Sammendrag fra NorSK: Norsk språkteknologikonferanse 1



9.oktober 2026
Universitetet i Oslo

arrangert med hjelp fra Norsk senter for fagspråk i bruk (NORBRUK)

Innholdsfortegnelse

Arbeid med tekstekstraksjon og -klassifisering i WebData-prosjektet: En arbeidsrapport
Mark Anderson, Magnus Breder Birkenes, Ildikó Pilán

Automatisk avidentifisering av papirbaserte pasientjournaler fra Norsk Helsearkiv
Pierre Lison

Automatisk korrigering og kategorisering av feil i norske andrespråkstekster
Ildikó Pilán, Peder Mæhle Førland, Mark Anderson

Bokmål og Nynorsk i HPLT-prosjektet
Mariia Fedorova, Andrey Kutuzov

Evaluering av bildesøk og norsk språk og kultur i vektormodeller for bilde og tekst
Tita Enstad, Marie Roald

Evaluering av språkmodeller for revisjonsoppgaver
Ove Haugland Jakobsen, Carolin Prabhu, Anita-Catrin Eriksen

Fra intuisjon til empiri: Verdier i språkmodeller
Oliver Ruste Jahren

Fra vilkårlig til forankret: Et rammeverk for valg og evaluering av språkmodeller i norsk offentlig forvaltning
Robindra Prabhu, Jacob Sjødin, Sarah Brodwall

Fra vondt til velvære: aspektbasert sentimentanalyse på norske pasienterfaringer
Lilja Charlotte Storset, Elma Jelin, Rebecka Maria Norman, Øyvind Bjertnæs, Lilja Øvrelid, Erik Velldal

Glisne vektorer, tette representasjoner: algebra før geometri i språkmodeller
Lars G Bagøien Johnsen

Humit-Taggeren: Morfologisk tagger og lemmatisator
Ahmet Yıldırım, Kristin Hagen, Dag Trygve Truslew Haug

Hver smitt og smule er ikke alltid best: verdien av målrettet seleksjon fremfor mengde av data
Victoria Handford, Lucas Georges Gabriel Charpentier

Lingvistiske utfordringer for frasestrukturgrammatikker
Petter Haugereid

Luminoner: fragmentbasert fortolkning med språkmodeller
Lars G Bagøien Johnsen

Norsk Bloggkorpus som et kasus for en KI-drevet pipeline for datarengjøring, anonymisering og tagging av målform og diskursemne
Olaf Mikkelsen

Norsk Fagspråkkorpus
Helene Brodin, Ahmet Yıldırım, Anders Nøklestad, Dag Trygve Truslew Haug

Om Benchmarking av Sammendrag og Informasjonsgjenfinning for Norsk på Frontier-Språkmodeller og Norske Språkmodeller
Ryan Marinelli

Radikale språkbrukere blant konservative prateroboter
Petter Mæhlum, Yves Scherrer

Revesans: effektivt søk i kunnskaps- og styringsdokumenter for forvaltningsrevisjon
Emanuele Lapponi, Murhaf Fares, Mimi Uyen Dan Nguyen

Sentimentanalyse på entitetsnivå
Egil Rønningstad

Skjevheter i store språkmodeller ved CEFR-basert vurdering av tekster fra norske andrespråksinnlærere
Gloria Kasangu, Samia Touileb

Sosiopolitiske konflikthendelser: Utfordringer, muligheter og resultater av automatisk uthentelse
Helene Bøsei Olsen

Språkteknologiske verktøy for Norchron, et diakront norsk korpus
Anders Nøklestad, Kari Kinn, Ahmet Yıldırım, Dag Trygve Truslew Haug

Uthenting av klinisk informasjon fra medisinske journaler
Tobias Jovall Wessel, Samia Touileb

Arbeid med tekstekstraksjon og -klassifisering i WebData-prosjektet: En arbeidsrapport

WebData-prosjektet skal utvikle en forskningsinfrastruktur for tilgang til data fra det norske internett som er blitt samlet inn av Nasjonalbiblioteket siden slutten av 1990-tallet. Arbeidet har særlig relevans for utvikling av språkteknologi fordi nettdata er råstoffet i store språkmodeller. Vi presenterer foreløpig arbeid med uthenting og klassifisering av tekst fra norske nettsider og vil se nærmere på to temaer vi har jobbet med i det første prosjektåret.

1) Datasett for evaluering av algoritmer for uthenting av norskspråklig tekst fra HTML ("boilerplate removal")

Store språkmodeller trenes i vesentlig grad på ren tekst ekstrahert fra HTML-sider, en prosess som gjerne omtales som “boilerplate removal” og som går ut på å skille hovedinnhold fra annet innhold på en nettside. Derfor er det overraskende at programvaren som brukes til denne oppgaven, eksempelvis Trafilatura (Barbaresi 2021) eller jusText (Pomikálek 2011), er lite testet på andre språk enn engelsk (se Bevendorff et al. 2023 for en oversikt over evalueringsdatasett).

Vi jobber derfor med å lage et åpent evalueringsdatasett for slike systemer på norskspråklige HTML-sider. Oppgaven går ut på å hente ut det som er HTML-sidens hovedinnhold (eksempelvis en nyhetsartikkel) og fjerne topp- og bunntekst, samtykkebannere, lenker til relatert innhold og annet. To annotører har sett på det samme settet med sider og vurdert hva som er hovedinnholdet og kodet teksten (overskrifter, avsnitt, lister, tabeller) i Markdown etter et felles sett med instruksjoner inspirert av CLEANEVAL (Baroni et al. 2008). Foreløpig har vi dobbelannotert totalt 119 HTML-sider fra 2025, omtrent jevnt fordelt mellom bokmål og nynorsk.

2) Klassifisering av dokumenter for tilgangsstyring

Tilgangen til nettdata samlet inn ved Nasjonalbiblioteket reguleres i *Lov om avleveringsplikt for allment tilgjengelege dokument*. For å styre adgangen til dokumentene i infrastrukturen, har vi utviklet en klassifikator som plasserer dokumentene i gitte tilgangskategorier basert på tilhørighet til domene og forekomst av personopplysninger.

Siden nettarkivet består av flere petabyte med data, er vi avhengige av effektive filtreringsmetoder som et første trinn, mens mer beregningsintensive metoder forbeholdes en mindre delmengde av dokumentene. Klassifikatoren kombinerer bruk av regulære uttrykk (f.eks. fødselsnummer, bankkontonummer) med konfigurerbare terskelverdier og kontekstuell validering (som bruker NER i konteksten av slike identifikatorer for å skille eksempelvis person og organisasjon), ikke ulikt Presidio-verktøyet (Mendels et al. 2018). Løypa produserer en binær beslutning om sensitivitet og bruker “early exit” ved funn av sensitive opplysninger. Beslutningen inneholder informasjon om utløste regler og identifiserte treff for videre manuell eller automatisk validering.

Referanser

Barbaresi, Adrien. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint

Conference on Natural Language Processing: System Demonstrations, pages 122–131, Online. Association for Computational Linguistics.

Baroni, Marco, Francis Chantree, Adam Kilgarriff og Serge Sharoff (2008): Cleaneval: a Competition for Cleaning Web Pages. I: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Janek Bevendorff, Sanket Gupta, Johannes Kiesel og Benno Stein (2023): An Empirical Comparison of Web Content Extraction Algorithms. I: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). Association for Computing Machinery, New York, NY, USA, 2594–2603. https://doi.org/10.1145/3539618.3591920

Mendels, Omri and Peled, Coby and Vaisman Levy, Nava and Hart, Sharon and Rosenthal, Tomer and Lahiani, Limor and others (2018): Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images. Microsoft. https://presidio.dataprivacystack.org/

Pomikálek, Jan (2011): Removing Boilerplate and Duplicate Content from Web Corpora. Doktorgradsavhandling. Masaryk University Brno.

Automatisk avidentifisering av papirbaserte pasientjournaler fra Norsk Helsearkiv

Pierre Lison
Norsk Regnesentral
plison@nr.no

8. september 2026

1 Introduksjon

Fra 2020 til 2024 pågikk forskningsprosjektet "CLEANUP" (**Ma**chine **L**earning for the A**n**onymisation of **U**nstructured **P**ersonal Data), med finansiering fra Norges Forskningsråd. Prosjektet tok sikte på å utvikle nye metoder for automatisk å gjenkjenne og sladde personopplysninger i tekstdokumenter.

En av dokumenttypene prosjektet fokuserte på, var pasientjournaler fra Helsearkivregisteret. Registeret inneholder journaler fra avdøde pasienter behandlet i spesialisthelsetjenesten. Norsk Helsearkiv, som har ansvaret for forvaltningen av registeret, har samlet inn over 1,7 millioner pasientjournaler fra 45 helseforetak. Mange av disse journalene er opprinnelig papirbaserte og må derfor skannes for digital lagring. Helstad et al. (2026) beskriver nærmere stegene i prosessen med å digitalisere og forvalte pasientjournalene.

Pasientjournalene fra Helsearkivregisteret kan etter søknad utleveres for bruk i medisinsk og helsefaglig forskning, under strenge vilkår for taushetsplikt og datasikkerhet. Blant annet bør direkte personidentifikatorer som navn eller fødselsnummer sladdes før avleveringen. Manuell sladding er imidlertid svært arbeidskrevende, og CLEANUP utviklet derfor et system for å sladde personidentifikatorer i skannede pasientjournaler. Nedenfor oppsummerer vi kort metoden og hovedfunnene fra dette arbeidet.

2 Metode

Systemet, utviklet i tett samarbeid med Norsk Helsearkiv, har som mål å gjenkjenne seks kategorier av personidentifikatorer: *pasientens navn*, *fødselsdato*, *personnummer*, *telefonnummer*, *hjemmeadresse* og *navn på nære pårørende*.

Systemet ble utviklet i to steg:

Steg 1: Annotering Prosjektet fikk tilgang til 2 800 pasientjournaler[1]. Siden journalene ikke inneholdt annoteringer, utviklet vi en rekke heuristikker for automatisk markering av deler av personidentifikatorene i materialet. For eksempel er pasientens navn ofte innledet av uttrykk som "*Pasient:*" eller "*Navn:*". Basert på disse heuristikkene ble det opprettet et syntetisk treningssett (*silver data*).

[1] Etter godkjenning fra Den nasjonale forskningsetiske komité for medisin og helsefag (NEM).

Steg 2: Finjustering Datasettet ble deretter brukt til å finjustere en språkmodell, nærmere bestemt LayoutXLM (Xu et al., 2021), som kan dra nytte av både tekstlig og visuell informasjon, blant annet dokumentets oppsett. Modellen tar en side som input og markerer personidentifikatorer med avgrensningsbokser (*bounding boxes*) og tilhørende sannsynligheter.

3 Evaluering og konklusjon

Vi utførte en innledende evaluering på et sett med 20 pasientjournaler som ble manuelt annotert for personidentifikatorer. Systemet oppnådde en presisjon på 69 % og en tilbakekalling på 74 % (F_1-score: 71 %).Modellen viser med andre ord lovende resultater, men er fortsatt ikke tilstrekkelig nøyaktig til å brukes uten manuell kontroll.

Hovedårsaken til feil gjenkjenning er i stor grad feil fra optisk tegngjenkjenning (OCR – *Optical Character Recognition*) i de skannede pasientjournalene. Nyere KI-modeller, slik som visjonspråkmodeller (Bai et al., 2025), vil antageligvis kunne redusere effekten av slike OCR-feil.

Referanser

S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

G. Helstad, P. Lison, E. Tuveng, and K. Nytrøen. Digitising health history: The creation, function and implementation of the Norwegian Health Archives Registry. *Health Information Management : Journal of the Health Information Management Association of Australia*, 55(1):166–172, 2026. doi: 10.1177/18333583251389095.

Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florencio, C. Zhang, and F. Wei. LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021.

Automatisk korrigering og kategorisering av feil i norske andrespråkstekster

Ildikó Pilán[‡,*], Peder Mæhle Førland[‡,§], Mark Anderson[‡]
[‡]Norsk Regnesentral, [§]Institutt for informatikk, Universitet i Oslo

Automatisk feilretting har som mål å oppdage, korrigere og kategorisere språklige, særlig grammatiske feil (Bryant, Yuan, et al. 2023). Dette er ikke bare verdifullt som skrivestøtte, men det er også et nyttig steg for videre prosessering med språkteknologiske verktøy som syntaktisk analyse, ettersom etablerte verktøy ofte har redusert ytelse når de brukes på ikke-standard tekst (Kyle and Eguchi 2024).

Vi presenterer et pågående arbeid med automatisk korrigering og kategorisering av feil i tekster skrevet av andrespråksinnlærere av norsk. Vi beskriver tilpasningen av det regelbaserte verktøyet, ERRANT[1] (Bryant, Felice, and Briscoe 2017) til norsk (NorERRANT), som vi har evaluert på et eksisterende korpus av innlærertekster, ASK (Tenfjord, Meurer, and Hofland 2006). I tillegg beskriver vi preliminære resultater fra en evaluering av store språkmodeller for automatisk feilretting av tekster med høy tetthet av andrespråksfeil fra et korpus under innsamling i IncluEdS-prosjektet[2]. Målet med arbeidet er å hente ut språklige trekk som kan brukes til å automatisk predikere språknivå og undersøke språklige mønstre i tastetrykksloggede tekster skrevet av andrespråksinnlærere i norsk.

Feilkategorisering. For å tilpasse det engelske ERRANT-verktøyet til norsk har vi både modifisert modulen for å ekstrahere endringer mellom en korrigert og en original tekstversjon og modulen for feilklassifisering. NorERRANT har totalt 28 feiltyper som omfatter rettskriving, morfologi, syntaks, m.m. Sentrale tillegg i NorERRANT inkluderer bestemthet av substantiv, V2-regler (verb i syntaktisk posisjon to) for subjekt-verb og verb-adverbial, samt stavekontroll basert på Norsk ordbank.[3] Vi har evaluert NorERRANT opp mot det manuelt annoterte ASK-korpuset ved å koble NorERRANTs feiltagger til feilkategoriene i ASK. For å identifisere riktig tekstsegment og feilkategori, oppnådde den aktuelle versjonen av systemet 0.796 presisjon, 0.774 tilbakekalling og 0.791 $F_{0.5}$-score, som gir noe høyere vekt til presisjon. For noen morfologiske og syntaktiske kategorier som inversjon av subjekt-verb hadde systemet god ytelse ($F_{0.5} > 0.8$), men vi fant noen begrensninger for andre, bredere kategorier med mer heterogene feil, som

*E-postadresse til korresponderende forfatter: pilan@nr.no
[1]`https://github.com/chrisjbryant/errant/tree/main`
[2]`https://inclueds.org/`
[3]`https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-5/`

ASK-kategorien *W* for feil ordvalg.

