Maskininlärning inom innehållsmodererering

Erik Wihlman 41860

Kandidatavhandling i datateknik

Handledare: Sepinoud Azimi

Fakulteten för naturvetenskaper och Teknik

Åbo Akademi

2021

Referat

Innehållsmoderering har på under år 2020 varit ett mycket diskuterat ämne, särskilt på grund av presidentvalet i USA och koronapandemin. Spridningen av falsk information samt annat skadligt material är lättare än någonsin förr tack vare sociala medier. Avsikten med denna avhandling är att redogöra vad innehållsmoderering går ut på och hur maskininlärning kan vara till nytta, hur vissa prominenta plattformar redan utnyttjar maskininlärning, samt problem och utmaningar som förekommer med automatiserad innehållsmoderering.

Sökord: maskininlärning, innehållsmoderering

Innehåll

1. Inledning

I och med att alltmer av diskussionen kring vårt samhälle flyttar till sociala medier, ökar också behovet för innehållsmoderering på de olika sociala plattformarna. Sociala mediernas användning har ökat stadigt sedan början av 2000-talet, med Facebook och YouTube som har kring 2 miljarder användare var [7]. Informationsflödet är nu friare än någonsin, vilket innebär vissa problem. Särskilt under koronapandemin och presidentvalet i USA har falsk information varit ett stort problem [9, 5]. Bland annat Twitter är en av plattformarna som har hamnat vidta nya åtgärder mot spridningen av falsk information [27].

Med dessa ökande användarmängder dyker också frågan om skalbarhet upp. Hur plattformarna modererar innehåll har också blivit en viktig del av diskussionen [12, 14]. Hur kan de olika plattformarna hålla jämna steg den växande arbetsmängden? Det är dock inte endast skalbarhet som är ett problem inom innehållsmoderering, utan även vissa hälsorelaterade risker finns, särskilt då det gäller moderatorernas psykiska hälsa. I många fall blir dessa människor utsatta för traumatiska bilder och videor, som i värsta fall orsakar långvariga psykiska problem [16].

Denna avhandling utforskar hur maskininlärning kan vara till nytta för att lösa detta problem, inte endast med skalbarhet i åtanke, utan också hur vi kan minimera skadan för människor. I kapitel 2 beskrivs innehållsmoderering, vad det innebär, vem modererar, samt vissa risker som moderatorerna hamnar ut för. Kapitel 3 går igenom vad maskininlärning är i allmänhet samt vissa relevanta exempel för modeller och träning som kan utnyttjas för moderering. Kapitel 4 beskriver tillämpningar med specifik fokus på prominenta plattformar som oftast befinner sig i nyheter och diskussion kring innehållsmoderering. Användning av maskininlärning för innehållsmoderering innebär vissa problem och risker som bör tas i beaktande, dessa beskrivs i kapitel 5.

2. Innehållsmoderering

2.1 Bakgrund

Innehållsmoderering är sociala plattformarnas sätt att reglera materialet som användarna av plattformen kan ladda upp och publicera. Stilen för moderering kan variera drastiskt mellan de olika plattformarna, allt från just och just raderandet av olagligt material till mer stränga regler om vad man kan uttrycka. Vem som upprätthåller reglerna kan också variera mellan plattformar, beskriver Caplan i Content or Context Moderation [18]. På vissa plattformar kan modereringen vara mera community driven, det vill säga de olika grupperna som användarna bildar ansvarar för att upprätthålla förutom sina egna regler, också en del av hela plattformens regler. Exempel av detta är plattformar som Reddit och Discord, dessa plattformar har också sina egna innehållsmoderatorer som ansvarar för att se till att alla de olika grupperna upprätthåller regler på ett tillfredställande sätt, samt samarbeta med olika rättsväsenden då det är aktuellt.

Till största delen handlar det ofta om att en moderator går systematisk igenom en kö av material som användarna har flaggat som skadligt eller annars obehagligt. Modereringen är alltså reaktiv till en stor del, dock används vissa modeller för proaktiv moderering, det vill säga innehållet blockeras innan den kan laddas upp och publiceras. Dessa varierar från enkla filter som reagerar på vissa ord på foruminlägg till mer komplexa modeller som klassificerar helheter.

