1. Introduktion

At være borger i et demokrati betyder at tage aktiv del i det samfund, vi alle – høj som lav – er en del af. Selve demokratiets grundessens er samtalen. Som Hal Koch forklarer er det gennem mødet og samtalen at vi ''*formår at træffe en afgørelse, som ikke alene tjener en enkelt person eller klasse, men som tager billigt hensyn til helheden.*'' (Koch, 1945).

I opgaven analyserer vi indlæg, kronikker, ledere og artikler fra dagbladet Informations debatsektion. Herigennem identificerer vi tendenser og mønstre, der viser hvordan mænd og kvinder adskiller sig fra hinanden i den offentlige debat.

For at kvantificere hvor meget af spaltepladsen på informations debatsektion, som tilfalder henholdsvis mænd og kvinder og hvad denne bliver brugt til, har vi scrapet alt indhold fra Informations debatsektion fra 1998 til 2016, svarende til ca. 28.000 indlæg. Disse opdeler vi yderligere til ca. 500.000 paragraffer. Vi har efterfølgende læst og klassificeret et tilfældigt udpluk af paragrafferne. Herudfra har vi brugt hhv. logistisk regression og naïve bayes estimatorer til at klassificere hvilket emne de resterende paragraffer tilhører.

Med det forarbejdede data undersøger vi, om mænd og kvinder har lige meget spalteplads i Informations debatsektion. Herefter blotlægger vi hvilke emner hhv. mænd og kvinder vælger at dedikere deres spalteplads til. Til sidst fremlægger vi i hvilket omfang de to køn hver især formår at generere læserrespons på deres debatbidrag.

Vi finder at skønt halvdelen af Danmarks befolkning som bekendt er kvinder, fylder de iøjnefaldende lidt på Informations debatsider; under en fjerdedel af spaltepladsen i Informations debatsektion er således forfattet af kvinder. Vi finder videre, at mænd oftere end kvinder skriver om statsforvaltning, teknologi og internationale anliggender, hvorimod kvinderne bruger deres spalteplads på familie og identitet.

2 Kodning af data

Web Scraping

For at hente alle artikler fra Informations debatsektion, har vi lavet et script, der automatisk henter og arkiverer brødtekst fra alle ønskede artikler samt relevant meta-information, som forfatterens navn, overskriften og antallet af kommentarer.

Informations webarkiv er organiseret på en sådan måde, at man i URL'en kan indtaste et kronologisk sidenummer. Hvert sidenummer indeholder 10 debatartikler. For at hente alle artikler, har vi loopet over 3.810 sidenumre i arkivet og hentet de tilhørende 10 artikler i subsets af 100, der efterfølgende samles. Vi har valgt at benytte subsets for at forebygge mod uregelmæssigheder, der kunne udfordre stabiliteten af vores script. Scriptet indeholder en error-handler for at undgå kodefejl, men for at sikre os mod netværksproblemer og lignende benytter vi os af subsets til løbende at arkivere det hentede data. Scriptet findes i bilag 1.

Datasortering og tilfældige udpluk

Vi har i alt hentet et korpus på i alt ca. 40.000 artikler af forskellig karakter. Heraf sorterede vi ca. 12.000 artikler fra, da de enten manglede forfatternavn, eller var opsat på en måde, så det ikke umiddelbart var muligt at isolere afsenderens navn – og dermed køn.

Ved at sammenholde forfatterens navn med en liste over alle godkendte drenge- og pigenavne i Danmark (Ankestyrelsen 2017) bestemmer vi forfatterens køn. For kode se bilag B.4. Det har været nødvendigt manuelt at sortere blandt navne, der er godkendt til begge køn. Fx vil Maria i vores datasæt være en kvinde og Kim en mand. Derudover har vi tilføjet en række navne, der ikke findes i Ankestyrelsens lister.

Herfra opdeles artiklerne på paragraf-niveau ved at tokenize teksterne på linjeskift. Således konstrueres et korpus på knapt 500.000 individuelle paragraffer. Af disse har hvert gruppemedlem fået tildelt et tilfældigt udpluk på 1.100 paragraffer, hvor 100 paragraffer heraf er tilfældigt udplukket fra de øvrige gruppemedlemmers paragraffer. Dette giver et samlet tilfældigt udtryk på 4000 unikke paragraffer. Formålet af manuelt at klassificere paragraffernes emner. For kode se bilag B.5.

