Denne artikel er udgivet i det elektroniske tidsskrift
Artikler fra Trafikdage på Aalborg Universitet
(Proceedings from the Annual Transport Conference at Aalborg University)
ISSN 1603-9696
https://journals.aau.dk/index.php/td

trafikdage
NY VIDEN & NETVÆRK

AI for trafiksikkerhed
Sprogmodeller og topic modelling til analyse af trafikdata

Mitta Hage (mitta.hage@gmail.com)
Paula Granlund (paulagranlund@hotmail.com)
Carlos M. Lima Azevedo (climaz@dtu.dk)
Mette Møller (mette@dtu.dk)
Technical University of Denmark (DTU), Department of Technology, Management and Economics, Intelligent Transportation Systems

Abstrakt
Politiets uheldsregistrering indeholder både strukturerede variable og en fritekstbeskrivelse om uheldet. Fritekstbeskrivelserne rummer detaljer, som de strukturerede felter ikke fanger, men de bruges i dag næsten kun ved manuel gennemgang. Denne artikel undersøger, om åbne, domænetilpassede sprogmodeller kan gøre de danske fritekstbeskrivelser mere anvendelige i uheldsanalysen. Vi opstiller to analytiske spor baseret på 898.677 politiskrevne fritekstbeskrivelser fra Vejdirektoratets uheldsdatabase (Vejman): en dansk BERT-model, der klassificerer uheldssituationen, og en BERTopic-model, der grupperer fritekstbeskrivelserne efter betydning. Klassifikationen opnår en nøjagtighed på 85,8 % og en makro-F1 på 83,5 %, hvilket viser, at fritekstbeskrivelserne rummer information om uheldet, og at modellen kan
genskabe den registrerede uheldstype ud fra teksten alene. Emnemodelleringen afdækker mønstre, der kan understøtte validering af dataene i politirapporten, berige brede uheldskategorier og identificere nye mønstre, som de eksisterende strukturerede felter ikke fanger, herunder flugtuheld. Resultaterne peger på, at teksterne kan supplere de strukturerede data, ikke erstatte dem, i et lokalt miljø, der kører på ens egen computer og imødekommer databeskyttelseskravene.

Introduktion
Trafiksikkerhedsarbejdet i Danmark bygger i høj grad på viden fra politiregistrerede færdselsuheld. Når politiet registrerer et uheld, bliver oplysningerne en del af Vejdirektoratets officielle uheldsstatistik, som anvendes i arbejdet med lovgivning, planlægning og forebyggelse [Vejdirektoratet, 2017]. Værdien af registreringerne afhænger derfor ikke kun af, at data indsamles, men også af, hvor effektivt de udnyttes.

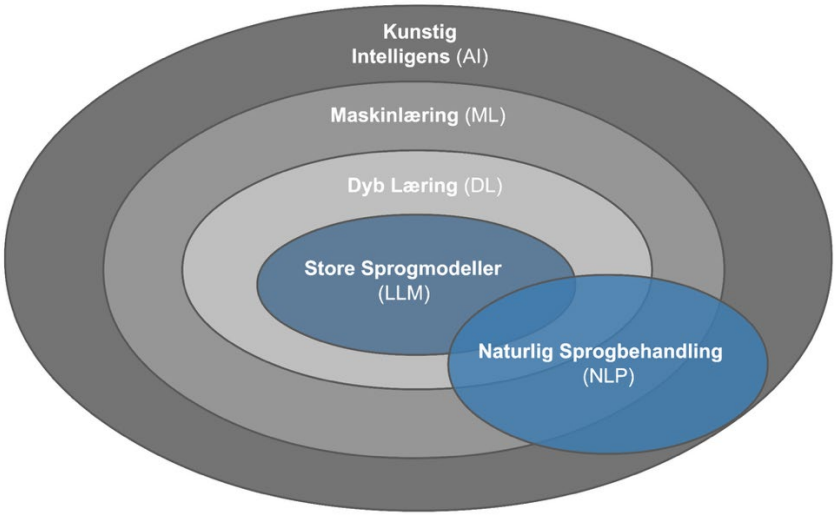
Hver uheldsregistrering kombinerer strukturerede variable med en fritekstbeskrivelse. De strukturerede variable beskriver uheldet gennem faste kategorier, såsom uheldssituation, vejforhold og oplysninger om trafikanterne, mens den registrerende betjent i fritekstbeskrivelsen beskriver hændelsesforløbet med egne ord [Vejdirektoratet, 2017]. Teksten kan dermed rumme kontekstuel information, som de faste felter ikke indfanger, herunder begivenhedernes rækkefølge og detaljer om de involverede parter.

Netop forskellen i format betyder dog, at de to datatyper ikke er lige lette at anvende i større analyser. Da de strukturerede variable allerede er organiseret i foruddefinerede kategorier kan de derfor behandles direkte og systematisk. Fritekstbeskrivelserne består derimod af almindeligt sprog, som først skal omsættes til en numerisk repræsentation, der bevarer tekstens betydning og kontekst, før de kan analyseres systematisk. I praksis gennemgås teksterne derfor hovedsageligt manuelt og fra sag til sag, ofte ved hjælp af ordsøgninger, og fritekstbeskrivelserne anvendes sjældent systematisk, selv om flere strukturerede felter er forbundet med registreringsusikkerhed, og visse uheldstyper ikke har et tilsvarende felt. Fritekstbeskrivelserne rummer således et informationspotentiale, som i dag kun udnyttes i begrænset omfang.

Fremskridt inden for sprogteknologi, særligt transformerbaserede sprogmodeller, gør det muligt at udnytte dette potentiale mere systematisk. Hvor en almindelig ordsøgning kun finder tekster, der indeholder foruddefinerede ord, kan sprogmodeller analysere teksten ud fra dens betydning. De kan derfor identificere beskrivelser af lignende hændelser, selv når betjentene har formuleret sig forskelligt [Wang et al. 2025]. En del af de mest udbredte sprogmodeller er dog lukkede og tilgås via eksterne platforme, hvor teksten sendes til og behandles på udbyderens servere. Det er et problem her, fordi fritekstbeskrivelserne kan indeholde personoplysninger såsom navne og adresser. Sådanne oplysninger er beskyttet af databeskyttelsesreglerne og må ikke uden videre deles med en ekstern part. I dette arbejde anvender vi derfor åbne sprogmodeller, der kan installeres og køres lokalt, så beskrivelserne kan analyseres uden at forlade det sikre datamiljø.

