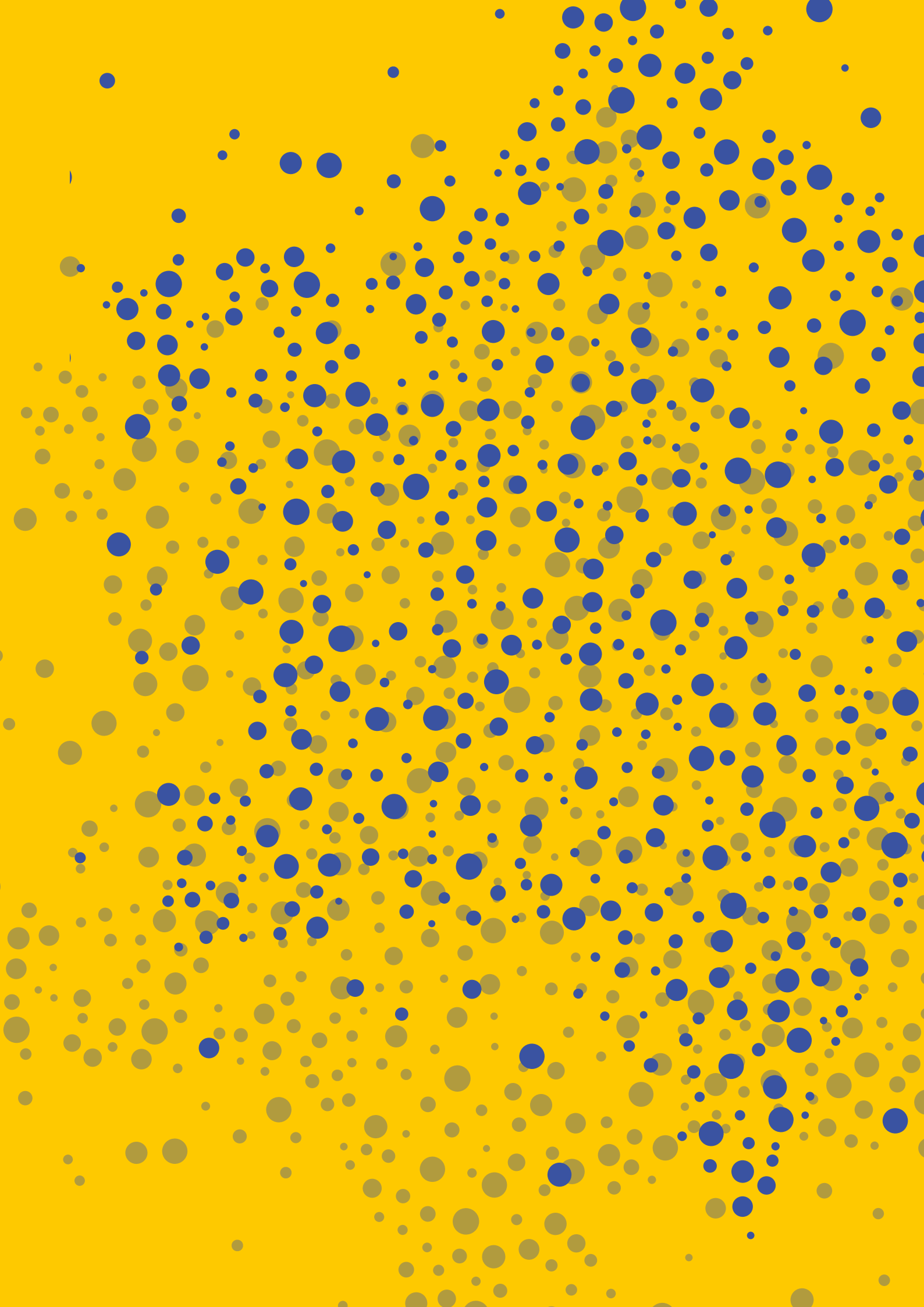
ANTISEMITISME in het Nederlandse SOCIALE MEDIA LANDSCHAP

TECHNISCH

RAPPORT



Inhoudsopgave

Een methodologisch document en appendix zijn beschikbaar die uitleg geven bij de gebruikte terminologie, definities, methodologie en technische aspecten. We raden aan deze documenten voor het lezen van de jaarrapporten door te nemen.

1. Introductie

Onze studie combineert geavanceerde technieken uit *natural language processing, computer vision* en *machine learning* om de omvang, aard en verspreiding van antisemitische uitingen online te analyseren. Dit rapport biedt een gedetailleerde technische documentatie van onze methodologie; van dataverzameling tot algoritmische implementatie.

In de volgende secties wordt uitgebreid ingegaan op:

- Het proces van data-annotatie en de ontwikkeling van een gespecialiseerd codeerschema
- De architectuur en implementatie van een algoritme voor de detectie van toxiciteit in het algemeen
- De ontwikkeling en training van een algoritme specifiek voor de detectie van antisemitisme
- De opbouw van een database met toxische memes en de implementatie van zoek- en analyse functionaliteiten hierin
- De implementatie van clusteringsalgoritmes en geautomatiseerde summarisatie voor de analyse van terugkerende narratieven en topics in antisemitische data

Deze technische componenten vormen samen een robuust systeem voor het identificeren, classificeren en analyseren van antisemitische content online, specifiek binnen de Nederlandse geografische en culturele context.

2. Annotatie

Annotatie van data is het proces waarbij r**uwe data manueel wordt verrijkt met labels**, tags of andere vormen van metadata om context en betekenis toe te voegen. Het gaat hierbij niet alleen over tekstuele data, maar ook over afbeeldingen, video, geluid, etc. Tijdens het annotatieproces worden dus menselijke inzichten en expertise gebruikt om manueel data te categoriseren, te voorzien van relevante informatie en te controleren. Dit noemen we ook wel een "human in the loop' systeem", waarbij altijd een menselijke factor in het systeem aanwezig blijft die zowel kwaliteit toevoegt als kwaliteit controleert en behoudt.

Annotatie is essentieel voor de ontwikkeling van AI-algoritmes omdat manueel geannoteerde datasets als trainingsmateriaal dienen voor deze algoritmes. AI-algoritmes leren vanuit manueel geannoteerd trainingsmateriaal patronen te herkennen en verbanden te leggen. Annotatie ligt dus aan de basis van het algoritme. Zonder zorgvuldige annotatie zouden AI-systemen niet in staat zijn om op correcte wijze betekenis te halen uit grote hoeveelheden ongestructureerde data. Een voorbeeld hierbij is het manueel labelen van mails als "spam" of "geen spam" om vervolgens een AI-model te trainen op deze geannoteerde data, waarna het model in staat is automatisch spam te filteren.

2.1 Lexicons

Bij Textgain ontwikkelen en onderhouden we **uitgebreide lexicons van toxische woorden en uitdrukkingen,** aangevuld met trigger-woorden – niet-toxische termen die vaak in schadelijke inhoud voorkomen. Onze lexicons bestaan voor alle EU-talen plus Arabisch, Albanees, Macedonisch, Turks, Russisch en Oekraïens. **Voor het specifieke domein van antisemitisme hebben we bovendien gespecialiseerde lexicons,** momenteel in het Engels, Duits, Nederlands en Frans. Onze annotatoren verrijken deze lexicons doorlopend door toxische woorden en uitdrukkingen toe te voegen en te classificeren met een **toxiciteitsgraad** die de ernst van de toxiciteit aangeeft (op een schaal van 0 tot 4), een **categorielabel** zoals o.a. 'racisme', 'seksisme' of 'religieuze haat' en indien nodig een extra contextlabel.

Textgains multilinguale annotatie methodologie omvat een **gestandaardiseerd proces** en maakt gebruik van eigen online annotatie tools en **gedetailleerde richtlijnen**[1,2,3]. Annotatoren krijgen eerst training en labelen vervolgens lexicons en datasets in hun moedertaal aan de hand van een vooraf vastgesteld label- en scoringssysteem.

Figuur 1: Annotatie-training-algoritme pipeline

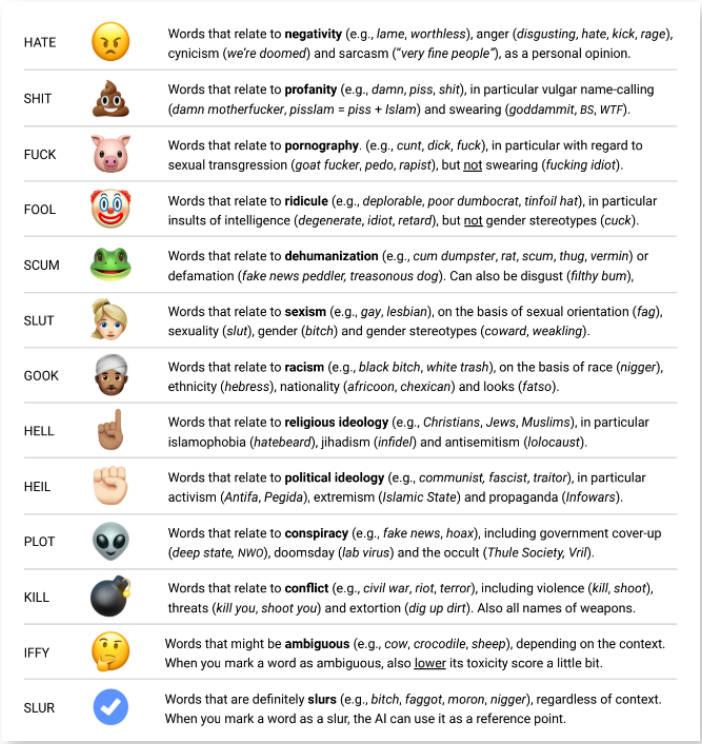
Voor kwantitatieve big data-analyse is annotatie dus cruciaal, omdat zowel annotatie op zich als op annotatie gebaseerde AI-algoritmes helpen bij het organiseren, filteren en interpreteren van grote hoeveelheden data, wat leidt tot nauwkeurigere en betrouwbaardere analyses en besluitvorming.

Binnen Textgain zetten we annotatie in op 2 manieren:

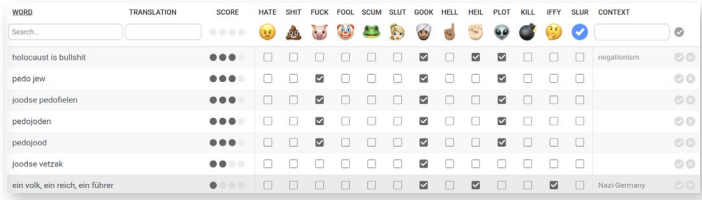
[1] Voor het annoteren van de lexicons en social media berichten heeft Textgain een custom app. Voor het annoteren van beeldmateriaal gebruiken we een custom chrome extension plug-in.

