NIFU

Innsikt | NR. 4 – 2024

Policy Brief 24.04.2024

Kunstig intelligens: Supplement eller erstatning for kunnskapsoppsummeringer?

Sabine Wollscheid, Henrik Karlstrøm og Lone W. Fossum. Forskningsgruppe bibliometri og kunnskapsoppsummeringer

Ny teknologi gjør det mulig å effektivisere og automatisere deler av forskningsprosessene. Det gjelder også arbeidet med å syntetisere resultater fra forskning, såkalt kunnskaps-oppsummering. Digitale verktøy som kunstig intelligens kan derimot ikke erstatte den kompetansen og kunnskapen som trengs for å gjøre gode kunnskapsoppsummeringer i samfunnsvitenskap. Men de kan effektivisere prosessen, så lenge de brukes riktig og i kombinasjon med fag- og metodeekspertise.

I dette notatet drøfter vi hvordan ekspertise på kunnskaps-oppsummering kan spille sammen med ny teknologi og digitale verktøy.

Digitalisering av informasjon har bidratt til en voldsom vekst i internasjonale forskningspublikasjoner.

Denne trenden vil fortsette som følge av ny teknologi som bidrar til å effektivisere deler av forskningsprosessen. Et eksempel er store språkmodeller som ChatGPT. De kan for eksempel bistå med spørsmålsformulering, og identifisering og sammenstilling av store mengder informasjon. Når mengden forskningslitteratur øker, øker også behovet for kunnskapsoppsummeringer. Slike oppsummeringer brukes i økende grad også innen samfunnsvitenskap for å informere politiske beslutninger, for eksempel innen forskning-, utdanning- og arbeidsmarkedspolitikk.

Når oppsummeringer av kunnskap brukes som grunnlag for politiske beslutninger må de utføres på en transparent og redelig måte og unngå systematiske feil. Det er også viktig å ha et bevisst forhold til hva slags formål oppsummeringene skal tjene:

- Noen kunnskapsoppsummeringer kan ha en bred rekkevidde og adressere mer åpne problemstillinger, såkalte systematiske kartlegginger (Munthe et al., 2022). Et eksempel på dette er å kartlegge hva som finnes av nordisk forskning om diskriminering, trakassering og likestilling (Aksnes et al., 2021), eller nordisk forskning på vurderingssystemet av eksamen (Hovdhaugen et al., 2022). Slike kartlegginger oppsummerer som regel ikke funnene fra de inkluderte studiene.
- Andre kunnskapsoppsummeringer er tenkt å besvare mer lukkede og spisse problemstillinger om effekt av tiltak, slik som effekter av opplæringstilbud for tospråklige elever og kompetansehevingstiltak for voksne innvandrere (Wollscheid et al. 2017).
- I tillegg finnes det såkalte metasynteser som oppsummerer kvalitativ forskningslitteratur, for eksempel ved å sammenlikne likheter og forskjeller i oppfatninger og erfaringer for å finne essensielle egenskaper ved et fenomen. (Munthe et al., 2022).

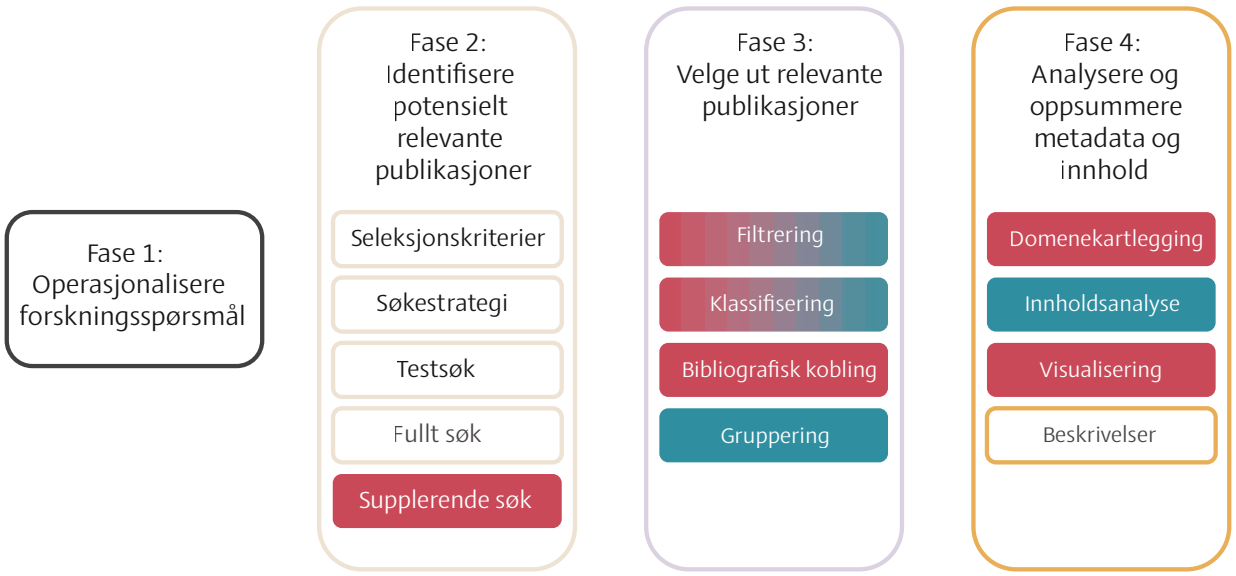
Hurtigoversikter og raske kunnskaps-oppsummeringer kan være problematisk

