

UPPSALA
UNIVERSITET

Språkmodeller och estimation

Slutrapport: Transformerbaserade språkmodeller och aktiv inlärning för skattning av ändliga populationer av textuella data, med tillämpning inom rättsstatistik

Måns Magnusson

Informationsklass: K1

Abstract

This project explores the use of transformer-based language models to estimate finite populations from textual data, focusing on crime statistics. In collaboration with the Swedish National Council for Crime Prevention (Brå), the project investigates the feasibility of applying pre-trained language models for statistical estimation in crime statistics. Initial results demonstrate that these models can significantly enhance classification accuracy, particularly in areas with extensive labelled data that can be used for training, such as Swedish hate crime statistics. The idea builds upon prediction-powered estimation, and the results in hate crime statistics are promising. The prediction model used to predict hate crime is efficient and is better or on par with the police manual annotations. This prediction can then be used together with sampling estimators, making estimators almost 80 times as efficient as using estimators built upon simple random sampling. We also examine the theoretical foundation of prediction-powered estimators in this setting, and these results underscore the importance of calibration in transformer models to mitigate biases and enhance statistical reliability. However, the effectiveness of these estimators builds upon good performant prediction models. Hence, the performance is limited when training data is sparse, prompting further exploration of Bayesian Active Learning for efficient selection of observations for manual annotation. Finally, the project also examines the ethical implications of using transformer-based language models in criminal statistics, particularly concerning bias, transparency, and legal accountability. One key concern is that these models cannot handle complex judicial problems or complex contextual information, leading to systematic under- or over-representation of certain groups in hate crime classification. The project explores ways to mitigate such biases, including using unbiased statistical estimators and thoroughly vetting developed models.

Innehållsförteckning

Introduktion 4
Projektets frågeställningar 5
Arbetets utförande 5
Rapportens omfattning och upplägg 6
Resultat 8
Artikel 1: “Prediction-powered estimators for finite population statistics in highly imbalanced textual data: public hate crime estimation” 8
Metod 8
Resultat 9
Diskussion och slutsats 9
Artikel 2: “Legal and Ethical Implications of Transformer-Assisted Hate Crime Classification and Estimation in Sweden” 10
Hatbrottsbedömningar i Sverige 10
Språkmodell för klassificering av hatbrott 10
Systematiska fel i klassifikation av hatbrott 11
Juridiska och etiska implikationer 11
Slutsatser och rekommendationer 12
Påbörjad artikel 3: Bayesian Active Learning using Pareto Smoothed Importance Sampling 13
Bayesiansk aktiv inlärning 13
Metod 13
Resultat 14
Slutsatser och fortsatt arbete 14
Övriga resultat: Godkänd etikprövning och forskningsmiljö för AI-forskning vid Brå 15
Betydelse för Länsförsäkringar 16
Slutsatser och diskussion 17
Framtida och fortsatt arbete 18
Referenser 20

Introduktion

Under de senaste åren har utvecklingen av transformerbaserade neurala nätverk revolutionerat bearbetningen av textdata (Vaswani et al., 2017), både genom encoderbaserade språkmodeller (Devlin et al., 2019) och generativa/decoderbaserade språkmodeller (Brown et al., 2020). Encoderbaserade språkmodeller, ibland kallade förtränade språkmodeller, har visat sig vara särskilt effektiva för textklassificering och har redan börjat förändra hur vi analyserar stora textmängder (Bonikowski et al., 2022; Do et al., 2024). I Sverige är delar av den officiella statistiken baserad på textuella data, exempelvis beskrivningar av brott i anmälningar eller domslut, yrken i yrkesstatistiken, eller dödsorsaker. Trots detta har textanalys inom statistikproduktion varit en utmaning, främst på grund av de stora mängder data och den komplexitet som textdata medför. Traditionellt har statistik från textuella källor därför ofta producerats genom manuell klassificering, vilket är resurskrävande och begränsar möjligheterna till mer detaljerad och kontinuerlig statistikproduktion (Tillé et al., 2022).

Detta forskningsprojekt har fokuserat på att undersöka möjligheten att använda svenska språkmodeller, framförallt encoderbaserade modeller (e.g. Malmsten et. al, 2020), för att effektivisera och förbättra produktionen av statistik från textuella data. Det initiala tillämpningsområdet i projektet var officiell statistik, då denna ofta har de högsta kraven på korrekthet och träffsäkerhet. Genom att kombinera moderna neurala nätverk och statistisk urvalsteori kan vi studera hur maskininlärning och AI kan användas för att effektivisera och förbättra statistiska estimat av central betydelse. Detta område, som är relativt nytt, kallas ofta för prediktionsförstärkt estimation (“prediction-powered inference”, Angelopoulos et al., 2023). Idén med denna estimationsstrategi är att använda prediktiva modeller, vanligtvis från maskininlärning, för att predicera egenskaper av intresse att studera. Dessa prediktioner kan sedan användas för statistiska ändamål, som att dra slutsatser från olika datamaterial och populationer.