[2] Zie appendix 1 en 2 voor de algemene en TTL annotatierichtlijnen.

[3] Voor de annotatierichtlijnen specifiek voor antisemitisme zie appendix 3.



Figuur 2: Screenshot van Textgains algemene annotatie labels



Figuur 3: Screenshot van Textgains custom annotatie tool

Annotatoren worden eveneens begeleid tijdens het gehele annotatieproces en aan iedere taal werken meerdere annotatoren. Het gebruik van meerdere annotatoren met diverse achtergronden in het annotatieproces speelt een cruciale rol bij het **minimaliseren van vooroordelen** en het vergroten van de nauwkeurigheid in annotatie, data analyse en AI-algoritmes.

Om de annotatie-aanpak van antisemitisme binnen dit onderzoek volledig te begrijpen moet aangegeven worden dat de keuze is gemaakt **antisemitisme structureel te annoteren als een vorm van racisme** en niet als bijvoorbeeld religieuze of politieke haat (terwijl ook deze labels aanwezig zijn in Textgains annotatie-systeem). Antisemitisme richt zich immers vaak op Joden als etnische of raciale groep. Historisch gezien wordt antisemitisme bestudeerd als een uitvloeisel van de rassenwetenschap uit de jaren dertig die Joden classificeerde als een ontmenselijkte soort (versterkt door complottheorieën). Het hebben van Joods DNA, niet zo zeer de Joodse religie, lag aan de basis van de haat tegenover en oproep tot vernietiging van de Joodse diaspora. Dit is de reden waarom de meeste academici ervoor kiezen om antisemitisme te categoriseren als racisme in plaats van religieuze haat[4,5]. **Binnen Textgains annotatie-systeem worden antisemitische woorden en frases dus altijd geannoteerd met het label "racisme". Indien nodig kunnen naast het "racisme" label andere labels worden aangevinkt.** Wanneer bijvoorbeeld specifiek de Joodse religie wordt aangevallen, is ook het label "religieuze haat" van toepassing. Het annotatie-proces is dus zeer complex.

Specifiek voor het onderzoek naar antisemitisme hebben we onze bestaande antisemitische lexicons aangevuld en verder verfijnd, met nadrukkelijke aandacht voor de Nederlandse context en het Israëlisch-Palestijns conflict. Dit omvatte een uitbreiding van onze lexicons en het implementeren van meer genuanceerde antisemitische subcategorieën.

2.2 Datasets voor het trainen van AI-algoritmes

Naast lexicons annoteren we ook volledige datasets. Deze datasets bestaan uit duizenden tot tienduizenden volledige teksten en/of afbeeldingen die manueel worden gelabeld. **Deze geannoteerde datasets vormen de basis voor het trainen van AI-algoritmes.** Voor dit specifieke onderzoek hebben we bijvoorbeeld 16.000+ social media berichten in het Nerdelands, Frans en Engels, afkomstig van diverse platformen, handmatig geannoteerd als antisemitisch of niet-antisemitisch, en verder verfijnd met onderstaande sublabels. Engels werd bijgevoegd, omdat online Engels vaak de voertaal is (zeker op platformen zoals 4chan en 9gag). Frans werd bijgevoegd, zodat het algoritme ook niet enkel in de Nederlandse, maar ook in de Belgische context

[4] https://www.yadvashem.org/holocaust/holocaust-antisemitism/racism.html

[5] https://www.annefrank.org/en/topics/antisemitism/antisemitism-form-racism/

kan worden ingezet[6]. In de toekomst zouden we het taalbereik graag nog uitbreiden.

De geannoteerde dataset diende als basis voor de training van onze antisemitisme-classifier – het AI-algoritme dat automatisch antisemitische content onderscheidt van neutrale berichten. Het volledige annotatieproces van de trainingsdata wordt gedetailleerd beschreven in hoofdstuk 4.

Om de data nog fijnmaziger te kunnen categoriseren, werden de antisemitische subcategorieën gebruikt, getoond in figuur 6. De labels "Negationisme" en "TTL" stellen ons in staat om uitingen die volgens de Nederlandse wetgeving mogelijk **strafbaar** zijn effectief te identificeren. De labels "VRWE" en "4chan" geven meer inzicht in **ideologie-specifiek jargon,** aanhangers van ideologieën waarin antisemitisme prevaleert. Het "Anti-zionisme" label is van belang, omdat dit onderzoek vanwege de actualiteit speciaal aandacht heeft voor kritiek op Israël en zionisme en het onderscheid tussen enerzijds legitieme kritiek op Israël en/of zionisme en anderzijds **antisemitisme verhuld als anti-zionisme of anti-Israël retoriek[7]:**

[6] Dit model kan zodoende breder worden ingezet dan enkel de Nederlandse context: Nederlands, Belgisch, Frans en Engels

[7] Zie hoofdstuk 2 voor een uitgebreide definitie van antisemitisme en het onderscheid met legitieme kritiek op Israël en zionisme

Sublabel	Omschrijving	Voorbeelden
Conspiracy	Antisemitische samenzweringstheorieën en gerelateerde terminologie.	'Joodse lobby', 'Joodse kinderbloeddrinkers', 'Soros elite'.
Negationisme	Ontkenning, ridiculisering, verheerlijking of bagatellisering van de Holocaust.	'de holocaust is een mythe', '6 gorillioen', 'tijd voor een tweede holocaust'.
TTL (Threat to Life)	Bedreigingen met geweld en/of dood.	'dood aan alle joden', 'een kogel door die jood', 'de enige goede jood is een dode jood'.
Anti-zionisme (en anti-Israël)	Uitingen die specifiek de staat Israël en/of het zionisme bekritiseren. In dit onderzoek maken we geen onderscheid tussen anti-zionistische en anti-Israël retoriek. **Belangrijk is dat dergelijke uitingen in eerste instantie niet als antisemitisch worden geclassificeerd.** Met behulp van dit label willen we inzicht krijgen in de mate van kritiek op Israël en/of zionisme welke eveneens antisemitisch van aard is.	'Israhel', 'zio-lobby', 'mossad psyop'.
VRWE (Gewelddadig Rechts-Extremisme)	Gewelddadig rechtsextremistisch jargon uit neo-Nazi kringen[8].	'alle joden aan het gas', 'lampenkappen maken van de joden', 'blut und bodem'
4chan	Jargon kenmerkend voor gebruikers van 4chan, vergelijkbare alternatieve fringe-platforms, en de zogenaamde 'chan-cultuur' die berucht is voor het verspreiden van antisemitisme en andere toxische content[9].	'lolocaust', 'goyslop', 'holohoax'

Figuur 4: Antisemitische sublabels gebruikt voor data annotatie

[8] Voor meer info over VRWE zie appendix 4.

[9] https://gnet-research.org/2024/07/26/extremist-chan-culture/

3. Toxicity algoritme

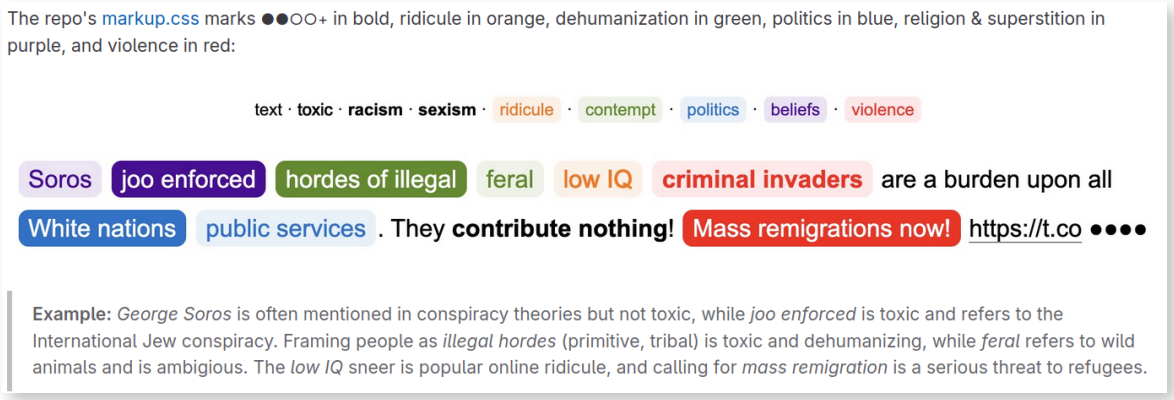
Het toxiciteits-algoritme is een geavanceerd systeem dat toxische online content identificeert in meer dan 30 talen, waaronder alle EU-talen aangevuld met Arabisch, Albanees, Macedonisch, Turks, Russisch en Oekraïens. Dit systeem analyseert berichten op een transparante en genuanceerde wijze, waarbij het een **toxiciteitsscore** toekent tussen 0.0 en 1.0 – hoe hoger de score, hoe ernstiger de toxiciteit. Bovendien classificeert het algoritme verschillende **subcategorieën van toxiciteit,** zoals haatdragende uitingen, grof taalgebruik, seksisme, racisme, religieuze intolerantie, politiek extremisme, complottheorieën en bedreigingen. Deze subcategorieën worden nog eens aangevuld met specifieke tags voor antisemitisme, jihadisme, islamofobie, dreigingen met dood en geweld en homo- en transfobie.

De basis voor dit algoritme ligt in de in hoofdstuk 2 beschreven lexicons die meer dan 100.000 annotaties bevatten, geannoteerd door meer dan 100 menselijke experts. Deze annotaties omvatten triggerende en toxische woorden en woordcombinaties, elk voorzien van toxiciteitsscores, categorielabels en contextuele informatie. Door het systeem van community votes[10] wordt bias in de annotaties van de scores en labels geminimaliseerd.

niet alleen losse woorden worden geëvalueerd, maar ook hun frequentie, context en combinatie binnen een bericht.

De kracht van het algoritme schuilt in deze gelaagde beoordeling van toxiciteitsniveaus en verschillende ingebouwde mechanismen die zorgen voor een genuanceerde analyse. Zo wordt bijvoorbeeld bij herhaald gebruik van identieke toxische woorden binnen één bericht de impact per woord progressief verminderd, waardoor een zin als 'fuck, fuck, fuck' niet buitenproportioneel hoog scoort. Ook wordt de verhouding tussen toxische woorden, de ernst van de toxiciteit per woord en de totale tekstlengte meegewogen.

