RTL

tegen online haat

Data-onderzoek naar de hoeveelheid online haat op onze Instagram kanalen

2024

Inhoud

Inleiding

Op 3 augustus dit jaar vond de jaarlijkse Amsterdam Canal Parade plaats en RTL was erbij. Onze boodschap was ‘Be Sweet!’: een oproep aan iedereen in Nederland om wat liever voor elkaar te zijn, en te blijven. We ervaren een toename van online haat en maken ons zorgen over de effecten ervan.

De boodschap werd langs de grachten in Amsterdam met liefde omarmd. Het contrast met het geluid online was echter groot: ondanks onze oproep tot verbinding werd daar pijnlijk duidelijk waarom het nog zo erg nodig is om bij dit onderwerp stil te staan. In reactie op deze haat gaan we online het gesprek aan en blijven we zoeken naar de verbinding. In sommige gevallen lukt dat niet en zijn we genoodzaakt comments te verwijderen of, in een uitzonderlijk geval, commentsecties te sluiten. Actieve webcare is onderdeel van onze werkwijze, wat betekent dat onze social redacteuren zich vaak moeten verhouden tot dit onderwerp. Veel van hen geven aan dat online haat, intimidatie en negativiteit inmiddels een dagelijks onderdeel van hun werk is geworden.

De afgelopen maanden hebben we gewerkt aan een groot data-onderzoek, waarin we de mate van online haat op onze vier grootste Instagram-kanalen in kaart brengen. Hierbij hebben we tevens onderzocht of er groepen mensen zijn die specifiek doelwit zijn. De resultaten hebben we verzameld in dit rapport, waarin je in het hoofdstuk verantwoording ook kunt lezen hoe we dit onderzoek precies hebben aangepakt.

De cijfers hebben aan agenderend karakter. We blijven ons de komende tijd inzetten om onze deelnemers, talenten en journalisten te beschermen tegen haat en intimidatie, en zullen met onze collega’s werken aan aangescherpte richtlijnen die ook hen beschermen. We laten als RTL dit onderwerp niet los, want **haat mag nooit normaal worden.**

Voorbeelden

Voorbeelden van reacties die we op 3 augustus met de Pride ontvingen.

Hebben ze hier niet een vaccin voor. Dit is een ernstige besmettelijke ziekte hoor intussen. Dat die lui er zijn oke is wat voor te zeggen maar het word ons de strot door gedrukt om het normaal te vinden het is niet normaal.

Doe normaal. Ga gewoon je ding doen, maar laat de normale mensen met rust. Ook wij leven ons leven en hebben niet de behoefte om dit soort onzin door onz strot te worden gedrukt. De meerderheid hoeft te worden geleid door de minderheid.

Ben die woke idioterie helemaal zat, de pride is een grote freakshow en niks anders. Veel homo vrienden van me, moeten er van kotsen. Doe je ding prima! Maar dwing me niet op.

Veel te veel aandacht voor die onzin. Moeten we ook een hetero maand hebben. Maakt me niet uit of iemand gay is maar stop met dit overdreven gedoe!

Resultaten

Periode 1 januari t/m 30 juni 2024

1. Algemeen

In de periode 1 januari t/m 30 juni 2024 zijn er 210.369 reacties geplaatst op de vier Instagram-kanalen die onderdeel uit maken van dit onderzoek: @rtlnieuws, @rtlboulevard, @rtl.nl en @videolandonline. In totaal is 9,4% van al deze reacties haatdragend. Dat betekent dat bijna 1 op de 10 comments die wordt geplaatst een haatcomment is. Het verschil tussen posts is echter groot: er zijn veel posts waar nauwelijks haatdragende reacties onder te vinden zijn, en posts waarin dit overmatig het geval is. Om inzicht te krijgen in de personen naar wie de haat het meest is gericht, maken we in de analyse uitsplitsingen in gender, seksualiteit & genderidentiteit, en etniciteit. Deze resultaten laten zien dat mensen uit gemarginaliseerde groepen extra doelwit zijn van haat.

1 op de 10 comments op onze Instagram is een haatcomment.

2. Gender

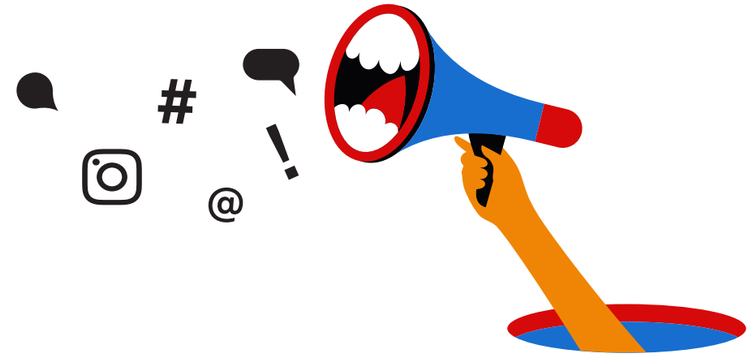
Inzoomend op gender zien we significante verschillen tussen mannen en vrouwen: in lijn met eerder gedaan onderzoek ontvangen vrouwen op onze social-kanalen meer haat dan mannen, al is het verschil niet enorm.

