Universiteit van Amsterdam

Onderzoeksvoorstel

Complotvideo's op YouTube

19 maart 2021

Student:
Roan Schellingerhout (12399779)

Docent:
Dr. Maarten Marx

1 Doelstelling en relevantie

Het doel van dit onderzoek is om te kijken hoe snel (oftewel: na hoeveel gekeken video's) het YouTube-aanbevelingsalgoritme een voorkeur krijgt voor complotvideo's. YouTube is met 34.6 miljard maandelijkse gebruikers de één na meestbezochte website op het internet, waardoor de impact van de website op de maatschappij niet onderschat kan worden (Neufeld, 2021). Er wordt content van alle categorieën geproduceerd en geconsumeerd. Echter, complotcontent begint een hoofdrol te spelen op YouTube. Alt-right (ook wel 'far-right') en complotkanalen krijgen een steeds grotere aanhang, wat negatieve gevolgen kan hebben voor de samenleving. Zo zijn er, mede dankzij complotcontent op YouTube, steeds meer mensen die beginnen te twijfelen aan de wetenschap. Wanneer zulke twijfels zich voordoen bij belangrijke onderwerpen, zoals het wel of niet innemen van het vaccin tegen het coronavirus, kan dit gevaarlijke gevolgen hebben voor de maatschappij. Zo is meer dan de helft van de Amerikaanse populatie twijfelachtig over - of zelf definitief tegen - het nemen van het coronavaccin (Rosenbaum, 2021). Dit onderzoek is gebaseerd op een aflevering van VPRO's *Zondag met Lubach*, waarin eenzelfde idee op kleinere schaal werd uitgevoerd (Lubach, 2020). De uiterst interessante resultaten van dat experiment hebben mij gemotiveerd om het in grotere mate te onderzoeken.

Er is al onderzoek gedaan naar filterbubbels en complot-content op YouTube, maar deze onderzoeken houden zich niet bezig met hoe snel een dergelijke bubbel ontstaat. Daarnaast kijken deze onderzoeken specifiek naar aanbevelingen op content, waarbij de aanbevelingen gebaseerd worden op de gelijkenis van de video's en niet op het kijkgedrag van gebruikers.

2 Vraagstelling

Het onderzoek zal zich richten op de volgende hoofdvraag: *Wat is de invloed van verschillende kijkstrategieën op het aantal complotvideos dat gekeken moet worden voordat de YouTube-aanbevelingen van een gebruiker de voorkeur krijgen voor conspiracy content?*

Hierbij zal de situatie als 'voorkeur' worden geteld zodra het percentage complotvideo's binnen de gebruikers aanbevelingen significant hoger is dan bij de baseline. Om deze hoofdvraag te beantwoorden, zullen er eerst drie deelvragen beantwoord moeten worden. Deze luiden als volgt:

- Bij welke kijkstrategie komt een gebruiker het snelst in een filterbubbel van complotvideo's terecht op YouTube?

- Hoe lang duurt het voor een YouTube-gebruiker om uit een filterbubbel te komen, wanneer deze zich erin bevindt?
- Welk type classifier werkt het beste om complotvideo's op YouTube te labelen?

3 Literatuur

Wanneer een gebruiker van een website zich bevindt in een eigen informatie-'universum', waarin de content en aanbevelingen inspelen op diens bestaande meningen en overtuigingen, zit deze in een filterbubbel (Pariser, 2011). Gebruikers zijn alleen in zulke bubbels; iedereen heeft een eigen, unieke bubbel. Er kan overlap zijn tussen de bubbels van verschillende personen, maar elke bubbel is precies aangepast op het individu. In traditionele media kan een persoon ervoor *kiezen* welk type meningen deze wil aanhoren, door bijvoorbeeld naar een specifieke nieuwszender te kijken. Online is deze keuze niet expliciet, op basis van het gedrag van de gebruiker wordt deze automatisch, zonder toestemming, een bepaald filter voorgelegd. Uit voorgaand onderzoek is gebleken dat het YouTube-algoritme dat verantwoordelijk is voor het geven van aanbevelingen gevoelig is voor het ontstaan van dergelijke filterbubbels. Roth et al. (2020) zijn tot deze conclusie gekomen nadat zij YouTube-aanbevelingen op basis van content hebben geanalyseerd. Op YouTube zijn er twee soorten aanbevelingen: aanbevelingen gebaseerd op het kijkgedrag van de gebruiker en aanbevelingen gebaseerd op de content van de huidige video. In hun onderzoek hebben Roth et al. de focus gelegd op aanbevelingen op basis van content. Hieruit bleek dat dergelijke aanbevelingen snel konden leiden tot lage diversiteit (oftewel: filterbubbels) en dat deze bubbels sneller optraden bij video's met meer weergaven; des te meer weergaven een video had, des te minder divers de gerelateerde aanbevelingen. Zij speculeren dat dit verklaard kan worden door het feit dat YouTube meer informatie opslaat over video's met veel weergaven, waardoor het algoritme betere aanbevelingen kan geven. Ook voorspellen zij dat, naarmate het algoritme meer informatie verkrijgt over de gebruiker, het deze informatie kan combineren met de informatie over video's, wat zou leiden tot een nog sterkere beperking van de aanbevelingen. Volgens Ledwich and Zaitsev (2019) is het kijkgedrag van de gebruiker verantwoordelijk voor ongeveer 70% van de aanbevelingen; kijkgedrag zou daarom veel invloed kunnen hebben op het ontstaan van filterbubbels op YouTube.