I detta projekt var fokus på att använda moderna språkmodeller som prediktiva modeller för att skatta totaler och proportioner av intresse. Dessa storheter är centrala i många områden där stora mängder text behöver analyseras. Det huvudsakliga tillämpningsområdet var inom kriminalstatistiken och projektet har skett i nära samarbete med Brottsförebyggande rådet (Brå). Brå har identifierat ett samhälleligt behov av mer detaljerad statistik kring exempelvis hatbrott, bedrägeribrott och våld i nära relationer. Syftet i projektet har varit att studera hur avancerade språkmodeller kan användas för att producera sådan statistik mer effektivt och samtidigt med hög kvalitet. Med mer och detaljerad kriminalstatistik kan vi förbättra kunskapsunderlaget för brottsförebyggande arbete och rättspolitiska beslut.

Projektets frågeställningar

Projektets övergripande mål har varit att undersöka möjligheten att använda förtränade språkmodeller för att skatta ändliga populationer av textuell data. Detta har operationaliserats i följande frågeställningar:

1. Kan förtränade språkmodeller användas för att skatta ändliga populationer av textuell data? Är kvaliteten i dessa skattningar tillräcklig för att de ska kunna användas i officiell statistik?

2. Hur kan osäkerheten i dessa skattningar beräknas korrekt? Officiell statistik kräver hög precision och korrekt osäkerhetsuppskattning, men neurala nätverk saknar ofta inbyggda metoder för att skatta osäkerhet (Gawlikowski et al., 2023). Är det möjligt att hantera dessa problem?

3. Vilka risker för social bias finns i modellerna? Tidigare forskning har visat att språkmodeller kan förstärka strukturell ojämlikhet genom att fånga sociala mönster i träningsdata (e.g. Bolukbasi et al., 2016). Påverkar detta skattningarna av kriminalstatistiska mått och kan vi utveckla metoder för att minimera dessa problem?

Genom att adressera dessa frågor har projektet bidragit både till den teoretiska utvecklingen inom statistik och maskininlärning samt till den praktiska tillämpningen av neurala språkmodeller inom kriminalstatistik.

Arbetets utförande mellan 2022-2025

Projektet påbörjades under hösten 2022. Materialet som användes i projektet är beskrivande text från samtliga polisanmälningar i Sverige under åren 2007-2022. Då detta material är känsligt, krävdes en hel del administrativa insatser för att få tillgång till relevant data. Detta inkluderade en utförlig etikprövning som godkändes under 2023 (se bilaga 1). Därefter genomfördes en omfattande dialog med Brottsförebyggande rådets chefsjurist kring vilka data som kunde lämnas ut och hur denna utlämningen kunde ske. Dessutom behövdes också en hel del arbete för att praktiskt få ut all relevant data från Brås system, vilket tog flera månader. Till sist behövde också Brå införskaffa kraftfulla datorer med moderna grafikkort för träning av språkmodeller, då detta var första gången dessa typer av större språkmodeller tränades hos Brå. I slutet av 2023 fanns all data på plats och projektets forskningsfas påbörjades.

När projektet påbörjades under 2022 var transformerbaserade neurala nätverk fortfarande främst ett forskningsområden inom akademin. Men under 2022 och 2023 ökade intresset explosionsartat i och med lanseringen av transformerbaserade generativa språkmodeller som ChatGPT (Zhao et al., 2023). Denna snabba utveckling har påverkat både akademisk och industriell användning av språkmodeller och skapat nya möjligheter och utmaningar, inte minst för detta projekt. Flera av projektets initiala

idéer publicerades oberoende av oss, av ledande forskningsgrupper vid Harvard, Berkey och Stanford (e.g. see Egami et al., 2023; Angelopuolos et al., 2023) vilket också påverkat projektets arbete. Forskningsframstegen innebar att projektet inriktning fick anpassas något för att möjliggöra vetenskaplig publicering. Framstegen inom forskningsfronten har dock varit positiva och motiverande för projektet då vi insett att vi arbetar med samma vetenskapliga frågeställningar som andra forskare vid världens främsta universitet.

Under 2023 togs den första vetenskapliga artikeln i projektet fram (se nedan samt bilaga 2). I detta delprojekt inleddes arbetet med att träna språkmodeller på svenska polisanmälningar och anpassa dessa för klassifikation på hatbrottsanmälningar. Redan under våren 2024 insåg vi att dessa modeller var så pass bra att de var bättre eller på samma nivå som polisens manuella bedömningar. Vi insåg då att detta mycket väl kan vara av intresse att använda, både vid Brå men även hos Polisen. Det gjorde att vi under hösten 2024 påbörjade ett arbete med att analysera dessa modeller enligt frågeställning 3 för att se vad det kan finnas för potentiella etiska och juridiska begränsningar av att använda dessa typer av modeller både i statistikproduktion hos Brå samt hos Polisen i det fall de skulle vilja använda denna klassificerare. Detta gjordes i samarbete med Holli Sargeant, doktor i juridik vid Cambridge Universitet, som specialiserat sig på etiska och juridiska aspekter vid automatiserat beslutsfattande och användandet av AI.

