Kjersti Haukaas

Automatisk kategorisering av nyhetsartikler fra to norske aviser

Masteroppgave
Avdeling for journalistikk, bibliotek og informasjonsfag

Sammendrag

I oppgaven har jeg kjørt forsøk på automatisk kategorisering av artikler fra de to norske avisene Aftenposten og Klassekampen. Jeg har valgt å bruke Support Vector Mashine (SVM) som bygger på teoriene fra maskinlæring. SVM er lært opp med artikler fra Aftenposten der kategorier er satt på manuelt. Det er å forvente at resultatet av kategoriseringen er bedre når artiklene som kategoriseres, er fra samme kilde som læringsdokumentene. Det ene forsøket mitt gikk derfor ut på å kategorier artikler fra en avis som ikke var brukt i opplæringen (Klassekampen) , og så sammenlikne resultatene herfra med resultatene fra kategorisering av Aftenposten. Forsøket viste et betraktelig dårligere resultat for Klassekampen. Dette var som forventet, det er flere faktorer som spilte inn og disse er diskutert i oppgaven.
Det andre forsøket jeg kjørte var for å se om kvaliteten på kategoriseringsmetoden ville holde seg over tid, eller om effekten reduseres etter som det blir større avstand mellom tidspunkt for opplæring og tidspunktet for kategoriseringen. Jeg forsøkte dette både for Aftenposten og Klassekampen. Tidsgapet var på tre måneder. For Aftenposten viste resultatene en liten nedgang i kvaliteten, dette var som forventet. For Klassekampen var resultatene betraktelig bedre etter tre måneder, men her virket nok et litt lite datagrunnlag inn på resultatet.

Masteroppgave ved Høgskolen i Oslo, Avdeling for journalistikk, bibliotek- og informasjonsfag
Oslo 2008

Forord

Jeg vil med dette rette en takk til alle som har bidratt til at oppgaven har blitt noe av. Spesielt en stor takk til Alf Endre Magnussen i Aftenposten som både har sendt meg tusenvis av artikler og som velvillig har svar på spørsmål og gitt meg tips om andre jeg kunne kontakte. Artiklene fra Klassekampen hentet jeg ut sjøl fra A-tekst, men også her har de velvillig svart på spørsmål, takk for det.

En arbeidskrevende del av oppgaven var å manuelt kategorisere artiklene fra Klassekampen. Til den jobben hadde jeg to gode hjelper, Mette Skeie og Tom Torfoss. De gikk på med friskt mot og nedla adskillige timer, en stor takk!

Til sist og ikke minst vil jeg takke veilederen min Ragnar Nordlie for mange god råd og mye støtte underveis.

Det hadde vært flott om noen ville følge opp forsøkene jeg har gjort, med nye forsøk. Både artikler og programmer ligger klart til videre bruk. Isåfall kan jeg kontaktes ved å sende epost til kjersti.haukaas@gmail.com

Kolbotn 18. juni 2008
Kjersti Haukaas

Innholdsfortegnelse

Tabelloversikt

Oversikt over figurer

1 Innledning

Digitaliseringen av tidligere trykt materiale foregår i stor skala i store deler av verden. Materiale som tidligere var forbeholdt et fåtall, blir nå tilgjengelig for alle enten via generelle søkemotorer eller via digitale tjenester fra bibliotek og dokumentsentre. Felles for mye av dette materialet er at det ikke er tilrettelagt for digital søking, det er ikke påført metadata som kunne være til hjelp både ved søking og ved presentasjon av søkeresultatene. Hvis en idag søker på innholdet i dokumentet, er dette ofte et fritekstsøk. Med store dokumentmengder kan det føre til mange treff med til dels dårlig presisjon. Søkemotorer som Google og Fast, bidrar til at søkeresultatene blir gode til tross for dette. Med bedre metadata i form av kategorier/emneord vil presisjonen kunne øke. I tillegg kan emneordene bidra til økt meningsinnhold på nettet ved at de inngår i emnekart der det opprettes assosiasjoner mellom emnene.

Til en viss grad påføres metadata på nytt materiale i dag, både av forfatterne sjøl og av informasjonspesialister. I tillegg er det blitt vanlig, men ikke veldig utbredt, at leserne sjøl kan påføre emneord, såkalt sosial tagging. Fordi mengden av dokumenter er enorm, er imidlertid manuell tildeling av emneord umulig. Dette gjelder både på den åpne verdensveven og på bedriftenes interne nettsteder. I 2002 ble det produsert mer enn 5 exabytes[1] med informasjon, dette tilsvarer 37 000 biblioteker på størrelse med Library of Congress (Lyman & Varian 2003). Økningen hvert år er formidabel, i følge Gartner Group øker det ustrukturerte innholdet innenfor en virksomhet med 93% hvert år (gjengitt etter IBM 2007).

I portaler på nettet der det skal søkes i dokumenter fra forskjellige kilder vanskeliggjøres fellessøk både ved at dokumentene mangler emneord og at det ofte mangler en felles emneordsliste. Dette gjelder for eksempel for nyhetsportaler. I forbindelse med utvikling av den portugisiske nyhetsportalen NewsSearch påpekte Nuno Maria og Mario J. Silvia (2000) viktigheten av en felles kategoriliste. Problemet var at det ikke fantes standarder for dette, med det resultat at hver enkelt avis brukte sin egen. Vi ser det samme i den norske nyhetsportalen Atekst, www.retriever-info.com Her er kun 9 av 35 aviser påført kategorier og ingen bruker samme liste. The International Press Telecommunications Council (IPTC), www.iptc.org har idag en internasjonal standard for metadata for nyheter, men denne foreligger ikke offisielt på norsk. NTB bruker en norsk versjon, og andre aviser, som Aftenposten, bruker en variant av denne. Å få en felles kategoriliste og å ta denne i bruk

1 $5*10^{18}$

for alle aviser, er et godt stykke fram. Gitt at ¾ av avisene ikke er tilordnet kategorier i det hele tatt, vil det å kunne tilordne kategoriene automatisk være en viktig faktor for å lykkes.

Automatisk tildeling av kategorier bygger på metodene fra maskinlæring. Kategoriseringsverktøyet læres opp på et sett av dokumenter som har emneord påført, og når verktøyet er opplært, brukes den for å tildele emneord på nye artikler. Utfordringen er at det ofte ikke eksisterer dokumenter med emneord påført i den dokumentmengde en ønsker å kategorisere. Verktøyet kan da læres opp med dokumenter fra en annen kilde, for eksempel kan begge være riksdekkende nyhetsaviser. Til tross for at de tilhører samme domene, kan kvaliteten bli dårligere enn når verktøyet læres fra samme kilde som senere skal kategoriseres. Faktorer som kan spille negativt inn, er for eksempel skribentenes forskjellig ordbruk og for aviser forskjellig redaksjonell vinkling av nyhetene, noe som kan påvirke både innhold og språkbruk. I denne oppgaven ser jeg nærmere på dette og prøver å få svar på følgene spørsmål:

- Vil kvaliteten av den automatisk kategorisering være like god på en dokumentsamling der kategoriseringsverktøyet er lært opp ved å bruke dokumenter fra en annen samling, som når dokumentene en bruker til læring kommer fra samme kilde som de som skal kategoriseres?

En dagsavis har nyheter som er gamle dagen etter at de er trykt, og over tid kan nyhetsbildet være helt annerledes. Hvem kunne på forhånd ane at Valla-saken ville komme opp som en stor sak med bred dekning i alle aviser over lang tid. Et annet eksempel er okkupasjonen og krigen i Midtøsten. Idag medfører situasjonen der at artikler som omhandler disse områdene, blir kategorisert under «Krig og konflikter». Når okkupasjonen en dag tar slutt og det ikke lenger er krig der, skal artiklene istedet kategorisert under dagligdagse emner som «Økonomi» og «Politikk».
Det kan være interessant å se om effekten av kategoriseringsmetoden blir dårligere over tid, og jeg vil derfor undersøke og prøve å få svar på følgende:

- Vil kvaliteten til kategoriseringsmetoden holde seg over tid, eller vil effekten reduseres etter som det blir større avstand mellom tidspunkt for opplæring og tidspunktet for kategoriseringen?

For å få svar på begge disse spørsmålene brukte jeg avisartikler fra to norske riksdekkende dagsaviser: Aftenposten og Klassekampen. Aftenposten er en av de forholdsvis få avisene som har påført kategorier/emneord idag, og den ville jeg derfor bruke til å lære opp verktøyet som skulle

brukes for kategorisering. Det var flere aviser jeg kunne valgt for å teste ut påstandene mine, kriteriet var at den var en riksdekkende generell nyhetsavis og at den ikke var alt for stor i omfang. Det siste tenkte jeg var viktig for lett å kunne få en oversikt over saksmengden i avisa. At akkurat Klassekampen ble valgt, var litt tilfeldig, men jeg har lest avisa i flere år og kjente godt til den.

I kapittel 2 gir jeg en beskrivelse av automatisk kategorisering og viktig forskning og litteratur innenfor fagområdet både internasjonalt og i Norge.
I kapittel 3 beskrives testkorpuset, dvs. avisene Aftenposten og Klassekampen, med fokus på det som kan ha betydning for forsøkene.
Kapittel 4 omhandler metodene som skal brukes i forsøkene og ved evaluering av resultater.
Forsøkene er beskrevet i kapittel 5, dette kapittelet er delt i fire underkapitler. I 5.1. beskrives forsøkene knyttet til opplæring og test av kategoriseringsverktøyet med artiklene fra Aftenposten. I 5.2 omhandles den manuelle kategoriseringen som ble gjort for Klassekampen. I 5.3 beskrives forsøkene for kategorisering av Klassekampen og resultatene sammenlikenes med testen gjort for Aftenposten i i 5.1. Forsøkene for å se om resultatene blir dårligere når kategoriseringen og læring er adskilt i tid beskrives i 5.4. Her er forsøkene gjort for både Aftenposten og Klassekampen.
I kapittel 6 vurderes resultatene av forsøkene og det gis forslag til videre arbeid.
Det er 4 vedlegg. I vedlegg A ligger en full oversikt over Aftenpostens kategorier og emner. Vedlegg B inneholder stoppordliste som er benyttet under forsøkene. I vedlegg C ligger tabeller og annen bakgrunnsinformasjon som har vært nødvendig for å komme fram til resultatene som presenteres. De viktigste programmer for oppbygging av matrise, vekting og sammenlikning av resultater ligger i vedlegg D.

2 Automatisk kategorisering av tekst

Tekstkategorisering går ut på å tilordne tekst i naturlig språk til en eller flere forhåndsdefinerte kategorier basert på tekstens innhold (Dumais, Hecherman, Platt & Sahami 1998). Det bygger på den antagelse at tekstens mening best representeres ved ordene i teksten og sammenhengen mellom ordene der. Så tidlig som på 1950 tallet, konkluderte H.P. Luhn (1961) med å bruke det naturlige språket da han foretok en vurdering av naturlige kontra kunstige språk for informasjonsgjenfinning. At dette så er blitt adoptert for kategorisering, er naturlig all den stund tekstkategorisering bygger på og bruker metodene utviklet for gjenfinning av informasjon. (Sebastiani 2002).

For å få gode resultater med tekstkategorisering er det viktig at kategoriene er så tydelig definert og adskilt fra hverandre som mulig. Her vil tekster som entydig omhandler en kategori, være bedre egnet enn tekster som omhandler flere emner. I informasjonsgjenfinning har det vært debatt om en får bedre resultater hvis en søker i deler av dokumentet. For 40 år siden konkluderte Lesk og Salton (1968) at større tekstmengder nok er bedre enn sammendrag, men at økningen i effektivitet ikke er stor nok til entydig å kunne trekke konklusjonen at full tekst er best. Med bruk spesielt av XML er mulighetene nå store for å strukturere dokumentene bedre. Om dette skal ha en positiv effekt på den automatiske kategoriseringen, fordrer det at teksten struktureres utfra innhold på en sånn måte at de enkelte deler av dokumentet mest mulig entydig omhandler et enkelt emne.

Så tidlig som på begynnelsen av sekstitallet begynte en å kategorisere maskinelt, men først 30 år etter fikk det noe særlig omfang (Sebastiani 2002). Da var datamaskinene blitt raske nok til å kunne håndtere de store datamengdene det her er snakk om. I dag er automatisk kategorisering av tekst et viktig emne innenfor informasjonsbehandling. Det tilbys kommersielt av flere store aktører, for eksempel har både Oracle(2005) og IBM(2008) dette som et produkt i sin portefølge.

I begynnelsen ble automatisk kategorisering foretatt ved at eksperter bygde opp regelsett som ble brukt i kategoriseringen. I en pilotversjon av det som senere skulle bli et system[2] for kategorisering av nyhetsartikler, baserte de seg på teknikker hentet fra prosessering av naturlige språk (Cellio, Hayes & Knecht 1988). De bygde opp regelsett hvor hver regel ble tildelt en vekt som anga hvor

2 Dette systemet er høyst sannsynlig det som i 1990 ble presentert som CONTRUE/TIS på « the Second Conference on Innovative Applications of Artificial Intelligence» http://www.aaai.org/Library/IAAI/iaai90contents.php

sannsynlig det er at regelen sa noe om innholdet i teksten. For eksempel betyr «probable» i regelen

(titanium) -> probable

=titanium

at en forekomst av ordet «titanium» betyr at teksten sannsynligvis omhandler dette. Mer kompliserte regler ble bygd for eksempel for å koble sammen eller utelate termer. En ulempe ved en regelbasert kategorisator er i følge Sebastiani (2002) at den er helt avhengig av eksperter, ikke bare i den initielle oppbyggingen, men også når nye kategorier kommer til senere og reglene da må endres. Til tross for dette er regelbaserte systemer i bruk også idag, en av de største kommersielle aktørene her er Autonomy med produktet IDOL, se http://www.autonomy.com/content/Products/IDOL Dette produktet brukes blant annet av den danske mediaportalen Infomedia (www.infomedia.dk)

Siden tidlig på 1990 tallet har metoder fra fagområdet maskinlæring i stor grad tatt over. Her er ideen at kategoriseringsmaskinen læres opp med et sett av dokumenter der kategoriene er påført av eksperter på forhånd. Senere kan så det opplærte verktøyet brukes til å tilordne riktig kategori til nye dokumenter. Det er flere metoder tilgjengelig og mange av verktøyene som bruker disse, er også gratis å bruke. En god oversikt over metoder, algoritmer og forskningen fram til år 2002 finner en hos Sebastiani (2002).

Etter 2002 har forskningen i stor grad gått ut på å optimalisere og å forbedre metodene og algoritmene. For eksempel organiseres de forhåndsdefinerte kategoriene i strukturer for å kunne ta i bruk den informasjonen som ligger implisitt i relasjonene i hierarki og nettverkstrukturer. Ching, Hao &Tu (2007) har beskrevet et forsøk der kategoriene organiseres i hierarkier vha klustering og kategoriseringen så skjer mot disse hierarkiene. De konkluderer med at kategorisering der kategoriene er strukturert i et hierarki, gir bedre resultat spesielt når antall kategorier og termer er stort. Da kan kategoriseringen bli delt opp i mindre prosesser, der hvert subhieraki håndteres for seg. De enkelte prosessene optimaliseres til å bruke kun de termer som passer best med subhierarkiets innhold. Undersøkelsen bygger på tilsvarende undersøkelser gjort tidligere, blant annet en utført av Chen & Dumais (2000). Her prøvde de ut hierarkisk kategorisering av innhold på Verdensveven. Konklusjonen deres er at en hierarkisk struktur gir noe bedre resultat enn flat struktur.

De aller fleste forsøk som er utført innenfor området automatisk tekstkategorisering, har hatt fokus

på uttesting av metoder og varianter av algoritmer. For å kunne sammenlikne resultater på tvers av forsøkene har mange brukt en samling med nyhetartikler. Spesielt har Reuter-21578 (Lewis 2004) og Reuters Corpus Volume l [RCV1] (Lewis, Li, Rose & Yang 2004) vært brukt. Disse består av et sett avisartikler fra Reuter som har blitt kategorisert manuelt. Det har, så langt jeg kan se, vært få forsøk der det er gjort sammenlikning mellom resultat av kategorisering på forskjellige datakilder. For at en sammenlikning skal kunne foretas, må de forskjellige datakildene tilhøre samme domene og de må bruke samme læringsmodell. Maria&Silva (1999, 2000) foretok et sånt forsøk da de lagde en prototyp av nyhetsportalen NewsSearch. Her samlet de inn en dags artikler fra 6 forskjellige portugisiske aviser (ca 1000 artikler) og brukte Support Vector Machine [SVM] for å tilordne kategorier automatisk. Det var 14 kategorier, og SVM var lært opp med artikler fra en nyhetsbase de sjøl hadde bygd opp. Fordi avisartikler svært ofte omhandler mer enn et emne, valgte de å tillate mer enn en kategori pr artikkel og de oppnådde da en så høy presisjon som 94,5 %. (Maria & Silva 2000). Nyhetartikler dekker dagsaktuelle saker, og hva dette er vil forandre seg over tid. Maria og Silva undersøkte derfor også metodens effekt over tid. I et eksempel trente og testet de SVM med artikler innenfor samme tidsperiode. Dette ble sammenliknet med et eksempel der de trente SVM med artikler fra en måned og testet ut den på artikler fra påfølgende måneder. Det siste ga et noe dårligere resultat, spesielt på dynamiske kategorier som «Sport» og «Politikk». Andre mer statiske kategorier viste ikke samme reduksjon, de nevner «Kultur og forretningsliv» som et eksempel her. NewsSearch er nå i drift, se www.newssearch.pt.
I NewSearch definerte de kategoriene sjøl og kategoriserte artiklene i læringssettet etter disse. (Maria & Silva 1999) . Et alternativ kunne vært å lære opp SVM med artikler fra en annen datakilde, der artiklene er kategorisert etter en allerede definert kategoriliste. Dette ville tilsvare forsøk jeg har kjørt. Litteratur som beskriver et sånt forsøk, har jeg ikke vært i stand til å finne.

Kategorisering ved bruk av maskinlæring baserer seg på at det eksisterer dokumenter som på forhånd er kategorisert manuelt. Disse dokumentene brukes både til å lære opp kategoriseringsverktøyet og til å teste ut dette når det er opplært. Utfordringen ved all kategorisering av tekst som utføres manuelt, er at den sjelden kan tolkes kun på en måte. Hvordan den forstås kommer an på leserens bakgrunn, erfaring og kunnskap. Forsøk viser at også godt trent personale kategoriserer samme dokument forskjellig. Harold Borko (1964) viser til flere undersøkelser der resultatene ligger så lavt som 50% konsistens i kategorisering. Dette gjelder både når en ser på konsistensen mellom forskjellige personer som kategoriserer samme dokument, og når samme person kategoriserer samme dokumenter på nytt etter en måned. Borko kjørte eget

forsøk der han lot tre psykologistudenter kategorisere 997 sammendrag av psykologirapporter. Resultatet her var et gjennomsnitt på 75% enighet mellom de tre. Victoria Uren (2000) gjennomførte forsøk der hun blant annet så på kategorisering av 9 dokumenter gjort av 4 svært erfarne personer. Konsistensen mellom dem viste seg å være så lav som 41% . Studien er liten, men i følge Uren (2000 s.3) er resultatet sannsynlig sett i forhold til andre forsøk.

2.1 Forskning på norskspråklig materiale

Det har vært forholdsvis lite forsking og forsøk på automatisk kategorisering av norskspråklig materiale. Mye av det som er gjort, er innenfor det lingvistiske fagfeltet for eksempel i det skandinaviske prosjektet Nomen Nescio (http://g3.spraakdata.gu.se/nn/) I april i år disputerte Åsne Haaland for dr.art.-graden med avhandlingen *A Maximum Entropy Approach to Proper Name Classification for Norwegian* (2008). Her undersøkte hun ulike tilnærminger for å kunne tilordne et egennavn til forhåndsdefinerte kategorier som person, organisasjon, lokasjon, tilsammen seks forskjellige kategorier. Hun så blant annet på effekten av å ta hensyn til ordene før og etter egennavnet i setningen, på effekten av akronymer og på bruk av oppslagstabeller. Forsøkene hun har gjort, kan være nyttig også for automatisk kategorisering. En kan da for eksempel skille mellom Jordan brukt som tannbørsteprodusent, Jordan som land og Jordan som person, noe som kan være med på å avgjøre hvilken kategori dokumentet bør tilordnes.

På høyskolen i Oslo er det tidligere skrevet to diplomoppgaver som omhandler automatisk kategorisering. Jørn Helge B. Dahl (2002) tok utgangspunkt i svartjenesten til Deichman og utredet og kjørte forsøk på automatisk kategorisering av e-posten som sendes til «Spør biblioteket». Konklusjonen hans var at med noen konkrete forbedringer, kan kategoriseringen benyttes for eksempel til statistikkgenerering. Av forbedringer han foreslo var en bedre stoppordliste, og også mer gjennomgripende endringer som «å basere seg på [...] helt andre kilder en dokumentene selv for trening av klassifikatoren» (Dahl 2002 s. 28) . Trond Strøm (2001) kategoriserte artikler fra Aftenpostens nettavis, og så på hvordan dette kan brukes for å bedre gjenvinningsmulighetene i et digitalt arkiv over artiklene i norske nettaviser.

En utfordring når en bruker norskspråklig materiale, er variasjonene i språkform både pga dialekter, bokmål og nynorsk samt bruken av radikalt eller konservativt språk. Dette gir seg spesielt til kjenne i materiale der det er flere forfattere, for eksempel i aviser der redaktøren stiller journalisten

forholdsvis fritt i valg av språkform. Den norske språknormen er inkluderende og tillater blant annet flere variasjoner av endelser på substantiv (a/en) og verb (a/et), mens roten av ordet er likt. Derfor vil en kunne få bedre gjenfinning ved å ta i bruk en stemmer.(fra eng. Stem - stamme). Denne fjerner endelser og lar kun ordstammen stå igjen. For engelskspråklig materiale er det utviklet gode stemmere, Porters stemmer (Porter 2006) er den mest kjente. Porter har utviklet nettstedet www.snowball.tartarus.org der han har implementert stemmere for flere språk, blant annet norsk. Denne ble prøvd ut i Pedersens (2002) fordypningsoppgave på NTNU, der han implementerte en norsk stemmer basert på Porters algoritme. Den samme stemmeren ble brukt av studenter på NTNU da de i faget 'kundestyrt prosjekt', så på automatisk kategorisering av dokumenter i emnekart på oppdrag fra firmaet Bouvet AS (Børke, Grythe, Løset, Mørch-Storstein & Vistnes 2005). Pedersens undersøkelse viste begrenset vinning ved å bruke en stemmer. Den medførte få nye treff, noe som ga lite utslag på resultatet. I Børke etl.al's oppgave hadde de ingen spesifikke tester der de vurderte effekten av stemmeren, og derfor heller ingen resultater å vise til.

3 Testkorpus

I mine forsøk har jeg brukt artikler fra to riksdekkende dagsaviser, Aftenpostens morgennummer og Klassekampen. For begge avisene har jeg et fullstendig sett med artikler for perioden september-desember 2007, samt alle artiklene fra mars 2008. Aftenposten har kategorier satt manuelt og skal derfor brukes til opplæring av kategoriseringsverktøyet. Klassekampen har ikke kategorier og skal i forsøket få tildelt kategorier med verktøyet lært opp på Aftenpostens artikler.

3.1 Aftenposten

Aftenpostens morgennummer er riksdekkende og er landets nest største avis, med et opplag på ca 250 000 og 732 000 lesere (Mediebedriftenes landsforbund 2007, 2008). Den er eid av Schibsted og regnes som en konservativ avis, og utkommer på riksmål. Fra å bruke et konservativt riksmål er språkbruken nå mer normalisert. I 1991 tok de i bruk rettskrivningen av 1986, der f.eks nå, etter og språk ble godkjente former for tidligere nu, efter og sprog. Daglig inneholder morgennummeret ca 170-180 artikler. Artiklene er forholdsvis korte, i perioden september-desember 2007 var gjennomsnittlig artikkellengde på 307 ord. Avisa dekker alle emner, og har for eksempel mye stoff om økonomi og næringsliv. I følge Karen Thorshaug i Aftenpostens dokumentasjonssenter (muntlig kommunikasjon 19 mai 2008) er kategorier påført manuelt i to nivåer. På det øverste nivå, som består av 17 kategorier, har de lagt seg nær opp til den internasjonale IPTC-standarden (International Press Telecomunications Council [IPTC] 2008), se tabell 1 på neste side.

Aftenpostens kategori	IPTC- standarden
Økonomi og næringsliv	04 Economy, bsiness & finance
Ulykker og naturkatastrofer	03 Disasters & Accidents
Sport	15 Sport
Forsvar krig og konflikter	16 Unrest, conflict & war
Politikk	11 Politics
Sosiale forhold	14 Social issues
Personalia	08 Human interest
Kriminalitet og rettsvesen	02 Crime, law & Justice
Vitenskap og teknologi	13 Science & Technology
Kultur og underholdning	01 Arts, culture & entertainment
Medisin og helse	07 Health
Utdanning	05 Education
Fritid	10 Lifestyle & Leisure
Arbeidsliv	09 Labour
Natur og miljø	06 Environmental issues
Religion og livssyn	12 Religion & belief
Diverse	

Tabell 1: Aftenpostens kategorier og IPTC

I følge Thorshaug (19 mai 2008) er kjente avvik fra IPTC-standarden som følger:

- IPTC kategori 17 «Weather» er utelatt, her bruker de istedet kategorien «Natur og miljø» med emneordet «Vær»
- IPTC kategori 08 «Humen interest» er begrenset til personalia
- Kategorien «Diverse» eksisterer ikke i IPTC, her har Aftenposten samlet emneord som vanligvis brukes i alle andre kategorier
- I Aftenposten er forsvar med i kategorien «Forsvar, krig og konflikter», i IPTC er forsvar en del av «Politikk»

Aftenposten bruker kategoriene både for artikler og bilder. Denne felles bruken av kategorier tilpasset IPTC ble tatt i bruk i november 2006 i forbindelse med innføringen av et nytt multimedialt arkiv internt i Aftenposten. Dette tolker jeg dithen at hovedhensikten med kategoriene er at Aftenpostens ansatte finner igjen artikler og bilder, de er altså ikke primært satt for søking og gjenfinning i A-tekst.