YouTube heeft beperkte regels tegen het verspreiden van complottheorieën (YouTube, 2021). Zolang de content niet direct aanzet tot geweld of de volksgezondheid in gevaar brengt (denk hierbij aan misinformatie over het COVID-19-virus), mogen ook objectief onjuiste ideeën worden verspreid via YouTube. Dit zorgt ervoor dat er op YouTube meerdere complotgemeenschappen huisvesten. Complottheorieën als 'de aarde is plat en de overheid probeert dat te verstoppen', 'de wereld zal binnenkort vergaan en alleen aanhangers van een bepaalde religie zullen overleven', en 'de wereld wordt geregeerd door kannibalistische, satanische pedofielen' (beter bekend als QAnon), worden op de website door miljoenen mensen bekeken (Paolillo, 2018; Miller, 2021). Hoewel dergelijke video's schadelijk kunnen zijn voor de maatschappij, worden deze door YouTube niet onderdrukt. Als een gebruiker interesse toont in zulke video's, zal de gebruiker soortgelijke video's aanbevolen krijgen, ook als YouTube ervan op de hoogte is dat het mogelijk schadelijke content is (Ledwich and Zaitsev, 2019).

Het YouTube-algoritme probeert aanbevelingen te doen op basis van de verwachtte kijktijd (watch time) en niet op basis van de kans dat een gebruiker op een video klikt (Covington et al., 2016). Dit is gedaan om misleidende video's (beter bekend als 'clickbait') een lagere kans te geven om aanbevolen te worden. Maar, het verkrijgen van feedback over een video op basis van kijktijd heeft te kampen met veel ruis, waardoor het lastig is om tevredenheid te meten. Zo blijkt dat, ook wanneer een persoon een video interessant vindt, deze de video vaak niet helemaal afkijkt. Gemiddeld kijken personen ongeveer 50-60% van een video voordat zij hiervan wegklikken (Park et al., 2016). Echter, video's die goed ge-

structureerd, of erg interessant zijn, kunnen dit percentage omhoog halen tot 70-80%, waarin ongeveer de helft van de kijkers de video zelfs volledig afkijkt (Lang, 2018). Nadat een video is afgekeken, is er een 41.6% kans dat de gebruiker een aanbevolen video zal bekijken. Welke video dit wordt, volgt een Zipf's verdeling ($\alpha = 0.78$) op basis van de positie in de lijst van aanbevelingen (Zhou et al., 2010).

In voorgaand onderzoek is er dus vernomen dat filterbubbels zich voordoen op YouTube en dat het algoritme geneigd is om complot-content aan te bevelen. Ook wordt er gespeculeerd dat het algoritme beslissingen maakt op basis van het kijkgedrag van gebruikers en dit koppelt aan de content van video's. Om de gebruiker zo lang mogelijk op de website te houden, wat winstgevend is voor YouTube, probeert het algoritme video's aan te bevelen die de gebruiker waarschijnlijk lang zal kijken, welke soms schadelijke content bevatten. Op basis van deze informatie kan er dus verder onderzoek worden gedaan naar het ontstaan van filterbubbels en het verspreiden van complot-content op YouTube. Zo is er nog weinig bekend over hoe snel de aanbevelingen van een gebruiker zich aanpassen, terwijl dit erg belangrijk kan zijn in het ontstaan van zogeheten 'rabbit holes'. Ook is er nog geen onderzoek gedaan naar hoe het soort video invloed heeft op het algoritme. Wanneer een gebruiker veel aanbevelingen kijkt, zou dit mogelijk gezien kunnen worden als impliciete positieve feedback, waardoor er een sneeuwbaleffect kan ontstaan.