Under hösten 2024 presenterade Hannes Waldetoft, doktorand i projektet, presenterat resultaten från den första vetenskapliga artikeln vid ett flertal institutioner. Dels har arbetet presenterats i vid ett s.k. “review”-seminarium där arbetet granskades av en extern granskare, Professor Dan Hedlin vid Stockholms universitet. Därefter har arbetet presenterats på seminarier vid Stockholms universitet, Umeå universitet, Linköpings universitet och för Demoskop som visat intresse för projektet. Likaså har projektet presenterats vid Länsförsäkringars Trygghetsdagar under våren 2024.

Parallellt under hösten 2024 påbörjades även arbetet med frågor om hur aktiv inlärning kan användas för att ytterligare effektivisera manuell annotering. I ett första steg tittade vi på Bayesianska metoder för att effektivisera uppdatering av Bayesianska modeller inom aktiv inlärning. Arbetet är fortsatt pågående och beräknas bli klart senare under 2025.

Rapportens omfattning och upplägg

Denna slutrapport sammanfattar de huvudsakliga resultaten från forskningsprojektet och dess vetenskapliga bidrag. Forskningsprojektet har resulterat i utkast till två vetenskapliga artiklar. Artikel 1 i bilaga 2 kommer i närtid att skickas in till antingen Journal of the American Statistical Association eller The Journal of the Royal Statistical Society, båda starka tidskrifter i statistik. Artikel 2 är i en pågående bedömningsprocess för konferensen ACM Conference on Fairness, Accountability, and Transparency, den internationellt främsta konferensen för ansvarsfull användning av AI. Artiklarna har

också båda granskats vid statistiska institutionens vetenskapliga seminarier med en extern granskare.

1. **Prediktions-förstärkt statistisk estimation med hjälp av transformermodeller**
   Den första artikeln fokuserar på hur förtränade språkmodeller kan kombineras med statistiska skattningsmetoder för att förbättra skattningar av ändliga populationsparametrar. Särskilt fokus läggs på situationer med skeva populationer, alltså där variabeln av intresse endast förekommer sällan, där vanliga slumpmässiga urval inte fungerar särskilt väl. Tillämpningsområdet är skattning av hatbrott i Sverige.

2. **Utvärdering av klassificeringsmodellen samt etiska och juridiska aspekter**
   Den andra artikeln analyserar den praktiska tillämpningen av den utvecklade klassificeringsmodellen och hur den kan användas. I samarbete med Dr. Sargeant från juridiska fakulteten Cambridge har vi studerat de etiska och juridiska aspekterna av att använda dessa modeller inom kriminalstatistik eller inom rättsväsendet.

Utöver dessa två färdigställda vetenskapliga artiklar pågår också arbete med en artikel om Bayesiansk aktiv inlärning som kommer färdigställas under 2025.

Resultat

Här presenteras populärvetenskapligt de huvudsakliga resultaten i de två artiklar som producerats (se bilaga 2 och 3). För tekniska detaljer och referenser, vänligen se de vetenskapliga artiklarna i bilagorna.

Artikel 1: "Prediction-powered estimators for finite population statistics in highly imbalanced textual data: public hate crime estimation"

Att uppskatta förekomsten av hatbrott är en central uppgift inom kriminalstatistiken. Idag görs detta genom att Polisen markerar brott de bedömer har hatbrottmotiv och att dessa bedömningar sedan manuellt granskas av experter vid Brå. Manuell klassificering av polisanmälningar är resurskrävande och begränsar möjligheterna att skapa en heltäckande och aktuell statistik. I denna artikel undersöker vi hur transformerbaserade språkmodeller i kombination med klassisk statistisk urvalsteori kan effektivisera och förbättra skattningar av hatbrott i Sverige. Genom att använda en prediktionsförstärkt metod minskar behovet av omfattande manuell annotering och samtidigt förbättras skattningarnas precision och effektivitet.

Metod

Studien kombinerar djupinlärningsmodeller med traditionella statistiska urvalsmetoder för att uppskatta det totala antalet hatbrott i Sverige. Modellen tränades på tidigare annoterade polisanmälningar där hatbrottsmotiv har identifierats av experter vid Brottsförebyggande rådet (Brå). Genom att använda en transformerbaserad språkmodell, finjusterad på svenska polisrapporter, kan man predicera sannolikheten för att en given anmälan innehåller ett hatbrottsmotiv.

Tre huvudsakliga urvalsmetoder testas:

1. Prediktionsförstärkt Hansen-Hurwitz-estimator (H2P2) – en estimator där prediktioner från modellen används som hjälpinformation för att skapa ett effektivt slumpmässigt urval.
2. Stratifikation-efter-prediktion – där polisanmälningarna delas in i grupper baserade på modellens klassifikation i hatbrottsmotiv eller inte. I det strata där modellen bedömmer att hatbrottsmotiv inte finns undersöks två typer av estimation:
    a. Vanligt obundet slumpmässigt urval (OSU)
    b. OSU med differensskattning – en metod som utnyttjar information från prediktionsmodellen för att justera skattningarna.

