HON2200: Sentimentanaltyse

Henrikke Gedde Rustad, Tor Magnus Næsset, Didrik Sten Ingebrigtsen

May 16, 2021

Sentimentanalyse er en prosess hvor man basert på tekstmateriale forsøker å fange opp affektive tilstander hos tekstens avsender. NLP (Natural Language Processing) og maskinlæringsalgoritmer har gjort det mulig å automatisere denne prosessen slik at store mengder tekst-data kan analyseres svært effektivt. I kjølvannet av dette arbeidet har man kunnet utvikle modeller som analyserer tekst og gir tilbakemeldinger til brukeren om tekstens affektive score. Det finnes ulike modeller, noen klassifiserer tekst som negativ eller positiv, mens andre har mer nyanserte kategorier som sint, trist, glad, redd osv. Sentimentanalyse brukes i stor grad av bedrifter i markedsanalyse, som en indikator på kundenes oppfatning av deres produkter og tjenester. På bakgrunn av fremveksten av internett og sosiale medier er tilfanget av tilgjengelig tekst-data fra ulike avsendere blitt mangedoblet. Dette gjør det gunstig å bruke sentimentanalyse som en prediksjon for stemninger, oppfatninger og bevegelser i folkeopinionen. På den annen side, hvis vi som samfunn skal basere beslutninger på hvordan en algoritme analyserer verden, bør vi forsikre oss om at modellen er treffsikker, og at den ikke inneholder skjulte fordommer mot noen grupper i befolkningen. Derfor har vi utforsket et datasett av norske anmeldelser som kan brukes til å lage en sentimentanalyse-modell, og undersøkt om vi finner bestemte kjønnsstrukturer i datasettet. Vi har også trent en enkel sentimentanalysemodell på datasettet for å se om en slik modell utviser kjønnsbias i sine prediksjoner.

For å lage sentimentanalyse-modellen vår har vi brukt FastText sin supervised classifier [3] og trent denne på The Norwegian Review Corpus [6], heretter kalt NoReC, et datasett bestående av norske film-, bok-, spill- og musikkanmeldelser o.l. fra norske nettsteder og tidsskrift. Datasettet består av mer enn 35 000 fullformat-tekster med en tilhørende numerisk verdi for hver tekst i form av terningkast fra 1-6, gitt av anmelderen selv. Vi ser det som en fordel at datasettet slik sett er lablet direkte av tekstenes avsendere, og at vi dermed unngår usikkerheten og potensialet for bias det innebærer å la en tredjepart vurdere den emotive scoren av tekstene.

Metoden vi bruker for å vurdere bias tilknyttet kjønn i modellen vår er delvis inspirert av the Equity Evaluation Corpus (EEC) [4] og tar utgangspunkt i et sett av setnings-maler hvor en undergruppe disse inneholder en åpen plass for emotive ord og alle inneholder åpne plasser for subjekt-ord, samt et sett av kjønnede ord som er parvis koblet sammen (f.eks. "hun"-"han", "min datter" - "min sønn"), to sett av navn og et sett med emotive ord, hvor noen beskriver en følelsestilstand og andre beskriver situasjoner. Ved å beregne den gjennomsnittlige differansen mellom den affektive scoren modellen gir for spesifikke emotive ord innsatt i setningsmalene for de korresponderende, kjønnede ordene, kan vi få et mål på hvor mye modellen tenderer til å oppjustere eller nedjustere spesifikke affektive tilstander hos et kjønn kontra et annet. Men i og med at vi har utviklet en modell som er trent på et lengre tekstformat (document level) fremfor det mer vanlige, på setninger (sentence level) eller kortere tekster (entity level), ønsket vi også å teste kjønnsbias ved å la modellen vurdere anmeldelser i test-settet før og etter vi forandret kjønnede ord i anmeldelsen, for så å sammenligne scoren modellen ga.