2.2 Mänsklig skada

I och med att innehållsmoderering ofta handlar om att radera olagligt och skadligt material, finns det vissa risker för moderatorerna. Främst handlar dessa risker om psykisk hälsa och deras långvariga effekter hos moderatorerna, särskilt de som jobbar länge inom moderering [16]. Företag som Facebook har ibland varit i nyheterna angående deras innehållsmoderering och hur de behandlar sina innehållsmoderatorer, eller hur deras leverantörer behandlar sina moderatorer. I synnerhet har moderatorer berättat om sina upplevelser med dåligt stöd för upprätthållandet av psykisk hälsa efter lång exponering till traumatiskt material.

De som jobbar inom innehållsmoderering blir ofta tvungna att gå med i sekretessavtal, som begränsar vem de kan diskutera med. Detta är problematisk särskilt om företaget inte ger tillgång till professionella psykologer eller motsvarande för att upprätthålla psykisk hälsa hos moderatorerna. Ibland blir

omständigheterna såpass omåttliga att moderatorerna bryter mot sekretessavtalen och publicerar sina berättelser om deras upplevelser.

I många fall leder dessa berättelser ingenvart, men år 2017 meddelade Mark Zuckerberg i ett blogginlägg att Facebook skulle börja använda artificiell intelligens för innehållsmoderering för att identifiera skadligt och riskabelt material, som terrorism, våld, mobbande och hot om självskada[1].

2.3 Skalbarhet och mänskliga faktorn

Psykisk hälsa är inte det enda problemet med de växande mängderna innehåll som laddas upp, utan skalbarhet blir snabbt ett problem också. Med varje ny användare som registrerar sig blir arbetsmängden större, särskilt om användaren är en aktiv medlem som deltar i diskussioner och publicerandet av material. Skalbarhet blir en alltmer pressande fråga i och med att lagstiftare överallt i världen kräver allt snabbare respons och ställning mot terrormaterial och spridning av falsk information. Efter stormningen av Capitol byggnaden i Washington D.C. blev flera sociala plattformars verkställande direktörer framkallade framför representanthuset, skriver Rebecca Kern för Bloomberg Law [20]. Temat för förhöret var hurudant ansvar sociala plattformar bör ta för spridningen av falsk information. Naturligtvis kommer mer ansvar att leda till större krav på prestanda och därmed blir skalbarhet ett relevant problem.

Ett anmärkningsvärt problem med människocentrerad moderering är tiden det tar för vissa material att raderas, i synnerhet då det är fråga om bland annat våld och terrorism. Särskilt Facebook är en av plattformarna som oftast är i nyheterna på grund av långsam moderering och uppenbarlig tillåtelse av materialet. I värsta fall hålls material upp i flera dagar innan det tas ned. Dessa problem kunde lösas, åtminstone delvis, med hjälp av maskininlärning.

[1] BBC artikel om Facebooks meddelande https://www.bbc.com/news/technology-38992657
Zuckerbergs inlägg på Facebook https://www.facebook.com/notes/3707971095882612/

3. Maskininlärning, en överblick

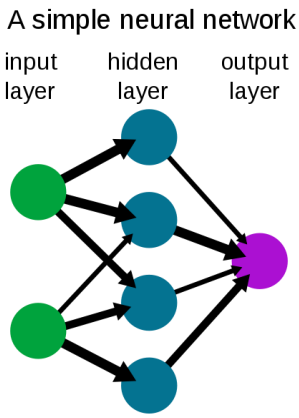
I The Hundred Page Machine Learning Book definierar Andriy Burkov maskininlärning som en maskin som på basis av inmatade data löser ut en formel som kan användas för att lösa liknande problem [1]. En stor del av maskininlärning handlar om att identifiera statistiska likheter i data och sedan med en viss sannolikhet förutspå någonting annat givet liknande data. Detta betyder också att maskiner inte egentligen lär sig, åtminstone inte så som vi förstår lärandet. Maskinerna har ingen intuition, vilket i sin tur betyder att vissa, specifika ändringar i data kan förvränga det slutliga resultatet på ett oönskat sätt. Exempelvis givet en modell som kan identifiera figurer ur en bild, kan vridningen på bilden ge olika, eller inga resultat, om inte modellens träningsdata har haft förvrängda bilder med.