På nivå 2 er det emneord, disse er ulike for artikler og bilder. For artiklene er det ca 650, se en

fullstendig liste over alle kategorier og emner i vedlegg A. Sjøl om emneordene i følge Thorshaug (9 mai 2008) i hovedsak tilhører en kategori, brukes de i praksis under flere kategorier, se tabell 5. Artiklene kategoriseres av Aftenpostens dokumentasjonssenter. I følge Lene Li Dragland (muntlig kommunikasjon 26 mai 2008) tar de utgangspunkt i artikkelens innhold og plasserer først i kategori, så emne. Antall emneord begrenses, sjøl om artikkelens innhold tilsier at de kunne dekkes av mange. Det finnes ikke skriftlige retningslinjer bortsatt fra lista over kategorier og emner. I tillegg kommer epost og informasjonsutveksling dem i mellom. Lengden på opplæringen varierer, den er personavhengig og avhengig av hva som finnes av ressurser. Kvalitetskontrollen er uformell. «Det er svært lite kontroll - fordi vi blir stadig færre... vi har med andre ord ikke tid til å drive kvalitetskontroll slik vi gjorde den gang vi var perfeksjonister [...] vi forsøker å ha prinsippdiskusjoner når kinkige saker oppstår.» (Dragland muntlig kommunikasjon 29 mai 2008).

3.2 Klassekampen

Klassekampen (http://www.klassekampen.no/om_oss) er idag en vel anerkjent landsdekkende dagsavis tilhørende på venstresida i norsk politikk. Den kom ut som dagsavis i 1977 etter i 8 år å først ha vært en månedsavis og så en ukeavis for det som da var AKP-ml. I tiden etter 1990 gikk den fra å være et rent talerør for marxist-leninistene med AKP-ml og Rød Valgallianse som hovedeiere, til å bli en avis for hele venstresida. Den eies nå av partiet Rødt, Fagforbundet, forlagene Pax og Oktober og av Klassekampens venner. Den har bidragsytere med forskjellige politiske meninger og tilhørigheter. Opplaget er på ca 11000, og den blir lest av ca 55 000 (Mediebedriftenes landsforbund, 2007, 2008).

Avisa er riksdekkende og har en god utenriksdekning. Det skrives mye om fagforeningspolitikk, miljøvern og globalisering. Dessuten har den en god kulturdekning, med jamnlige spesialsider med bokanmeldelser. Sport og kjendisstoff er nesten helt fraværende. Det er ukentlige petitartikler, og innimellom trykkes det dikt og andre prosatekster.

De første årene var Klassekampen kjent for å ha en svært radikal språkbruk, dette har moderert seg. Redaktør Bjørgulf Braanen (i muntlig kommunikasjon 19 mai 2008) opplyser at de «har en bokmålsnorm som i utgangspunktet skal gjelde for alle fast ansatte journalister. Eksterne bidragsytere står imidlertid fritt til å bruke sin egen målform». Det er imidlertid også ansatte journalister som har nynorsk som målform og får skrive dette. I perioden september-desember 2007 er i underkant av ca 20 %[3] artikler på nynorsk. En undersøkelse medieanalytiker Tommy H. Brakstad har utført for Dagbladet, viser at Klassekampen ligger på topp i antall stavefeil (Myhr

3 Dette er kun et anslag, artiklene er identifisert ved at de inneholder minst et av ordene 'ikkje', 'ein', 'eit' , 'dei'

2008). Han har undersøkt feilstaving av 14 vanlige norske ord og Klassekampen har henholdsvis 17.3 % (nett) og 14.8% (papir) feil. «Klassekampen er kanskje mer opptatt av sak enn av språk, men årsaken til den dårlige plasseringen kan også være så enkel som at de har en liten redaksjon med relativt små ressurser», sier Brakstad i en kommentar til undersøkelsen (Myhr 2008).

Hver dag inneholder avisa ca 30 artikler, på lørdagen er det opp mot 50. Artiklene er forholdsvis lange, i perioden september-desember 2007 var de gjennomsnittlig på 617 ord. Til sammenlikning er gjennomsnittlig artikkellengde i Aftenposten 307 ord.
Klassekampen er i større grad enn mange aviser basert på bidrag utenfra, dette er en tradisjon helt fra 70 tallet, da avisa la sin ære i å ha mange innlegg fra aktivister på arbeidsplassene og i protestbevegelsene. I dag er bidragene både fra skribenter og grupper av skribenter som har fast spalte en gang i uka, og artikler som tidligere er trykt i andre aviser og tidsskrifter. Antall leserinnlegg er også svært stort sett i forhold til avisas størrelse, de utgjør som oftest 2-3 av totalt 24 sider. I september-desember 2007 er det noe over 500 forskjellige skribenter, av disse er kun ca 50 tilknyttet avisa[4]. De står imidlertid for nesten 2000 av artiklene, dette utgjør 65% av den totale artikkelmengden for perioden.

4 Identifisert ved at de har «klassekampen» som del av domenenavnet i epost-adressen

4 Valg av metode

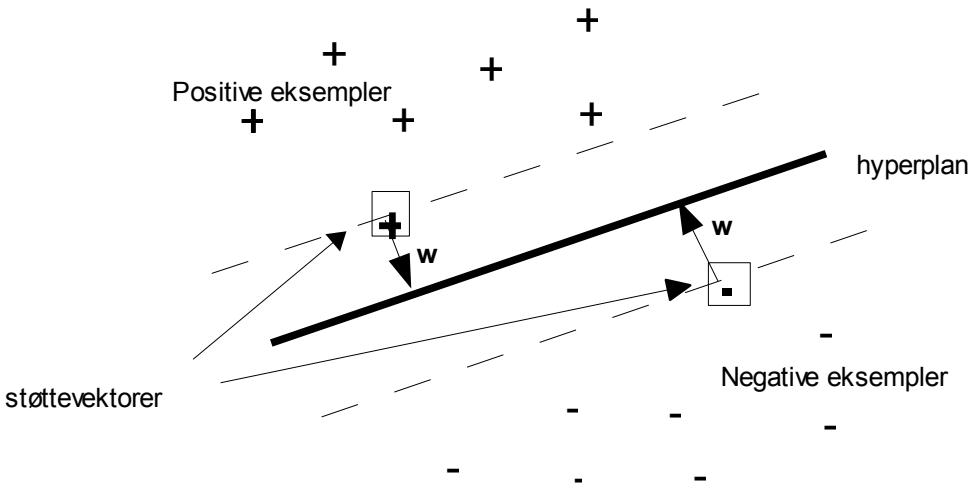
Metoder og algoritmer som kan benyttes for automatisk tekstkategorisering, er mange, og de fleste er vel utprøvd (Sebastiani, 2002). I de fleste forsøk har SVM gitt det beste resultatet, og det er denne som er brukt i forsøkene mine.

4.1 Størrelsen på læringssettet

Som mange av de andre metoder for automatisk kategorisering, bygger SVM på metodene fra fagområdet maskinlæring. Dokumentmengden deles i to, der en del brukes for opplæring av verktøyet og en del brukes for uttesting av resultatet av læringen. I forsøk som benytter dokumentsamlingen Reuters-21578, anbefales det å bruke ModApte split (Lewis 2004). Mange forsøk har fulgt denne anbefalingen (Dumais et.al. 1998, Joachims 2002). Hvis andre dokumentsamlinger skal brukes for læring, er den kritiske faktoren størrelsen på læringssettet og antall dokumenter pr kategori. Dumais et.al (1998) gjorde forsøk ved å variere antall positive eksempler pr kategori. Konklusjonen var at det minst må være 20 dokumenter for hver kategori for å få et stabilt resultat.

4.2 Support Vector Machine

Support Vector Machine [SVM] ble utviklet av Vladimir N. Vapnik innenfor fagområdet statistisk læring. Videre forskning, spesielt av Torsten Joachims (2002), tilpasset metoden til bruk for automatisk kategorisering og viste at den er godt egnet. Ideen med SVM er at det ved hjelp av læringssettet, opprettes et hyperplan som separerer positive eksempler fra negative på en sånn måte at avstandene mellom dem blir maksimal. Eksemplene nærmest til hyperplanet blir kalt støttevektorer, derav navnet Support Vector Machine, se figur 1 på neste side.



Figur 1 SVM's hyperplan og støttevektorer

Metoden ble for alvor tatt i bruk etter at Joachims utviklet SVM^{light} (Joachims 1998). Denne versjonen av SVM er tilpasset store datamengder, noe som ble oppnådd både ved å utelate termer som ikke var med i dokumentet fra dokumentvektoren, ved å optimalisere læringsalgoritmen og ved å innføre caching.

De fleste metoder og det meste av forskningen har basert seg på binær kategorisering, det vil si at teksten sjekkes mot en kategori for å se om den kan tilordnes denne eller ikke. Ved flere kategorier må kategoriseringen da foretas en gang pr kategori, for så å sammenstille resultatene til slutt. Mens SVM^{light} kun håndterer binær kategorisering kan *SVM*^{multiclass} håndtere flere kategorier av gangen. Begge versjoner kan brukes på flere plattformer og kan lastes ned fra http://svmlight.joachims.org/ I forsøkene mine har jeg valgt å bruke *SVM*^{multiclass} . Årsaken er at jeg har mange kategorier som hvert dokument kan knyttes opp til, og det er arbeidsbesparende å bruke et verktøy som ikke må kjøres for hver enkelt kategori, men som kan kjøres en gang for alle. Ved en binær kategorisering måtte jeg i tillegg ha sammenstilt resultatet fra de enkelte kjøringene for å finne riktig kategori for dokumentet. *SVM*^{multiclass} gir informasjon om beste kategori. Multippel kategorisering har vist gode resultater i forsøk (Crammer & Singer 2001). I dette forsøket brukte Crammer &Singer rett nok en noe annen algoritme enn den som ligger til grunn for *SVM*^{multiclass} (Joachims 2007).

SVM kan parameterstyres, men standardverdiene har også vært brukt i forsøk med SVM^{light} og gitt godt resultat (Alpha, Dixon & Liao s.a). Jeg har ikke funnet tilsvarende forsøk med informasjon om parametersetting for *SVM*$^{multiclass.}$. Derfor har jeg valgt å bruke standardverdier, med unntak for

parameter *c,* denne angir hvor tolerant læringsalgoritmen skal være for feil. Jo lavere *c*, jo mindre nøyaktig blir hyperplanet og støttevektorene. I retningslinjene som følger *SVM*^{multiclass} angis det at standardverdi for *c* er 0.01, se http://svmlight.joachims.org/ I et eksempel Joachims har lagt ut på samme sted setter han imidlertid *c*= 5000. Fordi det er så stor forskjell i størrelsen på disse to verdiene, er læringsalgoritmen blitt testet med flere verdier for *c,* se tabell 2.

Verdi for *c*	F-verdi[5]
0.01	0.6025
1	0.6025
5	0.6318
10	0.6870
50	0.7486
100	0.7478
1000	0.7518
2500	0.7558
5000	0.7592
10000	0.7550

Tabell 2: resultat av kategorisering av Aftenposten med forskjellige verdier for *c*
læringssettet består av 6574 artikler, testsettet av 6026
stoppord er fjerna

Det en merkbar bedring av resultatene fra *c* = 1 til *c* = 50, mens det etter det stiger sakte med en liten topp for *c* = 5000. Det bør derfor velges en verdi som er større enn 50, men hvor mye større den bør være, ser ikke ut til å spille noen særlig rolle. Jeg har valg å bruke *c* = 5000 siden dette var den verdien Joachims brukte i eksemplet sitt.

SVM fordrer at begge datasettene er på matriseform, der antall kolonner i matrisen er antall termer i total dokumentmengde. Hver rad i matrisen er et dokument og hver enkelt celle angir termens vekt. Termer som ikke forekommer i et dokument, utelates fra denne raden i matrisen, og på den måten reduseres størrelsen på datasettet drastisk. SVM^{multi} benytter seg av cashing og derfor er det viktig med stort minne. Jeg kjørte på en maskin med AMD 2 GHz prosessor og 2 GB RAM. Operativsystemet var Windows Vista™ .

5 F-verdien er beskrevet og definert i kapittel. 4.4.1

4.3 Termer og vekting

En viktig faktor ved automatisk kategorisering er valg av termer og vektingsalgoritmer. Termene trekkes ut av dokumentenes tekst. Som oftest er termene enkeltord adskilt med blanke i teksten, men en kan også ha sammensatte termer. Ved enkeltord kan en miste informasjonen som ligger i ordets plassering i setningen. Hvorvidt dette har noe å si for resultatet av den automatiske kategoriseringen, har vært diskutert lenge, men flere forsøk viser liten eller ingen forbedring i å bruke sammensatte ord. (Alpha et.al, Dumais et.al 1998). Jeg valgte derfor å trekke ut enkeltord som termer.

Termene bør bidra til å skille dokumenter og kategorier fra hverandre. Ord som ikke i seg sjøl er meningsbærende og som forekommer hyppig innenfor alle dokumenter, kalles stoppord og kan utelates som termer. Stoppordene omfatter i hovedsak preposisjoner, konjunksjoner, adverb og artikler. I de fleste forsøk utelates stoppord fra termlista. Joachims (2002 s. 110) konkluderer imidlertid i forsøk han gjorde for SVM at for Reuters-21578 blir resultatet noe bedre uten bruk av stoppordliste. Jeg valgte å prøve både med og uten stoppord.

Når termene er enkeltord vil det totale antall termer i en stor dokumentmengde bli svært høyt. For eksempel inneholder Aftenposten morgennummer i perioden september-desember 2007 ca 140 000 forskjellige ord, dvs termer. Flere forskere har sett på muligheter for å redusere termmengden ved å fjerne termer som er brukt sjelden, men Joackims (2002 s. 47) advarer mot dette da det kan medføre tap av informasjon siden språket ikke er entydig og ingen termer er felles for alle relaterte dokumenter. Det viser seg at dette også stemmer i praksis. Forsøk gjort med bruk av SVM for Reuter-21578 gir lite utslag på resultatet når antall termer reduseres (Bekkerman, El-Yaniv, Tishby, & Winter 2003). Jeg har kjørt forsøk både med og uten reduksjon i termer.

Samme ord vil i en tekstsamling ha flere former, f.eks genitivs form og forskjellige bøyningsformer av spesielt substantiv og verb. For å minske dette problemet, kan en bruke en stemmingsalgoritme. Her er, som nevnt tidligere, Potter's algoritme mest brukt, og den har også en norsk versjon. I forsøkene gjort av Halvorsen (2002) konkluderte imidlertid han med at stemming hadde liten innvirkning på resultatene. Den samme konklusjonen har Joachims (2002 s. 110). Han har ved forsøk oppnådd dårligere resultat ved stemming enn uten. Jeg brukte ikke stemmingalgoritme i mine forsøk.

Ikke alle ord, dvs termer, i en tekst bidrar like bra til å skille dokumentet fra andre dokumenter. Derfor er det vanlig å vekte termene utfra hvor ofte de forekommer i dokumentet og i samlingen som helhet. En term som forekommer hyppig i et dokument, sier sannsynligvis mer om

dokumentets innhold enn en term som forekommer kun en gang. En term som forekommer i mange av dokumentene i samlingen, er lite egnet til å skille dokumenter fra hverandre. Den vanligste vektingsalgoritmen er

$$w_{ij} = tf_{ij} * \log\frac{|d|}{df_{ij}}$$

Denne ble først beskrevet for bruk i informasjonsgjennfinning av Buckley & Salton (1987). tf_{ij} er termfrekvensen og $\frac{|d|}{df_{ij}}$ angir dokumentfrekvens hvor |d| er totalt antall dokumenter.

Når dokumentene i en samling har forskjellig lengde, kan vekten normaliseres med hensyn til dette:

$$w_{ij} = \frac{tf_{ij} * \log\frac{|d|}{df_j}}{\sum_{j=1}^{t} [tf_{ij} * \log\frac{|d|}{df_j}]^2}$$

Normaliseringen bidrar til å nøytralisere påvirkningen dokumentlengden vil ha på vektingen sånn at lange og korte dokumenter behandles likt. Fordi lengden varier sterkt både innenfor hver avis og avisene i mellom, vil jeg normalisere for lengden.

4.4 Evaluering av resultatene

4.4.1 Fullstendighet, presisjon og F-verdi

I gjenfinning har det vært vanlig å måle resultater i fullstendighet og presisjon. I følge Sebastiani (2002 s. 37) er dette også en aktuell målemetode for automatisk kategorisering. Her sammenliknes resultatene fra den automatiske kategoriseringen med en fasit som består av de samme dokumenter med kategorier påført, som oftest manuelt. Det tas utgangspunkt i en binær krysstabell der hver celle angir resultatet av sammenlikningen mellom den automatisk satte kategorien og den manuelle for et dokument, se tabell 3 på neste side.

Kategori		Manuelt satt kategori	
		Ja	Nei
Resultat av automatisk kategorisering	Ja	Funnet/Relevant **A**	Funnet/ikkeRelevant **B**
	Nei	IkkeFunnet/Relevant **C**	IkkeFunnet/Ikke relevant **D**

Tabell 3 Relevanse

Fullstendigheten er her antall riktig automatisk kategoriserte artikler i forhold til antall artikler som har denne kategorien satt manuelt

$$\rho = \frac{A}{A + C}$$

Presisjon er antall riktig kategoriserte artikler i forhold til totalt antall artikler som automatisk har fått denne kategorien

$$\pi = \frac{A}{A + B}$$

Presisjon og fullstendighet har en tendens til å utelukke hverandre, en høyere verdi av den ene gir en lavere verdi av den andre. Mekanismer som for eksempel innføres for å øke presisjon minker fullstendigheten.

Et felles mål for fullstendighet og presisjon er F-verdien (Sebastiani 2002 s. 42) . Den er gitt ved

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho}$$

der β angir vekten en vil gi fullstendighet i forhold til presisjon. Er $\beta = 0$ beregnes F-verdien kun utfra presisjon, mens hvis β er uendelig stor beregnes F-verdien kun utfra fullstendighet. Vanligvis settes β til 1, da har begge variabler lik betydning.

Programmet jeg har skrevet for beregning av fullstendighet, presisjon og F-verdi ligger i vedlegg D.

4.4.2 Beregning av gjennomsnitt

To metoder kan brukes for å beregne gjennomsnittsverdiene for presisjon og fullstendighet:

1. *'Macroaveraging'* der presisjon og fullstendighet først beregnes pr kategori. Gjennomsnittsverdiene beregnes så ut fra dette. Metoden er vanlig i statistiske beregninger og er den vanligste måten for å beregne gjennomsnitt. Alle kategoriene vektes likt.

$$\hat{\pi}^{m} = \frac{\sum_{i=1}^{|k|} \pi_i}{|k|} \qquad \hat{\rho}^{m} = \frac{\sum_{i=1}^{|k|} \rho_i}{|k|}$$

2. *'Microaveraging'* der gjennomsnittlig presisjon og fullstendighet beregnes på samlingen som helhet. Med denne metoden får store kategorier mer betydning enn små.

$$\hat{\pi}^{\mu} = \frac{\sum_{i=1}^{|k|} A_i}{|\sum_{i=1}^{|k|} (A_i + B_i)} \qquad \hat{\rho}^{\mu} = \frac{\sum_{i=1}^{|k|} A_i}{|\sum_{i=1}^{|k|} (A_i + B_i)}$$

For begge metoder angir |k| totalt antall kategorier.
Fordi kategoriene jeg vil bruke er forskjellig i størrelse og resultatene for små kategorier er like interessant som for de store, vil jeg bruke 'macroaveraging'. Dette er også det som anbefales i litteraturen når en har kategorier med svært forskjellig størrelse (Manning, Raghavan & Schütze 2008) . Lewis (1991) påpeker at det er en utfordring med 'macroaveraging' når det er kategorier som ikke er brukt for noen artikler, og både fullstendighet og presisjon får verdien 0. Dette slår forholdsvis hardt ut på gjennomsnittet. For å motvirke dette, foreslår han å regne ut gjennomsnittet kun for de kategorier som har verdier større enn 0. Samme utvalg må da brukes for hele forsøket.

En annen utfordring er det når det er få artikler pr kategori. Da vil kategoriseringen av en enkelt artikkel, få stort utslag på resultatet. Hvis kun en artikkel er tilordnet en kategori, vil resultatet være binært, 0 eller 1, avhengig av om artikkelen ble kategorisert riktig eller ikke. Med macroaveraging gir dette store utslag. I enkelte forsøk kan det derfor være aktuelt å ikke se på gjennomsnittverdiene, men beholde resultatene fra hver enkel kategori for seg.

4.4.3 Utvalget av artikler for sammenlikning av resultater

Artiklene i Klassekampen har ikke påført kategorier manuelt. For å evaluere kvaliteten av den automatiske kategoriseringen, må dette gjøres. Det beste ville være å manuelt kategorisere en like representativ mengde av Klassekampens artikler som i Aftenpostens testsett. Dette er imidlertid en ressurskrevende oppgave som er umulig med den tiden og de ressurser jeg hadde til rådighet. Istedet kan det trekkes ut et mindre utvalg av artikler. For å få en best mulig evaluering av den automatiske kategoriseringen, er det viktig at alle kategorier er representert blant artiklene som velges ut. Ved et helt tilfeldig utvalg kan en risikere at små grupper ikke kommer med, derfor kan en bruke dispropososjonal stratifisering (Ringdal, 2001 s. 145). Her trekker en ut et tilnærmet likt antall

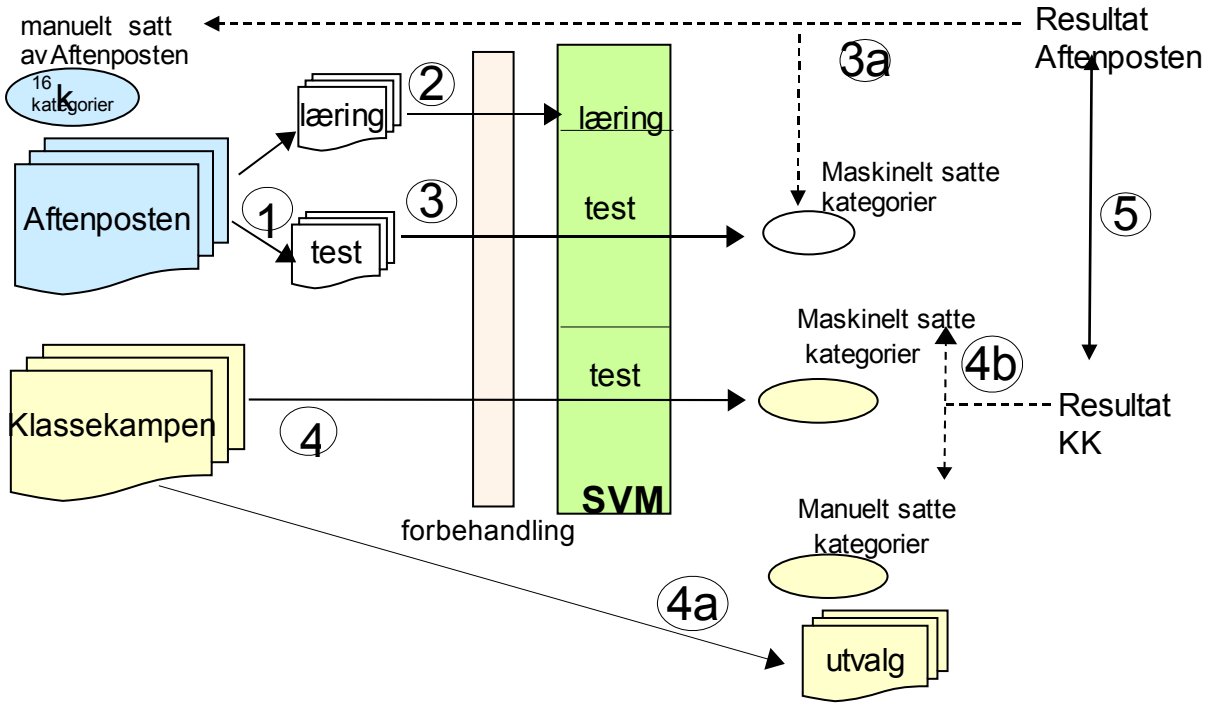
artikler innenfor hver gruppe, uavhengig av gruppas størrelse i total populasjon. Dette er aktuelt i mitt tilfelle siden antall artikler pr kategori varierer sterkt. Etter at artikkelsamlingen til Klassekampen er tildelt kategorier automatisk, tas derfor et tilnærmet likt antall tilfeldig utvalgte artikler innenfor hver kategori ut for manuell behandling og evaluering.
En utfordring med dette er at sjøl om fordelingen er jamn utfra en automatisk kategorisering, kan den igjen bli skjev når de samme artikler kategoriseres manuelt. Dette kan resultere i at det allikevel blir få artikler pr kategori, noe som kan få konsekvens for beregning av gjennomsnitt, se kapittel 4.4.2.

4.4.4 Statistisk signifikans

Ved kvantitativ analyse bør det statisk bevises at resultatene en får i forsøkene ikke kommer av tilfeldigheter, men har statistisk signifikans. Dette kan måles ved χ^2 (kjikvadrat). I henhold til KristenRingdal (2001 s. 334) har χ^2 testen som forutsetninger at utvalget skal være trukket tilfeldig og at antall forventede frekvenser skal overstige 5 i alle ruter i krysstabellen. For Klassekampen er dette ikke tilfelle. Utvalget har, som beskrevet i 4.4.3. en skjev fordeling og kan derfor ikke sies å være trukket helt tilfeldig. Det er også ruter i tabellen som har lavere verdi enn 5. Utfra dette kan statistisk signifikans ikke beregnes med χ^2 for Klassekampen og jeg bruker derfor ikke χ^2 i oppgaven.

5 Forsøk og erfaringer

I dette kapittelet redegjør jeg for forsøkene jeg kjørte for å få svar på forskningsspørsmålene nevnt i innledningskapittelet. Jeg beskriver de viktigste valgene jeg tok underveis, hvilke erfaringer jeg gjorde og hvilke resultater forsøkene ga. I figur 2 er en skisse over forsøket knyttet til spørsmålet om kvaliteten på den automatiske kategoriseringen vil være like god om læringsdokumentene kommer fra en annen kilde enn dokumentene en kategoriserer, som om de kom fra samme kilde.