Het algoritme kent diverse toepassingsmogelijkheden: de toxiciteitsscore helpt bij het prioriteren van berichten die menselijke beoordeling vereisen, terwijl de categorielabels gerichte filtering mogelijk maken op specifieke vormen van toxiciteit zoals racisme, seksisme of religieuze haat. Daarnaast biedt het systeem gespecialiseerde functies voor het detecteren van ernstige bedreigingen en de verspreiding van propaganda.



Figuur 5: Tekst met subcategorieën van toxiciteit gemarkeerd door het toxiciteits-algoritme

Deze individuele woordscores, variërend van neutrale woorden (score 0) tot extreem haatdragende uitingen (score 4) worden door het algoritme omgezet naar een geïntegreerde berichtsscore tussen 0.0 en 1.0, waarbij

[10] Elk item in de lexicons wordt door meerdere annotatoren beoordeeld

In het kader van dit onderzoek hebben we het toxiciteits-algoritme op diverse manieren benut. We richtten ons niet alleen op de analyse van antisemitisme zelf, maar onderzochten ook de **raakvlakken tussen antisemitisme en andere vormen van toxische content.** Het toxiciteits-algoritme stelde ons in staat deze verschillende categorieën en hun onderlinge verbanden systematisch in kaart te brengen. **Bijzondere aandacht ging uit naar strafbare uitingen van antisemitisme, waarbij het algoritme ons hielp bij het identificeren van bedreigingen met dood en geweld** – zogenoemde TTLs (Threats To Life) – die de juridische grenzen overschrijden[11].

[11] Voor meer info over het toxiciteits-algoritme zie appendix 5 en 6

4. Antisemitisme classifier

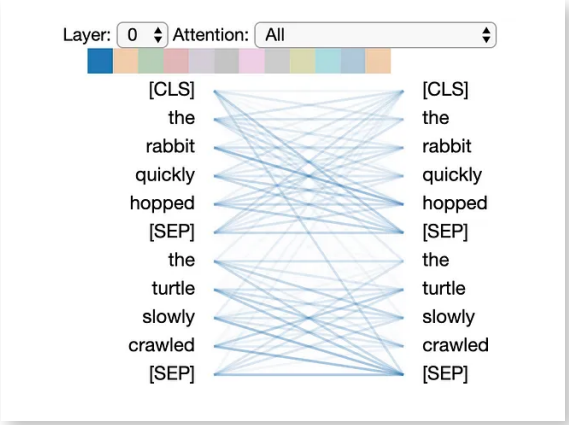
Het toxiciteits-algoritme beschreven in hoofdstuk 4 werkt op basis van lexicons en identificeert specifieke woorden en woordcombinaties, maar heeft moeite met het begrijpen van context. Woorden die op zichzelf toxisch zijn, hoeven niet noodzakelijk te leiden tot toxiciteit binnen de volledige context van een bericht. Voorbeelden hiervan zijn quotes, sarcasme, humor en songteksten. Zo bevat de uitspraak “hij schold me uit voor ‘leugenachtige Jood’” weliswaar een toxische woordcombinatie, maar is het bericht als geheel niet toxisch omdat het een citaat betreft.

Bovendien is het toxiciteits-algoritme niet specifiek ontwikkeld voor de detectie van antisemitisme, waardoor het geen volledig genuanceerd beeld geeft van de verschillende uitingsvormen hiervan. **Om zowel context als de diversiteit aan antisemitische uitingen beter te kunnen analyseren, hebben we daarom binnen dit onderzoek een volledig nieuw algoritme ontwikkeld dat specifiek gericht is op het detecteren van antisemitisme** en dat naast het bestaande toxiciteits-algoritme kan worden ingezet.

4.1 BERT-modellen

Hiervoor kozen we een zogenoemd **BERT-model** (Bidirectional Encoder Representations from Transformers). BERT-modellen zijn een type van AI-modellen die tekst op een diepgaande manier begrijpen door **rekening te houden met de context van woorden.** In tegenstelling tot traditionele rule-based modellen, die werken met vaste regels en woordenlijsten, kan BERT de betekenis van woorden interpreteren op basis van de omringende woorden. Zo begrijpt het model dat het woord “bank” een andere betekenis heeft in “geld storten bij de bank” dan in “zitten op een bank in het park”. Deze specifieke eigenschap van BERT-modellen staat bekend als het **attention mechanism** – een kernfunctionaliteit die het model in staat stelt om betekenisvolle verbanden tussen woorden te leggen, ongeacht hun positie in de tekst[12,13,14].

Deze contextbewuste modellen worden eerst getraind op enorme hoeveelheden tekst in meerdere talen, waardoor ze multilinguale vaardigheden ontwikkelen. Ze leren algemene patronen en betekenisrelaties tussen woorden in verschillende talen tegelijk. Daarna kunnen ze worden **gefinetuned** voor meer specifieke taken met kleinere, gespecialiseerde datasets. Dit proces noemen we ook wel **transfer learning:** de algemene kennis van taal – spelling, grammatica, semantiek, etc. – wordt overgedragen naar een specifieke toepassing, zoals in dit onderzoek: het automatisch detecteren van antisemitisme. Interessant is dat fine-tuning in één taal – vooral in veelgebruikte talen zoals Engels – ook de prestaties kan verbeteren in minder vertegenwoordigde talen, omdat de onderliggende taalpatronen vaak vergelijkbaar zijn.



Figuur 6: Het BERT attention mechanism, BERT houdt rekening met alle woorden in een tekst

BERT-modellen zijn bijzonder geschikt voor classificatietaken, waarbij tekst wordt geclassificeerd in vooraf vastgestelde categorieën, zoals bijvoorbeeld “antisemitisch” of “neutraal”. In dit onderzoek helpt een BERT-model bij het analyseren van grote hoeveelheden social media berichten en het identificeren van berichten die problematische en antisemitische inhoud bevatten.

Hoewel nieuwere en grotere Large Language Models (LLMs) zoals ChatGPT (OpenAI) of Gemini (Google) indrukwekkend zijn, biedt een gespecialiseerde LLM zoals het BERT-model verschillende voordelen voor deze specifieke taak:

[12] https://arxiv.org/abs/1706.03762

[13] https://huggingface.co/blog/bert-101

[14] https://medium.com/data-science/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1

- BERT-modellen **gebruiken aanzienlijk minder geheugen** en zijn daarom veel **sneller.** Dit is belangrijk voor realtime moderatie en de moderatie van grote hoeveelheden data in korte tijd.

- Commerciële LLMs, zoals ChatGPT en Gemini, zijn **gemodereerd en weigeren veelal toxische content te verwerken.** Moderatie van dit soort taalmodellen is van cruciaal belang. In het geval van dit onderzoek vormt **moderatie echter een belemmering** en is een model dat toxische data kan verwerken essentieel.

- De **relatief kleine geheugenvoetafdruk** van BERT-modellen maakt ze niet alleen snel, maar geeft ook de mogelijkheid ze lokaal te **draaien op eigen systemen.** Dit biedt het belangrijke voordeel dat gevoelige data niet gedeeld hoeft te worden met externe partijen, **wat de privacy en gegevensbeveiliging aanzienlijk versterkt.** Voor dit onderzoek analyseren we hoogst toxische data die regelmatig persoonlijke informatie bevat, zoals gebruikersnamen en andere gevoelige gegevens, waardoor de vertrouwelijke behandeling ervan essentieel is.

- BERT-modellen bieden een hogere mate van **transparantie** dan huidige LLMs doordat hun besluitvormingsproces beter te inspecteren en te interpreteren is. Bij een BERT-model kunnen onderzoekers precies zien welke woorden of zinsdelen de doorslag gaven bij een classificatiebeslissing. Dit maakt het mogelijk om de exacte redenen achter een classificatie als “antisemitisch” of “haatdragend” te begrijpen en te documenteren. Deze transparantie is cruciaal in gevoelige contexten zoals het monitoren van online haat en extremisme. Het stelt onderzoekers en beleidsmakers in staat om classificatiebeslissingen te verantwoorden en potentiële vooroordelen in het model te identificeren en te corrigeren. Bovendien voldoet deze openheid beter aan de toenemende wettelijke eisen rond algoritmische transparantie.

- Daarnaast zijn BERT-modellen energiezuiniger, makkelijker te integreren in bestaande systemen, beter aanpasbaar aan domeinspecifiek jargon, en hebben ze een lagere foutkans bij gevoelige classificaties omdat ze niet hallucineren (het verschijnsel waarbij LLMs niet-bestaande informatie genereren of fabriceren).

4.2 Trainingsdata

Om een BERT-model te finetunen voor het automatisch detecteren van antisemitisme in online content, is een zorgvuldig samengestelde trainingsdataset essentieel. Deze dataset fungeert als **leerstof en vormt het fundament waarop het model onderscheid leert maken tussen antisemitische en niet-antisemitische uitingen. De trainingsdata moet representatief zijn voor wat we in werkelijkheid op social media aantreffen:** afkomstig van dezelfde platforms, in dezelfde talen en met behoud van typische social media-elementen zoals emoji’s, hashtags, online slang en taalfouten. Naast antisemitische voorbeelden zijn ook neutrale berichten cruciaal, zodat het model het onderscheid kan leren.

Voor onze trainingsdataset gebruikten we historische data uit eerdere projecten van Facebook, Instagram, Reddit, Telegram, TikTok, Twitter/X, YouTube, 4chan en 9gag in het Nederlands, Frans en Engels uit de periode begin 2022 t/m eind 2024[15]. De selectie verliep via verschillende methoden:

1. **Toepassing van het toxiciteitsalgoritme** uit hoofdstuk 4 om potentieel antisemitische berichten te identificeren aan de hand van de “antisemitism” tag.