Resultat

De empiriska resultaten visar att den utvecklade modellen har en mycket hög träffsäkerhet och kan identifiera hatbrott med en precision som är jämförbar eller till och med bättre än Polisens egen bedömning. Den prediktionsbaserade metoden möjliggör också en betydligt effektivare uppskattning av hatbrottens totala antal jämfört

med ett vanligt obundet slumpmässigt urval (OSU). H2P2-estimatorn, som presenteras i artikeln, minskar det nödvändiga urvalsstorleken med upp till 97 % jämfört med att uppnå samma precision med ett vanligt OSU, vilket innebär att färre rapporter behöver granskas manuellt. Med ett så pass litet urval som bara 200 anmälningar kan antalet hatbrott i Sverige skattas väntevärdesriktigt till 6051 med ett osäkerhetsintervall på cirka 5000-7000 under 2022. Detta är högre än de hatbrott som identifierades genom den manuella annoteringsprocessen vid polisen och Brå. Studien finner således också att polisens egen flaggning av hatbrottsmotiv missar flera hatbrott.

Diskussion och slutsats

Studien visar tydligt att kombinationen av maskininlärning och statistiska urvalsmetoder kan ge mer precisa och resurseffektiva skattningar av hatbrott, och i princip även inom många andra liknande områden inom kriminalstatistiken eller andra områden. Genom att använda en prediktionsförstärkt strategi är det möjligt att minska den manuella arbetsbördan och samtidigt säkerställa hög kvalitet i de skattningar som görs. Dock finns vissa begränsningar. Modellen är exempelvis sämre på att identifiera mer subtila hatbrott, särskilt de som saknar explicita nyckelord.

Resultaten utgör ett viktigt steg mot att kombinera statistisk estimation och moderna maskininlärningsmodeller, här exemplifierat i att förbättra den svenska hatbrottsstatistiken. Genom att utnyttja AI och maskininlärning kan skattningar göras snabbare, mer effektivt och med högre precision och kvalitet. Studien öppnar också för vidare tillämpningar inom andra områden där kriminalstatistik baserad på textuella källor är avgörande, såsom bedrägeribrott och våld i nära relationer.

2025-03-11

Artikel 2: “Legal and Ethical Implications of Transformer-Assisted Hate Crime Classification and Estimation in Sweden”

Hatbrott är brott som begås med ett hatbrottsmotiv gentemot specifika, ofta utsatta, grupper. Dessa brott påverkar inte bara de direkta offren utan riskerar också att skapa otrygghet och minska förtroendet för rättsväsendet. En utmaning i arbetet mot hatbrott, både när det gäller att uppskatta dess omfattning och vid polisens hantering av enskilda anmälningar, är att korrekt identifiera och klassificera en given polisanmälan som ett hatbrottsmotiverat. Det finns flera utmaningar vid klassifikation av hatbrottsanmälningar, såsom underrapportering och avsaknad av information och juridiskt komplexa motivbilder. Sammantaget gör detta att det är utmanande att klassificera hatbrott.

Denna studie analyserar hur en transformerbaserad språkmodell, som utvecklades i den första vetenskapliga artikeln, kan användas för klassificering av hatbrott, och då särskilt vilka problem och utmaningar som finns med att använda den föreslagna och utvecklade språkmodellen. Studien undersöker därför även de juridiska och etiska konsekvenserna av att använda en sådan modell inom rättsväsendet, både om den skulle användas av Polisen eller om den skulle användas av Brå i produktion av hatbrottsstatistik.

Hatbrottsbedömningar i Sverige

Sverige har sedan 1990-talet prioriterat kampen mot hatbrott. Trots detta kvarstår stora utmaningar, och Sverige har vid flera tillfällen fått internationell kritik för sin hantering av dessa brott. Hatbrott är inte ett brott i sig i den svenska lagstiftningen. Istället finns en straffskärpningsregel som gäller för brott där motivet är att kränka en person eller grupp baserat på exempelvis etnicitet, religion, sexuell läggning eller könsidentitet.

Polismyndigheten och Brå har olika metoder för att identifiera och klassificera hatbrott. Polisen använder en “pop-up”-funktion i sitt anmälningssystem för att markera potentiella hatbrott, men tidigare studier har visat att deras klassificeringar kan vara inkonsekventa eller felaktiga. Samtidigt finns det ett stort behov av att producera löpande statistik över hatbrott för att kunna följa utvecklingen över tid. För att hantera felaktiga bedömningar vid Polisen gör Brå en djupare analys av de polisanmälningar som Polisen bedömt vara hatbrott och gör en egen bedömning av de texter som finns i dessa. Detta har inneburit att Brå under senare år har omklassificerat nästan hälften av de brott som polisen initialt bedömt som hatbrott som ej hatbrott.

Språkmodell för klassificering av hatbrott

I denna artikel görs ett fördjupat arbete av en transformerbaserad språkmodell som är tränad på svenska polisanmälningar. Modellen tränades för att identifiera hatbrottsmotiv genom att analysera språkbruk, kontext och mönster i polisanmälningar.

