amedia

User Needs at Amedia

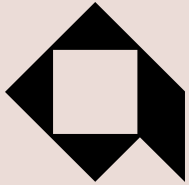
Training a User Needs 2.0 predictive model for Norwegian

Emiliano Guevara
Data Scientist, Computational Linguist
2024/09/19 - WAN-IFRA Data Science Meet-up

About me...

- Born and raised in Argentina
- Lived and studied in Italy, University of Bologna
- In Norway since 2010

- Background in theoretical linguistics and computational linguistics
- Experimental work, resource building, news corpora and web corpora, distributional semantics
- Academic career in Italy 2005-2010 and Norway 2010-2013
- Private sector in Norway since 2014, scientific programming, NLP, machine learning
- Senior Data Scientist at Amedia since 2018



User Needs at Amedia

Amedia

User Needs

Training a Norwegian model

Results and analysis

Status and future work

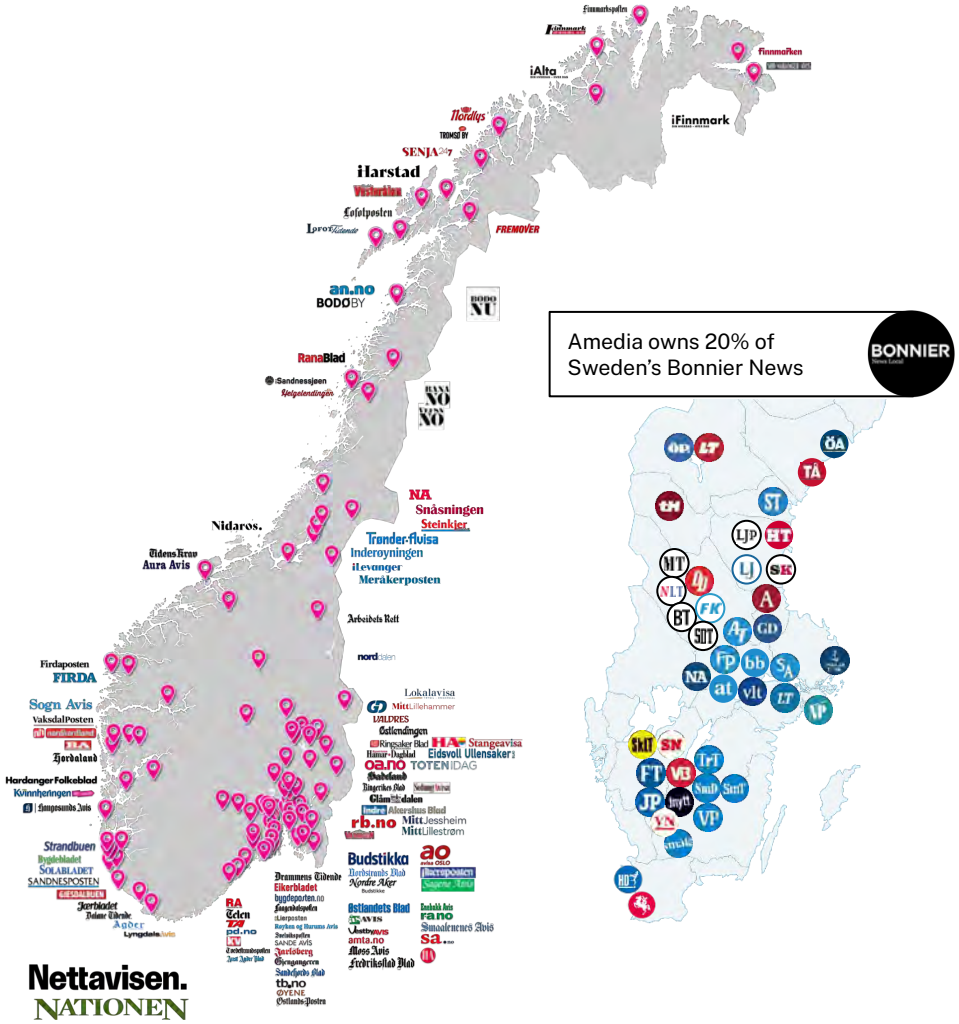
Amedia

About Amedia

Top publisher of independent media in Norway

- Over 100 local, regional and national publications in Norway
- 130 locations
- 2.4M daily readers
- 0.73M subscribers
- 2500 employees
- 1100 journalists

- Part-owner of 55 local/regional papers in Sweden via Bonnier News Local




Agder
Amtstidende
Akersposten
Arbeidets Rett
Aura Avis
Aust Agder Blad
GAULA
an
ao avisa OSLO
BA
BODØBY
BODO NU
Bondebladet
Budstikka
Bygdebladet
BygdePosten
Dalane Tidende
Drammens Tidende
Enebakk Avis
Fanaposten
Finnmark Dagblad
Finnmarken
Finnmarksposten
FIRDA
Firdaposten
Fredriksstad Blad
FREMOVER
Gjengangeren
GJESDALBUEN
Glåmdalen
Hadeland
Hammerfestingen
HA
Hardanger Folkeblad
Hordaland
Harstad
Indre Akershus Blad
Sandnessjøen
Jærbladet
Jarlsberg
KARMØYNYTT
KIRKENES BY
KV
Kvinnheringen
Laagendalsposten
Haugesunds Avis
iFinnmark
KRS
Lierposten
Lofotposten
Lyngdals Avis
Min energi
MITT HAMMERFEST
MittJessheim
MittLillestrøm
Moss Avis
NATIONEN
Nettavisen.
Nidaros.
nordhordland
Nordlys
Nordre Aker
Nordstrands Blad
Norsk Landbruk
Oppland
RakkestadAvis
RanaBlad
RANA NO
PD
RADIO2
Ringerikes Blad
RB
RA
RHA
Sagene Avis
SANDE AVIS
Sandefjords Blad
SANDNESPOSTEN
sa.
SENJA24/7
Smaalenenes Avis
Sogn Avis
SOLABLADET
SolungAvisa
Strandbuen
Svelvikposten
SØR-VARANGER AVIS
SYDVESTEN
TA
Telen
Tidens Krav
Tønsbergs Blad
TOTEN IDAG
Traktor
Tvedestrandsposten
VARINGEN
VestbyAVIS
Østlandets Blad
Østlands-Posten
Østlendingen
ØYENE
ÅSANE TIDENDE
AS AVIS

Editorial independence and liability in Norway

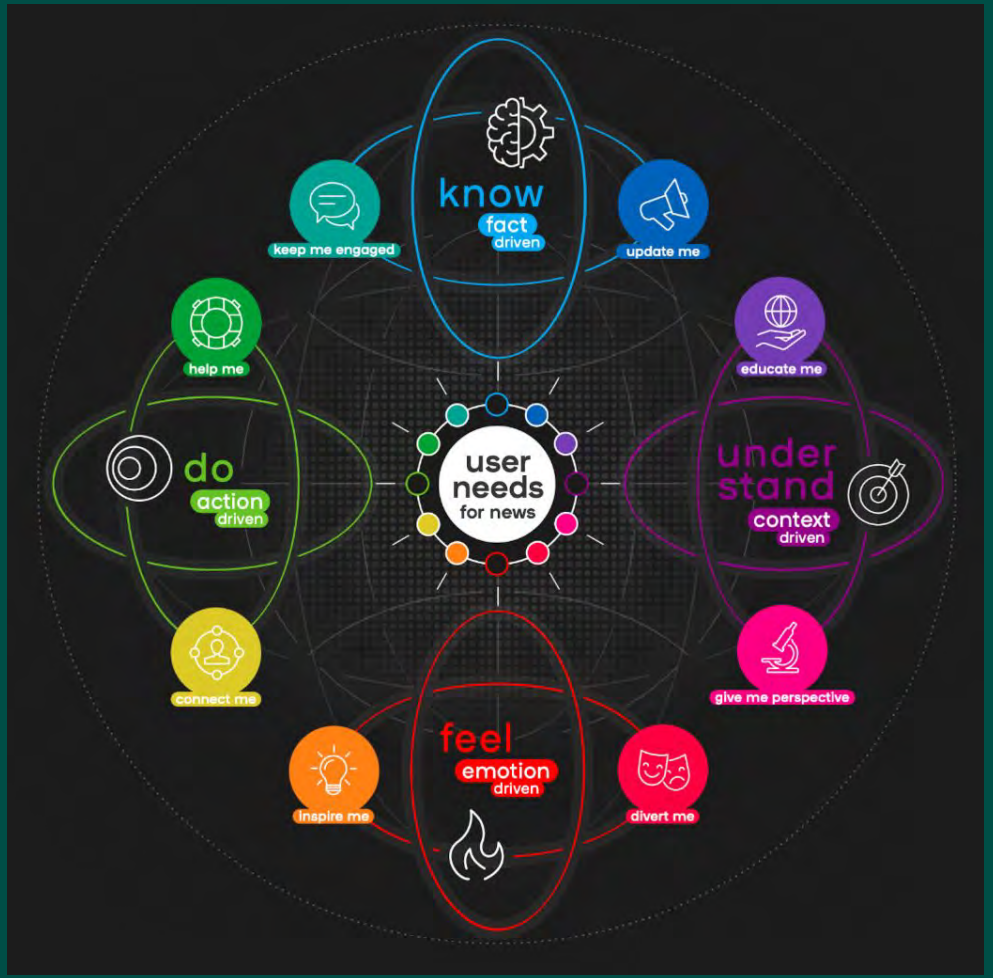
Media Liability Act:

The publisher, owner or other company management cannot instruct or overrule the editor on editorial issues, nor can they demand to see print, text or pictures, or demand to hear or see programme material before it is made available to the public.

https://lovdata.no/dokument/NL/lov/2020-05-29-59

All our centralized AI, NLP, analysis work must respect editorial independence

User Needs 2.0



What are User Needs?