Sprogmodeller og encoder-arkitektur

Moderne sprogmodeller er en del af kunstig intelligens (AI) og bygger på maskinlæring (ML), hvor modeller lærer mønstre fra data frem for at følge faste regler. En del af ML er dyb læring (DL), som anvender neurale netværk med mange lag og danner grundlag for moderne transformerbaserede modeller. Store sprogmodeller (LLM) er en type af disse modeller, mens naturlig sprogbehandling (NLP) er det bredere fagområde, der omfatter metoder til at analysere og behandle menneskeligt sprog. NLP overlapper derfor med LLMs, men omfatter også andre tilgange til tekstanalyse. Inden for området har transformerbaserede modeller fået særlig betydning, fordi de kan fortolke ord ud fra den sammenhæng, de indgår i. Figur 1 viser forholdet mellem de centrale begreber.



Figur 1: Sammenhæng mellem kunstig intelligens, maskinlæring, dyb læring, store sprogmodeller og naturlig sprogbehandling.

Før transformerbaserede modeller blev udbredt, byggede tekstanalyse ofte på ordforekomster. Metoder som TF-IDF og emnemodellen LDA er frekvensbaserede og kan fremhæve karakteristiske ord og identificere overordnede mønstre i store tekstmængder, men de tager ikke højde for ordenes betydning i den kontekst, de optræder i. Et ord får derfor som udgangspunkt samme betydning, uanset hvordan det indgår i teksten, selv om betydningen i praksis kan ændre sig med konteksten [Janstrup et al., 2023].

Det er netop denne begrænsning, transformerbaserede modeller søger at håndtere. I stedet for at betragte ordene isoleret fortolker de hvert ord i relation til resten af teksten [Vaswani et al., 2017]. Transformer-modeller kan overordnet opdeles i encoder- og decoder-modeller, som anvender denne kontekstuelle forståelse på forskellige måder. Encoder-modeller som BERT omsætter tekstens samlede indhold til en matematisk repræsentation, der blandt andet kan bruges til klassifikation og gruppering [Devlin et al., 2018]. Decoder-modeller bruger derimod konteksten til at generere ny tekst. Denne type model ligger blandt andet bag tjenester som ChatGPT og Claude og egner sig til mere åbne opgaver, eksempelvis at uddrage oplysninger eller formulere forklaringer.

I dette arbejde anvender vi encoder-modeller, fordi formålet er at genkende mønstre i uheldsbeskrivelserne og gruppere dem efter deres indhold, ikke at generere ny tekst. Encoder-modeller er desuden typisk mindre og kræver færre computerressourcer end store decoder-modeller, hvilket gør dem velegnede til lokal behandling i et sikkert datamiljø. En decoder-model kunne potentielt uddrage mere komplekse og detaljerede oplysninger, men til klassifikation af en enkelt uheldslabel vurderes denne merværdi ikke at opveje de større krav til beregning og databehandling. Tabel 1 sammenfatter forskellene mellem de tre tilgange.

Tabel 1: Sammenligning af de tre tilgange til tekstanalyse

Tilgang	Hvordan behandles teksten?	Typisk anvendelse	Typisk udfordring
Klassisk text mining (TF-IDF, LDA)	Teksten beskrives ud fra, hvilke ord der forekommer, og hvor ofte de optræder	Ordsøgning, simple emner	Mister sproglig sammenhæng
Encoder-sprogmodel (BERT)	Teksten omsættes til en matematisk repræsentation af dens samlede betydning	Klassifikation og gruppering	Svær at gennemskue
Decoder-sprogmodel (stor sprogmodel)	Modellen fortolker teksten og genererer et svar	Informationsudtræk og forklaringer	Kan hallucinere

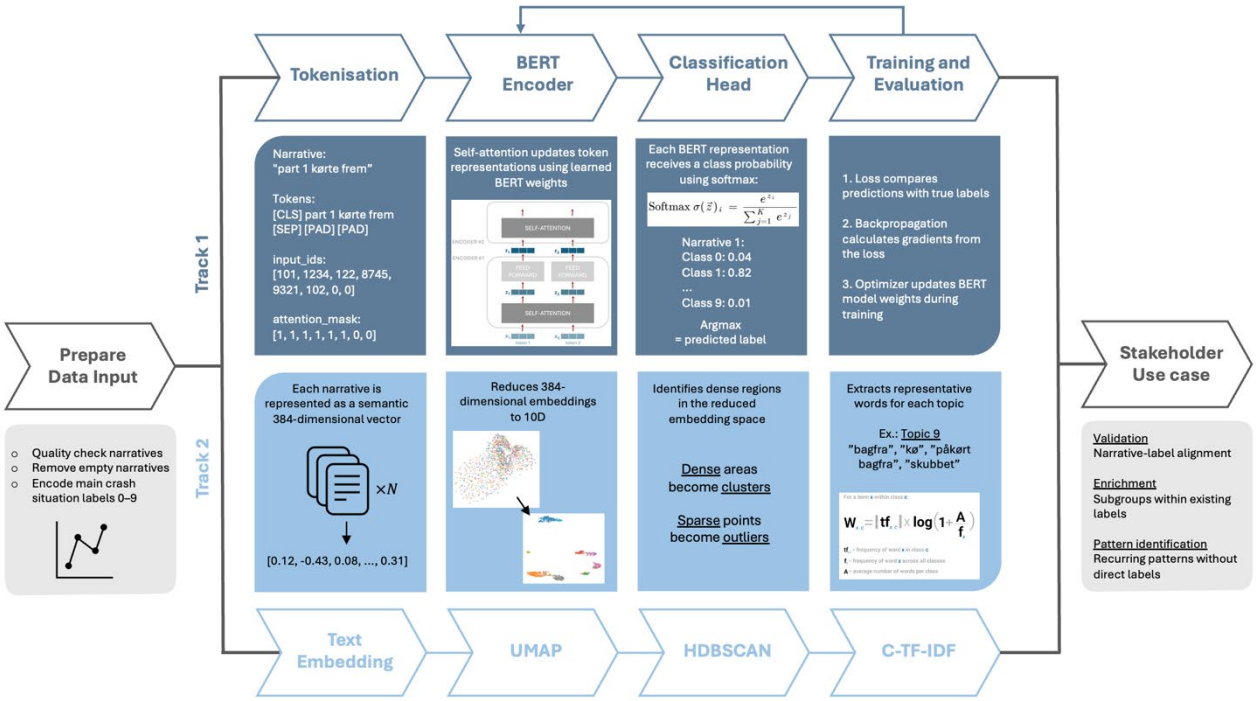
De to spor: klassifikation og topic modelling

Analysen er opbygget i to metodiske spor, som begge tager udgangspunkt i politiets fritekstbeskrivelser, men stiller forskellige spørgsmål til dem. Figur 2 viser det samlede analyseforløb fra dataforberedelse til fortolkning af resultaterne.