3.1 Modeller

Inom maskininlärning beskriver en modell det som maskinen lärt sig från träningsdata. Dessa varierar i komplexitet och storlek beroende på vilken algoritm som använts samt hur mycket data som använts under träningen. Många av modellerna baserar sig på neurala nätverk eller en variant av neurala nätverk, dock vissa andra modeller kan sporadiskt komma upp.

3.1.1 Neurala nätverk

Neurala nätverk är nätverk av neuroner, där varje neuron har en specifik uppgift att identifiera någon del av en större helhet. Dessa nätverk byggdes ursprungligen för att lösa problem lik en människohjärna, där vissa grupper av neuroner aktiveras under de rätta omständigheterna. Neurala nätverk kan ha varierande mängder av nivåer och neuroner, dock är de oftast begränsade till 2 mellanlager, i och med att större nätverk stöter på problem med prestanda, se figur 1 för en enkel modell.



Figur 1: Förenklad illustrering av ett neuralt nätverk [28]

Varje neuron har en aktiveringsfunktion som får data från den föregående nivån, och i vissa fall tillbaka som feedback. Den aktiveringsfunktion som används i den sista nivån bestämmer vilken sorts modell det är fråga om, alltså antingen en klassificerings- eller regressionsmodell.

Devlin, Chang, Lee och Toutanova introducerade år 2019 en ny modell för klassificering av text som bygger på äldre modeller av neurala nätverk [11]. Idén med Bidirectional Encoder Representations from Transformers (BERT) är att bygga upp en kontext baserad modell för klassificering av språk. Traditionellt flödar neurala nätverk i en riktning, vilket var motiveringen för utvecklingen av BERT, i och med at kontext från båda riktningarna är viktiga för den slutliga klassificeringen. Denna sort av språkteknologi är ännu inte allmänt utspridd, men Google har rapporterat att alla sökströngar körs igenom BERT [17].

Wang, Weeds och Comley undersöker hur olika modeller kunde utnyttjas inom innehållsmoderering på forum, varav en var BERT [6]. Träningsdata bestod av 22 487 inlägg från 1000 olika användare med dataetikett med bland annat kön, ålder och en anonymiserad identifierare. Utöver dessa etiketter, var varje inlägg markerad med antingen ”reject” eller ”accept”. Undersökningens målsättning var att demonstrera användbarheten av 3 olika modeller med dataförstoring, logistisk regression, Long Short Term Memory (LSTM) och BERT. Dataförstoringen i detta fall går ut på att skapa nya inlägg genom att ersätta vissa ord som inte ändrar betydelsen av hela inlägget på ett märkvärdigt sätt med synonymer. I över lag lyckades modellerna klassificera inläggen med hög noggrannhet, med BERT som marginellt nådde den största noggrannheten. Forskarna påpekar dock att dessa modeller kräver stora mängder data för bra resultat.

3.1.2 Stödvektormaskin

Stödvektormaskiner fungerar utifrån principen att data kan delas upp i två grupper. Fastän stödvektormaskiner egentligen är byggda för övervakad inlärning, kan de användas med oövervakad inlärning ifall träningsdata inte har färdiga klassificeringar. I dessa fall söker man ofta efter gemensamheter inom den inmatade data via klusteranalys. Dessa maskiner fungerar oftast bäst med stora mängder data som kan generaliseras.

Stödvektormaskiner kan nå liknande resultat för noggrannhet som ett neuralt nätverk, dock är deras problemställning olika. En stödvektormaskin som klassificerar inlägg binärt på Twitter kan klassificera inläggen som hatfulla med 97% noggrannhet [23]. Modellen är alltså effektiv om målsättningen är att endast klassificera inlägg med en enda klassificering.

3.2 Träning

Träning av en modell kan ske på flera olika sätt, varav de mest bemärkta är förstärkningsinlärning, övervakad inlärning, oövervakad inlärning samt en hybridmodell av de föregående, halv övervakat lärande. Det finns flera utöver dessa, men de flesta metoderna är inte relevanta för denna avhandling. För att träning kan ske över huvud taget, måste träningsdata vara i rätt form, alltså i vektorform där datas särdrag är representerade, oftast i numerisk form, dock vissa modeller kan och måste behandla data i textform, särskilt då det är fråga om NLP (språkteknologi, Natural Language Processing).