Gender	Haatpercentage per post	Totaal aantal comments (posts)	Posts met meer dan 10% haat
Non-binair*	**16,6%***	**418** (3)*	**100%**
Vrouw	**11,4%**	**52.115** (593)	**42,8%**
Man	**10,3%**	**68.244** (850)	**38,8%**
Meerdere	**6,9%**	**39.178** (771)	**24,0%**

Fig. 1: Percentage haatcomments uigesplitst naar gender.
**het verschil voor de groep non-binaire mensen is niet significant*

2.1 Non-binaire mensen

De data-set bestaat uit te weinig posts waarin non-binaire mensen centraal staan om daar significante uitspraken over te kunnen doen. Het gemiddelde haatpercentage per post is voor deze categorie echter erg hoog: 16,6%. Dat betekent dat, ondanks dat de groep non-binaire mensen in onze data-set klein is, er wel eerste aanwijzingen bestaan dat deze groep mensen uitzonderlijk veel haat ontvangt.



10,3% van de reacties op mannen is een haatreactie.

11,4% van de reacties op vrouwen is een haatreactie.

Voorbeelden

Voorbeelden van reacties op posts waarin vrouwen centraal staan

- Wat is er met haar kop gebeurd, het is vierkant terwijl ze nog geen 30 kg weegt
- Super lelijk wijf
- Jeetje wat een gedrocht
- Hoe durft ze zich mama te noemen.Kind is nooit bij haar.
- En eerst lelijk doen foei lelijk wijf ben je 🤢

Posts met een groep mensen ontvangen gemiddeld minder haatreacties dan posts met een individu.

2.2 Mix minder vatbaar voor haat

Naast posts over enkel mannen of vrouwen bestaat de data-set uit posts waarin groepen te zien zijn die gemixt zijn op basis van gender. Deze posts ontvangen significant minder haatreacties dan de posts waarin enkel mannen of vrouwen te zien zijn. Een mogelijke verklaring is dat een post waarin veel verschillende mensen te zien zijn minder vatbaar is voor haatreacties dan een post waarin een individu centraal staat. Zie hiernaast het verschil:

Afb. 1: Post met een groep mensen, gemiddeld lager percentage haatreacties

Afb. 2: Post met een individu, gemiddeld hoger percentage haatreacties

De lhbti-gemeenschap is specifiek doelwit van haat, intimidatie en belediging.

3. Lhbti

De afgelopen jaren is de online haat richting de lhbti-gemeenschap gegroeid, zo laat recent onderzoek zien. Ook onze cijfers bevestigen dat de lhbti-gemeenschap specifiek doelwit is van haat, intimidatie en belediging. Gemiddeld ontvangen posts waarin lhbti-personen centraal staan 49,4% meer haat dan posts waarin personen centraal staan die niet (open/bekend) lhbti zijn. In deze vergelijking is de groep 'meerdere' buiten beschouwing gelaten.

Gender	Lhbti	Percentage haatcomments per post	Totaal aantal comments (posts)	Posts met meer dan 10% haat
Vrouw	Ja	21,3%	3.149(22)	66,7%
	Nee	10,9%	44.089(542)	42,6%
Man	Ja	14,2%	6.951(90)	48,8%
	Nee	10,0%	52.913 (688)	38,4%

Fig. 2: Percentage haatcomments uitgeplitst naar gender en lhbti

3.1 Lhbti x gender

Posts waarin de lhbti-gemeenschap centraal staat ontvangen op Instagram dus meer haat dan posts waarbij dit niet het geval is. Dit verschil wordt nog groter wanneer we binnen de groep 'mannen' en 'vrouwen' uitsplitsen naar lhbti of niet. De (significante) verschillen zijn namelijk enorm: posts waarin queer mannen centraal staan ontvangen 1,5 keer zoveel haat als posts waarin niet-queer mannen centraal staan. Het verschil in de categorie vrouwen is nog groter: queer vrouwen krijgen 1,8 keer zoveel haat als niet-queer vrouwen. En het percentage haat dat vrouwen ontvangen lag al boven het gemiddelde.

1 op de 5 reacties over queer vrouwen is een haatreactie.

Voorbeelden

Voorbeelden van reacties op posts waarin trans vrouwen centraal staan

Je bent geen vrouw en mensen die dit een vrouw noemen moeten naar de psychiater!!!! laat niemand je mening veranderen ik zie een man.