I det første spor anvendes supervised classification med en dansk BERT-model. Modellen trænes til at forudsige uheldets registrerede hovedsituation ud fra fritekstbeskrivelsen alene. Formålet er at undersøge, om en fortrænet sprogmodel, altså en model, der allerede har lært almindeligt dansk sprog, kan tilpasses uheldsteksterne, og i hvilken grad den kan genskabe en allerede kendt, struktureret label. Sporet fungerer dermed som et proof of concept for, at information i teksten kan omsættes til et struktureret output.

Det andet spor undersøger fritekstbeskrivelserne uden at tage udgangspunkt i én bestemt label. Med BERTopic grupperes de efter semantisk lighed, så mønstre kan træde frem på baggrund af teksternes indhold. En semi-supervised udvidelse lader udvalgte strukturerede variable påvirke grupperingen, uden at de bestemmer den fuldstændigt. På den måde kan modellen både bygge på eksisterende viden om uheldene og bevare muligheden for at afdække undergrupper, uoverensstemmelser og mønstre, som ikke er direkte repræsenteret i de strukturerede data.

De to spor besvarer derfor forskellige dele af det samlede spørgsmål. Klassifikationen undersøger, om kendt uheldsinformation kan genfindes i fritekstbeskrivelserne, mens topic modelling undersøger, hvordan de kan anvendes mere udforskende. Resultaterne fra begge spor vurderes efterfølgende gennem tre anvendelsestilfælde: validering af eksisterende registreringer, berigelse af brede uheldskategorier og identifikation af mønstre uden et tilsvarende struktureret felt. De følgende afsnit beskriver metoden bag de to spor. Først gennemgås klassifikationen i spor 1 og derefter topic modelling i spor 2.



Figur 2: Det samlede analyseforløb. Efter dataforberedelse behandles beretningerne i to spor. Spor 1 klassificerer teksten med en BERT-model. Spor 2 danner tekstindlejringer, der reduceres, grupperes og beskrives med c-TF-IDF. Resultaterne fortolkes gennem tre anvendelsestilfælde.

Spor 1: Klassifikation med Dansk BERT

Klassifikation går ud på at lære en model at placere en tekst i en af flere på forhånd kendte kategorier. Modellen får under træningen et stort antal eksempler, hvor både teksten og det rigtige svar er givet, og lærer derved at genkende de sproglige træk, der kendetegner hver kategori. Her er teksten fritekstbeskrivelsen, og det rigtige svar er uheldets registrerede hovedsituation, som dermed fungerer som målvariabel, altså den værdi modellen skal lære at forudsige. Klassifikationen bygger på den danske BERT-model danish-bert-botxo, som er trænet på dansk tekst og derefter finjusteret til den konkrete klassifikationsopgave [Certainly, 2019].

For at kunne vurdere, om modellen har lært noget generelt og ikke blot har genkaldt de eksempler, den så under træningen, opdeles datasættet i tre dele. Modellen trænes på 72 % af data, 8 % bruges undervejs til at justere opsætningen, og de sidste 20 % holdes helt tilbage og bruges kun til den afsluttende evaluering. Opdelingen sker, så den oprindelige fordeling mellem de registrerede hovedsituationer bevares i alle tre dele. Da nogle kategorier forekommer væsentligt oftere end andre, vægtes fejl i de sjældnere kategorier højere under træningen.

Den færdige model evalueres på testsættet med nøjagtighed og makro-F1. Hvor nøjagtigheden viser den samlede andel af korrekte forudsigelser, beregner makro-F1 resultatet for hver kategori separat og vægter derefter alle kategorier lige. Netop makro-F1 er derfor det vigtigste mål her, fordi det viser, om modellen også rammer de sjældne uheldssituationer og ikke kun de hyppige.

Spor 2: Emnemodellering med BERTopic

Hvor klassifikation placerer teksten i kendte kategorier, går topic modelling den modsatte vej og leder efter mønstre, uden at kategorierne er givet på forhånd. Modellen får altså ikke det rigtige svar at træne på, men grupperer i stedet fritekstbeskrivelserne efter, hvor ens de er i indhold, og lader mønster, altså emnerne, træde frem af teksten selv.

Topic modelling gennemføres med BERTopic [Grootendorst, 2022]. Hver fritekstbeskrivelse omsættes først til en numerisk repræsentation af dens betydning ved hjælp af en dansk sentence encoder [Martinsen, 2024]. Repræsentationen er en talrække, der placerer beskrivelser med lignende betydning tæt på hinanden, så nærhed svarer til indholdsmæssig lighed. UMAP reducerer derefter repræsentationernes dimensionalitet, så de kan grupperes mere effektivt [UMAP Documentation, 2024]. På dette grundlag samler HDBSCAN beskrivelser med lignende indhold, mens beskrivelser, der ikke passer tydeligt ind i en gruppe, registreres som outliers [HDBSCAN Documentation, 2016]. For hver gruppe anvendes c-TF-IDF til at udpege de ord, der bedst kendetegner emnet.