1 Artiklene fra Aftenposten ble delt i en læringsdel og en testdel

2. SVM ble lært opp vha artiklene i læringssettet. Flere forsøk ble kjørt for å se konsekvenser av endret termstørrelse, bruk av stoppord, reduksjon av læringssett

3. Aftenpostens testsett ble påført kategorier automatisk

3a. De manuelle og de maskinelle kategoriene ble sammenliknet (Aftenposten)

4. Klassekampens artikler ble påført kategorier automatisk

4a. 310 av Klassekampens artikler ble påført kategori manuelt

4b. De manuelle og de maskinelle kategoriene ble sammenliknet (Klassekampen)

5. Resultatene fra Aftenposten og Klassekampen ble sammenliknet

Figur 2: oversikt over forsøk

5.1 Opplæring av kategoriseringsverktøyet

SVM, ble opplært og testet med bruk av artikler fra Aftenposten. Artiklene er uttrekk fra Aftenpostens morgennummer i tidsrommet september-desember 2007. Totalt utgjorde dette 20874 artikler. 4030 av disse var imidlertid uten kategori og ble derfor forkastet. Dette er artikler som blant annet inneholder førstesideoppslagene, leserinnlegg og kronikker. Ca 1100 artikler hadde mer enn en kategori. Et mål for metoden er at kategoriene er tydelig avgrenset fra hverandre. En artikkel knyttet opp mot to kategorier er med på å minske avstanden mellom disse to kategoriene, jeg valgte derfor å fjerne dem fra læringsmodellen. Aftenposten har 17 kategorier på øverste nivå, men kategorien «Diverse» var bare brukt alene i 7 artikler. Den er for liten til å være med i forsøkene og både kategorien og de 7 artiklene ble forkastet. Tilbake var da 16 kategorier som vist i tabell 4.

	Antall artikler	
Kategori	Læring	test
Økonomi og næringsliv	1653	1068
Ulykker og naturkatastrofer	218	145
Sport	2190	1374
Forsvar og krig og konflikter	359	221
Politikk	669	385
Sosiale forhold	238	142
Personalia	522	324
Kriminalitet og rettsvesen	669	395
Vitenskap og teknologi	189	103
Kultur og underholdning	1670	1045
Medisin og helse	272	172
Utdanning	172	117
Fritid	310	201
Arbeidsliv	249	143
Natur og miljø	233	143
Religion og livsyn	71	48
sum	9684	6026

Tabell 4 Fordeling mellom læring og testsett Aftenposten september-desember 2007

Læringssettet og testsettet fra Aftenposten ble delt i henholdsvis 9684 artikler for læring og 6026

for test. Størrelsen av læringssettet ble valgt for å sikre at det inneholdt mange nok artikler for hver kategori, også de minste kategoriene. Det samme gjaldt for testsettet, her var det viktig å sikre at ingen av kategoriene var knyttet til så få artikler at resultatene ville bli usikkert pga dårlig testgrunnlag.

Artiklene er på Retrievers XML-format, http://nyheter.no/RetXML/retxml.dtd .
Formatet har en fri form der de fleste elementer er valgfrie å bruke. Artikkelens tekst ligger samlet under XML-taggen Story, og en del av artiklene starter med en kort ingress adskilt fra resten av teksten med journalistens navn skrevet i store bokstaver. Fordi det ikke er en gjennomgående bruk av ingress i artiklene, har jeg ikke behandla ingressen spesielt, men kun som en del av teksten i artikkelen. Jeg har skrevet programmer for å pakke ut XML'en, og for å danne de vektormatrisene for læring og test som brukes av SVM. De viktigste programmene for oppretting av matrise og vekting av termer ligger i vedlegg D.

5.1.1 Valg av kategorier

Som beskrevet i kapittel 3.1 har Aftenposten kategorier på to nivåer. På øverste nivå er det, når en ser bort fra kategorien «Diverse», 16 kategorier. Fordi de er så få, blir de svært omfattende. For eksempel dekker kategorien «Politikk» både utenriks og innenrikspolitikk med blant annet emner som partipolitikk, ytringsfrihet, menneskerettigheter og diskriminering. På neste nivå var det i perioden september-desember 2007 brukt 592 forskjellige emner. Disse gir en mer detaljert kategorisering enn de overordnede kategoriene.

Emnene er ikke unike, men kan inngå i flere kategorier, se tabell 5 på neste side.
I læringssettet inngår hvert emne i gjennomsnitt i 3.2 kategorier. To emner, 'Administrasjon' og 'Undersøkelser', forekommer under alle de 16 kategoriene, mens 209 emner er unike for en kategori. I den 4 måneders perioden artiklene er hentet fra, forekommer 16% av emnene kun en gang totalt på tvers av kategoriene.

Kategori	Antall artikler	Antall emner brukt i kategorien	Antall emner kun brukt i denne kategorien
Økonomi og næringsliv	1653	300	47
Ulykker og naturkatastrofer	218	83	11
Sport	2190	156	45
Forsvar og krig og konflikter	359	112	15
Politikk	669	150	18
Sosiale forhold	238	132	7
Personalia	522	31	1
Kriminalitet og rettsvesen	669	169	13
Vitenskap og teknologi	189	79	2
Kultur og underholdning	1670	186	21
Medisin og helse	272	101	4
Utdanning	172	74	11
Fritid	310	97	4
Arbeidsliv	249	94	2
Natur og miljø	233	101	7
Religion og livsyn	71	39	0
Sum	9684		209

Tabell 5 Manuelt påførte kategorier og emner i læringssettet, 9684 artikler

Det er to tilnærmingsmåter hvis Aftenpostens emner skal benyttes i forsøkene:

1. Enkelte av de største kategoriene kan erstattes av enkelte av emnene knyttet til kategorien. Da vil en oppnå en jamnere fordeling på kategoriene.
2. Emner kan brukes i tillegg til kategoriene. Da ville kategoriene og emnene utgjøre et hierarki, og en kan i følge Chiang et.al (2007) dele kategoriseringen opp i mindre prosesser, der hvert subhieraki håndteres for seg. Da ville det spille mindre rolle om et emne er brukt under flere kategorier, fordi kategorisatoren læres opp for hvert enkelt subhierarki. Det er sannsynlig at for eksempel artikler som tilhører emnet «Administrasjon» under kategorien «Sport», har en annen termsammensetning enn artikler som tilhører samme emne under kategorien «Religion og livssyn».

Et problem med å bruke begge disse metodene for Aftenpostens uttrekk, er at kun 10% av emnene i læringssettet har nok artikler til å kunne brukes. Dumais et.al (1998) har konkludert med det læringssettet minst bør inneholde 20 positive eksempler pr kategori.
Et annet problem er at over halvparten av artiklene er tilordnet mer enn ett emne. Disse artiklene bidrar således dårligere til å skille emnene fra hverandre og bør utelates. Tas det hensyn til dette vil antall emner med mer enn 20 artikler reduseres til 2% av antall emner i læringssettet. Dette er et så lite antall at jeg ikke anser det som relevant å bruke emnene i forsøkene. Jeg vil derfor holde meg til kategoriene på øverste nivå.

Skulle emner vært brukt, måtte læringssettet bygges opp av artikler fra en mye lenger tidsperiode, men også da vil en nok erfare at enkelte emner er for lite brukt. Dette er en utfordring med å bruke metoder fra maskinlæring. Det kan være vanskelig å bygge opp bra nok læringssett når antall kategorier/emner er så stort som i Aftenposten, og dette kan vel være en grunn til at metoder som bygger på regelsett, forsatt er i bruk.

Størrelsen på kategoriene varierer sterkt , noe som virker inn på resultatene av den automatiske kategoriseringen. Se tabell 6 på neste side.

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0.9054	0.7525	0.8219
Ulykker og naturkatastrofer	0.6414	0.9300	0.7592
Sport	0.9789	0.9525	0.9655
Forsvar og krig og konflikter	0.8281	0.8356	0.8318
Politikk	0.7740	0.7801	0.7771
Sosiale forhold	0.4085	0.7073	0.5179
Personalia	0.7809	0.9336	0.8504
Kriminalitet og rettsvesen	0.8506	0.8463	0.8485
Vitenskap og teknologi	0.3398	0.7447	0.4667
Kultur og underholdning	0.9292	0.7658	0.8396
Medisin og helse	0.6977	0.8392	0.7619
Utdanning	0.7521	0.8627	0.8037
Fritid	0.5025	0.8632	0.6352
Arbeidsliv	0.4965	0.8256	0.6201
Natur og miljø	0.5105	0.7849	0.6186
Religion og livsyn	0.3333	0.7273	0.4571
gjennomsnitt	**0.6706**	**0.8220**	**0.7235**

Tabell 6[6] Fullstendighet og presisjon Aftenposten september-desember 2007
9684 artikler er brukt for opplæring og 6026 artikler for uttestingen.
stoppord er fjernet, c-verdi = 5000

Det er stor variasjon i fullstendighet og presisjon avhengig av kategori. Jamt over har de store kategoriene, som «Økonomi og næringsliv», «Sport» og «Kultur og underholdning», svært god fullstendighet, mens presisjonen er noe lavere. En utfordring med kategorisering av avisartikler er at artiklene ofte ikke omhandler kun ett tema, men ofte grenser opp til andre. Dette kan være årsaken til at de små kategoriene taper i forhold til de store. For eksempel vil mange av artiklene under kategorien «Arbeidsliv» også omhandler økonomiske forhold, 29% av disse artiklene har automatisk fått tilordnet kategorien «Økonomi og næringsliv», se vedlegg C.1.1. Dette medfører at fullstendigheten blir dårlig for «Arbeidsliv», og presisjonen lavere for «Økonomi og næringsliv». Sport skiller seg positivt ut med både god presisjon og fullstendighet, noe som kan forklares med at Sport er et tydelig avgrenset tema og dermed er lite overlappende med andre kategorier.

For å justere ned innvirkningen de største kategoriene har på testresultatet, har jeg kjørt tester der

6 Grunnlagsdata for beregning av resultatene ligger i vedlegg C.1.1

læringssettet kun inneholder opp til 801 artikler pr kategori. Resultatene vises i tabell 7. Læringssettet er redusert til 6574 artikler.

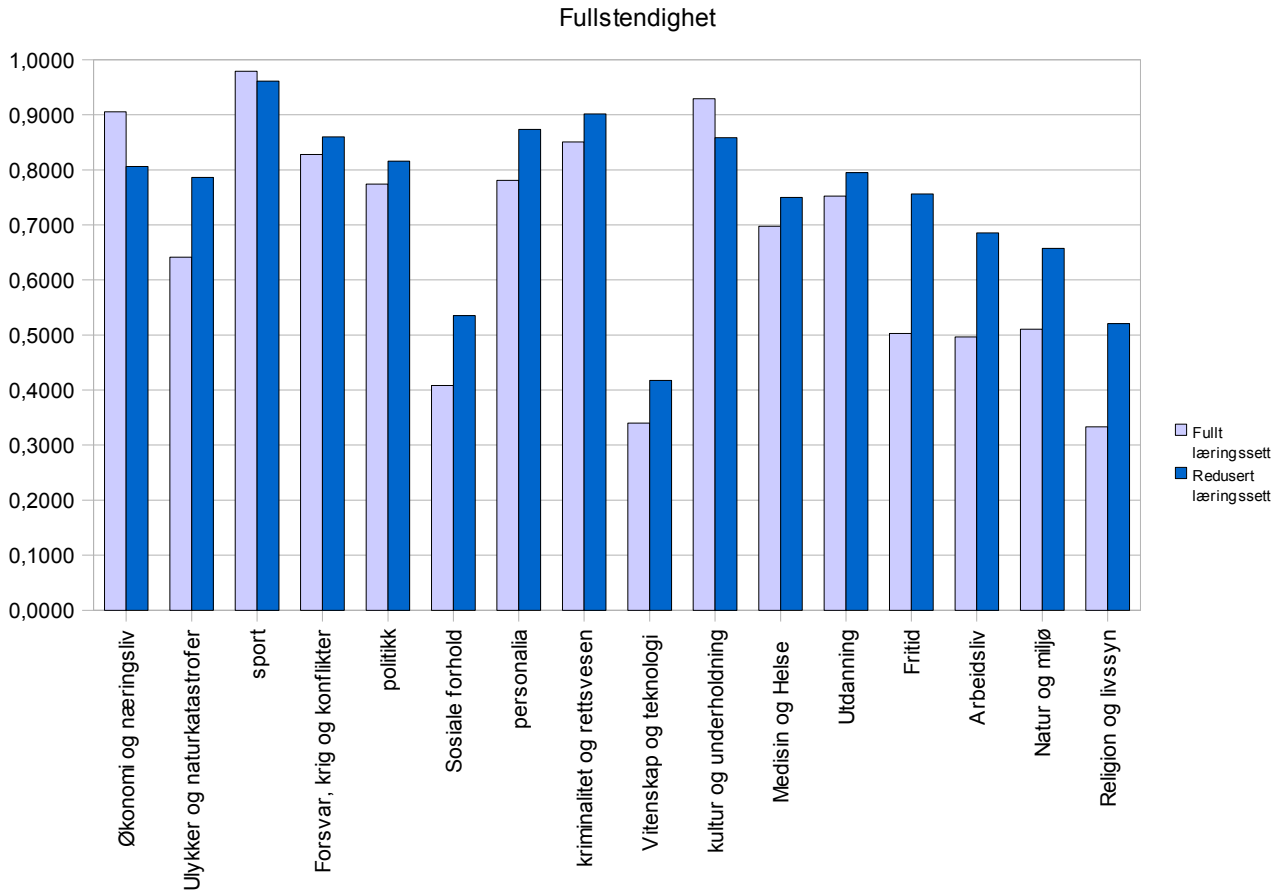
Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0.8062	0.8416	0.8235
Ulykker og naturkatastrofer	0.7862	0.8444	0.8143
Sport	0.9614	0.9763	0.9688
Forsvar og krig og konflikter	0.8597	0.7950	0.8261
Politikk	0.8156	0.6611	0.7302
Sosiale forhold	0.5352	0.6786	0.5984
Personalia	0.8735	0.8179	0.8448
Kriminalitet og rettsvesen	0.9013	0.7574	0.8231
Vitenskap og teknologi	0.4175	0.6825	0.5181
Kultur og underholdning	0.8584	0.8667	0.8625
Medisin og helse	0.7500	0.7866	0.7679
Utdanning	0.7949	0.8017	0.7983
Fritid	0.7562	0.8000	0.7775
Arbeidsliv	0.6853	0.7259	0.7050
Natur og miljø	0.6573	0.6861	0.6714
Religion og livsyn	0.5208	0.7576	0.6173
gjennomsnitt	**0.7487**	**0.7800**	**0.7592**

Tabell 7[7] Fullstendighet og presisjon Aftenposten september-desember 2007
6574 artikler er brukt for opplæring og 6026 artikler for uttestingen.
stoppord er fjernet, c-verdi = 5000

Dette gir store endringer for den automatiske kategoriseringen. Det er kategoriene «Økonomi og næringsliv», «Sport» og «Kultur og underholdning» som hadde mer enn 801 artikler, se tabell 4, og som har blitt redusert i dette læringssettet. Konsekvensen for alle disse er at antall artikler som tilordnes dem automatisk, går ned. For eksempel er antall artikler som er tilordnet «Økonomi og næringsliv», redusert fra 1285 ved fullt læringssett til 1023 artikler med redusert læringssett. «Økonomi og næringsliv» og «Kultur og underholdning» mister artikler til alle de andre kategoriene, mens artikler som var under «Sport», i hovedsak er flyttet til «Politikk», «Personalia» «Kriminalitet og rettsvesen» og «Fritid». De andre kategoriene får større oppslutning.

7 Grunnlagsdata for beregning av resutatenen ligger i vedlegg C.1.2

Fullstendigheten er nå jamnere fordelt mellom kategoriene, se figur 3. Den har gått ned for de store kategoriene «Økonomi og næringsliv», «Sport» og «Kultur og underholdning», og har gått opp for de andre.



Figur 3 Fullstendighet Aftenposten september-desember 2007
Resultater av automatisk kategorisering basert på fullt og redusert læringssett
stoppord er fjernet, c-verdi = 5000

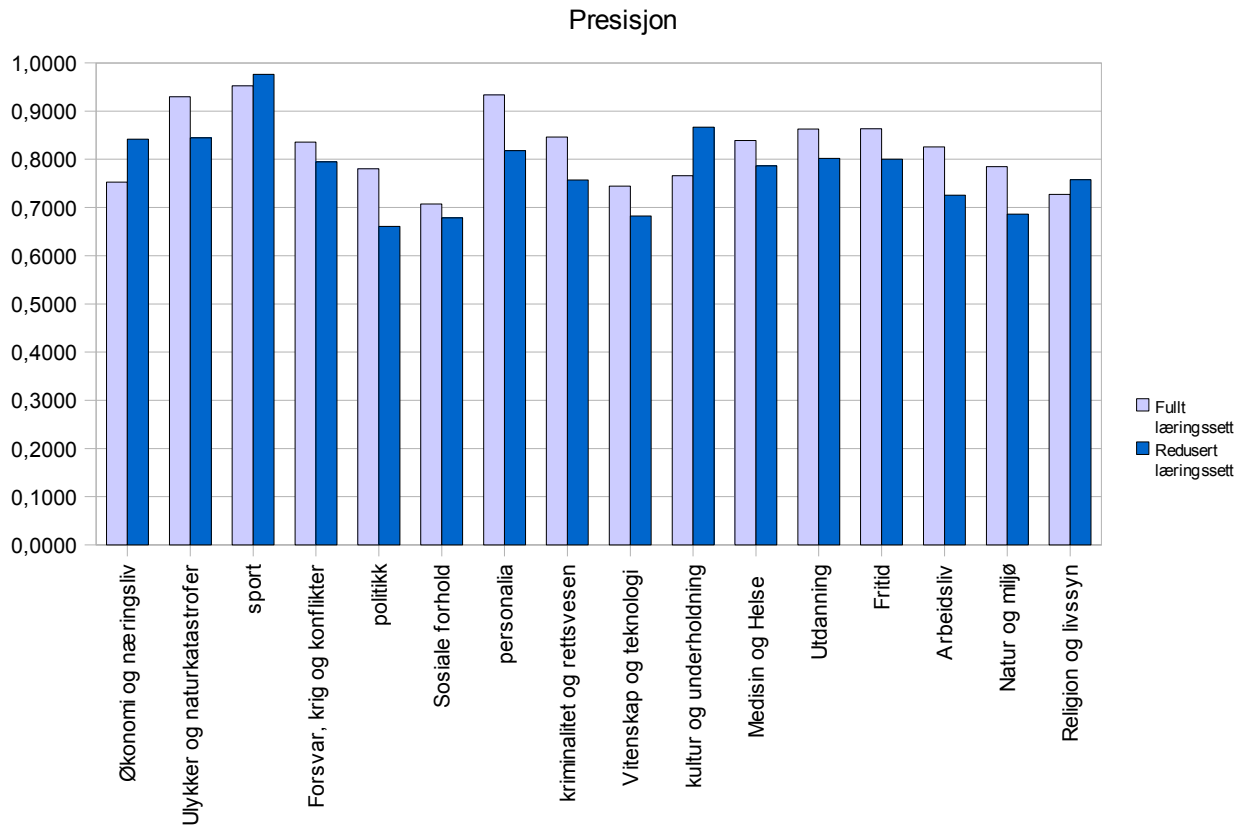
«Arbeidsliv» har økt fullstendigheten til 0.69, og nå har kun 11 % av artiklene som tilhører «Arbeidsliv» automatisk blitt tilordnet «Økonomi og næringsliv», se vedlegg C.1.2. Tilsvarende gjelder for «Religion og livssyn» Her er nå 19% av artiklene automatisk kategorisert under «Kultur og underholdning», mot 48% med fullt læringssett. Fullstendigheten har med det gått opp fra 0.33 til 0.52.

Kategorien «Sport» viser svært gode resultater med begge læringssett. Årsaken er nok som skrevet tidligere, at denne kategorien er klarere avgrenset. Dessuten har disse artiklene forholdsvis likt innhold, for eksempel er 49% av sportsartiklene i det reduserte læringssettet om fotball.

Kategorien «Vitenskap og teknologi» har tilsvarende en dårlig fullstendighet både med fullt og

redusert læringssett. En stor del av de artiklene hvor det ikke var samsvar mellom manuell og automatisk kategorisering for denne kategorien, er artikler om 'dagen idag'. Disse omtaler begivenheter som hendte på denne dagen bakover i historien og har derfor blitt tilordnet kategorien «Vitenskap og teknologi» med emne «Historie» manuelt. Automatisk er artiklene tilordnet kategori etter begivenheten som er omtalt. Årsaken til dette er at tilsvarende artikler som 'dagen idag', finnes i læringssettene. Artiklene har svært ulikt innhold og det er dermed få felles termer, noe som bidrar til at SVM blir dårlig opplært for «Vitenskap og teknologi». Termene brukt i disse artiklene forekommer oftere i kategorier som «Kultur og underholdning» og «Politikk» og det er derfor disse kategoriene som blir tildelt automatisk.

Presisjon har tilsvarende økt for de største kategoriene og minket for de små, se figur 4. Endringene er ikke så markante her som for fullstendigheten, og gjennomsnittlig har presisjonen gått noe ned.



Figur 4 Presisjon Aftenposten september-desember 2007
Resultater av automatisk kategorisering basert på fullt og redusert læringssett
stoppord er fjernet, c-verdi = 5000

Den gjennomsnittlige F-verdien er 0.76 for redusert læringssett mot 0.72 for det fulle.
En test jeg utførte på et utdrag av artikler fra Klassekampen viste det samme, her økte den gjennomsnittlige F-verdi fra 0.47 til 0.59, fullstendigheten økte tilsvarende og presisjonen holdt seg tilnærmet lik. Resultatene fra Klassekampen med det reduserte læringssettet er nærmere beskrevet i kapittel 5.3.1, resultatene med det fullstendige læringssettet ligger i vedlegg C.2.1.
Det er en såpass stor forbedring i resultatet ved bruk av det reduserte læringssettet at jeg valgte å bruke dette videre i forsøkene.

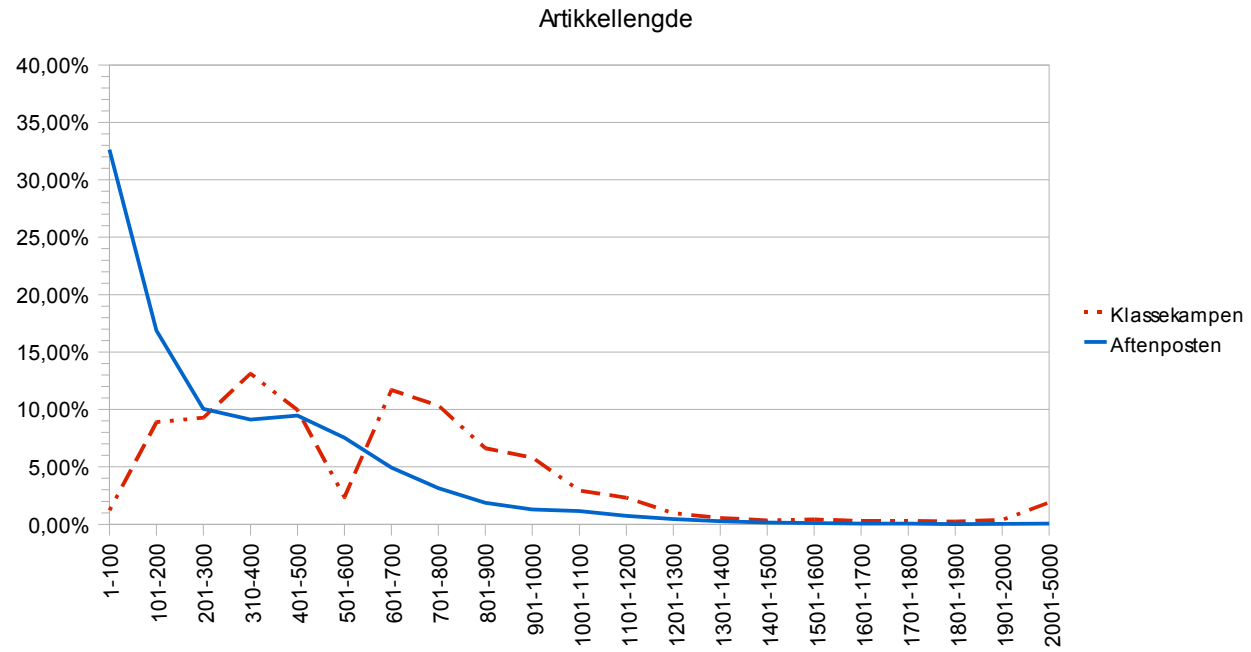
5.1.2 Termer og vekting

Termene er trukket ut fra de artiklene i Aftenposten som utgjør læringssettet. Termene er enkeltord som hentes fra artikkelteksten (XML-taggen 'Story'), spesialtegn[8] og tall fjernes. Med redusert artikkelmengde i læringssettet er antall termer her ca 115 000. 65% av termene forekommer kun en eller to ganger i hele dokumentmengden. 11% har mer enn 10 forekomster.
Til tross for at Joackims konkluderer med at fjerning av stoppord har liten innvirkning på resultatet av kategoriseringen (Joackims 2002 s. 110) har jeg utarbeidet en egen stoppordliste. Den bygger på listene fra http://www.ranks.nl/stopwords/norwegian.html (119 ord) og http://snowball.tartarus.org/algorithms/norwegian/stop.txt (175 ord) og er i tillegg beriket med ikke meningsbærende termer som forekommer ofte i termlista. Tilsammen inneholder den 368 termer, se vedlegg B. Jeg har kjørt tester både med og uten bruk av denne lista og resultatet samsvarer med Joachims konklusjon. Resultatet er likt om stoppordliste brukes eller ikke, med en F-verdi for begge på 0.76. Jeg har valgt å lære opp SVM med bruk av stoppordliste.