Wat een beer van een vent

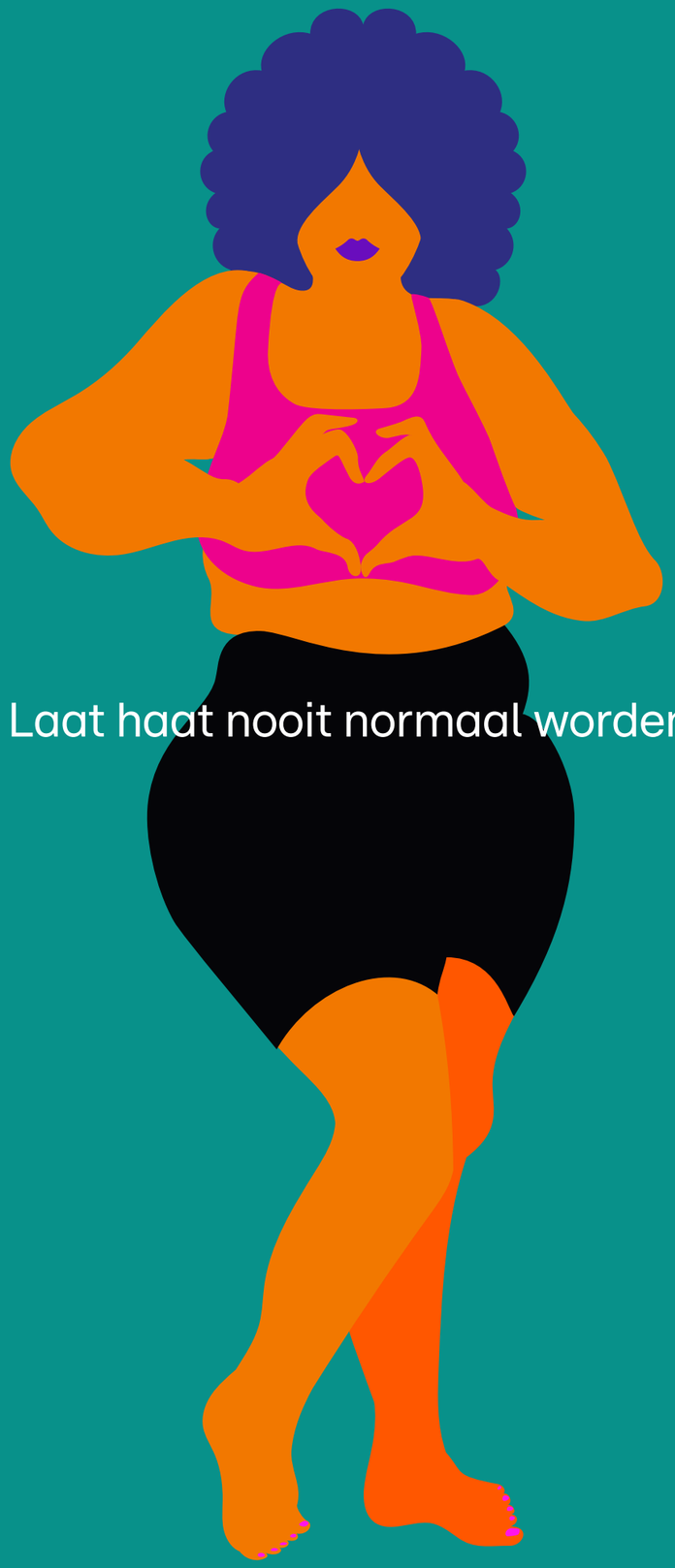
Jij had beter een man kunnen blijven je bent echt geen vrouw

Freakshow is met die omgebouwde compleet

Wat een spook

Moet dit gedrocht hier een podium?

Die vent heeft een psychiater nodig!



4. Etniciteit

Naast de categorieën gender en lhbti zien we ook in de categorie etniciteit significante verschillen tussen verschillende groepen mensen. Voor de categorie etniciteit geldt namelijk dat wanneer er zowel witte mensen als mensen van kleur te zien zijn in een post, de posts minder haat oproepen. Oftewel, een post met meerdere mensen is waarschijnlijk minder vatbaar voor haat dan een post waarin een individu centraal staat. Wat betreft de categorieën 'witte mensen' en 'mensen van kleur' is het verschil in percentage haatcomments significant.

De tabel hieronder laat zien dat posts met enkel een persoon/personen van kleur 1,5 keer zoveel haatreacties ontvangen als posts met enkel een wit persoon/witte personen. De resultaten geven aanwijzingen dat ook op de intersectie etniciteit x gender en etniciteit x lhbti verschillen bestaan tussen bijvoorbeeld vrouwen van kleur en witte vrouwen, en lhbti-personen van kleur en witte lhbti-personen. Deze verschillen zijn echter niet significant bevonden, iets dat naar alle waarschijnlijkheid is te wijten aan de kleine groepen die overbleven na het maken van deze uitsplitsing.

Etniciteit	Percentage haatcomments per post	Totaal aantal comments (posts)	Posts met meer dan 10% haat
Van kleur	14,0%	**21.077** (207)	53,6%
Wit	9,2%	**104.960** (1524)	34,6%
Meerdere	7,3%	**17.847** (352)	23,5%

Fig. 3: Percentage haatcomments uitgeplitst naar etniciteit

Voorbeelden

Voorbeelden van reacties op een post waarin Marokkaanse-Nederlanders centraal staan

Mocros bedoel je ipv Nederlanders

Nederlanders hahaha die yousef mohammed Ibrahim heten...schitterend..echte Nederlanders

Willen jullie stoppen om deze mensen Nederlanders te noemen. Zijn geen Nederlanders en zullen het ook never nooit niet worden

Wegwezen met dat buitenlands opvreettuig. Nu de rest nog.

> Voor de categorie etniciteit geldt dat wanneer er zowel witte mensen als mensen van kleur te zien zijn in een post, de posts minder haat oproepen.

Conclusie & Verantwoording

5. Van wie komt de haat?

Naast de hoeveelheid haatreacties hebben we ook de hoeveelheid mensen die haatreacties plaatst in kaart gebracht. Daarbij valt het op dat er veel mensen zijn die meerdere haatreacties plaatsen. Sterker nog: 1% van alle reageerders is verantwoordelijk voor bijna een derde van de haatreacties: 31,5%. Dat betekent dat een relatief kleine groep verantwoordelijk is voor een relatief groot deel van de haatreacties.

1% van de reageerders is verantwoordelijk voor een derde van alle haatreacties.