Feilkorrigering. Vi har også begynt arbeidet med å evaluere norske språkmodeller for korrigering av språkfeil i innlærertekster med lavere språknivåer. Vi har gjennomført eksperimenter der språkmodellen, uten ytterligere oppgavespesifikk trening, fikk 5 eksempler som illustrerte prinsippet om minimal feilkorrigering, også brukt i ASK-korpuset. Preliminære resultater for *normistral-11b-warm* modellen (Samuel et al. 2025) var 0.6337 GLEU-score (Wang et al. 2018) og 0.6486 $F_{0.5}$ på 12 manuelt korrigerte IncluEdS tekster. Generelt viste modellen god evne til å korrigere grunnleggende feil som påvirker grammatikalitet i stor grad, med noen tendenser til å underkorrigere mer sofistikerte aspekter knyttet til flyt, og en viss grad av overkorrigering i form av unødvendige tillegg av, for eksempel, modalverb.

References

Bryant, Christopher, Mariano Felice, and Ted Briscoe (July 2017). "Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 793–805.

Bryant, Christopher, Zheng Yuan, et al. (2023). "Grammatical error correction: A survey of the state of the art". In: *Computational Linguistics* 49.3, pp. 643–701.

Kyle, Kristopher and Masaki Eguchi (2024). "Evaluating NLP models with written and spoken L2 samples". In: *Research Methods in Applied Linguistics* 3.2, p. 100120.

Samuel, David et al. (2025). "Small languages, big models: A study of continual training on languages of Norway". In: *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pp. 573–608.

Tenfjord, Kari, Paul Meurer, and Knut Hofland (May 2006). "The ASK Corpus - a Language Learner Corpus of Norwegian as a Second Language". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). URL: https://aclanthology.org/L06-1345/.

Wang, Alex et al. (2018). "GLUE: A multi-task benchmark and analysis platform for natural language understanding". In: *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355.

Finansiering

Dette arbeidet ble støttet av IncluEdS-prosjektet, finansiert av Norges forskningsråd (prosjektnummer: 358153), og Norsk Regnesentral.

Bokmål og nynorsk i HPLT-prosjektet

Mariia Fedorova, Andrey Kutuzov

Universitetet i Oslo, {mariiaf,andreku}@ifi.uio.no

HPLT[1] er et prosjekt for å tilby åpne, svært store, rikt annoterte tekstdatasett av høy kvalitet for nesten 200 språk, samt verktøy for å bearbeide dataene, og modeller trente på dataene (Oepen et al., 2026; Burchell et al., 2025).

Med 30 billioner (10^{12}) token er dette en av de største tilgjengelige flerspråklige samlingene av pretreningsdata for store språkmodeller (SSM). Datasettene har blitt hentet fra nettkravler og inkluderer tekster fra ulike kilder.

Det har også blitt offentliggjort en komplett pipeline med åpen kildekode[2] for å velge dokumenter fra nettarkiver, trekke ut tekst fra HTML, identifisere språk, kjøre eksakt og nær deduplisering (fjerning av like tekster), annotere med blant annet registeretiketter, tekstkvalitetsestimater og personidentifiserbar informasjon.

Bokmål og nynorsk er representerte i dette prosjektet med følgende ressurser:

- Menneskelig verifisering av et datautvalg på bokmål
- Bokmål- og nynorsk-benchmark for SSM-evaluering, med særlig vekt på egenproduserte oppgaver og mekanismer for å redusere promptsensitivitet

[1] https://hplt-project.org/

[2] https://github.com/hplt-project

- Monospråklige encoder–decoder (NorT5), encoder (GPT-BERT) og decoder (GPT-lignende LLama 2B, 13B NorOlmo) modeller, samt finjusterte UD-parsere for bokmål og nynorsk[3]

- En stor samling parallelle engelsk–bokmål- og engelsk–nynorsk-tekster automatisk utvunnet fra HPLT-data, sammen med et nytt parallellkorpus syntetisert gjennom maskinoversettelse

- Språkidentifiseringsdata med én etikett og en modell for 194 språk, inkludert bokmål og nynorsk (OpenLID-v3[4](Fedorova et al., 2026a)); som et selvstendig prosjekt har vi også offentliggjort fleretikett språkidentifiseringsdata og modeller med fokus på skandinaviske språk (SLIDE[5] (Fedorova et al., 2025)).

- DHPLT[6], en åpen samling av diakrone korpus på 41 språk. DHPLT er basert på de nettkravlete HPLT-datasettene; vi bruker tidsstempler fra nettkravler som et signal på dokumentenes produksjonstidspunkt. Bokmål korpus dekker tre tidsperioder: 2011 – 2015, 2020 – 2021 og 2024 – nå, og har 1 million dokumenter per tidsperiode) (Fedorova et al., 2026b).

Referanser

Burchell, L., de Gibert, O., Arefyev, N., Aulamo, M., Bañón, M., Chen, P., Fedorova, M., Guillou, L., Haddow, B., Hajič, J., Helcl, J., Henriksson, E., Klimaszewski, M., Komulainen, V., Kutuzov, A., Kytöniemi, J., Laippala, V., Mæhlum, P., Malik, B., Mehryary, F., Mikhailov, V.,

[3] https://huggingface.co/HPLT/collections

[4] https://huggingface.co/HPLT/OpenLID-v3

[5] https://huggingface.co/collections/ltg/slide

[6] https://data.hplt-project.org/three/diachronic/

Moghe, N., Myntti, A., O'Brien, D., Oepen, S., Pal, P., Piha, J., Pyysalo, S., Ramírez-Sánchez, G., Samuel, D., Stepachev, P., Tiedemann, J., Variš, D., Vojtěchová, T., and Zaragoza-Bernabeu, J. (2025). An expanded massive multilingual dataset for high-performance language technologies (HPLT). In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Fedorova, M., Arefyev, N., Buljan, M., Helcl, J., Oepen, S., Rønningstad, E., and Scherrer, Y. (2026a). OpenLID-v3: Improving the precision of closely related language identification – an experience report. In Scherrer, Y., Aepli, N., Blaschke, V., Jauhiainen, T., Ljubešić, N., Nakov, P., Tiedemann, J., and Zampieri, M., editors, *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 275–292, Rabat, Morocco. Association for Computational Linguistics.

Fedorova, M., Frydenberg, J. S., Handford, V., Langø, V. O. C., Willoch, S. H., Midtgaard, M. L., Scherrer, Y., Mæhlum, P., and Samuel, D. (2025). Multi-label Scandinavian language identification (SLIDE). In Holdt, Š. A., Ilinykh, N., Scalvini, B., Bruton, M., Debess, I. N., and Tudor, C. M., editors, *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 179–189, Tallinn, Estonia. University of Tartu Library, Estonia.

Fedorova, M., Kutuzov, A., and Umarova, K. (2026b). DHPLT: large-scale multilingual diachronic corpora and word representations for semantic change modelling. In Tahmasebi, N., Cassotti, P., Montariol, S., Kutuzov, A., Huebscher, N., Spaziani, E., and Baes, N., editors, *The Proceedings for the 6th International Workshop on Computational Approaches*

to Language Change (LChange'26), pages 87–96, Rabat, Morocco. Association for Computational Linguistics.

Oepen, S., Arefyev, N., Aulamo, M., Bañón, M., Buljan, M., Burchell, L. V., Charpentier, L. G. G., Chen, P., Fedorova, M., de Gibert, O., Haddow, B., Hajič, J., Helcl, J., Kutuzov, A., Laippala, V., Li, Z., Malik, B., Mikhailov, V., Myntti, A., O'Brien, D., Polakova, L., Ramírez-Sánchez, G., Siewert, J., Stepachev, P., Tiedemann, J., Vahtola, T., Varis, D., Vitiugin, F., and Zaragoza, J. (2026). HPLT 3.0: Very large-scale multilingual resources for LLMs and MT. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 1409–1434, Palma de Mallorca, Spain. ELRA Language Resource Association.

Evaluering av bildesøk og norsk språk og kultur i vektormodeller for bilde og tekst

Multimodale vektormodeller som kan representere bilde og tekst i samme vektorrom er svært populære, og brukes til bildesøk på nettsidene til bl.a. Nasjonalarkivet, DigitaltMuseum og Nasjonalbiblioteket. Det er imidlertid gjort lite forskning på evaluering av slike modeller for norsk bruk. Selv modeller som er trent på flerspråklig data er i hovedsak evaluert på engelsk data. Vi jobber derfor med å lage datasett for å evaluere slike modeller for norsk arkiv, bibliotek og museum (ABM)-sektor, og vil presentere erfaringene vi har så langt.

I presentasjonen vil vi bl.a. vise spesifikke eksempler på hvordan modellenes evne til å lese tekst i bildet og en tilsynelatende preferanse for engelsk, selv i flerspråklige modeller, kan føre til pussige søkeresultater. Vi vil også vise resultater fra kvantitative eksperimenter som måler søkekvalitet for norsk data.

Siden vi er interessert i modellers ytelse for norsk ABM-sektor er datasettene våre basert på bilder, beskrivelser og annen metadata som er tilgjengelig på nettsidene til norske kulturinstitusjoner. Vi har laget et datasett basert på verkskatalogen til Edvard Munch, fra Munchmuseet. Datasettet består av rundt 10 000 bilder av kunstverk, tegninger og notatbøker, samt klassifikasjon og informasjon om bildene. Et annet datasett er basert på DigitaltMuseum, som er et samarbeid mellom 248 norske museer. Det er stor variasjon i bildematerialet, som rommer bl.a. fotografier, portretter, bruksgjenstander, båter, skulpturer, bygninger og plakater, og rikelig med metadata. Totalt består datasettet av ca. 200 000 bilder med tilhørende kategorisering, beskrivelse, emner, m.m.

Basert på disse datasettene har vi laget forskjellige oppgaver for å evaluere modellene. Blant annet simulerer vi søk ved å måle om vektorrepresentasjonen av tekstbeskrivelsen til et bilde "treffer" bildet, om bildet havner innenfor topp k resultater ved rangering på cosinuslikhet. Vi undersøker også om det er en korrelasjon mellom modellers tokenisatorfrodighet (eng: tokenizer fertility) på norsk representativ tekst, og ytelse på våre bildesøkoppgaver.

I vårt innlegg vil vi presentere noen foreløpige resultater, lærdom fra arbeidet så langt, og hvordan vi vil jobbe videre med prosjektet.

Evaluering av språkmodeller for revisjonsoppgaver

Vi presenterer Riksrevisjonens pågående arbeid med utvikling av evalueringsdata for å vurdere språkmodellers egnethet til å løse enkelte revisjonsoppgaver.

Riksrevisjonen har begynt å utforske bruk av språkmodeller for oppgaver som oppsummering eller informasjonsuthenting fra store mengder tekst. Kunnskap om påliteligheten til modellene i vår kontekst er essensielt før de kan settes i produksjon. For mange våre oppgaver er det nødvendig å bruke lokal infrastruktur, noe som øker behovet for en empirisk avveiing av modellstørrelse mot modellytelse.

Eksisterende sammenligninger er ofte basert på generelle oppgaver med en bias mot tekst på engelsk. Det er ikke åpenbart hvordan disse referanseverdier kan oversettes til ytelse på norsk fagspråk i oppgaver relevant i kontekst av revisjonsarbeid. I dag finnes det liten evaluering rettet mot offentlig forvaltning som domene [2] eller mot norsk språk [4].

For å vurdere forventet ytelse har vi begynt med å bygge opp et eget testsett for

(1) uthenting av informasjon fra styringsdokumenter i norsk offentlig forvaltning,

(2) oppsummering av bl.a. revisjonskonklusjoner, og

(3) dikotome spørsmål om norsk forvaltning

(1) I en forvaltningsrevisjon av samarbeid i helsetjenestene gjennomførte vi en systematisk evaluering av klassifisering med språkmodeller. Modellene klassifiserte 1 356 sider fra norske sykehusplaner på tvers av ni ulike tema basert på relevans. Et utvalg av sidene ble manuelt vurdert av revisjonsteamet, og 556 annoterte eksempler dannet et valideringssett. Opprinnelig ble 22 modeller testet, og dette utvalget er løpende utvidet i takt med at nye modeller har blitt tilgjengelige.

Vi ønsket å undersøke egnetheten til språkmodeller for *one-shot*-klassifisering, og hvorvidt nøyaktigheten øker med størrelsen (antall parametere) på modellen. Vi fant at det ikke er en konsekvent sammenheng med størrelsen på modellen og nøyaktighet, og at modellene var mer sensitive for definisjonen av tema.

(2) Språkmodeller kan egne seg til oppsummering av dokumenter fra offentlig sektor, men tilgjengelige evalueringsdata er i stor grad basert på nyhetsartikler [3]. En typisk årsrapport, tildelingsbrev eller revisjonsrapport er avviker både i struktur og språkbruk. Vi har sammenstilt evalueringsdata basert på forvaltningsrevisjonsrapporter, som inneholder et avsnitt med oppsummering av konklusjonene før et lengre avsnitt med utdyping av konklusjonene. Som i [2], bruker vi deretter en kombinasjon av semantisk likhet og evalueringsmål fra LLM-as-a-judge for å vurdere språkmodellens ytelse.

(3) Modeller kan ha utfordringer med å svare korrekt på dikotome spørsmål [5]. Vi har utforsket åpne data for å bygge et valideringsdatasett [6], men problemer med datakvalitet og relevans for Riksrevisjonen har gjort det vanskelig å benytte disse. I stedet bygger vi egne datasett hvor konteksten er hentet fra norsk forvaltning hvor hovedvekten av teksten er hentet fra Riksrevisjonen og Stortinget.

