ympäristöjä, joissa yksityisten käyttäjien on mahdollista esiintyä itse luomansa avatarin kautta. Tämän kaltaisia sivustoja ovat esimerkiksi Second Life ja Habbo Hotel. Virtuaaliset pelimaailmat, kuten World of Warcraft, ovat toinen tietokoneavusteisten alustojen ryhmä. Sosiaalisiin virtuaalimaailmoihin verrattuna suurin erottava tekijä on alustojen pelillisyys. Virtuaalisissa pelimaailmoissa käyttäjien toimintaa rajoittaa tarkat säännöt ja pelilliset tavoitteet, kun taas sosiaaliset virtuaalimaailmat keskittyvät enemmän käyttäjien väliseen kanssakäymiseen realistisessa maailmassa. [10][11]

Yhteistoiminnalliset projektit mahdollistavat sisällön yhteisluomisen useiden käyttäjien kesken. Tunnetuin yhteistoiminnallinen projekti on Wikipedia. [10]

3. MATEMAATTINEN MALLINTAMINEN INFORMAATION LEVIÄMISEN ENNUSTAMISESSA

Tässä luvussa käydään läpi työn matemaattinen osuus. Luku pohjautuu Leskovecin ja Yangin artikkeliin Modeling Information Diffusion in Implicit Networks [12].

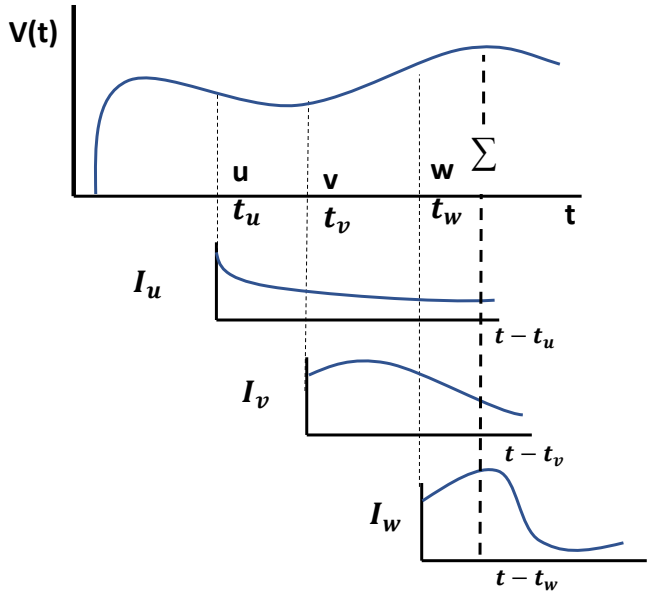
3.1 Mallin peruselementit

Sosiaalinen media on verkkoviestintäympäristö, joka mahdollistaa reaaliaikaisen tiedon tuottamisen ja levittämisen. Tyypillisesti tämän kaltaisen tiedon leviäminen mielletään yksinkertaisena leviämisprosessina sosiaalisten verkostojen yli, vaikka todellisuudessa se on seuraus lukuisien osallistujien välisistä vuorovaikutuksista monimutkaisissa verkostoissa.

Tätä leviämisprosessia voidaan kuvata **lineaarisen vaikutusmallin** *(Linear Influence Model)* avulla, jossa keskitytään tarkastelemaan tietyn solmukohdan vaikutusta informaation diffuusionopeuteen sosiaalisen median verkostoissa. Lineaarisessa vaikutusmallissa pyritään siis tarkastelemaan, milloin jokin informaatio on saavuttanut tietyn solmukohdan sen sijaan, että tarkasteltaisiin mikä solmukohta informaation on levittänyt. Mallissa passiiviset solmukohdat muuttuvat aktiivisiksi infektoituessaan eli saavuttaessaan tietyn informaation. Tämän jälkeen ne voivat levittää informaatiota edelleen eteenpäin infektoiden muita solmukohtia verkostossa.

Matemaattisten mallien käyttämiseksi ja soveltamiseksi todellisuudessa tulee tehdä muutamia oletuksia: täydellinen data verkostosta on saatavilla, informaation leviäminen voi tapahtua vain verkoston reunoja pitkin ja verkoston rakenteen mukainen informaation leviäminen riittää selittämään havaitut ilmiöt. Lineaarisen vaikutusmallin lähtökohtana toimii oletus, että jo infektoituneiden solmukohtien määrä riippuu aikaisemmin infektoituneiden solmukohtien määrästä. Tämän lähtökohdan avulla voidaan johtaa funktio tulevien solmujen informaation saavuttamisen diffuusionopeudelle.

Määritelmä 3.1. Olkoon $I_u(l)$ influenssifunktio *(influence function)*, joka kuvaa kuinka monta tietystä infektiosta seuraavaa infektiota voidaan liittää tähän aikaisemmin infektoituneeseen solmukohtaan u aikayksikön l aikana. Ensimmäinen informaation maininta käynnistää seuraavien mainintojen ketjun siten, että $I_u(1), \ldots I_u(n)$.



Kuva 3.1. *Solmukohtien influenssifunktioiden diffuusionopeus. Perustuu lähteeseen [12].*

Määritelmä 3.2. Merkitään infektoituneiden solmukohtien kokonaisvolyymia $V(t)$.Tällöin solmukohtien u influenssifunktioiden summa on

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u), \tag{3.1}$$

missä $A(t)$ on joukko aktivoituneita solmukohtia u ennen ajanhetkeä $t, 0 \leq t_u \leq t$.

Funktion (3.1) avulla saadaan piirrettyä kuva 3.1, jossa ylin käyrä kuvaa summafunktion $V(t)$ arvoja ajan t funktiona. Muuttujat t_u, t_v ja t_w kuvaavat ajanhetkiä, jolloin yksittäiset solmukohdat u, v ja w infektoituvat eli saavuttavat tietyn informaation. Nämä solmukohdat lisäävät $I_u(t-t_u)$, $I_v(t-t_v)$ ja $I_w(t-t_w)$ verran infektioiden lopullista määrää ajanhetkellä t.