Resultaten visar att språkmodellen identifierar hatbrott mer konsekvent än Polisen. Till exempel klassificerar modellen en större andel hatbrott korrekt, särskilt i de fall där hatfullt språk används. I dessa fall kan Polisen ibland missa eller felbedöma polisanmälningarna. Däremot har modellen svårt att identifiera mer subtila hatmotiv, exempelvis i brott mot samer där kontext och kulturella referenser spelar stor roll samtidigt som antalet observerade hatbrott är få. Likaså har både Polisen och språkmodellen svårt vid juridiskt komplexa fall, som vid de koranbränningar som skedde under 2022. Beroende på hur en sådan koranbränning sker, kan det antingen falla in under hatbrottmotiv eller inte, vilket beror på detaljer kring själva koranbränningen. I fall likt detta är det uppenbart att det krävs bedömningar av experterna på Brå för att en korrekt bedömning ska kunna göras.

Systematiska fel i klassifikation av hatbrott

Artikeln identifierade flera systematiska fel i både Polisens och modellens felklassificeringar. Polisen underrapporterar vissa hatbrott, särskilt där motivet inte är uppenbart. Modellen å sin sida överrapporterar vissa fall, exempelvis vid förekomst av specifika ord som den associerar med hatbrott, men där kontexten inte stödjer det. Detta kan vara polisanmälningar som innehåller ord som “dumskalle” då modellen tränats på många fall av anmälningar med ord som “svartskalle” och modellens språkstruktur fångar upp delordet “skalle” som relevant. Vissa grupper, som samer och transpersoner, underidentifieras av modellen på grund av begränsad träningsdata. I fallet när det gäller antisamiska hatbrott kan svårigheterna också bero på en mer specifik kontext i dessa hatbrott. Språkmodellen kan således förbättra hatbrottsklassificeringen men kräver fortsatt utveckling för att hantera dessa svårigheter vid en fullskalig implementation.

Juridiska och etiska implikationer

Användningen av språkmodeller inom rättsväsendet väcker juridiska och etiska frågor vilka analyseras som en del av studien. Det finns två typer av användningsområden där en språkmodell skulle kunna användas, antingen endast för produktion av statistik, eller som systemstöd för polisen. Artikeln sammanfattar dessa etiska och juridiska utmaningarna i följande delar:

1. *Datasäkerhet och integritet*: GDPR och svensk lag kräver strikt hantering av personuppgifter, vilket innebär att endast behöriga myndigheter får använda språkmodellen och den data som behövs för träning.
2. *Rättssäkerhet och icke-diskriminering*: Om modellen systematiskt missgynnar vissa grupper kan den bryta mot principer om rättvisa och likabehandling. Här kan exempelvis det faktum att modellen har svårare att bedöma samiska hatbrott innebära ett problem om detta inte adresseras genom ytterligare utveckling av modellen.
3. *Transparens och förklarbarhet*: Prediktiva modeller måste vara begripliga och möjliga att granska för att säkerställa att beslut kan motiveras och ifrågasättas vid behov. Språkmodeller är inte särskilt transparenta på grund av det stora antalet parametrar som används i modellen.

4. *Automationsbias*: Människor tenderar att lita på automatiserade beslut även när de är felaktiga, vilket kan leda till felaktig bedömning om modellen används slarvigt.

Sammanfattningsvis är det möjligt att konstatera att användningen av den utvecklade språkmodellen är mindre problematisk när det gäller produktion av hatbrottsstatistik eller i statistiksammanhang än om den skulle användas som systemstöd inom Polisen.

Slutsatser och rekommendationer

Den utvecklade språkmodellen har potential att förbättra Sveriges hatbrottsstatistik och minska Brås arbetsbörda genom att identifiera fler faktiska hatbrott och filtrera bort felaktiga bedömingar. Samtidigt krävs försiktighet vid implementering för att undvika rättssäkerhetsproblem, särskilt om den skulle användas utanför statistiksammanhang. Språkmodeller för automatisk klassifikation kan spela en viktig roll i hatbrottsklassificering, men noggrann tillsyn och ansvarsfull implementering är avgörande för att tekniken ska användas på ett etiskt sätt.

2025-03-11

Påbörjad artikel 3: Bayesian Active Learning using Pareto Smoothed Importance Sampling

Denna artikel är påbörjad, men ännu inte avslutad. Den är ett första steg i arbetet med att använda aktiv inlärning för effektiv träning av (probabilistiska) prediktiva modeller. Nedan sammanfattas de nuvarande resultaten i artikeln. Detta är dock preliminära resultat och är inte ännu vetenskapligt granskat. Därför kan flera delar komma att ändras i den slutliga artikeln. Artikeln fokuserar på enklare modeller, som ett första steg för effektiv aktiv inlärning med hjälp av språkmodeller.