Det er to argumenter for å ta i bruk maskinlæring og språkmodeller i kunnskaps-oppsummeringer: For det første kan det redusere tid, kostnader og arbeid. For det andre kan slike metoder gi raskere tilgang til oppsummert kunnskap for å informere politikkutforming (European Centre for Disease Prevention and Control, 2022). Men flere studier har pekt på at automatisering av prosessen kan påvirke kvaliteten på forskningsoppsummeringer (de la Torre-Lopés, et al. 2023; Marshall et al., 2019). Dette skyldes blant annet at bruk av maskinell tekstbehandling i mange tilfeller gjør prosessen mindre transparent og reproduserbar, også for dem som gjennomfører kunnskapsoppsummeringen. Det kan også argumenteres for at automatisering av oppsummeringen strider mot idealene for metodisk stringens (de la Torre-Lopés, et al. 2023; Marshall et al., 2019). Dette gjelder også for hurtigoversikter som er basert på anerkjente vitenskapelig retningslinjer og standarder for å identifisere, analysere og vurdere et utvalg av eksisterende forskningsbidrag for å besvare en bestemt problemstilling (Munthe et al., 2022).

Det er også forskjeller mellom fagområder når det gjelder på dette området. Bruk av store språkmodeller i screeningen og datauttrekk kan for eksempel egne seg for i fag som preges av standardisert og kodifisert språk i fagets vitenskapelige publikasjoner, slik som biomedisin. Innenfor samfunnsvitenskap er dette derimot mer utfordrende, selv om det er store forskjeller også innenfor samfunnsvitenskap, og det har skjedd en automatisering i forskningsprosessen i store deler av dette fagområdet også. I det følgende vil vi derfor rette fokus mot fordeler og utfordringer med å ta i bruk ny teknologi og verktøy for kunnskapsoppsummeringer innenfor samfunnsvitenskapene.

Hvordan kan vi bruke ny teknologi og verktøy i kunnskapsoppsummeringer i samfunnsvitenskap, og samtidig opprettholde god kvalitet?

I figuren under illustrerer vi de fire hovedfasene av en kunnskapsoppsummering, som vi i det videre skal beskrive nærmere.



Fase 1: Operasjonalisering av forskningsspørsmål

En kunnskapsoppsummering starter med at forskningsspørsmål operasjonaliseres. Her kan store språkmodeller som ChatGPT bidra med å generere eller revidere forslag til forsknings-spørsmål og konsepter. Derimot kan det være tidkrevende å lage gode spørringer som bidrar til at man får nyttige svar fra språkmodellen. På det nåværende tidspunkt er store språk-modeller neppe i stand til å erstatte sparring med andre personer. Store språkmodeller kan dermed brukes til inspirasjon og idégenering, men dette krever innsikt i behovene man har med tanke på å operasjonalisere problemstillinger. Dette innebærer fagekspertise, slik at man kan luke ut eventuelle «irrelevante» forslag med hensikt til avgrensning og spesifisering av problemstillinger.

Fase 2: Identifisere relevante publikasjoner

Systematiske søk i flere databaser kan være en tidkrevende prosess fordi det ofte innebærer å teste ut begreper, både alene og i kombinasjon med andre og med og uten boolske operatører for å undersøke om de treffer riktig (f.eks., AND, OR, NOT). I tillegg må søkestrenger tilpasses de ulike databasene man bruker. Denne iterative utprøvingen er viktige fordi mangelfulle søkestrenger kan utelate sentrale forskningsbidrag og skape skjevhet i datamaterialet.