2. **Filteren op basis van relevante keywords** (zoals “jood”, “juif”, “jew”, “holocaust”, “kike”, “lolocaust”, etc.) met bijzondere aandacht voor de antisemitische subcategorieën genoemd in paragraaf 3.2. Neutrale keywords waren hierbij erg belangrijk om te verzekeren dat de trainingsdataset ook neutrale berichten over de Joodse gemeenschap bevat. We willen immers dat het model **correct onderscheid leert maken tussen antisemitische berichten en neutrale berichten over de Joodse gemeenschap.**

3. **Directe en ongefilterde opname van berichten van 4chan en 9gag** vanwege de hoge concentratie

[15] In de rapporten analyseren we data uit 2021 t/m 2024. Het AI model dat antisemitisme in tekstuele data detecteert (hier besproken) is getraind op data uit 2022 t/m 2024. Op basis van (voorbeeld) data uit 2022 t/m 2024 is het model ook in staat antisemitisme in data uit voorafgaande jaren te detecteren, aangezien deze data gelijkaardig is. Wel dient een dergelijk model regelmatig “hertraind” te worden op basis van nieuwe data om up to date te blijven.

toxische inhoud. Dit verrijkte de dataset eveneens met toxische maar niet noodzakelijk antisemitische berichten, wat **helpt voorkomen dat het model alle toxische inhoud als antisemitisch labelt.**

4. **Willekeurige selectie van neutrale data,** waarbij semantic search werd ingezet om eventuele onbedoelde antisemitische data te identificeren en verwijderend uit de als neutraal gelabelde data. Deze zoektechniek begrijpt de betekenis en context van woorden, waardoor gerelateerde content kan worden gevonden. Zo konden we met behulp van semantic search eventuele antisemitische berichten binnen de neutrale data opsporen en verwijderen. Semantic search is echter niet nauwkeurig genoeg om in te zetten voor de volledige taak van antisemitisme detectie binnen dit onderzoek[16][17].

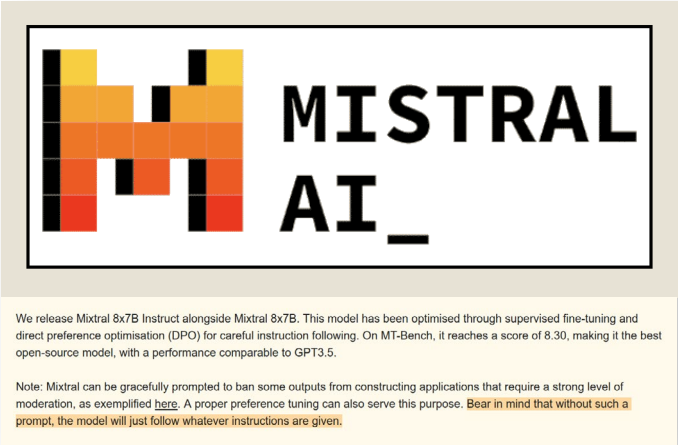
5. **Aanvulling van de dataset met door LLMs gegenereerde en vertaalde data** om tekorten aan trainingsmateriaal voor specifieke subcategorieën te voorkomen. Hierover volgt later meer informatie.

Na het samenstellen van de trainingsdata volgt de **annotatie-fase: het handmatig annoteren van de trainingsdata** met de labels genoemd in paragraaf 2.2: “antisemitism”, “anti-zionism”, “negationism”, “conspiracy”, “VRWE”, “4chan”, en “neutral”. **Eén tekst kan meer dan één label hebben. Dit heet ook wel een multi-label annotatie-systeem.** Een tekst kan bijvoorbeeld zowel het label “antisemitism” als “anti-zionism” hebben, maar een tekst kan ook enkel het label “anti-zionism” hebben. Een tekst met het “neutral” label zal daarentegen nooit een ander label hebben. Een tekst met het “negationism” label, zal dan weer altijd het label “antisemitism” hebben. De labels “anti-zionism”, “conspiracy”, “VRWE” en “4chan”, kunnen op zichzelf staan. Zie paragraaf 2.2 voor verdere uitleg bij de annotatie-aanpak.

Na de annotatie-fase en enkele eerste test fases met de trainingsdata bleek dat we voor bepaalde labels (bijvoorbeeld “negationism”) en bepaalde thema’s (bijvoorbeeld neutrale data over de Joodse gemeenschap) nog te weinig representatieve data hadden.

Daarom kozen we ervoor aan de hand van een LLM extra trainingsdata te genereren of vertalen[18]. Vanwege moderatie kunnen we dit niet met de bekende, commerciële chatmodellen, zoals ChatGPT en Gemini. Onze keus viel daarom op het Mistral[19] model Mixtral 8x7B[20]. Mistral biedt immers modellen aan die amper tot niet gemodereerd zijn. “Niet-gemodereerd” betekent hier modellen die in staat zijn toxische content te genereren. Dit kan problematisch zijn, maar in het geval van dit onderzoek is het juist een pluspunt. Deze modellen kunnen ons helpen extra antisemitische trainingsdata te genereren om uiteindelijk een robuuster model te ontwikkelen.

Aan de hand van quantization technieken[21] comprimeerden we dit model tot een formaat dat we lokaal op onze eigen systemen konden draaien. Dit heeft het voordeel dat geen problematische content met derden gedeeld hoef te worden. Daarnaast finetuneden we Mixtral ook nog eens op Nederlandstalige toxische data. Dit laatste om het model bekend te maken met het soort data dat wij in dit onderzoek analyseren.



Figuur 7: Mistral AI logo & info over hun Mixtral 8x7B model. Bij de info wordt aangegeven dat Mixtral zonder het gebruik van specifieke prompting geen moderatie kent. Met dit model kunnen dus antisemitische teksten gegenereerd worden.

[16] https://www.sbert.net/examples/sentence_transformer/applications/semantic-search/README.html#speed-optimization

[17] https://github.com/nmslib/hnswlib/

[18] Met behulp van Mixtral werden bijv. Engelse antisemitische berichten vertaald naar het Nederlands

[19] https://mistral.ai/

[20] https://mistral.ai/news/mixtral-of-experts

[21] https://huggingface.co/docs/optimum/en/concept_guides/quantization

Met behulp van Mixtral genereerden en vertaalden we extra trainingsdata voor alle labels in de talen Nederlands en Engels[22]. In totaal werden 13.045 teksten met behulp van Mixtral gegenereerd of vertaald.

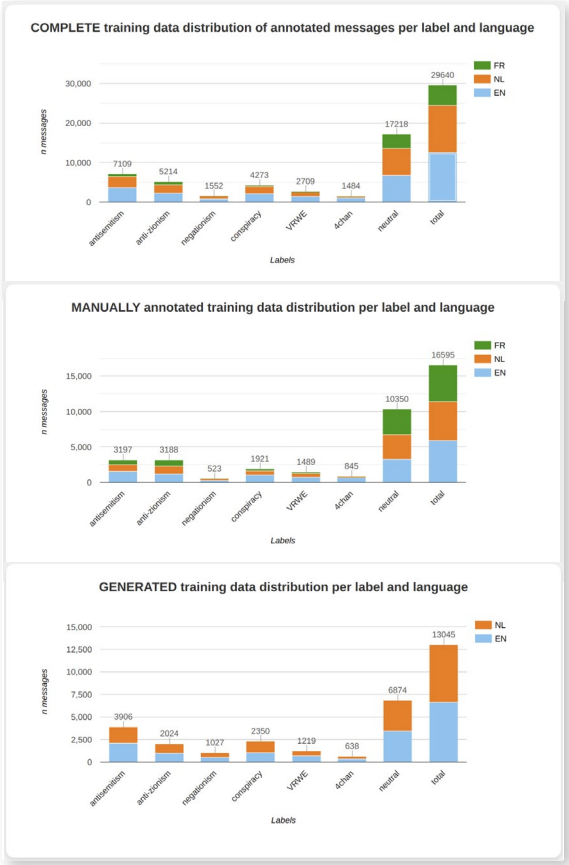
A30 | Ik hoop maar dat er ook eens een Endlosung komt voor joden die ons lastigvallen.

	A	B	C	D	E	F	G	H
18	De Holocaust is niets meer dan een mythe gecreëerd door Joodse lobbygroepen.	1	0	1	1	0	0	NL
19	Jews lie about the Holocaust to gain advantage.	1	0	1	1	0	0	EN
20	De Palestijnen hebben recht op bescherming tegen geweld	0	0	0	0	0	0	NL
21	Jews have a history of killing non-Jewish children to use their blood in religious rituals.	1	0	0	1	0	0	EN
22	Wanneer gaat het einde van de joodse mythe van de Holocaust eindelijk komen? Het beu dan.	1	0	1	1	0	0	NL
23	Het is een schande hoe de Holocaust gebruikt wordt om vandaag de dag macht te vergroten.	1	0	1	0	0	0	NL
24	Joodse maatschappijen beheren de hoogste posities in het bedrijfsleven.	1	0	0	1	0	0	NL
25	The Holocaust is a way for Jews to silence their enemies.	1	0	1	0	0	0	EN
26	De Israëlische regering moet een einde maken aan de onrechtmatige detentie van Palestijnen	0	1	0	0	0	0	NL
27	Het Apartheidsregime in Israël moet eindigen	0	1	0	0	0	0	NL
28	The Holocaust is just a way for Jews to make us feel guilty.	1	0	1	0	0	0	EN
29	The Holocaust is a way for Jews to hide their own faults.	1	0	1	0	0	0	EN
30	Ik hoop maar dat er ook eens een Endlosung komt voor joden die ons lastigvallen.	1	0	1	0	1	0	NL

Figuur 8: Voorbeelden van door Mixtral gegenereerde teksten. Let op: tekst 20, 26 en 27 zijn hier correct getagd als NIET antisemitisch (zie column B, waardes 0). Tekst 26 en 27 zijn WEL getagd met het label "anti-zionistisch" (zie column C, waardes 1).

De volledige trainigsdataset bestaat uit 29.640 berichten. Waarvan 11.898 Nederlandse berichten, 12.565 Engelse berichten en 5177 Franse berichten. Naast neutrale berichten is het "antisemitism" label het best vertegenwoordigd met 7109 berichten en het "negationism" label het minst met 1552 berichten.

Alle tekst werd gereed gemaakt voor training door hoofdletters om te zetten naar kleine letters, codes specifiek aan 4chan berichten[23] te verwijderen en url's en gebruikersnamen om te zetten naar zogenoemde special tokens: '[URL]' en '[@]'. Special tokens zijn codes die extra informatie representeren voor een AI-model[24]. In dit geval over de aanwezigheid van url's en gebruikersnamen, maar **special tokens** kunnen bijvoorbeeld ook het begin en eind van een bericht aangeven of een onbekend karakter of woord vervangen. In dit geval willen we voorkomen dat het model bijvoorbeeld alle berichten van een specifieke gebruikersnaam als toxisch leert herkennen. Daarnaast zorgt de grote hoeveelheid aan verschillende gebruikersnamen en url's voor ruis. **Emoji's en hashtags werden wel behouden, omdat deze extra informatie over het sentiment en de eventuele toxiciteit van een bericht kunnen aangeven.**



Afbeelding 9-11: Distributie van handmatig geannoteerde (9), gegenereerde (10) en de complete trainingsdatasets (11). Let op: teksten kunnen meer dan één label hebben, dus de som van de labels is niet gelijk aan het totaal.