Jeg har også kjørt tester der jeg kun bruker de termer som forekommer mer enn 2 ganger i dokumentmengden, da utgjør termlista ca 39000 termer. Resultatet er tilnærmet likt om det kjøres med full eller redusert termmengde. F-verdien er henholdsvis 0.76 og 0.75. I begge tilfeller kjøres med stoppordliste. Resultatene stemmer bra overens med andre forsøk gjort med SVM (Bekkerman et.al. 2003). Fordi Joackims (2002 s. 47) advarer mot å redusere termmengden har jeg derfor valgt å lære opp SVM med å bruke alle termene (unntatt stoppord) .

Lengden på hver enkelt avisartikkel varierer sterkt, og det er stor forskjell mellom Aftenposten og Klassekampen, se figur 5 på neste side.

8 .,:;*$/\&%+–…¦{}§=£][()-?!<>«»"'



Figur 5 fordeling av antall ord pr artikkel
Aftenposten[9] og Klassekampen september-desember 2007

Hele 33% av artiklene i Aftenposten har en artikkellengde på under 100 ord, mens kun 1% av artiklene i Klassekampen er så korte. Gjennomsnittlig artikkellengde i Aftenposten er på 307 ord, i Klassekampen er den 617 ord. Fordi lengden varier sterkt både innenfor hver avis og avisene i mellom, er vektingsalgoritmen som normaliserer for dokumentlengde, brukt.

5.2 Erfaring fra manuell kategorisering

Artiklene i Klassekampen har ikke påført kategorier manuelt. Derfor måtte det trekkes ut et utvalg av artikler som ble kategoriseres manuelt, og som kunne utgjøre testsettet. Fordi det er ressurskrevende å kategorisere, måtte utvalget begrenses mest mulig. Jeg tok derfor utgangspunkt i den kategorien artiklene fikk ved automatisk kategoriseringen og trakk ut et redusert datasett med 20 artikler innenfor hver kategori. Ved å gjøre dette, sikret jeg at alle kategorier ble vurdert. Et rent tilfeldig utvalg, uavhengig av kategori, ville lett, med den skjeve fordelingen som var mellom kategoriene, medført at små kategorier ble for dårlig representert. For månedene september-desember 2007 var kun 10 artikler blitt tilordnet kategorien 'Ulykker og katastrofer» så derfor består testsettet totalt av 310 artikler. En utfordring med denne måten å foreta utvalget på, er at sjøl om

9 Artikler som ikke er tilordnet noen kategori er utelatt

fordelingen er jamn utfra en automatisk kategorisering, blir det skjev fordeling når disse artiklene kategoriseres manuelt. For eksempel er kun tre artikler manuelt tildelt kategorien «Ulykker og katastrofer», noe som er litt for lite grunnlag å trekke slutninger fra. Andre kategorier er tilordna mellom 15-25 artikler, mens «Politikk» skiller seg ut med 43. Resultatet av den manuelle kategoriseringen ligger i vedlegg C.2.2. Jeg tok også et uttrekk for mars 2008, dette skulle brukes til å få svar på om effekten av kategoriseringen går ned over tid. Her var det flere kategorier med få artikler, så utvalget besto av kun 237 artikler totalt, se vedlegg C.4.1.

Hver artikkel ble tildelt kategori manuelt av to personer. Jeg var den ene og jeg kategoriserte alle artiklene i begge utvalg. To andre personer, en norsklærer med erfaring også innenfor bibliografisk arbeid, og en IT-spesialist med god kjennskap til søking på Verdensveven, delte artikkelmengden mellom seg og kategoriserte halvparten hver. Som utgangspunkt hadde vi at vi skulle tilordne artiklene til kategorier som vi ville finne nyttige når artiklene skulle søkes opp igjen på nettet. Dette var ikke alltid enkelt siden avisartikler ofte handler om sammensatte emner. Allikevel holdt vi oss til å tilordne kun en kategori. I de tilfeller vi ikke var enig, valgte vi en som hovedkategori og den andre som alternativ (og i de fleste tilfeller var dette den kategorien den andre personen ville ha påført hvis hun/han skulle påført en til).

Kategoriene brukt i Aftenposten er begrepsmessig ikke entydig definert og de blir derfor tildels overlappende med hverandre. «Politikk» er for eksempel overlappende med de fleste andre kategorier, og en kategori som «Sosiale forhold» kan forstås forskjellig fra person til person. I en avis som Klassekampen, har de aller fleste artiklene en forholdsvis tydelig politisk vinkling. Kategorien politikk kunne derfor godt vært tilordnet de fleste artikler. Artikler om Palestina og Midtøsten er et eksempel på emne som både kan tilordnes «Politikk» og «Krig, forsvar og Konflikter». Et annet emne er EU's tjenestedirektiv som også kan tilordnes «Arbeidsliv». Før jeg startet forventet jeg å finne få artikler i Klassekampen om sport og religion. Dette stemte for sport, det er knapt en ren sportsartikkel blant artiklene vi kategorisert. Religion var det overraskende mange artikler om, både knyttet til islam og til kristendommen. Mye av dette hadde en politisk vinkling, som for eksempel bruk av hijab i skolen og homofile prester i kirken.

I artikkelutvalget på 310 artikler var vi enig i kategoriseringen av 2/3 av artiklene. Noe av uenigheten gikk på emnene Midtøsten og EU's tjenestedirektiv nevnt over, men også emner som AFP var gjenganger. Årsaken til uenigheten skyldes både forskjellig tolkning av teksten og hva vi vektla mest, men også uklarhet på hvordan kategoriene skulle tolkes. Det siste gjaldt også artikler

om sjakk og brigde, her var innholdet helt entydig og det var ikke rom for tolkning, det var kun uenighet i bruk av kategoriene som spilte inn. Dette førte til at sjakk og bridgeartikler ble kategorisert både under «Fritid» og «Kultur og underholdning». Det var ingen merkbar forskjell på kvaliteten på kategoriseringen oss i mellom.

I artikkelutvalget fra mars 2008, som inneholder 237 artikler, måtte 4 artikler utgå fordi de inneholdt feil og var så tvetydige at de var helt umulige å tilordne til en kategori. Vi var enig i kategoriseringen av 161 artikler, noe som er litt i overkant av 2/3 av utvalget. Siden vi hadde snakket sammen om kategoriseringen av artiklene fra september-desember 2007, var uenigheten denne gangen lite relatert til kategorien «Politikk», men mer til «Sosiale forhold». Likestilling var et emne der artiklene litt vekselvis ble kategorisert under «Sosial forhold» og »Arbeidsliv». Vi hadde også problemer med petitartikler og dikt. Tildels ble disse artiklene tilordnet de kategoriene de omhandlet mest, tildels ble de tilordnet «Kultur og underholdning» uansett emne. Resultat av kategoriseringen ligger i vedlegg C.4.2.

For begge utvalgene oppnådde vi over 65% konsistens i kategorisengen. Dette må ses på som bra utfra at resultatene i andre undersøkelser ofte ligger under 50% og bare unntaksvis så høyt som 75%. (Borko 1964, Uren 2000). Sett i forhold til de som kategoriserer artiklene i Aftenposten hadde vi både mindre erfaring og opplæring da vi startet. Dette har spesielt gitt seg utslag i usikkerheten vi hadde med å tolke hvordan kategoriene skulle brukes. For å utelukke denne usikkerheten burde artiklene i Klassekampen vært kategorisert av de samme personene som gjør dette i Aftenposten, eller i det minste burde vi ha gått gjennom samme opplæring som dem.

Artiklene vi kategoriserte manuelt ble sammenliknet med den automatiske kategoriseringen. For de artiklene der vi ikke var enig i kategoriseringen, valgte vi det ene alternativet som hovedkategori og det andre som alternativ. I enkelte av forsøkene ble så resultatet av den automatiske kategoriseringen sammenliknet med både hovedkategori og med hovedkategori/alternativ, se de enkelte forsøk beskrevet under.

5.3 Resultat av kategorisering når læring og test kommer fra forskjellig kilde

Dette forsøket skal gi svar på det første forskningsspørsmålet nevnt i innledningskapittelet:

- Vil kvaliteten av den automatisk kategorisering være like god på en dokumentsamling der kategoriseringsverktøyet er lært opp ved å bruke dokumenter fra en annen samling, som når dokumentene en bruker til læring kommer fra samme kilde som de som skal kategoriseres?

For å teste ut dette brukte jeg artikler fra Klassekampen i perioden september-desember 2007. Dette er samme periode som læringssettet og testsettet fra Aftenposten. Jeg hentet ut artiklene fra A-tekst i rent tekstformat og skrev program for å danne vektormatrisen som brukes av SVM, se vedlegg D. Som for Aftenposten brukte jeg hele teksten inkludert ingress, i Klassekampen er den adskilt fra teksten med en blank linje. Totalt var det 3033 artikler. Disse ble tilordnet kategorier automatisk med det reduserte læringssettet, der ingen kategori har mer enn 801 artikler.

Resultatene fra den automatiske kategoriseringen ble så sammenliknet med de manuelt påførte kategoriene. For dette forsøket brukte jeg uttrekket på 310 artikler hentet fra september-desember 2007. Resultatene ble sammenliknet både med den manuelt satte hovedkategorien og med hovedkategori/alternativ. Dette fordi det var interessant å se om resultatet ble mye forbedret ved å ha to alternativer for kategori. Avisartikler omhandler ofte mer enn et emne, og derfor kan det være vanskelig å bestemme kun en kategori som er riktig.

Eksterne skribenter bidrar med en stor andel av artiklene i Klassekampen, blant annet er alltid minst 2-3 sider av avisas 24 sider fylt med leserinnlegg. Jeg har sett på om det er noen forskjeller avhengig av skribentens tilhørighet. Et lite antall artikler i Klassekampen er på nynorsk, disse er vurdert spesielt for å se om språk har en stor betydning for effekten av den automatiske kategoriseringen.

5.3.1 Resultatet av den automatiske kategoriseringen for Klassekampen

I tabell 8 på neste side vises resultater når den automatiske kategoriseringen sammenliknes med den manuelt satte hovedkategorien.

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0.6000	0.4500	0.5143
Ulykker og naturkatastrofer	1.0000	0.3000	0.4615
Sport	1.0000	0.4000	0.5714
Forsvar og krig og konflikter	0.5484	0.8500	0.6667
Politikk	0.3256	0.7000	0.4444
Sosiale forhold	0.4400	0.5500	0.4889
Personalia	0.4091	0.4500	0.4286
Kriminalitet og rettsvesen	0.6000	0.3000	0.4000
Vitenskap og teknologi	0.6923	0.4500	0.5455
Kultur og underholdning	0.4167	0.5000	0.4545
Medisin og helse	0.8421	0.8000	0.8205
Utdanning	0.7273	0.8000	0.7619
Fritid	0.8235	0.7000	0.7568
Arbeidsliv	0.6429	0.4500	0.5294
Natur og miljø	0.7619	0.8000	0.7805
Religion og livssyn	0.7826	0.9000	0.8372
gjennomsnitt	0.6633	0.5875	0.6231

Tabell 8[10] Fullstendighet, presisjon og F-verdi. Klassekampen september-desember 2007
Redusert læringssett er brukt for opplæring, testsett med 310 artikler
stoppord er fjernet, c-verdi = 5000

Den gode fullstendigheten for «Ulykker og naturkatastrofer» og «Sport» skyldes det lave antall artikler som er tilordnet disse kategoriene manuelt, henholdsvis tre og åtte artikler. Fullstendigheten på 1.0 betyr at alle disse artiklene fikk riktig tilordnet kategori automatisk.

125 av de 310 artiklene ble feil kategorisert. Ved en gjennomgang viser det seg at for ca 40% av artiklene er den automatisk satte kategorien akseptabel sjøl om den ikke er den beste. Spesielt gjelder dette for kategoriene «Natur og miljø», «Politikk» og «Sosiale forhold». Vurderes de feilaktig kategoriserte artiklene ut fra den manuelt satte kategorien, viser det seg at av de 29 artiklene som manuelt var tilordnet «Politikk» har 72% en akseptabelt satt kategori automatisk. Dette betyr nok at vi har brukt denne kategorien for ofte manuelt, noe som samsvarer med det jeg skrev i kapittel 5.2 om at de aller fleste artikler i Klassekampen kan vurderes som politiske.

Resultatene ved bruk av både hovedkategori og alternativ, gir et merkbart bedre resultat enn ved

10 Grunnlagsdata for beregning av resultatet ligger i vedlegg C.2.3

kun å bruke hovedkategori. Både fullstendighet og presisjon har økt og den gjennomsnittlige F-verdien har økt til 0.65. Mer informasjon om disse resultatene ligger i vedlegg C.2.4.
Enda bedre blir resultatet hvis også artiklene som automatisk var tildelt en akseptabelt, men ikke den beste kategorien automatisk, blir vurdert som riktige. Da er F-verdien så høy som 0.78.

Dikt og petitartiklene, som var vanskelig å kategorisering automatisk og manuelt, ga ikke utslag på resultatene, Til det var antallet for lite.

Skribentens tilhørighet

Antallet forskjellige skribenter er høyt i Klassekampen, i perioden september-desember 2007 var det så mange som 500, og av disse var kun ca 10% tilknyttet avisa[11]. Så mye som ¾ av de eksterne skribentene har kun skrevet en artikkel i fire måneders perioden. Gitt den store mengde leserinnlegg som er i avisa hver dag, er det stor sannsynlighet for at en skribent som skriver kun en artikkel, skriver et leserinnlegg. Eksterne som skriver hyppigere, er blant annet ansvarlig for ukentlige artikler og petitartikler, og flere er ivrige debattanter på debattsidene. I uttrekket på 310 artikler er 115 av artiklene skrevet av eksterne. Resultatene av kategoriseringen viser en gjennomsnittlig F-verdi lik, 0.59, for begge grupper. Resultatet er ikke overraskende siden mange av de eksterne som skriver i avisa, er skriveføre folk, som har lang erfaring i å skrive innlegg og artikler i aviser. Det må bemerkes at antall artikler spesielt for de små kategoriene er lavt, og derfor har stor usikkerhet. Uansett, tendensen er klar i at det ikke er stor forskjell på kategoriseringen avhengig av skribentenes tilhørighet. Detaljerte resultater for begge grupper ligger i vedleggene C.2.5 og C.2.6 .

Språkform

Av utvalget på 310 artikler var det 36 på nynorsk. Dette utgjør ca 12% av utvalget, altså en noe mindre andel enn nynorskandelen av alle artikler fra september-desember 2007. Fordi utvalget var så lite som 36, var flere av kategoriene kun brukt for en artikkel og to av kategoriene hadde verken artikler tilordnet manuelt eller automatisk. Derfor kontrollerte jeg den automatisk satte kategorien for 100 andre nynorskartikler. Disse ble tilfeldig trukket ut fra perioden september-desember 2007. Også her er det noen kategorier med få artikler og fire av kategoriene har ingen. Det er derimot så mange som 31 artikler tilknytta «Kultur og underholdning». Fordelingen er så skjev at det ikke er aktuelt å beregne gjennomsnittsverdier for hele utvalget. For de 8 kategoriene som har 5 eller flere artikler, er gjennomsnittlig F-verdi på 0.62. Flere detaljer om resultatet ligger i vedlegg C.2.7.

Fordi nynorskartiklene var såpass få og hadde en så skjev fordeling, er det vanskelig å komme med noen klar konklusjon på resultatene av denne testen. Det er sannsynlig at resultatet hadde blitt bedre

11 Identifisert ved at de har «klassekampen» som del av domenenavnet i epost-adressen

hvis stemming var brukt. Mye av forskjellen mellom nynorsk og bokmål ligger i bøyningsmønsteret. Siden SVM er lært opp på bokmålstekster er ord med nynorske bøyninger stort sett fraværende i termlista som brukes for kategoriseringen.

5.3.2 Sammenlikning Aftenposten-Klassekampen

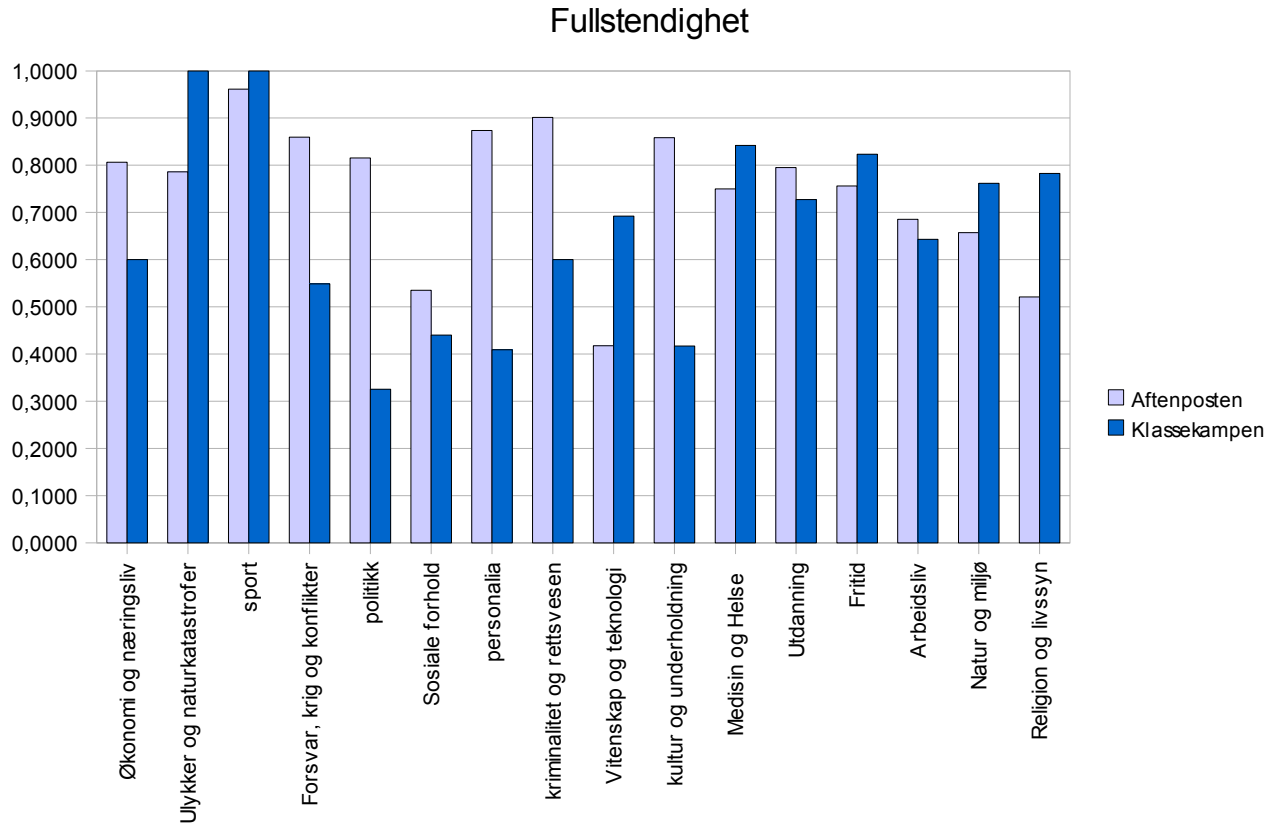
Ikke overraskende viser forsøkene at den automatiske kategoriseringen gir bedre resultater for Aftenposten enn Klassekampen. Dette var ventet fordi SVM jo er lært opp med artikler fra Aftenposten, fra samme tidsperiode som testsettet er tatt fra. Sammenlikningen mellom avisene er gjort med resultatene fra Klassekampen der den automatiske kategoriseringen er sammenliknet med kun hovedkategori. Årsaken til at alternativ kategori ikke er benyttet, er fordi artiklene i testsettet fra Aftenposten kun har manuelt tilordnet en kategori.

Gjennomsnittsverdiene for Klassekampen er usikre siden enkelte av kategoriene finnes i få artikler. Som skrevet i kapittel 4.2.2, vil kategorisering av en enkelt artikkel da kunne få stor konsekvens for denne kategoriens resultater og dermed stor betydning for gjennomsnittsverdien.
De 4 kategoriene «Ulykker og naturkatastrofer», «Sport», «Kriminalitet og rettsvesen» og «Vitenskap og teknologi» er alle manuelt blitt knyttet til færre enn 15 artikler. Om disse utelates fra gjennomsnittsberegningen blir resultatene noe bedre. Fullstendigheten går ned, mens presisjon går betydelig opp, se siste rad i tabell 9 under.

	Fullstendighet	Presisjon	F-verdi
Aftenposten	0.7487	0.7800	0.7592
Klassekampen	0.6633	0.5875	0.6231
Klassekampen – kun 12 kategorier	0.6100	0.6624	0.6352

Tabell 9 fullstendighet, presisjon og F-verdi
Aftenposten og Klassekampen september-desember 2007
stoppord er fjernet, c-verdi = 5000

Figurene 6 og 7 viser fullstendighet og presisjon fordelt på kategori for hver av de to avisene.



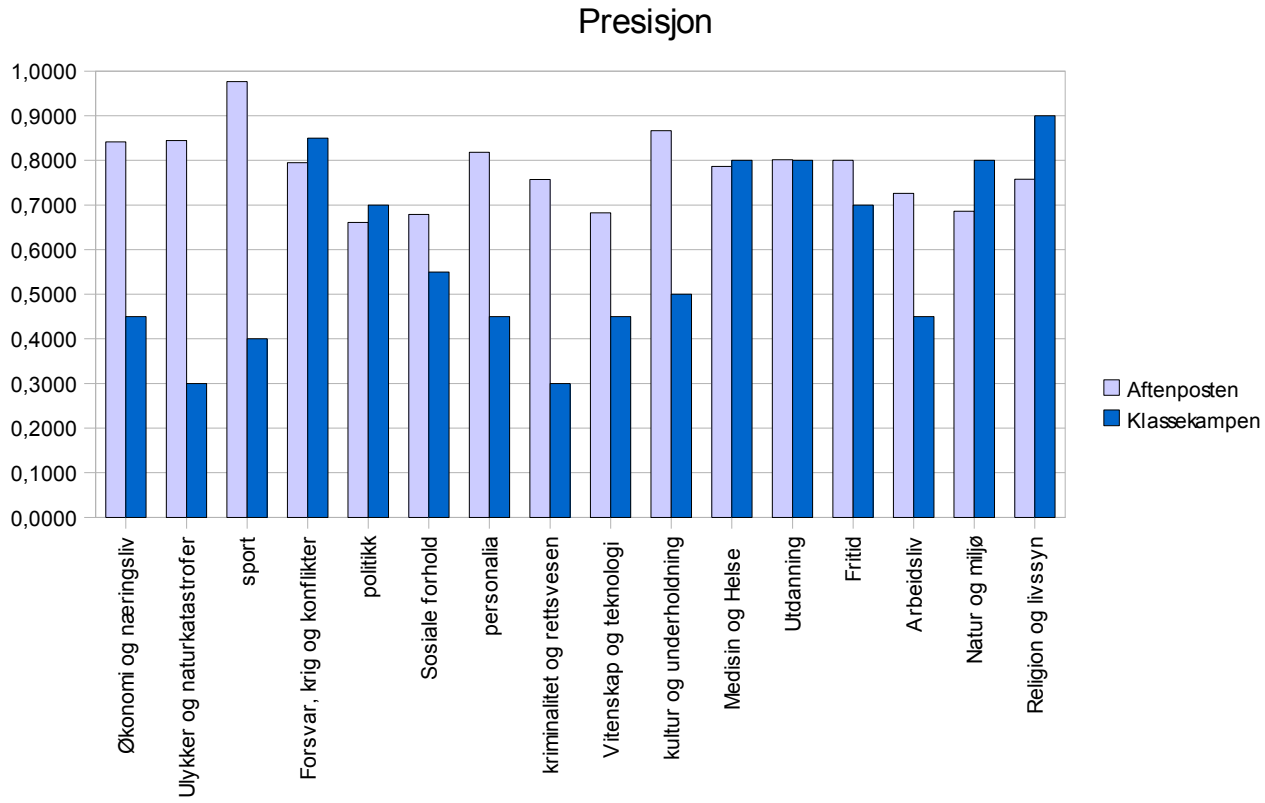
figur 6 Fullstendighet
sammenlikning Aftenposten og Klassekampen september-desember 2007

For enkelte kategorier er resultatet i Klassekampen betraktelig dårligere enn for Aftenposten. Størst forskjell har kategoriene «Personalia», «Politikk» og «Kultur og underholdning». Artikler som manuelt er kategorisert som «Personalia» i Klassekampen, er automatisk tilordnet det emnet den omtalte personen er kjent for. At dette ikke skjer like ofte for Aftenpostenartiklene skyldes sannsynligvis at artiklene her oftere omhandler personer som også er omtalt i læringssettets artikler. Artiklene fra Klassekampen som har fått denne kategorien satt manuelt, inngår i flere tilfeller i en serie som omtaler dagens dato. Her er det informasjon om historiske hendinger eller jubileer, for eksempel om flygeren Whitehead, dikteren August Wilhelm von Schlegel og juristen Anita Augspurg.

For kategorien «Politikk» er det ikke en enkelt årsak som kan forklare den dårlige kategoriseringen i Klassekampen. Som jeg skrev i kapittel 5.2, var «Politikk» en av de kategoriene vi fant vanskelig å bruke siden den er så altomfattende. Det er sannsynlige at årsaken til det bedre resultat i Aftenposten er at læring og testsettet er kategorisert manuelt av de samme personene.

Det dårlige resultatet for «Kultur og underholdning» har heller ikke én enkelt forklaring. Mens 72% av artiklene som manuelt var kategorisert under «Politikk», hadde en akseptabelt satt kategori automatisk, var tilsvarende tall for «Kultur og underholdning» kun 29%. De resterende artiklene var for eksempel to artikler om mulig nedlegging av bokbåten som automatisk har blitt kategorisert under «Ulykker og naturkatastrofer» og en omtale av en Riverton-biografi som var kategorisert under «Sport». Årsaken er mest sannsynlig at SVM ikke er opplært med artikler fra Klassekampen, og at ordvalget i Klassekampens artikler ikke er typisk for ordvalget i tilsvarende artikler i Aftenposten. Kategoriene «Vitenskap og teknologi» og «Religion» har gitt et bedre resultat i Klassekampen. Årsaken her er like mye at resultatet i Aftenposten er dårlig som at Klassekampens er så bra.

Presisjonen viser samme tendenser, se figur 7.