4 Onderzoeksplan

Er zal in python een script worden geschreven dat YouTube-accounts video's laten kijken en na elke video bijhoudt voor welk deel hun 'aanbevolen'-pagina uit complotvideo's bestaat. Het laten kijken van YouTube-video's door dit script, zal het mogelijk maken om de eerste deelvraag te beatwoorden (*Bij welke kijkstrategie ontstaat er het snelst een complot-filterbubbel op YouTube?*). In totaal zullen er twintig YouTube-accounts worden aangemaakt, welke zullen worden onderverdeeld in vier kijkstrategieën:

- Compleet willekeurige video's (controlegroep);
- Willekeurige complotvideo's uit een dataset;
- Eerst een willekeurige complotvideo en vervolgens telkens een aanbevolen video naast die video;
- Eerst een willekeurige complotvideo en vervolgens telkens een aanbevolen video op de YouTube-homepage van het account.

Elke kijkstrategie zal door vijf verschillende accounts worden uitgeprobeerd, en elk account zal 25 video's kijken.

Om het gedrag van echte gebruikers zo realistisch mogelijk te simuleren, zal de gemiddelde kijkduur van video's normaal verdeeld worden, met een gemiddelde van 55% en een standaardafwijking van 25% (Park et al., 2016). Ook het klikgedrag van gebruikers zal zo accuraat mogelijk worden nagebootst. De kans dat een video op een bepaalde positie k wordt aangeklikt (diens click-through rate: CTR), zal worden berekend met de volgende formule:

$$CTR(k; N, \alpha) = \frac{1/k^{\alpha}}{\sum_{n=1}^{N}(1/n^{\alpha})} \tag{1}$$

Waarin N het totaal aantal video's in de lijst is en α de exponent-waarde die de distributie bepaalt. Op basis van de waardes gevonden door Zhou et al. (2010), zal de video op de eerste positie een klikkans van ongeveer 20.6% krijgen, waarna de klikkans snel af zal lopen, tot 1.9% bij de twintigste aanbeveling.

Vervolgens zullen alle accounts compleet willekeurige video's gaan kijken. Dit wordt gedaan om deelvraag 2 (*Hoe lang duurt het voor een YouTube-gebruiker om uit een*

filterbubbel te komen, wanneer deze zich erin bevindt?) te beantwoorden. Er zal gekeken worden hoe lang het duurt voor de gebruikers in filterbubbels om een statistisch gelijke aanbevelingsdiversiteit te krijgen aan de gebruikers in de controlegroep. Zodra het aantal complotvideo's in de aanbevelingen niet meer significant verschilt met die van de controlegroep, zit de gebruiker niet meer in een filterbubbel.

Voor beide voorgaande onderzoeksvragen is het van belang om te kunnen bepalen of video's tellen als complotvideo's of niet. Om dit te doen, zijn er twee mogelijkheden. Als eerst is er een dataset met daarin 3000 YouTube-kanalen die door mensen zijn gelabeld als complotkanalen (Ledwich and Zaitsev, 2019), dus een video die geüpload is door een dergelijk kanaal zal tellen als complotvideo. Daarnaast, voor video's die buiten deze kanalen vallen, zal een ensemble van machine learning classifiers gebruikt worden, om zo te bepalen welke video's wel en niet als complotvideo's gezien kunnen worden. Dit leidt tot de derde onderzoeksvraag (*Welk type classifier werkt het beste om complotvideo's op YouTube te labelen?*). Om deze vraag te beantwoorden, en daarmee ook de prestatie van de classifier te optimaliseren, zal er gekeken worden welk type machine learning classifier het beste is in het identificeren van complotvideo's. De volgende algoritmes zullen worden getest: k-nearest neighbors, support-vector machine, neural network, logistic regression, en ridge regression. Al deze algoritmes zullen worden getraind op een dataset bestaande uit 40.000 YouTube-video's, welke zijn gelabeld als wel/geen complotvideo's. Voor elke video zal de titel, beschrijving en ondertiteling, plus de beschrijving en keywords van het bijbehorende YouTube-kanaal beschikbaar zijn. De dataset zal schoongemaakt worden op verschillende manieren: de twee klassen zullen worden gebalanceerd, niet-Engelse teksten zullen worden vertaald, woorden zullen gestemd worden en stopwoorden zullen verwijderd worden. Vervolgens zal de tekst gevectoriseerd worden op basis van TF-IDF waardes. Door middel van een train/test/validation-split van de data zal het mogelijk zijn om voor elk algoritme de hyperparameters te optimaliseren, om zodanig de best mogelijke prestatie te verkrijgen. Deze prestatie zal geëvalueerd worden op basis van vier verschillende waardes: de accuracy, die het aandeel correcte voorspelling uitdrukt; de recall, die aangeeft welk deel van de daadwerkelijk-positieve waardes ook als positief is voorspeld; de precision, die aangeeft welk deel van de positief-voorspelde waardes daadwerkelijk positief is; en de F1-score, die het harmonisch gemiddelde van de recall en precision uitdrukt (Sokolova and Lapalme, 2009). Door een zo hoog mogelijke, maar gebalanceerde, waarde te vinden voor al deze waardes, zal het mogelijk zijn om de optimale hyperparameters te kiezen en vervolgens te bepalen welke classifier de beste prestatie levert.

