Kan vi identificere fingeraftrykket fra generativ kunstig intelligens i en dansk kontekst?

En analyse af OpenAIs tekstklassifikators utilstrækkelighed med 15 danske politikeres Facebooksider som case

Analyse & Tal

Analyse & Tal, marts 2023

Udarbejdet af

Analyse & Tal F.M.B.A
Hejrevej 34A
2400 København NV
www.ogtal.dk

For mere information kontakt

Edin Lind Ikanovic
Tlf. +4530134787
edin@ogtal.dk

el.

Mikkeline Thomsen
Tlf. +4561607302
Mikkeline@ogtal.dk

Databehandling, analyse & tekst

Edin Lind Ikanovic
Mikkeline Thomsen

INDHOLD

Politikeres og borgeres brug af AI-genereret indhold på Facebook – ifølge OpenAI

I denne analyse vil vi illustrere, hvordan kunstigt intelligente værktøjer som ChatGPT og GPT4 endnu mangler autenticitetsværktøjer, der kan tøjle dem. Vi har analyseret 2.844 opslag fra de mest aktive politikeres sider på Facebook ved hjælp af OpenAIs AI Text Classifier[1], et værktøj der er udviklet til at detektere, om en tekst er skrevet af en kunstig intelligens eller ej. Resultaterne viser, at det i skrivende stund er umuligt at fælde dom over politikeres såvel som borgeres brug af AI-genereret tekst på dansk, medmindre vi tror på, at mindst 2/3 af alt indhold på politikernes facebookssider er genereret af AI. Resultaterne inviterer til en spændende diskussion: Hvordan undgår vi, at vores nyeste teknologiske redning ikke viser sig at være en Pandoras æske af ulykker, som er svær at få kontrol med?

Baggrund

Indholdsgeneratorer tager verden med storm

Halvvejs inde i 2020 lancerede OpenAI den kunstige intelligens GPT-3, og internettet eksploderede. En kunstig intelligens trænet på stort set alt offentligt indhold på internettet, der kunne komponere musik, skrive kode og generere gruk af Piet Hein med så høj kvalitet, at det kan være svært at skelne computerens arbejde fra menneskets. Hvilket af følgende gruk, er f.eks. hr. Heins værk, og hvilket er GPT-3's?[2]

Det kræver lidt erfaring,
At være pessimist
Hvis du vil leve til gamle dage,
Så er det klogt
At være optimist

Den virkelige vise
Er den som formår
At forstå også den
Som han ikke forstår

Siden 2020 er generativ AI kun blevet bedre, og senest har OpenAI lanceret efterfølgerne ChatGPT og GPT4, der på ny har forbløffet verden med deres imponerende resultater. Mange er begyndt at eksperimentere med generativ AI's potentiale til alt fra hjælpsomme chatbots, mere effektiv forvaltning, kunst og arbejdsassistenter. Microsoft og Google kæmper en kamp om hurtigst muligt at integrere værktøjerne i deres eksisterende søgemaskiner og andre produkter. Flest bruger foreløbig teknologien som intelligent søgemaskine og faglig sparringspartner. Skræddersyede overbygninger gør det nemmere at bruge teknologien til specifikke formål, f.eks. så hele to nye danske AI-baserede værktøjer til at skrive jobansøgninger: https://www.jobbutler.ai/ og https://20sekunder.dk/.

Flere er dog også gået ind i samtalen om de potentielle udfordringer ved, at intelligente maskiner som ChatGPT og GPT4 blev sluppet ud i verdenssamfundet. Hvordan har undervisere mulighed for at vide, om opgaver og regnestykker er udarbejdet af de studerende selv? Hvordan sikrer vi, at ChatGPT ikke bruges

[1] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[2] Svaret findes her: https://nyheder.tv2.dk/tech/2020-09-30-ny-kunstig-intelligens-testes-i-danmark-kan-du-kende-forskel-paa-menneske-og-maskine

til at væbne bothære og falske profiler på sociale medier med autogeneret politisk indhold, der er umuligt at skelne fra autentiske input til den offentlige debat?
Allerede nu har vi set eksempler på spredning af disinformation og hadtale i en sådan skala, at vi må formode, at generativ AI har været anvendt i processen[3].