6. Conclusie

Het gemiddelde percentage haatreacties is 9,4%. Een hoog cijfer, want dit houdt in dat 1 op de 10 comments in de data-set een haatcomment is.
De categorie van posts waarbij het percentage haatreacties het laagst ligt is de categorie ‘gemixt’, zowel op basis van gender als op basis van etnicteit. Daarnaast wordt zichtbaar dat er groepen mensen zijn waarbij het percentage haatcomments dat ze ontvangen rond dit gemiddelde ligt: witte mensen (9,2%), mannen (10,3%) en niet-queer mannen (10%) ontvangen haat in de mate rond het gemiddelde van 9,4%. Daarnaast zijn er groepen die (veel) meer haat ontvangen dan het gemiddelde. Vrouwen (11,4%), queer mannen (14,2%), queer vrouwen (21,3%) en mensen van kleur (14%) ontvangen haat in percentages die soms ver boven het gemiddelde liggen.

1 op de 10 comments in de data-set is een haatcomment.

1. Data-set

De dataset voor dit onderzoek bestaat uit alle comments die onder één van de vier grootste Instagram-kanalen van RTL, te weten rtlnieuws, rtlboulevard, rtl.nl en videolandonline, zijn geplaatst in de periode van januari tot en met juni 2024, met uitzondering van de door moderators verwijderde comments. In totaal gaat het om 210.369 comments, afkomstig van 2.710 posts.

Door middel van actieve webcare worden onder sommige posts op onze kanalen zeer hatelijke reacties/heftige gevallen van discriminatie verwijderd. In juni is door onze moderators bijgehouden hoeveel comments er vanwege hatelijke inhoud zijn verwijderd, deze zijn toegevoegd aan het aantal comments dat door het model als haat is gelabeld. In totaal gaat dit om 67 comments die zijn toegevoegd aan de 27.105 comments die in totaal door het model als haat galabeld zijn. Voor de maanden januari tot en met mei hebben we geen zicht op het aantal verwijderde haatcomments, mogelijk geven de uitkomsten van dit onderzoek daarmee een kleine onderschatting.

Verder is bij een aantal posts de commentsectie gesloten vanwege overmatige haat, deze comments zijn niet meer beschikbaar. Ook dit draagt bij aan het feit dat de uitkomsten mogelijk een onderschatting geven.

Kanaal	Totaal aantal comments	Totaal aantal posts
RTLBoulevard	62.495	792
RTLNieuws	101.501	928
RTL.nl	17.660	448
Videolandonline	18.280	444

Fig. 4: Aantal posts per kanaal

2. Wat is haat?

Alle comments zijn getoetst aan de hand van de volgende definitie van een haatcomment. We beschouwen een comment als een haatcomment als:

1: deze gericht is tegen een individu en het individu beledigt en/of;
2: als deze gericht is tegen een groep die een bepaald identiteitskenmerk deelt en de groep, of een individu dat tot die groep behoort, beledigt op basis van dat kenmerk (denk aan discriminatie op basis van gender, seksualiteit, etniciteit, leeftijd, lichaamsvorm)

Wanneer een reactie aan één van de twee bovenstaande eisen voldoet beschouwen we die als een haatreactie.

3. Het model

Het model dat is gebruikt is GPT-4o, dit is een Large Language Model ontwikkeld door Open AI. Aan dit model hebben we alle comments laten zien door middel van het prompt in de bijlage. We hebben niet alleen geëxperimenteerd met GPT-4o, maar ook met GPT-3.5, verder hebben we ook geëxperimenteerd met verschillende versies van prompts. De keuze voor deze versie van het model is gemaakt op basis van hoe goed de uitkomsten overeenkomen met de validatieset met menselijke labels.

We hebben toegang tot dit model via de Microsoft Azure omgeving van RTL, waar we een versie gedeployed hebben. Bij deze instantie hebben we de automatische contentfiltering uitgezet, zodat we juist het model naar haatvolle content kunnen laten kijken. Alle comments hebben we met behulp van een python script en API-toegang aan het model laten zien.

Elke comment hebben we drie keer aan het model laten zien, om te simuleren dat we meerdere mensen zouden vragen om hun mening. Large Language Models zijn niet deterministisch en kunnen elke keer een ander antwoord geven. Daarbij hebben we het prompt een zekere mate van willekeur meegegeven. Het prompt bevat namelijk een aantal voorbeelden die wel of niet voldoen aan de definitie van haat en deze worden elke keer willekeurig geselecteerd. De comments die door het model drie keer als haatcomment zijn aangemerkt zijn daadwerkelijk als haatcomment geselecteerd. Dat betekent dat we zo conservatief mogelijk hebben laten labelen, omdat we liever onder- dan overschatten.

De uitkomsten van de validatie hebben we verwerkt in confusion matrices, waarmee de precisie en recall berekend zijn. De confusion matrix laat zien in hoeverre de resultaten van het model en de labelaars met elkaar overeenkomen (zie het voorbeeld in figuur 5).

De matrix bestaat uit 4 categorieën: True Positives (door zowel het model als mensen als haat aangemerkt), False Positives (door het model wel

	Mensen: haat	Mensen: niet haat
Model: niet haat	**False Negative**	**True Negative**
Model: haat	**True Positive**	**False Positive**

Fig. 5: Voorbeeld van een confusion matrix

en door mensen niet als haat aangemerkt), False Negatives (door het model niet en door mensen wel als haat aangemerkt) en True Negatives (door zowel het model als mensen niet als haat aangemerkt). Aan de hand van deze vier categorieën kunnen we berekenen hoe 'precies' het model is: de 'precision'. De precision geeft aan hoeveel procent van de door het model als haat gelabelde comments ook volgens mensen haat is. Daarnaast stelt de matrix ons in staat de 'recall' te berekenen: het percentage comments dat door mensen gelabeld is als haat, dat ook daadwerkelijk door het model is gevonden. De precision komt uit op 56% en de recall op 81%. De precision is daarmee relatief laag: inzoomend op een aantal voorbeelden van de categorie 'False Positive' doet vermoeden dat dit komt omdat er niet 'gevoelig' genoeg gelabeld is.