Referanser

[1] J. Sandnes, O. H. Jakobsen, T. Næss, *Sifting for Gold: Mining relevant content in a large document collection*, chapter 14 in *From Algorithms to Evidence: Using GenAI in Evaluation Practice, From Algorithms to Evidence: Using GenAI in Evaluation Practice* https://doi.org/10.4324/9781003799139, https://www.taylorfrancis.com/chapters/oa-mono/10.4324/9781003799139-14/sifting-gold-joachim-sandnes-ove-haugland-jakobsen-tom-næss?context=ubx&refId=bae6a3b8-ff7e-441b-9c78-9e8f3b54abaa

[2] C. Dalerci, T. Michael, R. Schaefer, D. Weinland, *MÖVE: A Holistic LLM Benchmark for the German Public Sector*, https://arxiv.org/abs/2606.13111

[3] S. Touileb, V. Mikhailov, M. Kroka, L. Øvrelid, E. Velldal, *Benchmarking Abstractive Summarisation: A Dataset of Human-authored Summaries of Norwegian News Articles*, https://arxiv.org/abs/2501.07718

[4] V. Mikhailov, T. Enstad, D. Samuel, H. C. Farsethås, A. Kutuzov, E. Velldal, L. Øvrelid, *NorEval: A Norwegian Language Understanding and Generation Evaluation Benchmark*, https://arxiv.org/abs/2504.07749

[5] Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). *BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions*. arXiv preprint arXiv:1905.10044. https://arxiv.org/abs/1905.10044

[6] Ivanova, S., Andreassen, F. A., Jentoft, M., Wold, S., & Øvrelid, L. (2023). NorQuAD: Norwegian Question Answering Dataset. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa 2023), Tórshavn, Faroe Islands.
https://huggingface.co/datasets/ltg/norquad

Fra intuisjon til empiri: Verdier i språkmodeller

Abstract til 'NorSK: Norsk SpråkteknologiKonferanse' 2026

Oliver Ruste Jahren

8. september 2026

All teknologi er politisk [4]—det innebefatter også store språkmodeller. Men det er ikke gitt at de som bruker språkmodeller, er viss på deres politiske natur. Det har også blitt vist at språkmodeller kan forfekte desinformasjon, selv uten at det er en villet handling fra deres bakmenn [3]. Følgelig finnes det en samfunnsmessig interesse i å oppklare bias og politiske slagsider i språkmodeller. Dette favner om noe bredere enn bare hallusinasjonsfenomenet [6, s. 59], men at de kan ha verdimessige holdninger som man kanskje ikke nødvendigvis ønsker at de *skal*, spesielt uten at man er klar over det. Å ta ibruk språkmodeller som har verdier vi hverken (1) vet om, eller (2) kan stå inne for, minner om å slippe Trojanske hester løs i samfunnet og kan tenkes å ha store konsekvenser for vår digitale suverenitet og demokratiske legitimitet.

Å sikre at språkmodellen oppfører seg sånn vi vil, kan sees på som en instans av det mer generelle problemet knyttet til verdijustering. Verdijustering handler ikke bare om å sikre at språkmodellen oppfører seg 'sånn vi vil'—som er et vanskelig problem å løse all en tid det ofte er *uenighet* om hva som er etisk riktig [2, s. 411–412]—det handler også om å komme frem til nettopp hva dette 'riktige' er. Et vesentlig fundament for å kunne komme frem til hvorvidt en gitt språkmodell har de verdiene vi ønsker at den skal, er å få på det rene hvilke verdier den *fakitsk har*. Men hva er så egentlig en verdi? Og hvordan kan vi si at en språkmodell skal kunne 'ha' sånne?

Yao, Yi og Xie [5, s. 2] la frem et forslag til en løsning på dette problemet: De analyserer LLM-output, ved hjelp av en annen LLM. Men dette er ikke en holdbar løsning all den tid alle de grunner som får oss til å ville sjekke verdiene i den første LLMen, også vil motivere oss til å sjekke verdiene i den andre, da de naturligvis vil kunne prege resultatet av verdievurderingen. Ved å skulle introdusere en *tredje* LLM inn i dette bildet ser vi hvordan dette er sirkulært og gir opphav til en regress[1]. Altså trenger vi en måte å måle verdier som ikke *selv* belager seg på en LLM eller

[1]Hvilken LLM skal være den første? *Denne* må vi analysere på en annen måte.

tilsvarende 'ugjennomsiktig' [1] teknologi.

Ved å få på det rene hvordan man skal (1) finne, og (2) måle disse verdiene uten å lene seg på en annen språkmodell, vil vi kunne styrke demokratiet, og åpne for informert samtykke til bruk av språkmodellene. Man vil ikke lenger måtte ta de i bruk og bare satse på at den bare *forhåpentligvis* har verdier som er i tråd med ens egne.

Jeg vil derfor foreslå et prosjekt som utvikler et paradigme for å besvare disse filosofiske spørsmålene, samt et språkteknologisk rammeverk for å gjennomføre målinger innenfor det, for vestlige språk.

Referanser

[1] Jenna Burrell. «How the machine 'thinks': Understanding opacity in machine learning algorithms». I: *Big data & society* 1 (2016). ISSN: 20539517. DOI: 10.1177/2053951715622512.

[2] Iason Gabriel. «Artificial Intelligence, Values, and Alignment». I: *Minds and machines (Dordrecht)* 3 (2020), s. 411–437. ISSN: 09246495. DOI: 10.1007/s11023-020-09539-2.

[3] Mykola Makhortykh mfl. «Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine». I: *Harvard Kennedy School Misinformation Review* 4 (5 2024). DOI: https://doi.org/10.37016/mr-2020-154.

[4] Langdon Winner. «Do Artifacts Have Politics?» I: *Daedalus (Cambridge, Mass.)* 1 (1980), s. 121–136. ISSN: 00115266. URL: https://www.jstor.org/stable/20024652.

[5] Jing Yao, Xiaoyuan Yi og Xing Xie. «CLAVE: An Adaptive Framework for Evaluating Values of LLM Generated Responses». I: *Advances in Neural Information Processing Systems*. Red. av A. Globerson mfl. Bd. 37. Curran Associates, Inc., 2024, s. 58868–58900. DOI: 10.52202/079017-1877.

[6] Wayne Xin Zhao mfl. *A Survey of Large Language Models*. 2025. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223.

Fra vilkårlig til forankret

Et rammeverk for valg og evaluering av språkmodeller i norsk offentlig forvaltning

Robindra Prabhu Jacob Sjødin Sarah Broadwall

Språkmodeller finner veien inn i stadig flere løsninger i det offentlige. I motsetning til tradisjonelle maskinlæringsmodeller kan de utvikles av eksterne aktører, baseres på data utenfor virksomheten, og likevel gi nyttige anvendelser internt. Denne allsidigheten og anvendeligheten, evnen til å bearbeide kompleks tekstlig informasjon og høye digitaliseringsambisjoner [1] gjør det rimelig å forvente at bruken av språkmodeller i offentlig forvaltning vil tilta.

Teknologi i offentlig forvaltning må oppfylle strenge krav til utforming og bruk [2]. Valg av språkmodell berører spørsmål om kvalitet, etterlevelse av regelverk, kostnader, bærekraft, digital suverenitet, demokrati og rettsstatsprinsipper. Likevel er slike hensyn sjelden styrende. Modellvalg i det offentlige fremstår ofte vilkårlig, drevet av lokale, praktiske og forretningsmessige hensyn fremfor å være forankret i en analyse av krav og føringer i forvaltningen. Fraværet av tydelige vurderingskriterier bidrar vilkårlighet og reiser spørsmål om konsekvenser på systemnivå.

Det finnes mange referansemålinger for språkmodeller [3–6], men disse måler gjerne tekniske ferdigheter (som f.eks. evnen til å klassifisere tekst) eller isolerte egenskaper (f.eks. språkbeherskelse). Alene gir de liten veiledning og det er i beste fall uklart hvordan de fanger opp de nyanserte juridiske, prosedyremessige og strukturelle kravene i offentlig forvaltning. Det finnes få helhetlige rammeverk som vurderer hvorvidt språkmodeller er egnet for slik bruk [7,8], og ingen som er rettet spesifikt mot norsk forvaltning. Det offentlige Norge mangler dessuten en kvalitetssikret oversikt over eksisterende referansemålinger og rangeringslister: hva de måler, deres styrker og svakheter, hvilke som er relevante og hvordan resultatene bør fortolkes i lys av krav og føringer for offentlige løsninger. Dette arbeidet søker å bøte på denne mangelen ved å foreslå et helhetlig rammeverk med en prinsipiell tilnærming til å vurdere språkmodeller i lys av kravene til løsninger i norsk offentlig forvaltning. Rammeverket vurderer modeller langs tre dimensjoner:

- **Kunnskap om virkeområdet**: Hvilken kunnskap har modellen om virksomhetens fagområde, og hvor godt håndterer den relevante fagspørsmål? Vi bruker Nav som eksempel, gjennom et egenutviklet evalueringsdatasett basert på spørsmål og svar fra Nav Kontaktsenter.

- **Ferdigheter**: Hvordan presterer modellen på oppgaver som er typiske for anvendelser av teknologien i offentlig forvaltning?
- **Produksjonsprosess og verdiorientering**: Hvordan er modellen laget, og hvor godt samsvarer modellen med virksomhetens verdier og prinsipper? Vi illustrerer med evalueringer knyttet til Navs prinsipper for ansvarlig kunstig intelligens [9].

Rammeverket trekker dels på evalueringer basert på data fra Navs egne systemer, og dels på eksterne evalueringer. Til førstnevnte benyttes en kombinasjon av statistiske metrikker, metoder basert på vektorrepresentasjon («embedding») og «språkmodell-som-dommer» tilnærminger. Til sistnevnte foreslår vi å hente evalueringer fra et kvalitetssikret utvalg av eksterne rangeringslister («leaderboards»). På denne måten forankrer rammeverket vurderinger på tvers av offentlige etater i et felles sett med referansemålinger der dette er hensiktsmessig (f.eks. ferdigheter på norsk), samtidig som det åpner for lokal kalibrering til virksomhetens eget virkeområde, behov og rammer.

Vi presenterer det pågående arbeidet med rammeverket, viser en prototype og drøfter utfordringer og muligheter i videre arbeid, og håper slik å bidra til diskusjon om hvilke vurderingskriterier som bør være førende for språkmodeller i offentlig forvaltning.

[1] Digitaliserings- og forvaltningsdepartementet. Ny nasjonal digitaliseringsstrategi 2024.
[2] Schartum DW. Lovgivning i et digitalt samfunn – CompLex 1/25 2025.
[3] Eurolingua. European LLM leaderboard 2024.
[4] National Library of Norway. AI leaderboard n.d.
[5] Language Technology Group, University of Oslo. NoReEval – LLM Dashboard n.d.
[6] EuroEval. EuroEval – the robust european language benchmark n.d.
[7] Dalerci C, Michael T, Schaefer R, Weinland D. MÖVE: A holistic LLM benchmark for the German Public Sector 2026.
[8] Rystrøm J, Schmitz C, Korgul K, Batzner J, Russell C. Agent benchmarks fail Public Sector Requirements 2026.
[9] Nav. Prinsipper for ansvarlig KI n.d.

Fra vondt til velvære:
aspektbasert sentimentanalyse på norske pasienterfaringer

Hvert år sender FHI (Folkehelseinstituttet) ut nasjonale spørreundersøkelser til pasienter for å samle inn deres erfaringer med helsetjenesten. Disse undersøkelsene genererer titusenvis av svært verdifulle pasientkommentarer som potensielt kan brukes til å forbedre helsetjenesten. På tross av dette, blir de sjeldent analysert fordi dette etter dagens manuelle metoder krever store mengder tid og ressurser. Gjennom SA-PREMS-prosjektet samarbeider UiO og FHI med å undersøke hvordan språkteknologi kan bidra i ulike prosesser der manuelt arbeid tidligere har vært gjeldende. Til dette formålet bruker vi blant annet aspektbasert sentimentanalyse (ABSA), og introduserer i sammenheng med det et menneskelig annotert datasett bestående av pasienterfaringer innen fastlegetjenesten og psykisk helsevern. Annotasjonene dekker både setnings- og kommentarnivå, og tar i bruk 25 finkornede, helserelaterte aspekter med tilhørende polaritet. Vi utfører eksperimenter på både enkoder- og dekodermodeller, og viser resultater for forskjellige modelleringsteknikker.

Videre fører resultatet av arbeidet ikke bare til en effektivisert prosess med mulighet for å lytte til alle pasienter, men åpner også for analyse på flere nivåer som kan hjelpe både fastleger og andre aktører å forstå pasienters erfaringer på et dypere nivå enn det kvantitative metrikker ellers kan vise. Som et ytterligere nivå av kvalitativ analyse, utforsker vi nå generative sammenfatninger av pasienterfaringer basert på ABSA-annotasjonene vi allerede har.

Referanser:

From Pain to Praise: Aspect-Based Sentiment Analysis for Norwegian Patient Feedback
(Storset et al., HeaLing 2026)

Mixed Feelings: Cross-Domain Sentiment Classification of Patient Feedback
(Rønningstad et al., NoDaLiDa 2025)

Glisne vektorer, tette representasjoner: algebra før geometri i språkmodeller

Lars G. B. Johnsen

Nasjonalbiblioteket

Språkmodeller forklares ofte geometrisk: Ord og setninger plasseres i et semantisk rom, avstander uttrykker likhet, og retninger svarer til betydningsforskjeller. Denne språkbruken kan være nyttig, men også fremmedgjørende. Den gir lett inntrykk av at det finnes et uavhengig rom som ordene plasseres inn i, og at ordlister tilhører et eldre stadium av språkteknologien.

Et vektorrom er i utgangspunktet en algebraisk struktur definert ved addisjon og multiplikasjon med skalarer. I et fysisk rom kan origo flyttes uten at rommet forandres. I et vektorrom er nullpunktet derimot bestemt som identiteten for addisjon. Representasjonene befinner seg derfor ikke i et språklig landskap som eksisterer uavhengig av dem.

En frekvensliste kan forstås som en funksjon fra ord til tall. Etter vekting med for eksempel relativ frekvens, TF–IDF eller PMI er den en vektor på samme måte som en embedding. Frekvensvektoren er høydimensjonal og glissen, med tolkbare koordinater. Embeddingen er lavdimensjonal og tett, med latente koordinater. Ord- og frekvenslister er dermed ikke før-vektorielle representasjoner, men eksplisitte og lettere forståelige vektorer.

Den samme frekvensinformasjonen kan sammenfattes med ulike likhetsmål. Cosinus normaliserer skalarproduktet og omtales gjerne som vinkelen mellom to vektorer. Jaccard sammenligner snitt og union, mens vektet Jaccard summerer komponentvise minima og deler på summen av komponentvise maksima. Jaccard-målene kan forklares direkte som delt ordmasse i forhold til samlet ordmasse, uten at observasjonene først må forestilles som punkter eller piler. Målene bruker samme grunnlagsdata, men kan rangere likhet forskjellig. Naboskapet bestemmes derfor av operasjonen, ikke av vektorene alene.