I boken Reinforcement Learning: An Introduction beskriver Sutton och Barto förstärkningsinlärning som ett samtida problem och mängd lösningar [21]. Idén med förstärkningsinlärning är att maskinen får agera i en omgivning som ger feedback för varje åtgärd maskinen vidtar. Om maskinen gör någonting bra, får den positiv feedback som indikerar att det lönar sig att göra samma sak på nytt nästa gång, naturligtvis blir feedbacken negativ ifall maskinen gör någonting dåligt.

Hu et al. visar att förstärkningsinlärning kan användas för att förutse popularitet av foruminlägg [12], vilket i många fall blir nyttigt för moderatorer [4]. Kännedom om vilka inlägg som kommer att få mycket trafik kan hjälpa moderatorer med dirigering av resurser, särskilt i fall där inläggets innehåll kan anses vara upphetsande.

Övervakad inlärning går ut på att den inmatade data är parad ihop med den önskade klassificeringen som maskinen skall nå. Algoritmen kan då räkna ut en sådan funktion som uppfyller kriteriet. Beroende på modellen, kan klassificeringen i träningsdatan vara text eller siffror. Denna metod används bland annat i [6], där inläggen i träningsdata fick slutliga klassificeringen.

Under oövervakad inlärning saknar träningsdata däremot den slutliga klassificeringen och det är maskinens uppgift att hitta gemensamheter i data.

4. Tillämpningar och effektivitet

Hur maskininlärning kan tillämpas inom innehållsmoderering varierar enligt vad man vill att maskinen gör och hur. Vilken modell och metod för träning som lönar sig beror på målsättningen, de olika modellerna och metoderna klarar av vissa uppgifter bättre och andra sämre, dock i många fall kan flera modeller göra samma sak.

Maskininlärning används redan nu på plattformar som Facebook, Twitter och YouTube för att kämpa mot terrorism och våld. Enligt Global Internet Forum to Counter Terrorism (GIFCT) [8] organisationen lyckades dessa plattformar förhindra eller flagga uppladdningen och publicerandet av terrormaterial med lovande resultat. På YouTube lyckades maskininlärningsmodellen flagga 98% av det material som moderatorerna bekräftade och raderade. Twitter däremot rapporterar att mellan juli och december år 2017 hade de stängt 274 460 konton som tillhörde extremister, varav 93% var flaggade av deras algoritmer och 74% av dem var stängda innan de kunde posta någonting. På Facebook raderas 99% av all ISIS och Al Qaeda terrormaterial före en användare flaggar det.

Gorwa, Binns och Katzenbach [19] beskriver några framträdande implementationer av automatiserad moderering. Vissa av dessa implementationer är hash baserade och har inte direkt någonting att göra med maskininlärning, men är ändå anmärkningsvärda på grund av deras påverkan inom innehållsmoderering. De hash baserade implementationerna är YouTubes Content ID, GIFCT:s Shared-industry hash database och Microsofts PhotoDNA. Var och en av dessa har olika uppgifter men fungerar på den samma underliggande principen att matcha något material mot en hash databas. Om en likadan hash hittas i databasen bör materialet flaggas eller tas ner. Google, Twitter och Facebook däremot använder sig av maskininlärning för Perspective API[2], kvalitets filter respektive hatfull textklassificerare. Skillnaden mellan de maskininlärningsbaserade implementationerna och hash implementationerna är hurudant material de skall matcha. Hash systemen används för ljud, video och bilder, medan maskininlärningen används främst för text och konton.