Figur 7 Presisjon
sammenlikning Aftenposten og Klassekampen september-desember 2007

Artiklene om bokbåten bidrar også til at presisjonen for «Ulykker og naturkatastrofer» blir svært dårlige. Andre artikler plasseres feilaktig under denne kategorien fordi artiklene er forholdsvis korte og har termer som sterkt er knyttet til kategorien, som skyting og ulykke. Skyting er brukt i en artikkel om miljøkriminalitet og skadeskyting av dyr, ulykker i en biografisk artikkel om Grace

Keller. Dette er det lite å gjøre med all den stund den automatiske kategoriseriene baserer seg på representasjonen av termene i dokumentet.
Det er også en stor forskjell i presisjonen for kategorien «Sport». I Aftenposten har denne kategorien svært god både fullstendighet og presisjon. I Klassekampen er mange artikler som manuelt er kategorisert andre steder, automatisk plassert her. Det er ingen innlysende årsak til dette, så også her er det sannsynlig at den automatiske innplasseringen er gjort fordi artiklene er forholdsvis korte og samtidig inneholder termer som er mye brukt i sportsartikler, for eksempel rekord, poeng og plass. Så lenge enkeltord er brukt som termer er det lite en kan få gjort med dette. Hadde en i tillegg tatt hensyn til termenes naboer, ordene før og etter, kunne resultatene blitt bedre (Haaland 2008) . Da ville for eksempel *første plass* indikerer at artikkel handler om sport, mens *Alexander Kiellands plass* kunne indikerer at det var en artikkel om kultur.

5.4 Resultat av kategorisering når læring og kategorisering er adskilt i tid

Det andre spørsmålet som skulle undersøkes, var om kvaliteten på kategoriseringsmetoden ville holde seg over tid, eller om effekten reduseres etter som det blir større avstand mellom tidspunkt for opplæring og tidspunktet for kategoriseringen. Her brukte jeg både artikler fra Aftenposten og Klassekampen for mars 2008. Artiklene var på samme formater og ble behandlet på samme måte som i forrige forsøk. Artiklene fra Aftenposten ble kategorisert automatisk og sammenliknet med resultater fra Aftenposten i september-desember 2007. Det samme ble gjort for Klassekampen.

5.4.1. Resultater fra kategorisering av Aftenposten mars 2008

I Aftenpostens morgenutgave for mars 2008 var det 5573 artikler, av disse var 1044 uten kategori. Artikler med mer enn en kategori ble fjernet og datasettet som ble automatisk kategorisert besto da av 4188 artikler.

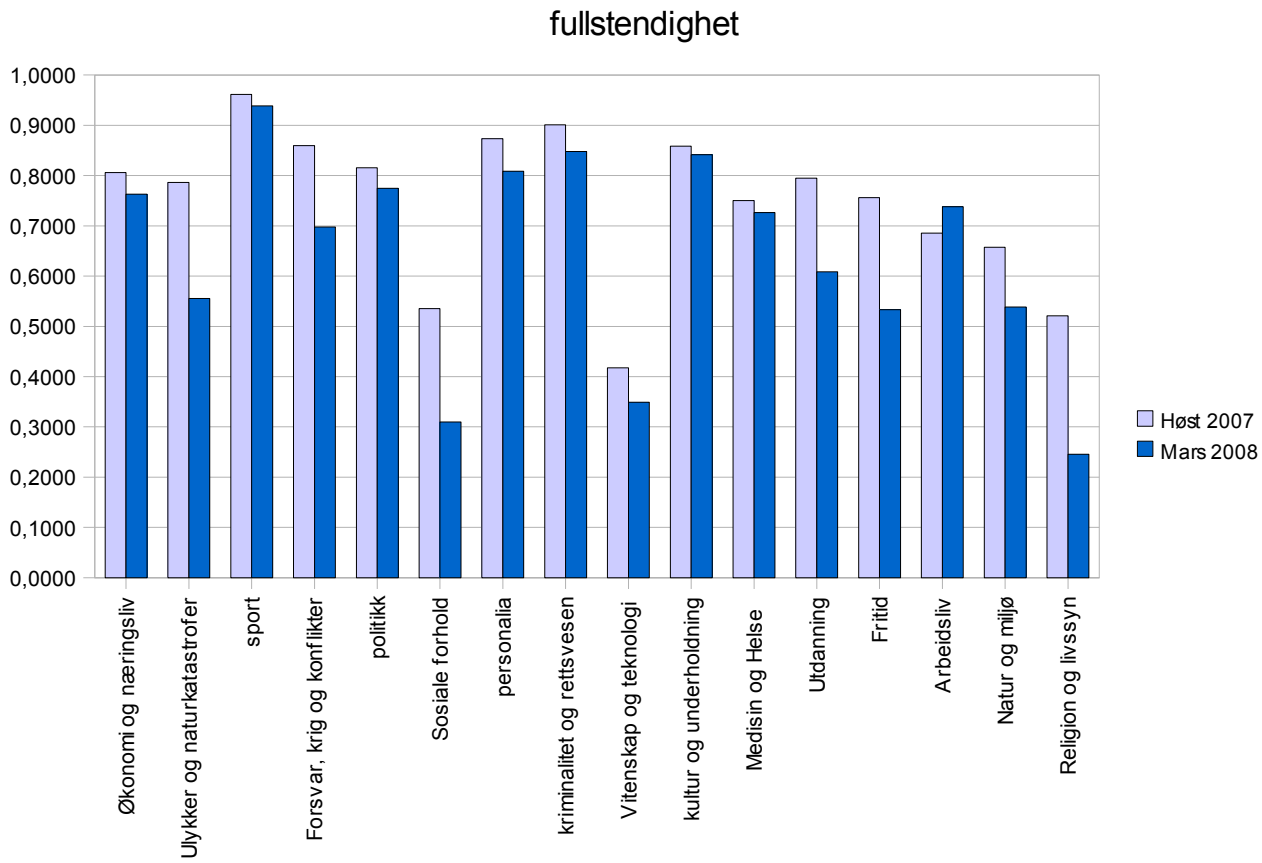
Resultatene av kategoriseringen er dårligere enn resultatene fra september-desember 2007. Den gjennomsnittlige F-verdien er redusert fra 0.76 til nå 0.66. Det samme gjelder for både gjennomsnittlig fullstendighet og presisjon, som nå er 0.64 og 0.70 mot tidligere 0.75 og 0.78.
Se tabell 10 på neste side.

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0.7631	0.7867	0.7747
Ulykker og naturkatastrofer	0.5556	0.8152	0.6608
Sport	0.9386	0.9622	0.9503
Forsvar og krig og konflikter	0.6978	0.6978	0.9678
Politikk	0.7745	0.5233	0.6246
Sosiale forhold	0.3095	0.5652	0.4000
Personalia	0.8087	0.6578	0.7255
Kriminalitet og rettsvesen	0.8478	0.6424	0.7309
Vitenskap og teknologi	0.3492	0.5946	0.4400
Kultur og underholdning	0.8415	0.8354	0.8384
Medisin og helse	0.7262	0.8472	0.7821
Utdanning	0.6087	0.8046	0.6931
Fritid	0.5333	0.6780	0.5970
Arbeidsliv	0.7383	0.6991	0.7182
Natur og miljø	0.5385	0.6125	0.5731
Religion og livssyn	0.2456	0.5185	0.3333
gjennomsnitt	0.6423	0.7025	0.6587

Tabell 10[12] Fullstendighet,presisjon og F-verdi. Aftenposten mars 2008
Redusert læringssett er brukt for opplæring, testsett med 4188 artikler
stoppord er fjernet, c-verdi = 5000

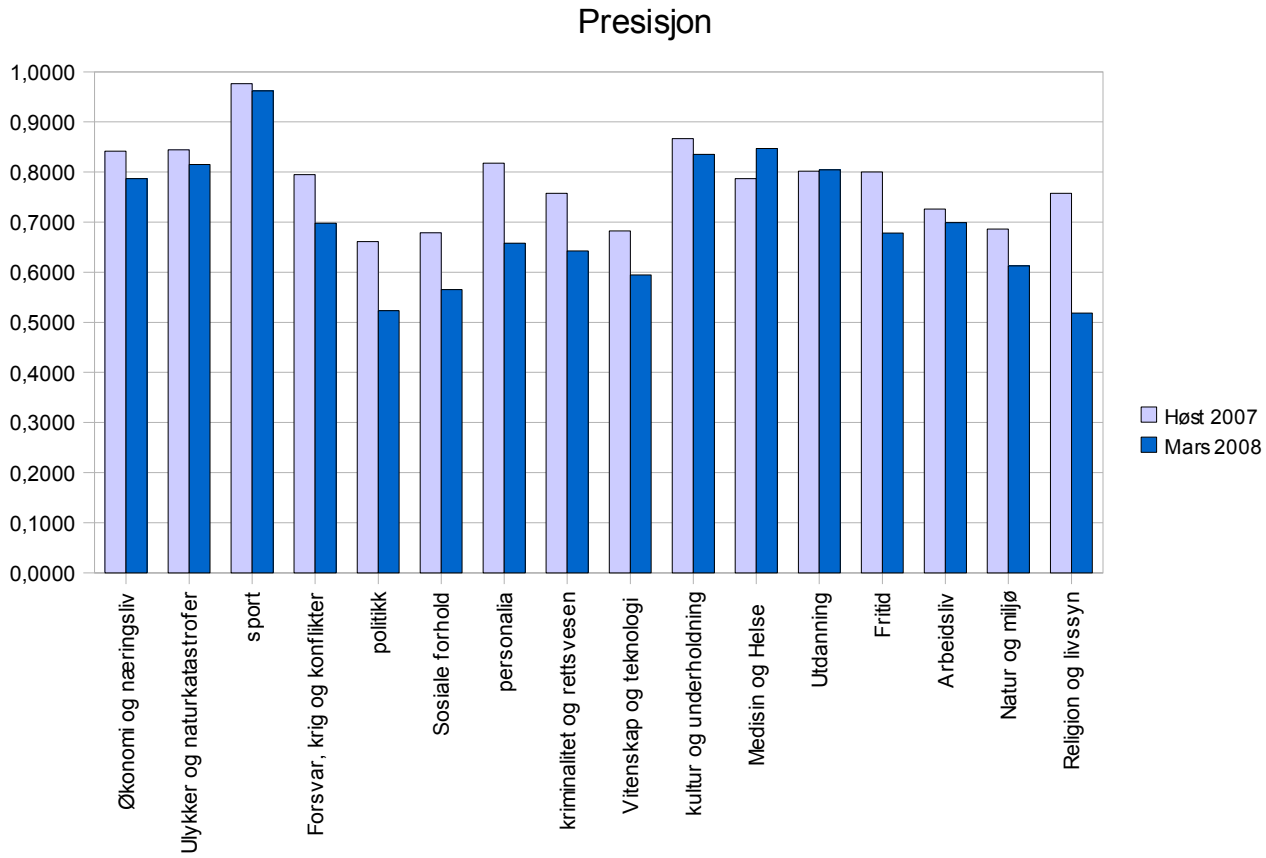
Nedgangen er forholdsvis lik for alle kategorier. Alle kategorier, med unntak av «Arbeidsliv», har fått redusert fullstendighet, se figur 8 på neste side. Den største reduksjonen er for «Religion og livssyn». 2/3 av artiklene som manuelt er kategorisert her, er automatisk kategorisert under «Politikk» og «Kultur og underholdning». En årsak til dette kan være at «Religion og livssyn» er en liten kategori og kun 71 artikler ble brukt ved opplæringen, se tabell 4. Dessuten omhandler flere artikler emner som er i grenselandet med «Politikk» så den automatiske kategoriseringen er sånn sett ikke helt feil. For de store kategorier er fullstendigheten tilnærmet lik resultatene fra høsten 2007.

12 Grunnlagsdata for beregning av resultatene ligger i vedlegg C.3.1



Figur 8 Fullstendighet
Aftenposten høst 2007 og mars 2008

Presisjonen viser samme tendens, se figur 9 på neste side. For de fleste kategorier er reduksjonen liten, og igjen størst for Religion og Livssyn. 11 av 27 artikler som automatisk var kategorisert her, var manuelt plassert under «Politikk» og de fleste av disse var artikler som debatterte pristildelingen til Mohammad Usman Rana i Aftenpostens kronikkkonkurranse. Debattinnleggene var jamt over korte og alle inneholdt ord som Mohammad, muslimer og andre termer som er mye brukt i artikler om religion.



Figur 9 Presisjon
Aftenposten høst 2007 og mars 2008

5.4.2. Resultater fra kategorisering av Klassekampen mars 2008

Klassekampen hadde 787 artikler i mars 2008. Et utvalg av disse ble trukket ut og behandlet på samme måte som for utvalget fra september-desember 2007, se kapittel 5.2 Utvalget var på kun 233 artikler fordi flere kategorier hadde færre enn 20 artikler tilordnet automatisk, se vedlegg C.4.1. Resultatet etter sammenlikningen med den manuelt satte hovedkategorien, vises i tabell 11 på neste side.

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0.7333	0.5789	0.6471
Ulykker og naturkatastrofer	1.0000	0.3333	0.5000
Sport	1.0000	0.5000	0.6667
Forsvar og krig og konflikter	0.7273	0.8000	0.7619
Politikk	0.5238	0.5500	0.5366
Sosiale forhold	0.6071	0.8500	0.7083
Personalia	0.5000	0.4286	0.4615
Kriminalitet og rettsvesen	0.9000	0.4500	0.6000
Vitenskap og teknologi	0.2500	0.2857	0.2667
Kultur og underholdning	0.5417	0.6842	0.6047
Medisin og helse	0.8889	0.8889	0.8889
Utdanning	0.7143	0.9091	0.8000
Fritid	0.9000	0.7500	0.8182
Arbeidsliv	0.7619	0.8000	0.7805
Natur og miljø	0.9091	0.6667	0.7692
Religion og livsyn	0.5789	0.9167	0.7097
Gjennomsnitt	0.7210	0.6495	0.6834

Tabell 11[13] Fullstendighet og presisjon Klassekampen mars 2008
Redusert læringssett er brukt for opplæring og testsett har 233 artikler
stoppord er fjernet, c-verdi = 5000

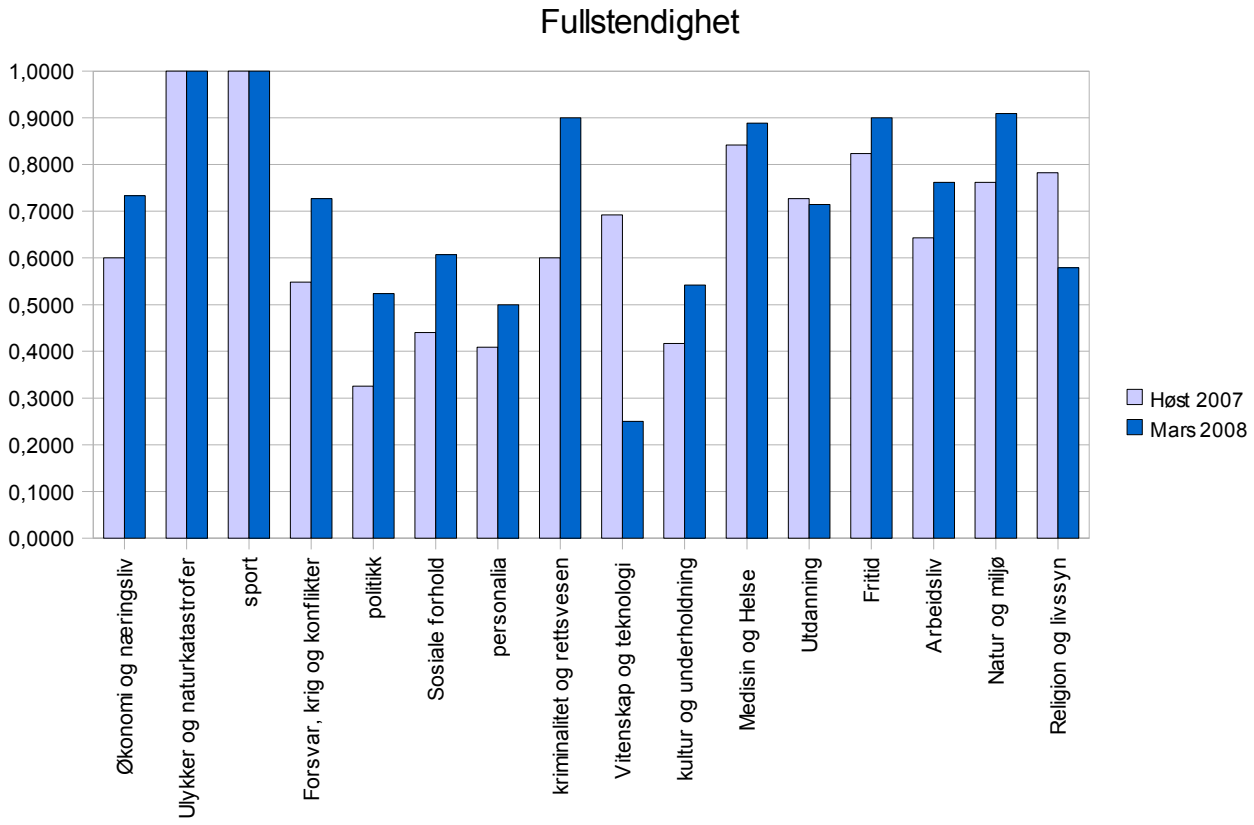
Fordi det er flere kategorier som har få artikler, er gjennomsnittsverdiene usikre, se kapittel 4.2.2. Fjernes kategoriene «Ulykker og naturkatastrofer», «Sport», «Personalia» og «Vitenskap og Teknologi», som alle har færre en 10 artikler, øker gjennomsnittsverdiene, se tabell 12 på neste side. I samme tabell vises tilsvarende resultater fra september-desember 2007.

13 Grunnlagsdata for beregning av resultatene ligger i vedlegg C.4.3

	Fullstendighet	Presisjon	F-verdi
september-desember 2007	0.6633	0.5875	0.6231
september-desember 2007 – 12 kategorier[14]	0.6259	0.6500	0.6377
mars 2008	0.7210	0.6495	0.6834
mars 2008 – 12 kategorier	0.7322	0.7370	0.7346

Tabell 12 fullstendighet, presisjon og F-verdi
Klassekampen september-desember 2007 og mars 2008
stoppord er fjernet, c-verdi = 5000

Resultatet er noe overraskende siden det var forventet at resultatet heller ville gå ned enn opp når det ble avstand i tid til opplæringen. Forskjellene i fullstendighet varierer sterkt mellom kategoriene, se figur 10 under.



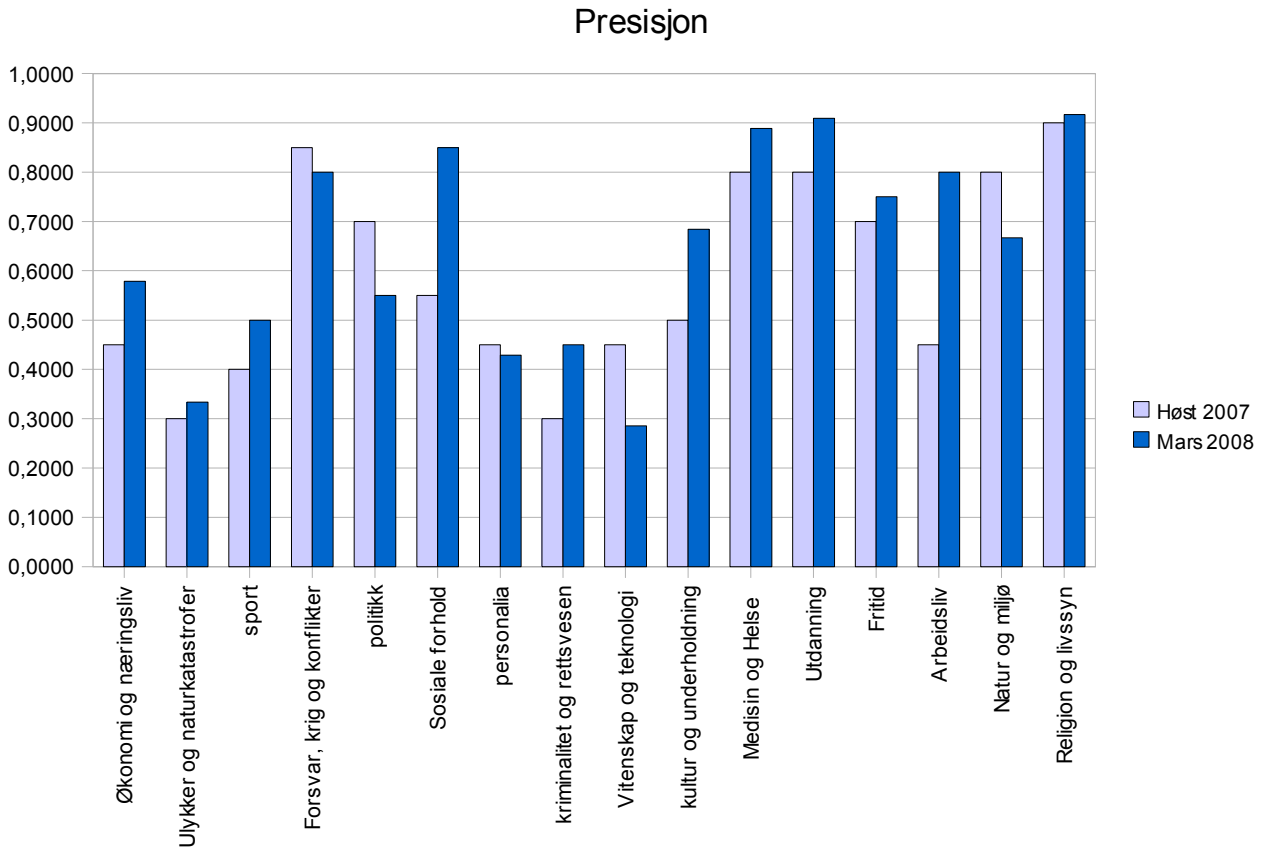
Figur 10 Fullstendighet
Klassekampen september-desember 2007 og mars 2008

For de fleste kategorien har fullstendigheten økt , aller mest for kategorien «Kriminalitet og rettsvesen». Antall artikler som er tildelt denne kategoriene manuelt, er lavt, 10 artikler i begge

14 Vær obs på at dette *ikke* er de samme 12 kategorier som ble brukt i sammenlikningen med Aftenposten i kapittel 5.3.4

perioder. Det dårlige resultatet for september-desember skyldes delvis at 2 av de 4 artiklene som manuelt var kategorisert her, var på nynorsk. I mars var kun 1 artikkel kategorisert feil. Fullstendigheten har sunket mest for «Vitenskap og teknologi», der den nå er så langt nede som 0.25. Årsaken er nok i hovedsak av vi som kategoriserte disse artiklene manuelt var usikker på hvordan kategorien skulle brukes, resultatet øker derfor til 0.75 hvis en tar hensyn også til alternativ kategori. Det må bemerkes at for enkelte av kategoriene er antall artikler svært lavt, sånn at kategoriseringen av hver enkelt artikkel får stor konsekvens for resultatene.

Presisjon viser samme tendens, se figur 11. Resultatene er bedre for 11 av kategoriene, her skiller «Sosiale forhold» seg ut spesielt, presisjonen har økt fra 0.55 til 0.85. Som nevnt før er denne kategorien noe uklar, og dette er grunnen til at presisjonen er såpass lav i september-desember 2007. For hele 6 av de 9 artiklene som automatisk var kategorisert under denne kategorien, men manuelt et annet sted, er «Sosiale forhold» en akseptabel kategori, om enn ikke den beste.



Figur 11 Presisjon
Klassekampen høst 2007 og mars 2008

Det samme kan sies om «Arbeidsliv». For 6 av de 11 artiklene som automatisk var kategorisert her,

var dette en akseptabel kategori, sjøl om det ikke var den som var satt manuelt.
Utover det som kan bemerkes for enkelte av kategoriene, er det ingen åpenbar forklaring på hvorfor resultatene jamt over er så mye bedre for mars 2008. Mest sannsynlig skyldes dette at utvalget av artikler var lite for begge uttrekkene fra Klassekampen og at dette medfører noe tilfeldige resultater. Enkelte kategorier er brukt i svært få kategorier. Dette gjør at kategoriseringen av hver enkeltartikkel får uforholdsmessig stor konsekvens for disse kategorienes resultater.

6 Oppsummering og konklusjon

I innledningskapittelet ble det stilt to spørsmål jeg har prøvd å finne svar på. Det er kjørt flere forsøk, og resultatene fra disse gir ikke entydige svar på alt. Avgjørende for forsøkene var opplæringen av SVM og de valg som ble gjort der. Det viste seg at valg av kategorier og antall artikler i læringssettet for hver av kategoriene var de faktorene som spilte sterkest inn på effekten av kategoriseringsmetoden.

En utfordring med Aftenpostens kategorier, og med IPTC-standarden, er at de er svært omfattende i omfang og tildels overlappende i innhold. Dette viser også resultatene av forsøkene. Kategoriene som er vanskeligst å forstå meningsinnholdet for, er også blant de som gir dårligst resultat. Det ville derfor kunne hjelpe å øke antall kategorier. Med det kan en oppnå at hver kategori får en mer entydig definisjon. Alternativt kan en ta i bruk emnene Aftenposten og IPTC-standarden har definert på neste nivå. Ved manuell kategorisering er disse med på å klargjøre de overordnede kategoriene. For bruk i automatisk kategorisering er imidlertid antall emner for stort. Det ville stille svært store krav til datagrunnlaget som utgjør læringssettet. For alle kategorier/emner bør det være minst 20 artikler. Dette er vanskelig å oppnå med det store antall emner det her er snakk om. I den 4 måneders perioden jeg hentet artiklene fra, var kun 10% av emnene brukt mer enn 20 ganger og hele 15% var kun brukt en gang i perioden. Dette betyr at læringssettet må bygges opp av artikler fra en mye lenger tidsperiode, men også da vil en nok erfare at enkelte emner er for lite brukt.