Et skalarprodukt består selv av komponentvis multiplikasjon etterfulgt av summering. Matrisemultiplikasjon og attention bygger videre på den samme mekanismen. Glisne frekvens- og samforekomstmatriser kan dessuten komprimeres tapsfritt eller omformes til kortere, tette representasjoner ved dimensjonsreduksjon og faktorisering. Komprimeringen gjør lagring og beregninger mer håndterlige, men koordinatene mindre direkte fortolkbare.

Bidraget demonstrerer kontinuiteten fra ordlister og glisne frekvensvektorer til tette embeddings og transformeroperasjoner. Hensikten er ikke å avvise geometriske analyser, men å vise at geometrispråket kan skjule en enklere algebraisk sammenheng. Dette gir et mer tilgjengelig begrepsapparat og større frihet til å velge representasjon og likhetsmål etter oppgaven.

Humit-Taggeren: Morfologisk tagger og lemmatisator

Vi introduserer Humit-taggeren, en morfologisk tagger og et lemmatiseringsverktøy som kan behandle tekster på bokmål og nynorsk. Taggeren implementerer fire forskjellige oppgaver: 1) tokenklassifisering for setningsgrensedeteksjon, 2) sekvensklassifisering for målformdeteksjon, 3) tokenklassifisering for morfologiske tagger og 4) tokenklassifisering for deteksjon av lemmatiseringsregel. Alle oppgavene løses gjennom finjusteringer av Norbert3-modellene (Samuel et al. 2023) i fire forskjellige størrelser. Lemmatisatoren (Yıldırım et al. 2026) bruker også oppslag i Norsk ordbank.

Vi har brukt Norsk dependenstrebank (Solberg et al. 2014) for å lage treningsdatasett for oppgave 1, 3 og 4. For oppgave 2 (målformdeteksjon) har vi benyttet et ekstra datasett hentet fra Nasjonalbibliotekets API, bestående av om lag 16 000 Kudos-dokumenter fra tidsperiodene 1999-2002 og 2015-2018. Vi hentet ut tekstbiter på omtrent 25 ord per eksempel sammen med tekstbitens målform.

Vi konverterer alle datasettene til representasjoner egnet til å trene klassifikasjonsmodeller. For morfologisk analyse lar vi hver taggkombinasjon (f.eks. “subst appell be ent mask”) være en klasse. For lemmaene bruker vi den regelbaserte løsningen av UDpipes-lemmatisator (Straka et al. 2019). Her representerer en klasse en regel som genererer lemmaet til ordet. For eksempel gir ordet «planlagt» og lemmaregelen «lower;-+e→-+ge» lemmaet «planlegge». Her betyr den første delen av lemmaregelen at lemmaet skal skrives med liten bokstav, og den andre delen betyr “legg til e på slutten, erstatt den nest siste bokstaven med g, og erstatt den fjerde siste bokstaven med e.” For setningsgrenseidentifikasjon klassifiserer vi tokenene som enten “slutt på setning” (siste token) eller “ikke slutt på setning” (alle andre token). For målformsdeteksjon genererer vi datasettet slik at det første tokenet “[CLS]” klassifiseres som “nynorsk” eller “bokmål”.

For hver oppgave delte vi dataene i et treningssett (80%), et utviklingssett (10%) og et testsett (10%). For de tre oppgavene som bruker NDT er testsettet det samme for å unngå lekkasje av setninger fra trenings- eller utviklingssettet for en oppgave til testsettet for en annen oppgave. Oppdelingen i treningsdata og utviklingsdata er forskjellig for hver oppgave og tar sikte på balansert representasjon av klassene i den aktuelle oppgaven.

Vi bruker Norbert3-modellene i fire forskjellige størrelser og legger til fire lineære klassifiseringslag, ett for hver oppgave, etter det siste laget. Vi trener modellene for alle oppgavene samtidig ved å blande treningssettene tilfeldig for alle oppgavene. Vi beregner tapsfunksjonen for hver oppgave basert på de eksemplene som hører til hver oppgave i hver “bunt” (batch). Tapsfunksjonen for hele bunten er summen av tapsfunksjonene for hver oppgave. Gjennom treningen øker nøyaktigheten for oppgavene i målsformidentifikasjon og klassifisering av setningsgrensedeteksjon raskt fordi disse oppgavene bare har to klasser. For å

kompensere for skjevhet mot disse oppgavene vekter vi tapsfunksjonene for disse oppgavene lavere enn for de to andre oppgavene. Vi trener modellene i 50 epoker med en buntstørrelse på 16. Læringsraten ble satt til 5e-5 med et lineært vektforfall på 0,01. Tabellen viser nøyaktigheten på testsettet, som inneholder både bokmål og nynorsk.

	xs	small	base	large
Lemmatisering	99.08	99.36	99.48	99.51
Morfologisk taggdeteksjon	98.19	98.58	98.77	98.89
Setningsgrensedeteksjon	99.58	99.78	99.93	99.96
Målformsdeteksjon	99.36	99.30	99.90	99.90

Referanser

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian dependency treebank. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 789–795, Reykjavik, Iceland. European Language Resources Association (ELRA).

Milan Straka, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morpho- 103 logical categories, corpora merging. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 95–103, Florence, Italy. Association for Computational Linguistics.

Ahmet Yıldırım, Kristin Hagen, Dag Trygve Truslew Haug, Automatic Lemmatisation for Norwegian, Proceedings of the Workshop on Structured Linguistic Data and Evaluation (SLiDE 2026), pages 93–103, May 11, 2026, ELRA Language Resources Association (ELRA).

Hver smitt og smule er ikke alltid best: verdien av målrettet seleksjon fremfor mengde av data

Victoria Handford
The National Library of Norway

Lucas Charpentier
The National Library of Norway

1 Introduction

Seksjonen for språkmodeller i Nasjonalbiblioteket har fått i oppdrag å trene store språkmodeller for norsk og samisk innenfor opphavsrettslige rammer. Dette sammendraget handler spesifikt om valg av data for å trene embeddingmodeller. Et sentralt mål for moderne embeddingmodeller er høy ytelse på informasjonsgjenfinning (retrieval), så vi prøver å optimisere særlig for retrieval, men ekskluderer ikke andre typer oppgaver.

2 Data

Mens tidligere embeddingmodeller fra biblioteket har blitt trent på NLI-data (∼500k eksempler på bokmål og engelsk[1] (Williams et al., 2018)), ser man nå at de beste flerspråklige embeddingmodellene er trent på mye mer data. For å øke datamengden samlet vi inn en bred portefølje på om lag 10 millioner tekstpar og tripler fordelt på 11 datasett (Beskrivelsen av datasettene finnes i A.1). Etter noe eksperimentering innså vi at vi trengte å se mer på hvilken data, ikke bare mengden data, som inkluderes.

3 Eksperiment

Det er upraktisk å kjøre det fulle treningsløpet for modeller på alle datakombinasjoner. For å unngå ustrukturert intuisjon og unødvendig ressursbruk formulerte vi valg av dataset som et *feature selection problem*. Vi benyttet en nedskalert modell og implementerte en *Greedy Forward Selection* metode med eliminering:

1. Modellen finjusteres iterativt på hvert datasett i kombinasjon med den forrige beste delmengde.

2. Modellen evalueres på MTEB(Scandinavian) benchmarken (Muennighoff et al., 2023). Vi er mest interessert i gjennomsnittsscore (MeanTask) og informasjonsgjenfinning (Retrieval).

3. Datasett som tilfører støy, gir marginal forbedring relativt til treningskostnad, eller svekker ytelsen, fjernes fra mulige treningsdatasett.

[1]Datasettet er tilgjeneglig på Hugging Face https://huggingface.co/datasets/NbAiLab/mnli-norwegian

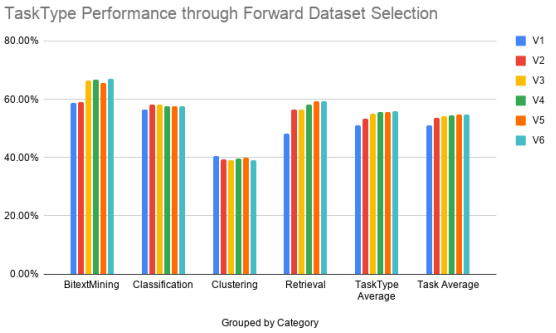


Figure 1: Ytelsesutvikling på MTEB (Scandinavian) gjennom iterasjonene (V1–V6). Prosessen gir en markant fremgang i både informasjonsgjenfinning (Retrieval) og gjennomsnittlig oppgaveytelse (Mean Task), uten stor nedgang på øvrige oppgavetyper. Bildet er forstørret i A.2.

4 Resultater

Denne tilnærmingen reduserte søkerommet fra 2048 kombinasjoner til kun 34 faktiske treningskjøringer. Eksperimentene avdekket at rå datamengde ikke førte til høyere representasjonskvalitet. Store datasett som er fokuserte på spesifikke oppgaver kan svekke signalen av andre datasett og mer data av samme type er ikke garantert til å hjelpe.

Etter seks iterasjoner landet vi på en mengde bestående av 5 dataset. Dette reduserte den samlede datamengden fra ca. 10 millioner til 3,5 millioner dokumenter (en reduksjon på 65%). Samtidig økte retrieval fra 48,30% til 59,30% (+11,00%)

og MeanTask fra 51,10% til 54,80% sammenlignet med å bare trene på det opprinnelige 500k NLI datasettet. Figure 1 viser resultatene på MTEB(Scandinavian) etter hver iterasjon og en oversikt av resultatene finnes i A.2.

5 Konklusjon

Resultatene viser at systematisk, målrettet seleksjon fører til bedre resultater enn ufiltrert datamengde for treningen av vår embeddingmodell.

References

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Datasett

Table 1: Oversikt over de 11 kandidatdatasettene vurdert i seleksjonsprosessen (totalt $\approx$ 10 mill. eksempler), fordelt på oppgavetype, størrelse og beskrivelse.

Datasett	Datatype / Format	Størrelse	Beskrivelse
NLI	Sentence Triplet	527 098	NbAiLab/mnli-norwegian
Interne artikler	Query-to-Document	553 697	Norske aviser
OPUS100	Bitext Mining	1 628 961	Engelsk/norsk setningspar fra sentence-transformers/parallel-sentences-opus-100
DDSC Retrieval	Query-to-Document	158 148	DDSC/nordic-embedding-training-data retrieval eksempler
Wiki	Query-to-Document	617 937	Norske Wikipedia articler fra lightonai/embeddings-pre-training
NLI V2 (Old Mix)	Sentence Triplet	557 850	Intern maskinoversettelse av sentence-transformers/all-nli med engelsk/norsk distribusjon fra V1
DDSC Triplets	Sentence Triplet	176 443	DDSC/nordic-embedding-training-data unit-triple eksempler
NLI V2	Sentence Triplet	1 114 219	Intern maskinoversettelse av sentence-transformers/all-nli med bare engelske og norske triplets
FineWiki	Query-to-Document	792 184	Norske Wikipedia articler fra HuggingFaceFW/finewiki
NLLB	Bitext Mining	3 721 296	Engelsk/norsk setningspar fra hotchpotch/nllb-english-bitext-hq
NDLA	Query-to-Document	189 652	Parallell Bokmål - Nynorsk avsnittspar fra NbAiLab/ndla_parallel_paragraphs
Totalt evaluert		**10 037 485**	11 datasett totalt ($\sim$ 10 mill. eksempler)
Kuratert delmengde		**3 485 841**	**5 datasett beholdt** (65 % reduksjon i datamengde)

A.2 Resultater

Versjon	Valgt datasett	Ny datastørrelse	Mean (Task)	Retrieval	Beskrivelse / Aksjon
V1	NLI	527 098	51,10 %	48,30 %	Startpunkt; NLI V2 (Old Mix) fjernes.
V2	+ Interne aviser	1 080 795	53,70 %	56,40 %	Markant løft (+8,1 pp ret). DDSC Trip og NLI V2 elimineres.
V3	+ OPUS100	2 709 756	54,20 %	56,60 %	Sterkt løft på Bitext Mining (+7,3 pp).
V4	+ DDSC Ret	2 867 904	54,60 %	58,20 %	Positiv for retrieval. FineWiki og NLLB elimineres.
V5	+ Wiki	3 485 841	**54,80 %**	**59,30 %**	Høyeste oppnådde retrieval-score.
V6	+ NDLA	3 675 493	54,70 %	59,20 %	NDLA testet, men fjernes (negativ mean-effekt; bare forbedring i BitextMining på bekostning av andre oppgaver).

Table 2: Oversikt over iterasjonsstegene i Greedy Forward Selection på MTEB(Scandinavian).

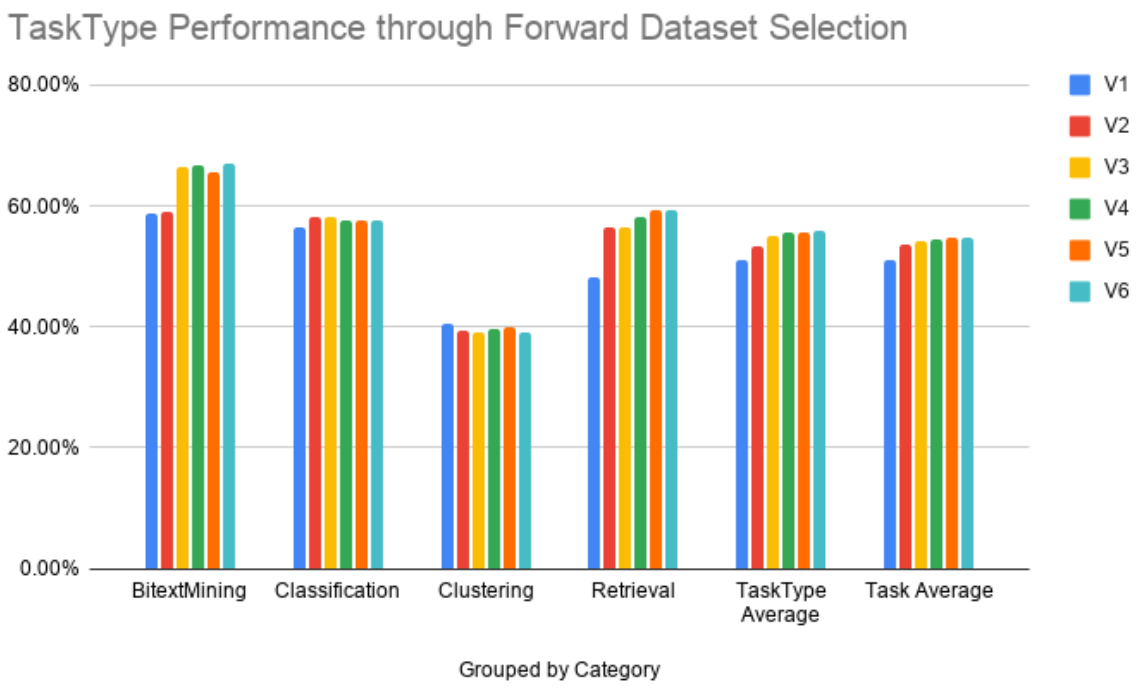


Figure 2: Ytelsesutvikling på MTEB (Scandinavian) gjennom iterasjonene (V1–V6). Prosessen gir en markant fremgang i både informasjonsgjenfinning (Retrieval) og gjennomsnittlig oppgaveytelse (Mean Task), uten stor nedgang på øvrige oppgavetyper.