- Original idea: 2016/2017, Dmitry Shishkin, BBC World Service Language

- Readers consume news for many reasons, not just staying informed...

  “audiences expect much more from newsrooms - they want to be updated, yes, but they also want to be educated, kept on trend, inspired, amused and given perspective, all on the key news topics of the day”

https://www.linkedin.com/pulse/user-needs-content-publishing-slide-started-all-five-years-shishkin/

What are User Needs?

- Basic question:
  What do people want from news???

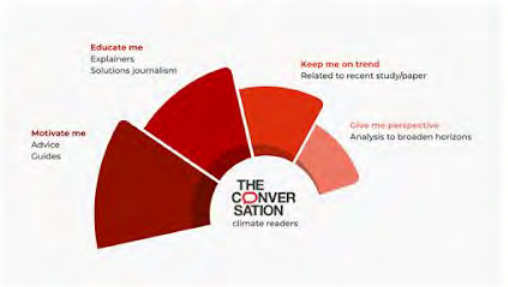
- Each medium/audience will answer differently: there are many possible models, all of them valid
  - Some alternatives:
    https://tinyurl.com/49s4zxt4
    https://tinyurl.com/49vvtwwm (images)

What are User Needs?

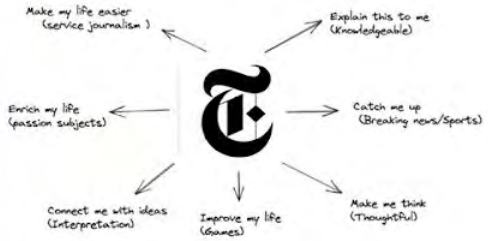
VOGUE

BuzzFeed

THE CONVERSATION



The New York Times



The Atlantic

Reader & listener needs

- Give me deeper clarity & context
- Help me discover new ideas
- Challenge my assumptions
- Let me take a meaningful break
- Introduce me to writers at the top of their craft

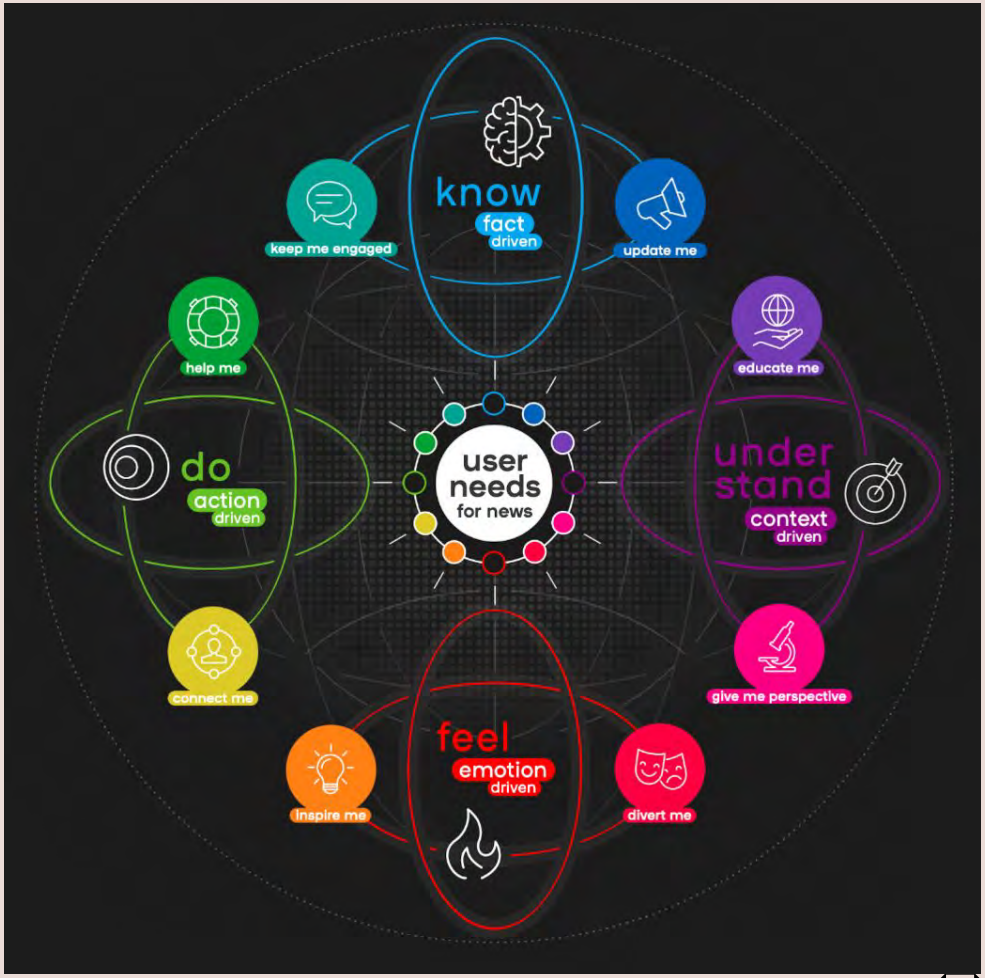
Vox

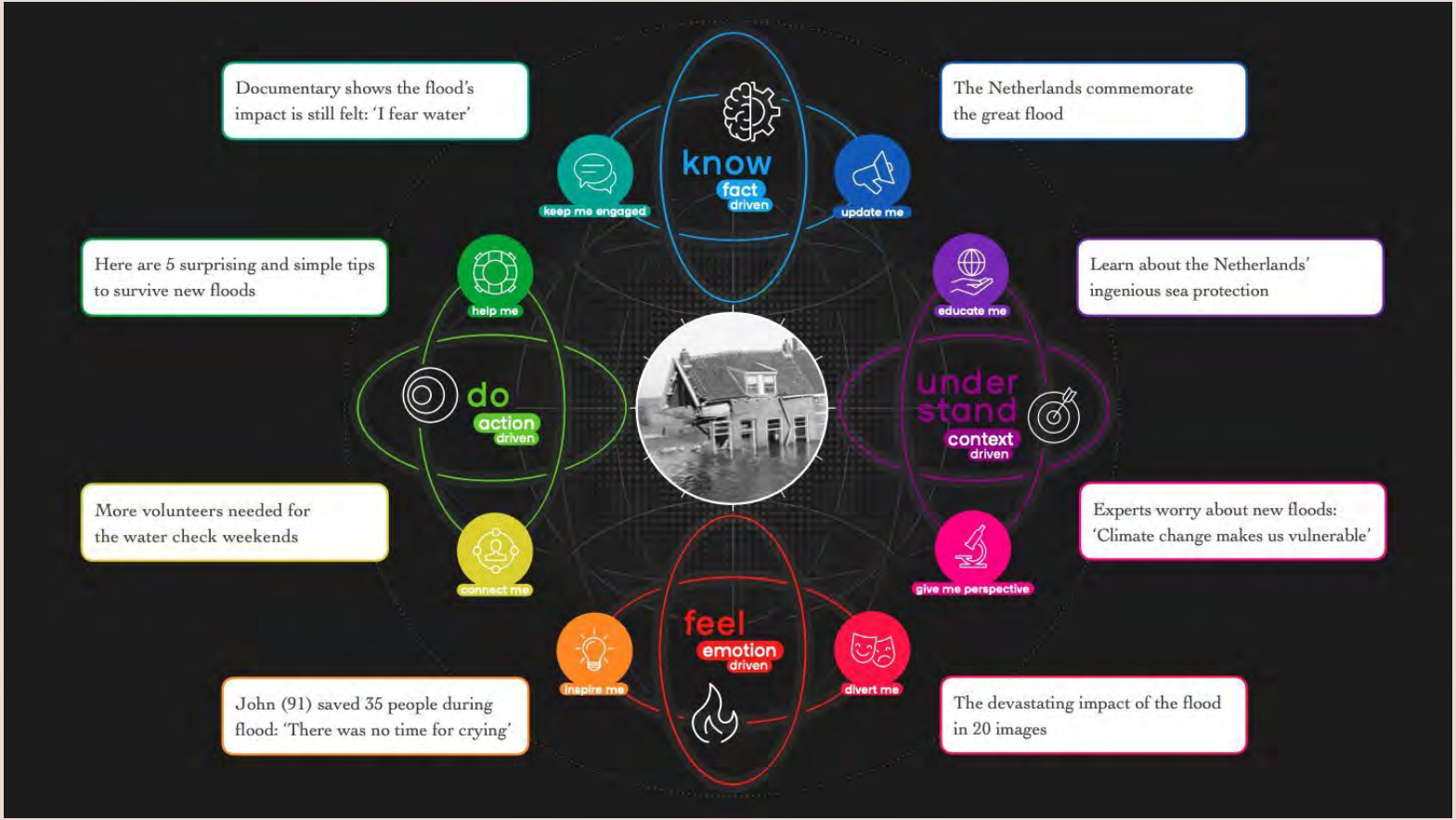
User Needs 2.0

- 2023, latest iteration by Shiskin/Smartocto
  - 4 macro categories
  - 8 micro categories