Als laatste zal er gekeken worden naar de toegevoegde waarde van het gebruik van een machine learning ensemble. Aangezien er twee mogelijke labels zijn voor de data, zal dit ensemble bestaan uit twee verschillende classifiers (Bonab and Can, 2016, 2017). Voor dit ensemble zullen wederom de hyperparameters worden getuned (waaronder welke twee classifiers gebruikt zullen worden), om op die manier te onderzoeken of een ensemble beter presteert dan losse classifiers op zichzelf. Het resultaat hiervan zal antwoord bieden op de eerste deelvraag.

5 Begeleider en EC

Dr. M. J. Marx
18 EC

6 Planning en afspraken

Er zal elke week in principe een vaste, wekelijkse Zoom-afspraak zijn. Verder zal de begeleiding zich voordoen aan de hand van issues op GitHub wanneer er vragen en/of onduidelijkheden zijn. Wanneer deze punten te groot zijn om via GitHub te overleggen, kan ervoor worden gekozen om ze telefonisch af te handelen.

Tabel 1: Planning

Week	Handelingen
28/03-03/04	Hyperparamters optimaliseren voor de classifiers
03/04-10/04	Classifier-ensemble optimaliseren
11/04-17/04	Afronden classifiers en inleiding uitbreiden
18/04-24/04	Deelvraag 1 schrijven
25/04-01/05	Theoretisch kader afmaken, bronnen zoeken voor deelvraag 2
02/05-08/05	Deelvraag 2 schrijven
09/05-15/05	SIM-kaarten kopen en Google-accounts aanmaken
16/05-22/05	Uitvoering experiment, beginnnen met schrijven van deelvraag 3
23/05-29/05	Deelvraag 3 afschrijven
30/05-05/06	Discussie en abstract schrijven
06/06-12/06	Laatste aanpassingen en verbeteringen
13/06-19/06	Inleveren scriptie en verdediging

Referenties

Bonab, H. and Can, F. (2017). Less is more: a comprehensive framework for the number of components of ensemble classifiers. *arXiv preprint arXiv:1709.02925*.

Bonab, H. R. and Can, F. (2016). A theoretical framework on the ideal number of classifiers for online ensembles in data streams. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2053–2056.

Covington, P., Adams, J., and Sargin, E. (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Ledwich, M. and Zaitsev, A. (2019). Algorithmic extremism: Examining youtube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.

Lubach, A. (2020). De online fabeltjesfuik — zondag met lubach (s12).

Miller, D. T. (2021). Characterizing qanon: Analysis of youtube comments presents new conclusions about a popular conservative conspiracy. *First Monday*.

Neufeld, D. (2021). The 50 most visited websites in the world.

Paolillo, J. C. (2018). The flat earth phenomenon on youtube. *First Monday*.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Park, M., Naaman, M., and Berger, J. (2016). A data-driven study of view duration on youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Rosenbaum, L. (2021). Escaping catch-22—overcoming covid vaccine hesitancy.

Roth, C., Mazières, A., and Menezes, T. (2020). Tubes and bubbles topological confinement of youtube recommendations. *PloS one*, 15(4):e0231703.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Zhou, R., Khemmarat, S., and Gao, L. (2010). The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 404–410.