[2] Vidare material om Perspective https://perspectiveapi.com/how-it-works/

4.1 Facebook

Facebook beskriver hur de närmar sig problemet med maskininlärning i ett blogginlägg [15], anmärkningsvärt handlar största delen av Facebooks ansträngningar om terrormaterial. Systemet utvärderar och poängsätter varje inlägg på plattformen. Ifall poängsättningen överskrider en gräns sätts inlägget i en kö för vidare utvärdering. Desto högre poängsättningen är, desto större prioritet får inlägget och i vissa fall skickas de vidare för specialister. I fallen då systemet har väldigt hög konfidens i utvärderingen kan inlägget till och med raderas automatiskt. Facebook påpekar också att varken människor eller maskiner är helt felfria, utan misstag händer och det är viktigt att ha något sätt att överklaga ett beslut om bestraffningar på plattformen. Mer om hurudana misstag maskiner är mottagliga för i nästa kapitel.

Facebook rapporterar i samma blogginlägg också en stor förökning i mängden raderat material som resultat av de nya systemen. I första kvartalet av 2018 raderades 1,9 miljoner inlägg, varav 640 000 var gamla. I andra kvartalet togs den nya modellen i bruk och Facebook rapporterar att de tog ner 9,4 miljoner inlägg av terrormaterial, dock var största delen av materialet gammalt som söktes upp med vad Facebook kallar specialiserade tekniker. I tredje kvartalet rapporterar Facebook att de raderade ytterligare 3 miljoner inlägg, varav 800 000 var gamla. Enligt rapporten hittade Facebook 99% av materialet själv under båda kvartalen, inte med hjälp av användarnas flaggor. Jämfört med resultat från det gamla systemet blir det självklart att med hjälp av maskininlärning har modereringen blivit effektivare. Facebook betonade att det är viktigare att fokusera på hur snabbt materialet sprids snarare än hur snabbt det raderas från plattformen.

4.2 YouTube

Utöver Content ID har YouTube också klassificeringsmodeller för moderering [19]. Lik Facebook, är en stor del av YouTubes ansträngningar mot terrormaterial. YouTube har dock haft problem med sin moderering, särskilt med modellens överreaktion till material som inte tillhör terrormaterial eller hatfull klassificeringen, skriver Internet Creators Guild [10]. 2017 skrev YouTubes VD, Susan Wojcicki, ett blogginlägg kring YouTubes ansträngningar för förbättring av deras dåtida system för moderering [25]. I inlägget betonar Wojcicki betydelsen av mänsklig insats både för träningen av maskininlärningsmodeller och för raderandet

av material, ”[h]uman reviewers remain essential to both removing content and training machine learning systems because human judgment is critical to making contextualized decisions on content”.

Enligt Wojcickis inlägg har moderatorerna på YouTube utvärderat nästan 2 miljoner videor med terrormaterial under ungefär ett halvår, varav 150 000 raderades. Det är dock inte endast videorna som ställer till problem, utan kommentarerna under videorna måste också behandlas och träning för en modell för kommentarer började snart därefter. Ett bekant tema dyker upp igen, maskininlärningen kompletterar mänskliga moderatorer genom att flagga material för utvärdering. Fastän inlägget betonar maskininlärningens roll, påminner Wojcicki om den mänskliga insatsens betydelse för kampen mot oönskat material.

4.3 Twitter

Twitter däremot har haft problem med spridning av falsk information mer än något annat. Särskilt under 2020 presidentvalen i USA och koronapandemin, som slutade med att före detta president Donald Trump blev spärrad från plattformen[3] efter flera falska påståenden samt uppmaning till våld. I början av mars publicerade Twitter ett nytt inlägg som beskriver hur de kommer att bekämpa spridningen av falsk information kring korona vacciner med maskininlärning [27]. Målet är att ha en maskininlärningsmodell som jobbar tillsammans med moderatorerna för att identifiera falsk information samt att dirigera publiken till relevanta artiklar med verklig information.

Enligt inlägget börjar detta med att moderatorerna stämplar inlägg med falsk information som sedan matas in i maskinen för träning. Twitter säger att målet kommer att vara att få detta system att fungera på flera språk, men att det kommer att ta sin tid. Det är dock ännu inte klart hur effektivt detta system kommer att vara, men i och med att Facebook och YouTube har haft framgångar med detta kommer Twitter antagligen snart efter.