Bayesiansk aktiv inlärning

Artikeln presenterar en ny metod för Bayesian Active Learning (BAL) där en modell uppdateras sekventiellt men en ny observation i taget genom uppdatering av modellen med s.k. Pareto Smoothed Importance Sampling (Vehtari et al., 2024, PSIS). Syftet med metoden PSIS-BAL är att effektivisera urvalet av de mest informativa observationerna utan att behöva träna om hela modellen efter varje ny observation, vilket annars kan vara kostsamt i beräkningstid av större modeller, ett problem också för stora neurala nätverk.

Metod

Metoden bygger på att använda PSIS för att vikta om posteriorfördelningar med varje ny observation, istället för att träna om hela modellen. I artikeln tillämpas metoden på ett enkelt datamaterial med miljögifter i sik (Coregonus sp.) från svenska vatten, där toxinnivåer relateras till fettinnehåll och fisklängd. Valet av empiriskt datamaterial är för att primärt studera egenskaperna i metoden för enklare problem, innan mer komplexa data, som textuella dokument, kan hanteras.

Två enkla prediktiva modeller studeras:

1. En Bayesiansk linjär regression med toxinnivåer som beroende variabel och fettinnehåll som förklarande variabel.
2. En utökad modell som inkluderar även fiskens längd som en extra prediktor i modellen.

Båda modellerna tränas initialt på en slumpmässig delmängd (20%) av datamaterialet och uppdateras sedan iterativt genom att BAL identifierar vilken nästa observation som ska inkluderas i datamaterialet. Därefter viktas modellen om baserat på den nya observationen med hjälp av PSIS och standarddiagnostik (Yao et al., 2018) för att studera om omvägningen är numeriskt stabil eller om modellen behöver skattas om. Syftet är att studera effektiviteten i att uppdatera modellen efter varje enskild observation jämfört med slumpmässigt urval samt att undersöka möjligheten till en effektiv uppdatering av modellen efter varje ny observation.

Resultat

De första initiala resultaten visar att PSIS-BAL väljer observationer mer informativt och resulterar i en mer koncentrerad posteriordistribution jämfört med slumpmässigt urval eller urval då istället observationerna väljs flera i taget. Posteriorosäkerheten för regressionskoefficienterna minskar kontinuerligt vid varje iteration, vilket visar att metoden korrekt identifierar de mest informativa observationerna. Samtidigt sker bara omvägning av posteriorfördelningen med PSIS i varje steg, vilket är väldigt effektivt.

Efter varje ny observation undersöks att omvägningen är numeriskt stabil efter att en ny observation har lagts till i materialet. Generellt sett är PSIS väldigt stabil, dock visade diagnostik att när en mer komplex modell används behövdes modellen skattas om ett fåtal tillfällen.

Slutsatser och fortsatt arbete

De preliminära resultaten i studien visar att PSIS-BAL är en lovande metod för aktiv inlärning med Bayesianska modeller, då den minskar behovet av att träna om modellen för varje ny observation. Den väljer också observationer mer effektivt än slumpmässiga urval eller när flera observationer väljs i taget. Generellt sett fungerar PSIS väldigt bra för aktiv inlärning i dessa enklare modeller och enbart ett fåtal gånger behöver modellen skattas om för den utökade modellen. I det vidare arbetet kommer fokus ligga på mer komplexa modeller, som neurala nätverk, och större dataset. Dessutom kommer arbete ske för en utveckling av mer avancerade urvalsstrategier som bättre tar hänsyn till olika typer av osäkerhet.

Syftet med detta arbete är att effektivt kunna identifiera vilka observationer som är de bästa för att effektivt träna större maskininlärningsmodeller. I fallet med hatbrott fanns nästan 100 000 observationer modeller kunde tränas på, men i andra områden är antalet observationer mycket färre. Ett exempel på detta är när det finns behov av ny statistik, exempelvis kring detaljer vid våld i nära relationer eller avseende detaljer i bedrägeribrott.

Övriga resultat: Godkänd etikprövning och forskningsmiljö för AI-forskning vid Brå

Utöver de vetenskapliga arbetena som producerats i projektet har det administrativa och juridiska förarbetet inneburit ytterligare leverabler. Arbetet med etikprövning, datatillgång och juridiska processer vid Brå har resulterat i att det nu finns en mindre forskningsmiljö för AI-forskning vid Brottsförebyggande rådet. Etableringen av den här forskningsmiljön lägger en god grund för fortsatt arbete efter projektets slut och nya projekt inom området maskininlärning för förbättring av rättsstatistiken.

Betydelse för Länsförsäkringar

Försäkringsbolag som Länsförsäkringar hanterar dagligen stora mängder ostrukturerad textuell data, särskilt inom sakförsäkringsärenden där skadeanmälningar innehåller detaljerad information som är avgörande för bedömningen av försäkringsfall. Ett exempel på detta är hanteringen av bilskador, där varje skadeanmälan innehåller beskrivningar av händelseförlopp, typ av skador, och eventuella omständigheter kring olyckan.