Analysen anvender en semi-supervised udvidelse, hvor uheldets registrerede hovedsituation indgår som vejledning i UMAP-trinnet. Formålet er at lade den eksisterende viden om uheldene styre grupperingen en smule, uden at den overtager den helt. Den strukturerede variabel bestemmer dermed ikke grupperingen, men påvirker placeringen af beskrivelserne, så beskrivelser med samme label i nogen grad trækkes tættere sammen. Hvor meget den får lov at trække, styres af en vægt mellem 0 og 1 [Grootendorst, 2024]. På baggrund af en følsomhedsanalyse blev vægten sat til 0,3. Denne konfiguration gav tydeligere adskilte og mere forskellige emner end de øvrige relevante vægte, om end en større andel af beskrivelserne blev placeret som outliers.

Resultater

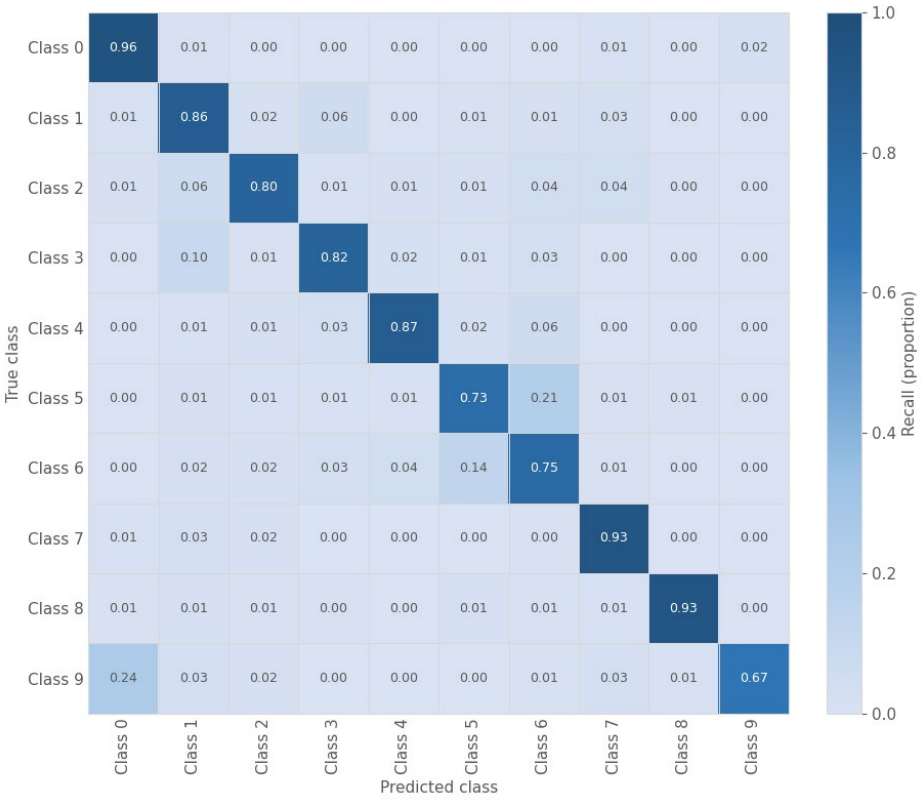
Resultaterne falder i to dele, der svarer til de to spor. Først viser klassifikationen, at fritekstbeskrivelserne bærer genkendelig information om uheldet. Det er forudsætningen for resten: kan en model ikke genfinde det kendte, er der ingen grund til at lede efter det ukendte. Dernæst bruger emnemodelleringen den samme tekst til at afdække mønstre, der rækker ud over de strukturerede kategorier, samlet i tre anvendelsestilfælde.

Spor 1: Klassifikationen som Proof of Concept

Spor 1 er ikke et mål i sig selv, men en test af grundantagelsen. Hvis en model kan genskabe uheldets registrerede hovedsituation ud fra fritekstbeskrivelsen alene, viser det, at teksten indeholder uheldsinformation i en form, en model kan arbejde med. Først når det er fastslået, giver det mening at bruge teksten til at finde mønstre, de strukturerede felter ikke fanger.

Den finjusterede danske BERT-model opnår en nøjagtighed på 85,8 % og en makro-F1 på 83,5 % på testmængden. En tilsvarende engelsksproget model klarer sig dårligere, hvilket bekræfter værdien af en dansk model. Modellen genskaber de brede uheldssituationer fra teksten alene, og grundantagelsen holder dermed: fritekstbeskrivelserne bærer den information, de strukturerede variable indeholder.

Fejlene fordeler sig ikke tilfældigt. De samler sig mellem kategorier, der også i praksis er svære at holde adskilt. Eksempelvis forveksles klasse 0, som omfatter eneulykker, ofte med klasse 9, der dækker eneulykker med faste genstande på kørebanen. Tilsvarende ses en markant forveksling mellem klasse 5 og 6. Figur 3 viser dette. Modellen fejler dermed netop dér, hvor selve registreringen også er forbundet med usikkerhed. En uenighed mellem model og registreret label er derfor ikke nødvendigvis en modelfejl, men kan lige så vel pege på en usikker registrering.



Figur 3: Forvekslingsmatrix for de ti uheldssituationer. Hver række er normaliseret, så værdierne viser andelen af en klasses fritekstbeskrivelser fordelt på modellens forudsigelser

Spor 2: Emnemodellering og tre Anvendelsestilfælde

Emnemodelleringen grupperer fritekstbeskrivelser med lignende betydning og identificerer de ord, der kendetegner hver emne. Emnerne er ikke definerede kategorier, men datadrevne grupperinger, der udspringer af mønstre i fritekstbeskrivelserne.

Modellen dannede 40 emner og grupperingerne varierer med hensyn til, hvad de fanger. Nogle svarer til en uheldsmekanisme (fx påkørsel bagfra), andre til måden en rapport er skrevet på, og andre igen til en omstændighed ved uheldet. Ikke alle emner lader sig fortolke som en uheldstype, og de bør derfor ses som datadrevne grupperinger, ikke som færdige kategorier. Frem for at gennemgå de enkelte emner samler vi resultatet i tre anvendelsestilfælde, der viser, hvad grupperingerne kan tilføre uheldsanalysen. Det første bruger et emne til at validere de strukturerede registreringer. Det andet deler en bred struktureret kategori i mere præcise undergrupper. Det tredje finder en uheldstype, der ikke har noget struktureret felt.