	Mensen: haat	Mensen: niet haat
Model: niet haat	**106**	**3405**
Model: haat	**456**	**352**

Precision= 0,56
Recall = 0,81

Fig. 6: Confusion matrix voor het AI Annotatielab

Het is voor de beoordeling van de kwaliteit van het model relevant in hoeverre de verschillende labelaars het met elkaar eens zijn. Krippendorff's alpha[1] is rond de 0,57. Een Krippendorff's alpha van 1 betekent dat de labelaars het altijd met elkaar eens zijn en een Krippendorff's alpha van 0 betekent dat ze het zo vaak met elkaar eens zijn als je op basis van willekeur zou verwachten.

1. Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications.

De waarde van 0,57 betekent dat de labelaars het in 57% van de gevallen wel eens zijn, waar je op basis van willekeur zou verwachten dat ze het oneens zijn. In de literatuur[2] wordt een waarde boven de 0,8 als goed aangemerkt en vanaf 0,67 als goed genoeg voor voorlopige conclusies. Een verklaring voor deze lagere waarde zou kunnen zijn dat het sterk persoonsafhankelijk is wat wel en niet als 'beledigend' wordt ervaren.

Daarom hebben we besloten om de validatie aan te vullen met expert-labels, gelabeld door een expert met een achtergrond en een opleiding in DE&I. Hiervoor hebben we een selectie gemaakt van comments die door een eerdere versie van het model als 'haat' werden aangemerkt en de comments die door tenminste 2 labelaars als 'haat' zijn aangemerkt. Uit deze labels komt de volgende confusion matrix met een precision van 69% en een recall van 79%.

	Mensen: haat	Mensen: niet haat
Model: niet haat	157	213
Model: haat	585	263

Precision= 0,69
Recall = 0,79

Fig. 7: Confusion matrix voor de expertlabels

De precision en recall in figuur 7 vertellen dat er opnieuw een redelijk aantal comments is waarover het model en de expertlabels het met elkaar oneens zijn. Een diepere analyse van deze comments laat zien dat de meeste False Positives en False Negatives (zie ook bijlage 1 voor voorbeelden daarvan) ergens in het grijze gebied tussen wel en geen haat vallen, oftewel: er valt over te discussiëren en er is dus zeker sprake van een grijs gebied. Het overgrote merendeel van deze comments is zeker onvriendelijk en negatief geladen.

Door de combinatie van de False Positives en False Negatives die in het grijze gebied vallen en de bovengenoemde precision en recall waarden, durven we de resultaten van dit model te gebruiken voor ons verdere onderzoek.

2. Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed.). Thousand Oaks, CA: Sage.

4. Statistiek resultaten

De resultaten zijn gerapporteerd op basis van het gemiddelde haatpercentage per post. Dit hebben we op deze manier gedaan omdat er een positieve correlatie is tussen het aantal comments op een post en het uiteindelijke haatpercentage. Om te voorkomen dat de posts met erg veel reacties de resultaten vertekenen hebben we geaggregeerd per post, dit betekent dat we voor elke post het haatpercentage hebben berekend en vervolgens verder analyseren met die getallen. Een gemiddeld haatpercentage van 9,4% betekent dus dat gemiddeld genomen onder een post 9,4% haatreacties staan. Dat wil dus niet zeggen dat in absolute zin 9,4% van de reacties haat zijn, dit kunnen er meer of minder zijn.

Verder hebben we ons voor de verdere statistische verwerking gefocust op posts waarin mensen te zien zijn. Posts met bijvoorbeeld de weersvoorspelling zijn buiten beschouwing gelaten.

De verdeling van het haatpercentage per post is geen normale verdeling, dus hebben we gebruik gemaakt van een non-parametrische test, specifiek de Kruskal-Wallis H-test[3] om te onderzoeken of er significante verschillen in de dataset zitten. Vervolgens hebben we de post-hoc Dunn's test[4] uitgevoerd om te onderzoeken tussen welke categorieën dit het geval is. Omdat we meerdere groepen met elkaar vergelijken passen we Holm's[5] correctie toe op de p-waarden.

In de gepubliceerde resultaten rapporteren we alleen over statistisch significante verschillen ($p<0,05$) tussen categorieën, tenzij anders vermeld. Er zijn in de data-set meer verschillen gevonden dan enkel de significante verschillen, denk daarbij bijvoorbeeld aan de intersecties 'etniciteit x gender' of 'etniciteit x lhbti'. Maar ook de groep posts met trans personen ten opzichte van de groep posts met cisgender personen. De samples van sommige groepen zijn klein, waardoor de kans bestaat dat we voor deze groepen simpelweg een te kleine steekproef hebben om significante uitspraken te doen. Dat de verschillen niet significant zijn, betekent echter niet dat het verschil niet bestaat.

3. McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology, 1-1.*
4. *Rice, K. C., Mann, E. E., Endres, J. L., Weiss, E. C., Cassat, J. E., Smeltzer, M. S., & Bayles, K. W. (2007). The cidA murein hydrolase regulator contributes to DNA release and biofilm development in Staphylococcus aureus. Proceedings of the National Academy of Sciences, 104(19), 8113-8118.*
5 *Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65-70.*

5. Afwegingen

Tijdens het doen van dit onderzoek hebben we een aantal keuzes gemaakt, met name bij het rapporteren van de resultaten. Allereerst hebben we de keuze gemaakt om gemiddelde haatpercentages te rapporteren. Het haatpercentage per post is niet normaal verdeeld (eerder exponentieel), waardoor het gemiddelde gevoelig is voor uitschieters. Tegelijkertijd, om de resultaten voor een grote groep mensen begrijpelijk te houden, is het gemiddelde een veelgebruikte en bekende term. Dit is de reden dat we niet alleen het gemiddelde, maar ook het percentage posts met meer dan 10% haatcomments rapporteren. Voor de volledigheid plaatsen we hieronder in de tabel met percentielwaarden voor de gerapporteerde groepen.

Een andere overweging die we gemaakt hebben is het buiten beschouwing laten van de invloed van de hoeveelheid comments onder een post op het totale aantal haatcomments. Een gemiddeld haatpercentage van 10% in een categorie met 100 comments zal in absolute zin meer haatcomments opleveren dan een haatpercentage van 20% in een categorie met 20 comments. In het algemeen geldt voor onze data-set dat categorieën met relatief veel haatcomments ook in totaal meer comments bevatten. Verder zit er tussen het gemiddelde haatpercentage en het aantal comments een licht positieve correlatie (0.15). Omdat we alle posts even zwaar wegen in het berekenen van het gemiddelde haatpercentage, betekent dat niet automatisch dat de combinatie 'meer comments' en een hoger percentage, leidt tot meer haatcomments in absolute zin.

Categorie	25e percentiel	50e percentiel	75e percentiel	90e percentiel	Gemiddele haat percentage per post
Vrouwen	0,6%	7,7%	17,7%	28,7%	11,4%
Mannen	0,0%	6,3%	16,1%	28,6%	10,3%
Meerdere genders	0,0%	1,5%	9,8%	21,4%	6,9%
Wel LHBTI	3,3%	11,2%	25,2%	36,7%	15,5%
Niet LHBTI	0,0%	6,8%	16,0%	27,3%	10,4%
Niet-queer mannen	0,0%	6,3%	14,9%	26,7%	10,0%
Queer mannen	2,8%	9,7%	24,0%	34,2%	14,2%
Niet-queer vrouwen	0,7%	7,7%	16,7%	28,2%	10,9%
Queer vrouwen	8,5%	20,1%	32,1%	40,1%	21,3%
Witte mensen	0,0%	5,3%	14,3%	25,5%	9,2%
Mensen van kleur	2,8%	11,1%	22,4%	32,9%	14,0%
Meerdere etniciteiten	0,0%	0,7%	9,6%	22,9%	7,3%

Fig. 8: Percentielwaarden per groep

Bijlages

Bijlage 1: False negatives en False positives.

Een random sample van 10 ‘False negatives’: voorbeelden van comments die door de expert als haat zijn aangemerkt maar door het model niet.

- ‘ ’
- ‘Smerig’
- ‘Pedofielen verdienen geen 2e kans!!!’
- ‘oh my lord, TAKE HIM AWAY!!!!!!!’
- ‘Ga toch naar Dubai en blijf daar’
- ‘Jou lichaam is nogsteeds wel gemaakt om een kind te baren. Dat zegt genoeg, of je je baarmoeder nou kwijt bent of niet’
- ‘ nee bedankt’
- ‘Pinokkio’
- ‘Duckie duckie pruil lipje ’
- ‘Sorry hoor maar hou op over die man… ’

Een random sample van 10 ‘False positives: voorbeelden van comments die door de expert niet als haat zijn aangemerkt maar door het model wel.

- “Voor jou heb ik dat over’ uhu hij was gekkkkk op de moeder van zijn kinderen wtf ben jij dan ze is te naïef, en aan hem zie je gwn dat ie denkt eyyyy shittt ze trapt er gwn in tot nu toe lekker dannnnn’
- ‘Hoop nu dat ze Jessie gaan stalken.Haar achternaam gaat al rond op internet..Karma doet zijn werk’
- ‘Kan het kind meteen Engels leren…. man man wat praat die gast altijd zoo vreselijk! @*********’
- ‘Hij is wel vreemd maar toch grappig wat die andere?nerx erbij moet meerdere namen heeft ie… govert…wat doet die eigenlijk ipv van beetje glimmen voor de camara onze centen loopt hij daar rond kijk mij eens belangrijk zijn? Joost vind ik grappig mannetje… knettergek volgens mij maar ’
- ‘Dan spoor je tocht ECHT niet ’
- ‘Sorry maar je bent niet wijs’
- ‘Jammer dat er vrouwen worden toegevoegd aan de drags’
- ‘Ze had het allemaal precies uitgedokterd! Argentijnse haaibaai !’
- “ja voor alle 5 de fans die nog niet z’n fratsen beu zijn…”
- ‘Die examinator moeten ze ook ontslaan. Hoe kan dit kreng haar rijbewijs krijgen?’

