



AI for trafikssikkerhed:

Domænetilpassede sprogmodeller og semantisk topic modelling til analyse af trafikdata

Mitta Hage og Paula Granlund

Supervisors: Carlos M. Lima Azevedo og Mette Møller

DTU Management, Department of Technology, Management and Economics

Trafikuheld koster hvert år mange liv. Data findes allerede, så hvorfor har vi ikke svarene?

Betjentens egne ord

De strukturerede felter reducerer et uheld til nogle få koder. Fritekstbeskrivelsen fortæller, hvad der faktisk skete, med forløbet, konteksten og de detaljer, som skemaet ikke har plads til.

Det, der ikke har et felt

Nogle mønstre bliver aldrig kodet. Om en fører *flygtede* fra stedet, eller hvad der gik forud for sammenstødet. Beskrivelsen registrerer det alligevel, i klartekst.

Fire årtier, næsten uberørt

Næsten hvert uheld siden 1985 har en beskrivelse, tilsammen omkring 900.000. Det er langt mere, end nogen kunne læse manuelt, og en stort set uudnyttet kilde til viden om trafiksikkerhed.

Dagens Agenda

A stylized illustration of a winding grey road with white dashed lines on a light green background. The road starts from the left, curves upwards, then downwards, then upwards again, and finally curves to the right. Several cars are on the road: a blue car with motion lines trailing behind it, followed by a pink car, a cyan car, an orange car, a purple car, a green car, a yellow car, a blue car, and a red car. Further along the road, there is another purple car and a red car. At the end of the road on the right, a small blue runner is visible. In the bottom center, a blue car is shown with a person standing next to it, and a bicycle is lying on the ground nearby.

1

Motivation og mål

3

Proof of Concept

5

Tre anvendelsestilfælde

2

Forskellen mellem
dette projekt og
ChatGPT

4

Metode

6

Potentiale og næste skridt

Forskellen mellem kendte sprogmodeller og den valgte metode

GENERATIVE SPROGMODELLER

ex. ChatGPT, Gemini

VS

VORES MODEL

BERT-encoder

Vores data forlader huset

Input sendes til en ekstern leverandør, kan bruges til at træne deres model til andre kunder og gemmes på servere uden for EU, et alvorligt GDPR-problem.

Black box & hallucination

En decoder genererer frit tekst, så modellen kan finde på indhold, og vi kan ikke spore, hvor et output stammer fra.

Vores data bliver hos os

Vi hoster modellen selv på egne/danske servere, data forlader aldrig huset, bruges ikke til andre, og overholder GDPR.

Gennemsigtig & sporbar

Encoderen klassificerer i stedet for at generere, så der sker ingen hallucination, og vi ved præcis hvor hvert resultat kommer fra.

Klassifikation er proof of concept

Topic modelling opdager mønstrene

86% nøjagtighed



Friteksten indeholder et signal

Teksten alene indeholder tilstrækkelig information til at forudsige uheldssituationen



Sprogtilpasning forbedrer modellen

Træning på dansk gør modellen bedre til at opfange information i de danske uheldsbeskrivelser.



Fejlene er ikke tilfældige

Modellen forveksler især uheldskategorier, der også er svære at adskille i registreringen.

Klassifikation er proof of concept

Topic modelling opdager mønstrene

1.

Beskrivelse af udheldsforløbet:

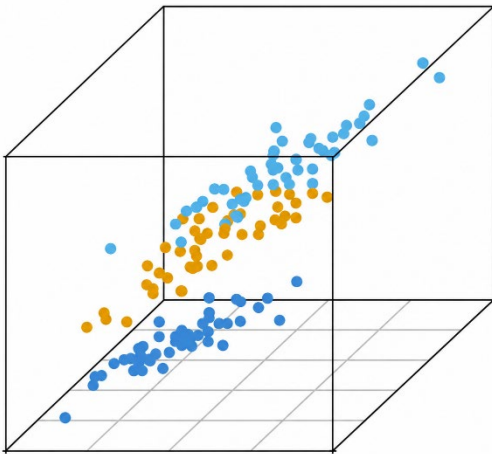
Part 1 personbil kørte i nordlig retning ad den pågældende vejstrækning. Part 2 personbil befandt sig forud for part 1 i samme færdselsretning. Part 1 påkørte part 2 bagfra på kørebanen.