Traditionellt innebär bearbetning av dessa data för statistikändamål en omfattande manuell insats, vilket är tidskrävande, kostsamt och beror på enskilda handläggares bedömningar. Genom att använda förtränade språkmodeller, såsom de som studerats i detta forskningsprojekt, kan Länsförsäkringar och andra försäkringsbolag effektivisera processen avsevärt.

Metoden med prediktionsförstärkt estimation innebär att om det finns träningsdata kan modeller automatiskt användas för att klassificera skadeärenden och identifiera relevanta attribut i anmälningarna. Exempelvis kan man snabbt identifiera ärenden med särskilda riskfaktorer eller ovanliga skadeorsaker, såsom specifika skadetyper eller misstänkta bedrägerier. Genom att kombinera dessa modeller med mindre urval och prediktionsförstärkta estimatorer är det då också möjligt att snabbt och effektivt uppskatta volymer av ex. misstänkta bedrägerier.

Automatisk klassifikation, exempelvis av enklare ärenden, kan frigöra värdefull tid för skadehandläggarna att fokusera på komplexa fall som verkligen kräver mänsklig bedömning samtidigt som det också är möjligt att uppskatta omfattningen baserat på den textuell data som finns tillgängligt. För försäkringsbolag som Länsförsäkringar innebär detta även förbättrade möjligheter att samla in och analysera statistik över skadeärenden. Sådan statistik kan användas strategiskt för riskhantering, produktutveckling, och för att identifiera trender och mönster som kan vara avgörande för skadeförebyggande insatser och policyutveckling.

Sammanfattningsvis möjliggör kombinationen av förtränade språkmodeller och prediktionsstärkta estimatorer en kraftfull och effektiv metod för försäkringsbolag att effektivt använda textuella data i skadeanmälningar, vilket kan ge både operativa effektiviseringar och strategiska insikter.

Slutsatser och diskussion

Nedan besvaras forskningsprojektets tre övergripande frågeställningar. Dessa slutsatser följs sedan av hur detta arbete kan utvecklas vidare i olika riktningar.

Kan förtränade språkmodeller användas för att skatta ändliga populationer av textuell data?
Förtränade språkmodeller, såsom transformerbaserade neurala nätverk (exempelvis BERT och roBERTa), har visat sig vara effektiva verktyg vid estimation av ändliga populationer av textuell data. Dessa modeller kan kraftigt förbättra klassificeringen av textdokument och därigenom minska behovet av manuella annoteringar. Genom att kombinera dessa modeller med traditionella statistiska estimatorer, såsom Hansen-Hurwitz estimatorn (HH) och stratifierade urval, kan skattningarna göras mer effektiva och noggranna, särskilt i skeva populationer där oberoende slumpmässiga urval (OSU) är ineffektiva. Ett tillämpningsområde där detta visat sig effektivt är vid skattning av antalet hatbrott i Sverige. De tränade språkmodellerna har visat sig vara lika bra eller bättre än polisens manuella klassificeringar. Genom att kombinera dessa modeller med klassiska statistiska estimatorer är det möjligt att skapa prediktionsförstärkta estimatorer. Användningen av dessa metoder kan inte bara förbättra effektiviteten utan även möjliggöra en mer detaljerad statistikproduktion både gällande hatbrott och annan statistik som bygger på textuell data.

Hur kan osäkerheten i dessa skattningar beräknas?
En central utmaning med att använda språkmodeller för statistiska uppskattningar är att de ofta saknar inbyggda metoder för att kvantifiera osäkerheten i sina skattningar. Genom att använda prediktionsförstärkta designbaserade estimatorer, såsom H2P2 som presenteras i den första vetenskapliga artikeln, kan vi beräkna variansen baserat på klassisk urvalsteori och bidra till att skapa mer tillförlitliga konfidensintervall.

Vilka risker för social bias finns i modellerna?
Trots att språkmodeller kan förbättra skattningarna finns det även risker för den social bias som finns i dessa modeller också påverkar skattningar som baseras på dem. En av de största riskerna är att modellerna speglar och förstärker bias från träningsdata, vilket kan leda till systematiska fel i klassificeringen av vissa grupper eller brottstyper. Till exempel kan vissa typer av hatbrott, såsom de riktade mot samer eller transpersoner, vara underrepresenterade i träningsdata, vilket gör det svårare för modellen att korrekt identifiera dem. Samtidigt finns det en risk för överrapportering av hatbrott om modellen felaktigt associerar vissa ord eller fraser med denna kategori utan att ta hänsyn till kontexten.

För att hantera dessa risker krävs flera åtgärder. För det första behöver modellerna kontinuerligt valideras och justeras för att minska dessa och liknande systematiska fel. Det är också viktigt att inkludera mer representativ träningsdata för att säkerställa att underrepresenterade grupper inte missgynnas. Dock är dessa problem också ett starkt

argument för kombinationen av designbaserade estimatorer och språkmodeller genom prediktionsförstärkta estimatorer då dessa är just väntevärdesriktiga även om modellerna har skevheter.