- CC BY-SA 4.0
  https://creativecommons.org/licenses/by-sa/4.0/
  - Free to share
  - Free to adapt

- User-centric
- Not determined by content
- Not topics / document classification
- Different modes of communication





The flood: headlines meeting the 8 user needs

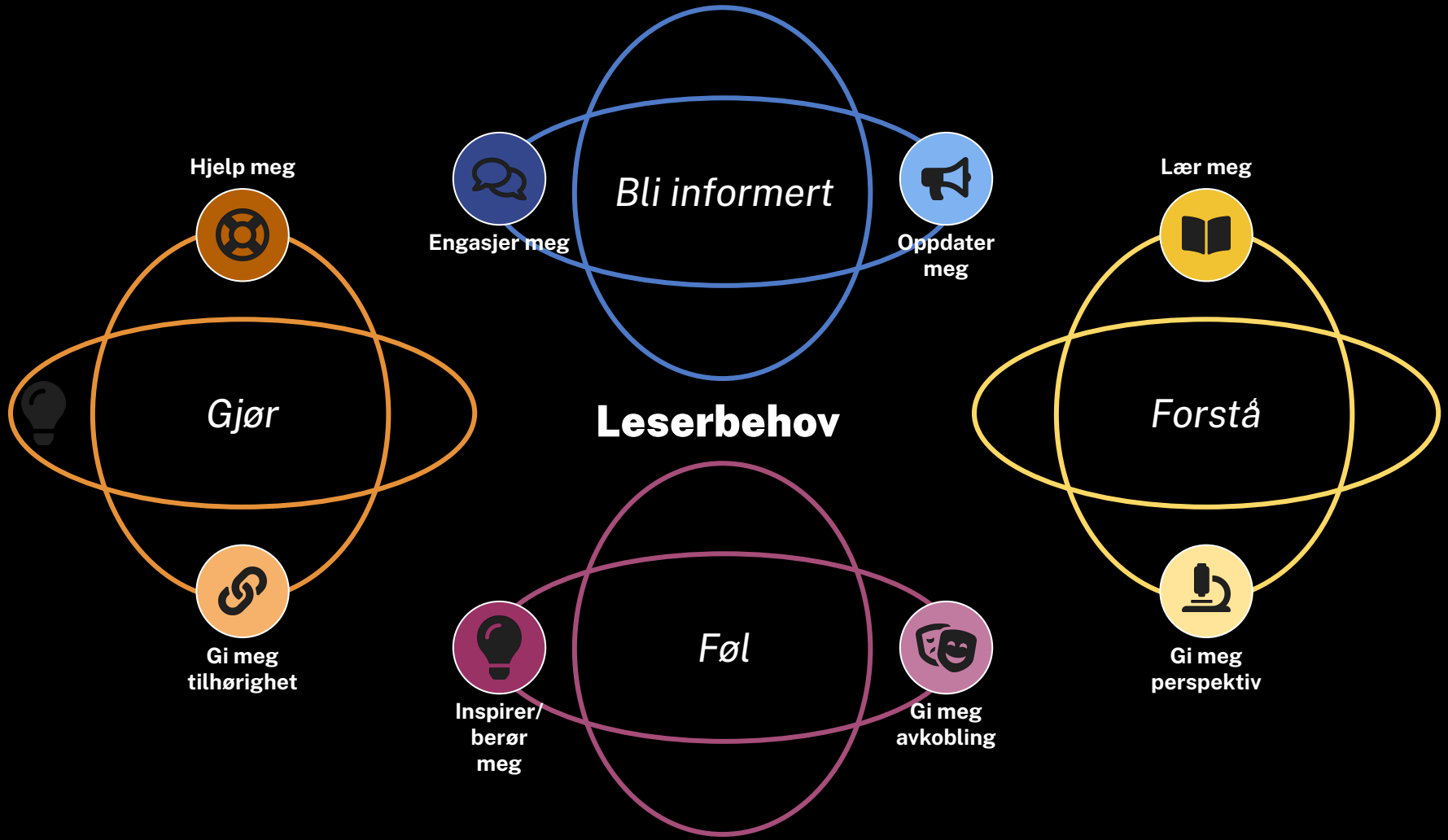
Starting User Needs 2.0 at Amedia

- 2023/2024
- Very much work in progress

- Adapting the Smartocto model
  - 7 or 8 categories???
  - color scheme???

- Translating the labels
- Writing guidelines for annotation and interpretation
- Disseminating the model amongst 100 news desks

- Work by colleagues at Amedia's Content Development department


Leserbehov
Bli informert
Engasjer meg
Oppdater meg
Gjør
Hjelp meg
Gi meg tilhørighet
Forstå
Lær meg
Gi meg perspektiv
Føl
Inspirer/ berør meg
Gi meg avkobling

Bli informert

Engasjer meg

Saker som viser hvordan en aktuell hendelse fører til (positive eller negative) **reaksjoner** og **engasjement** hos mennesker som berøres.

Dekker brukerens behov for å delta i samtaler og diskusjoner om hva som skjer i (lokal)samfunnet.

Oppdater meg

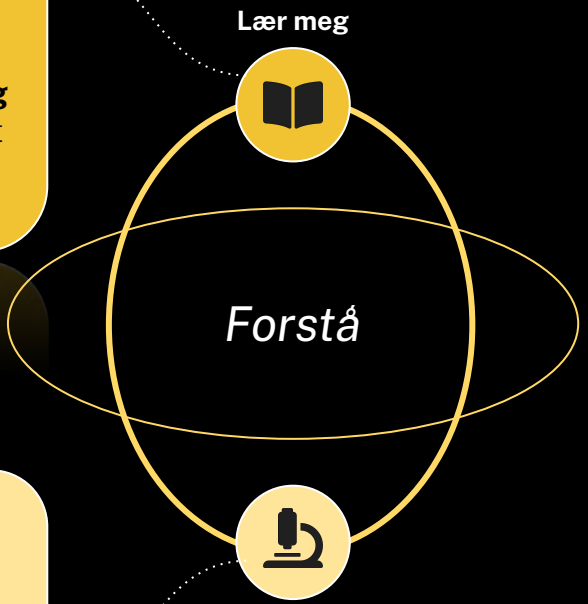
'Oppdater meg'-artiklene er de klassiske "**hvem, når, hvor, hva**"-sakene:

Ofte tradisjonelle **breaking-pregede** nyhetssaker hvor hovedformålet er å informere leseren om aktuelle hendelser. Ofte ulykkes-/krimstoff, men finnes i alle temakategorier.

Setter enkeltsaker inn i **større sammenhenger**. De svarer på behovet til dem som ikke bare ønsker å lese om "hvem, når, hvor, hva", men også om "**hvordan**" og "**hvorfor**".

Saker som dette har typisk **lang lesetid** og **lang levetid**. Trenger ikke nødvendigvis være knyttet til noe aktuelt, og kan tvert i mot omhandle en historisk hendelse.

Saker med en **analytisk** tilnærming til aktuelle problemstillinger, for eksempel i form av en kommentar eller intervju med en ekspertkilde. Her får ikke leseren svar på konkrete spørsmål ("Lær meg"), men derimot hjelp til å danne egne meninger og **se en kompleks sak fra flere sider**.