Sjøl om resultatet nok kunne blitt bedre med å bruke flere kategorier/emner, viser forsøkene mine akseptable resultater med bruk av kategoriene sånn de er definert hos Aftenposten idag. Det er viktig at artiklene som inngår i læringssettet, er gode eksempler for kategorien de representerer. Derfor er det viktig hvordan kvaliteten er på den manuelle kategoriseringen som ligger til grunn. Fordi avisartikler ofte omhandler sammensatte emner, kommer en ikke unna at manuell kategorisering vil blir personavhengig og avhengig av hver enkelts erfaringer og bakgrunn. God opplæring, felles retningslinjer og rutiner og formelle kvalitetgjennomganger vil bidra både til at kvaliteten blir bedre og at kategoriseringen blir mer ensartet mellom de som gjør jobben.

I kapittel 4.2.3 nevner jeg forskjellige alternativer for utvalget av artikler fra Klassekampen. Fordi disse måtte kategoriseres manuelt og jeg hadde begrenset tid til rådighet, valgte jeg å trekke ut et

forholdsvis lite antall artikler, 310 artikler fra høsten 2007. Artiklene var ikke helt tilfeldig trukket ut. Jeg tok hensyn til kategorien som var satt automatisk for å sikre at alle kategorier var representert i utvalget. Forsøkene viser at utvalget med fordel kunne vært større. I den manuelle kategoriseringen av artiklene ble enkelte av kategoriene tilordnet svært få artikler. Dette medførte at feil kategorisering av én artikkel ga store utslag på forsøkets resultater.

Det første spørsmålet jeg forsøkte å finne svar på, var om det var forskjell å kategorisere artikler fra en kilde som ikke var brukt for opplæring av kategoriseringsverktøyet. Jeg lærte opp SVM med artikler fra Aftenposten og kategoriserte artikler fra Klassekampen med denne. Dette ga en gjennomsnittlig F-verdi for Klassekampen på 0.62, adskillig dårligere enn med tilsvarende kategorisering for artikler fra Aftenposten. Her var gjennomsnittlig F-verdi 0.76. Det er ingen enkeltstående årsak til Klassekampens resultater, men flere faktorer har sannsynligvis spilt inn.

Avisartikler omhandler som tidligere nevnt, sammensatte emner. Dette gjør at artiklene ofte passer inn under flere kategorier. I kategoriseringen av resultatene fra Aftenposten og Klassekampen har jeg kun satt en kategori automatisk. Dette har spesielt konsekvenser for artiklene fra Klassekampen, siden mange artikler omhandler flere emner og derfor burde ha hatt mer enn en kategori. Sammenliknes den automatisk satte kategorien i Klassekampen både med hovedkategori og et alternativ, blir den gjennomsnittlige F-verdien 0.68. Artiklene fra Aftenposten har kun en kategori satt manuelt, de forholdsvis få som hadde flere kategorier, ble på et tidlig tidspunkt utelatt fra utvalget. Derfor sammenliknes resultatene fra de to avisene kun for en manuelt satt kategori, noe som altså slår mer uheldig ut for Klassekampen enn Aftenposten. Dette er en faktor jeg kunne gjort noe med hvis jeg var blitt klar over den på et tidlig tidspunkt, ved enten å åpne for flere kategorier i begge aviser, eller ved å utelate de artiklene i Klassekampen som var tvetydige.

Som nevnt er enkelte av kategoriene svært omfattende i omfang, og tildels overlappende i innhold. Dette medfører at de kan forstås forskjellig fra person til person. Artikler i Aftenposten, både de som er brukt for læring, og de som er brukt for uttesting, er tildelt kategori manuelt av den samme gruppe mennesker. Dette er en viktig årsak til det gode resultatet for Aftenposten, og en like viktig årsak til at resultatet for Klassekampen er dårligere.

Artiklene i Klassekampen skiller seg ut fra artiklene i Aftenposten med at de i større grad er skrevet av eksterne bidragsytere og at de i større grad er på nynorsk. Enkelte av artiklene er også dikt og petitartikler. Denne type tekster er så langt jeg kan se, fraværende i Aftenpostens artikkelutvalg.

Om skribent er ekstern eller internt ansatt i Klassekampen har så langt mine forsøk har vist, ingen påvirkning på resultatene. Det har heller ikke dikt og petitartiklene, til det var nok antallet artikler for lite. At artiklene er skrevet på nynorsk ser ut til å spille en større rolle. Hovedårsaken til at nynorskartiklene gir dårlig resultat er at SVM er lært opp på bokmålsartikler. Jeg vil tro bruk av stemmer vil virke positivt inn her, dessverre fikk jeg ikke tid til å prøve ut dette.

Gitt alle disse faktorene er resultatene fra Klassekampen akseptable, til tross for at de er såpass mye dårligere enn for Aftenposten. De har også i seg endel usikkerhet pga det forholdsvis lite antall artikler som ble kategorisert. Hvis de fire kategoriene med færrest antall artikler fjernes, øker F-verdien til 0.64. Ser en på de enkelte kategorier har Aftenposten betraktelig bedre resultat på de kategorier som har flest antall artikler i læringssettet. For små kategorier er resultatet jamnere og for enkelte har også Klassekampen best resultat.

Det andre spørsmålet jeg forsøkte å finne svar på, var om effekten av kategoriseringsmetoden blir mindre med tiden. Her brukte jeg artikler fra mars 2008, både for Aftenposten og Klassekampen. For Aftenposten var resultatene dårligere enn for høsten 2007. F-verdien er redusert fra 0.76 til 0.66. Dette er et forventet resultat siden det nå er et gap i tiden mellom tidspunkt for opplæring og tidspunkt for kategoriseringen. For Klassekampen var resultatet nå merkbart bedre enn for høsten 2007, F-verdiene er økt fra 0.62 til 0.68. Dette er det ingen annen forklaring på enn at utvalget av artikler var lite for begge uttrekkene fra Klassekampen, noe som sannsynligvis har medført at resultatene er noe tilfeldige. Enkelte kategorier er brukt i svært få artikler, noe som medføre at . kategoriseringen av hver enkeltartikkel får uforholdsmessig stor konsekvens for resultatet.

I et foredrag holdt av Lennart Christensen og Hasse Helstrøm (2004) , begge fra danske Infomedia, påpeker de at 100% automatisk kategorisering er en utopi, men at 80% er et mål. Resultatene jeg har i mine forsøk ligger et godt stykke under dette, men det er ikke uoppnåelige gitt at

- antall kategorier økes og de har en entydig definisjon
- artiklene som inngår i læringssettet gjennomgår kvalitetsikring
- det åpnes for mer enn en kategori pr artikkel
- det tas i bruk en stemmer, i hovedsak pga artikler på nynorsk

Det er vel verdt å foreta flere forsøk. Både med å ta hensyn til punktene nevnt over, og også å bruke andre aviser. Klassekampen har en spesiell plass i norsk avisflora, og det vil være interessant å se mer på hvor mye stoffvalg og redaksjonell vinkling av nyhetene har å spille for resultatet. Det er

også interessant å ta et nytt utdrag fra Aftenposten med halvt års mellomrom for å se hvordan effekten reduseres ytterligere med tiden.

7 Etterord

Som et lite forsøk for seg sjøl, kategoriserte jeg herværende oppgave, med bruk av SVM og samme læringssett og parameter som ble brukt i forsøkene i oppgaven. Resultatet ble kategorien «Kultur og underholdning». Om dette er en passende kategori eller om det er en annen som kunne passet bedre ut fra tema og innhold i oppgaven overlater jeg til leseren å finne ut av ;-)

8 Litteraturliste

Alpha, S., Dixon, P. & Liao, C. (s.a) *Feature preparation in text categorization*. Lokalisert 16. april 2008 på Verdensveven: http://www.oracle.com/technology/products/text/pdf/feature_preparation.pdf

Bekkerman,R., El-Yaniv, R., Tishby, N. & Winter, Y. (2003) Distributional word clusters vs. words for text categorization. *Journal of Machine Learning research 3(2003),* 1183-1208. Lokalisert 8. april 2008 på Verdensveven: http://jmlr.csail.mit.edu/papers/volume3/bekkerman03a/bekkerman03a.pdf

Borko, H. (1964) Measuring the reliability of subject classification by men and machines. *American Documentation, October 1964*, 268-273.

Buckley, C. & Salton,G. (1987) *Term weighting approaches in automatic text retrieval.* New York*:* Cornell university. Lokalisert 4. januar 2008 på Verdensveven: http://ecommons.library.cornell.edu/bitstream/1813/6721/1/87-881.pdf

Børke, M.A., Grythe, K.A., Løset, T.K., Mørch-Storstein, O.K. & Vistnes, R. (2005) *Paperprism: TDT4290 kundestyrt prosjekt gruppe 10 Bouvet.* [Trondheim] : Institutt for datateknikk og informasjonsvitenskap, NTNU. Lokalisert 2. april 2008 på Verdensveven: http://www.idi.ntnu.no/emner/tdt4290/Rapporter/2005/PaperPrism.pdf

Cellio, M.J., Hayes, P.J. & Knecht, L.E. (1988) A news story categorization system. I:*Second conference on applied natural language processing (9-17)* Lokalisert 2. mai 2008 på Verdensveven: http://acl.ldc.upenn.edu/A/A88/A88-1002.pdf

Chen, H. & Dumais, S. (2000) Hierachical classification of web content. *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. July 24-28, 2000 Athen, Hellas (s.256-263).* Lokalisert 15. mai 20008 på Verdensveven: http://www.cs.ucdavis.edu/~hchen/paper/sigir00.pdf

Chiang, J., Hao, P. & Tu, Y. (2007) Hierarchically SVM classification based on support vector clustering method and its application to document categorization *Expert systems with applications 33 (2007),* 627-635.

Christensen, L. & Hilstrøm, H.(2004) *Kategorisering nu og i fremtiden: Human vs automatisk kategorisering.* Lokalisert 10. juni 2008 på Verdensveven: http://www.mediegruppendk.dk/InfoMedia.pdf

Crammer, K. & Singer, Y. (2001) On the algorithmic implementation of multiclass kernel-based vector machines *Journal of Machine Learning Research 2 (2001),* 265-292 .Lokalisert 16. april 2008 på Verdensveven: http://www.cis.upenn.edu/~crammer/publications/crammer01a.pdf

Dahl, J.H.B. (2002) *Automatisk kategorisering av e-post til tjenesten «spør biblioteket»*. Oslo:Høgskolen i Oslo, avdeling for journalistikk, bibliotek og informasjonsfag
Diplomoppgave

Dumais, S., Hecherman, D., Platt, J. & Sahami, M.(1998) Inductive learning algorithms and representations for text categorization. *Proceedings of the 1998 ACM CIMK International conferance on information and knowledge managment, November 3-November 7, 1998. Bethesda, Maryland, USA. (s.148-155)* Lokalisert 8.april 2008 på Verdensveven : http://robotics.stanford.edu/users/sahami/papers-dir/cikm98.pdf

Haaland, Å. (2008) *A maximum entropy approach to proper name classification for norwegian.* Oslo: Universitetet i Oslo (Faculty of humanities; 335)
Doktoravhandling

International Business Machines [IBM] (2007) *IBM Classification module***.** Lokalisert 11.april 2008 på Verdensveven.:
http://demos.dfw.ibm.com/on_demand_illustrated/Download/IBM_Demo_IBM_Classification_Module-1-Sep07.pdf

International Business Machines [IBM] (2008) *IBM Classification module***.** Lokalisert 11.juni 2008 på Verdensveven: http://www-306.ibm.com/software/data/enterprise-search/classification/

International Press Telecomunications Council [IPTC] (2008) *The IPTC NewsCodes: Metadata taxonomies for the news industry* Lokalisert 19 mai 2008 på Verdensveven:
http://www.iptc.org/NewsCodes/

Joachims, T. (1998) Making large-scale SVM learning practical. I: C. Burges, B. Schölkopf & A. Smola (ed.) *Advances in kernel methods - Support Vector Learning*, Cambridge: MIT-Press. Lokalisert 3. januar 2008 på Verdensveven:
http://www.cs.cornell.edu/People/tj/publications/joachims_99a.pdf

Joachims,T. (2002) *Learning to classify text using Support Vector Machines: Methods, theory and algorithms.* Boston: Kluwer Academic

Joachims, T. (2007) *Multi-class Support Vector Machine.* Lokalisert 2 mai på Verdensveven:
http://svmlight.joachims.org/

Lesk, M.E & Salton, G.(1968) Computer evaluation of indexing and text processing. I: K Sparck Jones & P. Willett (ed) *Readings in information retrieval*. (s.60-84) San Francisco: Morgan Kaufmann

Lewis, D.D .(1991) *Evaluating text categorization* : lokalisert 14. mai 2008 på Verdensveven: http://www.aclweb.org/anthology-new/H/H91/H91-1061.pdf

Lewis, D. (2004) *Reuters-21578 text categorization test collection*. Lokalisert 2. mai 2008 på Verdensveven: http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt

Lewis, D, .Li, F., Rose, T.G. & Yang, Y. (2004) RCV1 : A new benchmark collection for text categorization research *Journal of Machine Learning research 5,* 361-397. Lokalisert 14.april på verdensveven: http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lewis04a.pdf

Luhn, H.P. (1961) The automatic derivation of information retrieval encodements from machine-readable texts. I: K Sparck Jones & P. Willett (ed) *Readings in information retrieval* (s. 21-24) San Francisco: Morgan Kaufmann

Lyman, P. & Varian, H.R. (2003) How much information 2003? Lokalisert 8. juni 2008 på Verdensveven : http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Manning, C.D, Raghavan, P. & Schütze, H. (2008) *Introduction to information retrieval.* Cambridge University Press. Lokalisert 14. mai 2008 på Verdensveven: http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html

Maria, N. & Silva, M.J. (1999) *NewsSearch: An architecture for information retrieval of online news*. Lisboa: Faculdade de Ciências da Universidade de Lisboa . Lokalisert 14.april 2008 på Verdensveven: http://old.di.fc.ul.pt/sobre/documentos/tech-reports/99-3.pdf

Maria, N. & Silva, M.J. (2000) Theme-based Retrieval of web news. I: D. Suciu & G. Vossen (ed) *The world wide web and databases: Third International workshop WebDB 2000*. Lokalisert 16. juni på Verdensveven: http://www.softlab.ntua.gr/facilities/public/AD/Text%20Categorization/Theme-based%20Retrieval%20of%20Web%20News2.pdf

Mediebedriftenes landsforening (2007) *Opplagstall*. Lokalisert 2. april 2008 på Verdensveven: http://www.mediebedriftene.no/novus/upload/pdf/Lesertall/Opplag07.xls

Mediebedriftenes landsforening (2007) *Lesertall*. Lokalisert 2. april 2008 på Verdensveven: http://www.mediebedriftene.no/novus/upload/pdf/Lesertall/F&M%2008-1.xls

Pedersen, J.A.S (2002) *Automatisk klassifisering ved hjelp av søkemotor tilpasset norsk språk: Fordypningprosjekt høsten 2002*. [Trondheim]: Institutt for Datateknikk og Informasjonsvitenskap, NTNU. Lokalisert 4. januar 2008 på Verdensveven: http://www.idi.ntnu.no/grupper/if/publikasjoner/Topicalyzer.pdf

Myhr, K (2008, 11.mai) Trykkleif-toppen *Dagbladet.no*. Lokalisert 11. mai 2008 på Verdensveven: http://www.dagbladet.no/kultur/2008/05/11/534960.html

Oracle (2005) Classifying Documents in Oracle Text I: Text Application Developer's Guide Lokalisert på verdensveven 10 april 2008:
http://download.oracle.com/docs/cd/B19306_01/text.102/b14217/classify.htm

Porter, M. (2006) *The Porter stemmer algorithm.* Lokalisert 26 mai 2008 på Verdensveven:
http://tartarus.org/~martin/PorterStemmer/

Ringdal, K. (2001) *Enhet og Mangfold: Samfunnsvitenskaplig forskning og kvantitativ metode*. Bergen; Fagbokforlaget

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys, 34(1),* 1–47. Lokalisert 11. november 2007 på Verdensveven:
http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf

Strøm, T.J (2001) *Automatisert tekstkategorisering av artikler fra norske nettaviser.* Oslo: Høgskolen i Oslo, avdeling for journalistikk, bibliotek og informasjonsfag Diplomoppgave

Uren, V (2000) An evaluation of text categorisation errors. O*ne-day workshop on evaluation of information management systems, London 15. september 2000.* Lokalisert 19. mai 2008 på Verdensveven:
http://kmi.open.ac.uk/people/victoria/publications/eval200.pdf

9 Vedlegg

Vedlegg A: Aftenpostens kategori og emneliste

Lista under er mottatt fra Karen Thorshaug i Aftenpostens dokumentasjonssenter (muntlig kommunikasjon 19 mai 2008)

KATEGORIER OG EMNEORD GYLDIGE I A-PORTEN FRA 01.01.2007

KATEGORIER:

- ARBEIDSLIV
- DIVERSE
- FORSVAR KRIG OG KONFLIKTER
- FRITID
- KRIMINALITET OG RETTSVESEN
- KULTUR OG UNDERHOLDNING
- MEDISIN OG HELSE
- NATUR OG MILJØ
- PERSONALIA
- POLITIKK
- RELIGION OG LIVSSYN
- SOSIALE FORHOLD
- SPORT
- ULYKKER OG NATURKATASTROFER
- UTDANNING
- VITENSKAP OG TEKNOLOGI
- ØKONOMI OG NÆRINGSLIV

EMNEORD SORTERT I DEN KATEGORI DE I HOVEDSAK TILHØRER

ARBEIDSLIV

- Arbeidsgiverforeninger
- Arbeidskonflikter
- Arbeidsmarked
- Arbeidsmiljø
- Arbeidstid
- Barnearbeid
- Fagforeninger
- Ferie
- Lønninger
- Offentlig ansatte
- Oppsigelser
- Pensjoner
- Pensjonsalder
- Permisjoner
- Personell
- Sommerjobber
- Sykefravær
- Tariffoppgjør

DIVERSE

- Aprilspøk
- Besøk
- Billetter
- Foreninger
- Forhandlinger
- Informasjon
- Konkurranser
- Langtidsplaner
- Materiell
- Metafysikk
- Møter
- Registre
- Samarbeid
- Statistikk
- Undersøkelser

FORSVAR OG KRIG OG KONFLIKTER

- Annen verdenskrig
- Atomvåpen
- Demonstrasjoner
- Etterretning
- Flyvåpen
- Forsvar
- Fredsstyrker
- Garden
- Heimevern
- Hæren
- Kjemiske våpen
- Krig
- Krigsfanger
- Krigsforbrytelser
- Krigsfunn
- Krigsveteraner
- Marine
- Militæranlegg
- Militærøvelser
- NATO
- Nedrustning
- Sikkerhetspolitikk
- Soldater
- Spionasje
- Terrorisme
- Ubåter
- Verneplikt
- Våpen

FRITID

- Boliger
- Bomiljø
- Camping
- Friluftsliv
- Reisebeskrivelser
- Ski
- Småbåter
- Spill

- Hobbyer
- Hytter
- Jakt
- Leker
- Matoppskrifter
- Sportsfiske
- Tipping
- Turisme
- Turistskip

KRIMINALITET OG RETTSVESEN

- Amnesti
- Bankran
- Bedrageri
- Bøter
- Datakriminalitet
- Dom
- Dommere
- Drap
- Drapsforsøk
- Dødsstraff
- Erstatninger
- Fengsler
- Flukt
- Forfalskning
- Fri rettshjelp
- Havrett
- Heleri
- Hjemmebrenning
- Hærverk
- Ildspåsettelse
- Industrispionasje
- Injurie
- Innsidehandel
- Jurister
- Kapringer
- Kidnapping
- Korrupsjon
- Kriminalitet
- Lover
- Miljøkriminalitet
- Misligheter
- Overvåking
- Politi
- Postran
- Promillekjøring
- Ran
- Rettssaker
- Rettsvesen
- Råkjøring
- Samfunnsstraff
- Seksualforbrytelser
- Skattesnyteri
- Skyting
- Smugling
- Tortur
- Trusler
- Tyveri
- Utvisninger
- Vaktselskaper
- Vold
- Økokrim

KULTUR OG UNDERHOLDNING

- Anmeldelser
- Antikviteter
- Arkeologi
- Arkitektur
- Aviser
- Biblioteker
- Bokanmeldelser
- Dans
- Festivaler
- Film
- Filmanmeldelser
- Flagg
- Folkeminne
- Forfattere
- Fotografering
- Internett
- Kino
- Monumenter
- Museer
- Musikk
- Nasjonaldrakter
- Navn
- Omgangsformer
- Opera
- Orkestre
- Oversettelser
- Plater
- Plateanmeldelser
- Pressebyråer
- Radio
- Sangkor
- SceneanmeldelserSensur
- Skuespillere
- Skulptur

- Konserter
- Konsertanmeldelser
- Kultur
- Kulturminner
- Kultursentre
- Kunst
- Kunstkritikk
- Litteratur
- Media
- Merkedager
- Språk
- Subkultur
- Teater
- Tegneserier
- Tidsskrifter
- Trender
- TV
- TV og radioprogrammer
- Underholdning
- Utstillinger

MEDISIN OG HELSE

- Abort
- Barnløshet
- Blod
- Feilbehandling
- Forgiftning
- Fysioterapi
- Fødsler
- Førstehjelp
- Genetikk
- Helse
- Helsepersonell
- Hjernen
- Idrettsskader
- Kiropraktikk
- Kirurgi
- Kosthold
- Kroppspleie
- Legemidler
- Leger
- Lesevansker
- Naturmedisin
- Næringsmiddelkontroll
- Prevensjon
- Psykiatri
- Psykisk utviklingshemmede
- Rehabilitering
- Rusmidler
- Røyking
- Stress
- Svangerskap
- Sykdommer
- Sykehus
- Søvn
- Tannpleie
- Yrkesskader

NATUR OG MILJØ

- Dyr
- Dyreparker
- Dyrevern
- Elver
- Farlige stoffer
- Fisker
- Forurensning
- Fugler
- Havområder
- Hunder
- Hvaler
- Insekter
- Isbreer
- Kalendre
- Kart
- Kjæledyr
- Klima
- Miljøvern
- Naturreservater
- Parker
- Planter
- Radioaktiv stråling
- Reinsdyr
- Renhold
- Resirkulering
- Rovdyr
- Samer
- Seler
- Skadedyr
- Skalldyr
- Sommertid
- Støy
- Veterinærer
- Været
- Økologi
- Øyer
- Årstider

PERSONALIA

- Gaver
- Kongehus
- Monarki
- Stipendier
- Utmerkelser

POLITIKK

- AKP
- Ap
- Boikott
- Bydeler
- Bypolitikk
- Byråkrati
- Bystyre
- Demokrati
- Diplomati
- Diskriminering
- Distriktspolitikk
- EU
- FN
- Folkeavstemninger
- Fredsbevegelser
- Fredspriser
- Fremmedkontroll
- Frp
- Fylker
- Fylkesting
- Grenser
- Grunnloven
- Høyre
- Høyreekstremisme
- Ipol
- Kommuner
- Kommunestyre
- Konsesjoner
- KrF
- Landsmøter
- Menneskerettigheter
- Opposisjonelle
- Partibarometre
- Partistøtte
- Pass
- Personvern
- Politikere
- Politikk
- Privatisering
- Regjeringen
- RV
- Sametinget
- Samfunnsdebatt
- Småpartier
- Sp
- Statlige virksomheter
- Statsborgerskap
- Statsstøtte
- Stemmerett
- Stortinget
- Stortingsmeldinger
- SV
- Uland
- Upol
- Valg
- Valgkamp
- Venstre
- Ytringsfrihet

RELIGION OG LIVSSYN

- Bibelen
- Etikk
- Frikirker
- Høytider
- Kirken
- Livssyn
- Prester
- Religioner
- Seremonier
- Statskirke

SOSIALE FORHOLD

- Adopsjon
- Arv
- Barn
- Barnebidrag
- Barneomsorg
- Befolkning
- Innvandrere
- Krisesentre
- Kvinner
- Levestandard
- Likestilling
- Menn

- Død
- Ekteskap
- Eldre
- Eldreomsorg
- Enslige
- Familie
- Flyktninger
- Funksjonshemmede
- Helse og sosialsentre
- Homofili
- Husokkupasjon
- Innsamlinger
- Omsorgsarbeid
- Pornografi
- Prostitusjon
- Samboere
- Seksualspørsmål
- Selvmord
- Skilsmisse
- Sosialhjelp
- Sykehjem
- Trygder
- Ungdom

SPORT

- Alpint
- Badminton
- Bandy
- Baseball
- Basketball
- Bedriftsidrett
- Biljard
- Bilsport
- Bob
- Boksing
- Bordtennis
- Bowling
- Brettseiling
- Bryting
- Bueskyting
- Båtsport
- Casting
- Curling
- Doping
- Dykking
- EM
- Fallskjermhopping
- Fekting
- Fjellklatring
- Fotball
- Friidrett
- Golf
- Handicapidrett
- Hanggliding
- Hestesport
- Hopp
- Hundekjøring
- Håndball
- Idrett
- Idrettsanlegg
- Innebandy
- Ishockey
- Kampsport
- Kappgang
- Kombinert
- Kroppsbygging
- Kunstløp
- Landhockey
- Langrenn
- Maraton
- NM
- Olympiader
- Orientering
- Padling
- Rekorder
- Roing
- Rugby
- Sandvolleyball
- Seiling
- Skateboard
- Ski
- Skiorientering
- Skiskyting
- Skyting
- Skøyter
- Sportsgymnastikk
- Squash
- Stup
- Svømming
- Sykling
- Tennis
- Trenere
- Triathlon
- Trim
- Turn
- Vannpolo
- Vannski
- Vektløfting
- VM
- Volleyball

ULYKKER OG NATURKATASTROFER

- Arbeidsulykker
- Branner
- Brannvesen
- Eksplosjoner
- Flom
- Flyulykker
- Jordskjelv
- Hjelp
- Katastrofeberedskap
- Nestenulykker
- Ras
- Savnede
- Sikkerhet
- Skipsulykker
- Snøras
- Sult
- Togulykker
- Trafikkulykker
- Tørke
- Ulykker
- Uvær
- Vannlekkasjer
- Vulkaner