Yksittäisen influenssifunktion $I_u(l)$ mallinnusta voidaan tarkastella erilaisten lähestymistapojen avulla. Lähestymistapa voi olla parametrinen, jolloin funktion $I_u(l)$ oletetaan noudattavan tiettyä parametrista muotoa. Funktiota $I_u(l)$ voitaisiin tarkastella esimerkiksi eksponentiaalisesti muodossa $I_u(l) = c_u e^{-\lambda_u l}$ tai potenssilain avulla muodossa $I_u(l) = c_u l^{-\alpha_u}$.

Vaikkakin parametrinen lähestymistapa olisi yksinkertainen tapa lähestyä asiaa, sisältää se muutamia perustavanlaatuisia ongelmia. Parametrinen lähestysmistapa olettaa, että kaikki solmukohdat noudattavat samaa parametrista muotoa. Tämä liiankin pelkistetty oletus jättää kuitenkin huomioimatta sosiaalisen median kompleksisen luonteen, jossa jokaisella käyttäjällä on varsin erilainen vaikutus informaation leviämiseen. Tämän monimuotoisuuden huomioon ottamiseksi tulee tarkastelussa käyttää parametritonta lähesty-

mistapaa. Parametriton lähestymistapa lineaariseen vaikutusmalliin ei tee mitään oletuksia yksittäisen influenssifunktion muodosta, mikä tuo joustavuutta lineaarisen vaikutusmallin mallintamiseen.

3.2 Mallin parametrinen arviointi

Seuraavaksi esitellään tehokas tapa arvioida lineaarisen vaikutusmallin parametreja. Olkoon N joukko solmukohtia ja K joukko erilaisia informaatioita eli tartuntoja, jotka ovat levinneet verkoston solmukohtien välillä ajan kuluessa. Oletetaan ajan kasvavan epäjatkuvina intervalleina, kuten tunteina siten, että $T = 0, 1, \ldots, T$. Olkoon $V_k(t)$ tartunnan k kokonaisvolyymi ajanhetkellä t.

Määritelmä 3.3. Olkoon $M_{u,k}(t)$ *indikaattorifunktio* ajanhetkellä t. Muuttujan $M_{u,k}(t)$ arvoja voidaan tarkastella ajanhetkellä t seuraavasti:

- Jos solmukohta u on saavuttanut informaation k, saa $M_{u,k}(t)$ arvon 1.
- Jos solmukohtaa u *ei* ole saavuttanut informaatiota k, saa $M_{u,k}(t)$ arvon 0.

Lause 3.1. *Merkitään kokonaisvolyymia $V_k(t+1)$ yksittäisten solmukohtien u influenssifunktioiden summana, jotka ovat saavuttaneet tietyn informaation ennen ajanhetkeä $(t+1)$. Tällöin voidaan merkitä*

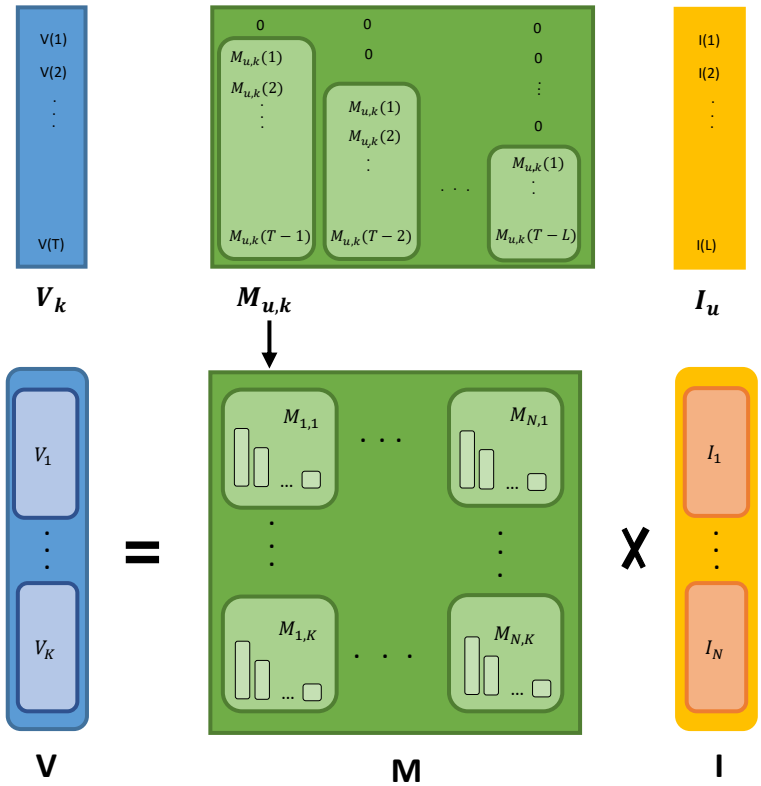
$$V_k(t+1) = \sum_{u=1}^{N} \sum_{l=0}^{L-1} M_{u,k}(t-l) I_u(l+1), \tag{3.2}$$

missä ulompi summafunktio käy läpi kaikki solmukohdat ja sisempi summafunktio kaikki influenssifunktiot määritetyllä aikavälillä.

Funktion 3.2 avulla voidaan tarkastella saavuttiko solmukohta u informaation k l-aikayksikköä sitten annetulla ajanhetkellä t. Jos informaatio saavutti solmukohdan, saa $M_{u,k}(t-l)$ arvon 1. Tällöin siis solmukohdan u influenssifunktio $I_u(l)$ vaikuttaa kokonaisvolyymiin $V_k(t)$ influensifuntkion $I_u(l)$ verran.

Seuraavaksi esitellään tapa arvioida yksittäisiä influenssifunktioita $I_u(t)$, jonka avulla voidaan ennustaa tarkemmin volyymin $V_k(t+1)$ käyttäytymistä. Kaikkien solmukohtien yksittäisten influenssifunktioiden tarkastelu ei ole oleellista, vaan tavoitteena on tarkastella kokonaisvolyymia $V(t)$ pienemmän joukon solmukohtien N funktiona. Mallin toimimiseksi tehdään lisäksi oletus, että K on suurempi kuin N.

Koska ajan t oletetaan kuluvan epäjatkuvina intervalleina, voidaan $V_k(t)$ ja $I_u(l)$ esittää vektoreina ja $M_{u,k}(t)$ matriisina. Näin saadaan muodostettua pienimmän neliösumman menetelmän lähtötilannetta vastaava ongelma, missä jokaiselle solmukohdalle u pyritään arvioimaan L-kappaletta influenssifunktion arvoja siten, että $I_u(1), \ldots, I_u(L)$.