Tilfælde 1: validering af registreringer

Det første tilfælde bruger et emne til at kontrollere de strukturerede registreringer. Vejdirektoratets gennemgang og rettelse af registreringer er ressourcekrævende og bygger på manuelle procedurer. Emnemodellen kan pege på sager, hvor teksten og det strukturerede felt trækker i hver sin retning. Vi

illustrerer dette med variablen Vigepligt, der angiver vigepligtsforholdet for hver part i uheldet. Modellen dannede et emne på 2.213 uheld kendetegnet ved omtale af vigelinjer. Cirka 500 af disse uheld var dog ikke registreret med en tilsvarende kategori i variablen Vigepligt. Nedenfor præsenteres et eksempel på et sådant tilfælde. Her angiver beskrivelsen af uheldet, at krydset var reguleret med ubetinget vigepligt og vigelinje, mens den tilhørende politiregistrering angiver, at vigepligten var højre vigepligt. Af hensyn til fortrolighed er vejnavnene anonymiserede og erstattet med fiktive navne.

" Part 1 kører i sin bil ad Skovvej i sydgående retning, hvor han stopper ved T-krydset Skovvej/Åvej på grund af ubetinget vigepligt og vigelinjer. Efter kun at have orienteret sig mod venstre påkører han part 2, en cyklist, som kørte ad Åvej i nordøstlig retning på cykelstien i den forkerte side af vejen. Part 2's cykel blev næsten fuldstændig ødelagt."

Dette betyder ikke nødvendigvis, at registreringen er forkert, men eksemplet viser, hvordan topic modelling kan anvendes til at identificere uheld, der potentielt kræver nærmere gennemgang. I stedet for at gennemgå alle registreringer manuelt én for én kan metoden dermed fungere som et værktøj til at pege på tilfælde, hvor en manuel kontrol kan være relevant.

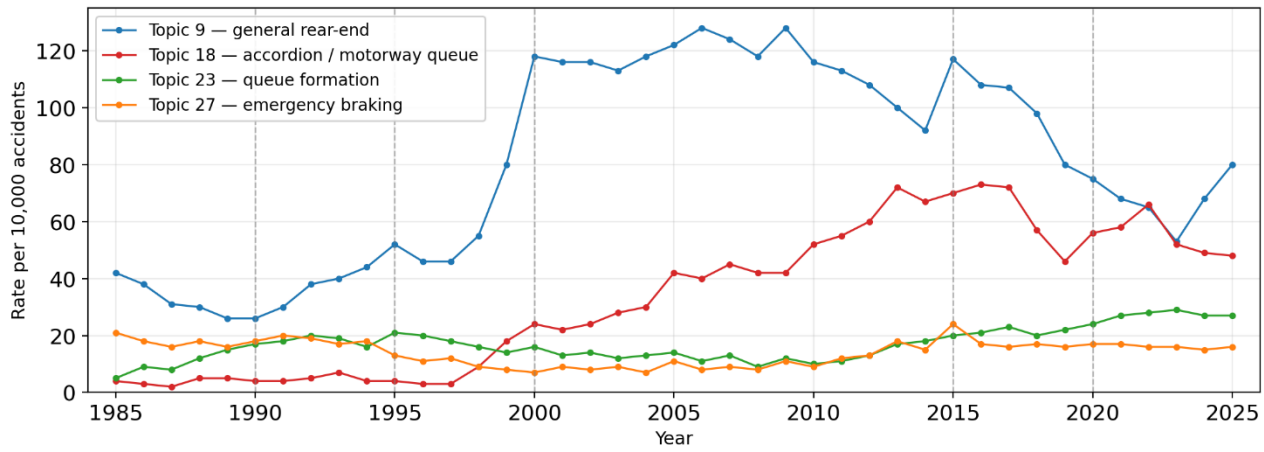
Tilfælde 2: berigelse af brede kategorier

Det andet tilfælde viser, hvordan emnemodellen kan opdele en bred struktureret kategori i mere præcise undergrupper. Den strukturerede kategorien "påkørsel bagfra mellem ligeudkørende – samme retning" dækker 94.413 uheld. Emnemodellen identificerer fire emner, som alle er repræsenteret inden for denne kategori: påkørsel bagfra mellem to køretøjer (emne 9), harmonikasammenstød (emne 18), kødannelseskollision (emne 23) og nødbremsningskollision (emne 27).

Tabel 2: Profil for de fire undergrupper af bagfrakollisioner sammenlignet med kategorien som helhed

Klynge	Antal	Motorvej	Typisk geografi	Flere end 2 køretøjer	Alkohol
Emne 9 (to køretøjer)	6.850	34,8 %	Spredt, byområder	8,1 %	4,4 %
Emne 18 (harmonika)	2.542	59,2 %	E45/E20-korridoren	61,4 %	0,5 %
Emne 23 (kødannelse)	1.545	74,4 %	Motorring 3, København	24,0 %	0,9 %
Emne 27 (nødbremsning)	1.361	51,6 %	Blandet	12,7 %	2,3 %
Bagfrakollision (alle)	94.413	38,9 %	Hele landet	10,5 %	7,7 %