UTDANNING

- Barnehager
- Eksamener
- Folkehøyskoler
- Førskolelærere
- Grunnskoler
- Høyskoler
- Karakterer
- Kurs
- Leirskoler
- Læremidler
- Lærere
- Mobbing
- Privatskoler
- Russ
- Skole
- Skolealder
- Skoleelever
- Skolefag
- Skolefravær
- Skolefritidsordninger
- Spesialundervisning
- Studenter
- Studiefinansiering
- Universiteter
- Utdannelse
- Utenlandsstudenter
- Videregående skoler
- Voksenopplæring
- Yrkesopplæring

VITENSKAP OG TEKNOLOGI

- Astronomi
- Biologi
- Bioteknologi
- Ekspedisjoner
- Filosofi
- Forskning
- Fysikk
- Geografi
- Geologi
- Havforskning
- Historie
- Kjemi
- Matematikk
- Oppfinnelser
- Psykologi
- Romforskning
- Samfunnsforskning
- Satellitter
- Standardisering

ØKONOMI OG NÆRINGSLIV

- Administrasjon
- Aksjer
- Apoteker
- Atomkraft
- Auksjoner
- Avgifter
- Landbruk
- Ligning
- Luftfart
- Lys og varme
- Lån
- Matvarer

- Bensinstasjoner
- Betalingskort
- Biler
- Bokhandel
- Broer
- Budsjetter
- Busser
- Byggevirksomhet
- Bygninger
- Børs
- Data
- Data og elektroindustri
- Design
- Drivstoff
- Drosjer
- Eiendomshandel
- Eiendomsselskaper
- Ekspropriasjoner
- Emballasje
- Energi
- Energiøkonomisering
- Ferger
- Finansnæring
- Fiskerinæring
- Fjærfeavl
- Fly
- Flyplasser
- Flyselskaper
- Fonds
- Forbrukerspørsmål
- Forlag
- Forsikringer
- Frisører
- Fusjoner og oppkjøp
- Førerkort
- Gass
- Glassindustri
- Godstransport
- Grafisk industri
- Grensehandel
- Gruveindustri
- Gull
- Hagebruk
- Handel
- Havner
- Helikoptre
- Hjemmeelektronikk
- Hoteller
- Husdyr
- Håndverk
- Inkasso
- Jernbaner
- Jordbruksoppgjør
- Journalister
- Meglere
- Meierier
- Metallindustri
- Motorsykler
- Møbler
- Nedleggelser
- Næringsliv
- Næringsmiddelindustri
- Offshore
- Olje
- Oljeinntekter
- Opphavsrett
- Parkering
- Pelsdyravl
- Penger
- Plasser
- Post
- Priser
- Rederier
- Regnskaper
- Reguleringer
- Reinsdyr
- Reklame
- Renter
- Restauranter
- Revisorer
- Rivning
- Samferdsel
- Skatter
- Skip
- Skipsfart
- Skogbruk
- Småfly
- Snørydding
- Sparing
- Sportsartikler
- Sprengstoff
- Stasjoner
- Statsdrift
- Steinindustri
- Tekoindustri
- Telefon
- Telekommunikasjon
- Trafikk
- Trafikkontroll
- Treforedlingsindustri
- Tunneler
- Utleie
- Valuta
- Vannforsyning
- Varehandel
- Varemerking
- Vedlikehold
- Veier

- Kjemisk industri
- Kollektivtrafikk
- Konkurser og akkorder
- Konserner
- Konsulentfirmaer
- Kraftforsyning
- Verkstedindustri
- Vin
- Våpenindustri
- Økonomi
- Åpningstid
- Årsrapporter

Vedlegg B: stoppordliste

a
absolutt
aldri
alle
allerede
alltid
alt
altså
andre
annet
as
at
atten
av
b
bare
bedre
begge
beste
blant
ble
blei
bli
blir
blitt
bort
bra
bruke
bør
både
båe
c
d
da
de
deg
dei
deira
deires
del
dem
den
denne
der
dere
deres
derfor
derimot
dermed
dersom
dessuten
dessverre
det
dette
di
din
disse
ditt
du
dykk
dykkar
då
e
eg
egen
ei
ein
eit
eitt
eks
eksempel
eller
elles
elleve
en
én
ene
eneste
enhver
enn
er
et
ett
etter
etterpå
f
fem
femten
fikk
finne
finnes
fire
fjor
fjorten
flere
folk
for
fordi
forteller
fortsatt
fra
frem
før
først
første
få
får
fått
g
gang
gi
gikk
gir
gjennom
gjerne
gjorde
gjort
gjør
gjøre
god
gode
godt
grunn
gå
går
h
ha
hadde
ham
han
hans
har
hatt
hele
heller
helt
hennar
henne
hennes
her
hjå
ho
hoe
honom
hos
hoss
hossen
hun
hva
hvem
hver
hvert
hvilke
hvilken
hvis
hvor
hvordan
hvorfor
i
ifølge
igjen
ikke
ikkje
ingen
ingi
inkje
inn
innen
inni
j
ja
jeg
jo
k
kan
kanskje
kom
komme
kommer
korleis
korso
kun
kunne
kvar
kvarhelst
kven
kvi
kvifor
l
lage
lang
langt
ligger
lik
like
likevel
lite
litt
m
man
mange
me
med
medan
meg
meget
mellom
men
mener
mens
mer
mest
mi
min
mindre
mine
mot
mulig
mye
mykje
må
måte
måtte
n
ned
neste
nesten
nettopp
ni
nitten
no
noe
noen
nok
noka
noko
nokon
nokor
nokre
ntb
ny
nye
nå
når
o
og
også
om
opp
oss
over
overfor
p
part
per
på
q
r
rett
riktig
rundt
s
sa
samme
sammen
samtidig
sant
satt
se
seg
seks
seksten
selv
ser
sett
si
sia
sidan
siden
sier
sin
sine
sist
siste
sitt
sju
sjøl
skal
skulle
slik
slike
slutt
so
som
somme
somt
stor
store
større
største
størst
står
svært
synes
sytten
syv
søtten
så
sånn
t
ta
tar
tatt
ti
tid
tidligere
til
tilbake
tillegg
tjue
to
tok
tolv
tre
tretten
tror
tyve
u
um
under
upp
ut
uten
v
var
vart
varte
ved
vel
veldig
vere
verte
vet
vi
via
viktig
vil
ville
viser
vite
vore
vors
vort
være
vært
vår
våre
w
www
x
y
z
æ
ø
ønsker
å
år
åtte

Vedlegg C: Resultat fra forsøkene

I flere av tabellene under er ikke kategorinavnet med, kun et nummer som representerer kategorien :

Nummer	Kategori
1	Økonomi og næringsliv
2	Ulykker og naturkatastrofer
3	Sport
4	Forsvar og krig og konflikter
5	Politikk
6	Sosiale forhold
7	Personalia
8	Kriminalitet og rettsvesen
9	Vitenskap og teknologi
10	Kultur og underholdning
11	Medisin og helse
12	Utdanning
13	Fritid
14	Arbeidsliv
15	Natur og miljø
16	Religion og livsyn

C.1: Forsøkene med Aftenposten september-desember 2007

C.1.1 Bakgrunnstall for beregning av resultater - fullt læringssett

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	967	32	14	3	25	30	9	18	18	28	18	8	29	41	43	2	1285
	2	1	93	0	0	0	0	0	1	0	0	0	0	2	0	3	0	100
	3	9	3	1345	0	4	3	9	4	3	7	7	0	11	4	3	0	1412
	4	0	3	0	183	17	0	1	4	3	3	2	1	1	0	0	1	219
	5	18	1	1	15	298	7	5	11	5	7	2	1	0	3	6	2	382
	6	0	0	0	0	9	58	0	2	1	2	3	2	3	2	0	0	82
	7	4	0	0	0	1	2	253	1	1	4	2	0	2	0	0	1	271
	8	10	11	2	9	5	6	2	336	0	3	1	3	1	5	1	2	397
	9	2	0	0	1	0	0	0	1	35	5	1	1	0	0	1	0	47
	10	39	0	9	10	15	27	41	15	32	971	11	9	49	8	9	23	1268
	11	3	0	1	0	2	3	2	1	3	0	120	2	0	5	1	0	143
	12	1	0	1	0	1	3	1	0	0	2	1	88	0	4	0	0	102
	13	4	0	1	0	0	0	0	0	0	7	1	1	101	0	2	0	117
	14	4	0	0	0	3	2	0	0	0	1	3	1	0	71	0	1	86
	15	6	2	0	0	5	0	1	0	2	2	0	0	2	0	73	0	93
	16	0	0	0	0	0	1	0	1	0	3	0	0	0	0	1	16	22
	sum	1068	145	1374	221	385	142	324	395	103	1045	172	117	201	143	143	48	6026

9684 artikler for opplæring og 6026 artikler for uttestingen, stoppord er fjernet, c-verdi = 5000

C.1.2 Bakgrunnstall for beregning av resultater - redusert læringssett

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	861	10	18	1	13	14	5	7	8	23	9	6	13	16	17	2	1023
	2	9	114	0	0	0	0	2	2	0	0	0	0	3	0	5	0	135
	3	6	0	1321	0	0	2	3	2	1	12	1	0	3	1	1	0	1353
	4	3	2	0	190	19	0	1	4	3	12	2	0	1	0	1	1	239
	5	42	2	5	17	314	15	8	13	10	23	4	3	0	7	8	4	475
	6	5	1	1	0	9	76	0	2	1	3	6	2	3	2	1	0	112
	7	13	0	3	0	2	2	283	2	3	28	4	0	3	1	0	2	346
	8	31	13	11	11	6	9	1	356	4	12	2	3	1	6	1	3	470
	9	4	1	1	1	0	0	0	1	43	5	2	1	0	0	4	0	63
	10	33	0	4	1	10	8	16	4	19	897	4	5	16	4	5	9	1035
	11	11	0	2	0	2	5	1	1	4	1	129	2	0	4	2	0	164
	12	3	0	1	0	1	4	1	1	1	5	3	93	0	3	0	0	116
	13	12	0	7	0	0	2	2	0	1	9	1	0	152	1	3	0	190
	14	13	1	0	0	3	4	0	0	1	6	4	2	1	98	0	2	135
	15	22	1	0	0	6	0	1	0	4	3	1	0	5	0	94	0	137
	16	0	0	0	0	0	1	0	0	0	6	0	0	0	0	1	25	33
	Sum	1068	145	1374	221	385	142	324	395	103	1045	172	117	201	143	143	48	6026

6574 artikler for opplæring og 6026 artikler for uttestingen, stoppord er fjernet, c-verdi = 5000

C.2: Forsøkene med Klassekampen september-desember 2007

C.2.1 Resultater - fullt læringssett

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0,8667	0,2063	0,3333
Ulykker og naturkatastrofer	0,6667	0,5000	0,5714
Sport	1,0000	0,2424	0,3902
Forsvar og krig og konflikter	0,4839	0,8824	0,6250
Politikk	0,2791	0,8000	0,4138
Sosiale forhold	0,2800	0,5833	0,3784
Personalia	0,0455	0,1667	0,0714
Kriminalitet og rettsvesen	0,3000	0,2500	0,2727
Vitenskap og teknologi	0,3077	0,5714	0,4000
Kultur og underholdning	0,7500	0,2432	0,3673
Medisin og helse	0,6316	0,9231	0,7500
Utdanning	0,5909	0,8125	0,6842
Fritid	0,4706	0,8889	0,6154
Arbeidsliv	0,3571	0,5556	0,4348
Natur og miljø	0,3810	1,0000	0,5517
Religion og livssyn	0,4783	0,9167	0,6286
gjennomsnitt	0,4931	0,5964	0,4680

9684 artikler er brukt for opplæring av modellen og 310 artikler for uttestingen.
stoppord er fjernet, c-verdi = 5000

Bakgrunnstall for beregning av resultater – fullt læringssett

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	**13**	1	0	2	11	3	0	4	4	1	5	2	1	5	11	0	63
	2	0	**2**	0	1	0	0	0	0	0	0	0	0	0	0	1	0	4
	3	0	0	**8**	3	1	2	8	0	0	3	1	0	3	2	1	1	33
	4	0	0	0	**15**	1	0	0	0	0	0	0	1	0	0	0	0	17
	5	0	0	0	1	**12**	0	1	0	0	0	0	0	0	0	0	1	15
	6	0	0	0	0	5	**7**	0	0	0	0	0	0	0	0	0	0	12
	7	1	0	0	1	0	0	**1**	0	0	1	0	2	0	0	0	0	6
	8	1	0	0	3	3	1	0	**3**	0	1	0	0	0	0	0	0	12
	9	0	0	0	0	1	0	1	0	**4**	0	0	1	0	0	0	0	7
	10	0	0	0	5	6	9	11	1	4	**18**	1	3	5	1	0	10	74
	11	0	0	0	0	0	0	0	0	0	0	**12**	0	0	1	0	0	13
	12	0	0	0	0	1	0	0	1	1	0	0	**13**	0	0	0	0	16
	13	0	0	0	0	0	1	0	0	0	0	0	0	**8**	0	0	0	9
	14	0	0	0	0	2	1	0	1	0	0	0	0	0	**5**	0	0	9
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	8
	16	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**11**	12
	Sum	15	3	8	31	43	25	22	10	13	24	19	22	17	14	21	23	310

9864 artikler for opplæring og 310 artikler for uttestingen, stoppord er fjernet, c-verdi = 5000

C.2.2 Resultat av den manuelle kategoriseringen

Kategori	Antall artikler
Økonomi og næringsliv	15
Ulykker og naturkatastrofer	3
Sport	8
Forsvar og krig og konflikter	31
Politikk	43
Sosiale forhold	25
Personalia	22
Kriminalitet og rettsvesen	10
Vitenskap og teknologi	13
Kultur og underholdning	24
Medisin og helse	19
Utdanning	22
Fritid	17
Arbeidsliv	14
Natur og Miljø	21
Religion og livsyn	23

Resultat av den manuelle kategoriseringen
310 artikler fra Klassekampen september-desember 2007

C.2.3 Bakgrunnstall for beregning av resultater – redusert læringssett

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	**9**	0	0	0	2	0	0	1	1	1	1	0	1	0	3	1	20
	2	0	**3**	0	1	0	0	1	0	1	2	0	0	0	0	2	0	10
	3	1	0	**8**	2	1	1	2	0	0	3	1	0	1	0	0	0	20
	4	0	0	0	**17**	2	0	0	0	0	0	0	1	0	0	0	0	20
	5	0	0	0	3	**14**	0	1	0	0	0	0	1	0	0	0	1	20
	6	0	0	0	0	5	**11**	0	1	0	3	0	0	0	0	0	0	20
	7	1	0	0	2	0	1	**9**	0	0	2	0	3	0	0	0	2	20
	8	1	0	0	3	3	3	2	**6**	0	2	0	0	0	0	0	0	20
	9	0	0	0	1	4	1	2	0	**9**	1	1	1	0	0	0	0	20
	10	0	0	0	1	3	1	3	0	0	**10**	0	0	0	1	0	1	20
	11	0	0	0	0	0	0	1	0	0	0	**16**	0	0	3	0	0	20
	12	0	0	0	0	2	0	0	1	1	0	0	**16**	0	0	0	0	20
	13	0	0	0	1	2	1	1	0	0	0	0	0	**14**	1	0	0	20
	14	2	0	0	0	3	4	0	1	0	0	0	0	1	**9**	0	0	20
	15	1	0	0	0	2	0	0	0	1	0	0	0	0	0	**16**	0	20
	16	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	**18**	20
	Sum	15	3	8	31	43	25	22	10	13	24	19	22	17	14	21	23	310

6574 artikler for opplæring og 310 artikler for uttestingen, stoppord er fjernet, c-verdi = 5000

C.2.4 Resultater inkludert alternativ kategori et – redusert læringsett

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0,7500	0,6000	0,6667
Ulykker og naturkatastrofer	1,0000	0,3000	0,4615
Sport	1,0000	0,5000	0,6667
Forsvar og krig og konflikter	0,8095	0,8500	0,8293
Politikk	0,4000	0,9000	0,5538
Sosiale forhold	0,4800	0,6000	0,5333
Personalia	0,4500	0,4500	0,4500
Kriminalitet og rettsvesen	0,7500	0,3000	0,4286
Vitenskap og teknologi	0,7857	0,5500	0,6471
Kultur og underholdning	0,3824	0,6500	0,4815
Medisin og helse	0,8500	0,8500	0,8500
Utdanning	0,8000	0,8000	0,8000
Fritid	0,9333	0,7000	0,8000
Arbeidsliv	0,6667	0,5000	0,5714
Natur og miljø	0,8500	0,8500	0,8500
Religion og livssyn	0,7500	0,9000	0,8182
gjennomsnitt	0,7286	0,6438	0,6505

6574 artikler er brukt for opplæring og 310 artikler for uttestingen, inkl alt. ktegori stoppord er fjernet, c-verdi = 5000

Bakgrunnstall for beregning av resultat inkl alternativ kategori – redusert læringssett

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	**12**	0	0	0	0	0	0	0	2	2	1	1	1	0	1	0	20
	2	1	**3**	0	0	0	0	1	0	0	4	0	0	0	0	1	0	10
	3	1	0	**10**	2	0	2	2	0	0	2	1	0	0	0	0	0	20
	4	0	0	0	**17**	1	0	1	0	0	0	0	1	0	0	0	0	20
	5	0	0	0	0	**18**	0	1	0	0	1	0	0	0	0	0	0	20
	6	0	0	0	0	4	**12**	0	0	0	3	0	0	0	0	0	1	20
	7	1	0	0	0	2	1	**9**	0	0	3	0	2	0	1	0	1	20
	8	1	0	0	1	5	2	2	**6**	0	2	1	0	0	0	0	0	20
	9	0	0	0	0	3	1	2	0	**11**	2	0	0	0	0	1	0	20
	10	0	0	0	1	1	0	1	0	0	**13**	0	0	0	1	0	3	20
	11	0	0	0	0	0	0	1	0	0	0	**17**	0	0	2	0	0	20
	12	0	0	0	0	2	0	0	1	0	0	0	**16**	0	0	0	1	20
	13	0	0	0	0	3	1	0	0	0	1	0	0	**14**	1	0	0	20
	14	0	0	0	0	4	4	0	1	0	1	0	0	0	**10**	0	0	20
	15	0	0	0	0	2	0	0	0	1	0	0	0	0	0	**17**	0	20
	16	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	**18**	20
	Sum	16	3	10	21	45	25	20	8	14	34	20	20	15	15	20	24	310

6574 artikler er brukt for opplæring og 310 artikler for uttestingen, inkl alt. ktegori stoppord er fjernet, c-verdi = 5000

C.2.5 Resultater for kategorisering av artikler skrevet av eksterne skribenter

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0,5000	0,3000	0,3750
Ulykker og naturkatastrofer	1,0000	0,2000	0,3333
Sport	1,0000	0,3750	0,5455
Forsvar og krig og konflikter	0,3636	1,0000	0,5333
Politikk	0,2667	0,5714	0,3636
Sosiale forhold	0,7000	0,7778	0,7368
Personalia	0,1667	0,5000	0,2500
Kriminalitet og rettsvesen	0,7500	0,3750	0,5000
Vitenskap og teknologi	0,5000	0,5000	0,5000
Kultur og underholdning	0,4000	0,4000	0,4000
Medisin og helse	0,8571	1,0000	0,9231
Utdanning	1,0000	0,8889	0,9412
Fritid	0,7500	0,7500	0,7500
Arbeidsliv	1,0000	0,5000	0,6667
Natur og miljø	0,7000	0,7000	0,7000
Religion og livssyn	0,8571	1,0000	0,9231
gjennomsnitt	0,6757	0,6149	0,5901

6574 artikler er brukt for opplæring og 115 artikler for uttestingen
stoppord er fjernet, c-verdi = 5000

Bakgrunnstall for beregning av resultat for eksterne skribenter

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	**3**	0	0	0	1	0	0	0	1	0	1	0	1	0	2	1	10
	2	0	**1**	0	1	0	0	0	0	1	1	0	0	0	0	1	0	5
	3	0	0	**3**	2	0	1	1	0	0	0	0	0	1	0	0	0	8
	4	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0	0	4
	5	0	0	0	2	**4**	0	1	0	0	0	0	0	0	0	0	0	7
	6	0	0	0	0	1	**7**	0	0	0	1	0	0	0	0	0	0	9
	7	1	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	2
	8	1	0	0	1	0	1	1	**3**	0	1	0	0	0	0	0	0	8
	9	0	0	0	0	3	0	0	0	**3**	0	0	0	0	0	0	0	6
	10	0	0	0	0	2	0	1	0	0	**2**	0	0	0	0	0	0	5
	11	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	6
	12	0	0	0	0	1	0	0	0	0	0	0	**8**	0	0	0	0	9
	13	0	0	0	1	1	0	1	0	0	0	0	0	**9**	0	0	0	12
	14	1	0	0	0	0	1	0	1	0	0	0	0	1	**4**	0	0	8
	15	0	0	0	0	2	0	0	0	1	0	0	0	0	0	**7**	0	10
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	6
	Sum	6	1	3	11	15	10	6	4	6	5	7	8	12	4	10	7	115

6574 artikler er brukt for opplæring og 115 artikler for uttestingen
stoppord er fjernet, c-verdi = 5000

C.2.6 Resultater for kategorisering av artikler skrevet av ansatte skribenter

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0,6667	0,6000	0,6316
Ulykker og naturkatastrofer	1,0000	0,4000	0,5714
Sport	1,0000	0,4167	0,5882
Forsvar og krig og konflikter	0,6500	0,8125	0,7222
Politikk	0,3571	0,7692	0,4878
Sosiale forhold	0,2667	0,3636	0,3077
Personalia	0,5000	0,4444	0,4706
Kriminalitet og rettsvesen	0,5000	0,2500	0,3333
Vitenskap og teknologi	0,8571	0,4286	0,5714
Kultur og underholdning	0,4211	0,5333	0,4706
Medisin og helse	0,8333	0,7143	0,7692
Utdanning	0,5714	0,7273	0,6400
Fritid	1,0000	0,6250	0,7692
Arbeidsliv	0,5000	0,4167	0,4545
Natur og miljø	0,8182	0,9000	0,8571
Religion og livssyn	0,7500	0,8571	0,8000
gjennomsnitt	**0,6682**	**0,5787**	**0,5903**

6574 artikler er brukt for opplæring og 195 artikler for uttestingen
stoppord er fjernet, c-verdi = 5000

Bakgrunnstall for beregning av resultat for ansatte skribenter

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	**6**	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	10
	2	0	**2**	**0**	0	0	0	1	0	0	1	0	0	0	0	1	0	5
	3	1	0	**5**	0	1	0	1	0	0	3	1	0	0	0	0	0	12
	4	0	0	0	**13**	2	0	0	0	0	0	0	1	0	0	0	0	16
	5	0	0	0	1	**10**	0	0	0	0	0	0	1	0	0	0	1	13
	6	0	0	0	0	4	**4**	0	1	0	2	0	0	0	0	0	0	11
	7	0	0	0	2	0	1	**8**	0	0	2	0	3	0	0	0	2	18
	8	0	0	0	2	3	2	1	**3**	0	1	0	0	0	0	0	0	12
	9	0	0	0	1	1	1	2	0	**6**	1	1	1	0	0	0	0	14
	10	0	0	0	1	1	1	2	0	0	**8**	0	0	0	1	0	1	15
	11	0	0	0	0	0	0	1	0	0	0	**10**	0	0	3	0	0	14
	12	0	0	0	0	1	0	0	1	1	0	0	**8**	0	0	0	0	11
	13	0	0	0	0	1	1	0	0	0	0	0	0	**5**	1	0	0	8
	14	1	0	0	0	3	3	0	0	0	0	0	0	0	**5**	0	0	12
	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	10
	16	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	**12**	14
	Sum	9	2	5	20	28	15	16	6	7	19	12	14	5	10	11	16	195

6574 artikler er brukt for opplæring og 195 artikler for uttestingen
stoppord er fjernet, c-verdi = 5000

C.2.7 Resultater for kategorisering av nynorskartikler[15]

Kategori	Fullstendighet	Presisjon	F-verdi
Økonomi og næringsliv	0,8333	0,4167	0,5556
Ulykker og naturkatastrofer	0,0000	0,0000	0,0000
Sport	0,0000	0,0000	0,0000
Forsvar og krig og konflikter	0,6000	0,5000	0,5455
Politikk	0,6429	0,9000	0,7500
Sosiale forhold	0,1250	1,0000	0,2222
Personalia	0,0000	0,0000	0,0000
Kriminalitet og rettsvesen	1,0000	0,2857	0,4444
Vitenskap og teknologi	0,0000	0,0000	0,0000
Kultur og underholdning	0,9355	0,7073	0,8056
Medisin og helse	1,0000	0,7500	0,8571
Utdanning	0,5556	0,7143	0,6250
Fritid	0,0000	0,0000	0,0000
Arbeidsliv	0,4000	1,0000	0,5714
Natur og miljø	0,0000	0,0000	0,0000
Religion og livssyn	0,0000	0,0000	0,0000

6574 artikler er brukt for opplæring og 100 artikler for uttestingen
stoppord er fjernet, c-verdi = 5000

Bakgrunnstall for beregning av resultat for nynorskartiklene

		Manuell kategorisering																
	Kategori	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	sum
Automatisk kategorisering	1	**5**	0	0	1	2	1	0	0	0	1	0	0	1	0	1	0	12
	2	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	**3**	0	1	0	0	0	1	0	0	0	1	0	0	6
	5	0	0	0	0	**9**	1	0	0	0	0	0	0	0	0	0	0	10
	6	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	1
	7	0	0	0	0	0	0	**0**	0	0	0	0	1	0	0	0	0	1
	8	0	0	0	0	1	3	0	**2**	0	0	0	0	0	1	0	0	7
	9	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
	10	0	0	0	1	1	1	1	0	2	**29**	0	3	0	0	1	2	41
	11	0	0	0	0	0	0	0	0	0	0	**6**	0	0	2	0	0	8
	12	0	0	0	0	1	0	0	0	0	0	0	**5**	0	1	0	0	7
	13	0	0	0	0	0	0	1	0	0	0	0	0	**0**	1	0	0	2
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	4
	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
	Sum	6	0	0	5	14	8	2	2	2	31	6	9	1	10	2	2	100