Ettersom kunnskaps-oppsummeringer bygger på de identifiserte publikasjonene (datamaterialet), er det avgjørende at man finner riktige søkebegreper, synonymer og andre avgrensninger. For å oppnå et representativt datautvalg er det viktig å finne den rette balansen mellom presisjon av søkestrengen og kompletthet («recall») (Buckland & Gey, 1994). Språkmodeller har et potensial for å effektivt genere fungerende og komplekse søkestrenger tilpasset de enkelte databasene. Studier av søkestrenger fra ChatGPT viser at den kan raskt redigere søk fra tidligere og automatiske søkestrenger og lage søkestrenger med høyere presisjon, men at dette kan gå utover bredden eller kompletthetен i søketrefflisten (Qureshi et al., 2023; Wang et al., 2023). I tillegg kan språkmodellens utvelgelse av begreper være utfordrende å granske, ettersom begrunnelsen for genererte begrepslister ikke presenteres av språkmodellen. Mangelfull transparens for utvelgelse av søkeord er også utfordrende i de tilfeller der like spørringer til språkmodellen gir *ulike* søkestrenger (Wang et al., 2023). Det er også verdt å merke seg at ChatGPT ikke har tilgang til data utover det som var synlig for dem på tidspunktet de ble trent opp, og ikke kan teste søkestrengene før de gis. Det er derfor viktig å teste ut de automatisk genererte søkestrengene selv.

På grunnlag av eksisterende forskning kan det konkluderes med at mange store språkmodeller som ChatGPT ikke innfrir kravet om transparens i systematiske kunnskapsoppsummeringer. De kan bidra til å utforme søkestrenger og kan tilpasse dem ulike databaser, men har ingen iboende egenskap til å vurdere kvaliteten på det som kommer ut. Det er derfor viktig med manuell validering av alle steg i prosessen, dvs. at man velger ut spesifikke begreper, slik at man vet hvorfor de ulike begrepene er med og i hvilken sammenheng.

Fase 3: Velge ut relevant forskningslitteratur

For å velge den relevante forskningslitteraturen som skal inngå i datamaterialet i kunnskapsoppsummeringer kan man i første omgang gjøre en såkalt screening, som innebærer at forskningsstudienes tittel og sammendrag blir lest for å vurdere om den skal inkluderes eller ikke. Screeningarbeidet er tid- og ressurskrevende, spesielt hvis man skal innfri kvalitetskravet om at to forskere screener materialet uavhengig av hverandre (dobbel screening). Verktøyene kan imidlertid potensielt redusere antall personer som trenger å være involvert i screeningen (Marshall & Wallace, 2019). Et velkjent eksempel er Elicit som tilbyr filtreringsmuligheter for å redusere søkeresultatene, å tilføye nøkkelord og filtrere etter studietype (Whitfield & Hofmann, 2023).

Med tanke på utvelgelse eller screening av relevant forskningslitteratur har automatisering kommet langt for å gjøre en del av jobben mer effektiv. Blant annet kan maskinlæring bidra til å sortere litteraturen etter relevans basert på læring av manuell screening av et utvalg. I tillegg til å arbeide raskere enn mennesker, kan verktøyet også screene mer objektivt – så fremt utvalget den baserer seg på er konsekvent. Jo mer standardisert språket i en fagdisiplin er, jo større er potensialet for en slik effektivisering. For samfunnsvitenskap, som har et lite standardisert språk, kan den automatiske screeningen fungere *dårligere* enn manuell screening når sentrale begreper brukes ulikt. Teknologien ser etter gjengående sammenhenger i tekstmaterialet, men kan ikke lese konteksten slik mennesker kan.

Det er mange verktøy for å effektivisere screeningsprosessen. Felles for screeningsverktøyene er at de kan redusere tiden man bruker på screeningen, men det trengs fortsatt forskerens fagekspertise for at utvelgelsen skal bli korrekt.

Fase 4: Analysere og oppsummere forskningslitteratur

Å analysere og oppsummere forskning anses som kjernen i en systematisk kunnskapsoppsummering. Dette er også den mest krevende fasen, ettersom avanserte tematiske synteser og oppsummeringer av mange studier er utfordrende. Kunstig intelligens og maskinlæring kan bidra til å effektivisere denne fasen. Tiden det tar å velge ut informasjon og lage en oppsummering kan potensielt reduseres ved å bruke teknologi som genererer automatisk innholdsanalyse gjennom å trekke ut tekst fra dokumenter og identifisere kjernebegreper, ord og temaer. Studier finner lovende forsøk i å identifisere og oppsummere relevant informasjon basert på et utvalg av tre sammendrag ved hjelp av verktøyet Elicit (Whitfield & Hofmann, 2023). En av styrkene ved denne type automatisk innholdsanalyse er å unngå å rette søkelys på anekdotisk bevis ved å trekke tekst fra store mengder dokumenter og identifiserer kjernebegreper og temaer (Smith & Humphreys, 2006). Ut over det kan økt automatisering understøtte forskersamarbeid i kunnskapsoppsummering på tvers av institusjoner og land, oversettelse av forskningsbidrag til andre språk enn engelsk og gi bedre muligheter for løpende oppdateringer av for eksempel en kartlegging av forskning som oppdateres løpende (European Centre for Disease Prevention and Control, 2022).