Kuva 3.2. *Matriisiyhtälön muodostamiseksi määritellyt vektorit* ***V*** *ja* ***I*** *sekä matriisi* ***M****. Perustuu lähteeseen [12].*

Arvoja $I_u(l)$ voidaan arvioida yksinkertaisella matriisiyhtälöllä, koska volyymi $V_k(t)$ on suoraan verrannollinen influenssifunktioon $I_u(l)$ kuten on esitetty funktiossa 3.2. Matriisiyhtälön muodostamiseksi määritellään volyymivektori **V**, influenssivektori **I** ja influenssin indikaattori-matriisi **M**, joita on esitelty kuvassa 3.2.

Määritelmä 3.4. Oletetaan ajan kasvavan epäjatkuvina intervalleina, kuten tunteina siten, että $T = 0, 1, \ldots, T$. Tarkastellaan yksittäisen tartunnan volyymifunktiota $V_k(t)$ vektorimuodossa $\mathbf{V_k}$ siten, että $\|\mathbf{V_k}\| = T$. Kasvattamalla eri tartuntojen määrää siten, että $k = 1, \ldots, K$, voidaan määritellä *volyymivektori* **V**. Olkoon **V** volyymivektori, jonka pituutta merkitään pistetulolla $K \cdot T$. Volyymivektori **V** on pystyvektori, missä $\mathbf{V} = [V_1, V_2, \ldots, V_K]^T \in \mathbb{R}^{K \times T}$.

Määritelmä 3.5. Tarkastellaan yksittäisen solmukohdan influenssifunktiota $I_u(l)$ vektorimuodossa $\mathbf{I_u}$ siten, että $\|\mathbf{I_u}\| = L$. Kasvattamalla verkoston solmukohtien määrää siten, että $u = 1, \ldots, N$, voidaan määritellä *influenssivektori* **I**. Olkoon **I** influenssivektori, jonka pituutta merkitään pistetulolla $N \cdot L$. Influenssivektori **I** on pystyvektori, missä $\mathbf{I} = [I_1, I_2, \ldots, I_N]^T \in \mathbb{R}^{N \times L}$.

Määritelmä 3.6. Olkoon **M** *influenssin indikaattori-matriisi*. Matriisi **M** koostuu $K \cdot T$ rivistä ja $N \cdot L$ sarakkeesta siten, että $\mathbf{M} = (M_{1,1}, \ldots M_{N,K}) \in \mathbb{R}^{K \times T \cdot N \times L}$.

Oletetaan, että solmukohta u saavuttaa informaation k ajanhetkellä t. Lisäksi matriisin **M** alkiolle (i, j) pätee, että $m_{i,j} = 1$, kun $i = kT+t+l$ ja $j = uL+l+1$ muuttujan l arvojen vaihdellessa välillä $l = 0, \ldots, \min(L-1, T-t)$. Matriisilla **M** on lohkorakenne, jossa jokainen lohko $M_{u,k}$ esittää solmukohdan ja informaation tartunnan muodostamaa paria. Solmukohdan u saavuttaessa informaation k ajanhetkellä t, lohkon $M_{u,k}$ diagonaaliset alkiot saavat arvon 1. Kun $l = l, \ldots, L-1$, voidaan merkitä

$$M_{u,k}[t+l, l+1] = 1. \tag{3.3}$$

Tällöin lohkon $M_{u,k}$ rivit T laskevat aikaa ja sarakkeet L mittaavat solmukohdan influenssin muutosta tartunnan jälkeen.

Seuraavana tavoitteena on ratkaista matriisiyhtälö **V** $=$ **M I**. Influenssivektorin **I** arvoja pyritään arvioimaan, kun volyymivektorin **V** ja influenssia indikoiva matriisin **M** arvot ovat tiedossa datan muodossa. Systeemin kompleksisuuden vuoksi täsmällisen tarkkaa vastausta ei voida olettaa löytyvän. Tarkoituksena on löytää ennuste influenssivektorille **I**, joka minimoi Euklidisen etäisyyden avulla mitatun ennustevirheen oikean ja ennustetun volyymin välillä.

Pienimmän neliösumman menetelmä positiivisille luvuille *(Non-negative least squares)* on eräänlainen rajoitettu sovellus pienimmän neliösumman menetelmästä, jota voidaan soveltaa tehokkaasti suurellekkin joukolle solmukohtia ja tartuntoja.

Lause 3.2. *Olkoon NNLS:n mukainen ongelma influenssifunktion ennusteiden $\hat{I}$ muodostamiseksi*

$$\hat{I} = \min_{\boldsymbol{I} \geq 0} \| \boldsymbol{V} - \boldsymbol{M}\,\boldsymbol{I} \|_2^2 \tag{3.4}$$

*missä $\| \ \|_2^2$ kuvaa Euklidisen normin neliötä jokaisen vektorin **I** alkion ollessa positiivinen.*

Soveltamalla heijastavaa Newtonin menetelmää *(reflective Newton method)* funktioon (3.4), saadaan arvioitua influenssifunktion arvoja [13]. Nämä arvot voidaan syöttää funktioon (3.2), jolloin saadaan ennustettua kokonaisvolyymi ajanhetkellä $t+1$.

4. TUTKIMUSDATA JA ENNUSTEVIRHE

Tässä luvussa kuvaillaan käytettyä dataa ja kokeellista järjestelyä sekä tarkastellaan lineaarisen vaikutusmallin ennustevirhettä. Alaluvut 4.1 ja 4.2 pohjautuu Leskovecin ja Yangin artikkeliin Modeling Information Diffusion in Implicit Networks [12].

4.1 Tutkimusdata

Kokeellisessa järjestelyssä mallinnetaan lyhyiden fraasien diffuusiota sosiaalisen median verkostoissa. Yang ja Leskovec ovat keränneet dataa 343:sta miljoonasta fraasista yli miljoonalta eri verkkolähteeltä välillä 1. syyskuuta 2008 ja 31. elokuuta 2009. Varmistuakseen koko fraasien elinkaarin taltioinnista, Yang ja Leskovec valitsivat tarkasteluun vain 5. syyskuuta 2008 jälkeen ilmestyneitä fraaseja.