Lingvistiske utfordringer for frasestrukturgrammatikker

Det er en rekke lingvistiske fenomener som er utfordrende for implementerte frasestrukturgrammatikker. I denne presentasjonen vil jeg ta for meg fire fenomener i norsk, nemlig VP-fronting, ellipser, frie komplementer og høyredislokering. Jeg vil presentere disse fenomenene og vise hvilke utfordringer de har for grammatikkimplementasjoner.

Det første fenomenet, VP-fronting, går ut på at en topikaliserer hele VP-er, som i (1a), hvor VP-en *spise den pizzaen* er topikalisert. Det er også mulig å topikalisere bare deler av VP-er, som i (1b) (delvis VP-fronting), hvor bare hovedverbet *spise* er topikalisert, mens objektet er realisert etter subjektet.

(1) a. Spise den pizzaen vil jeg ikke.

b. ? Spise vil jeg ikke den pizzaen.

Det andre fenomenet, ellipser, går ut på at en koordinerer to eller flere setninger eller verbfraser, hvor verbet mangler i konjunktene som kommer etter det første. (2a) er et eksempel på en ellipse hvor to setninger er koordinert, og (2b) er et eksempel på en ellipse hvor to VP-er er koordinert.

(2) a. Jeg spiser isen og du sjokoladen.

b. Hun ga meg isen og deg sjokoladen.

Det tredje fenomenet, frie komplementer, dreier seg om at enkelte gradsadverb (ofte *så*) åpner for at et fritt komplement (en at-setning) kan opptre helt til slutt i setningen (se (3)).

(3) Han er så sen at jeg må smile.

Det fjerde fenomenet er høyredislokering. Ved høyredislokering koordineres to setninger mens den siste delen av den første konjunkten realiseres som en del av den andre konjunkten. I eksempel (4a) gjelder dette objektet *teen*, mens i eksempel (4b) gjelder dette hele VP-en *drikke teen*.

(4) a. Jeg lager og du drikker teen.

b. Jeg kan og du må drikke teen.

Samlet utgjør disse fenomenene en stor utfordring for trekkbaserte frasestrukturgrammatikker, som LFG og HPSG, siden hovedverbet i setningen som regel er antatt å være opphavet til linkingen mellom semantikken og argumentstrukturen. Men hva skjer når verbet mangler i en konjunkt, som i (2) og (4b)? Og hva skjer når argumenter realiseres tilsynelatende uavhengig av verbet, som i (1b) og (3)? Jeg vil diskutere hvordan hvordan disse utfordringene kan løses i en inkrementell frasestrukturgrammatikk hvor konstituentstrukturen representeres av en stakk, og linkingen mellom syntaks og semantikk foregår i dekomponerte konstruksjoner.[1]

[1] GhatGPT ble brukt til å telle antall ord i sammendraget.

Luminoner: fragmentbasert fortolkning med språkmodeller

Lars G. B. Johnsen

Nasjonalbiblioteket

Sammendrag

Frekvenslister, kollokasjoner og konkordanser gjør store tekstsamlinger målbare, men sier ikke uten videre hvilken funksjon hvert enkelt tekstbelegg har. Nærlesing gir rikere fortolkninger, men skalerer dårlig. Vi presenterer luminoner som en mellomform: små, strukturerte og etterprøvbare fortolkninger av tekstfragmenter, produsert med støtte fra språkmodeller.

Et luminon er forankret i et konkret tekstfragment, vanligvis en konkordans med et målord og kontekst på begge sider. Språkmodellen klassifiserer fragmentet, vurderer eventuelle egenskaper og gir en kort begrunnelse. Resultatet kan aggregeres kvantitativt, men hvert datapunkt kan samtidig føres tilbake til teksten det bygger på.

Luminoner kan dannes både ovenfra og nedenfra. I en ovenfra-styrt analyse definerer forskeren kategoriene på forhånd og bruker modellen til å annotere forekomstene. I en nedenfra-styrt analyse kan modellen foreslå tilbakevendende egenskaper og kategorier i materialet. Disse kan deretter vurderes av forskeren, stabiliseres som et analytisk skjema og brukes i en ny gjennomgang. Luminonene produserer dermed nye, avledede data som ikke finnes eksplisitt i tekstsamlingen.

Til arbeidet har vi utviklet en annotasjonsapp der forskeren velger tekstmateriale, kategorier, kontekstlengde, modell og antall fragmenter per omgang. Appen sender fragmentene i grupper til blant annet GPT-4o-mini og GPT-5-mini og mottar strukturerte svar. Modellens forslag vises sammen med tekstbelegget og kan kontrolleres og korrigeres av forskeren.

Metoden er prøvd i flere undersøkelser av Nasjonalbibliotekets tekstsamlinger. I analyser av klima skilles det mellom meteorologisk omtale, inneklima, metaforisk bruk og forekomster uten tematisk relevans. I undersøkelser av «mindreårig» synliggjøres forskjellen mellom administrative og juridisk-seksuelle bruksmønstre. Andre forsøk omfatter maritime hendelser og kulturelle markører i barne- og ungdomslitteratur.

Bidraget diskuterer luminonet som språkteknologisk analyseenhet og demonstrerer appen som arbeidsredskap. Virkelig tekst fungerer som anker, forskerens kategorier som analytisk snitt og språkmodellen som relasjonell fortolker.

Norsk Bloggkorpus som et kasus for en KI-drevet pipeline for datarengjøring, anonymisering og tagging av målform og diskursemne

Olaf Mikkelsen, Universitetet i Sørøst-Norge

Norsk Bloggkorpus er et korpus på ca. 20 millioner ord bestående av 1001 norske blogger, publisert i perioden 2010-2021. Dataene ble innhentet høsten 2021 ved å mate en liste med startadresser inn i Trafilatura (Barbaresi 2021). Hver blogg er knyttet til følgende metadata for hver forfatter (der denne informasjonen var tilgjengelig): aldersgruppe (født tiår), kjønn og opprinnelse (landsdel). Korpuset er tagget med Humit-taggeren (Haug et al. 2023) og glossa-søkesystemet (Nøklestad et al. 2017) er integrert i brukergrensesnittet, som vil bli gjort tilgjengelig på www.bloggkorpus.uit.no. Før det kan gjøres tilgjengelig, må det imidlertid rengjøres og anonymiseres. Dette bidraget fokuserer på en KI-drevet pipeline for det pågående arbeidet med rengjøring og anonymisering og diskuterer fordelene ved et korpus som er tagget for målform og diskursemne.

Store språkmodeller har vist seg å være pålitelige til anonymisering (Deußer et al. 2025), og UiOs personverntrygge språkmodell GPT UiO, som kan brukes med røde data, ble valgt til å (1) rengjøre dataene for irrelevant informasjon (f.eks. navigasjonsmenyer) som ikke ble luket ut av Trafilatura, og (2) maskere sensitiv og personlig informasjon. Et prompt for å utføre disse oppgavene ble utformet iterativt ved at språkmodellen selv ble bedt om å finne svakheter i promptet, og ble så matet inn i språkmodellen sammen med tekstutdrag på ca. 30 000 tegn gjennom API-tilgang.

Siste skritt i pipelinen er tagging av målform og diskursemne. Da tekstene ble samlet inn, ble hver enkelt blogg knyttet til et emne og en målform. Dette viste seg å være en lite pålitelig måte å tagge dataene på, ettersom selv tematiske blogger kan inneholde svært ulike teksttyper, og enkelte forfattere veksler mellom ulike målformer med varierende innslag av dialektale trekk. Ved å tagge hvert enkelt blogginnlegg for emne og målform vil man kunne foreta mer presise søk som tillater å f.eks. inkludere innlegg på nynorsk som omhandler politikk, eller ekskludere alle innlegg som handler om matlaging eller hagearbeid. Potensielle fordeler ved en slik inndeling vil bli diskutert, og casestudier som illustrerer hvordan korpuset skiller seg fra andre eksisterende korpuser vil bli presentert.

Barbaresi, A. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. I *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 122–131.

Deußer, T., Sparrenberg, L., Berger, A., Hahnbück, M., Bauckhage, C., & Sifa, R. (2025, October). A survey on current trends and recent advances in text anonymization. I *IEEE 12th International Conference on Data Science and Advanced Analytics*, 1–9.

Haug, D. T. T., Yildirim, A., Hagen, K., & Nøklestad, A. (2023). I *Rules and neural nets for morphological tagging of Norwegian - Results and challenges*, 425–435.

Nøklestad, A., Hagen, K., Johannessen, J. B., Kosek, M. & Priestley, J. 2017. A modernised version of the Glossa corpus search system. I J. Tiedemann (red.): *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 251–254.

Norsk fagspråkkorpus

Norsk fagspråkkorpus er et korpus under utvikling gjennom Senter for norsk fagspråk i bruk (NORBRUK) ved Universitetet i Oslo (i samarbeid med Universitetet i Sørøst-Norge, OsloMet og Nasjonalbiblioteket.)

Senter for norsk fagspråk i bruk skal utvide kunnskapsgrunnlaget om, og ivareta bruken av, norsk som fagspråk i forskning, utdanning og forvaltning. Med dette ønsker senteret å bidra til å styrke demokrati, rettssikkerhet og deltakelse i akademia og samfunnet.

Gjennom Norsk fagspråkkorpus ønsker vi å tilgjengeliggjøre norske faglitterære tekster for forskere, noe som vil danne et sentralt empirisk grunnlag for forskning på norsk fagspråk i forskning, utdanning og forvaltning, og for analyse og utvikling av språkteknologiske ressurser for norsk fagspråk.

Korpuset er under utvikling, og vil bestå av fagbøker, tidsskrifter og vitenskapelige avhandlinger fra 1980 frem til i dag. De faglitterære tekstene omfatter mange ulike områder og fagfelt, og blir innhentet fra blant annet Nasjonalbiblioteket og Sikt. Vi jobber med å rense og øke kvaliteten på disse OCR-tekstene, og har som mål å oppnå høyest mulig datakvalitet på korpuset samtidig som det opprinnelige innholdet bevares.

Alle tekstfiler vil kategoriseres og tilgjengeliggjøres for forskere, hvor hver enkelt kan søke om tilgang for en tidsbegrenset periode på Universitetet i Oslos plattform for forskningsprosjekter og samarbeid i forskning (Educloud Research). Dette vil gjøre det mulig å forske komputasjonelt på tekstene og kunne finjustere modeller for oppgaver relatert til norsk fagspråk. Korpuset vil ikke gjøres tilgjengelig for trening av språkmodeller.

Deler av korpuset vil tilgjengeliggjøres gjennom søkeportalen Glossa, en allerede eksisterende tjeneste utviklet av UiO Tekstlab, som nå driftes av Humit. Glossa har et brukervennlig grensesnitt, der både utforming og metadatameny kan tilpasses etter egne behov. Søkegrensesnittet tilbys i flere varianter: en enkel versjon som ligner et vanlig Google-søk, en mer avansert versjon med valgbare funksjoner som lemma- og ordklassesøk, samt en løsning for søk ved hjelp av regulære uttrykk.

I forbindelse med Norsk SpråkteknologiKonferanse ønsker vi å få innspill som kan gi oss en bedre forståelse av hvordan korpuset vil bli brukt i praksis slik at korpuset kan utformes og tilgjengeliggjøres best mulig. Vi ønsker blant annet innspill på hvordan materialet bør kategoriseres og organiseres, hvilke filformater som er mest hensiktsmessige, og hvilke deler av korpuset det er ønskelig å ha tilgjengelig i Glossa.

Om Benchmarking av Sammendrag og Informasjonsgjenfinning for Norsk på Frontier-Språkmodeller og Norske Språkmodeller

Ryan Anthony Marinelli

SFI MediaFutures, University of Oslo

September 2026

1 Introduksjon

Målet med dette arbeidet er å måle evnen til norske språkmodeller sammenlignet med frontier-språkmodeller innenfor nyhetsdomenet. Vi evaluerer to benchmarker innenfor dette området. Først bruker vi NorQuAD, en informasjonsgjenfinningsbenchmark fra Ivanova et al. [2023]. Den består av 4 752 spørsmål-svar-par, hvorav 2 398 kommer fra nyhetsdomenet, basert på 403 annoterte nyhetsartikler. Den andre benchmarken er NorSumm fra Touileb et al. [2025]. Denne benchmarken består av 33 sammendrag på bokmål og nynorsk. Hvert sammendrag har et grunnsannhetssammendrag (ground truth) skrevet av norske morsmålsbrukere. Forskningsspørsmålet i dette arbeidet er følgende: *Er det et evnegap mellom norske språkmodeller og frontier-modeller på norsk i nyhetsdomenet?*

2 Oppsummeringsevne

Vi bruker ROUGE for å måle evnen til å generere sammendrag. ROUGE-1 måler overlappet av unigrammer mellom sammendraget og grunnsannheten, mens ROUGE-2 bruker bigrammer. ROUGE-L er det lengste overlappet delt med grunnsannheten, i samme rekkefølge.

Den beste modellen er Fable. Dette er ikke overraskende, gitt størrelsen på modellen. Den nest beste modellen kommer fra Anthropic, som vist i Tabell 1.