Den nye teknologi giver allerede helt konkrete udfordringer for vores lærere. I digitale fora for undervisere kommer flere og flere medlemmer med eksempler på, at de modtaget eksamensbesvarelser, som de mistænker, er genereret af kunstig intelligens. Det følgende udsnit er en anonymiseret og parafraseret samtale fra et forum for undervisere:

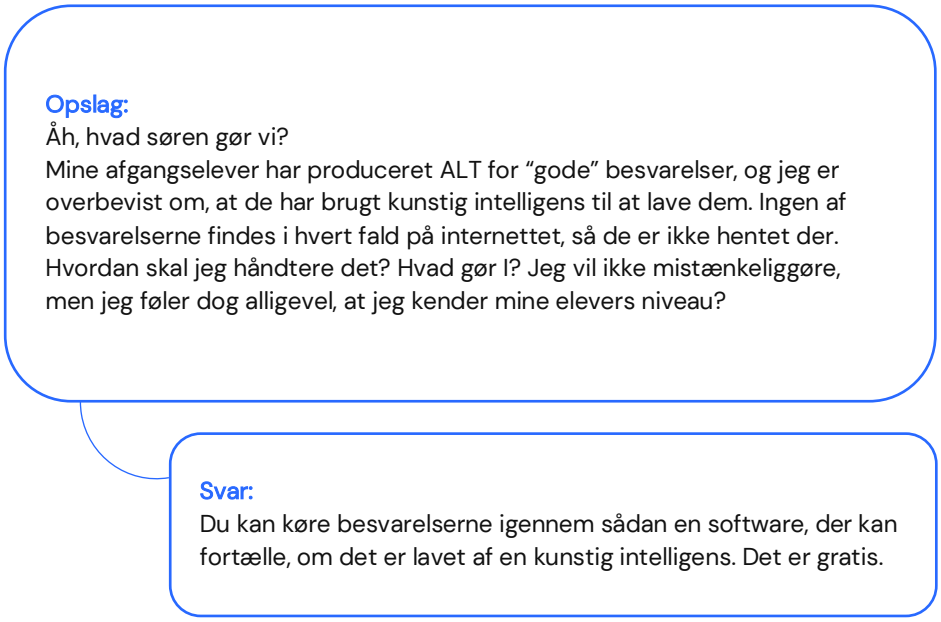


Illustration 1: Udsnit af de verserende debatter i digitale fora for lærere

Debatterne der pt udspiller sig i lærernes faglige sparringsfora vidner om konkrete og aktuelle problemer for en hel sektor, men de vidner også om, at vi allerede er begyndt at anvende og stole på løsninger, som vi ikke ved om virker. Følgende analyse vil argumentere for, at lærerne (og andre) foreløbig bør holde hestene i forhold til at anvende kunstig intelligens vurderinger til forfatteridentifikation.

Værktøjet, der skal læse dens fingeraftryk

OpenAI anerkender problemet. Men hvad gør vi? Æsken er jo åbnet, og teknologien frigivet. Én oplagt mulighed for at begrænse generativ AI er at tvinge teknologien til at bekendte kulør. OpenAI har derfor lanceret et værktøj kaldet AI Text Classifer[4], på dansk AI-tekstklassifikator, der skal afhjælpe det problem. Og så skulle man tro, at dét problem var løst. Hvis det er relativt nemt at finde ud af, om et stykke indhold er formuleret af et menneske eller en kunstig intelligens, tager det brodden af faren ved en rabiat politisk smædekampagne eller et autogenereret speciale.

Man skulle tro, at de kloge hoveder bag chatGPT relativt nemt kunne få deres intelligente teknologi til at pege indad – at teknologien ville have nemt ved at genkende sin egen skrivestil. Men intet tyder på det.

[3] https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan

[4] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/

Tekstklassifikatoren virker kun på engelsk og kan kun anvendes på tekster med mere end 1.000 tegn. Og selv her har den kun en succes rate på 26%[5]. Falske positiver (altså tilfælde hvor tekstklassifikator fejlagtigt estimerer, at en tekst skrevet af et menneske, er AI genereret) udgør ifølge OpenAI 9%.

Vi må derfor antage, at resultaterne vil være endnu mindre robuste på dansk. For det meste sprogteknologi gælder det, at virksomhederne ofte praler med flotte resultater, der dog mere end ofte kun gælder for engelsk indhold. Hvad end det handler om moderation af hadtale og disinformation eller detektion af AI-generet indhold er verdens fattigste og mindste sprogmarkeder altid bagerst i køen. Men hvor slemt? Er dette ene missilforsvar mod generativ AI overhovedet virksomt på dansk? Den følgende analyse er et første forsøg på en test i en dansk kontekst.