2.

$X =$

0.08	-0.51	0.23	0.06	0.14	-0.12	-0.03	0.78	0.26	0.46
------	-------	------	------	------	-------	-------	------	------	------

3.



4.

Emne 1:
Hit-and-run

Emne 3:
Harmonikasammen-
stød

Emne 2:
Hajtænder

Udvalgte user cases for brugen af modellen

Hvad fritekstbeskrivelserne tilføjer, på tre niveauer

1 VALIDERING

Måltet datakvalitetsarbejde

En udfordring er, at de strukturerede registreringer ikke altid stemmer overens med oplysningerne i fritekstbeskrivelsen. Modellen kan automatisk identificere disse uoverensstemmelser og målrette den manuelle gennemgang mod de registreringer, der kræver nærmere kontrol.

2 BERIGELSE

Gør brede uheldskategorier mere detaljerede

De strukturerede data grupperer uheld i foruddefinerede kategorier. Friteksten kan bruges til at identificere mere specifikke undergrupper og omstændigheder inden for den samme kategori. Det giver mulighed for mere nuancerede trafiksikkerhedsanalyser.

3 MØNSTERGENKENDELSE

Find mønstre, der ikke registreres i dag

De strukturerede data kan kun indeholde information, som der på forhånd er oprettet et felt til. Friteksten kan derimod indeholde hændelser og omstændigheder på tværs af de eksisterende kategorier. Topic modelling kan identificere disse mønstre og dermed gøre hidtil ustruktureret information tilgængelig for analyse.

Samme data, stigende udbytte: validering ligger tættest på skemaet, mens mønsterfund rækker længst væk fra det.

Case 1 Validering - Eksempel:

Ubetinget vigepligt



Hajtænder, vigelinje, vigepligt, ubetinget



Ca. 20% af ulykkerne er ikke klassificeret som ubetinget vigepligt i kategorien *Vigepligt*



Modellen finder uoverensstemmelserne – uden at alle beskrivelser skal gennemgås manuelt

Eksempel 1:

”... part 1 fortsatte ud over hajtænder og ind i part 2s højre bagskærm.”

Vigepligt: NaN

Eksempel 2:

”... part 2 holdt ved markeringen(hajtænder og tavler) for ubetinget vigepligt ... kører part 2 ud i krydset, hvor hun bliver påkørt af part 1...”