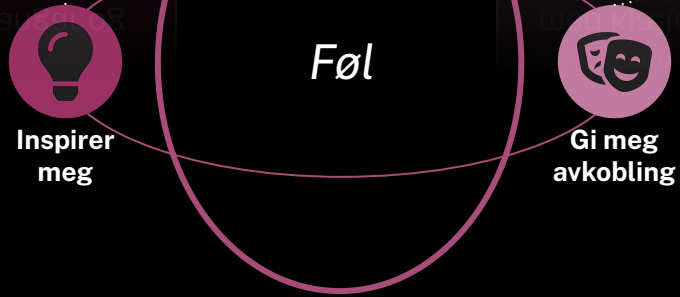


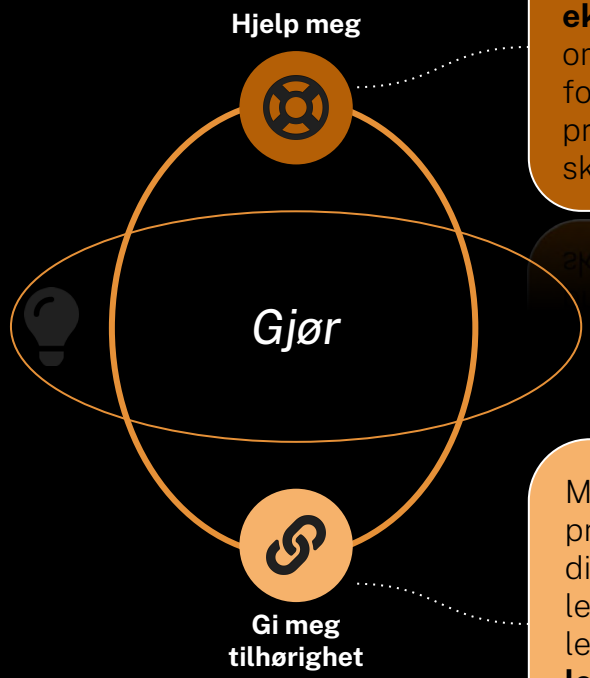
Typisk **menneskehistorier** om noen som har opplevd noe uvanlig, interessant eller overraskende. Ofte i intervjuform.

Leserbehovet handler om å **identifisere** seg med, og bli **inspirert** av, menneskene det angår og historien de forteller.

Typisk **lettbeinte** artikler om **kuriøse** forhold, men kan også ta utgangspunkt i tunge og alvorlige tema så lenge presentasjonen og vinklingen har et lettere tilsnitt.

Eksempel: Artikkel om barn som ble gjenforent med kjæledyret etter en brann.





Service-og forbrukerjournalistikk. “Du”- og “slik”-saker. Typisk saker hvor **ekspertkilder** gir **råd** og **anbefalinger** om hvordan de helt praktisk skal forholde seg til aktuelle problemstillinger (“Slik finner du de skjulte jobbene”, “Advarer mot svindel”).

Mens “hjelp meg”-sakene gjerne har et praktisk og personlig tilsnitt, fyller disse sakene et mer **sosialt behov** hos leserne. Tilhørighet-sakene hjelper leserne med å være en **aktiv deltaker i lokalsamfunnet**, for eksempel i form av restaurantanmeldelser eller saker om hva som skjer i helga.

NLP at Amedia

Previous experience

- Several applications of NLP are fully developed and integrated into Amedia's production systems
- Language technology pre-ChatGPT (pre-LLM)
- 2018-2023

POLITIKK ARBEIDSLIV RAYMOND JOHANSEN NORSK FOLKEHJELP ETTERLØNN

Byrådslederens kontor vurderer å avkorte etterlønnen til Raymond Johansen

Raymond Johansen vurderer å begynne i den nye jobben tidligere. Dette vil kutte ned på etterlønnen han mottar fra kommunen. Foto: Heiko Junge / NTB

Av Åsmund Swensen Høeg Publisert: 21.11.23 15:54 Del

Raymond Johansen fikk innvilget nesten 400.000 kroner over tre måneder i etterlønn fra jobben som byrådsleder. Nå vurderer han å begynne tidligere i Norsk Folkehjelp.

- Automatic document classification
- Automatic tagging

- Named entity recognition
  - Person
  - Organization
  - Location

- Automatic document classification
- Automatic tagging

- Named entity recognition
  - Person
  - Organization
  - Location

Other NLP systems in production

- Document classification
  - Cat20
  - politics, economy, religion, sports, etc.
- Keyword extraction
- Automatically transcribing journalistic notes (Whisper - OpenAI)
- Modelling reader interests
  - interested in buying a house
  - interested in buying a car
- Modelling user segments based on article keywords (ad-market)

User Needs in Norwegian: Training

Main challenge: no data available

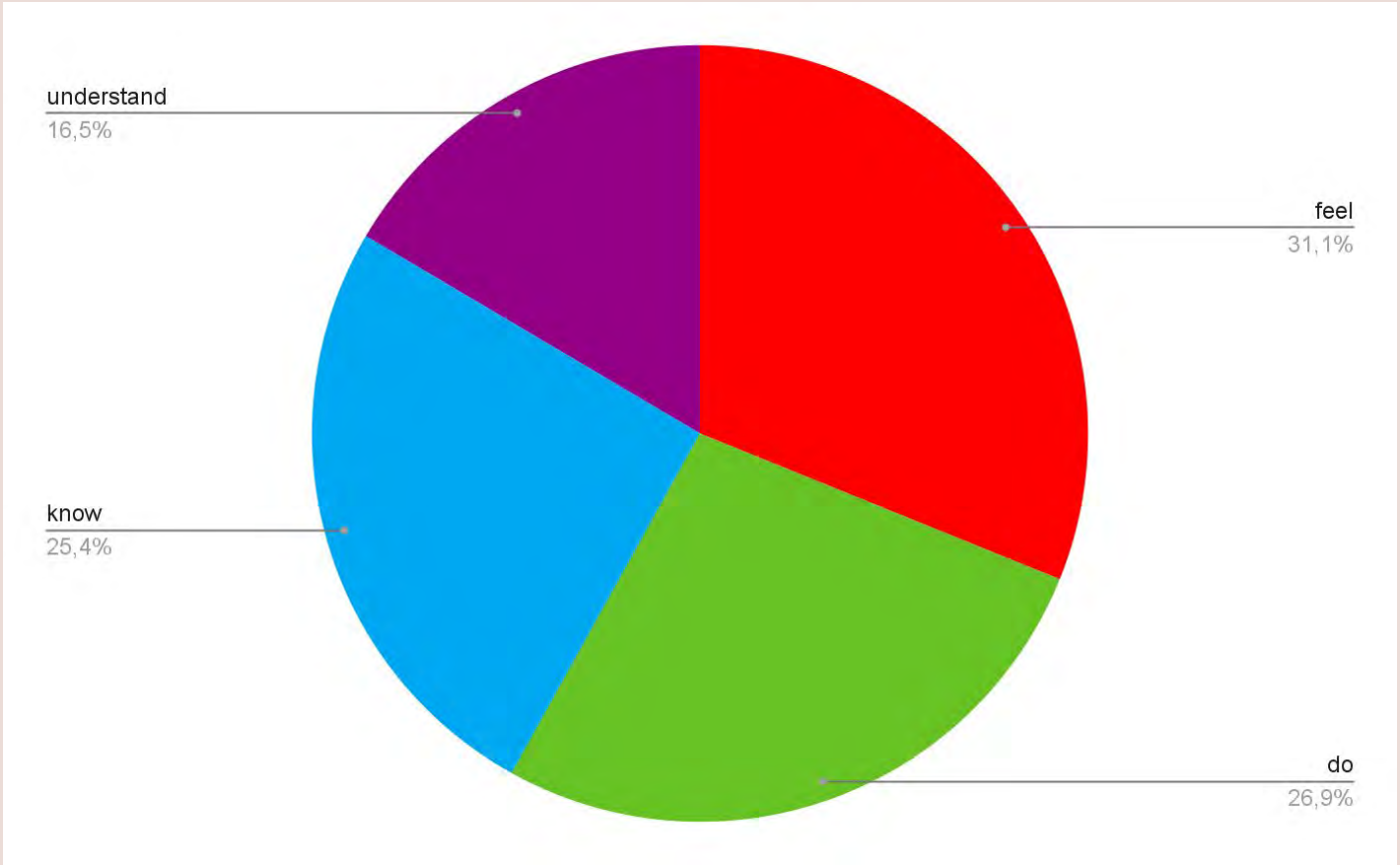
- There are no freely available datasets (neither English, nor Norwegian)
- A lot of interest in training AI-models able to tag/classify text automatically
  - check Smartocto demo/playground
- We organized an annotation campaign, target about 10k articles
  - lots of work...
  - recruiting and organizing annotators
  - writing guidelines
  - extracting representative and balanced candidate articles
- Results
  - finished after about 2 months, 2 annotators + 2 analysts
  - over 8k examples after cleaning and post-processing

