HØGSKOLEN I OSLO
OG AKERSHUS

Språkteknologiske verktøy i kartleggingen av hatefulle ytringer på Internett

Institutt for informasjonsteknologi

Hugo Hammer

06.10.2013

Disposisjon

— Hva er hatefulle ytringer?

— Interessante analyser ifm hatefulle ytringer

— Relevante språkteknologiske verktøy

— Ekstremismeprosjektet på HiOA

Hva er hatefulle ytringer?

”Vi i SIAN har som mål å vekke alle Breivikene i samfunnet sånn at muslimpakket kan utryddes sammen med alle sosialistene som har ødelagt landet vårt og de kristne tradisjonene det er bygget på”

”Vi har sendt spesialstyrker til Groruddalen hvor vi skyter alle med mørk hud som er ute etter 24:00”

”Det hadde ikke gjort meg 5 flate øre å skyte dritten ut av dritten der, måtte blitt om dagen, fordi vi ser de ikke om natta faen”

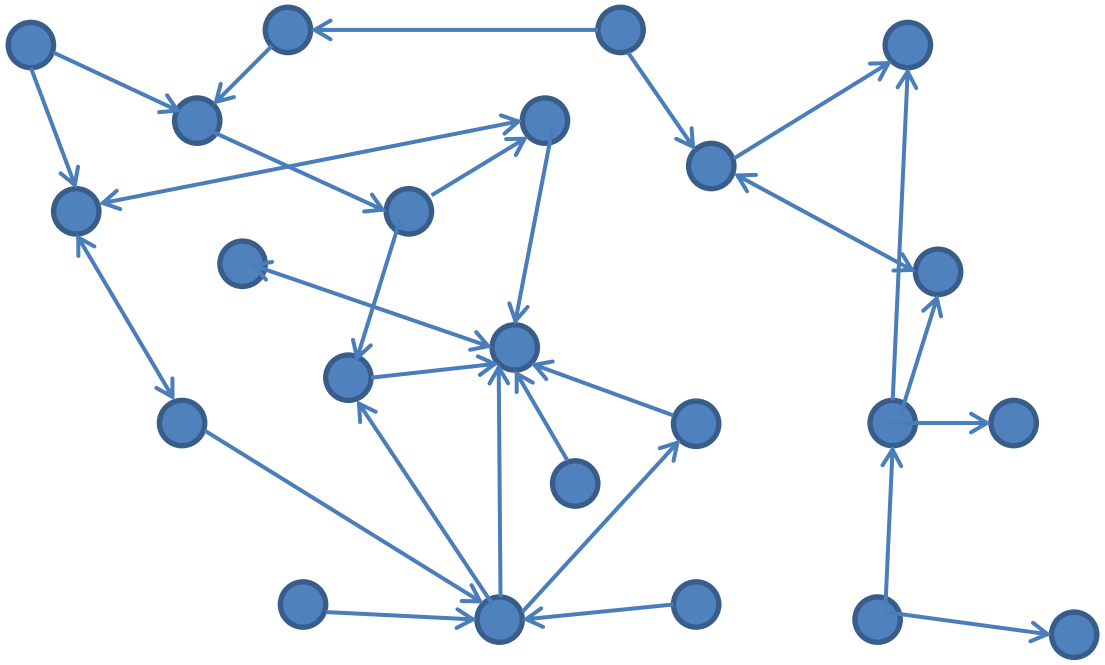
”Jeg personlig er for borgerkrig og borgervern mot utlendinger”

Kartlegging hatefulle ytringer: Eksempler på analyser

- Hvor omfattende er hatefulle ytringer ulike steder på nettet?
- Hvordan er utviklingen?
- Blir brukere mer aggressive etter lang tids deltagelse i hatefulle debatter?
- Endrer brukerne meninger?
- Hvilken effekt har ulike forebyggingstiltak?
- Hvilke nettsteder og brukere er sentrale i spredningen av hat og konspirasjonsteorier?

Relevante språkteknologiske verktøy

Web crawler



Måling aggressivitet. Enkel tilnærmelse

Glad	3	Krig	-3
Snill	2	Forræder	-4

Summer alle ord med følelse i dok. Del på antall ord.

Behover:

— Utvikle gode ordlister for måling av aggressivitet

— Mangelfullt på engelsk

— Analyse av film og produktanmeldelser

— Totalt fraværende på norsk

Hvor godt fungerer slike metoder?

”Eneste løsning er væpnet krig mot …”

”Jeg tar sterk avstand fra væpnet krig.”