Metode

Sådan gjorde vi

Vi har analyseret 2.844 opslag fra de 15 mest aktive politiker- og partisider på Facebook. Opslagene er fra hhv. 2020 og 2023. De to år er valgt sådan, at dataudtrækket repræsenterer en periode før og efter at generativ kunstig intelligens blev et almindeligt kendt og potentielt anvendt værktøj. 894 af opslagene er slået op af *partiet* eller *politikeren* bag siden. 1.950 af opslagene er slået op af *besøgende* på politikerens eller partiets Facebookside. Der er altså både professionelt og civilt genereret indhold i datasættet.

Opslagene skal som sagt bestå af mere end 1.000 karakterer, for at tekstklassifikatoren kan komme med sit bud på, om teksten er AI-genereret eller menneskeskabt. De i alt 2.844 opslag udgør alle opslag fra siderne og sidernes besøgende i 2020 og 2023 på mere end 1.000 tegn. De 2.729 opslag er fra 2020. De 115 er fra 2023.

Tekstklassifikatoren fungerer således, at man kopierer en tekst på mere end 1.000 tegn ind i en tekstboks, og får returneret en vurdering af sandsynligheden for, at teksten er genereret af AI[6]. Vurderingen foretages på at klassisk 5-trins likert skala:

The classifier considers the text to be:	The classifier considers the text to be:	The classifier considers the text to be:	The classifier considers the text to be:	The classifier considers the text to be:
Likely AI-generated	Possibly AI-generated	Unclear if it is AI-generated	Unlikely AI-generated	Very unlikely AI-generated

Tabel 2: Open AI tekstklassifikatorens outputskala

Datagrundlaget for undersøgelsen:

Sidens navn	Opslag fra sideindehaveren	Opslag fra besøgende
Alternativet	23	32
Brian Nielsen Nye Borgerlige Thisted	27	0
Christian Juhl	13	1

[5] https://cobusgreyling.medium.com/testing-openais-new-ai-text-classifier-for-identifying-ai-written-content-7b2ec3c3a35

[6] https://platform.openai.com/ai-text-classifier

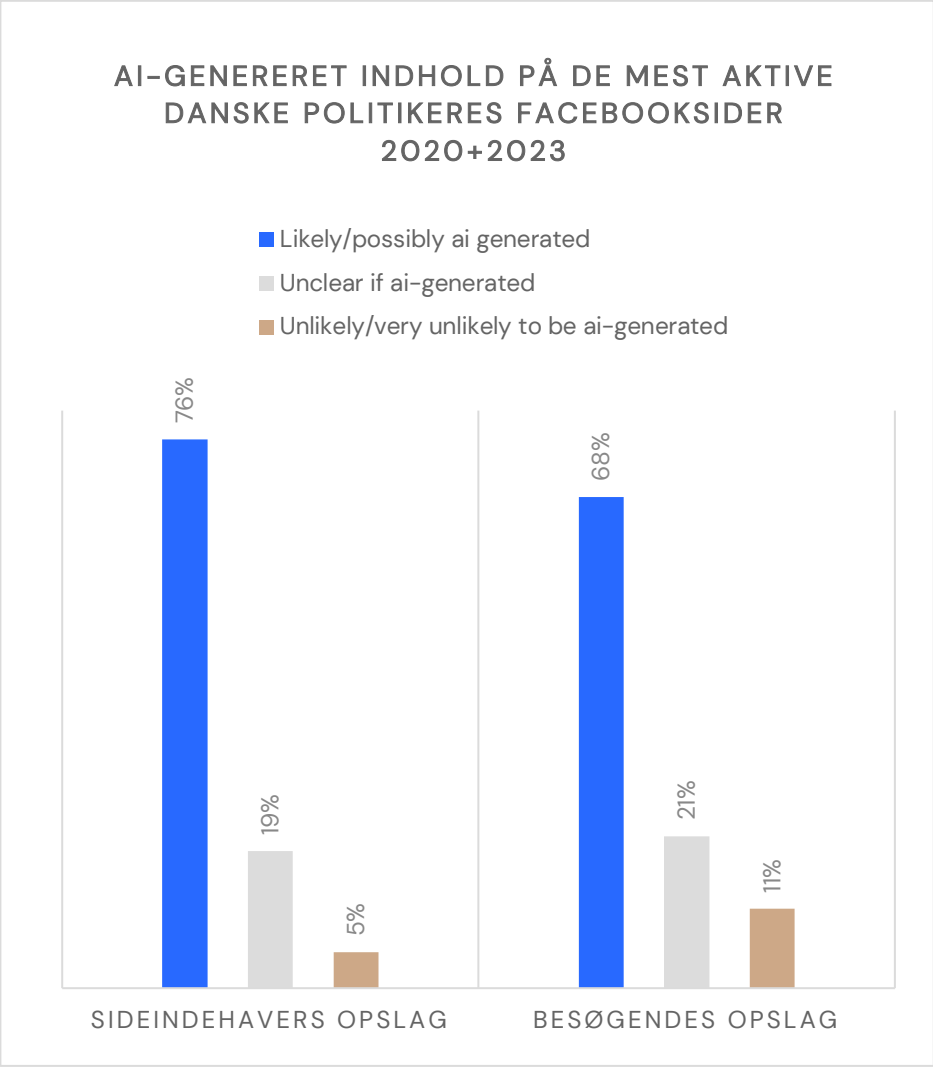
Enhedslisten	36	0
Hans Kristian Skibby	6	0
Kasper Sørmer-Sørensen	7	0
Mette Frederiksen	76	1778
Morten Messerschmidt	31	47
Naser Khader	50	0
Peter Sig Politik	308	0
Rasmus Prehn	52	6
Rasmus Stoklund	40	68
SF	22	0
Sysser Nørholm - Nye Borgerlige - Nordjylland	75	0
Victoria Velásquez	128	18