Kokeelliseen järjestelyyn valittiin mukaan yhteensä 1000 fraasia, joiden volyymi oli suurin tietyn ajanjakson aikana niiden volyymihuipun ympärillä. Tämän viiden päivän ajanjakson aikana kerättiin dataa verkkosivuista, jotka mainitsivat näitä kokeelliseen järjestelyyn valittuja fraaseja. Tämän työn toteutus perustuu Yangin ja Leskovecin keräämän datan tarkasteluun ja mallintamiseen. Kuvassa 4.1 esitettynä esimerkki tutkimusdatan tietomuodosta.

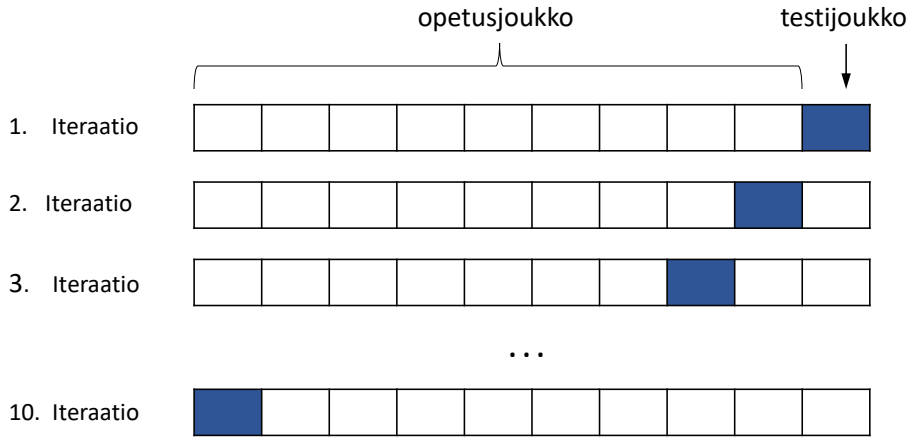
```
data=[<M_{1,1}(1)>,<M_{2,1}(1)>,...,<M_{N,1}(1)>,<V_1(1)>,<1>;
<M_{1,1}(2)>,<M_{2,1}(2)>,...,<M_{N,1}(2)>,<V_1(2)>,<1>;
<M_{1,1}(3)>,<M_{2,1}(3)>,...,<M_{N,1}(3)>,<V_1(3)>,<1>;
...
<M_{1,1}(T)>,<M_{2,1}(T)>,...,<M_{N,1}(T)>,<V_1(T)>,<1>;
<M_{1,2}(1)>,<M_{2,2}(1)>,...,<M_{N,2}(1)>,<V_2(1)>,<2>;
...
<M_{1,K}(T)>,<M_{2,K}(T)>,...,<M_{N,K}(T)>,<V_K(T)>,<T>;]
```

Kuva 4.1. *Esimerkki tutkimusdatan tietomuodosta.*

Kerätty data on tallennettu binääritiedoston muodossa. Kuvassa 4.1 toistuva rakenne M_{U,K}(t) kuvaa, kuinka monta kertaa solmukohta u saavutti tietyn informaation k ajanhetkeen t mennessä. Lisäksi binääritiedosto sisältää kaksi vektoria, johon on tallennettu käytetyt verkostossa leviävät fraasit je verkkosivustojen nimet.

4.2 Lineaarisen vaikutusmallin ennustevirhe

Informaation k kokonaisvolyymia $V_k(t)$ voidaan tarkastella ajanjaksojen avulla. Tavoitteena on arvioida lineaarisen vaikutusmallin kykyä ennustaa informaation leviämistä ajanjaksoissa. Etsitään siis ne solmukohdat, jotka saavuttivat informaation k ajanhetkeen t mennessä ja ennustetaan volyymi $V_k(t+1)$ informaation k leviämisestä tulevaisuuden ajanhetkellä $t+1$.



Kuva 4.2. *Havainnollistus 10-kertaisesta ristiinvalidoinnista.*

Sovelletaan lineaarisen vaikutusmallin toiminnan arvioimiseksi 10-kertaista ristiinvalidointia. Tätä ristiinvalidointia on havainnollistettu kuvassa 4.2. Aluksi data jaetaan 10:een mahdollisimman yhtä suureen osaan ristiinvalidointia varten. Ristiinvalidoinnissa suoritetaan 10 kierrosta, joissa kullakin kierroksella yksi osa toimii testijoukkona ja loput 9 osaa opetusjoukkona. Jokaisella kierroksella lasketaan testijoukosta ennustevirhe vertaamalla oikeita luokkatietoja ja ennustettuja luokkatietoja keskenään. Arvioidaan jokaiselle testijoukon tartunnalle k ennuste $\hat{V}_k$ ajanhetkellä $t+1$. Lasketaan seuraavaksi todellisen volyymin ja ennusteen välinen erotus

$$E_k(t+1) = V_k(t+1) - \hat{V}_k(t+1), \tag{4.1}$$

jonka avulla voidaan laskea mallin suhteellinen virhe

$$\text{RE} = \frac{\sqrt{\sum_{k,t} E_k(t)^2}}{\sqrt{\sum_{k,t} V_k(t)^2}}. \tag{4.2}$$

Soveltamalla 10-kertaista ristiinvalidointia, voidaan laskea ennustevirhe lineaariselle vaikutusmallille ja vertailla sitä muiden kirjallisuuden mallien ennustekykyyn. Yangin ja Leskovecin tuloksien mukaan, voidaan lineaarisen vaikutusmallin todeta olevan soveltuva sosiaalisen sosiaalisen median informaation leviämisen tarkasteluun.