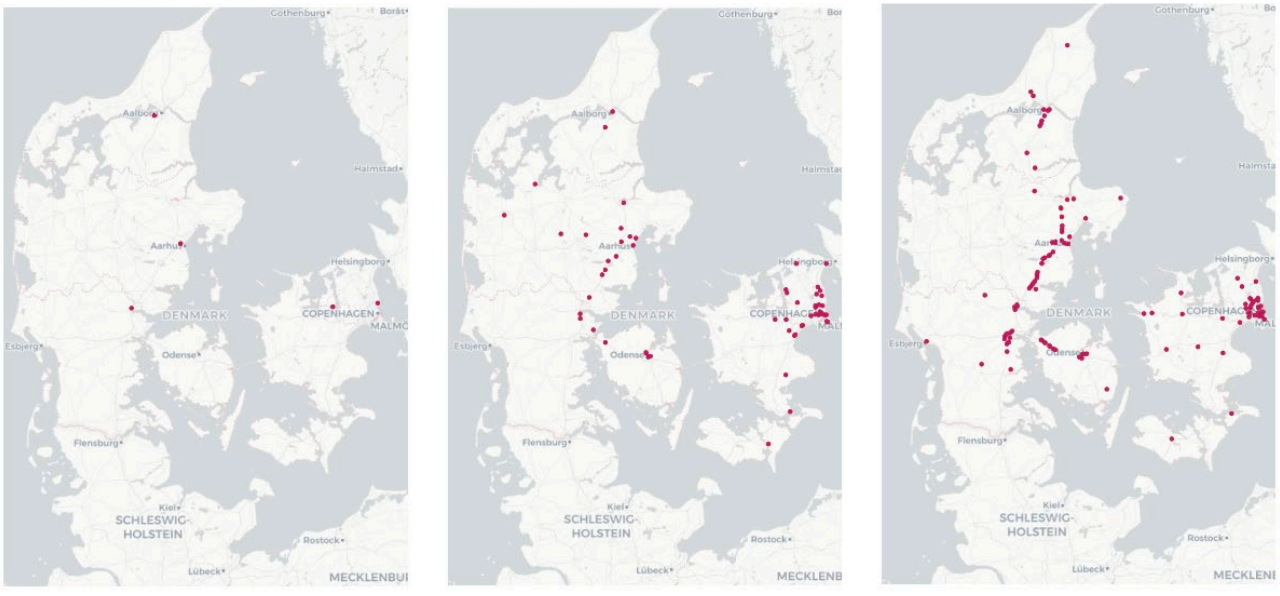
De fire grupper adskiller sig med hensyn til vejtype, geografi, antal køretøjer, alkohol og alder, som vist i Tabel 2. To af klyngerne, emne 18 og emne 23, har næsten samme andel af motorvejsuheld, men ligger geografisk helt forskelligt. Emne 18 følger E45/E20-korridoren gennem Jylland, mens emne 23 samler sig om Motorring 3 vest for København. Det strukturerede vejkategorifelt behandler begge som motorvejsuheld og kan derfor ikke skelne dem, mens fritekstbeskrivelserne gør forskellen synlig. Klyngerne udvikler sig også forskelligt over tid. Figur 4 viser hver klynges andel af uheldene år for år, med de skift i niveau, der er fundet med changepoint-metoden PELT [Truong et al., 2020]. Emne 18 er næsten fraværende før år 2000 og stiger derefter markant.



Figur 4: Udvikling i de fire emner over tid, normaliseret efter det samlede antal uheld per år. Det øverste panel viser hele bagfrakollisions-kategorien, de nederste hver enkelt klynge. Stiplede linjer markerer de skiftpunkter, der er fundet med PELT.

Figur 5 viser den geografiske udvikling for emne 18 i tre nedslagsår. I 1997 er uheldene få og spredte uden noget klart mønster. I 2006 samler de sig om de større byer og begynder at tegne en linje ned gennem Østjylland. I 2017 står mønsteret tydeligt: en næsten sammenhængende korridor langs E45 fra Aalborg over Aarhus mod trekantområdet, videre ad E20 over Fyn og ind mod hovedstadsområdet. Emne 18 dækker harmonikasammenstød, hvor flere køretøjer kører op i hinanden i tæt trafik med kødannelse, og den type uheld hører netop hjemme på strækninger med meget gennemkørende trafik.

Udviklingen falder tidsmæssigt sammen med to store udvidelser af det overordnede vejnet. Storebæltsforbindelsen åbnede i 1998 og Øresundsbroen i 2000, og begge ledte mere gennemkørende trafik ind på netop denne korridor. Sammenfaldet er dog tidsmæssigt og ikke nødvendigvis årsagsbestemt. Stigningen kan også afspejle, at trafikmængden voksede i perioden, eller at flere uheld blev registreret, og analysen kan ikke adskille disse forklaringer.



(a) 1997 *(b) 2006* *(c) 2017*

Figur 5: Geografisk fordeling af emne 18 (harmonikasammenstød) i 1997, 2006 og 2017. Korridormønsteret langs E45/E20 opstår efter år 2000 og står tydeligt i 2017.

Tilfælde 3: mønstre uden et struktureret felt

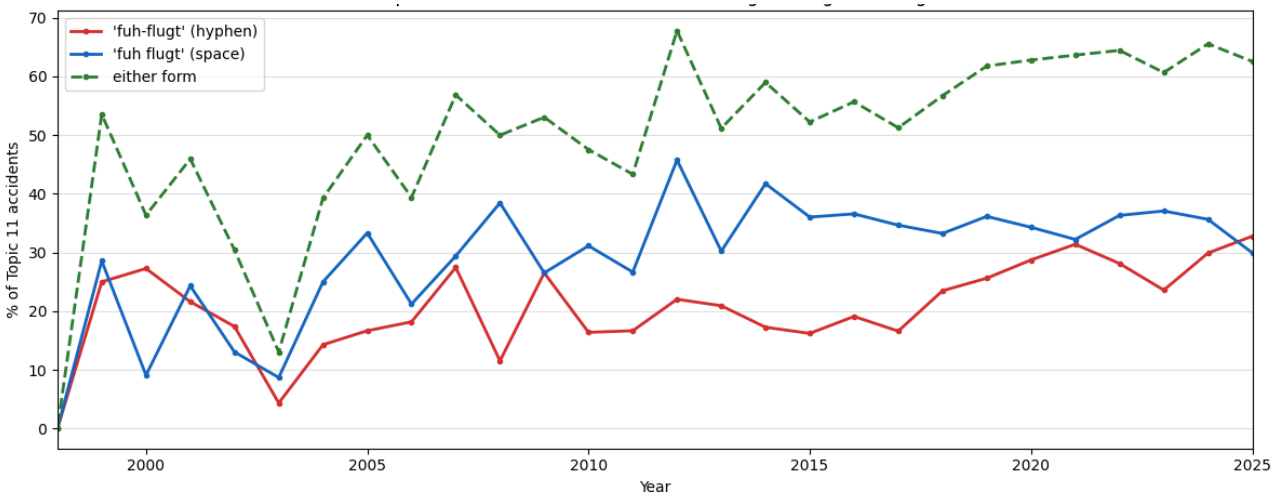
Det tredje tilfælde handler om mønstre, der ikke har nogen struktureret variabel. Vejdirektoratet har ingen variabel, der direkte registrerer, om en fører flygtede fra stedet. Alligevel dannede emnemodellen emne 11, der samler flugtuheld mellem kørende parter. Emnet er kendetegnet ved ord som flugt, flygtet og ukendt og strækker sig over 89 % af de registrerede uheldssituationer, hvilket viser, at ingen enkelt struktureret variabel isolerer denne uheldstype.