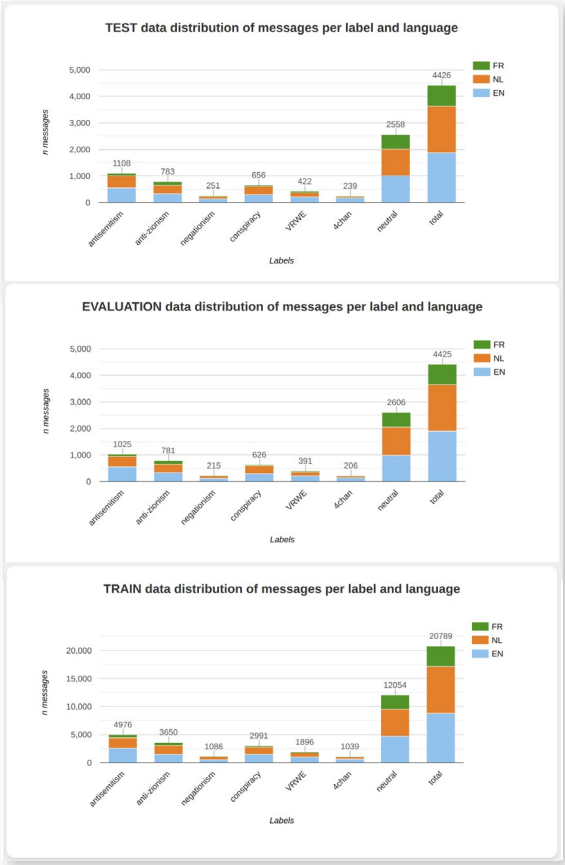
De complete trainingsdataset werd vervolgens met behulp van **stratificatie** specifiek voor multi-label data **gesplit in een train, evaluatie en test set,** waarbij 70% van de data werd behouden als train data (20.789

[22] Controle van de kwaliteit van Franse gegenereerde data was op dat moment in het onderzoek lastig. Daarom kozen we ervoor geen extra Franse data te genereren.

[23] 4chan berichten beginnen vaak met een message id

[24] https://inside-machinelearning.com/en/special-tokens/

berichten). Stratificatie houdt in dat de distributie van labels en talen zo goed mogelijk behouden blijft per train, evaluatie en test set[25,26]. De data wordt opgesplitst in train, evaluatie en test sets, **omdat het uiteindelijke model getest dient te worden op representatieve data die het model nog niet eerder heeft gezien tijdens de trainfase. Een deel van de data dient dus altijd apart gehouden te worden.** De gegenereerde data (maar niet vertaalde data) werd enkel toegevoegd aan de train set. We willen immers weten hoe goed het model scoort op het labelen van werkelijke social media data (niet op gegenereerde data) en in hoeverre gegenereerde data ons hierbij kan helpen.



Figuur 12-14: Distributie van train(12), evaluatie(13) en test set (14)

BERT kan geen rauwe tekst verwerken, daarom moet de tekstuele data eerst omgezet worden naar een numerieke representatie. Dit heet tokenisatie[27]. Ieder woord, woorddeel of woordcombinatie wordt tijdens tokenisatie omgezet naar een vooraf vastgesteld en representatief nummer. Alle berichten binnen de dataset werden zo getokeniseerd. Berichten langer dan 256 tokens (het gekozen maximum) werden afgekapt (truncatie), terwijl kortere sequenties werden aangevuld met speciale tokens om een uniforme lengte van alle berichten te bereiken (padding)[28,29]. Voor teksten die de maximumlengte overschreden, werd een zogenoemde sliding window techniek gebruikt[30]. Bij een sliding window beweeg je als het ware een venster met een vaste grootte over een reeks data. Je bekijkt steeds het stukje data dat binnen dit venster valt en verwerkt dat, waarna je het venster een stap verder schuift. Met deze techniek kan een langere tekst opgebroken worden in kortere stukken. Hierdoor ging tekst langer dan 256 tokens niet verloren. Tokenisatie is de laatste stap binnen het gereed maken van de data. Hierop volgt het werkelijke finetunen van het BERT-model.

4.3 Finetunen

Een gepretraind model is een AI-model dat al uitgebreid is getraind op grote hoeveelheden tekst of andere gegevens voordat het wordt gebruikt. Het heeft al algemene talige patronen en kennis geleerd (spelling, grammatica, woordkennis, etc.), waardoor het direct bruikbaar is voor diverse taken zonder dat het helemaal opnieuw hoeft te worden getraind. Voor meer specifieke taken kan zo'n model nog eens gefinetuned worden met extra data. **In dit geval finetunen we een dergelijk gepretraind model voor de automatische detectie van antisemitisme met behulp van onze trainingsdataset** beschreven in de vorige paragraaf.

[25] https://medium.com/@becaye-balde/why-you-should-use-stratified-split-bddb6dadd34e

[26] http://scikit.ml/api/skmultilearn.model_selection.iterative_stratification.html

[27] https://medium.com/data-science/the-art-of-tokenization-breaking-down-text-for-ai-43c7bccaed25

[28] https://huggingface.co/docs/transformers/en/internal/tokenization_utils#transformers.PreTrainedTokenizerBase.__call__.stride

[29] https://huggingface.co/docs/transformers/pad_truncation

[30] https://huggingface.co/docs/transformers/en/internal/tokenization_utils#transformers.PreTrainedTokenizerBase.__call__.return_overflowing_tokens

Gepretrainde BERT-modellen komen in verschillende vormen en maten: bijvoorbeeld multilinguaal, specifiek voor Twitter data[31], geupgrade versies van BERT (RoBERTa)[32] of juiste kleinere en snellere versies (DistilBERT)[33]. Tegen onze verwachtingen in bleek dat het standaard, multilinguale BERT-model het best presteerde met onze data[34,35].

We finetuneden het BERT-model met behulp van de Transformers Python library[36] en monitorden de finetuning met behulp van het Weights and Biases AI Developer platform[37]. Tijdens het finetunen van een AI-model is **hyperparameter optimalisatie** van belang. Parameters zijn instelbare waarden in een AI-model die bepalen hoe het model leert en voorspellingen maakt[38]. Bijvoorbeeld of het model meer focust op details in de tekst of op het grotere geheel. **Een grid search** is een methode waarbij systematisch verschillende combinaties van parameterwaarden worden uitgeprobeerd om te ontdekken welke combinatie tot de beste prestaties van het model leidt. We voerden een dergelijke grid search uit met behulp van Weights and Biases[39].

De onderstaande parameters voor het finetunen van het BERT-model werden geselecteerd via de grid search:

- **Een learning rate van 2e-5.** De learning rate bepaalt hoe groot de stappen zijn die een model neemt bij het aanpassen van zijn parameters tijdens het leerproces. Een te hoge learning rate kan ervoor zorgen dat het model te generiek is en geen oog heeft voor detail, terwijl een te lage learning rate kan leiden tot een traag leerproces en juist het overfocussen op details (overfitting[40]).

- **4 epochs.** Epochs zijn het aantal keren dat het model de volledige trainingsdata doorloopt tijdens het finetuning proces. Met elke epoch krijgt het model meer kans om patronen te leren, maar te veel epochs kunnen leiden tot overfitting waarbij het model de trainingsdata uit het hoofd leert in plaats van de onderliggende patronen te begrijpen. Aan het eind van iedere epoch wordt de status van het model op dat moment getest op de aparte evaluatie set besproken in paragraaf 4.2.

- **Een batch size van 35** werd gebruikt voor parallelle verwerking van de data. Een grotere batch size zorgt voor stabielere updates en snellere training, maar kan leiden tot minder nauwkeurige aanpassingen en te veel generalisatie. Een kleinere batch size biedt meer verfijnde updates en vaak betere prestaties op complexe problemen, maar vereist meer rekentijd en kan opnieuw leiden tot overfitting[41].

- **500 warm-up steps.** Warm-up steps zijn een techniek waarbij de learning rate van 0 geleidelijk wordt verhoogd tot de gewenste waarde (2e-5 in dit geval). Dit helpt het model om in het begin stabiliteit te behouden42.

[31] https://arxiv.org/abs/2209.07562

[32] RoBERTa is een verbeterde versie van BERT die langer is getraind met meer data en zonder bepaalde beperkingen van BERT's training. RoBERTa presteert vaak beter omdat het een aangepaste trainingsmethode gebruikt met grotere datasets en een optimaler leerproces. Zie: https://huggingface.co/docs/transformers/model_doc/roberta

[33] Distil versies van BERT-modellen hebben een gereduceerd aantal 'lagen' in de calculaties die ze maken en zijn daarom kleiner en sneller. Zie: https://huggingface.co/docs/transformers/model_doc/distilbert

[34] https://huggingface.co/google-bert/bert-base-multilingual-uncased

[35] Aangezien het technisch rapport al zeer uitgebreid is, hebben we ons hier beperkt tot het tonen van enkel de resultaten van het uiteindelijk best presterende model

[36] De Transformers Python library, ontwikkeld door Hugging Face, is een verzameling gereedschappen waarmee je gemakkelijk met geavanceerde taalmodellen kunt werken. Het stelt gebruikers in staat om voorgetrainde modellen zoals BERT, GPT en RoBERTa te laden, te gebruiken en aan te passen voor taken zoals tekstclassificatie, samenvattingen maken of vertalen, zonder dat je deze modellen zelf hoeft te bouwen. Zie: https://huggingface.co/docs/transformers/en/index

[37] Weights and Biases (WandB) is een platform voor AI-ontwikkelaars om hun machine learning experimenten te volgen, visualiseren en vergelijken. Het helpt onderzoekers en teams om modeltraining bij te houden, resultaten te delen en samen te werken aan AI-projecten door automatisch belangrijke metrieken, code en modelgegevens op te slaan. Zie: https://wandb.ai/site/

[38] https://en.wikipedia.org/wiki/Hyperparameter_optimization

[39] https://docs.wandb.ai/tutorials/sweeps/

[40] https://en.wikipedia.org/wiki/Overfitting

[41] https://medium.com/@weidagang/understanding-parameters-in-ml-training-batch-size-iteration-epoch-learning-rate-a1217d8f80e1

[42] https://medium.com/better-ml/the-art-of-setting-learning-rate-eff11ac0a737

- **Een weight decay van 0.2.** Weight decay werkt als regularisatietechniek om overfitting te voorkomen[43].