5. TULOKSET DATAN MATEMAATTISESTA MALLINTAMISESTA

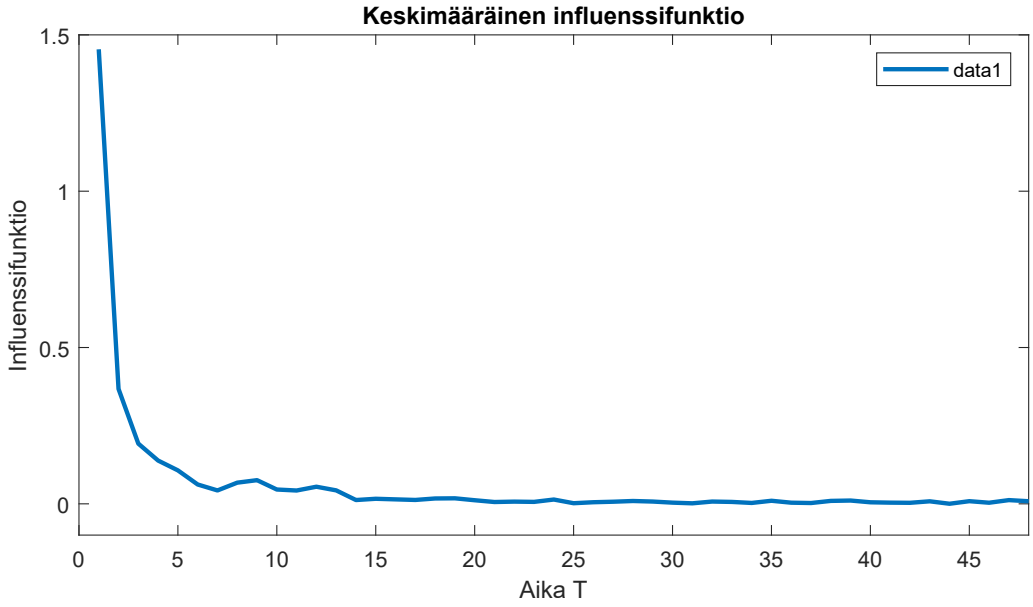
Tässä luvussa esitellään tulokset matemaattisista mallinnuksista. Luvussa esitellään koko dataa hyödyntämällä saatu influenssifunktio sekä vertaillaan lisäksi lineaarisen vaikutusmallin antamia ennusteita todellisiin arvoihin aihepiireittäin. Mallinnus perustuu influenssifunktion arvojen arvioimiseen. Näitä arvoja pyritään arvioimaan pienimmän neliösumman menetelmällä funktion (3.4) mukaisesti. Täten kokonaisvolyymia voidaan ennustaa funktion (3.2) mukaisesti.

Matemaattisessa mallintamisessa lähdetään liikkeelle lähtöarvojen asettamisesta. Tarkasteluissa käytetään eri tartuntojen määrän lähtöarvona $K = 1000$. Lisäksi asetetaan lähtöarvo $L = 10$, joka kuvaa solmukohdan influenssin laskemista arvoon nolla kymmenen tunnin jälkeen. Jokaisen tartunnan tarkastelu lähtee liikkeelle ajanhetkestä $t = 0$. Matemaattisessa mallinnuksessa tarkastellaan yhteensä $N = 100$ solmukohdan eli eri sivuston välistä informaation leviämistä. Alkuperäisessä datasetissä on kerätty tietoa yli miljoonalta verkkosivustolta, mutta tarkasteluun on valittu vain nämä 100 aktiivisinta sivustoa. Datasta analysoidut tulokset ovat mallinnettu MATLABIN versiolla R2019b.

5.1 Influessifunktio

Tarkasteltu ajanjakso T keskimääräistä influenssifunktiota tarkastellessa on 48, koska tutkimusten mukaan käyttäjien jakamalla sisällöllä on hyvin lyhyt elinikä [14]. Kuvassa 5.1 esitellään keskiarvo kaikkien solmukohtien influenssifunktiosta ajan funktiona.

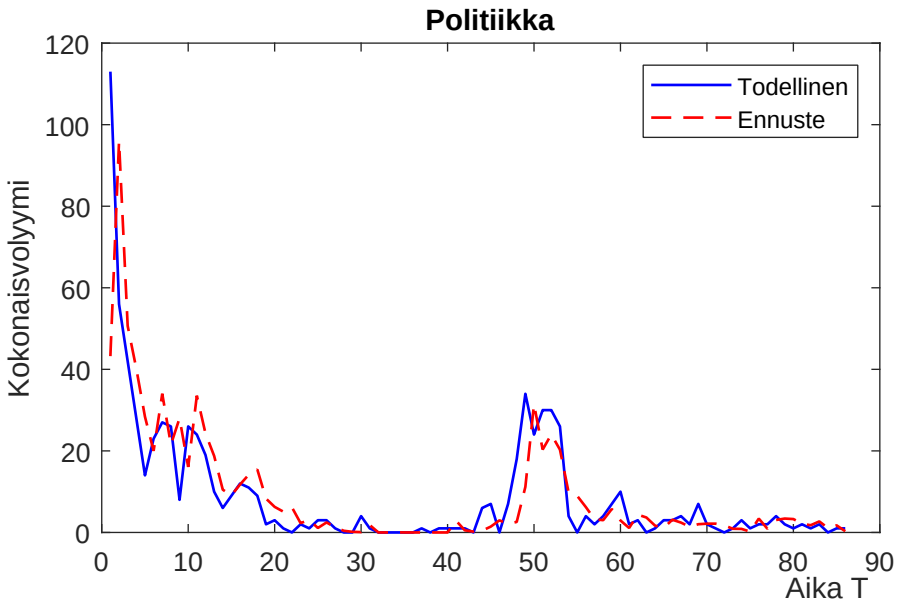
Informaation leviäminen lähtee liikkeelle tietystä verkkosivustosta, josta informaation leviäminen muille verkkosivustoille tapahtuu minuuttien sisällä. Alkuperäisen verkkosivuston ja siitä informaation saavuttaneiden verkkosivustojen informaation leviäminen ei tyypillisesti kuitenkaan pysähdy. Informaatiovirrat verkoston sisällä vain hidastuvat. Tämä hidastuminen tapahtuu havaittujen tutkimustulosten mukaan noin 12 tunnin kuluttua. [14] Informaation leviämisen hidastuminen on huomattavissa myös kuvassa 5.1, jossa kokonaisvolyymi saavuttaa viimeisen nousukohtansa noin 12 tunnin kohdalla. Tämä vahvistaa oletusta siitä, että sosiaalisen median käyttäjiä vetää puoleensa ajankohtaiset ja tuoreet informaatiot.