”Erna i full krig med FrP.”

Mer detaljerte metoder finnes for engelsk

Opinion mining. Enkel tilnærming.

— ”Islam er en voldelig ideologi, ikke en religion. Islam må utryddes fra Norge.”

s(voldelig)/3 + s(utrydde)/2 = $-3/3-4/2=-3$

— ”Du tar feil! Islam er ikke en voldelig ideologi, men en religion med et fredfullt budskap.”

$-1\cdot$s(voldelig)/4 + s(fredfullt)/11 $= -1\cdot(-3)/4+4/11=1{,}11$

Behover:

— Mer detaljerte metoder finnes for engelsk

— Analyse av film og produktanmeldelser

— Hvor overførbart til analyse av hatefulle ytringer?

— Hvor overførbart til norsk?

Oppsummering av store mengder tekst

— Hvilke temaer diskuteres i teksten vi har samlet inn?

Mer senere i presentasjonen!

Ekstremismeprosjektet HiOA

Prosjektbeskrivelse

— Utvikle verktøy til analyse og forebygging av hatefulle ytringer på nettet

— Tett samarbeide med samfunns- og terrorismeforskere

— Disposisjon videre:

Se på to konkrete oppgaver vi jobber med relatert til språkteknologi

Oppgave 1: Tekstoppsummering

Fem norske nettsteder

— Nordfront (nf)

— Frie ytringer (fy)

— Honest thinking (ht)

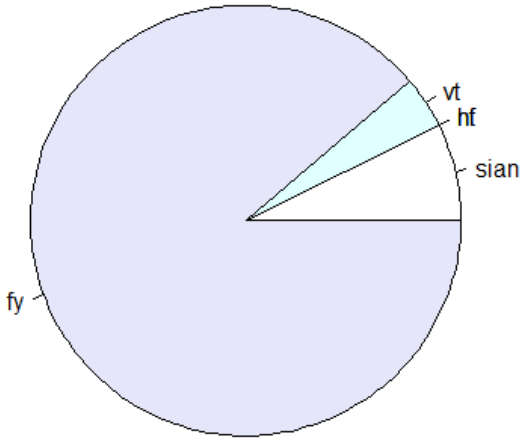
— Vigrid Tore Tvedt (vt)

— Stopp islamiseringen av Norge (sian)

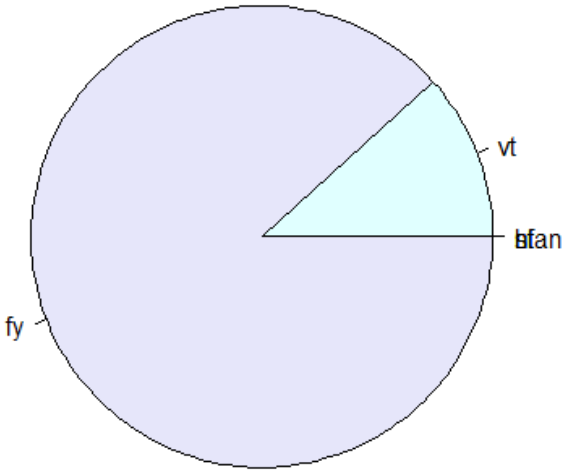
Web crawler innsamlet alle artiklene og diskusjonene fra disse 5 nettstedene (ca 10000 nettsider)

Oppsummering baserer seg på hvor ofte ulike ord brukes i de ulike nettsidene.