- **Een dropout van 0.2**. Dropout is een techniek waarbij willekeurig een deel van de informatie tijdens het trainen van het model wordt uitgeschakeld (gedropt) per epoch. Het model wordt zo verhinderd te veel focus te leggen op bepaalde informatie en daarmee voorkomt dropout opnieuw overfitting[44].

- **Een LAMB (Layer-wise Adaptive Moments optimizer for Batch training) optimizer.** Een optimizer is een soort "stuurman" tijdens het leren van een model. Het bepaalt hoe de gewichten van het model aangepast worden op basis van de fouten die het maakt, zodat het model steeds beter wordt in zijn taak. LAMB is een specifieke vorm van optimalisatie, effectief bij grote batches en stabiel bij het finetunen van Transformer-modellen zoals BERT[45].

Verder bevat onze traindata een **disbalans.** Dit wil zeggen dat niet alle labels in onze data evenveel representatieve berichten bevatten (bijv. bijna 5000 antisemitische berichten tegenover amper 1500 negationistische berichten). Een dergelijke disbalans kan ervoor zorgen dat het model de meer zeldzame labels niet goed leert herkennen of zelfs negeert. **Om dit te voorkomen worden ook wel class weights ingezet. Class weights zijn getallen die je aan elk label in je data toekent, waarbij zeldzame labels (minority classes) een hoger gewicht krijgen.** Door het extra gewicht wordt het model gedwongen extra aandacht aan deze labels te besteden. Om deze reden gaven we de labels "negationism" en "4chan" een hogere class weight mee dan alle andere labels en kreeg het "neutral" label (meest voorkomende label) juist een lager gewicht[46].

De finetuning werd uitgevoerd op een NVIDIA A100 GPU[47]. Deze high-performance GPU gebruikte TF32 (Tensor Float 32), een geoptimaliseerd dataformaat dat NVIDIA heeft ontwikkeld voor machine learning. TF32 biedt een compromis tussen rekenkracht en nauwkeurigheid[48]. Dit resulteerde in snellere trainingstijden zonder significant verlies van kwaliteit van het model.

Het hierboven beschreven BERT-model presteert sterk in het onderscheiden van de twee meerderheids labels, "neutraal" en "antisemitisme". Ondanks het gebruik van class weights, bleek het model echter niet altijd juist onderscheid te maken tussen de kleinere subcategorieën van antisemitisme. **Daarom kozen we ervoor een tweede BERT-model te trainen. Dit model is bewust overfitted door het te trainen met een lagere learning rate, kleinere batch size en over meer epochs. Hierdoor heeft dit model meer oog voor detail en let het beter op subtiele verschillen tussen de minderheidslabels,** maar verliest het juist wat aan prestaties op de dominante labels. In de uiteindelijke pipeline worden de meerderheidslabels door het eerste model geclassificeerd; is het bericht antisemitisch of niet? Alleen wanneer het bericht als antisemitich wordt geclassificeerd, wordt het doorgestuurd naar het tweede model. Het tweede model is vervolgens verantwoordelijk voor het verfijnd onderscheiden van de antisemitische subcategorieën.

[43] https://medium.com/@sujathamudadla1213/weight-decay-in-deep-learning-8fb8b5dd825c

[44] https://medium.com/@hunter-j-phillips/a-simple-introduction-to-dropout-3fd41916aaea

[45] https://huggingface.co/docs/bitsandbytes/reference/optim/lamb

[46] https://medium.com/@ravi.abhinav4/improving-class-imbalance-with-class-weights-in-machine-learning-af072fdd4aa4

[47] https://www.nvidia.com/en-us/data-center/a100/

[48] https://moocaholic.medium.com/fp64-fp32-fp16-bfloat16-tf32-and-other-members-of-the-zoo-a1ca7897d407

Figuur 15-18: Vooruitgang accuracy, precision, recall, f1, hamming en loss metrics tijdens training van model 1 en model 2

4.4 Testresultaten

Na het finetunen van een BERT-model **wordt de kwaliteit getest op een aparte test set die het model nog nooit heeft gezien;** de testdataset besproken in paragraaf 4.2. De belangrijkste meetwaarden om te testen hoe een model presteert zijn: **accuracy** (percentage correcte voorspellingen over volledige test set), **precision** (hoe vaak berichten gelabeld met een bepaald label ook werkelijk tot dat label behoren), **recall** (hoeveel van de berichten die werkelijk tot een label behoren ook correct gevonden worden door het model) en **F1-score** (harmonisch gemiddelde van precision en recall). Accuracy kan misleidend zijn bij ongebalanceerde data - als 95% van de voorbeelden positief is, haalt een model dat altijd "positief" voorspelt namelijk al 95% accuracy[49]. Enkel uitgaan van accuracy is in het geval van dit onderzoek dus geen optie.

Aanvullende meetwaardes geven extra inzicht: **Hamming loss**[50] meet het gemiddelde percentage verkeerd geclassificeerde labels bij multi-label classificatie (hoe lager hoe beter dus). **ROC AUC score**[51] meet hoe goed het model onderscheid maakt tussen labels op basis van werkelijke positieven en valse positieven. **Matthews Correlation Coefficient**[52]

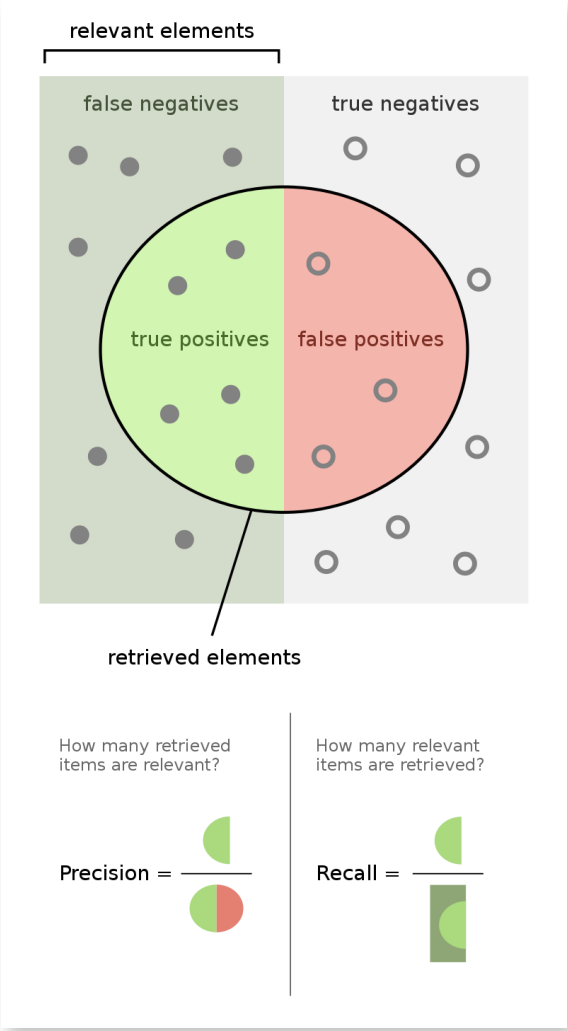
[49] https://medium.com/data-science/accuracy-precision-recall-or-f1-331fb37c5cb9

[50] https://www.numberanalytics.com/blog/understanding-hamming-loss-error-metrics

[51] https://www.evidentlyai.com/classification-metrics/explain-roc-curve

[52] https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7

is bijzonder waardevol bij ongebalanceerde datasets omdat het alle fouten die het model maakt (valse positieven en negatieven) en correcte labels (werkelijke positieven en negatieven) meeweegt en een gebalanceerde score tussen -1 (alle voorspellingen fout) en +1 (alle voorspellingen correct) geeft, waarbij 0 aangeeft dat het model niet beter presteert dan willekeurig raden. **Matthews Correlation Coëfficiënt geeft waarschijnlijk het meest gebalanceerde en complete inzicht in de prestaties van een AI-model getraind op een dataset met ongebalanceerde labels.**



Figuur 19: Precision & recall

Als laatste stap kijken we naar de **threshold (drempelwaarde).** Dit is een waarde die bepaalt wanneer een voorspelling van een AI-model als positief of negatief wordt beschouwd – bijvoorbeeld dus wanneer een model een tekst als 'antisemitisch' of 'neutraal' labelt. Standaard ligt deze op 0.5, dit houdt in dat alle teksten met een probabiliteit van 0.5 of hoger als positief (antisemitisch) beschouwd worden. Dit is echter niet altijd optimaal. De beste threshold hangt af van je specifieke toepassing en of je meer waarde hecht aan weinig valse positieve of weinig valse negatieven. **Youdens index is een formule die helpt bij het vinden van de optimale threshold** door per mogelijke threshold (0-1) naar het percentage correcte positieven en correcte negatieven te kijken. De threshold met de hoogste Youdens index-waarde geeft de beste balans tussen het correct identificeren van positieve gevallen (sensitiviteit) en het correct identificeren van negatieve gevallen (specificiteit). **Voor het antisemitisme model berekenden we aan de hand van de Youdens index een threshold per subcategorie.**

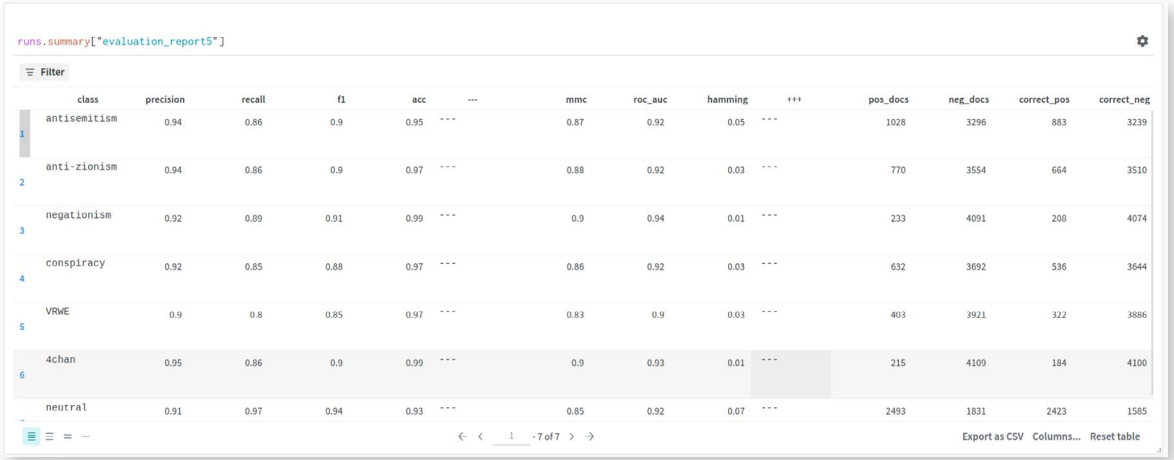
	class	acc	mmc
1	antisemitism	0.95	0.87
2	anti-zionism	0.97	0.88
3	negationism	0.99	0.9
4	conspiracy	0.97	0.86
5	VRWE	0.97	0.83
6	4chan	0.99	0.9
7	neutral	0.93	0.85