Table 1: NorSumm — leksikalske sammendragsmetrikker

Modell	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Norske modeller			
Borealis-27B-instruct-preview (bf16)	0.4560	0.2367	0.3100
Borealis-27B (bf16)	0.4252	0.2190	0.2972
Anthropic			
Claude-Fable-5	**0.5326**	**0.2626**	**0.3464**
Claude-Fable-5.1	0.5228	0.2494	0.3276
Claude-Opus-5	0.5137	0.2473	0.3224
Claude-Sonnet-5	0.4775	0.2103	0.3031
OpenAI			
GPT-6-Astra	0.4813	0.2133	0.3166
GPT-5.6-Sol	0.4517	0.1913	0.2904
Google			
Gemini-3.5-Flash	0.4594	0.2016	0.2963
Gemma-3-27B-it-base	0.4507	0.1744	0.2701
Mistral AI			
Mistral-Small-3.2-24B	0.4644	0.2064	0.3005
Alibaba			
Qwen3.6-27B	0.4475	0.1923	0.2844

3 Informasjonsgjenfinning

Anthropic har ikke et embeddings-API, så vi bruker ikke Opus og Fable med NorQuAD. Hver modell får en artikkel med et spørsmål og gir et svar. Grunnsannheten sammenlignes med svaret på en SQuAD-lignende måte. GPT 6 er den sterkeste modellen i disse oppgavene med en F1 på 0,91.

Table 2: Modellytelse etter selskap

Selskap	Modell	Ant. spørsmål	Ant. tomme pred.	EM	F1
Google	gemini-3.5-flash	300	0	0.7333	0.9068
	gemma-3-27b-it-base	300	0	0.6233	0.8575
Anthropic	claude-fable-5-1	300	0	0.7367	0.9034
	claude-fable-5	300	9	0.7233	0.8889
	claude-sonnet-5	300	0	0.6667	0.8832
	claude-opus-5	300	11	0.6567	0.8538
OpenAI	gpt-6-astra	300	0	**0.7467**	**0.9142**
	gpt-5.6-sol	300	0	0.6533	0.8779
NbAiLab	borealis-27b	300	0	0.6367	0.8343
	borealis-27b-instruct-preview	300	0	0.6267	0.8373

Når vi bruker konvensjonelle metoder for informasjonsgjenfinning med embeddings, finner vi resultatene i Tabell 3. Den beste modellen er Microsofts E5 med en recall@1 på 0,73.

Table 3: Evalueringsresultater for gjenfinning med embeddingmodeller (2 398 spørringer per modell)

Modell	recall@1	recall@5	recall@10	mrr@10	ndcg@10	sekunder
BAAI						
bge-m3	0.7031	0.8574	0.9008	0.7714	0.8028	691.1
Alibaba						
Qwen3-Embedding-0.6B	0.6159	0.7736	0.8253	0.6843	0.7182	3762.3
Microsoft						
multilingual-e5-small	0.6922	0.8528	0.8924	0.7623	0.7939	96.5
multilingual-e5-base	0.7077	0.8632	0.9133	0.7759	0.8091	237.6
multilingual-e5-large	**0.7323**	**0.8791**	**0.9162**	**0.7977**	**0.8266**	638.0
Nasjonalbiblioteket (NbAiLab)						
nb-sbert-base	0.4541	0.6735	0.7510	0.5489	0.5974	68.2
borealis-embed-212m	0.5905	0.7681	0.8240	0.6693	0.7067	1412.3
nb-sentence-bert-base-mnli-test	0.5375	0.7799	0.8515	0.6400	0.6910	240.4
Perplexity AI						
pplx-embed-v1-0.6b	0.6856	0.8490	0.8962	0.7563	0.7901	1343.7
sentence-transformers (UKP Lab)						
paraphrase-multilingual-mpnet-base-v2	0.4135	0.6193	0.6902	0.5040	0.5487	80.3

En gjenrangerer (reranker) brukes til å rangere gjenfunne dokumenter på nytt. Vi bruker en flerspråklig MiniLM cross-encoder trent på mMARCO Sentence-Transformers [2021]. NorSumm brukes til finjustering. Dataene brukes til å lage tre referansesammendrag. Dette gir 90 par til trening for en positiv gruppe. Negative sammendragspar brukes sammen med disse for 450 treningspar med binære etiketter (labels).

Table 4: Evaluering av gjenrangering på NorQuAD-gjenfinning (2 398 spørringer per rad)

Reranker	recall@1	recall@5	recall@10	mrr@10	ndcg@10
Bi-encoder: intfloat/multilingual-e5-large					
Ingen	0.7323	0.8791	0.9162	0.7977	0.8266
mmarco-mMiniLMv2-L12-H384-v1	0.8015	**0.9083**	0.9295	0.8491	0.8689
mmarco-mMiniLMv2 (NorSumm-finjustert)	**0.8103**	0.9074	**0.9308**	**0.8546**	**0.8732**
Bi-encoder: NbAiLab/borealis-embed-212m					
Ingen	0.5905	0.7681	0.8240	0.6693	0.7067
mmarco-mMiniLMv2-L12-H384-v1	0.7606	0.8495	**0.8649**	0.8001	0.8161
mmarco-mMiniLMv2 (NorSumm-finjustert)	**0.7673**	**0.8507**	0.8645	**0.8040**	**0.8189**

4 Resultater

Når de evalueres med norske benchmarker, er frontier-modeller bedre enn norske språkmodeller på begge oppgavene.

4.1 Informasjonsgjenfinning

Vi rapporterer recall@1. Gapet mellom multilingual-e5-large og borealis-embed-212m er 0,14 med embeddingmodeller alene og 0,04 med gjenrangerer. Gjenrangereren gir +0,18 for borealis-embed-212m og +0,08 for multilingual-e5-large.

4.2 Sammendrag

På sammendrag er gapet 0,08 i ROUGE-1.

5 Konklusjon

Det bør investeres mer i norske språkmodeller for å tette gapet til frontier-modeller og styrke digital suverenitet. For å forbli relevante må norske modeller utvikle større kapasitet. Gjenrangering er samtidig det mest lovende tiltaket vi har målt, og gevinsten er størst for den norske modellen. Vi trenger derfor mer utvikling av gjenrangerere for norsk.

References

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. NorQuAD: Norwegian question answering dataset. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.17/.

Sentence-Transformers. Cross-Encoder for multilingual MS MARCO. https://huggingface.co/cross-encoder/mmarco-mMiniLMv2-L12-H384-v1, 2021. Hugging Face model card.

Samia Touileb, Vladislav Mikhailov, Marie Ingeborg Kroka, Lilja Øvrelid, and Erik Velldal. Benchmarking abstractive summarisation: A dataset of human-authored summaries of Norwegian news articles. In Richard Johansson and Sara Stymne, editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 729–738, Tallinn, Estonia, March 2025. University of Tartu Library. ISBN 978-9908-53-109-0. URL https://aclanthology.org/2025.nodalida-1.73/.

Radikale språkbrukere blant konservative prateroboter

Denne presentasjonen legger fram ett avsluttet og to pågående arbeider i doktorgradsprosjektet mitt, som handler om hvordan endringer i skrivenorm påvirker utdata fra ulike maskinlæringsmodeller.

Den første delen handler om et rettearbeid som ble utført som del av et dugnadsprosjekt ("shared task") ved den internasjonale maskinoversettelseskonferansen i 2025 (WMT25)[1]. Sammen med to kollegaer korrigerte jeg en rekke feil i bokmålsdelen av FLORES-datasettet Mæhlum et al. (2025). Datasettet viste en rekke ugrammatiskheter, uoversatte termer og anglisismer. I tillegg omarbeidet vi det nye bokmålsmaterialet og lagde en parallell versjon på radikalt bokmål, basert på Bokmålsordboka[2] I eksempelet under er en setning gjengitt både på konservativt/moderat bokmål og på radikalt bokmål:

(1) [...] masse krem**fløte** (ikke **melke**skum) og te blir servert uten **melk**. (MBM)

[...] masse krem**fløyte** (ikke **mjølke**skum) og te blir servert uten **mjølk**. (RBM)

Dette radikale datasettet danner grunnlaget for undersøkelsene i del 2, der jeg ser på hvordan normvariasjon påvirker utdata i en maskinoversettelseskontekst. Et utvalg modeller ble undersøkt gjennom instruksbasert oversetting ved temperatur på 0,0 (nær-deterministisk). Undersøkelsen viser at omtrent en firedel av de oversatte setningene fra radikalt bokmål skiller seg fra de moderate på ulikt vis, blant annet ved andre leksikalske valg. Prosjektet går videre ut på å kartlegge variansens natur og spredning i takt med økt temperatur.

Til slutt presenterer jeg i del 3 arbeid på ulike prateroboters evne til å gi tilbakemelding på et lite sett tekster på radikalt bokmål, der rollen deres som aktører i diskursen er i fokus. Modellene brukes gjennom GPT UiOs tjeneste, der flere ulike modeller gjøres tilgjengelige for studenter og ansatte. Jeg viser at de ulike systemene ikke overraskende nok gjentar en del stereotypiske holdninger til radikale elementer, men at dette ikke er jevnt fordelt. Dette viser at prateroboter har en vei å gå for å fullstendig kunne dekke hele bokmålsnormen.

Presentasjonen er omformet basert på en lignende presentasjon ved ICLaVE 13 i Lausanne.

Kildehenvisning

Mæhlum, P., Næss Evensen, A., and Scherrer, Y. (2025). Improved Norwegian Bokmål translations for FLORES. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Tenth Conference on Machine*

[1] `https://www2.statmt.org/wmt25/]`

[2] `ordbokene.no`

Translation, pages 1124–1132, Suzhou, China. Association for Computational Linguistics.

Revesans: effektivt søk i kunnskaps- og styringsdokumenter for forvaltningsrevisjon

Murhaf Fares
Riksrevisjonen
mufa@riksrevisjonen.no

Emanuele Lapponi
Riksrevisjonen
emla@riksrevisjonen.no

Mimi Uyen Dan Nguyen
Riksrevisjonen
mudn@riksrevisjonen.no

Forvaltningsrevisjon bygger i stor grad på offentlig tilgjengelige dokumenter: stortingsmeldinger, tildelingsbrev, etatsrapporter, evalueringer og utredninger. Disse samles i søkemotoren Kudos (`https://kudos.dfo.no`), driftet av Direktoratet for forvaltning og økonomistyring (DFØ). Kudos tilbyr kun nøkkelordsøk, men en revisjon starter vanligvis med et spørsmål, ikke et begrep. Et nøkkelord bærer dessuten ingen intensjon: systemet skiller ikke tilfeldig omtale fra substansiell behandling. Dokumentene er også lange, slik at hver falsk positiv koster lesetid.

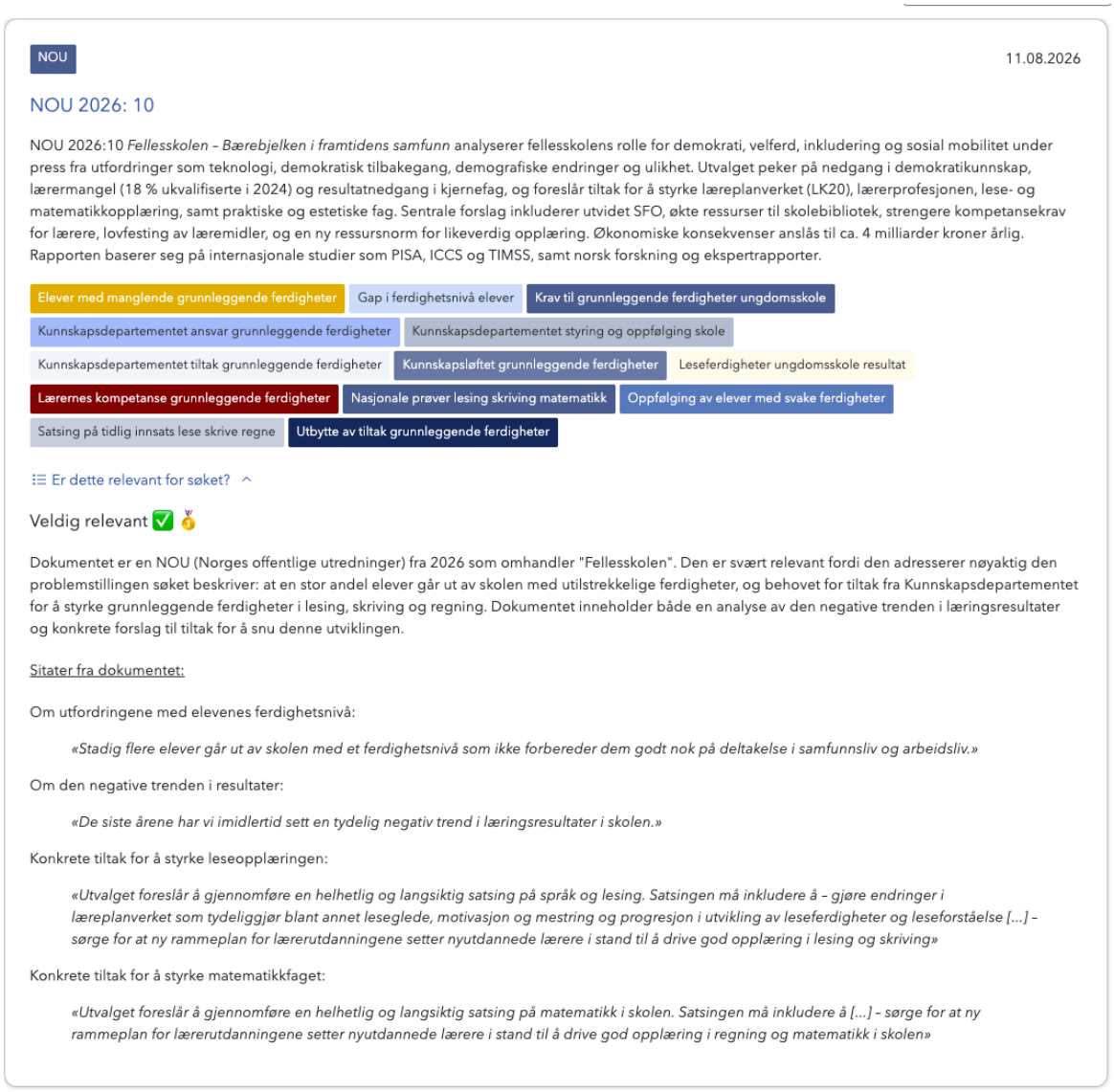
Revesans er et informasjonsgjenfinningssystem bygd på det samme korpuset, men utformet rundt revisjonsoppgaven framfor et nøkkelordsøk. Alle dokumenter blir stykket opp, vektorrepresentert, oppsummert og tilordnet emner og nøkkelord ved hjelp av språkmodeller. Dokumentene blir deretter indeksert i en søkedatabase for en hybrid av et semantisk og leksikalsk søk. Høsting og prosessering utføres kontinuerlig, slik at indeksen blir oppdatert når nye dokumenter publiseres.