Sammanfattningsvis visar resultaten i projektet att förtränade språkmodeller kan användas för att skatta ändliga populationer av textuell data på ett effektivt och träffsäkert sätt, särskilt genom att kombinera dem med statistiska metoder. Dock krävs korrekta statistiska metoder för att kvantifiera osäkerhet och bidra till korrekta skattningar som är användbara i praktiska och viktiga tillämpningar. Denna kombination kan då bidra till en mer effektiv och rättssäker produktion av kriminalstatistik, annan samhällsrelevant statistik eller utveckling av statistikproduktion inom exempelvis Länsförsäkringar.

Framtida och fortsatt arbete

Detta projekt har lagt grunden för den nödvändiga infrastrukturen för fortsatt forskning i gränslandet mellan AI, maskininlärning och statistiska tillämpningar inom rättsväsendet. Detta arbete inkluderar etikprövning, juridiska processer vid Brå, tillgång till data och tillgång till beräkningskraft vid Brå. Detta möjliggör fortsatta forskningsframsteg inom området efter projektets slut.

En viktig insikt från arbetet är att metoden med att kombinera prediktiva modeller och statistiska estimatorer fungerar särskilt bra när det finns riklig och högkvalitativ träningsdata, vilket har varit fallet vid klassificering av hatbrott. Samtidigt finns ett stort intresse från Brå att tillämpa dessa metoder på områden där mängden träningsdata är betydligt mindre, såsom våld i nära relationer. För att hantera detta problem är nästa logiska steg att arbeta mer med aktiv inlärning, där Bayesian Active Learning med Pareto Smoothed Importance Sampling (BAL-PSIS) är en möjlig väg framåt. I framtiden är tanken att kombinera Bayesiansk aktiv inlärning med transformerbaserade modeller, såsom BERT och RoBERTa, för att bättre hantera situationer där annoterade data är begränsade för att välja de mest informativa observationer för en given modell.

Vidare finns det mer teoretisk arbete att göra avseende prediktionsförstärkta estimatorer som utvecklats i detta projekt. Genom att studera estimatorernas egenskaper kan ytterligare förståelse skapas hur estimatorer och prediktiva modeller bäst kan kombineras.

En utmaning som Brå står inför är de tekniska begränsningarna i nuvarande beräkningskapacitet. Med nuvarande infrastruktur är det svårt att använda större språkmodeller som exempelvis GPT-baserade system. Dessa stora språkmodeller skulle kunna bidra till bättre kontextförståelse och hantering av mindre dataset, exempelvis gällande kontexten kring anti-samiska hatbrott. Eftersom känsliga data som polisanmälningar inte kan behandlas i externa molnlösningar, krävs investeringar i lokal beräkningskraft för att möjliggöra dessa experiment. Ett framtida forskningsområde kan

därmed vara att utvärdera huruvida mindre, mer specialiserade språkmodeller kan ge liknande resultat som större modeller utan att kräva lika mycket beräkningskraft. Alternativt kan det, särskilt efter AI-kommissionens rapport som lyfter vikten av myndigheters arbete med AI, bli möjligt att också rättsväsendet och Brå införskaffar bättre beräkningskraft för att möjliggöra mer kraftfulla språkmodeller.

Slutligen har preliminära teoretiska resultat visat att välkalibrerade modeller, d.v.s. modeller som korrekt uppskattar osäkerheten i sina prediktioner, tycks ha bättre statistiska egenskaper och ger mer tillförlitliga skattningar. Detta gör det naturligt att fortsätta undersöka hur kalibrering kan förbättras i transformerbaserade neurala nätverk, särskilt i relation till osäkerhetsmått och tillförlitlighet i beslutsfattande processer. En viktig forskningsfråga blir då hur man kan optimera träningsprocessen för att producera bättre kalibrerade språkmodeller för användning i statistiska tillämpningar.

Referenser

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. Science, 382(6671), 669-674.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in neural information processing systems, 29.

Bonikowski, B., Luo, Y., & Stuhler, O. (2022). Politics as usual? Measuring populism, nationalism, and authoritarianism in US presidential campaigns (1952–2020) with neural language models. Sociological Methods & Research, 51(4), 1721-1787.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Do, S., Ollion, É., & Shen, R. (2024). The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. Sociological Methods & Research, 53(3), 1167-1200.

Egami, N., Hinck, M., Stewart, B., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. Advances in Neural Information Processing Systems, 36, 68589-68601.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl 1), 1513-1589.

Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden--Making a Swedish BERT. arXiv preprint arXiv:2007.01658.

Ollion, E., Shen, R., Macanovic, A., & Chatelain, A. (2023). ChatGPT for text annotation? Mind the hype. SocArXiv preprint, 32.

Tillé, Y., Debusschere, M., Luomaranta, H., Axelson, M., Elvers, E., Holmberg, A., & Valliant, R. (2022). Some Thoughts on Official Statistics and its Future (with discussion). Journal of Official Statistics, 38(2), 557-598.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2024). Pareto smoothed importance sampling. Journal of Machine Learning Research, 25(72), 1-58.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018, July). Yes, but did it work?: Evaluating variational inference. In International Conference on Machine Learning (pp. 5581-5590). PMLR.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2).