6574 artikler er brukt for opplæring og 100 artikler for uttestingen
stoppord er fjernet, c-verdi = 5000

15 Gjennomsnittverdier er ikke vist fordi flere kategorier ikke har forekomster

C.3: Forsøkene med Aftenposten mars 2008

C.3.1 Bakgrunnstall for beregning av resultater – mars 2008

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	509	13	6	3	8	11	3	11	9	16	6	6	19	11	15	1	647
	2	6	75	0	0	0	0	0	5	2	0	0	0	3	0	1	0	92
	3	4	2	688	0	0	3	1	0	0	8	0	2	7	0	0	0	715
	4	3	2	4	127	20	1	5	11	2	4	2	0	0	0	1	0	182
	5	32	3	8	25	213	23	2	11	8	28	7	11	9	4	9	14	407
	6	3	2	0	0	4	39	0	3	0	2	5	3	0	1	0	7	69
	7	9	2	9	0	0	5	148	1	1	26	4	9	3	3	1	4	225
	8	37	20	5	23	7	15	3	273	2	19	7	0	7	5	1	1	425
	9	3	0	0	1	0	0	0	0	22	4	2	2	0	0	2	1	37
	10	26	3	6	2	8	10	21	3	14	685	2	8	15	3	1	13	820
	11	5	1	1	1	1	4	0	1	1	0	122	0	2	0	4	1	144
	12	1	0	0	0	2	1	0	1	0	7	2	70	0	0	2	1	87
	13	7	5	4	0	0	6	0	0	0	9	2	0	80	0	5	0	118
	14	11	1	0	0	0	7	0	0	1	4	5	4	1	79	0	0	113
	15	11	6	2	0	1	0	0	2	1	1	2	0	4	1	49	0	80
	16	0	0	0	0	11	1	0	0	0	1	0	0	0	0	0	14	27
	Sum	667	135	733	182	275	126	183	322	63	814	168	115	150	107	91	57	4188

6574 artikler for opplæring og 4188 artikler for uttestingen, stoppord er fjernet, c-verdi = 5000

C.4: Forsøkene med Klassekampen mars 2008

C.4.1 Uttrekket av testsettet mars 2008 basert på automatisk kategorisering

Kategori	artikler totalt	artikler trukket
Økonomi og næringsliv	86	19
Ulykker og naturkatastrofer	3	3
Sport	10	10
Forsvar og krig og konflikter	71	20
Politikk	181	20
Sosiale forhold	32	20
Personalia	7	7
Kriminalitet og rettsvesen	48	20
Vitenskap og teknologi	7	7
Kultur og underholdning	236	19
Medisin og helse	31	18
Utdanning	11	11
Fritid	12	12
Arbeidsliv	25	20
Natur og Miljø	15	15
Religion og livsyn	12	12
	787	233

Uttrekk av testsett
233 artikler fra Klassekampen mars 2008

C.4.2 Resultat av den manuelle kategoriseringen

Kategori	Antall artikler
Økonomi og næringsliv	15
Ulykker og naturkatastrofer	1
Sport	5
Forsvar og krig og konflikter	22
Politikk	21
Sosiale forhold	28
Personalia	6
Kriminalitet og rettsvesen	10
Vitenskap og teknologi	8
Kultur og underholdning	24
Medisin og helse	18
Utdanning	14
Fritid	10
Arbeidsliv	21
Natur og Miljø	11
Religion og livsyn	19

Resultat av den manuelle kategoriseringen
233 artikler fra Klassekampen mars 2008

C.4.3 Bakgrunnstall for beregning av resultater – mars 2008

		Manuell kategorisering																
	Kategori	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**sum**
Automatisk kategorisering	**1**	**11**	0	0	1	2	0	0	0	0	2	0	0	0	3	0	0	19
	2	0	**1**	0	0	0	0	0	0	0	0	0	0	1	0	1	0	3
	3	0	0	**5**	1	0	0	1	0	0	2	0	0	0	1	0	0	10
	4	0	0	0	**16**	1	1	0	1	0	1	0	0	0	0	0	0	20
	5	2	0	0	2	**11**	1	0	0	1	0	0	0	0	0	0	3	20
	6	0	0	0	0	1	**17**	0	0	1	0	1	0	0	0	0	0	20
	7	0	0	0	0	1	0	**3**	0	0	2	0	1	0	0	0	0	7
	8	1	0	0	0	3	1	1	**9**	0	1	0	3	0	0	0	1	20
	9	0	0	0	1	0	1	1	0	**2**	0	0	0	0	0	0	2	7
	10	0	0	0	0	0	2	0	0	2	**13**	0	0	0	0	0	2	19
	11	0	0	0	0	0	1	0	0	0	0	**16**	0	0	1	0	0	18
	12	0	0	0	0	0	0	0	0	0	0	1	**10**	0	0	0	0	11
	13	1	0	0	0	0	1	0	0	0	1	0	0	**9**	0	0	0	12
	14	0	0	0	0	0	2	0	0	1	1	0	0	0	**16**	0	0	20
	15	0	0	0	1	2	0	0	0	1	1	0	0	0	0	**10**	0	15
	16	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**11**	12
	Sum	15	1	5	22	21	28	6	10	8	24	18	14	10	21	11	19	233

6574 artikler for opplæring og 4188 artikler for uttestingen, stoppord er fjernet, c-verdi = 5000

Vedlegg D: Programmer

I dette vedlegget ligger et lite utvalg av programmene som er skrevet for forsøkene som er gjort.

- LagMatriseA.x oppretting av termmatrise fra Aftenpostens XML-format.
- LagMatriseKK.x oppretting av termmatrise fra Klassekampens txt-baserte filer
- fjernTegn.x subprogram som fjerner redigeringstagger og spesialtegn fra teksten
- Vekting.x vekting av termene
- Resutat.x. Beregning av fullstendighet, presisjon og F-verdi

Programmene er skrevet i språket Open Object Rexx. Det er åpen programvare som kan lastes ned fra http://www.oorexx.org/download.html
Editoren X2 er brukt for skriving av programmene. Dette er også åpen programvare som kan lastes ned fra http://www.tangbu.com/x2main.shtml

LagMatriseA.x

```
/* Oppretting av termmatrise fra Aftenpostens XML-format
    • gyldige termer leses  inn fra termfil
    • termmatrise lagres i tempfil, hver rad i matrisa inneholder termer for et dokument
    • termer med termfrekvens 0 lagres ikke
    • dokumentfrekvens skrives ut i dffil sammen med antall dokumenter
    • tempfil og dffil er input til programmet Vekting.x som vekter hver enkel term i matrisa
    •
*/
arg opt
Select
  When opt = 1 then matrise = 'TR'              /* lag matrise for læring*/
  When opt = 2 then matrise = 'TE'              /* lag matrise for test*/
  Otherwise
      'msg du må oppgi verdi 1 eller 2'
       exit
End

Select
   When matrise = 'TR' then do
      filliste = 'C:\Masteroppg\A2007\train\*.xml'
      tempfil='C:\Masteroppg\Vekting\A2007\temp_train.dat'
      dffil='C:\Masteroppg\Vekting\A2007\df_train.dat'
      filut='C:\Masteroppg\Vekting\A2007\logg_lagmatrise_train_ddmmaa.dat'
   end
   When matrise = 'TE' then do
      filliste = 'C:\Masteroppg\A2007\test\*.xml'
      tempfil='C:\Masteroppg\Vekting\A2007\temp_test.dat'
      dffil='C:\Masteroppg\Vekting\A2007\df_test.dat'
      filut='C:\Masteroppg\Vekting\A2007\logg_lagmatrise_test_ddmmaa.dat'
   end
   Otherwise
end
termfil='C:\Masteroppg\Termer\termliste.csv'
```

```
dokFrekv. = 0
term. = 0
funnetTerm. = 0

Address CMD '@ERASE' tempfil
Address CMD '@ERASE' dffil

/* gjør klar loggfil for skriving  */
'edit' filut
'top'
'delete *'

/* leser inn termlista */
q1=stream(termfil,'C','QUERY EXISTS')
if q1='' then 'input termfil eksisterer ikke'
overskrift = linein(termfil)                /* første linje er en overskrift*/
Do i1=1 while lines(termfil)
   linje=linein(termfil)
   parse var linje termnr ';' term ';' .
   term.termnr = term
   funnetTerm.term = 1
End
call lineout termfil
antTermer = i1-1

/* traverserer XML-filene og bygger matrisa */
fileopt='FS'
call SysFileTree filliste, "filer", fileopt
do i1=1 to filer.0                          /* filer.0 = antall xfiler*/
   parse var filer.i1 . . . . f1
   fil1.i1=strip(f1)
end

do i1=1 to filer.0
   artikkel=''
   drop antOrd.
   antOrd. = 0
   ordliste = ''
   pk = lastpos('\',fil1.i1)
   filnavn = substr(fil1.i1,pk+1)
   do while lines(fil1.i1)
      artikkel = artikkel strip(linein(fil1.i1))
   end
   call lineout fil1.i1                     /* fila er lest og kan lukkes*/
   do while artikkel<> ''
      /* behandler kun den delen av artikkelen som ligger i <story> */
      parse lower var artikkel A1 '<' tag'>' artikkel
     If tag = 'story' Then do
        parse var artikkel story '</story>' artikkel
        /* fjerner ![cdata[  som alltid ligger først i story <*/
        parse var story ord1 story
        If ord1 <> '<![cdata[' Then story = ord1 story
        /* kaller subrutine for å fjerne spesialtegn i teksten */
        tekst = fjernTegn(story)
```

```
        /* finn termer */
        do j1=1 to words(tekst)
            term = word(tekst,j1)
            If funnetTerm.term = 1 then do
              If antOrd.term = 0 then do
                ordliste = ordliste term
                dokFrekv.term = dokFrekv.term + 1
              end
              antOrd.term = antOrd.term + 1
            end
         end

         /* skriver ut dokumentets termfrekvenser i tempfil */
        ut=filnavn
        do j1=1 to antTermer
           term = term.j1
           If antOrd.term > 0 then do
             ut = ut j1||':'antOrd.term
           end
        end
        call lineout tempfil, ut
      end                              /* story */
   end                                 /* end artikkel */
end                                    /* end filtraversering */
call lineout tempfil

/* skriver ut antall dokumenter, termer og termens dokumentfrekvens i dffil*/
call lineout dffil, 'antDok' filer.0
call lineout dffil, 'antTerm' antTermer
ut='dokfrekv'
Do j1=1 to antTermer
   term = term.j1
   ut = ut j1||':'dokFrekv.term
end
call lineout dffil, ut
call lineout dffil

exit
```

LagMatriseKK.x

```
/* Oppretting av termmatrise fra Klassekampens txt-filer
```

- gyldige termer leses inn fra *termfil*
- termmatrise lagres i *tempfil*, hver rad i matrisa inneholder termer for et dokument
- termer med termfrekvens 0 lagres ikke
- dokumentfrekvens skrives ut i *dffil* sammen med antall dokumenter
- *tempfil* og *dffil* er input til programmet Vekting.x som vekter hver enkel term i matrisa
-

```
*/
ffilliste = 'C:\Masteroppg\K2007\test\*.xml'
tempfil='C:\Masteroppg\Vekting\K2007\temp_test.dat'
dffil='C:\Masteroppg\Vekting\K2007\df_test.dat'
filut='C:\Masteroppg\Vekting\K2007\logg_lagmatrise_test_ddmmaa.dat'
termfil='C:\Masteroppg\Termer\termliste.csv'
```

```
Address CMD '@ERASE' tempfil
Address CMD '@ERASE' dffil

dokFrekv. = 0
term. = 0
funnetTerm. = 0
antOrd. = 0
antLinje = 0
antArt = 0
antTermer = 0
fortekst. = 0

/* gjør klar loggfil for skriving  */
'edit' filut
'top'
'delete *'

/* leser inn termlista */
q1=stream(termfil,'C','QUERY EXISTS')
if q1='' then 'input termfil eksisterer ikke'
do i1=1 while lines(termfil)
    linje=linein(termfil)
   parse var linje termnr ';' term ';' .
   term.termnr = term
   funnetTerm.term = 1
end
call lineout termfil
antTermer = i1-1

/* leser og behandler artikkelfilene */
fileopt='FS'
call SysFileTree filliste, "filer", fileopt
do i1=1 to filer.0                          /* filer.0 = antall filer*/
    parse var filer.i1 . . . . f1
    fil1.i1=strip(f1)
end

do i1=1 to filer.0
     avisfil=fil1.i1
    /* leser inn og behandler avisfila  */
   do while lines(avisfil)
      linje=linein(avisfil)
      antLinje = antLinje + 1
      /* traverserer ned til start artikkel*/
      do until pos('---',linje) > 0
         linje=linein(avisfil)
         If linje  <> 0 & pos('---',linje) = 0 & pos('===',linje) = 0 then tittel = linje
         antLinje = antLinje + 1
      end
      /* hopper over de første linjene, antall linjer varierer, avhengig av om forfatter og epostadresse
          er med. Første linje er alltid 'Klassekampen dato og tid'
      */
      do j1=1 to 6
          fortekst.j1 = linein(avisfil)
          antLinje = antLinje+1
```

```
        If fortekst.j1 = '' | pos('@',fortekst.j1) > 0 then leave
      end
      linje = fortekst.j1
      artikkel = ''
      tekst1 = ''
      do while lines(avisfil)
         linje = linein(avisfil)
         antLinje = antLinje + 1
         if pos('(C) Klassekampen',linje) > 0 then leave
         artikkel = artikkel strip(linje)
      end
      If artikkel <> '' then do                    /* ny artikkel*/
        antArt = antArt + 1
       /* fjerner redigeringstagger i artikkelen  */
       do while artikkel<> ''
          parse var artikkel A1 '<' tag '>' artikkel
          tekst1 = tekst1 A1
       end
       /* kaller subrutine for å fjerne spesialtegn i teksten */
       tekst = fjernTegn(tekst1)
       /* finn termer */
       do j1=1 to words(tekst)
          term = word(tekst,j1)
          If funnetTerm.term = 1 then do
            If antOrd.term = 0 then  dokFrekv.term = dokFrekv.term + 1
            antOrd.term = antOrd.term + 1
          end
       end
       /* skriver ut dokumentets termfrekvenser i tempfil */
       ut='art'||antArt
       do j1=1 to antTermer
          term = term.j1
          If antOrd.term > 0 then do
             ut = ut j1||':'antOrd.term
             antOrd.term = 0
          end
       end
       call lineout tempfil, ut
     end                  /* end artikkel */
   end                    /* end lesing avisfil */
end                       /* end filtraversering */

/* skriver ut antall dokumenter, termer og termens dokumentfrekvens i dffill*/
call lineout dffil, 'antDok' antArt
call lineout dffil, 'antTerm' antTermer
ut='dokfrekv'
do j1=1 to antTermer
   term = term.j1
   ut = ut j1||':'dokFrekv.term
end
call lineout dffil, ut
call lineout dffil

exit
```

fjerntegn.x

```
/* subbrutine for å fjerne spesialtegn fra teksten */
parse arg inn
ut = ''
/* fjerner redigeringstagger i teksten */
Do while inn <> ''
  parse var inn A1 '<' tag '>' inn
  ut = ut A1
end
/* fjerner skilletegn og spesialtegn*/
ut = translate(ut, ' ','.,:;*$/\&%+–…¦{}§=£][()-?!<>«»"")
exit space(ut)
```

Vekting.x

```
/* programmet leser termene i termmatrisa tempfil som leveres inn til programmet
      ● vekter hver term med vektingsfunksjonen (tf*log idf) / (sum(tf*log idf))^2
      ● oppretter vektormatrise på formatet for SVM^multi
      ● når programmet kjøres for Train hentes kategori fra katfil , ved kjøring for test er katfil = 0
*/

call RxFuncAdd "MathLoadFuncs","rxmath","MathLoadFuncs"
call MathLoadFuncs

arg opt
Select
  When opt = 1 then matrise =  'TR'       /* læring*/
  When opt = 2 then matrise = 'TE'        /* test*/
  Otherwise
      'msg du må oppgi verdi 1-2'
      exit
End

Select
  When matrise = 'TR' then do
      tempfil='C:\Masteroppg\Vekting\A2007\temp_train.dat'
      katfil='C:\Masteroppg\Kategori\A2007\train.csv'
      dffil='C:\Masteroppg\Vekting\A2007\df_train.dat'
      svmfil='C:\Masteroppg\SVMmulti\train.dat'
      filut='C:\Masteroppg\SVMmulti\logg_train_ddmmaa.dat'
  end
  When matrise = 'TE' then do
      tempfil='C:\Masteroppg\Vekting\A2007\temp_test.dat'
      dffil='C:\Masteroppg\Vekting\A2007\df_test.dat'
      svmfil='C:\Masteroppg\SVMmulti\test.dat'
      filut='C:\Masteroppg\SVMmulti\logg_test_ddmmaa.dat'
  end
  Otherwise
end

dokFrekv. = 0
katfil. = 0
sumvekt. = 0
nevner. = 0
```

```
katnr. = 0
sumvekt. = 0

/* gjør klar loggfil for skriving  */
'edit' filut
'top'
'delete *'

/* leser inn antall dokumenter, termer  og dokumentfrekvens */
q1=stream(dffil,'C','QUERY EXISTS')
if q1='' then 'input dffil eksisterer ikke'
rad = linein(dffil)                          /* antall dokumenter*/
parse var rad xx antDok .
rad = linein(dffil)                          /* antall termer*/
parse var rad xx antTerm .
rad = linein(dffil)                          /* dokumentfrekvens*/
do j1=2 to words(rad)
    celle = word(rad,j1)
    parse var celle termnr ":" df
    If df > 0 then  dokFrekv.termnr = df
end
call lineout dffil

/* leser inn kategorilista med filene som skal behandles */
/* fila eksisterer ikke når jobben kjøres for test */
q1=stream(katfil,'C','QUERY EXISTS')
if q1='' then 'input katfil eksisterer ikke'
else do
   do while lines(katfil)
      linje=linein(katfil)
      parse var linje fn ';' katnr ';'.
      pk = lastpos('\',fn)
      filnavn = substr(fn,pk+1)
      katfil.filnavn = 1
      katnr.filnavn = katnr
   end
   call lineout katfil
end

/* leser dokumentenes termfrekvenser fra tempfil og beregner nevnerdelen av vektingsalgoritmen
    denne skal brukes når vekt etterpå beregnes for hver enkelt term */
do i1=1 while lines(tempfil)
   rad=linein(tempfil)
   do j1=2 to words(rad)
      celle = word(rad,j1)
      parse var celle termnr ":" tf
      df = dokFrekv.termnr
      If df = 0 then do
         'input FEIL: dokfrekvens er null for termnr:' termnr
          iterate
      end
      idf = antDok/df
      teller = tf  * RxCalcLog10(idf)
      sumvekt.termnr = sumvekt.termnr + teller
   end
```

```
end                              /* end traversering tempfil  */
call lineout tempfil
/* beregner nevner */
do i1=1 to antTerm
   termnr = i1
   If sumvekt.termnr > 0 then nevner.termnr = RxCalcSqrt(sumvekt.termnr)
end

Address CMD '@ERASE' svmfil

/* beregner vekt pr term i matrisa og opprettet SVM fil */
do i1=1 while lines(tempfil)
   rad=linein(tempfil)
   filnavn =  word(rad,1)
   ut = 0
   If matrise = 'TR'  then  ut=katnr.filnavn   /*Train  - kategorinummer skal settes, */
   do j1=2 to words(rad)
      celle = word(rad,j1)
      parse var celle termnr ":" tf
      df = dokFrekv.termnr
      If df = 0 then iterate
      idf = antDok/df
      teller =  tf  * RxCalcLog10(idf)
      If nevner.termnr = 0 then iterate
      vekt = teller/nevner.termnr
     ut = ut termnr||':'||vekt
   end
   call lineout svmfil, strip(ut)
end
call lineout tempfil
call lineout svmfil

exit
```

Resultat.x

```
/* Sammenlikning  av manuell og automatisk kategorisering
```

- fil.1 – fil med resultater fra automatisk kategorisering
- fil.2 – fil med resultater fra manuell kategorisering
- katfil - fil med kategorinavn
- avvikfil – fil med artikler som ikke har samsvar mellom automatisk og manuell kategori
- stat1 – matrise med reslutater fra sammenlikningen
- stat2 – fullstendighet, presisjon og F-verdi for hver kategori, samt gjennomsnitt (micro og macro)

```
*/

fil.1='C:\Masteroppg\K2007\autkat.csv'
fil.2='C:\Masteroppg\K20072\mankat.csv'
katfil= 'C:\masteroppg\kategori\katnavn.csv'
avvikfil= 'C:\Masteroppg\K2007\Avvik.csv'
stat1='C:\Masteroppg\K2007\stat_raa_tabell.csv'
stat2='C:\Masteroppg\K2007\stat_fverdi.csv'
utfil='C:\Masteroppg\K2007\logg_resultat_ddmmaa.txt'
```

```
Address CMD '@ERASE' stat1
Address CMD '@ERASE' stat2

'EDIT' utfil
'TOP'
'DELETE *'

maxKatnr=16
ant.=0
katnr.=''
tittel.=''
katnavn.=''

/* Les inn kategorifil for å få kategorinavnene */
q1=stream(katfil,'C','QUERY EXISTS')
if q1='' then 'input katfil ikke funnet'
do i2=1 while lines(katfil)
   linje=linein(katfil)
   parse var linje katnr ';' katnavn ';' .
   If datatype(katnr,'W')=0 Then iterate
   katnavn.katnr=katnavn
end
call lineout katfil

/* resultatet fra automatisk kategorisering*/
q1=stream(fil.1,'C','QUERY EXISTS')
if q1='' then 'input' fil.1 'ikke funnet'
do i2=1 while lines(fil.1)
   linje=linein(fil.1)
   parse var linje art ';' tittel ';' katnr ';' .
   artnr = strip(art)
   If datatype(katnr,'W')=0 Then iterate
   ant.1.katnr=ant.1.katnr+1          /* antall forekonmster pr kategori*/
   ant.1=ant.1+1                      /* antall forekomster totalt*/
   katnr.1.artnr=katnr
   tittel.artnr = tittel
   If katnr>maxKatnr Then maxKatnr=katnr
end
call lineout fil.1

filnr=0
Address CMD '@ERASE' avvikfil

/* resultatet av manuell kategorisering*/
q1=stream(fil.2,'C','QUERY EXISTS')
if q1='' then 'input' fil.2 'ikke funnet'
do i2=1 while lines(fil.2)
   linje=linein(fil.2)
   parse var linje artnr ';' katnr ';' dublnr .
   If datatype(katnr,'W')=0 Then do
      'input katnr mangler for artikkel' artnr
     iterate
   end
   If katnr.1.artnr = '' then 'input artnr' artnr      /* finnes ikke automatisk match*/
   ant.2.katnr=ant.2.katnr+1                           /* antall forekomster pr kategori*/
```

```
   ant.2=ant.2+1                                    /* antall forekomster totalt*/
   katnr1=katnr.1.artnr
   ant.3.katnr1.katnr=ant.3.katnr1.katnr+1
   If katnr>maxKatnr Then maxKatnr=katnr
   If katnr.1.artnr <> katnr then
     call lineout avvikfil, artnr ';' tittel.artnr ';' katnr.1.artnr ';' katnr
end
call lineout fil.2
call lineout avvikfil

/* stat1 – matrise  */
sum.=0
ut = ''
do k2=1 to maxKatnr
    ut=ut ';'k2
end
call lineout stat1, ut';sum;'
do k1=1 to maxKatnr
    ut=k1
    do k2=1 to maxKatnr
      ut=ut';'ant.3.k1.k2
      sum.1.k1=sum.1.k1+ant.3.k1.k2
      sum.2.k2=sum.2.k2+ant.3.k1.k2
    end
    call lineout stat1, ut';'sum.1.k1
end
ut='sum;'
do k2=1 to maxKatnr
     ut=ut';'sum.2.k2
end
call lineout stat1, ut
call lineout stat1

/* stat2 – Fullstendighet, presisjon og F-verdi  */
antKat = 0
sumFullst = 0
sumPresisjon = 0
gjf_nevner = 0
gjf_teller = 0
gjf_micro = 0
gjf_macro = 0
gjp_nevner = 0
gjp_teller = 0
gjp_micro = 0
gjp_macro = 0
gjfv_micro = 0
gjfv_macro = 0
antKat = 0

call lineout stat2, 'Katnr;fullstendighet;presisjon;fverdi'
Do i2=1 to maxKatnr
   fullst = 0
   presisjon = 0
   f = 0
   antKat = antKat + 1
```

```
   /* fullstendighet*/
   if ant.2.i2 > 0 then fullst = ant.3.i2.i2/ant.2.i2
   /* presisjon */
   if ant.1.i2 > 0 then presisjon = ant.3.i2.i2/ant.1.i2
   /* beregning av fverdi B = 1  */
   If fullst > 0 | presisjon > 0 then  f = 2*presisjon*fullst/(presisjon+fullst)

   /* for gjennomsnitt micro */
   gjf_nevner = gjf_nevner + ant.3.i2.i2
   gjf_teller = gjf_teller + ant.2.i2
   gjp_nevner = gjf_nevner
   gjp_teller = gjp_teller + ant.1.i2

   /* for gjennomsnitt macro  */
   sumFullst  = sumFullst + fullst
   sumPresisjon = sumPresisjon + presisjon

   call lineout stat2, i2';'format(fullst,,4)';'format(presisjon,,4)';'format(f,,4)';'katnavn.i2
end

/* gjennomsnitt  */
gjf_micro =  gjf_nevner/gjf_teller
gjp_micro = gjp_nevner/gjp_teller
gjfv_micro = 2*gjf_micro*gjp_micro/(gjf_micro+gjp_micro)
gjf_macro = sumFullst/antKat
gjp_macro = sumPresisjon/antKat
gjfv_macro = 2*gjf_macro*gjp_macro/(gjf_macro+gjp_macro)

call lineout stat2, 'micro;'format(gjf_micro,,4)';'format(gjp_micro,,4)';'format(gjfv_micro,,4)
call lineout stat2, 'macro;'format(gjf_macro,,4)';'format(gjp_macro,,4)';'format(gjfv_macro,,4)
call lineout stat2

exit
```