Det tætteste, de strukturerede data kommer, er et felt, der kan udfyldes, når den skyldige part stadig er ukendt ved rapportens afslutning. Feltet siger dog ikke, at en part flygtede, kun at ingen blev fundet, og det fanger derfor ikke de flugtuheld, hvor føreren senere blev identificeret. Det er udfyldt i 3,9 % af uheldene i emne 11 mod 0,67 % i hele datasættet, altså langt hyppigere blandt flugtuheld, men stadig kun i en mindre del af dem.

Emnet fanger derimod selve handlingen, uanset hvad der siden skete. Et flugtuheld dækker over, at en fører flygtede eller forsøgte at flygte, men føreren kan være blevet fanget af politiet, fundet efterfølgende eller aldrig fundet. En fører, der flygtede og senere blev fundet, giver samme type beskrivelse som en, der aldrig blev identificeret, og emnet rummer da også ord om eftersættelse og standsning. Netop derfor er teksten ofte det eneste sted, flugten fremgår. Følgende beskrivelse har ingen registrering i de strukturerede data, men afslører ved læsning et flugtuheld, hvor føreren blev fanget kort efter. Vejnavnet er fiktivt.

"Part 2 holdt for rødt, da part 1 påkørte bagenden af part 2's bil og herefter kørte fra stedet. Fører af part 1 blev kort efter eftersat og bragt til standsning på Ringvej."

Emnet kan ikke genfindes med en simpel ordsøgning. Knap 60 % af fritekstbeskrivelserne indeholder ordformen fuh-flugt, mens resten kun findes gennem den betydningsmæssige lighed, sætnings-encoderen fanger. En ordsøgning ville derfor overse omtrent hvert tredje flugtuheld. Figur 6 viser, at denne andel er stabil over hele perioden. Det illustrerer værdien af en betydningsbaseret repræsentation frem for eksakt ordmatch.

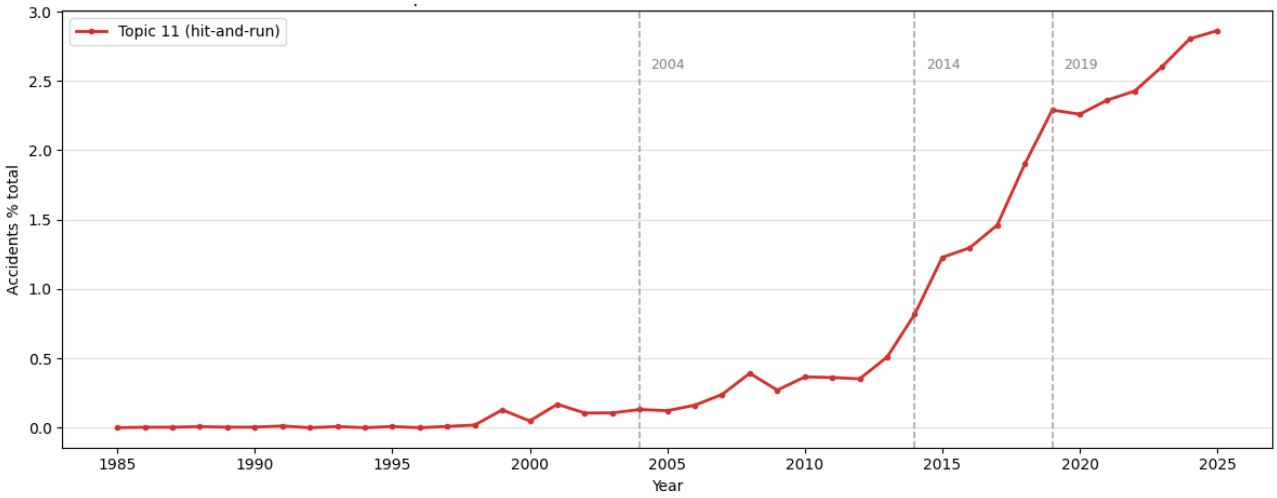


Figur 6: Andel af emne 11, der indeholder ordformen fuh-flugt, per år. Resten findes alene gennem semantisk lighed

Uheldene sker overvejende i lavhastighedsmiljøer. Kommuneveje og private fællesveje er overrepræsenterede, og skaderne er mindre alvorlige end i datasættet som helhed. Det kan hænge sammen med, at en fører lettere kan forlade stedet efter et uheld ved lav hastighed. Alkohol er registreret i 4,2 % af

uheldene mod 11,7 % i hele datasættet. Den forskel skyldes sandsynligvis, at en alkoholtest ikke kan gennemføres, når føreren er flygtet, og ikke at spirituskørsel er sjældnere i flugtuheld. Underrepræsentationen er altså en måleartefakt, ikke et adfærdsmønster.

Emnet er næsten fraværende før 2004 og stiger derefter støt. Figur 7 viser udviklingen med changepoints i 2004, 2014 og 2019. Stigningen falder sammen med flere ændringer i lovgivningen i perioden, herunder klippekortordningen fra 2005 og vanvidskørselsloven fra 2021. Ændringerne udgør mulige medvirkende faktorer, men analysen kan ikke fastslå, om de har påvirket udviklingen. Stigningen kan også afspejle, at flugtuheld er blevet beskrevet anderledes i fritekstbeskrivelserne over tid, snarere end at de er blevet hyppigere.



Figur 7: Emne 11 som andel af alle uheld per år, 1985–2025. Stiplede linjer markerer de changepoints, der er fundet med PELT, i 2004, 2014 og 2019.

Diskussion og konklusion

De to spor viser tilsammen, at politiets fritekstbeskrivelser indeholder information ud over de strukturerede variable. Klassifikationen fastslår, at informationen er til stede, og emnemodelleringen viser, at teksten kan afdække mønstre, som de faste kategorier ikke fanger, fra usikre registreringer over undergrupper i brede kategorier til uheldstyper uden et struktureret felt.