Modellen vår oppnår en accuracy-score på 56.8 %, så den gjør det rimelig godt i å predikere et terningkast. Når vi testet setningsmalene innsatt de korresponderende kjønnede ordene på modellen vår, beregnet vi gjennomsnittet av differansen mellom scoren til de kvinnelige og mannlige setningene. Denne gjennomsnittsdifferansen var svært liten, så algoritmen ser ikke ut til å utvise bias mot et kjønn, i hvert fall ikke når teksten er svært kort. Vi testet også modellen vår på tre tekster fra test-datasettet hvor vi hadde endret alle kjønnede ord til tilsvarende ord/navn for det andre kjønnet, og sammenlignet scoren vi fikk med scoren til de originale tekstene. Heller ikke her oppdaget vi noe avvik i score mellom de ulike tekstversjonene. En svakhet ved dette resultatet er at vi fikk testet for såpass få tekster. Å endre disse fullformat-tekstene er tidkrevende arbeid, men hvis det var gjort for hele test-datasettet hadde vi kunnet trekke tydeligere konklusjoner fra resultatet av denne testen. Modellen virker altså ikke å ha noe kjønnsbias som slår ut på innholdet i tekst. Når det gjelder kjønnet til avsenderen av teksten, ser vi at forskjellen mellom gjennomsnittlig predikert score og gjennomsnittlig faktisk score for de to kjønnskategoriene er liten, 0.06 for kvinnelige anmeldere og 0.04 for mannlige anmeldere. R2-scoren er noe høyere for kvinnelige enn for mannlige anmeldere, henholdsvis 0.362 og 0.320. Selv om denne forskjellen ikke er stor, finner vi det interessant at modellen vår gjør det noe bedre når den predikerer tekster skrevet av kvinner, til tross for at en stor andel av datasettet den er trent på består av tekster skrevet av menn.

For å se nærmere på eventuelle tendensiøse forskjeller i hvordan menn og kvinner uttrykker seg, undersøkte vi gjennomsnittlig setningslengde (i antall ord) og gjennomsnittlig ordlengde (i antall bokstaver) for de respektive

gruppene. Vi finner ingen signifikant forskjell i disse tallene. Til slutt så vi på de ordsammensetningene (lengde 1-4) hvor forholdstallet mellom frekvens hos kvinner og frekvens hos menn var størst, betinget at ordsammensetningen var brukt minst 10 ganger av den gruppen som brukte den minst. Vi legger merke til at tekniske betegnelser, ord relatert til musikkbransjen og produktnavn brukes mye oftere av menn enn av kvinner, noe som kan tyde på at menn i større grad anmelder musikk og tekniske produkter enn det kvinner gjør. Av kvinner brukes ordene "Ferrantes" og "Vigdis Hjorth" fire ganger oftere enn av menn. Dette er påfallende siden begge ordene er navn på svært anerkjente, kvinnelige forfattere. Er det slik at kvinner oftere skriver anmeldelser av kvinners verk? I så fall, kan dette ha noe å si for den prediktive kraften og eventuelle bias til en algoritme trent på et slikt datasett? Selv om ikke vår modell har vist et signifikant kjønnsbias, så kan vi ikke utelukke at en mer kompleks modell, med bedre presisjon, ville ha utvist bias. Muligens er det slik at det eventuelle kjønnsbiasen i tekstene er såpass subtil at det krever en kompleks og veltrent modell for at den skal bli påvirket av disse strukturene.

Å vurdere bias i maskinlærings-algoritmer, samt å finne metoder for å fjerne slike har blitt svært viktig fordi disse modellene påvirker livene våre i stadig større grad. Men når det gjelder å vurdere bias kan denne oppgaven være mer rett frem for noen modeller enn det er for andre. Dersom et bildegjenkjennings-verktøy systematisk klarer å identifisere lyse ansikter som menneske, mens mørke ansikter ofte identifiseres som andre pattedyr, er dette et klart bias i algoritmen. Det hersker stor enighet om hvordan man korrekt foretar en taksonomisk inndeling av objekter i verden. Derimot blir begrepet om bias mer komplisert når det er språklig mening som analyseres av en algoritme og ikke bilder. I sentimentanalyse kan bias mot spesifikke grupper manifestere seg på to måter, enten ved at tekster som omhandler èn gruppe blir systematisk vurdert til å ha en høyere sentimentverdi enn en tekst som omhandler en annen gruppe, eller at algoritmen systematisk vurderer tekster skrevet av en gruppe til å ha høyere sentimentverdi enn tekster skrevet av en annen gruppe. I det første tilfellet virker det implisitte kriteriet for å være fri for bias i metoden vår uproblematisk. Nemlig det at modellen vurderer like setninger, sett bort fra "kjønning", til å ha samme affektive ladning i seg. Dette er i tråd med hvordan EEC er brukt for å teste kjønns- og etnisitetsbias i sentimentanalyseverktøy av andre (kilde). Men i tilfellet hvor en algoritme har implisitt tilgang til avsenderens kjønn og differensierer basert på dette, er det ikke like klart at dette er et bias, i betydningen av at modellen har en iboende skjevhet i sin representasjon av verden. For er det gitt at ulike grupper alltid mener det samme i bruken av like ord og fraser, eller at de alltid vurderer verden likt?