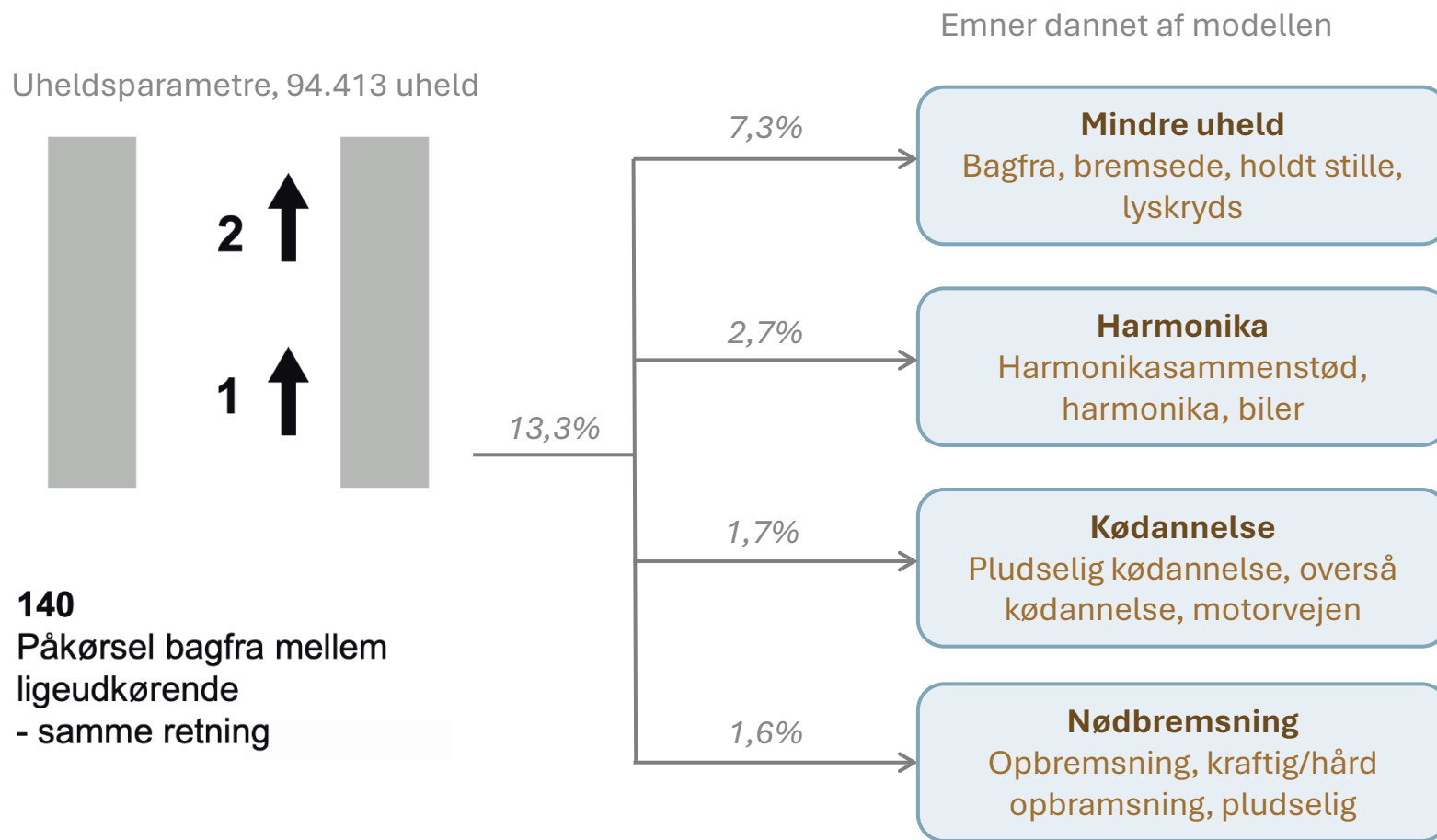
Vigepligt: Ingen

Eksempel 3:

”part 1 standser for ubetinget vigepligt hajtænder . efter kun at have set til venstre påkører han part 2 - en cyklist ... ”