Manuel klassificering

De 4.000 læste paragraffer bliver nu kategoriseret som tilhørende et eller flere af 6 emner, for at danne et træningssæt til senere klassificering af samtlige paragraffer. De 6 emner er konceptualiseret således:

Tabel 2.1: Klassificering af labels

Overblik	
Labels	**Kontekst/underemner**
Internationale anliggender	Internationale anliggender, begivenheder og kontekst
Konflikter og konsekvenser	Flygtninge, Integration, Ulandshjælp, Krig og konflikter
Statsforvaltningen og teknologi	Retspolitik, Økonomi, Forvaltning, It og teknologi
Miljø & klima	Miljø, Klima og Økologi
Kultur, religion, historie og medier	Kultur, Religion, Historie og Medier (Konventionelle såvel som sociale)
Familie & identitet	Køn, Børn, Familie, Ligestilling, Sexualitet og Identitetspolitik

En uddybende kodebog findes i bilag B.3. Hvis en paragraf kodes som internationale anliggender, kodes den som minimum også som en af de andre kategorier. Fx hvis paragraffen handler om politik i EU, da er den både kategoriseret som internationale anliggender samt statsforvaltning og teknologi.

Udover de 6 hovedkategorier findes kategorien "andet", som dækker over paragraffer, der ikke passer i førnævnte 6 kategorier, samt kategorien "Ignorer", der benyttes til rester af tekst, som ikke blev frasorteret i vores script.

Hvis en paragraf ikke entydigt kan klassificeres under ét emne, kan den tilskrives flere emner. Hvis en paragraf fx handler om bederum i folkeskolen, vil den både klassificeres som kultur og religion og/eller familie og identitet. Dette vidner om, at det måske havde været fordelagtigt at reducere antallet af kategorier, ved at samle nogle emner i større grupper. Dog kan de nuværende kategorier

ikke tilfredsstillende håndtere emner som sundhed eller uddannelse, hvorfor flere kategorier i andre tilfælde måske havde forbedret interkoder reliabiliteten.

Interkoder reliabilitet

For at teste interkoder reliabiliteten, som er et udtryk for enigheden blandt de kodere, der har foretaget den manuelle klassificering, har vi udregnet Krippendorffs Alpha, som mål herfor. Alpha-parameteren får værdien 1, hvis gruppemedlemmerne har kodet paragrafferne identisk. Hvis paragrafferne blev klassificeret tilfældigt ville parameteren antage værdien 0. Parameteren kan antage negative værdier, hvis paragrafferne systematisk klassificeres forskelligt (Krippendorff 2011). For kode for udregning af Krippendorffs alpha se bilag B.7[1].

Tabel 2.2: Krippendorffs alpha

	Int.	Miljø	Kultur	Forvalt.	konfl	Familie
Alpha	0,626	0,631	0.423	0,495	0.527	0.463

Tabel 2.2 viser Krippendorffs Alpha fordelt på emner. Vi ser, at der har været stor grad af enighed i klassificeringen af paragraffer omhandlende internationale anliggender og miljø, mens der er større uenigheder i klassificeringen af de øvrige emner. Der er ikke videnskabeligt sat et minimum for en tilfredsstillende alpha-værdie, men Krippendorff (1980) foreslår dog, at alpha parameteren helst skal være over 0,67.

Hvis vi havde mere tid kunne vi finde korrelations koefficienten mellem de forskellige kodere, hermed kan man finde ud af, om der er en koder, som systematisk afviger fra de andre. Med denne information ville man være i stand til at gå tilbage og skærpe koderegler. Den lave interkoder reliabilitet er med til at reducere modellens evne til at klassificere nye paragraffer og artikler. En løsning kan være at lade det være en iterativ proces, hvor man evaluerer på alpha-værdier og gentager processen med manuel kodning for at skabe en bedre reliabilitet.

Klargøring af data

Før vores data skal gennem en classifier, skal vores korpus af paragraffer bearbejdes. Først splitter vi paragrafferne op i enkelte ord ved at benytte en tokenizer. Alle store bogstaver ændres til små og vi fjerner alle kommaer og reducerer alle ordene til deres stammer. Herved mister vi noget

[1] Koden til udregning af alpha-værdierne er fra GitHub-bruger 'Grrrr' (2017).

information fra paragrafferne, men kompleksiteten af vores data reduceres så markant, at det bliver væsentligt lettere at finde sammenhænge mellem teksten og bestemte emner. Dette er et trade-off vi, givet vores relativt få observationer, accepterer (Hopkins and King 2010, Quinn et al. 2009). Vi har ikke fjernet stopwords, men i stedet brugt en TFIDF vægtning, som vægter ord, der optræder ofte, mindre, end ord, der optræder sjældent (Foster et. al. 2017, 192-193). Intuitionen er, at de ord, der optræder ofte, fx 'at' eller 'og', sandsynligvis optræder på tværs af alle emner og indeholder derfor ikke information, der er relevant for vores klassificering.