Dataset

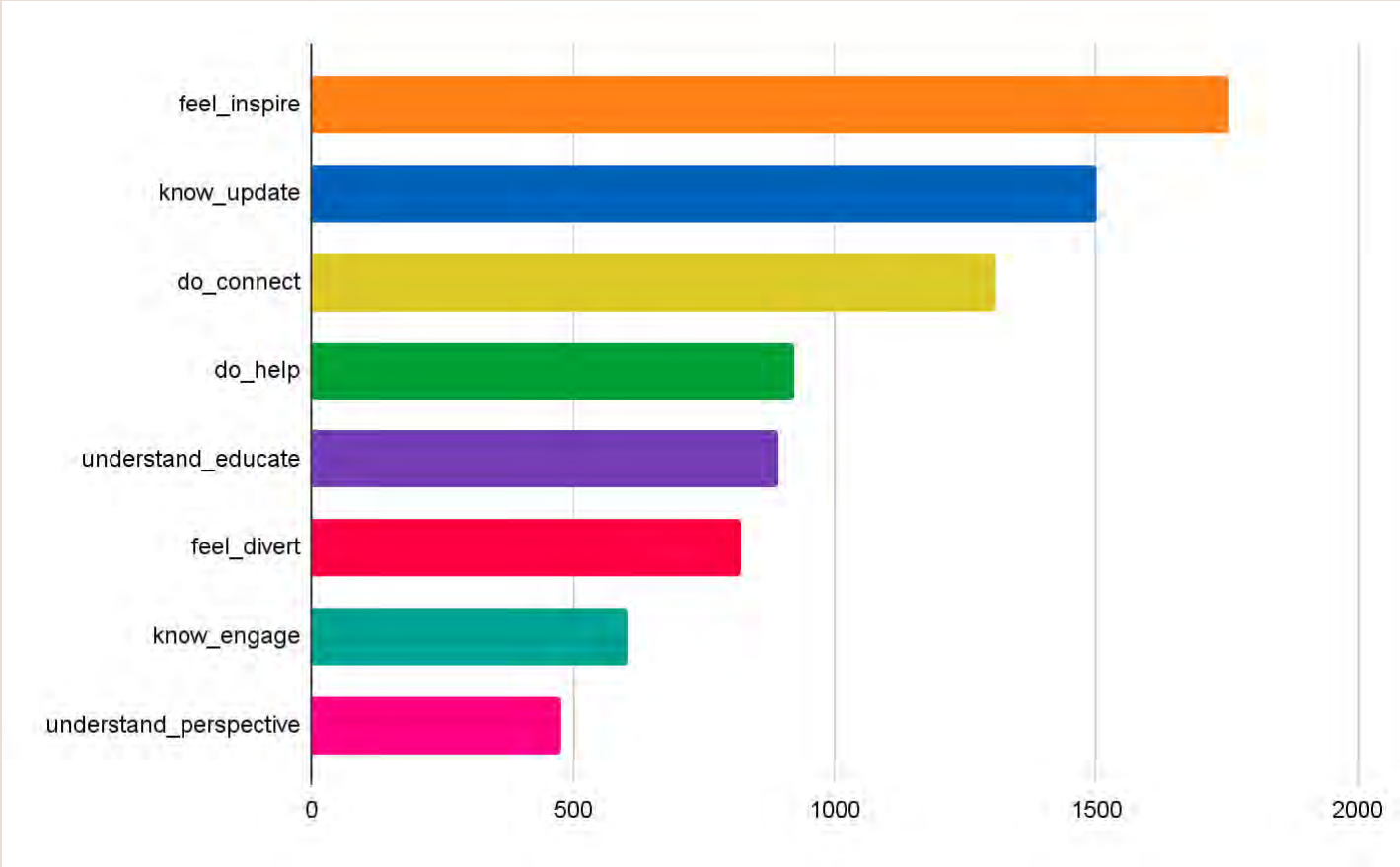
training:	6334
validation:	704
test:	1242
total:	8280

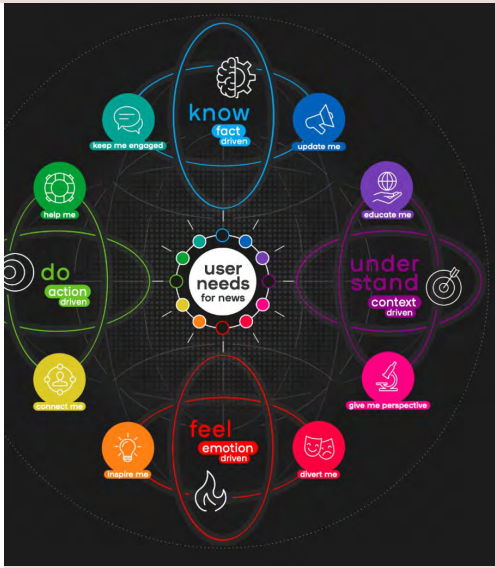

```
{
  'id': '5-34-1899397',
  'text': 'Sto i kø for bilde med ikonisk kjøretøy . Det var tidvis så folksomt rundt den velkjente semitraileren at folk måtte smøre seg med tålmodighet . – Kan du ta et bilde av meg med den , spør noen . Utenfor Jekta sto Coca-Colas juletrailer til manges store glede . – Vi fant ut at dette var en god anledning til å dra hit så han her får se traileren for første gang . Han liker store kjøretøy og sånt , forteller Eivind Thoresen med Alfred på armen . Noen meter unna står et firkløver og tar bilder . – Det er koselig , forteller Nathalie Gaare Bjørnstad . Hun erkjenner at det var hennes idé å stikke innom juletraileren . – Det var jeg som så det . Vi bor i Nordreisa og er på bytur , så da slo vi to fluer i en smekk , forteller hun . – Dette er starten på jula . Vi kjører rundt i hele landet frem til 20. desember og vi starter i Tromsø , forteller eventmanager hos Coca-Cola , Harald Berg . Han forteller om stor interesse den tiden de har stått utenfor Jekta torsdag ettermiddag . – Det er veldig masse folk innom som prøver VR-briller , sender julekort , tar seg en kald drikke og hilser på nissen , sier Berg . Det blir imidlertid kun med den ene dagen i Tromsø . – Torsdag kveld kjører vi videre til Bardufoss og skal stå der på fredag , sier Berg .',
  'labels': 0
}
```


Dataset: the 4 main categories



Dataset: the 8 user needs





Technology stack

- Google Cloud
- Vertex AI

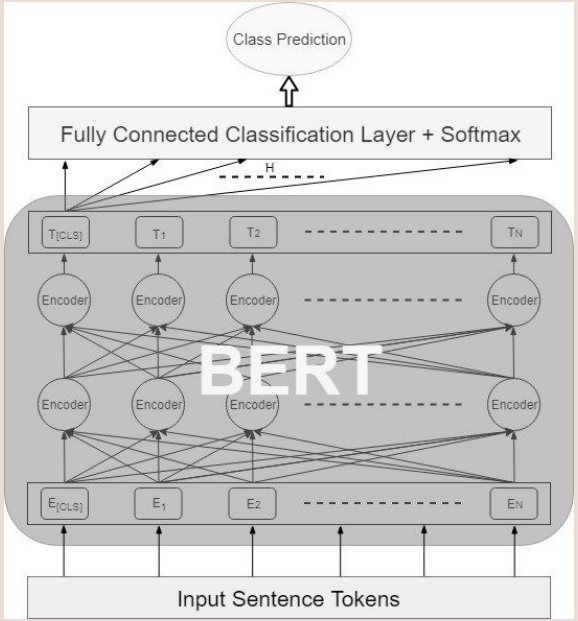
- PyTorch
- Hugging Face - Transformers - Datasets
- NorBERT 3 small - UiO
  https://huggingface.co/ltg/norbert3-small

Motivation / Methodology

- Our legacy models use
  - Tensorflow
  - Word embeddings and document embeddings trained in-house
  - Word2Vec / FastText
  - Dedicated neural networks
    (Word embeddings + 2 CNN/MaxPooling + Decision layers)
  - Good results, but requires large datasets (100k++ examples)

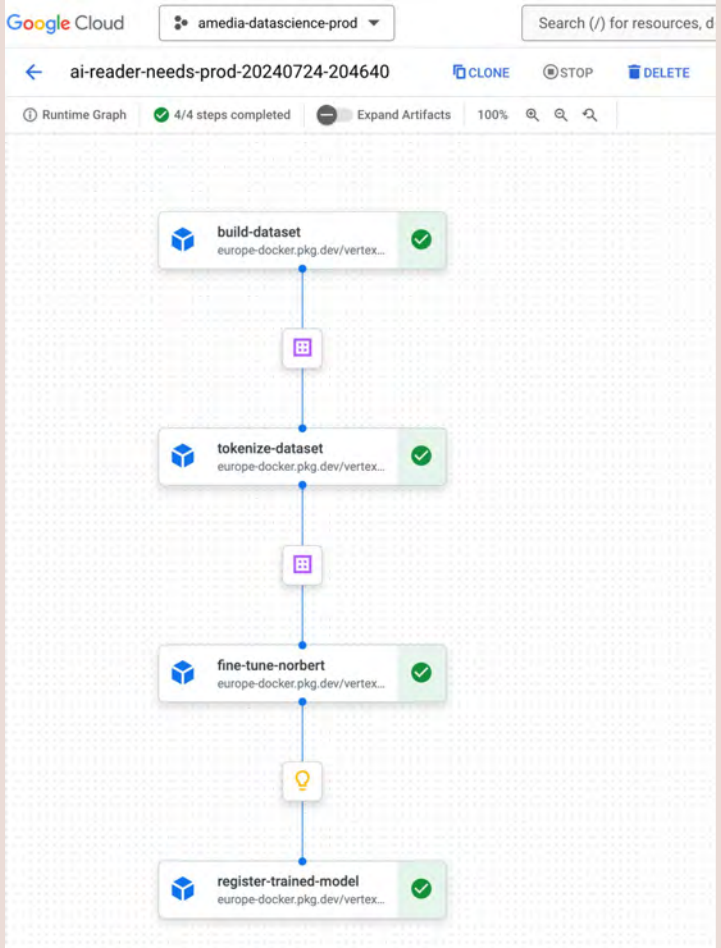
- We needed to update the methodology...