Tabel 3: Facebooksider og indhold inkluderet i undersøgelsen

Resultater

Ifølge tekstklassifikatoren er det mere reglen end undtagelsen, at politikerne har brugt AI til at generere indhold på Facebook

Ifølge OpanAIs tekstklassifikator er hele 76% af politikernes/partiernes opslag på Facebook i perioden "likely" eller "possibly" genereret af en AI. For 19% af opslagene er dommen "unclear", mens det kun er 5% af opslagene, som tekstklassifikatoren vurderer enten "unlikely" eller "very unlikely" er genereret af en AI. Blandt opslag som brugerne har efterladt på politikernes eller partiernes Facebooksider, er det 11%, altså mere end dobbelt så mange, der er skrevet af et menneske. Men blandt dem vurderer tekstklassifikator stadig at 68% "likely" eller "possibly" er formuleret af en AI. Vi anser det som usandsynligt, at politikere og borgere, særligt tilbage i 2020, anvendte generativ kunstig intelligens til at producere deres Facebookopslag i så høj grad.



Figur 4: Tekstklassifikatorens vurderinger af indholdet på politikernes Facebooksider

Ingen - ikke engang udviklerne bag tekstklassifikatoren - ved, hvad der gør udslaget ift. dens dom over, om et opslag er forfattet af en AI eller et menneske. I forsøget på at finde mønstre i tekstklassifikatorens vurderinger, har vi læst et repræsentativt og randomiseret udsnit i form af 10% af opslagene med henblik på at finde forskelle og ligheder i teksterne, der kunne give os et praj om, hvorfor tekstklassifikatoren dømmer, som den gør. Analysen gav tæt på ingen resultater.

- Retstavning lader ikke til at have betydning for tekstklassifikatorens vurdering. Det burde være relativt nemt for en kunstig intelligens, der er trænet på hele internettets tekster at genkende de hyppigst anvendte staveformer, hvilket ville tale for, at AI-genereret tekst som udgangspunkt vil være tæt på fejlfrit rent retskrivningsmæssigt. Der er dog både eksempler på fejlfyldte tekster, og tekster med selvopfundne ord, som tekstklassifikatoren vurderer er hhv. menneskeskabte og ai-genereret.
- Brug af emojis lader ikke til at være afgørende.
- Emnerne på opslagene lader heller ikke til at have indflydelse. Corona fylder selvsagt meget i datasættet, men der er også opslag om bl.a. EU, personsager, kriminalitet og uddannelse på tværs af tekstklassifikatorens dom.
- En tydelig afsender og modtager i teksten lader heller ikke til at være afgørende. Både opslag som vurderes at være AI-genereret og forfattet af mennesker indeholder specifikke og personlige detaljer om afsender eller modtager (f.eks. livshistorier eller egne oplevelser). Man kunne ellers

oplagt tænke, at en borgers eller politikeres personlige og konkrete erfaringer med jobcentre eller coronarestriktioner ville være sværere at generere med AI.