Gapet mellom språk og mening er det mang en tenker som har forsøkt å tette. Arven fra den tyske logikeren Gottlob Frege medfører at ord, og

spesielt setninger, anses å ha en bestemt logisk betydning, de refererer til noe bestemt [2]. Hvis vi følger en slik tankegang vil en algoritme som differensierer basert på avsenderens gruppetilhørighet, klart inneholde et bias. Men er det virkelig slik at setninger har en iboende, fast betydning som er uavhengig av den som ytrer dem? En som utfordret dette synet var Ludwig Wittgenstein som i sine senere arbeid mente at vi lar oss lure av våre ords uniforme fremtreden, men at ord er i en tilstand av flux hva angår deres mening, tett knyttet til bruken av dem, og at denne bruken er mangfoldig og foranderlig [1]. Hvis vi aksepterer dette kan det være at en algoritme som beskrevet over faktisk har plukket opp ulik språkbruk mellom grupper, og gjør en god statistisk prediksjon når den gir ulik sentimentverdi til like tekster, skrevet av ulike grupper. En studie av kjønnsbias gjort med utgangspunkt i et subset av NoReC-datsettet har for eksempel funnet at kvinnelige anmeldere gir lavere terningkast til kvinnelige forfattere enn til mannlige forfattere, samt gir lavere terningkast enn det mannlige anmeldere gir kvinnelige forfattere [5]. Et slikt "kryssbias" mellom grupper er også noe en sentimentanalysemodell kan inkludere i sine beslutninger, og er noe det kan være viktig å kontrollere for i et treningsdatasett.

I NoReC-datasettet finner vi at ord som brukes proposjonalt høyere av menn enn kvinner har med teknologi, musikk og produkter å gjøre, mens kvinner brukte bestemte kvinnelige forfatteres navn i vesentlig større grad enn menn. I kompleksitet, setningslengde og ordlengde, finner vi ingen signifikante forskjeller. Modellen vår utviser ikke kjønnsbias basert på innhold i tekst eller basert på anmelderens kjønn, men det er ikke dermed sagt at dette vil gjelde for en mer kompleks modell trent på samme datasett. Et viktig moment i videre utforskning vil være å utvikle et parallelt test-datasett hvor alle kjønnede ord er byttet om fra den originale utgaven, slik at en grundigere test av kjønnsbias i modellen kan foretas. I tillegg vil det være interessant å se hvordan mer komplekse modeller responderer på treningsdataene. Vil en slik modell få et mer betydelig kjønnsbias fra dette datasettet?

References

[1] Anat Biletzki and Anat Matar. "Ludwig Wittgenstein". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University, 2020.

[2] Simon. W. Blackburn. "Philosophy of language". In: *Encyclopedia Britannica* (2017). URL: `https://www.britannica.com/topic/philosophy-of-language`.

[3] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the*

Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Apr. 2017, pp. 427–431.

[4] Svetlana Kiritchenko and Saif M. Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems". In: *CoRR* abs/1805.04508 (2018). arXiv: 1805.04508. URL: http://arxiv.org/abs/1805.04508.

[5] Samia Touileb, Lilja Øvrelid, and Erik Velldal. "Gender and sentiment, critics and authors: a dataset of Norwegian book reviews". In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 125–138. URL: https://www.aclweb.org/anthology/2020.gebnlp-1.11.

[6] Erik Velldal et al. "NoReC: The Norwegian Review Corpus". In: *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*. Miyazaki, Japan, 2018, pp. 4186–4191.