Samtidig advares det om at ukritisk bruk av slike verktøy kan skape systemfeil. GPT-4.0 viste noen små tekniske forbedringer i denne fasen sammenlignet med den tidligere versjonen, men ingen forbedringer i andre faser (Qureshi et al., 2023). I automatisk generert innholdsanalyse kan dessuten informasjon om konteksten forsvinne. Whitfield og Hofmann (2023) påpeker at verktøyet Elicit har samme begrensninger i presisjon som andre verktøy. Mens Elicit utfører enkle oppgaver basert på store språkmodeller, er resultatene ikke alltid presise. Dette betyr at forskeren fortsatt trenger å sjekke presisjon av de oppnådde resultatene, noe som forutsetter fagkunnskap og tid. Det påpekes at Elicit ikke er i stand til å utføre kognitive funksjoner på høyt nivå som er nødvendig til å skape en forståelse av, og å syntetisere, litteraturen. Whitfield og Hofmann (2023) konkluderer med at verktøy som Elicit ikke kan erstatte forskningsaktiviteten ved en forskningsoppsummering, men erstatte noen av de enkle rutineoppgavene i prosessen.

Når det gjelder analyse og oppsummering av forskningslitteratur kan analyseverktøy bidra til å minske tiden det tar å gå gjennom store mengder forskningslitteratur, men det trengs i tillegg faglig ekspertise for å veilede analyseverktøyet gjennom prosessen (Marshall & Wallace, 2019).

Nye teknologier og verktøy er først og fremst nyttige verktøy for å effektivisere *prosessen* med kunnskapsoppsummeringer

Nye teknologier og digitale verktøy er i ferd med å endre måten å forske på og oppsummere forskning. Per i dag ser vi at det er mange fordeler med å ta i bruk ny teknologi for å oppsummere kunnskap. Tradisjonelt har kunnskapsoppsummeringer vært svært tids- og ressurskrevende å gjennomføre, men automatisering og kunstig intelligens kan bidra til å redusere tid og kostnader. Men som vi har skissert ovenfor, kan nye teknologier og verktøy ikke erstatte faglig ekspertise og metodekunnskap. De kan derimot effektiserte noen rutineoppgaver i forskningsprosessen, og på den måten frigjøre tid til dybdeanalyser og mer krevende prosesser i arbeidet med kunnskapsoppsummeringer.

Samtidig er det viktig å være bevisst på hva denne automatiseringen konkret innebærer for hver enkelt fase i en systematisk kunnskapsoppsummering. Automatiseringen har kommet lenger for noen delprosesser enn for andre, som for utformingen av litteratursøk og i screeningprosessen. Automatiseringen fungerer også bedre for noen fagområder enn for andre, da spesielt for fagdisipliner med et mer standardisert begrepsapparat, slik vi finner i for eksempel medisin, helse- og naturvitenskap sammenlignet med samfunnsvitenskap (og humanioria også). Automatisering av metasynteser, dvs. oppsummering av kvalitativ forskningslitteratur som er mer vanlig innen samfunnsvitenskap og humanioria, vil derfor være mer utfordrende enn oppsummeringer av kvantitativ forskningslitteratur med et relativt standardisert begrepsapparat og mindre metodemangfold som er mer utbredt innen medisin, helse- og naturvitenskap. Automatisert tematisk innholdsanalyse kan også egne seg for å kartlegge store datamengder og beskrive et forskningsfelt «ovenfra». Likevel kan ikke disse prosessene utføres av et verktøy alene. Verktøyene fungerer best når de blir veiledet av noen med faglig og metodisk ekspertise.

Vi advarer derfor mot lettvint og ukritisk bruk av ny teknologi og nye verktøy

Utviklingen av nye verktøy og teknologier skjær raskt, samtidig er det er mange alternativer av slike verktøy å velge inni mellom. Lettvint og ukritisk bruk av slikt frarådes å tas i bruk i kunnskapsoppsummeringer, særlig når det ikke har blitt testet ut og evaluert tilstrekkelig av forskningssamfunn. Et ukritisk valg og ukritisk bruk av et verktøy kan introdusere systema-

tiske skjevheter i prosessen som man nettopp ønsket å minimere med selve metoden. Det gjelder spesielt verktøy hvor de indre funksjonene er ukjent og hvor man ikke vet hva som ligger til grunn for materialet man får.