- Fine-tuning and deploying a Norwegian monolingual LLM: NorBERT 3 Small
  - Produces very good results with small datasets



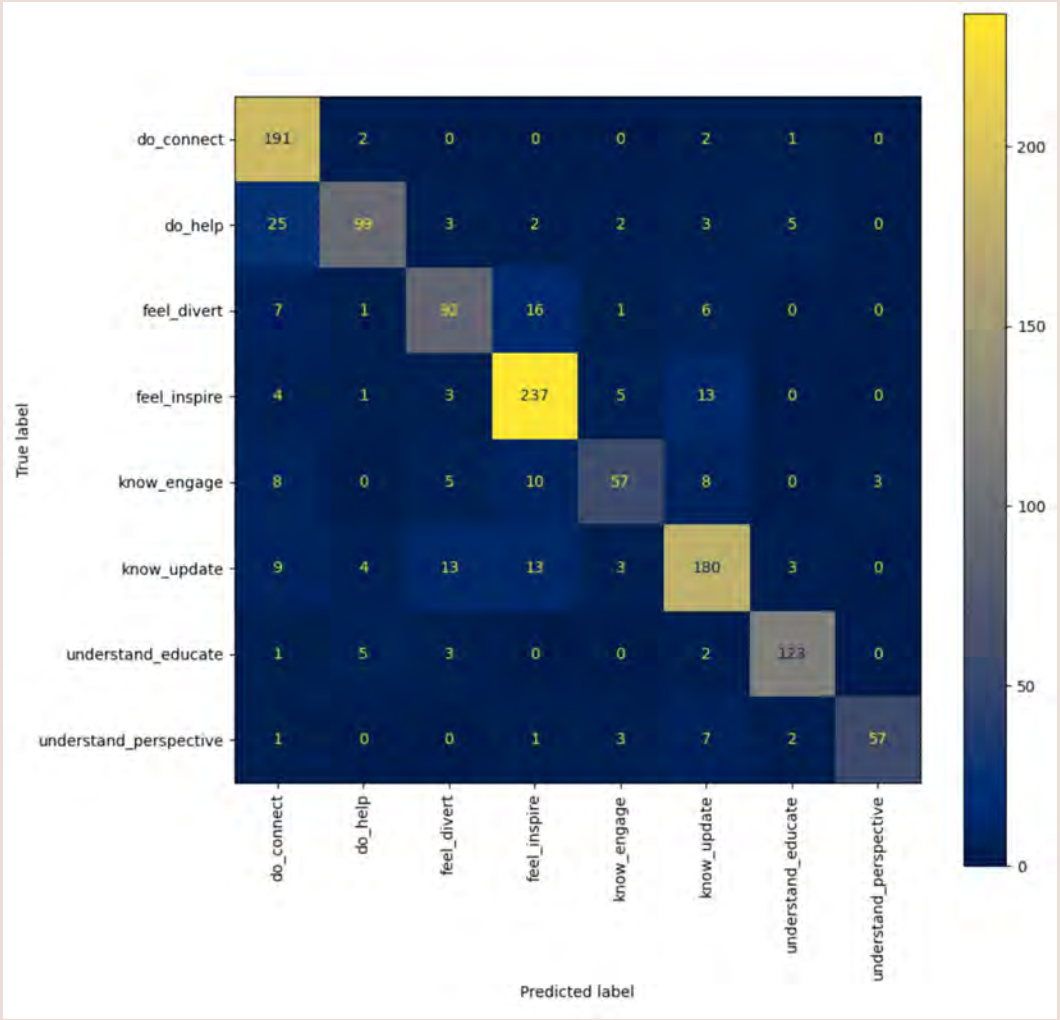
Training the prediction model

- Vertex AI - Kubeflow pipeline
- NorBERT 3 Small is fine-tuned with the task of predicting the User Needs label
- Training time: between 4 and 8 hours with 1 GPU
- Output model registered as a Vertex AI prediction object (endpoint ready to serve predictions)



Test results (unseen data, deployed model)

	precision	recall	f1-score	support
do_connect	0.78	0.97	0.86	196
do_help	0.88	0.71	0.79	139
feel_divert	0.77	0.75	0.76	123
feel_inspire	0.85	0.90	**0.87**	263
know_engage	0.80	0.63	0.70	91
know_update	0.81	0.80	0.81	225
understand_educate	0.92	0.92	**0.92**	134
understand_perspective	0.95	0.80	**0.87**	71
accuracy			0.83	1242
macro avg	0.85	0.81	0.82	1242
weighted avg	0.84	0.83	0.83	1242


do_connect
do_help
feel_divert
feel_inspire
know_engage
know_update
understand_educate
understand_perspective
True label
Predicted label
191 2 0 0 0 2 1 0
25 99 3 2 2 3 5 0
7 1 90 16 1 6 0 0
4 1 3 237 5 13 0 0
8 0 5 10 57 8 0 3
9 4 13 13 3 180 3 0
1 5 3 0 0 2 123 0
1 0 0 1 3 7 2 57
200
150
100
50
0

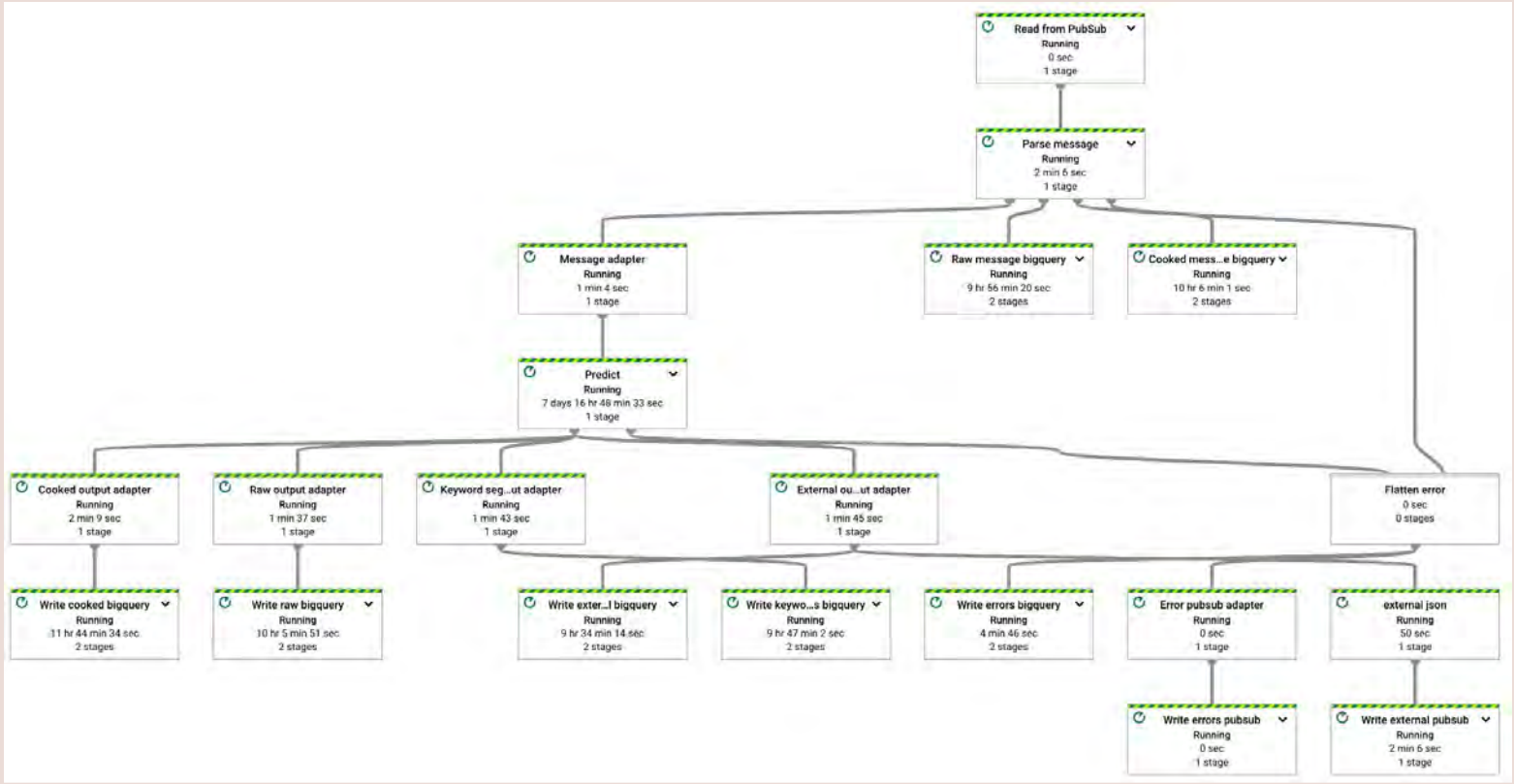
Inference: near real time / streaming

- Model trained and deployed in production
- Integrated into our main NLP streaming pipeline
- Vertex AI, Dataflow (Apache Beam), PubSub