I stedet for nøkkelord skriver revisoren for eksempel bakgrunnen for og formålet med selve revisjonsoppgaven. Systemet utleder et sett med spørringer fra denne beskrivelsen. Hver spørring gir en poengsum for likhet mot dokumentet, og den endelige rangeringen tar hensyn til både den samlede poengsummen og hvor mange spørringer som treffer over en gitt terskel. Hvert treff vises med den forhåndsgenererte oppsummeringen og med spørringene som traff, sortert etter best resultat.

Revisoren kan i tillegg be om en 'relevansvurdering' av resultater: systemet leser da dokumentet opp mot den opprinnelige beskrivelsen og returnerer en vurdering med sitater fra teksten, knyttet til den delen av søketeksten de besvarer. Figur 1 viser et eksempel av et søkeresultat med oppsummering, treff og relevansforklaring.

Hele prosesseringskjeden kjører på egen maskinvare med åpne modeller som vi drifter selv, og ingen del av den benytter eksterne tjenester. Det gir frihet til å reprosessere om lag 50 000 dokumenter for å f.eks. eksperimentere med forskjellige vektoriserings- og oppsummeringsstrategier, og det sikrer at bakgrunnen og formålet, som kan inneholde sensitiv informasjon, aldri forlater virksomheten.

Evaluering. Vi evaluerer mot allerede gjennomførte revisjoner: bakgrunn og formål (slik de opprinnelig ble formulert i revisjonsdokumentet) sendes til Revesans, og vi vurderer dette mot dokumentene som revisjonen endte med å sitere. Det gir et realistisk testsett, ettersom de siterte dokumentene er et resultat av at revisorer har vurdert dem som relevante etter lesing. Den kvantitative evalueringen er fortsatt under utvikling, og vil veilede videreutvikling og forbedring av systemet. Sånn sett vil Revesans kunne brukes til nedstrømsevaluering av f.eks. åpne språkmodeller, hvor bokmåls- og nynorskforståelse er særlig relevant.



Figur 1: Brukergrensesnittet viser oppsummering, treff blant de genererte spørringene, og relevansforklaring for det høyest rangerte dokumentet for søket «**Bakgrunn:** Stortinget har over tid hatt høge ambisjonar for dei grunnleggjande ferdigheitene til elevane. Kunnskapsdepartementet, som har det overordna ansvaret, har gjennomført mange tiltak for å leggje til rette for ei opplæring som styrkjer lese-, skrive- og rekneferdigheitene. Likevel går ein stor andel av elevane ut av grunnskulen utan tilstrekkelege ferdigheiter til å lese, skrive og rekne. **Formål:** Målet med undersøkinga er å vurdere i kva grad Kunnskapsdepartementet har bidrege til at grunnskulen gir elevane tilfredsstillande grunnleggjande ferdigheiter i lesing, skriving og rekning.»

Sentimentanalyse på entitetsnivå

Sammendrag av doktorgradsforskningen

Egil Rønningstad

Hva er ELSA?

Sentimentanalyse på entitetsnivå (ELSA) er sentimentklassifisering per entitet i lengre dokumenter, der entitetene er personer og organisasjoner (Rønningstad, Velldal og Øvrelid 2022). Oppgaven bryter med vanlig praksis, som enten er setningsavgrenset klassifisering (f.eks. Mitchell mfl. 2013; Øvrelid mfl. 2020) eller kun én generell sentimentklassifisering per dokument (f.eks. Maas mfl. 2011; Velldal mfl. 2018). Vår tilnærming tar hensyn til at samme dokument kan framstille ulike entiteter forskjellig, og at sentiment kan relateres til en entitet uten at navnet nevnes i samme setning.

Datasettet for ELSA

Vi har manuelt annotert det første datasettet som er spesielt utformet for ELSA: 412 norske profesjonelle anmeldelser med 2479 entiteter (Rønningstad, Klinger mfl. 2024). Fem norskspråklige annotatører med språkfaglig bakgrunn ble opplært gjennom flere runder med reviderte retningslinjer. Hvert dokument ble annotert av én annotatør og deretter kuratert. Alle fem annoterte testdelen uavhengig.

For hver entitet annoterte vi tekstens samlede polaritet og intensitet. Deretter identifiserte vi alle sentimentbærende relasjoner per entitet: hvilke setninger som bidrar til formidlingen av sentiment, det lokale sentimentuttrykket og relasjonen mellom setningen og entiteten. Slik kan vi analysere de lingvistiske fenomenene som oppstår når sentiment relateres til entiteter. Relasjonene omfatter direkte navneomtale, koreferanse og *bridging*, for eksempel når sentiment rettet mot et kunstverk overføres til skaperen eller når sentiment overføres mellom en organisasjon og et medlem.

Analysen viser hvorfor ELSA ikke kan reduseres til sentiment ved navneomtaler. Bare 36,7 % av de annoterte sentimentrelasjonene forekommer i setninger som nevner entiteten ved navn. Bridging utgjør om lag 35 % av relasjonene, nesten like mye som navneomtaler (Rønningstad, Klinger mfl. 2024).

Gjennomsnittlig Cohens κ mellom hver enkelt annotatør og den kuraterte fasiten var 0,71. Uenighet ble ikke bare behandlet som feil. Vi analyserte variasjonen mellom annotatørene som mulige uttrykk for ulike, men likeverdige lesninger av sentimentinnholdet i teksten. En sammenligning med GPT-4 viste lavere samsvar for modellens etiketter enn for hver av de fem menneskelige annotatørene (Rønningstad, Velldal og Øvrelid 2024).

Modellering

Vi sammenlignet forhåndstrente modeller, både internasjonale og norske, og vurderte finjustering opp mot prompting av store språkmodeller. For modeller med BERT-arkitektur og begrenset kontekstvindu utviklet vi metoder for å velge ut entitetsrelevante deler av teksten. For dekoder-

modeller ga finjustering bedre resultater enn prompting. Finjustering for *pairwise comparison* ga de beste og mest stabile samlede resultatene (Rønningstad, Klinger mfl. 2026).

Dette sammendraget er generert med KI basert på mine artikler og mine stikkord for hva sammendraget skulle inneholde. Sammendraget er sjekket og godkjent av meg.

Referanser

Mitchell, M., J. Aguilar, T. Wilson og B. Van Durme (2013). «Open Domain Targeted Sentiment». I: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Red. av D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu og S. Bethard. Seattle, Washington, USA: Association for Computational Linguistics, s. 1643–1654. URL: https://aclanthology.org/D13-1171/.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng og C. Potts (2011). «Learning Word Vectors for Sentiment Analysis». I: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Red. av D. Lin, Y. Matsumoto og R. Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, s. 142–150. URL: https://aclanthology.org/P11-1015/.

Rønningstad, E., R. Klinger, L. Øvrelid og E. Velldal (2024). «Entity-Level Sentiment: More than the Sum of Its Parts». I: *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Red. av O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc og S. Tafreshi. Bangkok, Thailand: Association for Computational Linguistics, s. 84–96. URL: https://aclanthology.org/2024.wassa-1.8/.

Rønningstad, E., R. Klinger, L. Øvrelid og E. Velldal (2026). «Entity-Level Sentiment Analysis with Sentence Relevance Detection». I: *Proceedings of the Fifteenth Language Resources and Evaluation Conference*. Red. av S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek og A. Toral. Palma de Mallorca, Spain: ELRA Language Resource Association, s. 8040–8055. URL: https://aclanthology.org/2026.lrec-1.638/.

Rønningstad, E., E. Velldal og L. Øvrelid (2022). «Entity-Level Sentiment Analysis (ELSA): An Exploratory Task Survey». I: *Proceedings of the 29th International Conference on Computational Linguistics*. Red. av N. Calzolari mfl. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, s. 6773–6783. URL: https://aclanthology.org/2022.coling-1.589/.

Rønningstad, E., E. Velldal og L. Øvrelid (2024). «A GPT among Annotators: LLM-based Entity-Level Sentiment Annotation». I: *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*. Red. av S. Henning og M. Stede. St. Julians, Malta: Association for Computational Linguistics, s. 133–139. URL: https://aclanthology.org/2024.law-1.13/.

Velldal, E., L. Øvrelid, E. A. Bergem, C. Stadsnes, S. Touileb og F. Jørgensen (2018). «NoReC: The Norwegian Review Corpus». I: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Red. av N. Calzolari mfl. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1661/.

Øvrelid, L., P. Mæhlum, J. Barnes og E. Velldal (2020). «A Fine-grained Sentiment Dataset for Norwegian». eng. I: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Red. av N. Calzolari mfl. Marseille, France: European Language Resources Association, s. 5025–5033. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.618/.

Skjevheter i store språkmodeller ved CEFR-basert vurdering av tekster fra norske andrespråksinnlærere

Gloria Anne Babile Kasangu og Samia Touileb

Sammendrag

Interessen for å bruke store språkmodeller (LLM-er) til automatisk essayvurdering (AES) har økt. Dette inkluderer vurdering av tekster skrevet av andrespråksinnlærere basert på det felles europeiske rammeverket for språk (CEFR). I motsetning til tradisjonelle AES-systemer er LLM-baserte systemer i stand til å gi mer fleksible og helhetlige vurderinger. Selv om noen studier har undersøkt bruk av LLM-er til AES med fokus på vurderingsytelse og samsvar med menneskelige sensorer, har svært lite oppmerksomhet blitt rettet mot hvordan demografisk informasjon kan påvirke modellgenererte vurderinger, særlig i en norskspråklig kontekst. Vi fyller dette kunnskapgapet ved å systematisk evaluere ytelsen til flerspråklige og norske LLM-er i CEFR-basert essayvurdering, samt undersøke om demografiske egenskaper påvirker modellgenererte vurderinger. Vi er spesielt interessert i skjæringspunktet mellom pedagogisk psykologi, kognitiv vitenskap, og språkteknologi.

Vi undersøker sosial skjevhet i LLM-er ved å analysere hvordan skjevheter knyttet til kjønn, alder, morsmål, nasjonalitet og yrke kan komme til uttrykk i automatisk CEFR-vurdering av norske andrespråksinnlæreres tekster. Vi fokuserer på fordelingsmessige skader (*allocational harms*) (Gallegos et al., 2024). Skjevhet vurderes gjennom kontrollerte eksperimenter der demografisk informasjon i inndata systematisk varieres, og endringer i modellresponsene analyseres. Vi fokuserer på to forskningsspørsmål: FS1) Hvordan varierer ytelsen til norske og flerspråklige LLM-er på en automatisert CEFR-vurderingsoppgave på tvers av ulike bruksstrategier? Og FS2) I hvilken grad påvirker demografiske egenskaper CEFR-prediksjonene til norske og flerspråklige LLM-er?

For å besvare disse spørsmålene anvender vi etablerte metoder for kontrafaktisk evaluering av skjevhet. Vi undersøker om genererte vurderingsprediksjoner varierer systematisk på tvers av demografiske egenskaper (alder, kjønn, nasjonalitet, morsmål og yrke). Vi evaluerer to flerspråklige modeller, Llama 3 (Grattafiori et al., 2024) og EuroLLM (Martins et al., 2025), samt en norsk modell, NorMistral (Samuel et al., 2025). Vi introduserer et stort kontrafaktisk datasett bestående av mer enn 1.700.000 demografiske varianter[1]. Resultatene viser at vår eksperimentellstrategi basert på CEFR-deskriptorer gjennomgående gir best ytelse på tvers av de testede modellene, men at ingen av modellene oppnådde et tilfredsstillende nivå for automatisk essayvurdering. Vi viser også at alle modellene utviser demografisk sensitivitet under kontrafaktiske perturbasjoner, noe som tyder på at demografisk informasjon kan påvirke CEFR-prediksjoner selv når essayteksten forblir uendret. NorMistral viser den sterkeste demografiske sensitiviteten, mens de øvrige modellene produserer statistisk signifikante endringer i prediksjonene knyttet til spesifikke demografiske egenskaper.

Funnene våre tyder sterkt på at ingen av de evaluerte modellene per i dag er egnet for bruk i automatiserte CEFR-vurderingssituasjoner. Utover den begrensede prediksjonsnøyaktigheten indikerer de observerte demografiske sensitivitetene at modellene risikerer å produsere systematisk skjeve vurderinger som er knyttet til bestemte demografiske profiler.

Referanser

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

[1]Kode benyttet i denne studien er offentlig tilgjengelige på GitHub: `https://github.com/GloriaABK/KRAKEN`

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., and et al. (2024). The llama 3 herd of models.

Martins, P. H., Alves, J., Fernandes, P., Guerreiro, N. M., Rei, R., Farajian, A., Klimaszewski, M., Alves, D. M., Pombal, J., Boizard, N., Faysse, M., Colombo, P., Yvon, F., Haddow, B., de Souza, J. G. C., Birch, A., and Martins, A. F. T. (2025). Eurollm-9b: Technical report.

Samuel, D., Mikhailov, V., Velldal, E., Øvrelid, L., Charpentier, L. G. G., Kutuzov, A., and Oepen, S. (2025). Small languages, big models: A study of continual training on languages of norway.

Sosiopolitiske konflikthendelser
Utfordringer, muligheter og resultater av automatisk uthentelse

Helene Bøsei Olsen
Språkteknologigruppa (LTG), Universitet i Oslo

Hendelser i tekst kan forstås som handlinger eller tilstandsendringer som involverer deltagere på et bestemt tidspunkt og sted. Hvordan hendelser defineres og representeres varierer mellom fag- og forskningsfelt.

I fred- og konfliktforskning er strukturerte data om sosiopolitiske hendelser, som væpnet konflikt og protester, en viktig ressurs for å analysere mønstre, utvikling og konsekvenser av konflikt og uro over tid og på tvers av geografiske områder (Raleigh et al., 2010; Sundberg and Melander, 2013). Manuell uthenting av slike data er imidlertid tid- og ressurskrevende, noe som skaper behov for automatiserte metoder.