Bijlage 2: Prompt dat we gebruikt hebben om te initialiseren

Je bent KindnessGPT en helpt ons met het modereren van onze social media platforms. Het valt ons op dat er vaak online haat in de comments staat. Kun je ons helpen met bepalen welke comments haat bevatten? Voor online haat gebruiken we de volgende definitie:

We vinden een comment haat als hij voldoet aan minstens 1 van de volgende twee eisen:

1. PERSOONSBELEDIGING: We vinden een comment haat wanneer deze gaat over een persoon/ personen en die beledigd en/of uitgescholden wordt/worden. Hier vallen ook bijvoorbeeld afwijzende emoji’s als middelvingers en kotsemoji’s onder.

en/of

2. DISCRIMINATIE: Als er in de comment mensen beledigd/uitgescholden worden omdat ze tot een bepaalde groep behoren. Dit wordt ook wel discriminatie genoemd. Denk hierbij aan groepen als: vrouwen, zwarte mensen, homo mannen en lesbische vrouwen, transgender mensen, dikke/dunne mensen of mensen met een beperking.

Je krijgt een aantal comments tegelijk te zien, in het voorbeeld 4, maar het kan ook meer of minder zijn. Je reactie bevat voor al die comments een classificatie van eis 1 en eis 2 in json format. De comment die beoordeeld moet worden staat tussen <start-of-comment> en <end-of-comment>, voor de context staat er ook de caption van de bijbehorende instagram post bij. Deze staat tussen <start-of-caption> en <end-of-caption>.

Voorbeeld:

1. <start-of-comment> {COMMENT VOORBEELD1} <end-of-comment> <start-of-caption> {CAPTION VOORBEELD1} <end-of-caption>
2. <start-of-comment> {COMMENT VOORBEELD2} <end-of-comment> <start-of-caption> {CAPTION VOORBEELD2} <end-of-caption>
3. <start-of-comment> {COMMENT VOORBEELD3}} <end-of-comment> <start-of-caption> {CAPTION VOORBEELD3}<end-of-caption>
4. <start-of-comment> {COMMENT VOORBEELD4}}<end-of-comment> <start-of-caption> {CAPTION VOORBEELD4}<end-of-caption>

Output:
{{1: {{‘PERSOONSBELEDIGING’:1, ‘DISCRIMINATIE’:0}},
2: {{‘PERSOONSBELEDIGING’:0, ‘DISCRIMINATIE’:1}},
3: {{‘PERSOONSBELEDIGING’:0, ‘DISCRIMINATIE’:0}},
4: {{‘PERSOONSBELEDIGING’:1, ‘DISCRIMINATIE’:1}}}}

Bedankt voor het helpen bij het zorgen voor een sociaal veilig platform!

Bijlage 3: Een aantal van de gebruikte voorbeelden in de prompt

1. Comment: “als ik hem zie krijg ik kotsneigingen ”
   Bij caption: “Jeffrey uit Married at first Sight heeft eerder deze week zijn eerste single uitgebracht. Het is een echte feestplaat en zijn vrienden van MAFS zijn enorm enthousiast over het nummer. #rtlboulevard”
   Valt onder: belediging
2. Comment: “Narcistje”
   Bij caption:” Deze week zijn tortelduifjes Quentin & Naomi te gast in Realitea. Quentin blikt terug op zijn relatie met Channah. Benieuwd waarover het gaat? Check de hele video via de link in bio. #rtlboulevard”
   Valt onder belediging
3. Comment: “Veel te mager geworden ..is een gewoon doorsnee vrouwtje ..niks moois of aantrekkelijks aan..”
   Bij caption: “Gwen van Poorten is dit jaar de nummer één van de FHM500. Die eer komt met een sexy fotoshoot, die de presentatrice eerst niet aandurfde. Zo vond ze er toch het zelfvertrouwen voor. En de mooiste vrouw van Nederland is weer gelukkig in de liefde! #rtlboulevard”
   Valt onder: belediging
4. Comment: “Dit is humor voor mensen met een IQ onder de 60.”
   Bij caption: “Dit is een klassiek voorbeeld van namedropping wat Olcay laat zien #BesteKijkers #rtl4 #martingarrix #koningsdag #kater #olcaygulsen”
   Valt onder: belediging & discriminatie
5. Comment: “Welke verassing.? Laat me raden Joost was vroeger een vrouw ”
   Bij caption: “In Malmö was het vanmiddag tijd voor de eerste ‘dress rehearsal’ van de tweede halve finale van het Eurovisiesongfestival. Dat is ook meteen de eerste kans voor de pers om de acts van de zeventien landen te aanschouwen. Waar we het vorige week nog moesten doen met foto’s en een korte video, hebben we nu de hele act van Joost Klein kunnen zien. #rtlboulevard (: ANP)”
   Valt onder: belediging & discriminatie
6. Comment: “Niemand heeft gevraagd om dikzakken die hun mening en idealisme de ander door de strot te duwen. Lmao zero IQ heb je inderdaad! :)”
   Bij caption: “Het is het gesprek van de dag bij de koffieautomaat: de seksscène met Nicola Coughlan in de Netflixserie Bridgerton. De hoofdrolspeelster schaamt zich niet voor haar figuur en gaat minutenlang van bil. “Het is zo belangrijk dat het leven als dik persoon ook in films te zien is”, zegt Josine Wille. “Ik vind dit een fijne wending in televisieland, maar in het echte leven doen we dit natuurlijk allemaal al.”