- Each time our journalists publish/save an article, we produce a new prediction based on the updated text
  - Asynchronous streaming job
  - After a few minutes, the prediction is saved in our Content Store

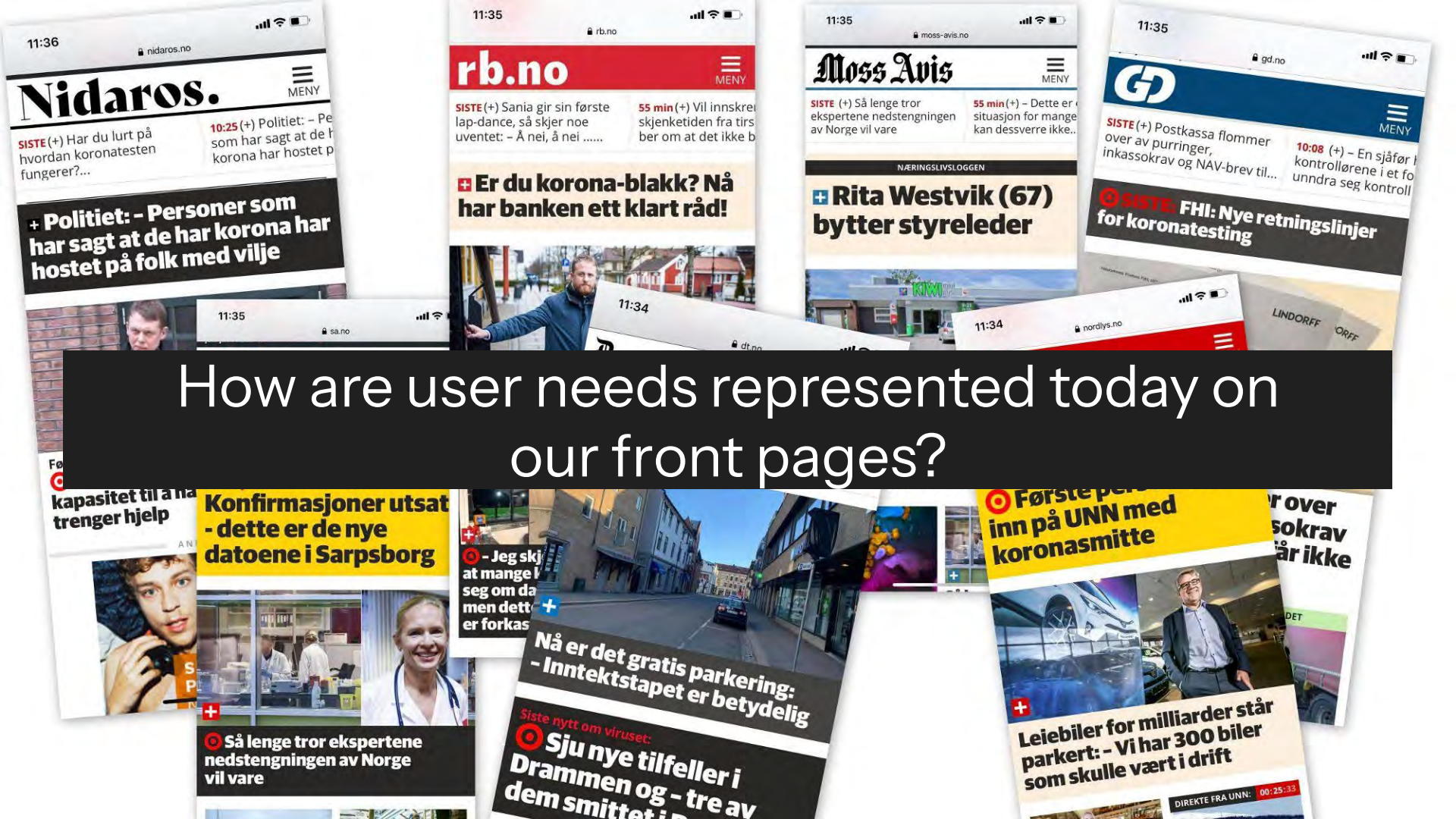
- Also, the endpoint is available for **on-demand / real time inference** (a few seconds per article at the moment)

Inference: near real time / streaming



Results and Analysis

By colleagues from
Content Development



How are user needs represented today on our front pages?

und_educate

know_update

politimann etter vill biljakt

und_perspective

Kraftig økning:

Smittes du av dette, har legen et klart råd til deg

Ny arealplan:

Slik skal Haugesund bygges og bevares

know_update

Snart går alarmen: Mobilen din skal ule og vibrere

know_update

und_perspective

Festfotball før kollaps, tre ekstremt viktige poeng og Riises superuke

Daan har arrangert Hardangervidda-turer i 13 år:

I år ble firmaet hans plutselig anmeldt

know_update

Politiet skal teste ut kroppskamera

know_update

Gikk inn som eier i februar:

Nå kjøper selskapet opp lokalt firma

know_update

Nå har den nye takterrassen åpnet i sentrum: – Spektakulær utsikt

know_engage

Forberedt på støy og reaksjoner:

– Det er lov å være uenige, men det får da være grenser

To personer hentet av ambulanse etter brann i silo

know_update

know_update

Brann i el-sykkel i Haugesund

h-avis.no, 14. juni 09:30

know_update

Nå

Alvorlig trafikkulykke på I
Mann i 40-årene sendt til sykehus

know_update

Her blokkerer gjestene rømningsveien. Nå risikerer det populære utestedet bot

know_update

togtrafikken inn mot Oslo

know_update

Derfor vil mobilen din pipe og vibrere i dag

do_connect

Nå har den nye takterrassen åpnet i sentrum: – Spektakulær utsikt

Kraftig økning:

Smittes du av dette, har legen et klart råd til deg

know_engage

Takbrannen skaper trøbbel for Ruter-passasjerer: – Frustrerende

und_perspective

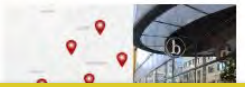
Sjekk planene: Her skal det bygges flere hundre nye eldreboliger

und_perspective

Et forbud vil være katastrofalt for boligprisene i Oslo

und_perspective

De rødgrønnes giftige boligcocktail



do_connect

17 bakerier på få år: – Bransjen kan takke seg se

know_update

Det kan bli tropenatt

ao.no, 14. juni 09:30

know_updat e

know_engag e

Magnor savner internett og TV: – Borte i 22 timer

MENINGER

und_perspe ctive

Blomsterbutikken blir døgnåpen: – En av de første i Norge

Foreslått utbygging på Gressbanen – et voldsomt inngrep!

know_updat e

Vil gjøre det attraktivt å være fastlege i Øvre Eiker: – Skal bli bedre enn Kongsberg og Drammen

know_engag e

– Lekeplassen har blitt så fin, men jeg tør ikke å la barna leke her

know_updat e

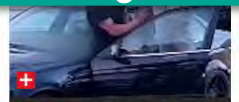
Disse boligene ble solgt i løpet av måneden som gikk