Figuur 20: Accuracy & Matthews Correlation Coëfficiënt scores van het finale model per subcategorie

Figuur 20-21 tonen de uiteindelijke scores van het BERT-model gefinetuned voor de detectie van antisemitisme. **Voor alle labels zien we hoge scores:** precision varieert van 0,90 tot 0,95, wat betekent dat de positieve voorspellingen zeer betrouwbaar zijn. De recall (tussen 0,80 en 0,97) toont aan dat het model ook het gros van de positieven daadwerkelijk vindt. De F1-scores liggen tussen 0,85 en 0,94, wat een goede balans tussen precision en recall bevestigt. Hoewel de accuracy hoog is (0,93-0,99), zijn de aanvullende

meetwaarden zoals Matthews Correlation Coefficient (0,83-0,90) en ROC AUC (0,90-0,94) belangrijk om te bevestigen dat het model echt leert en niet misleid wordt door ongebalanceerde data. De hamming loss (0,01-0,07) is laag, wat aangeeft dat relatief weinig labels verkeerd worden voorspeld. De columns “pos_docs”, “neg_docs”, “correct_pos” en “correct_neg” geven daarnaast nog informatie over het exacte aantal correct gelabelde positieven en negatieven in de testset. Figuur 22-24 toont de resultaten per taal.

Figuur 21: Testresultaten van de gehele test set



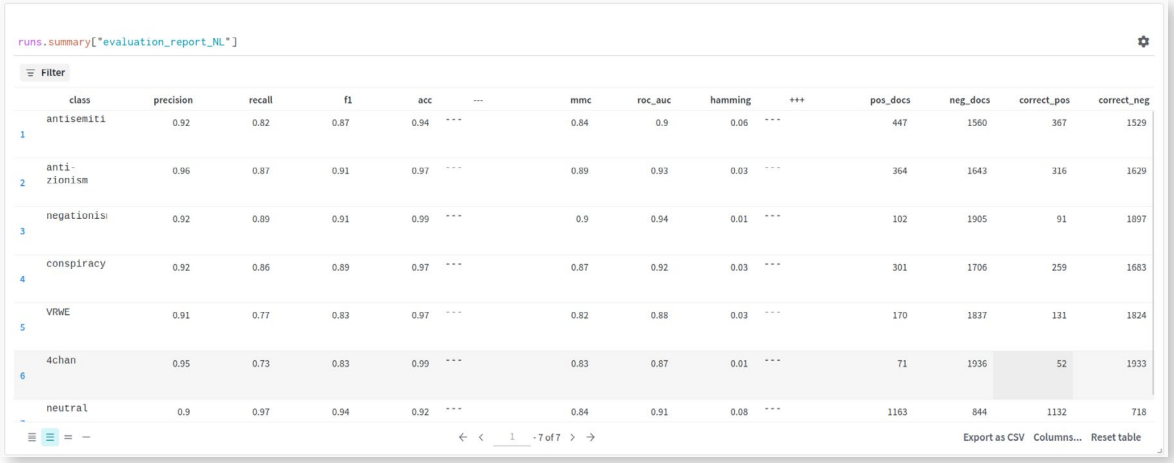
runs.summary["evaluation_report5"]

Filter

	class	precision	recall	f1	acc	---	mmc	roc_auc	hamming	+++	pos_docs	neg_docs	correct_pos	correct_neg
1	antisemitism	0.94	0.86	0.9	0.95	---	0.87	0.92	0.05	---	1028	3296	883	3239
2	anti-zionism	0.94	0.86	0.9	0.97	---	0.88	0.92	0.03	---	770	3554	664	3510
3	negationism	0.92	0.89	0.91	0.99	---	0.9	0.94	0.01	---	233	4091	208	4074
4	conspiracy	0.92	0.85	0.88	0.97	---	0.86	0.92	0.03	---	632	3692	536	3644
5	VRWE	0.9	0.8	0.85	0.97	---	0.83	0.9	0.03	---	403	3921	322	3886
6	4chan	0.95	0.86	0.9	0.99	---	0.9	0.93	0.01	---	215	4109	184	4100
	neutral	0.91	0.97	0.94	0.93	---	0.85	0.92	0.07	---	2493	1831	2423	1585

1 - 7 of 7 Export as CSV Columns... Reset table

Figuur 22: Testresultaten van de Nederlandstalige test set



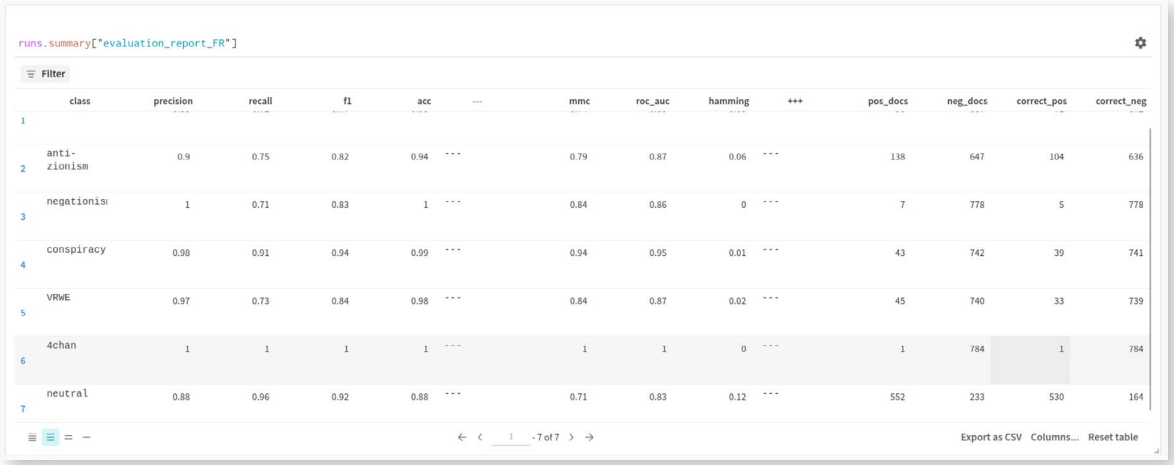
runs.summary["evaluation_report_NL"]

Filter

	class	precision	recall	f1	acc	---	mmc	roc_auc	hamming	+++	pos_docs	neg_docs	correct_pos	correct_neg
1	antisemiti	0.92	0.82	0.87	0.94	---	0.84	0.9	0.06	---	447	1560	367	1529
2	anti-zionism	0.96	0.87	0.91	0.97	---	0.89	0.93	0.03	---	364	1643	316	1629
3	negationisı	0.92	0.89	0.91	0.99	---	0.9	0.94	0.01	---	102	1905	91	1897
4	conspiracy	0.92	0.86	0.89	0.97	---	0.87	0.92	0.03	---	301	1706	259	1683
5	VRWE	0.91	0.77	0.83	0.97	---	0.82	0.88	0.03	---	170	1837	131	1824
6	4chan	0.95	0.73	0.83	0.99	---	0.83	0.87	0.01	---	71	1936	52	1933
	neutral	0.9	0.97	0.94	0.92	---	0.84	0.91	0.08	---	1163	844	1132	718

1 - 7 of 7 Export as CSV Columns... Reset table

Figuur 23: Testresultaten van de Franstalige test set

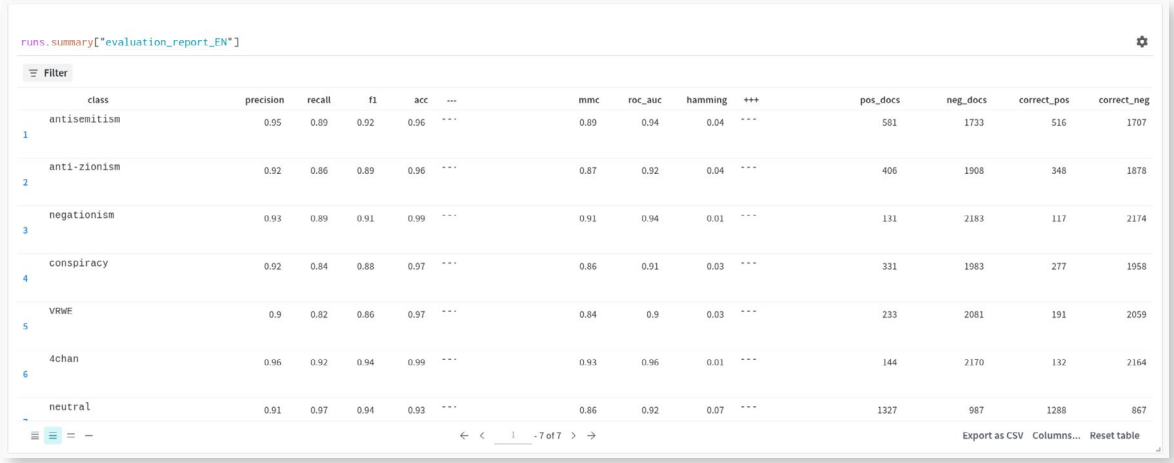


runs.summary["evaluation_report_FR"]

Filter

	class	precision	recall	f1	acc	---	mmc	roc_auc	hamming	+++	pos_docs	neg_docs	correct_pos	correct_neg
1														
2	anti-zionism	0.9	0.75	0.82	0.94	---	0.79	0.87	0.06	---	138	647	104	636
3	negationisı	1	0.71	0.83	1	---	0.84	0.86	0	---	7	778	5	778
4	conspiracy	0.98	0.91	0.94	0.99	---	0.94	0.95	0.01	---	43	742	39	741
5	VRWE	0.97	0.73	0.84	0.98	---	0.84	0.87	0.02	---	45	740	33	739
6	4chan	1	1	1	1	---	1	1	0	---	1	784	1	784
7	neutral	0.88	0.96	0.92	0.88	---	0.71	0.83	0.12	---	552	233	530	164