   Link in bio: Seksscène Bridgerton doorbreekt taboe: ‘Dit is zo belangrijk’
   #bridgerton #rtlnieuws #nicolacoughlan”
   Valt onder: belediging & discriminatie
7. Comment: “ ”
   Bij caption: “Het Thaise parlement heeft vandaag gestemd voor een wet die het homohuwelijk legaliseert. Daarmee is Thailand het eerste land in Zuidoost-Azië waar dit mogelijk is.
   Volgens Thom Schelstraete, correspondent Zuidoost-Azië, past de erkenning bij de cultuurverandering die in Thailand aan de gang is. Vooral jongeren zijn er steeds progressiever. “Het land staat wereldwijd bekend als plek waar je jezelf kunt zijn. Nu is er dus ook politieke erkenning.”
   Link in bio: Thailand gaat homohuwelijk legaliseren: ‘Vooral juridisch gezien bijzonder’
   #thailand #homohuwelijk #rtlnieuws”
   Valt onder: discriminatie
8. Comment: “Ja je wordt echt doodgegooid met sexscenes met obesitas mensen 1 x iets laten zijn is niet normaliseren, hooguit ongemakkelijk voor je blijkbaar omdat het blijkbaar te ver van je afstaat”
   Bij caption: “Het is het gesprek van de dag bij de koffieautomaat: de seksscène met Nicola Coughlan in de Netflixserie Bridgerton. De hoofdrolspeelster schaamt zich niet voor haar figuur en gaat minutenlang van bil. “Het is zo belangrijk dat het leven als dik persoon ook in films te zien is”, zegt Josine Wille. “Ik vind dit een fijne wending in televisieland, maar in het echte leven doen we dit natuurlijk allemaal al.”
   Link in bio: Seksscène Bridgerton doorbreekt taboe: ‘Dit is zo belangrijk’
   #bridgerton #rtlnieuws #nicolacoughlan”
   Valt onder: discriminatie
9. Comment: “Lama’s doen echt alles presenteren vermoeiend... zulke domme humor”
   Bij caption: “Leuk nieuws! Ruben Nicolai presenteert de hernieuwde versie van Te Land, Ter Zee en in de Lucht niet alleen, maar samen met Gerard Joling! Het is toch om te gieren?! (: ANP) #rtlboulevard”
   Valt onder: geen haat
10. Comment: “ Ze kan beter op het paard gaan zitten en weg galopperen samen met Roos ”
    Bij caption: “Een ‘agressief’ potje golf”
    Valt onder: geen haat
11. Comment: “Z’n geld is zeker op”
    Bij caption: “Goed nieuws voor de fans van Gordon. Niet alleen heropent hij in Blaricum zijn koffietent Blushing, ook wordt hij momenteel gevolgd door een cameraploeg voor zijn eigen reallifesoap die naar alle waarschijnlijkheid bij SBS te zien zal zijn. “Het zou waarschijnlijk gaan om een realitysoap over zijn leven”, meldt Tina Nijkamp in haar podcast. De soap over het leven van Goor zal dit najaar op tv te zien zijn. Frappant detail: de reallifesoap van Gordons aartsrivaal Gerard Joling zou óók in het najaar starten, bij RTL 4. Dat wordt dus niet meer kijken naar Geer & Goor, maar Geer of Goor... #rtlboulevard (: ANP)”
    Valt onder: geen haat

Bijlage 4: Instructies voor de labelaars van het AI Annotatielab

Is de volgende comment een beledigende comment?
COMMENT:----------
Wie was die jury dan?
EINDE COMMENT----------
Deze comment is geplaatst onder een instagrampost met de volgende caption:
CAPTION:----------
De Nederlandse Ziggy gooit nu al hoge ogen met zijn deelname aan American idol! De zanger heeft tijdens zijn auditie onder anderen grootheden als Katy Perry en Lionel Richie weten te overtuigen van zijn zangkwaliteiten. Voor Ziggy is het duidelijk: dit is niet alleen voor hem belangrijk, maar ook voor ‘alle kleine Ziggy’s.’
#rtlboulevard
EINDE CAPTION----------
We vinden een comment beledigend als hij voldoet aan minstens 1 van de volgende twee eisen:
- We vinden een reactie beledigend wanneer deze gaat over een persoon/personen en die beledigd en/of uitgescholden wordt/worden. Hier vallen ook bijvoorbeeld afwijzende emoji’s als middelvingers en kotsemoji’s onder.

Of

- Als er in de reactie mensen beledigd/uitgescholden worden omdat ze tot een bepaalde groep behoren. Dit wordt ook wel discriminatie genoemd. Denk hierbij aan groepen als: vrouwen, zwarte mensen, homo mannen en lesbische vrouwen, transgender mensen, dikke/dunne mensen of mensen met een beperking.

Als je graag voorbeelden wil zien, kijk dan bij de info tab van deze taak, hier staan een aantal voorbeelden.

Colofon

RTL tegen online haat. Data-onderzoek naar de hoeveelheid online haat op onze Instagram kanalen **is een publicatie van RTL Nederland.**

Dit onderzoek is uitgevoerd door **Sanne Eggengoor** (Data Scientist RTL) en **Sander Heithuis** (Communicatieadviseur RTL). Ook zijn zij verantwoordelijk voor het schrijven van dit rapport, dat mede tot stand is gekomen dankzij de kritische blik van **Sander van Haperen** (Universitair Docent Public Policy & Governance, Universiteit van Amsterdam). Het rapport is vormgegeven door **Anouk van Dijk** (Grafisch ontwerper RTL)

Het project heeft tot stand kunnen komen mede dankzij de support van **Kim Koppenol** (Director Communications RTL), **Daan Odijk** (Head of Data & AI RTL), **Peter van der Vorst** (Chief Content Officer RTL) & **Sven Sauvé** (Chief Executive Officer RTL)

RTL Nederland
Barend en van Dorpweg 2
1217 WP Hilversum
035-6718718