3 Automatisk klassificering

Kort introduktion til tekstklassificering

Sprog er et kompliceret system, hvor mening overføres fra person til person gennem lyde og tegn. På skrift kalder vi det tekst, og hvad én person skriver, kan en anden læse. Når mængden af tekst overstiger kapaciteten for hvad et menneske kan nå at processere, må vi ty til en computer for at kunne klassificere og forstå det sprog, der ligger i teksten. Automatisk klassificering – eller natural language processing[2] – er den kunst, at omforme tekst til tal og derigennem afdække mønstre af mening i teksterne. Indeværende kapital forklarer de metodiske overvejelser, der ligger til grund for vores analyse.

Test og træning

For senere at kunne udføre en test af modellens evne til korrekt at foretage en out-of-sample-prediction opdeles vores datasæt med de manuelt kodede paragraffer i et tilfældigt udvalgt test- og et træningsdatasæt på hhv. 25 og 75 pct. af de kodede paragraffer. Vores fremgangsmåde er først at estimere en model på baggrund af data fra træningssættet og derefter teste dennes evne til at klassificere paragrafferne i testsættet. Denne proces gentages indtil en tilfredsstillende model opnås. En vigtig pointe ved at opdele data i et trænings- og testsæt før der foretages vektorisering er, at modellen kun genkender ord, der forekommer i træningssættet, og derfor ikke snyder på vægten ved klassificering af testsættet.

[2] Natural language processing (NLP) benyttes som begreb på forskellige måder. Med NLP mener vi det automatisk at udtrække mening i tekst data.

Bag-of-Words

For at vores model skal kunne forstå tekst som data, skal teksterne omdannes til en såkaldt bag-of-words. Denne proces består af to trin, fitting og transformering. Hertil har vi valgt at bruge en såkaldt 'count vectorizer', der konverterer vores tokens til en matrix. Matricens har en række for hver observation (paragraffer) og hvert unikt ord i paragraffen repræsenteres af en kolonne. Matricens celler repræsenterer, hvor ofte det enkelte ord optræder (sklearn doc. 2017). Dette skaber et højdimensionalt feature rum, hvilket kan være en udfordring for klassificeringen af vores labels, især ved et lavt antal observationer. Dette forhold omtales nedenfor i diskussion af klassifikationsmodellerne.

Denne metode har været standard inden for tekst klassificering de sidste par årtier og er således en veldokumenteret og gennemprøvet metode (Baroni et al. 2014). Alternativt kunne man benytte en 'predict vectorizer' (word2vec), der i stedet for blot at tælle ordene, forsøger at forudsige hvilken kontekst ordene optræder i. Denne metode præsterer generelt bedre end en count-vectorizer og kan være simplere at indstille (Baroni et al. 2014). Årsagen til at vi ikke bruger denne metode er, at vi har afgrænsede teoretisk funderede labels, og dels at det kræver en stor mængde data at træne en predict-vectorizer. Prediction vectorizere benyttes ofte til unsupervised learning for at udrede hvilke klynger af emner, der findes i et givent tekstkorpus. Med vores relativt begrænsede tekstmateriale samt få afgrænsede fast definerede labels vælger vi derfor at benytte en count-vectorizer til at konstruere vores bag-of-words.

Udvælgelse af modellen

Da vi undersøger den offentlige debats karakter, ligger undersøgelsens fokus ikke på at afdække sammenhænge mellem features og labels, men udelukkende på at optimere modellens evne til at klassificere teksterne korrekt. Modellens kvalitet vurderes på en out-of-sample-prediction frem for eksempelvis Pearsons r^2.

Til at klassificere data benyttes en statistisk model, der kan fortolke vores bag-of-words til en sandsynlighed for at teksten omhandler det givne emne. Da vores features består af tekst og dermed har er høj dimensionalt, vil naïve bayes være det naturlige modelvalg[3]. Men da vores labels

[3] En anden mulighed Support-Vector-Machines, hvilket ikke har fundet plads i indeværende analyse.

er binære variable, har vi valgt også at fitte data med en logistisk regression for derefter at sammenligne de to modeller.