Kuva 5.1. *Influenssifunktio ajan funktiota informaation leviämisen tarkastelemiseksi*

5.2 Lineaarinen vaikutusmalli

Seuraavaksi vertaillaan lineaarisen vaikutusmallin antamia ennusteita todellisiin arvoihin aihepiireittäin. Datasta valitaan erilaisia fraaseja neljästä eri aihepiiristä: politiikasta, taloudesta, urheilusta ja viihteestä.

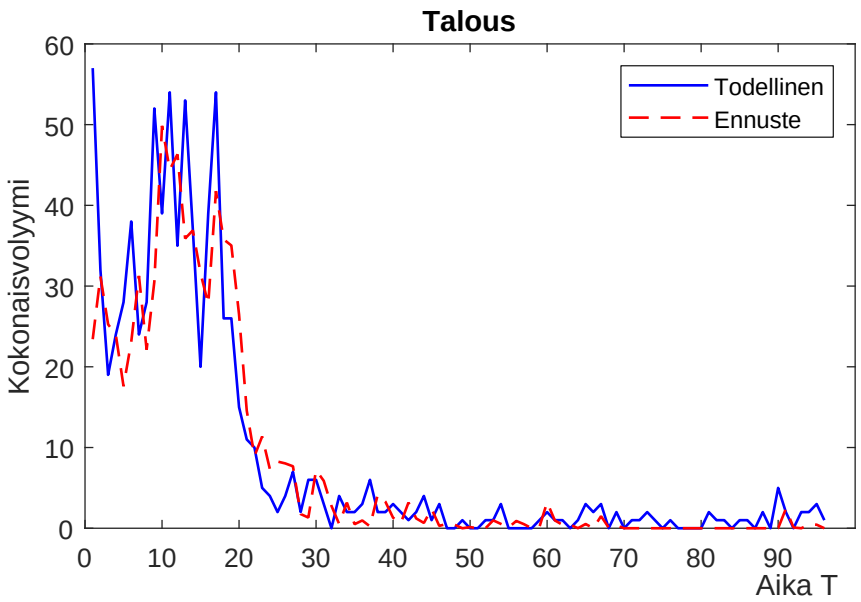


Kuva 5.2. *Poliitiikan aihepiirin fraasien kokonaisvolyymin ennusteet ja todelliset arvot*

Kuvassa 5.2 kuvataan poliittisten fraasien leviämistä sosiaalisen median verkostoissa. Poliittisten fraasien kokonaisvolyymia tarkastellaan suurimmalla asteikolla, koska näiden

fraasien leviäminen sosiaalisen median verkostoissa on tutkitusti kaikista suurinta. Dataa kerättiin välillä 1. syyskuuta 2008 ja 31. elokuuta 2009, jolloin politiikan suuri kokonaisvolyymi voidaan selittää vuoden 2008 Yhdysvaltain presidentinvaaleilla. Tutkimusten mukaan kaksi kolmasosaa amerikkalaisista sosiaalisen median käyttäjistä ovat harjoittaneet jonkinlaista kansalais- tai poliittista toimintaa sosiaalisessa mediassa. Lisäksi 22 prosenttia rekisteröidyistä yhdysvaltalaisista äänestäjistä käytti sosiaalista mediaa kertoakseen muille, miten he äänestivät presidentinvaaleissa. [15]

Poliittisten fraasien leviämistä tarkastellessa, voidaan huomata piikki kokonaisvolyymissa noin 48 tunnin kohdalla. Tämä voidaan selittää kaksiasteisella flow-mallilla (*two-step flow model*), jonka mukaan informaatio saavuttaa tyypillisesti ensin aktiiviset sosiaalisen median käyttäjät. Näiden aktiivisten käyttäjien levittäessä informaatiota, saavuttaa informaation myös vähemmän aktiiviset käyttäjät. Suurin osa ihmisistä siis saavuttaa informaation toisilta käyttäjiltä eikä perinteisistä massamedioista. [16]



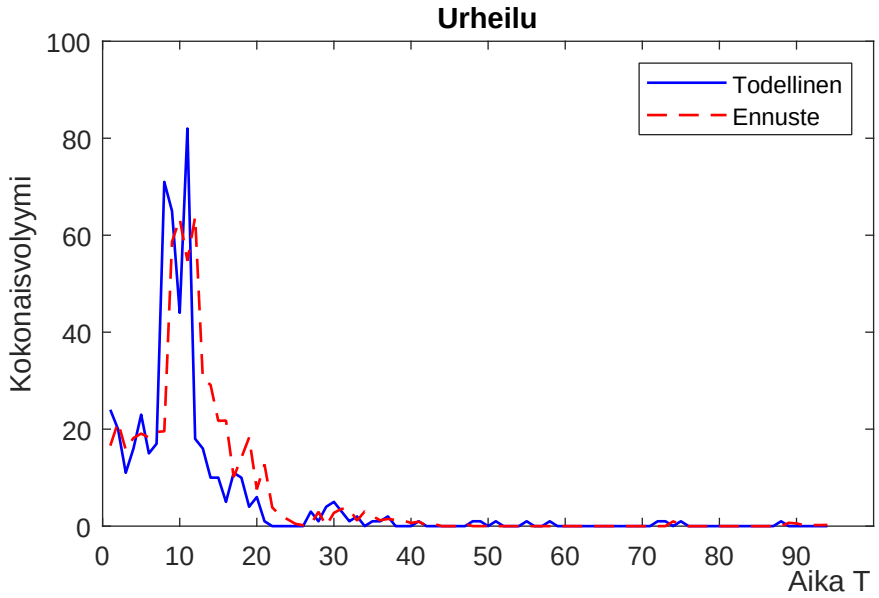
Kuva 5.3. *Talouden aihepiirin fraasien kokonaisvolyymin ennusteet ja todelliset arvot*

Kaksiasteisen flow-mallin vaikutus voidaan huomata myös kuvassa 5.3, jossa tarkastellaan talouden aihepiiriin liittyvien fraasien mainintojen kokonaisvolyymia. Vähemmän aktiivisten sosiaalisen median käyttäjien adoptoiman informaation vaikutukset kokonaisvolyymiin ovat kuitenkin havaittavissa aikaisemmin jo noin 10 tunnin jälkeen. Poliittisissa sosiaalisen median julkaisuissa on mukana usein vahvasti yksilön oman mielipiteen ilmaisu. Tähän yksilön oman mielipiteen muodostamiseen käytetään tutkimusten mukaan enemmän aikaa [17], jolloin politiikan kokonaisvolyymin toinen piikki on talouden kokonaisvolyymiin verrattaen myöhemmin.