Fastän maskininlärning och andra automatiserade system för tillfället används som komplement till mänskliga moderatorer finns det förhoppningar för fullt

[3] Twitter Inc. “Permanent suspension of @realDonaldTrump”, *Twitter blog*, 8 januari 2021 https://blog.twitter.com/en_us/topics/company/2020/suspension.html

automatiserade system inom industrin, skriver Gollatz, Beer och Katzenbach i Alexander von Humboldt Institute for Internet and Society (HIIG) Workshop rapporten [14]. Dock inte utan invändningar från andra experter som anser dessa förhoppningar vara en del av den bredare entusiasmen för maskininlärning och artificiell intelligens. För den förutsebara framtiden kommer moderering till största delen att vara en hybridmodell av automatiserad och mänsklig insats.

5. Risker och utmaningar

Användning av maskininlärning och automatisering i allmänhet för moderering innebär vissa risker och utmaningar som bör tas i beaktande vid implementering. Människor har alltid vissa fördomar som bör beaktas vid träning av maskiner, särskilt då de skall utvärdera inlägg från människor överallt i världen [3]. Det är ändå inte endast mänskliga fördomar som bör tas i beaktande, utan bias inom träningsdata är ett problem som kan förekomma. Kontextuell bedömning är också en viktig förmåga [18], exempelvis en video som laddas upp på YouTube kan få helt olika betydelse beroende på vem som laddat upp den och varför.

5.1 Bias

Det är inte sällan plattformar blir anklagade för att vara politiskt partiska då det gäller moderering. Bland annat Donald Trump och Republikanska partiet var mycket högljudda om sociala mediernas möjliga partiskhet mot konservativa röster [22]. Jiang, Robertson och Wilson undersökte om dessa klagomål hade någon verklig grund på YouTube [24]. Undersökningen kom till slutsatsen att kommentarer från konservativa individer blev modererade oftare, men inte på grund av politisk ideologi, snarare på grund av att dessa individer oftare skrev kommentarer som bröt mot reglerna.

Med maskininlärningsmodeller och artificiell intelligens i allmänhet kan statistisk bias bli ett problem då träningsdatan inte är tillräckligt representativt [3]. I fallen där vissa datamängder är för små, kan metoder som dataförstoring användas [6] för att nå mer representativa data. Vidare åtgärder som [3] föreslår är anonymisering av data, alltså ett inlägg separeras fullständigt från den person som publicerat det. På detta sätt försäkras maskinen från att göra beslut på individuella egenskaper och materialet bedöms endast på sitt eget värde.

Ett problem som plattformar bör ta i beaktande vid träning av modeller, särskilt med data som består av äldre modereringsbeslut är att i många fall ändrar reglerna på plattformarna. Detta betyder att dessa ändringar bör reflekteras i träningsdata, annars kan modellen lära sig föråldrade ställningar [26].

I och med att största delen av bias inom maskininlärning har att göra med identiteten av den utvärderade lyckas innehållsmoderering till stora delar undvika många av fallgroparna vid träning av maskininlärningsmodeller. Ytterligare är

plattformarna isolerade från varandra då det gäller moderering, alltså behöver modellen inte vara generaliserad för alla sociala medier. Problemet handlar därmed mindre om bias, utan problemet blir kontext och hur modellen borde bedöma någonting givet specifika omständigheter.

5.2 Kontextuell bedömning

Förmågan att bedöma kontext tillsammans med materialet är också kritiskt i många fall. Under COVID-19-pandemin blev YouTube kritiserad för deras modereringsstrategi gällande videor om pandemin. Enligt The Verge rapporterade YouTube ca 11 miljoner raderade videor [13], varav 320 000 överklagades. Detta representerar en ungefär 100% ökning i överklaganden och ungefär hälften av de överklagade videorna blev återställda. Den större än tidigare mängden överklaganden samt återställningar tyder på att YouTubes modell har överreagerat till COVID-19 relaterade videor. Om videon handlade om COVID-19 var sannolikheten för dess radering hög, oberoende på hur faktabaserad videon var. Videor från kvalificerade läkare var under samma risk för censurering som videor från konspirationsteoretiker.

Falsk information om pandemier och sjukdomar är inte det enda YouTubes modeller censurerar, utan många videor kring LGBTQ teman har i många fall drabbats av olika former av censur [2]. Modellen reagerade negativt till LGBTQ terminologi, oberoende av sammanhanget kring orden. Detta anknyter tillbaka till bias i träningsdata med för lite representation.