Vigepligt: Højre vigepligt

Case 2 Berigelse - Eksempel: Påkørt bagfra





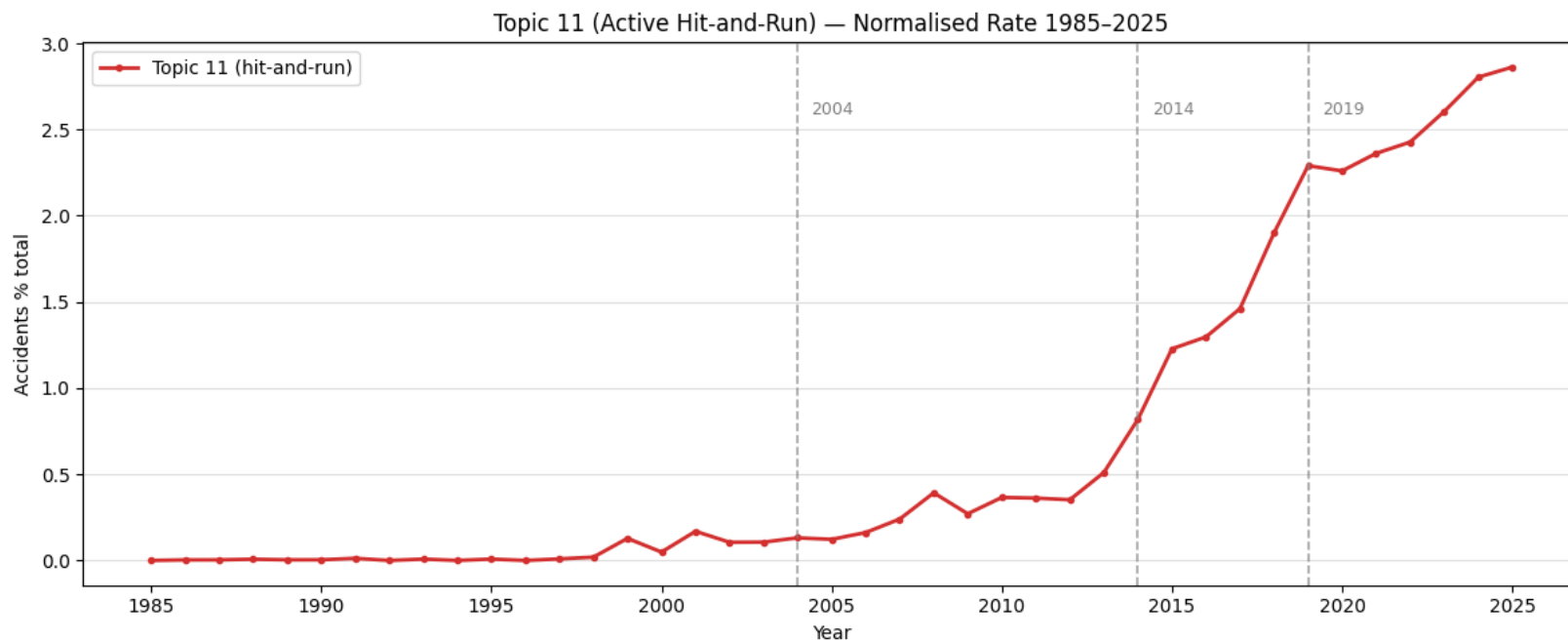
Case 3 Mønstergenkendelse – Eksempel: Manglende standsning efter sammenstød (“hit-and-run”)

5,572

Uheld i topic

41%

Inholder ikke ”fuh-flugt”*



* Different formats of the word is a part of this number

Case 3 Mønstergenkendelse – Eksempel:

Manglende standsning efter sammenstød (“hit-and-run”)

HVAD KUNNE VI UDELUKKE?

- % Omstændigheder: ingen forskel i vejr- eller lysforhold
- % Ingen forskel i målte promilleniveau
- % Ingen forskel i aldersprofil på den flygtende part (element 1)

HVAD KAN VI SÅ SIGE?

- + Høj koncentration i Hovedstaden
- + Høj koncentration på "Fællesvej"
- + Markant overrepræsentation af aldersgruppen 40-65 hos den ramte part (element 2)

Diskussion og konklusion

Et supplement til de strukturerede data, ikke en erstatning

Validering kan markere, ikke dømme

Begge spor lærer af de samme registreringer, de holdes op imod. En uoverensstemmelse markerer et uheld til manuel gennemgang, men kan ikke selv vise, om det er labelen eller modellen, der tager fejl.

Forstørrelsesglas, ikke optælling

Rear-end-clustrene fanger 10,6% af labelen, ikke forekomsten. Medlemskab er geometrisk tæthed i embedding-rummet, ikke en verificeret markering, så kortfattet beskrevne uheld falder udenfor.

Datakvalitet som fælles vilkår

Modellens fejl samler sig, hvor registreringen selv er inkonsistent (5/6 og 0/9), og sproget i beskrivelserne skifter over fire årtier. Fritekst og struktur deler de samme begrænsninger.

PERSPEKTIVERING / NÆSTE SKRIDT

Det største løft ligger i at domænetilpasse encoderen på beskrivelserne uden at lække labels. Genererede topic-navne kan lette fortolkningen, holdt lokalt af hensyn til GDPR, og det samme workflow kan rettes mod kendte svære områder som lyskryds og ny mobilitet.

Tak!