Logistisk regression er en parametrisk estimator, der normalt bruges til problemer, hvor man ønsker at forstå hvordan en feature påvirker en label. Ulempen ved parametriske estimatorer er, at de ved mange uafhængige variable har en tendens til at overfitte (Hastie et al. 2017, 220-230). Overfittingen kommer af, at modellen opfanger for meget af den stokastiske støj, med andre ord 'memorerer' modellen data frem for at blotlægge de egentlige underliggende processer.

Når vi tester vores model på vores testsæt risikerer vi overfitting af modellen. Der er her et trade-off mellem på den ene at kunne forklare variationen i træningssættet og på den anden side forudsige labels i testsættet korrekt. Hvis modellen *overfitter*, er den i stand til at forklare meget af variationen i træningssættet, men præsterer dårligt i prædiktionen af labels i testsættet. Hvis modellen derimod *underfitter* vil den ikke opfange den underliggende uobserverede 'sande' model. Dette trade-off kaldes også et *variance-bias* trade-off. Årsagen til dette navn skal findes i, at en model, der overfitter, typisk vil have en høj varians men være middelret, mens det omvendte er tilfældet for en model, der underfitter, der således være biased men med lav varians.

For at være sikre på at få den optimale out-of-sample prediction har vi valgt at træne begge modeller på vores sample, for hvert af vores 6 labels. For at optimere præstationen af naïve bayes benytter vi en automatisk gridsearch funktion til at estimere de optimale hyperparametre til modellen for hver eneste label. Vores fokus er at maksimere modellens F1 score, da denne er et harmonisk gennemsnit mellem modellens recall, antallet af true positives divideret med antallet af faktiske positives, samt modellens precision, antallet af true positives divideret med antallet af estimerede positives (Foster et. Al. 2017). Vi bruger derfor F1-scoren til at optimer under det trade-off, der ligger mellem modelles recall og precision.

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \cdot Recall \cdot \ Precision}{Recall + Precision}$$

Naïve bayes er en non-parametrisk estimatior, der beregner sandsynligheden for, at en observation har en given label på baggrund af en række betingede sandsynligheder – givet den angivne label.

Det naive består i, at givet den valgte label antager modellen, at alle features er uafhængige. Modellen tenderer derfor mod et bias men præsterer ofte godt ved mange features, idet modellens simplicitet ofte giver en relativ lav varians (Hastie 2017, s. 211). Estimatoren tildeler derefter en label alt efter højeste sandsynlighed.

For at optimere præstationen for den logaritmiske regression, benyttede vi os af et automatisk gridsearch på modellens parametre. Vi fandt dog, at vi ved at foretage en manuel gridsearch på vægtningen af observationer og ikke observationer alene, fik en bedre F1-score end ved den automatiske gridsearch. Et alternativ til at bruge vægte er at undersample på ikke-observationerne (Foster et. al. 2017, s. 182-183). Problemet ved dette er, at vi kun har kategoriseret 4.000 ud af 500.000 paragraffer, svarende til 0.8 pct. For at beholde så meget information som muligt vælger vi at korrigere for ubalancen gennem vægtning i stedet.

Modellernes præstationer er overskueliggjort i Tabel A1, hvor det fremgår, at den bedste f1 score fås ved den logistiske regression for langt de fleste af vores modeller med undtagelse af emnerne: "kultur og religion" og "Ignorer". Her præsterer naïve bayes bedre. Målt på f1 scoren har vi efterfølgende valgt de modeller, der præsterede bedst for hvert label.

Vort initiale udgangspunkt for at vurdere modellen var et fokus på recall fremfor precision. Motivationen hertil var dels det ubalancerede datasæt og dels et ønske om at fange så mange true positives som muligt (Foster et al. 2017, s. 176). Konsekvensen af dette var dog en stor mængde false positives, da næsten alle paragraffer under dette optimeringsregime i praksis blev klassificeret som omhandlende størstedelen af alle emner. Hvis recall vægtes marginalt højere end precision, reduceres mulighederne for at skelne mellem kategorierne. I vores endelige model, der som nævnt baseres på F1-scoren, accepterer vi således at både precision og recall vægtes for dermed at sikre en balance i modellen (Foster et al, 178).