DEBATTINNLEGG

und_perspe ctive

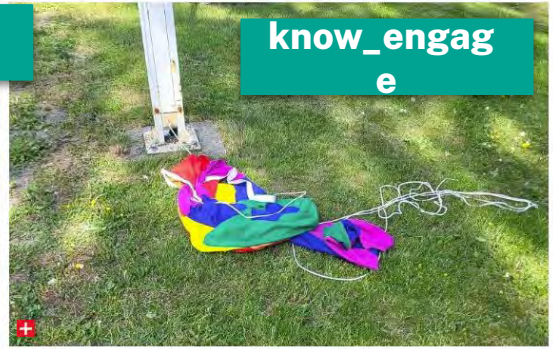
Bygg en svømmehall for fremtidens behov

know_engag e



Klapset til sjåfør som tok bilde: – Jeg driter i sånne som deg

know_engag e

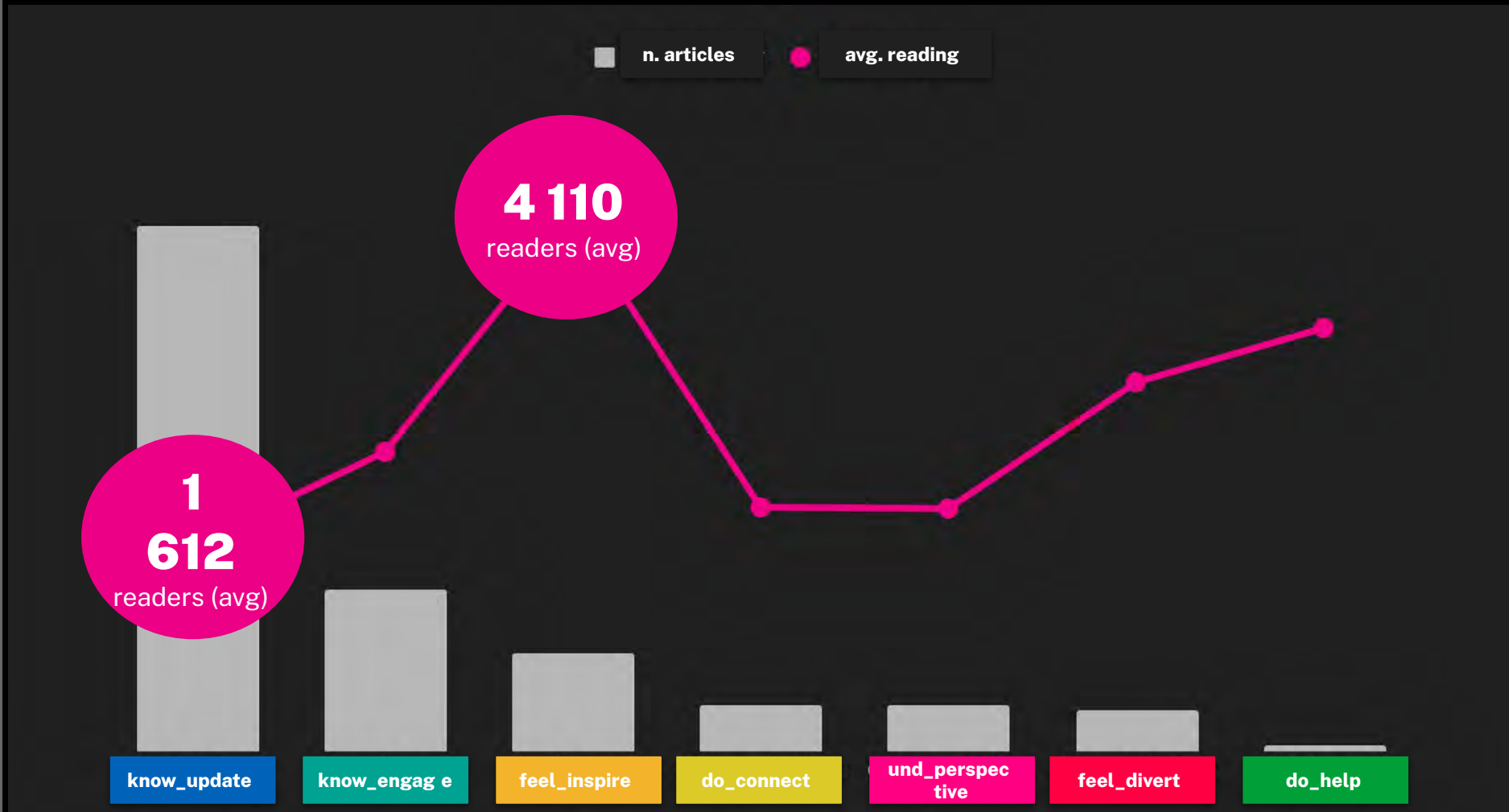


Nå har den nye takterrassen åpnet i sentrum: – Spektakulær utsikt

Skolens Pride-flagg revet i filler: – Flagget sk… igjen brenn…

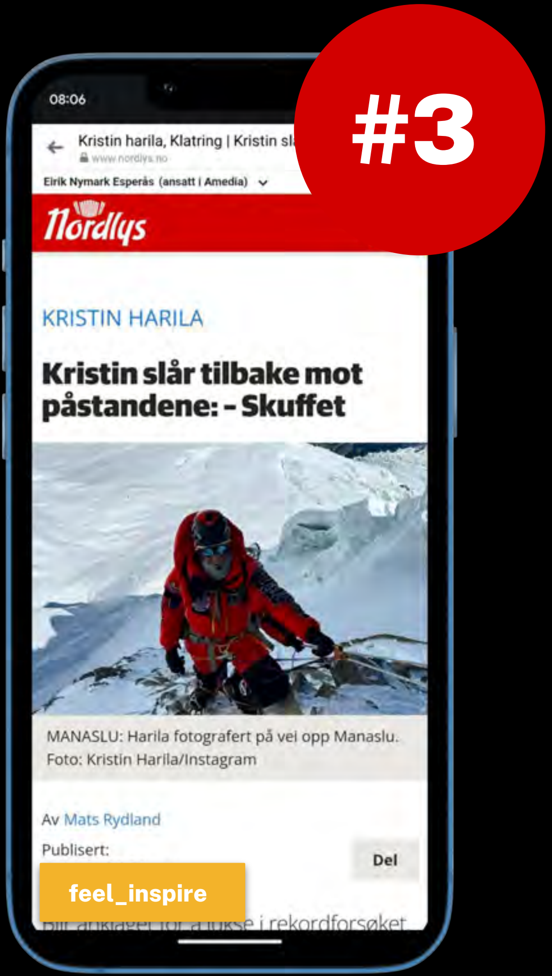
eikerbladet.no, 14. juni 12:30

Sports articles in Nordlys (Tromsø) according to User Needs


n. articles
avg. reading
4 110
readers (avg)
1
612
readers (avg)
know_update
know_engag e
feel_inspire
do_connect
und_perspec tive
feel_divert
do_help


#1
08:05
Sport, Fitness | Natalie fra N
www.nordlys.no
Eirik Nymark Esperås (ansatt i Amedia)
Nordlys
SPORT
Natalie fra Nordreisa: - Jeg er 34 år og trebarnsmor. Jeg hadde ikke trodd jeg kunne nå toppen
MOR OG BIKINIMESTER: Natalie Aarberg fra Nordreisa kombinerer livet som trebarnsmor for Lykke (14), Milla (10) og Lui (7) med toppsatsingen i bikinifitness. Foto: Privat
feel_inspire
Del


#2
08:06
UNN, TIL | Over natta valgte H
www.nordlys.no
Eirik Nymark Esperås (ansatt i Amedia)
Nordlys
UNN
Over natta valgte Håkon (31) et helt annet liv: - Det har vært litt for deilig
RØD, BLÅ - OG HVIT: Både TIL og TUIL-drakten er pakket vekk for Håkon Kjæves del. I dag er det hvit uniform som gjelder - som turnuslege på UNN. Foto: Torgrim Rath Olsen
feel_inspire
Rath Olsen


#3
08:06
Kristin harila, Klatring | Kristin sl
www.nordlys.no
Eirik Nymark Esperås (ansatt i Amedia)
Nordlys
KRISTIN HARILA
Kristin slår tilbake mot påstandene: - Skuffet
MANASLU: Harila fotografert på vei opp Manaslu. Foto: Kristin Harila/Instagram
Av Mats Rydland
Publisert:
Del
feel_inspire

Dashboard

- Visualization prototype
  https://amedia.eu.looker.com/dashboards/1114

Status and Future Work

Present status

- Two training iterations finished
- Training and inference infrastructure deployed in production settings
  - 100% of published articles, near real time
- Visualisation and reporting are still not ready, just prototypes
- We still do not have a clear picture of how the model will translate into actionables
  - no recipe for conversion, neither for CTR
  - the picture is more complex than what others seem to understand

Future work

- Use as input for representing user preferences
  - personalization of front-pages based on my typical pattern of consumption
- Integration with CMS: in-house development
  - the journalist can see how the content is classified as they are still writing it
  - adjust the desired user need in each article
  - control production style and balance
- Training LLMs to generate user needs specific versions of articles
  - if we produce too much **know_update**
  - make the LLM generate/modify it into other categories, **feel_inspire**, **know_engage**

Future work

- *Dear Santa:*
  merge user needs, different topic classes, personalization and
  AI-generated alternative versions for each article

- Each user experiences the content in a different way because it is adapted to their own taste/behavior

- User 1:
  Sport as **feel_inspire**, Nature/environment as **understand_educate**, rest as default
- User 2:
  No Sport at all, Nature/environment as **do_connect**, rest as default

amedia

Navn.Navnesen@amedia.no
Stillingstittel
Virksomhetsnavn
amedia.no