- Men ét mønster er tydeligt: En tredjedel af de opslag, hvor tekstklassifikatoren vurderer, at det er "højest usandsynligt/very unlikely", at en AI har haft fingrene med i spillet, er langt oftere opslag på engelsk. Sprog lader altså til at være den eneste afgørende faktor for, om et opslag bliver frikendt. Dette er selvfølgelig, fordi tekstklassifikatoren faktisk virker (bedre) på engelsk. Når tekstklassifikatoren bruges i en dansk kontekst betyder det dog bare, at alle engelske opslag lavet på danske politikeres Facebooksider bliver "clearet" som autentiske. Denne indsigt forklarer også, hvorfor tekstklassifikatoren vurderer, at borgernes opslag på politikernes Facebooksider i lidt højere grad er produceret af mennesker. Flere af borgernes opslag er på engelsk, og de "frikendes" i udpræget grad af tekstklassifikatoren.

Diskussion

Absurde resultater, der rejser en vigtig debat

Som resultaterne viser, er OpenAI's tekstklassifikator ikke i stand til at detektere om politiske tekster på dansk er skrevet af en AI. Det på trods af, at OpenAI's andre produkter som ChatGPT og GPT4 i stor stil kan autogenerere indhold på dansk og til en dansk politisk kontekst. ChatGPT og GPT4 kan generere budskaber og argumenter i relation til konkrete og aktuelle politiske udfordringer. De har genrekendskab og kan nemt skræddersy budskaberne til bestemte onlineplatforme og målgrupper. Autogenereret indhold kan altså nu indgå i og påvirke den politiske debat online, uden at vi har en chance for at vide, om budskaberne og fortolkningerne stammer fra mennesker eller maskiner.

Én af grundene til, at tekstklassifikatoren ikke dur på dansk er, at små sprog generelt er underprioriterede, når internationale virksomheder skal udvikle produkter med kunstig intelligens. Markederne er simpelthen for små. En anden forhindring ift. at bruge klassifikatoren, er, at der foreløbig ikke stilles en såkaldt API-adgang til rådighed, som gør det muligt at vurdere mange tekststykker ad gangen. API'er er til gengæld til rådighed ift. at generere indhold med ChatGPT og GPT4. Vi befinder os altså i en situation med ubegrænset produktion af indhold og ekstremt begrænset mulighed for forfatteridentifikation.

Situationen efterlader et hul i markedet til mere avanceret writeprint-teknologi (writeprint-metaforen refererer til, at alle mennesker har et sprogligt fingeraftryk, når de skriver). Der findes allerede noget teknologi, der, baseret på en persons tidligere tekster, kan vurdere, om vedkommende har forfattet en tekst. Denne teknologi anvendes især i retslingvistikken. Men bedre forfatteridentifkation er kun en lille returnering i et teknologikapløb, hvor der kun er kort vej hen til, at en AI på autentisk vis kan imitere din stil. Og så er klasselærerne på den igen, ligesom vi alle er yderligere i fare for identitetstyveri og svindel.

Konklusion

I denne mini-analyse har vi testet OpenAI's tekstklassifikators evne til at identificere, om en tekst er udarbejdet af AI eller mennesker. Vi har ved hjælp af værktøjet analyseret 2.844 opslag fra de mest aktive politikeres sider på Facebook. Tekstklassifikatoren vurderer, at mere end 75% af danske politikeres opslag og 68% af borgernes opslag på politikernes side formentlig er forfattet af AI. Vi anser det som usandsynligt, at politikere og borgere, særligt tilbage i 2020, anvendte generativ kunstig intelligens til at producere deres Facebookopslag i så høj grad (det ironiske er selvfølgelig, at vi ikke kan være sikre).

Resultaterne dokumenterer nok mest af alt, at tekstklassifikatoren er ubrugelig på dansk, og at vi på nuværende tidspunkt ikke har noget redskab, der virksomt kan hjælpe os med at detektere, om en tekst er skrevet af en AI eller et menneske. OpenAI er selv meget eksplicitte omkring, at tekstklassifikatoren ikke kan bruges til at fælde endelig dom over, hvorvidt en tekst har en kunstig intelligens eller et menneske

som sin forfatter. Men det er vigtigt at understrege, at den i en dansk kontekst ikke bare er en begrænset hjælp - den er ingen hjælp. At OpenAIs tekstklassifikator i praksis ikke virker, understreger, at vi nu står i en verden, hvor meget eksisterende arbejde og flow af information vil blive påvirket. F.eks. vil det blive sværere at detektere mis- og desinformationskampagner, phishing-forsøg, eller autogenererede eksamensopgaver og jobansøgninger. OpenAIs eget bud på en løsning kommer ikke til at hjælpe os foreløbig, og intet tyder på at bedre værktøjer er på vej, eller overhovedet kan udvikles. Og når hjælpen ikke kommer til at komme fra tech-virksomhederne selv, så må vi stille os det vigtige spørgsmål: Hvad gør vi så?