Kontextuell bedömning blir också relevant med geografi i tanke. Olika länder har olika lagar gällande uttryckande av åsikter och påståenden, bland annat Tyskland har bland de mest stränga lagarna om hatfulla uttryck, medan i USA kan man uttrycka ungefär vad som helst [29].

5.3 Etiska frågor

Gillespie lyfter fram frågan om vi ens borde automatisera innehållsmoderering fullständigt. Han argumenterar att fallen där det är fråga om att fullständigt spärra en människa från en plattform borde besluten göras endast av andra människor, inte en maskin [26]. En maskin förstår inte de värden vi som ett samhälle har, utöver detta ändrar dessa värden beroende på var i världen man vistas.

6. Slutsats

Innehållsmoderering är alltså en väsentlig del av upprätthållningen på nätsidor som tillåter användarna att ladda upp och publicera material. Den viktigaste delen av arbetet är ännu manuellt med människor som går igenom köer av flaggad material, både automatiska och manuella flaggor. Ofta är materialet som moderatorerna måste gå igenom förkastligt och till och med psykiskt skadande i värsta fall.

Neurala nätverk och dess varianter verkar vara populära val för innehållsmoderering, antagligen på grund av deras mångfaldighet, men detta är en fråga för någon annan. Det betyder ändå inte att andra modeller som inte kunde fungera för moderering, beroende på kraven som plattformen anger. Stödvektormaskiner har visat sig vara ett alternativ i vissa fall då mångfaldig klassificering inte är nödvändig.

Några av de största plattformarna har länge använt algoritmisk moderering, men under de senaste åren har många investerat i maskininlärning, med varierande resultat. I många fall har användningen av maskininlärning varit nyttigt, men vissa problem har ändå förekommit. Särskilt YouTube har blivit kritiserad för överdriven moderering.

7. Litteraturförteckning

[1] A. Burkov, *The Hundred-Page Machine Learning Book*, 1st ed. 2019 [E-book] PDF, http://themlbook.com/

[2] A. Romano, ”A group of YouTubers is trying to prove the site systematically demonetizes queer content”, *Vox*, 10 oktober 2019
https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report
[Hämtad: 26 mars, 2021]

[3] “Use of AI in online content moderation”, Cambridge Consultants, 18 juli 2019
https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-content-moderation
[Hämtad: 29 mars, 2021]

[4] C. Lampe, P. Resnick, “Slash(dot) and burn: distributed moderation in a large online conversation space”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, april 2004
https://dl.acm.org/doi/10.1145/985692.985761

[5] D. Orso, N. Federici, R. Copetti, L. Vetrugno, T. Bove, “Infodemic and the spread of fake news in the COVID-19-era”, *US National Library of Medicine*, 4 maj 2020
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7202120/
[Hämtad: 29 mars, 2021]

[6] D. Wang, J. Weeds, I. Comley, “Improving mental health using machine learning to assist humans in the moderation of forum posts”, *Health Informatics*, vol 5, mars 2020,
http://sro.sussex.ac.uk/id/eprint/89769/

[7] E. Ortiz-Ospina, “The rise of social media”, *Our World in Data*, 18 september 2019
https://ourworldindata.org/rise-of-social-media
[Hämtad: 29 mars, 2021]

[8] Global Internet Forum to Counter Terrorism, “About our Mission”, maj 2019
https://perma.cc/44V5-554U
[Hämtad: 22 mars, 2021]

[9] G. Miller, “As U.S. election nears, researchers are following on the trail of fake news”, *Sciencemag*, 26 oktober 2020
https://www.sciencemag.org/news/2020/10/us-election-nears-researchers-are-following-trail-fake-news
[Hämtad: 29 mars, 2021]

[10] Internet Creators Guild, “YouTube De-Monetization Explained”, *Medium.com*, september 2016
https://perma.cc/R7AQ-MG2F
[Hämtad: 23 mars, 2021]

[11] J. Devlin, M-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 24 maj 2019

https://arxiv.org/pdf/1810.04805.pdf
[Hämtad: 29 mars, 2021]