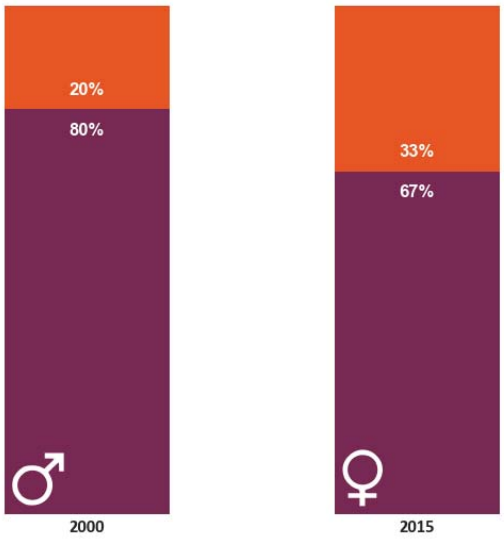
Tabel 3.1 viser fordelingen af paragraffer på emner og køn, samt antallet af paragraffer i alt og antallet af leder artikler efter den automatiske klassifikation.

Tabel 3.1 Paragraffer efter emner.

paragraffer	intern.	familie	ignorer	konflikter	kultur	miljø	andet	statsforv.	leder	male	female	all male	mixed	all female
473.792	136.511	32.997	29.964	81.725	38.910	10.599	74.289	105.271	47.661	333.879	103.694	16.674	12.881	6.664