Fritekstbeskrivelserne bør dog ses som et supplement til de strukturerede data, ikke som en erstatning. De registrerede kategorier gør det muligt at tælle og sammenligne uheld på tværs af år, steder og vejtyper, fordi alle uheld beskrives ud fra de samme faste valgmuligheder. Et bagfrauheld registreres for eksempel med samme kode, uanset hvilken betjent der udfylder rapporten, og uanset hvor i landet det sker. Teksten giver ikke det samme grundlag. To betjente kan beskrive samme uheld forskelligt, og samme ordvalg kan dække forskellige uheld. Fritekstbeskrivelserne kan derfor ikke give den ensartethed, som officiel statistik og sammenligning over tid hviler på.

Resultaterne skal fortolkes med forbehold. Analyserne bygger på politiregistrerede uheld og afspejler derfor ikke den fulde mængde af uheld i trafikken. Beskrivelsen er desuden betjentens gengivelse af hændelsen ud fra de oplysninger, der forelå på stedet, og ikke nødvendigvis et fuldstændigt billede. Endelig grupperer emnemodellen fritekstbeskrivelser efter, hvordan de er skrevet, og ikke kun efter, hvad der skete. To forhold følger af det. For det første kan et emne samle uheld, der ligner hinanden sprogligt, uden at de nødvendigvis udgør den samme uheldstype, og fortolkningen af et emne afhænger derfor af specifik viden om trafikuheld

og deres kontekst. For det andet kan ændringer i sprog og registreringspraksis over de fire årtier påvirke, hvilke mønstre der træder frem. En stigning i et emne over tid bør derfor ikke uden videre tolkes som en stigning i den underliggende uheldstype, sådan som flugtuheldsemnet illustrerer.

For Vejdirektoratet ligger den umiddelbare værdi i at målrette arbejdet med datakvalitet, analyse og opfølgning på uheldsdata. En narrativbaseret kontrol kan lede den manuelle gennemgang mod de registreringer, hvor model og label er uenige, frem for mod en tilfældig stikprøve. Emnerne kan desuden nuancere brede kategorier, så spørgsmål om eksempelvis trængsel, kapacitet eller bestemte uheldsforløb kan undersøges på det niveau, hvor de faktisk optræder i beskrivelserne. Samtidig kan metoden fremhæve nye mønstre, før de eventuelt formaliseres som nye strukturerede felter. I alle tre tilfælde fungerer teksten dermed som et supplement til de eksisterende data ved at pege på forhold, der bør undersøges nærmere, snarere end at levere et endeligt svar.

Samlet set viser analysen at åbne, domænetilpassede sprogmodeller kan gøre de danske fritekstbeskrivelser mere anvendelige i uheldsanalysen, uden at data forlader et sikkert miljø. Den største værdi ligger i at understøtte udforskende analyser som supplement til de strukturerede data og den faglige vurdering. Teksten kan pege på mønstre og uoverensstemmelser, som efterfølgende kan vurderes af en ekspert.

Det videre arbejde

Flere muligheder for videre arbejde udspringer af resultaterne. Flugtuheldsemnet hviler på en begrænset manuel gennemgang og bør vurderes af fageksperter, før der drages konklusioner om det. En sådan validering kan samtidig afklare, hvor godt emnet adskiller de reelle flugtuheld fra sprogligt lignende beskrivelser.

Klassifikationsmodellen kan afprøves i forbindelse med Vejdirektoratets kvalitetssikring af uheldsdata. Ved at anvende modellen på registreringer med forskellig grad af sikkerhed i de eksisterende labels kan det undersøges, om modellens uoverensstemmelser identificerer potentielle fejlregistreringer, og om tekstbaseret screening kan målrette den manuelle gennemgang.

Undergrupperne af bagfrakollisioner kan bruges som udgangspunkt for videre analyse. Skellet mellem harmonikasammenstød og kødannelseskollisioner knytter sig til trængsel og vejkapacitet og kan undersøges nærmere sammen med trafikdata for de berørte strækninger.

Metodisk kan emnernes kvalitet forbedres på to måder. Sætnings-encoderen kan tilpasses uheldsdomænet, så grupperingerne i højere grad følger, hvad der skete, frem for hvordan det er skrevet. Og fortolkningen af emnerne kan understøttes af en decoderbaseret model, der kan sammenfatte en klynge eller uddrage mere komplekse oplysninger, end en enkelt label rummer, forudsat at den kan køre lokalt af hensyn til databeskyttelsen.

Referencer

Certainly (2019). *Nordic BERT*. GitHub-repositorium. URL: https://github.com/certainlyio/nordic_bert.

Devlin, J. et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". I: *CoRR* abs/1810.04805. arXiv: 1810.04805.

Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". I: *arXiv preprint* arXiv:2203.05794.

Grootendorst, M. (2024). Semi-supervised Topic Modeling. BERTopic documentation. URL: https://maartengr.github.io/BERTopic/getting_started/semisupervised/semisupervised.html.

Han, Z. og J. Oke (2025). "Novel Typology of Fatal Crash Behaviors via Natural Language Processing". I: *Data Science for Transportation* 7.3, s. 22. DOI: 10.1007/s42421-025-00135-3.

HDBSCAN Documentation (2016). "How HDBSCAN Works". hdbscan 0.8.1 documentation. URL: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

Janstrup, K. H. et al. (2023). "Predicting injury-severity for cyclist crashes using natural language processing and neural network modelling". I: *Safety Science* 164, s. 106153. DOI: 10.1016/j.ssci.2023.106153.

Martinsen, K. T. (2024). *MiniLM-L6-danish-encoder*. Hugging Face model card. URL: https://huggingface.co/KennethTM/MiniLM-L6-danish-encoder.

Truong, C., L. Oudre og N. Vayatis (2020). "Selective review of offline change point detection methods". I: *Signal Processing* 167, s. 107299. DOI: 10.1016/j.sigpro.2019.107299.

UMAP Documentation (2024). "How UMAP Works". umap 0.5.8 documentation. URL: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html.

Vaswani, A. et al. (2017). "Attention Is All You Need". I: *CoRR* abs/1706.03762. arXiv: 1706.03762.

Vejdirektoratet (2017). *Indberetning af færdselsuheld – Kodeark.* København: Vejdirektoratet.

Wang, X. et al. (2025). *Domain-Adapted Pre-trained Language Models for Implicit Information Extraction in Crash Narratives.* arXiv preprint. DOI: 10.48550/arXiv.2510.09434.