Systematisk kunnskapsoppsummering som metode er basert på strenge vitenskapelig standarder som repliserbarhet, transparens, validitet og overførbarhet på like linje med forskning ellers. Dersom man tar i bruk verktøy som tar selvstendige valg, uten at man kjenner til datagrunnlaget og kriteriene for valgene, er det motstridende med metodikkens idealer og kan underminere funnene til oppsummeringen. Det anbefales derfor å bruke tid i forkant for å velge ut verktøy som har blitt testet ut med tanke på forskningsfeltets egenart og de ulike fasene i forskningsprosessen. Akkurat som i fase 1, som legger grunnlaget for hva man vil finne og basere kunnskapsoppsummeringen på, er forarbeidet med å lære seg muligheter og begrensninger ved verktøyene man tar i bruk avgjørende for sluttresultatet. Tiden det brukes for å velge et egnet verktøy i forkant kan dermed være like viktig eller enda viktigere enn tiden man sparer ved å bruke det i prosessen.

Ny teknologi og verktøy basert på kunstig intelligens er et forskningsfelt i dynamisk og rask utvikling. Selv om verktøyene ofte er lett å bruke, kan det være vanskelig å forstå hva som ligger bak funksjonene, i «black boksen». For å opprettholde høy vitenskapelig kvalitet i systematiske kunnskapsoppsummeringer, fordrer det høy bevissthet om hvordan man bruker ulike digitale verktøy i forskningsprosessen. Det fordrer også en forståelse av fordeler og svakheter ved verktøyene. På nåværende tidspunkt vil kunstig intelligens og maskinelle programmer hovedsakelig tilby et supplement til – og ikke en erstatning av – manuelle prosesser i systematiske kunnskapsoppsummeringer innenfor samfunnsvitenskapen.

Referanser

Aksnes, D. W., Bergene, A. C., Fossum, L. W., Wollscheid, S. (2021). Nordisk forskning om diskriminering, trakassering og likestilling: En forenklet kunnskapsoversikt. Nordisk institutt for studier av innovasjon, forskning og utdanning NIFU. Rapport.

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12-19.

de la Torre-López, J., Ramírez, A., & Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 1-24.

European Centre for Disease Prevention and Control. Use and impact of new technologies for evidence synthesis. Stockholm: ECDC; 2022.

Hovdhaugen, E., Flobakk-Sitter, F., Wollscheid, S., Fossum, L. W. & Korseberg, L. (2022) Kartlegging av nordisk forskning på eksamen. NIFU. Rapport.

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8, 1-10.

Munthe, E., Bergene, A. C., Braak, D. t., Furenes, M. I., Gilje, T. M., Keles, S., Ruud, E., & Wollscheid, S. (2022). Systematisk kunnskapsoppsummering utdanningssektoren. *Norsk Pedagogisk Tidsskrift*(2), 131-144.

Qureshi, R., Shaughnessy, D., Gill, K. A., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic reviews*, 12(1), 72.

Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior research methods*, 38, 262-279.

Wang, S., Scells, H., Koopman, B., & Zuccon, G. (2023). Can ChatGPT write a good boolean query for systematic review literature search? *arXiv preprint arXiv*:2302.03495.

Whitfield, S., & Hofmann, M. A. (2023). Elicit: AI literature review research assistant. *Public Services Quarterly*, 19(3), 201-207.

Wollscheid, S., Flatø, M., Hjetland, H. N., Smette, I. (2017). Effekter av opplæringstilbud for tospråklige elever og kompetansehevingstiltak for voksne innvandrere : En kunnskapsoversikt. NIFU. Rapport.

NIFU

Nordisk institutt for studier av innovasjon, forskning og utdanning

Nordic Institute for Studies in Innovation, Research and Education

NIFU er et uavhengig samfunnsvitenskapelig forskningsinstitutt som tilbyr handlings- og beslutningsorientert forskning til offentlig og privat sektor. Forskningen omfatter hele det kunnskapspolitiske området – fra grunnopplæring, via høyere utdanning til forskning, innovasjon og kompetanseutvikling i arbeidslivet.

NIFU
PB 2815 Tøyen, NO-0608 Oslo
www.nifu.no | post@nifu.no

NIFU-Innsikt
ISSN 2704-0771