[12] J. He, M. Ostendorf, X. He, J. Chen, J. Gao, L. Li, L. Deng, "Deep Reinforcement Learning with a Combinatorial Action Space for Predicting Popular Reddit Threads", *Microsoft*, November 2016
https://www.microsoft.com/en-us/research/publication/deep-reinforcement-learning-combinatorial-action-space-predicting-popular-reddit-threads/
[Hämtad: 29 mars, 2021]

[13] J. Vincent, "YouTube brings back more human moderators after AI systems over-censor", *The Verge*, 21 september 2020
https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns
[Hämtad: 26 mars, 2021]

[14] K. Gollatz, F. Beer, C. Katzenbach, "The Turn to Artificial Intelligence in Governing Communication Online", *Alexander von Humboldt Institute for Internet and Society*, september 2018
https://www.ssoar.info/ssoar/bitstream/handle/document/59528/ssoar-2018-gollatz_et_al-The_Turn_to_Artificial_Intelligence.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2018-gollatz_et_al-The_Turn_to_Artificial_Intelligence.pdf
[Hämtad: 24 mars, 2021]

[15] M. Bickert, B. Fishman, "Hard Questions: What Are We Doing to Stay Ahead of Terrorists", *Facebook*, november 2018
https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/
[Hämtad: 23 mars, 2021]

[16] O. Solon, "Underpaid and overburdened: the life of a Facebook moderator", *The Guardian*, 25 maj 2017
https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content
[Hämtad: 29 mars, 2021]

[17] P. Nayak, "Understanding searches better than ever before", *The Keyword*, 25 oktober 2019
https://www.blog.google/products/search/search-language-understanding-bert/
[Hämtad: 29 mars, 2021]

[18] R. Caplan, "Content or Context Moderation", *Data & Society*, 2018
PDF, https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf
[Hämtad: 22 mars, 2021]

[19] R. Gorwa, R. Binns, C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance", *Big Data & Society*, vol 7, nr. 1, januari 2020.
PDF, https://journals.sagepub.com/doi/pdf/10.1177/2053951719897945

[20] R. Kern, "House to Confront Tech CEOs Over Online Spread of False Info", *Bloomberg Law*, 22 mars, 2021
https://news.bloomberglaw.com/tech-and-telecom-law/house-to-confront-tech-

ceos-over-online-spread-of-false-info
[Hämtad: 24 mars, 2021]

[21] R. S. Sutton, A. G. Barto *Reinforcement Learning: An Introduction* 2nd ed. 2015 [E-book]
PDF,
https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

[22] Reuters Staff, "Republicans renew complaints Twitter sifles president, conservatives", *Reuters*, 9 juli, 2020
https://www.reuters.com/article/us-twitter-trump-republicans-idUSKBN24937F
[Hämtad: 25 mars, 2021]

[23] S. Agarwal, A. Sureka, "Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter", *Lecture Notes in Computer Series*, vol 8956, 2015
https://doi.org/10.1007/978-3-319-14977-6_47

[24] S. Jiang, R. E. Robertson, C. Wilson, "Reasoning about Political Bias in Content Moderation", *AAAI Conference on Artificial Intelligence*, vol 34, februari 2020
https://ojs.aaai.org/index.php/AAAI/article/view/7117

[25] S. Wojcicki, "Expanding our work against abuse of our platform", *YouTube Official Blog*, dec. 2017
https://blog.youtube/news-and-events/expanding-our-work-against-abuse-of-our
[Hämtad: 24 mars, 2021]

[26] T. Gillespie, "Content moderation, AI, and the question of scale", *Big Data & Society*, vol 7, nr 2, juli 2020
https://journals.sagepub.com/doi/full/10.1177/2053951720943234

[27] Twitter Safety, "Updates to our work on COVID-19 vaccine misinformation", *Twitter blog*, 1 mars, 2021
https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html
[Hämtad: 25 mars, 2021]

[28] Wikipedia, *Simplified view of a feedforward artificial neural network*,
https://en.wikipedia.org/wiki/Neural_network#/media/File:Neural_network_example.svg

[29] Z. Laub, "Hate Speech on Social Media: Global Comparisons", *Council on Foreign Relations*, 11 april 2019
https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons
[Hämtad: 30 mars, 2021]