Taloudellisen aihepiirin fraasien kokonaisvolyymi on suurempi huomattavasti pidempään muihin aihepiireihin verrattessa. Politiikan, teknologian, urheilun ja viihteen kokonaisvolyy-

miin vaikuttaa suuresti yksilöiden kiinnostuksenkohteet kun taas talouden ympärillä liikkuva informaatio koskettaa suurempaa osaa käyttäjistä kiinnostuksenkohteista riippumatta. Talouden informaation diffuusioon vaikuttaa suuresti esimerkiksi hallitseva työttömyystilanne sekä taloudelliset lasku- ja nousukaudet. [18]

Politiikan ja talouden aihepiirien osalta voidaan huomata, että kokonaisvolyymi saavuttaa pieniä kasvukohtia jatkuvasti. Tämä johtuu aihepiirien informaatioiden jatkuvuudesta. Esimerkiksi politiikan puolueelliset asiat soteuudistuksista ja talouden työttömyystilanteet ovat jatkuvia ja alati kehittyviä aiheita. Viihteen ja urheilun aiheet keskittyvät tyypillisesti yksittäisiin ja ajankohtaisiin tapahtumiin, jolloin pienet kasvukohdat kokonaisvolyymissa jäävät saavuttamatta.

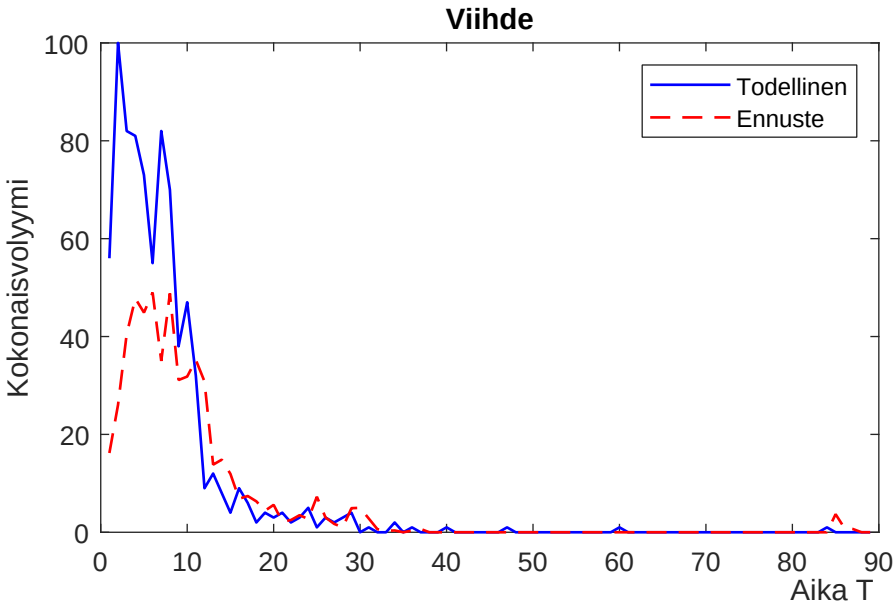


Kuva 5.4. *Urheilun aihepiirin fraasien kokonaisvolyymien ennusteet ja todelliset arvot*

Urheilun fraasien huippu saavutetaan muista aihealueista poiketen noin 10 tunnin kohdalla, kuten kuvasta 5.4 voidaan nähdä. Lisäksi kuvasta 5.4 huomataan kokonaisvolyymin laskeminen lähes nollaan noin 30 tunnin jälkeen. Koska tarkasteltavana on sosiaalisessa mediassa tapahtuva informaation leviäminen, niin voidaan perinteisinä medioina nähdä olevan niiden sosiaalisen median sivut, kuten Iltalehden Instagram-tili. Nämä perinteiset mediat muodostavat alle 15 prosentin osuuden informaation välittäjinä urheilun aihepiirissä. Urheilun aihepiirin informaation leviämisessa avainroolia näytteleekin yksilöiden, kuten sosiaalisen median vaikuttajien, kyky levittää informaatiota. [19]

Tässäkin aihealueessa kaksiasteisen flow-mallin ideologia on huomattavissa. Ensin informaation adoptoivat urheilun parissa työskentelevät vaikuttajat, jotka lähtevät levittämään informaatiota eteenpäin verkostolleen. Muiden käyttäjien adoptoidessa informaation, saavuttaa kokonaisvolyymi huippuarvonsa. Urheilun aihepiirin fraasien kokonaisvolyymi laskee lisäksi melko nopeasti lähelle nollaa. Tämä johtuu urheiluun liittyvien informaatioiden

luonteesta: uusia otteluita, kilpailuita ja urheilusuorituksia järjestetään päivittäin.



Kuva 5.5. *Viihteen aihepiirin fraasien kokonaisvolyymin ennusteet ja todelliset arvot*

Viimeisessä kuvassa 5.5 tarkastellaan kokonaisvolyymin ennusteita ja todellisia arvoja viihteen aihepiirin fraasien keskuudessa. Viihteen aihealueen informaation leviämisessä korostuu vahvasti ajankohtaisuus ja tuoreus. Informaatioiden huippu saavutetaan tyypillisesti nopeasti, mutta informaatioiden elinikä on on jokseenkin lyhyt. Viihteen saralla aiheita adoptoidaan tyypillisesti nopeammin. Aiheet keskittyvät tyypillisesti julkisten henkilöiden, kuten urheilijoiden, poliitikkojen tai näyttelijöiden, yksityiselämiin ja saavutuksiin. Tämän takia uutta informaatiota saadaan jatkuvasti, jolloin yksittäisen informaation elinikä jää lyhyeksi, eikä mainintoja ajan kuluessa enää ilmene yksittäisiä poikkeuksia lukuunottamatta.