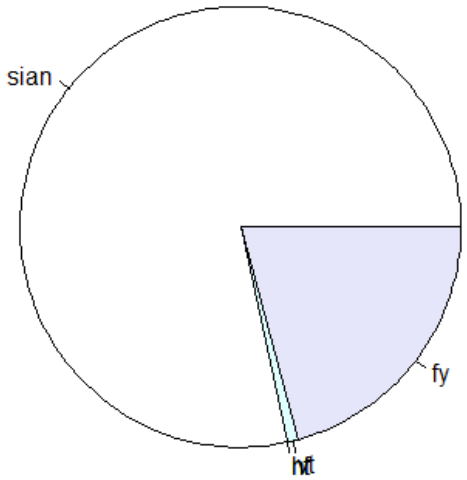
Klusteranalyse
Kluster 1



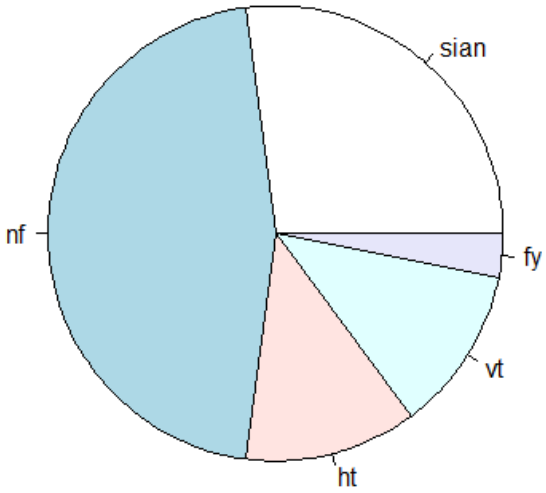
Klusteranalyse
Kluster 2



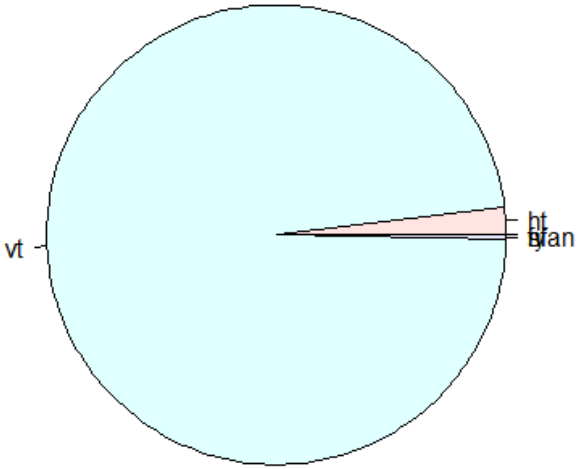
Klusteranalyse
Kluster 3



Klusteranalyse
Kluster 4



Klusteranalyse
Kluster 5



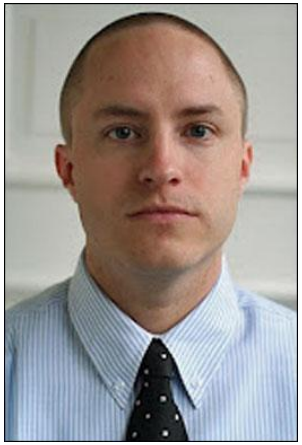
Oppgave 2:
Finne trusler om/sympati med vold

Straffeloven § 135 a

— «… straffes den som ved uttalelse eller annen meddelelse … **truer, forhåner eller utsetter for hat, forfølgelse eller ringeakt en person eller en gruppe av personer på grunn av deres trosbekjennelse, rase, hudfarge eller nasjonale eller etniske opprinnelse. Tilsvarende gjelder slike krenkelser overfor en person eller en gruppe på grunn av deres homofile legning, leveform eller orientering.** … »

Oppgave 2:
Finne trusler om/sympati med vold

Trusler om vold på internett har ført til arrestasjoner



—Eivind Berge:

«Jeg holder fast ved at politidrap er både etisk og taktisk riktig.»

—Ubaydullah Hussain (Profetens Ummah):

«Jeg skal gi dem beskyttelse jeg, inshallah. Så fort jeg har tatt jegerprøven og får tak i en AK47.»

Materiale og metode

Materiale

— Manuell klassifisering av ca 25.000 YouTube-komm.

— 1.469 trusler om/sympatier med vold

Metode

Hvilke sentrale ord forekommer sammen i en setning?

— Sanns. voldstr.<-> w1(jeg, drap, 5) + w2(drap, etisk, 3)

— Sanns. voldstr.<-> w3(jeg, ak47, 9)

Resultater

Beste prediktorer:

kill, burn, nuke, deport, exterminate, shoot, die, death, breivik, bomb, nukes, eradicated, deportation, we, exterminated, need, should

we-kill, you-die, should-killed, i-kill, hope-die, death-all, kill-yourself, i-hope, be-deported, be-executed, be-eradicated, deport-all, i-will, we-deport, be-killed, whould-die, breivik-hero, kill-them, we-will

Test:

- Unigram: 13% feilkl. (65.876 unike ord)
- Unigram+selektert bigram: 9% feilkl. (130.450 uni + bi)

Videre:

- Teste parametre (+ justeringer) på annet tekstmateriale
- Lingvistikk

Avsluttende kommentarer

Språkteknologiske verktøy

Styrker

— Mange nettsteder og store mengder tekst kan effektivt analyseres

— Objektivitet (underbygge påstander)

Utfordringer

— Ytringsfrihet og personvern

— Detaljforståelse

Behover

— Bedre verktøy for måling av aggressivitet og meninger rettet mot hatefulle ytringer. Spesielt på norsk.