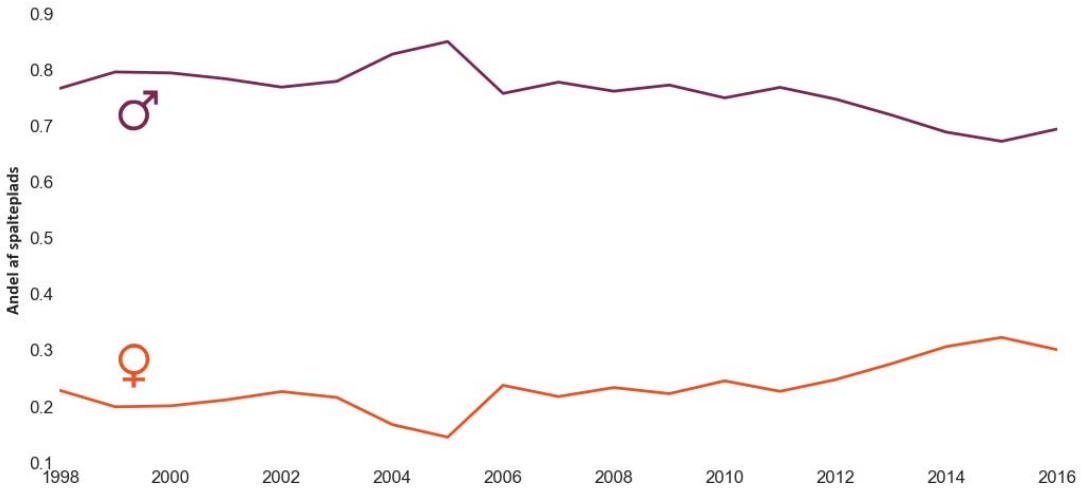
4 Resultater

Figur 4.1 Fordeling af forfatters køn



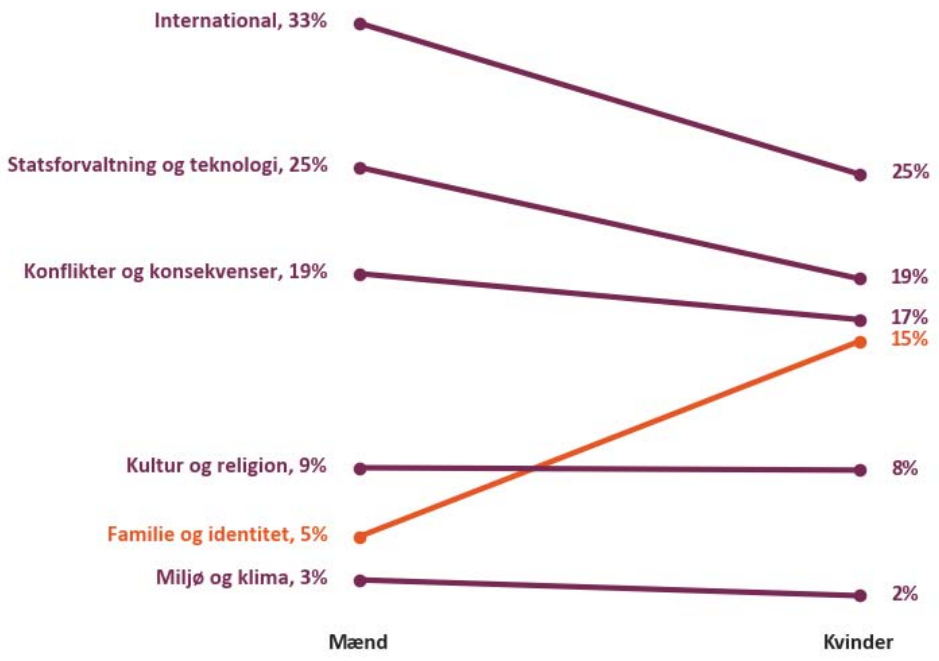
Halvdelen af Danmarks befolkning er kvinder, men hvad angår spalteplads på debatsiderne fylder de en iøjnefaldende lille andel. Under en fjerdedel af spaltepladsen i Informations debatsektion er forfattet af kvinder, jf. figur 4.1. Dog har der været en udvikling hen imod en mere lige fordeling og kvinder har særligt i de senere fået en større stemme i den offentlige debat, jf. figur 4.2.

Figur 4.2 Spalteplads efter forfatters køn



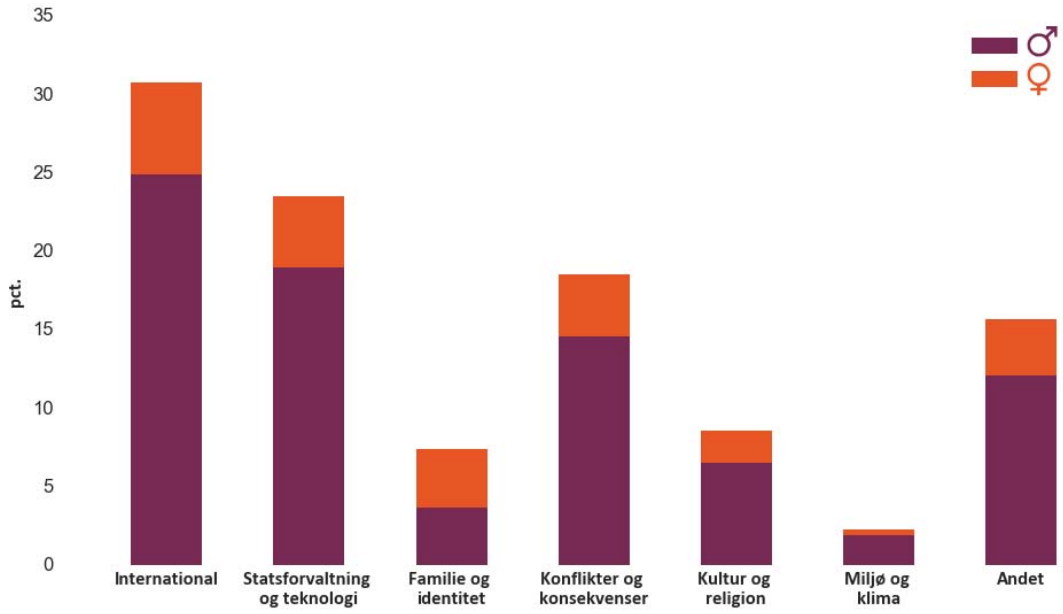
Når kvindernes indlæg fylder så lille en del af spalterne i debatsektionen, så risikere vi at miste vigtige nuancer af den fælles samtale, idet debatten bliver domineret af den ene halvdel, mænd! Kvinder og mænd beskæftiger sig med forskellige emner, et træk der tydeligt går igen i vores data. Figur 4.3 viser hvilke emner, der optager mænd og kvinder. Hvor mænd langt oftere end kvinder skriver om internationale anliggender, bruger kvinderne oftere deres spalteplads på at diskutere familie og identitet. Familie og identitet får således tre gange så meget spalteplads af kvindelige skribenter end tilfældet er for de mandlige.

Figur 4.3 Emner efter køn



På Informations debatsider optager internationale anliggender mange af skribenternes opmærksomhed. Ca. en tredjedel af alle indlæg havde en international vinkel, hvilket gør emnet det til hyppigste af alle – for både mænd og kvinder, jf. figur 4.4. Så selvom kvinderne går mere op i familie og identitet, så optager dette emne kun ca. 7 pct. af den samlede spalteplads. Når kvinderne altså er i så markant undertal, er konsekvensen altså den, at emner, der tiltaler mandlige skribenter, oftere komme til at styre den offentlige debat. Vi risikerer altså, at mænd overtager debatten og den demokratiske samtale kommer i ubalance.