Aihepiiri	RMSE	Keskiarvo todellinen	Keskiarvo ennuste
Politiikka	10,35	8,76	8,94
Talous	6,22	8,83	7,94
Urheilu	8,90	5,45	6,30
Viihde	12,72	9,08	6,20

Kuva 5.6. *Keskineliövirheen neliöjuuri (RMSE), todellisten arvojen ja ennusteiden kokonaisvolyymien keskiarvot*

Kuvassa 5.6 on kuvattuna jokaisen aihealueen keskineliövirheen neliöjuuri sekä todellisten ja ennusteiden keskiarvo. Tuloksista voidaan päätellä, että suurempia kokonaisvolyymin arvoja saavuttamat aihepiirit aiheuttavat suuremman keskineliövirheen neliöjuuren. Todellisten ja ennusteiden keskiarvot kuvaavat kokonaisvolyymien keskiarvoja. Keskineliövirheen neliöjuuri sekä keskiarvojen erotus viihteen aihepiirissä on suurin. Tämä ilmiö

voidaan selittää datan luonteella, jossa viihteen fraaseja oli kerättynä huomattavasti vähemmän kuin muiden aihepiirien fraaseja. Nämä tulokset toimivat yhteenvetona jo aikaisemmin perustelluille tuloksille.

6. YHTEENVETO

Kandidaatintyössä tutkittiin matemaattisten mallien tarjoamia mahdollisuuksia sosiaalisen median informaation leviämisen ennustamisessa. Päätutkimusongelmaa, miten sosiaalisen median informaation leviämistä voidaan ennustaa matemaattisen mallinnuksen avulla, lähestyttiin aluksi tutkimalla lineaarisen vaikutusmallin peruselementtejä. Mallinnusta lähestyttiin parametrisella arvioinnilla, jossa yhdisteltiin laajasti erilaisia matematiikan osa-alueita.

Hyödyntämällä tutkimusdataa, saatiin ratkaistua lineaarisen vaikutusmallin ennustevirhe, joka todisti lineaarisen vaikutusmallin soveltuvuuden informaation leviämisen ennustamisen tarkasteluun. Viimeisessä osassa työtä esiteltiin koko dataa hyödyntämällä saatu influenssifunktio. Lisäksi vertailtiin lineaarisen vaikutusmallin antamia ennusteita todellisiin arvoihin aihepiireittäin.

Työn tulokset olivat selkeitä ja ne onnistuivat vastaamaan asetettuun päätutkimusongelmaan. Kandidaatintyön avulla onnistuttiin luomaan yhtenäinen kirjallisuuskatsauksen ja matemaattisen tarkastelun kokonaisuus. Haasteita työn toteuttamiselle tuotti lineaarisen vaikutusmallin uutuuden takia ilmennyt tieteellisen lähdemateriaalin vähyys. Työn aiheen ja laajuuden vuoksi tiettyjä osa-alueita päädyttiin rajaamaan.

Lineaarisen vaikutusmallin tarkastelu kirjallisuudessa on vähäistä. Tämän seurauksena matemaattisen mallintamisen osiossa lähteinen määrä on vähäinen. Ottaen huomioon systemaattisesti toteutetun tieteellisesti laadukkaan kirjallisuuden hakuprosessin, voidaan tutkimuskentästä todeta löytyneen jatkotutkimuksille olennainen aukko. Näiden jatkotutkimusten avulla voitaisiin edistää konkreettisten mittarien luomista sosiaalisen median täyden potentiaalin saavuttamiseksi. Näitä matemaattisesti mallinnettuja mittareita voitaisiin käyttää esimerkiksi yritysten markkinoinnin työkaluina.

LÄHTEET

[1] Tranfield, D., Denyer, D. ja Smart, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management* 14.3 (2003), s. 207–222.

[2] Salminen, A. *Mikä kirjallisuuskatsaus? - Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin*. Vaasan Yliopiston Julkaisuja, 2011. 44 s.

[3] Snyder, H. Literature review as a research methodology: An overview and guidelines. *Journal of business research* 104 (2019), s. 333–339.

[4] Leskovec, J. *96 million memes from Memetracker*. URL: http://memetracker.org/data/memetracker9.html (viitattu 23. 03. 2021).

[5] Lietsala, K. ja Sirkkunen, E. *Social media.Introduction to the tools and processes of participatory economy*. University of Tampere, 2008. 197 s.

[6] O'Reilly, T. *What is Web 2.0*. 2005, s. 1–12. URL: https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=all (viitattu 27. 01. 2021).

[7] Wolcott, M. *What is Web 2.0?* 2008. URL: https://www.cbsnews.com/news/what-is-web-20/ (viitattu 27. 01. 2021).

[8] Moens, M.-F., Li, J. ja Chua, T.-S. *Mining User Generated Content*. Taylor Francis, 2014. 328 s.

[9] Berthon, P. R., Pitt, L. F., Plangger, K. ja Shapiro, D. Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons* 55.3 (2012), s. 261–271.

[10] Kaplan, A. ja Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. 53.1 (2010), s. 59–68.

[11] Karjaluoto, H. *Digitaalinen markkinointiviestintä: esimerkkejä parhaista käytännöistä yritys- ja kuluttajamarkkinointiin*. Docendo, 2010.

[12] Leskovec, J. ja Yang, J. Modeling Information Diffusion in Implicit Networks (2010), s. 599–608. DOI: 10.1109/ICDM.2010.22.

[13] Coleman, T. F. ja Li, Y. A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on Some of the Variables. eng. *SIAM journal on optimization* 6.4 (1996), s. 1040–1058.

[14] Haralabopoulos, G., Anagnostopoulos, I. ja Zeadally, S. Lifespan and propagation of information in On-line Social Networks: A case study based on Reddit. *Journal of network and computer applications* 56 (2015), s. 88–100.

[15] Halberstam, Y. *Homophily, Group Size, and the Diffusion of Political Information in Social Networks : Evidence from Twitter*. eng. National Bureau of Economic Research, 2014.

[16] Postelnicu, M. *Two-Step Flow Model of Communication*. eng. 2008.

[17] Leeper, T. J. ja Slothuus, R. Political Parties, Motivated Reasoning, and Public Opinion Formation. *Political psychology* 35.1 (2014), s. 129–156.

[18] Fogarty, B. J. Determining economic news coverage. *International journal of public opinion research* 17.2 (2005), s. 149–172.

[19] Vermeer, S. ja Araujo, T. Keep the ball rolling: Information diffusion within large sports-related networks through social mediators. *Communication and sport* 8.6 (2020), s. 803–824.