Automatisk uthenting av hendelser er en veletablert oppgave i språkteknologi (NLP), men oppgaveformuleringen bygger på antagelser om hvordan hendelser uttrykkes og representeres.

I fred- og konfliktforskning kan hendelser forstås som sosiopolitiske fenomener, og defineres ut fra samfunnsvitenskapelig teori og representeres med variabler som involverte aktører, lokasjon, og tidspunkt. I NLP identifiseres hendelser tradisjonelt gjennom tekstlige utløseruttrykk som knytter forhåndsdefinerte argumenter og attributter til hendelsen.

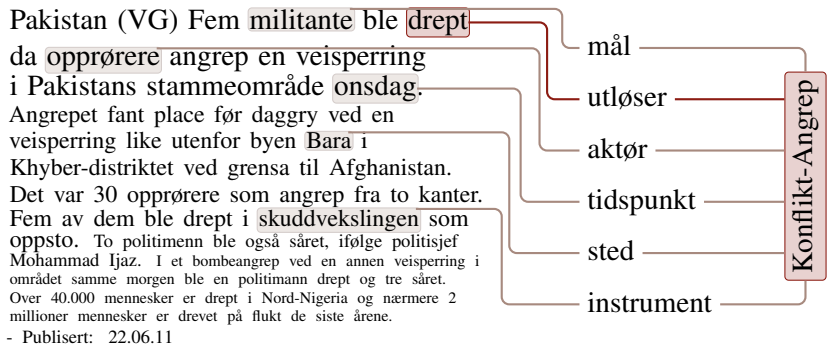
Figur 1 illustrerer hvordan disse forskjellene fører til at samme hendelse representeres på ulike abstraksjonsnivåer. Mens NLP-annotasjoner forankres i konkrete tekstsegmenter, representerer sosiopolitiske databaser ekspertfortolkninger som kan inneholde informasjon utover det som eksplisitt uttrykkes i teksten. For eksempel omtaler nyhetsartikkelen konfliktaktørene som "*militante*" og "*opprørere*", mens databasen normaliserer aktørnavnene basert på kontekst. Tilsvarende angir teksten tidspunktet som "*onsdag*", mens databasen utleder den konkrete datoen fra publiseringsdatoen.

Disse forskjellene skaper et abstraksjonsgap mellom den språklige realiseringen av hendelsen og den strukturerte representasjonen som kreves for samfunnsvitenskapelig analyse. En systematisk kartlegging av ressurser for sosiopolitisk hendelsesuthenting (Olsen et al., 2024) viser dessuten at sosiopolitiske representasjoner typisk er knyttet til dokumentene der kildetekstene sjelden er offentlig tilgjengelige.

På bakgrunn av disse forskjellene har vi, i samarbeid med fredsforskere ved Universitetene i Oslo og Uppsala, utviklet et åpent datasett med nyhetsartikler og annotasjoner fra domene-eksperter (Simon et al., 2025). Vi finjusterer store språkmodeller til å generere strukturerte representasjoner av konflikthendelser, der informasjonen i hovedsak ikke kan knyttes til eksplisitte tekstsegmenter - en oppgave vi kaller abstraktiv hendelsesanalyse.

Resultatene viser at modellene kan generere tilfredsstillende resultater for mange hendelser, men utfordres av nye konfliktaktører som ikke forekommer i treningsdataene, og datoer som avviker fra publiseringsdatoen. Videre eksperimenter bekrefter at generative modeller har utfordringer med både eksplisitt og implisitt informasjon, men at økt modellstørrelse gir forbedret ytelse (Olsen et al., 2026).

Arbeidet viser hvordan informasjonsbehov i konkrete anvendelser kan utfordre og utvide etablerte formuleringer av NLP-oppgaver, og bidrar med nye ressurser, metodiske perspektiver og empiriske analyser for automatisk uthenting av sosiopolitisk hendelsesinformasjon.



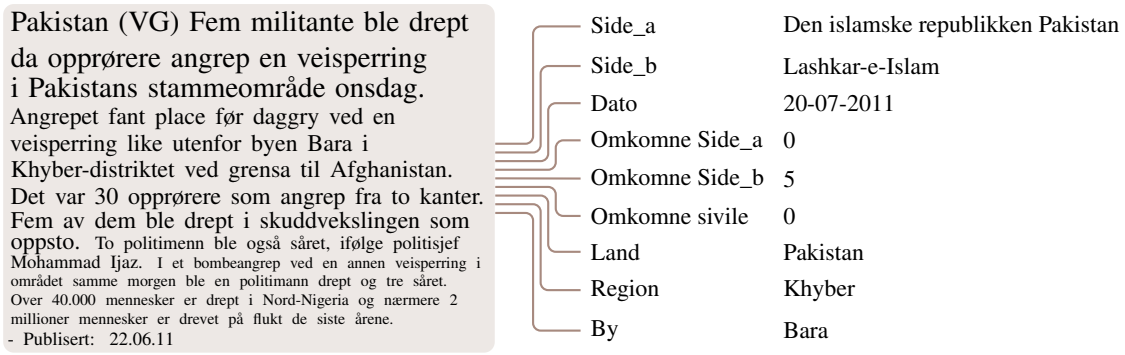


Figure 1: Eksempel på ulik hendelsesrepresentasjon i et NLP-dataset (øverst) og en databaser for fred- og konfliktforskning (nederst) NLP-annotasjonen er forankret i konkrete tekstsegmenter, mens den sosiopolitiske representasjonen inneholder normaliserte og utledede verdier. Annotasjonene er konstruerte og illustrerer hvordan artikkelen kunne vært representert dersom den var inkludert i de respektive ressursene. Nyhetsartikkelen er hentet fra VG her

Referanser

Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53, St. Julians, Malta. Association for Computational Linguistics.

Helene Bøsei Olsen, Erik Velldal, and Lilja Øvrelid. 2026. Muc-4 revisited: Document-level event analysis beyond span-based arguments. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7766–7780, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Clionadh Raleigh, Andrew Linke, Havard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Étienne Simon, Helene Bøsei Olsen, Ramón Carreño, Rahul Mishra, Nikolay Arefyev, Mert Can Yilmaz, Lilja Øvrelid, and Erik Velldal. 2025. Abstractive event analysis of armed conflicts: Introducing the UCDP-AEC dataset. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 104–119, Hannover, Germany. HsH Applied Academics.

Ralph Sundberg and Erik Melander. 2013. Introducing the UCDP georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

I prosjektet Norchron utvikler vi et korpus med norsk språk fra år 200 til 1814. Hele korpuset annoteres med ordklasser og lemma, og utvalgte deler (minst 200 000 ord) skal også annoteres med morfosyntaktiske trekk og syntaktiske funksjoner. I den nåværende fasen av prosjektet fokuserer vi på ordklasser og lemma.

Prosjektet utvikler språkteknologiske verktøy for eldre norsk språk, deriblant verktøy for automatisk grammatisk tagging og lemmatisering. I tillegg til å bli gjort allment tilgjengelig skal verktøyene også brukes for å lette annotasjonsarbeidet i prosjektet. Dette skjer ved en iterativ prosess der språkmodeller trenes på manuelt annotert tekst og anvendes på uannotert materiale, som korrigeres og legges til treningsmaterialet til bruk i neste iterasjon.

Vi kunne tatt utgangspunkt i BERT-modeller for moderne norsk, som NorBert (Kutuzov et al. 2021) eller NB-BERT (Kummervold et al. 2021), men siden det meste av språket i Norchron ligger nærmere islandsk enn moderne norsk, tar vi i stedet utgangspunkt i en islandsk BERT-modell, IceBERT (Snæbjarnarson et al. 2022). Modellen er trent på 2,7 milliarder islandske tokens fra ulike tidsperioder og varierende sjangrer, deriblant sagaer, moderne skjønnlitteratur, faglitteratur og web-tekst.

Modellen blir videretrent med “masked language modelling” (MLM) med konvensjonell maskeringsrate på 15% (Devlin et al. 2018) på 2,66M tokens av skandinaviske middelaldertekster, hovedsakelig tekster fra 1200–1350 produsert av Menota-nettverket (https://menota.org/forside.xhtml). 1M tokens er manuelt annotert med ordklasse og brukes til å finjustere modellen for ordklassetagging. Lemmatisering foregår ved å slå opp lemmaer med samme ordform og ordklasse i treningsmaterialet.

Rundt 11% av materialet har både diplomatariske former (bokstavtro manuskriptgjengivelser) og normaliserte former. Testing av MLM på usette data med normaliserte former viser betydelig bedre nøyaktighet (“accuracy”) enn med diplomatariske (88,7% vs. 76,6%), men siden en såpass liten del av materialet er normalisert, har det likevel størst hensikt å fokusere på diplomatariske former.

En utfordring er at det meste av treningsmaterialet består av eldre språk (klassisk norrønt), der ordformene ofte skiller seg fra formene i det senere mellomnorske materialet som skal annoteres. I tillegg er det betydelig inkonsekvens i den manuelle taggingen av treningsmaterialet. I tester av automatisk ordklassetagging på senere materiale øker nøyaktigheten fra 78,3% til 88,2% når vi anvender regler som gjør taggingen i treningsmaterialet mer konsekvent. Inkludering av testmaterialet i MLM-treningen (men ikke ordklassetreningen) for å gjøre modellen kjent med ordformene gir en liten ytterligere økning til 88,4%.

Referanser:

Devlin, Jacob, Ming-Wei Chang, Kenton Lee og Toutanova, Kristina. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

Kummervold, Per E., Javier De la Rosa, Freddy Wetjen og Svein Arne Brygfjeld. 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. I

Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), side 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Kutuzov, Andrey, Jeremy Barnes, Erik Velldal, Lilja Øvrelid og Stephan Oepen. 2021. Large-Scale Contextualised Language Modelling for Norwegian. I Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), side 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sverige.

Snæbjarnarson, Vésteinn, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson og Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. I Proceedings of the Thirteenth Language Resources and Evaluation Conference, side 4356–4366, Marseille, France. European Language Resources Association.

Uthenting av klinisk informasjon fra medisinske journaler

Tobias Jovall Wessel og Samia Touileb

Sammendrag

Klinisk dokumentasjon i Norge består av store mengder fritekst i pasientjournaler. Disse journalene inneholder mye klinisk relevant informasjon, men det ustrukturerte formatet gjør det utfordrende å systematisk gjenbruke informasjonen til blant annet analyse, informasjonsutveksling og beslutningsstøtte. Samtidig arbeider Helsedirektoratet med å ta i bruk den medisinske terminologien SNOMED CT for å etablere en felles klinisk terminologi i helsesektoren og dermed styrke informasjonsutvekslingen på tvers av aktører og systemer (Helsedirektoratet, 2026a,b).

I dette arbeidet, i samarbeid med Helsedirektoratet og Fagspråksenteret, undersøker vi hvordan store norske språkmodeller kan (videre) trenes eller finjusteres til automatisk uthenting av SNOMED CT terminologi og relevant kontekst fra norske pasientjournaler. Arbeidet tar for seg et fortsatt relativt begrenset forskningsfelt innen norsk språkteknologi. En nyere gjennomgang av klinisk språkteknologi for skandinaviske språk fant at bare 21 av 113 inkluderte studier fokuserte på norsk klinisk tekst (Woldaregay et al., 2025). Samtidig har domenetilpasning av norske språkmodeller allerede vist potensiale for automatisert klinisk koding av pasientjournaler. For eksempel, NorDeClin-BERT ble domenetilpasset for prediksjon av ICD-10 koder og ga bedre resultater enn de generelle norske språkmodellene den ble sammenlignet med (Ngo et al., 2025). I vårt arbeid, bygger vi videre på dette ved å utforske automatisk uthenting av den norske versjonen av SNOMED CT ved bruk av nyere og kraftigere språkmodeller.

Identifisering av begrepene er imidlertid bare en del av problemstillingen, de identifiserte begrepene må også tolkes ut fra konteksten de forekommer i. "Påvist pneumoni", "ingen tegn til pneumoni" og "mistanke om pneumoni" uttrykker eksempelvis tre ulike kliniske påstander. Tidligere forskning på klinisk tekst viser betydningen av kontekstuell informasjon som negasjon, temporalitet og hvem et utsagn gjelder (Harkema et al., 2009; Uzuner et al., 2011). Vi ønsker derfor å utforske hvordan resonneringsevnen til generative modeller kan utnyttes til både å trekke ut SNOMED CT-relevante kliniske begreper og identifisere relevant kontekstuell informasjon knyttet til begrepene.

Et annotert datasett med norske pasientjournaler er for øyeblikket under konstruksjon. Datasettet skal brukes til videre trening av norske språkmodeller, med mål om å trekke ut SNOMED CT-relevante kliniske begreper og tilhørende kontekst. Vi kommer til å sammenligne ulike norske modellfamilier og modellstørrelser. For å redusere risikoen for hallusinerte terminologikoder vil prosjektet undersøke en systemarkitektur der de identifiserte begrepene kobles mot en database med den offisielle norske SNOMED CT-terminologien, fremfor at språkmodellen selv genererer kodene.

Målet er å undersøke hvordan norske språkmodeller og verdens mest omfattende kliniske terminologi kan kombineres for å redusere manuelt arbeid for helsepersonell og bedre legge til rette for datadrevet bruk av klinisk informasjon i den norske helsesektoren.

Referanser

Harkema, H., Dowling, J. N., Thornblade, T., and Chapman, W. W. (2009). ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Helsedirektoratet (2026a). Felles språk.

Helsedirektoratet (2026b). Helsefaglig terminologi.

Ngo, P. D., Tejedor Hernández, M. A., Chomutare, T., Budrionis, A., Olsen Svenning, T., Torsvik, T., Lamproudis, A., and Dalianis, H. (2025). Domain-specific pretraining of NorDeClin-Bidirectional

Encoder Representations From Transformers for International Statistical Classification of Diseases, Tenth Revision, code prediction in norwegian clinical texts: Model development and evaluation study. *JMIR AI*, 4:e66153.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Woldaregay, A. Z., Lund, J. A., Ngo, P. D., Tayefi, M., Burman, J., Hansen, S., Sillesen, M. H., Dalianis, H., Jenssen, R., Ole, L. R., et al. (2025). Natural language processing for electronic health records in scandinavian languages: Norwegian, swedish, and danish. *arXiv preprint arXiv:2503.18539*.