Figur 4.4 Spalteplads efter køn og emne



Hvad angår den politiske debat i offentligheden så opstår der hermed et dobbeltsidet problem. For det første er kvinder generelt underrepræsenteret på Informations debatsider. For det andet skriver kvinder mindre om politiske emner end mænd. Resultatet bliver derfor at den politiske debat føres af mænd og hensynet til helheden kan derfor glemmes.

Formålet med et debatindlæg er at skabe debat og reaktioner, og vi har derfor valgt at kigge på antallet af kommentarer, som en proxy for læsernes interesse for en given artikel. Figur 4.5 viser hvor meget læserrespons artiklerne for fordelt efter køn på udvalgte emner. Vi definerer her læserrespons som gennemsnittet af det antal kommentarer, der knytter sig til den artikel, som paragraffen er taget fra.

Som figuren også viser, så er kvinder bedre end mænd til at generere interesse fra læserne, når indlægget handler om familie og identitet. Derimod forekommer indlæg forfattet af mænd at skabe mere læserinteresse, når debatten handler om politik eller miljøspørgsmål. Eksempelvis får kvinder en gennemsnitlig læserrespons på 15 indenfor familie og identitet, der hermed ligger over mændenes værdi på 13. Men når det drejer sig om politik ligger mændenes læserrespons på knap 10, mens kvindernes respons halter to under med en værdi på under 8.

Figur 4.5 Læserrespons efter køn og emne

I udgangspunktet vil Information være interesseret i at skabe debat og reaktioner på indholdet i deres debatsektion. Hvis dette var det eneste kriterie, som debatredaktionen udvælger indlæg på baggrund af, ville man forvente at mænd og kvinder i gennemsnit ville genere den samme mængde læserrespons indenfor et givent emne – uanset forfatterens køn. Men som vi beskriver ovenfor, er der tydelige forskelle på kønnenes læserrespons. Enten skyldes dette, at læserne reagerer anderledes på indholdet alt efter køn på forfatteren, ellers skyldes det, at debatredaktionen diskriminerer skribenter efter køn.

Hvis det sidste er tilfældet, vil der således være tale om positiv diskrimination af kvinder indenfor klima og miljø samt statsforvaltning og teknik. Den mulige diskrimination kan komme fra et ønske fra redaktionen om at diversificere debatten på bekostning af indlæggets evne til at generere læserrespons. Omvendt diskrimineres mænd positivt indenfor emnet familie og identitet, idet artikler med dårlige evne til at genere debat alligevel bliver udgivet.

Nedenstående citat fra debatredaktør på Politiken, Ditte Giese, bekræfter i nogen grad vores formodning om positiv diskrimination af kvinder på debatsiderne.

> *"Jeg sidder som jourhavende på Politikens debatredaktion i denne uge og må konstatere, at 90 procent af de indsendte indlæg kommer fra mænd. Jeg vil derfor gerne opfordre jer til at skrive noget og mene noget, meget gerne ud fra jeres faglighed, da langt de fleste indlæg, som vi modtager fra kvinder, handler om noget selvoplevet,"*
> (Schelde og Holm, 2016)

Om det samme gør sig gældende for debatredaktionen på Information, kan være vanskeligt at bekræfte.

5. Konklusion

Ovenstående analyse bygger på 28.000 artikler fra Informations debatsektion; 1998 til 2016. Heraf udtog vi et tilfældigt udpluk, som vi kategoriserede i 6 emner. Herefter benyttede vi krippendedorpfs alpha til at estimere interkoderreliabileten. Vi fandt at vores interkoder reliabilitet var i underkanten af det ønskede.

Efter rensning og preprocessing af data afprøvede vi to forskellige estimatorer, en logistisk regression og en naïve bayes. Vi fandt at den logistiske regression i de fleste tilfælde klarede sig bedre i out-of-sampel-prediction end naïve bayes, med undtagelserne af Kultur og religion samt kategorien 'Ignorer'. Her benyttede vi naïve bayes, og resten blev gennemført med den logistiske regression. Vores out-of-sample-predictions var af noget divergerende karakter, hvilke skyldes hhv. den tidligere nævnte interkoderreliabiltet men også at nogle af emner optræder meget sjældent. I tilfældet for de logistiske regressioner imødekom vi dette ved at vægte sjældne observationer højere.

Med de nu formulerede modeller klassificerede vi det fulde tekst korpus. På baggrund heraf fandt vi at kvinder fyldte væsentligt mindre i Informations debatsider end mændene, og kvinderne har i gennemsnit kun stået for omkring en fjerdedel. Vi fandt dog også at der især de sidste år ser ud til at være en tendens mod en udligning. Ikke desto mindre var det i 2016 stadigt kun en tredjedel af debat sektionens spalteplads der var besat af kvindelige forfattere.

Videre så vi, at hvor mændene især skriver om statsforvaltning, teknik og internationalt anliggende skriver kvinder i stedet om familie og identitet. I forlængelse heraf fandt vi, at der også var en systematisk forskel på hvor meget respons de to køn formåede at genere på tværs af udvalgte emner.

Det kunne have været en fordel, givet vores interkoder reliabilitet, at betragte opgaven som en iterativ process, hvor vi vendte tilbage til definitionen af labels og kodebogen, for at forbedre interkoder reliabiliteten. Alternativt kunne man komme uden om dette ved at forsøge med en unsupervised clustering.

Litteratur

Ankestyrelsen. 2017. Godkendte fornavne. Sidst tilgået 23. august 2017 https://ast.dk/born-familie/navne/navnelister/godkendte-fornavne

Grrrr. 2017. https://github.com/grrrr/krippendorff-alpha

Hopkins, Daniel J. and Gary King. 2010. "A method of automated nonparametric content analysis for social science". *American Journal of Political Science* 54(1): 229-247.

Krippendorff, K. 2011. "Computing Krippendorff's Alpha-Reliability". http://repository.upenn.edu/asc_papers/43/ Sidst tilgået 21. august 2017.

Krippendorff, K. 1980 Content Analysis: An introduction to its methodology, Beverly Hills CA: Sage, Chapter 12.

Schelde, Nanna og Ahrenkilde Holm, Thue. 2016. "Vanetænkning og frygt for chikane holder kvinder væk fra debatten.". Kristeligt Dagblad. https://www.kristeligt-dagblad.dk/danmark/vanetaenkning-og-frygt-chikane-holder-kvinder-vaek-fra-debatten

Koch, Hal. 1945. "Hvad er demokrati". Gyldendahl.

Quinn, K.M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2009. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1): 209–28.

Marco Baroni, Georgiana Dinu, Germán Kruszewski. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference* 1 (2014): 238-247.

Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds. 2017. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series. Boca Raton, FL: CRC Press Taylor & Francis Group.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer.

Appendiks A

Tabel A.1: Confusion Matrix samt f1 score, accuracy, recall og precision for logistisk regression.

	True Positive	False Positive	True Negative	False Negative	f1	accuracy	recall	precision
International	150	101	683	69	0,638	0,831	0,685	0,598
familie og identitet	42	16	836	109	0,930	0,875	0,981	0,885
Konflikter	68	61	809	65	0,519	0,874	0,511	0,527
Kultur og Religion	107	102	671	123	0,487	0,776	0,465	0,512
Miljø og Klima	10	9	965	19	0,417	0,972	0,345	0,526
Statforvaltning og Tech	143	33	590	237	0,814	0,731	0,947	0,713
Andet	54	116	785	48	0,397	0,836	0,529	0,318
Ignorer	69	24	878	32	0,711	0,944	0,683	0,742

Tabel A.2: Confusion Matrix samt f1 score, accuracy, recall og precision for Naïve Bayes.

	True Positive	False Positive	True Negative	False Negative	F1	Accuracy	Recall	Precision
International	157	147	637	62	0,600	0,792	0,717	0,516
Familie og id	14	14	838	137	0,917	0,849	0,984	0,859
Konflikter	84	158	712	49	0,448	0,794	0,632	0,347
Kultur og Religion	129	153	620	101	0,504	0,747	0,561	0,457
Miljø og Klima	16	102	872	13	0,218	0,885	0,552	0,136
Statsforvaltning og Tech	280	177	446	100	0,763	0,724	0,716	0,817
Andet	52	158	743	50	0,333	0,793	0,510	0,248
Ignorer	75	29	873	26	0,732	0,945	0,743	0,721

Figur A.1 Antal paragraffer efter år