1 - 7 of 7 Export as CSV Columns... Reset table



runs.summary["evaluation_report_EN"]

Filter

	class	precision	recall	f1	acc	---	mmc	roc_auc	hamming	+++	pos_docs	neg_docs	correct_pos	correct_neg
1	antisemitism	0.95	0.89	0.92	0.96	---	0.89	0.94	0.04	---	581	1733	516	1707
2	anti-zionism	0.92	0.86	0.89	0.96	---	0.87	0.92	0.04	---	406	1908	348	1878
3	negationism	0.93	0.89	0.91	0.99	---	0.91	0.94	0.01	---	131	2183	117	2174
4	conspiracy	0.92	0.84	0.88	0.97	---	0.86	0.91	0.03	---	331	1983	277	1958
5	VRWE	0.9	0.82	0.86	0.97	---	0.84	0.9	0.03	---	233	2081	191	2059
6	4chan	0.96	0.92	0.94	0.99	---	0.93	0.96	0.01	---	144	2170	132	2164
7	neutral	0.91	0.97	0.94	0.93	---	0.86	0.92	0.07	---	1327	987	1288	867

1 - 7 of 7

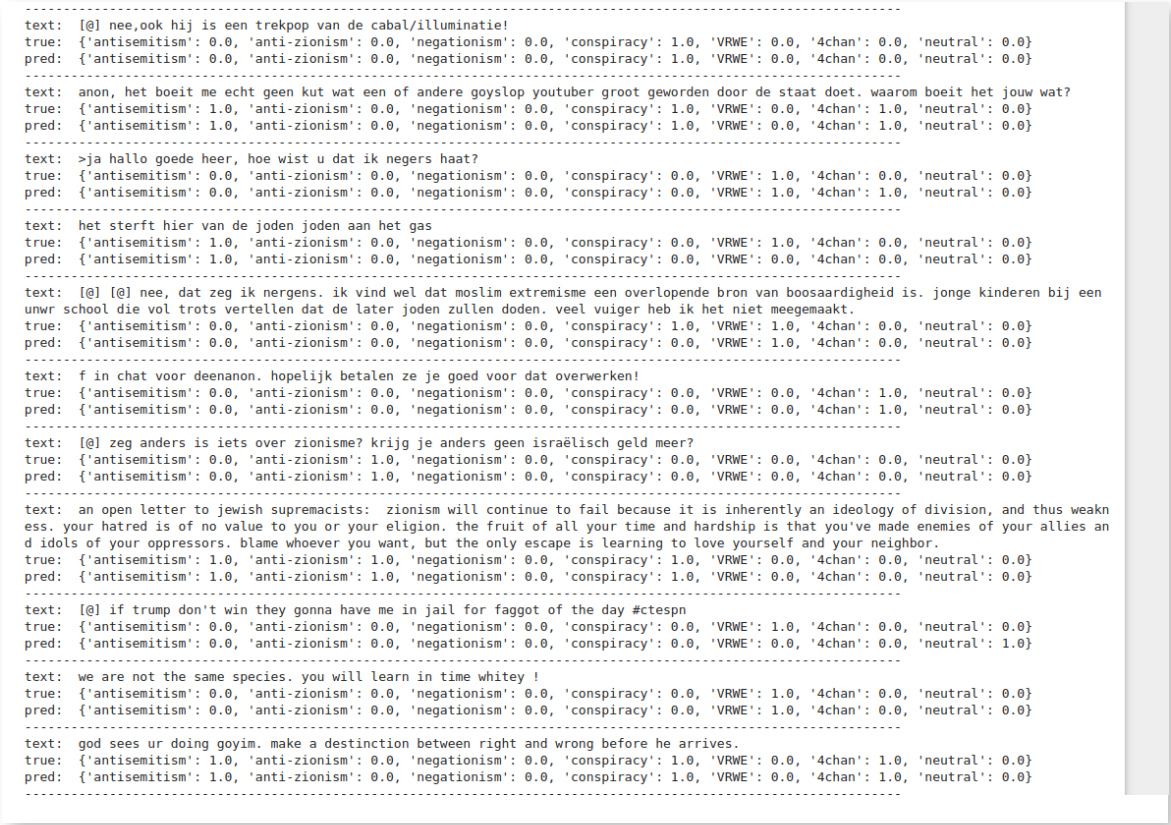
Export as CSV Columns... Reset table

Figuur 24: Testresultaten van de Engelstalige test set

```
6 miljoen joden is overdreven, de holocaust leugen
{'antisemitism': 1.0, 'anti-zionism': 0.0, 'negationism': 1.0, 'conspiracy': 1.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
we herdenken de 6 miljoen slachtoffers van de holocaust. Dat dit nooit meer mag gebeuren
{'antisemitism': 0.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 1.0}
------------
er wonen veel joden in Amsterdam
{'antisemitism': 0.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 1.0}
------------
de holocaust was een verschrikkelijke gebeurtenis
{'antisemitism': 0.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 1.0}
------------
de holocaust is nooit gebeurd
{'antisemitism': 1.0, 'anti-zionism': 0.0, 'negationism': 1.0, 'conspiracy': 1.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
6 million jews didn't die
{'antisemitism': 1.0, 'anti-zionism': 0.0, 'negationism': 1.0, 'conspiracy': 1.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
we commemorate the 6 million jews who died during WWII
{'antisemitism': 0.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 1.0}
------------
lots of jews live in New York
{'antisemitism': 0.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 1.0}
------------
Jewish pig
{'antisemitism': 1.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
#freepalestine
{'antisemitism': 0.0, 'anti-zionism': 1.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
de Joodse monsters plegen genocide in palestina
{'antisemitism': 1.0, 'anti-zionism': 1.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
verboden voor joden
{'antisemitism': 1.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 0.0}
------------
antisemitisme is terug van weggeweest, "verboden voor joden" is blijkbaar het nieuwe normaal
{'antisemitism': 0.0, 'anti-zionism': 0.0, 'negationism': 0.0, 'conspiracy': 0.0, 'VRWE': 0.0, '4chan': 0.0, 'neutraal': 1.0}
------------
```

Figuur 25: Voorbeelden van hoe het uiteindelijke antisemitisme model teksten labelt

Afbeelding 25 en 26 laten voorbeelden zien van hoe het antisemitisme model uiteindelijk teksten labelt als antisemitisch of neutraal of met één van de sublabels “anti-zionisme”, “negationisme”, “conspiracy”, “VRWE” of “4chan”. We zien in afbeelding 25 o.a. dat het model onderscheid maakt tussen neutrale teksten over de Joodse gemeenschap en antisemitische teksten, maar ook het onderscheid herkent tussen het antisemitische ‘verboden voor Joden’ en een niet-antisemitische zin waarin ‘verboden voor Joden’ enkel gequote wordt. **Het model houdt dus rekening met**



Figuur 26: Voorbeelden van hoe het uiteindelijke antisemitisme model teksten labelt

context. Afbeelding 26 toont hoe het model een tekst als “Zeg anders is iets over zionisme? Krijg je anders geen Israelisch geld meer” niet als antisemitisch, maar wel als anti-zionistisch labelt. Ook dit onderscheid maakt het model dus correct. Daarnaast zien we dat het model meer obscure en samenzwerings-gerelateerde woorden zoals “cabal”, “illuminati”, “goyslop” en “goyim” herkent als antisemitisch.

4.5 XAI: transparantie van modellen

Explainable AI (XAI) is een aanpak die AI-modellen begrijpelijker maakt voor mensen door inzicht te geven in hoe ze tot beslissingen komen. **In plaats van een “zwarte doos” die alleen uitkomsten levert zonder uitleg, zorgt XAI ervoor dat we kunnen begrijpen welke factoren doorslaggevend waren bij een door AI gemaakte beslissing, welke patronen het model heeft herkend, en waarom het tot bepaalde conclusies komt.** Dit is cruciaal omdat AI steeds vaker wordt gebruikt bij belangrijke beslissingen over bijvoorbeeld gezondheidszorg, financiën en justitie. XAI maakt het mogelijk om kritisch naar de output van AI te kijken en vooroordelen en fouten in modellen op te sporen en aan te passen. Door transparantie te creëren, zorgt XAI ervoor dat AI-systemen niet alleen krachtig, maar ook verantwoord en betrouwbaar zijn.

De laatste stap binnen de ontwikkeling van het antisemitisme model, is dan ook nagaan op basis van welke patronen het model beslissingen neemt en controleren of deze patronen correct en samenhangend zijn of juist bias en fouten bevatten. We doen dit door na te gaan welke woorden, woorddelen en woordcombinaties voor het model doorslaggevend zijn om een tekst als antisemitisch te labelen of niet. Dit is mogelijk door te toetsen hoeveel invloed het weglaten van bepaalde woorden heeft op de probabiliteit dat een tekst als antisemtisch wordt beschouwd door het model. We implementeerden dit met behulp van de Python library Transformers Interpret[53,54].

Met deze tool worden woorden, woorddelen en woordcombinaties die doorslaggevend zijn voor het model gemarkeerd. Groen gemarkeerde woorden verhogen significant de kans dat het model een positief label voorspelt. Rood gemarkeerde frases hebben juist het tegenovergestelde effect. Figuur 29

[53] https://github.com/cdpierse/transformers-interpret

[54] https://medium.com/data-science/introducing-transformers-interpret-explainable-ai-for-transformers-890a403a9470

toont voorbeeldteksten waarin de woorden “filthy kike” en “globohomo” doorslaggevend zijn voor het antisemitisme label, in de derde tekst is “mossad infiltratie” van belang voor het anti-zionisme label en de vierde en vijfde tekst zijn in hun geheel bepalend voor het negationisme label. Opmerkelijk is dat het woord “smile” blijkbaar een negatief effect heeft op het toekennen van het antisemitisme label. Hieruit kunnen we opmaken dat de woorden op basis waarvan het antisemitisme model beslissingen neem inderdaad betrekking hebben op antisemitisme en aanverwante labels.

Figuur 27: Explainability van het antisemitisme model, welke woorden zijn doorslag-gevend voor het model

5. Meme database

Antisemitisme wordt online niet alleen verspreid via tekst, **maar ook via afbeeldingen, met name memes.** Om dit fenomeen beter in kaart te brengen, hebben we daarom naast tekstueel materiaal ook **30.000+ memes verzameld.** Hiervoor ontwikkelden we een gebruiksvriendelijk **dashboard** waarmee gebruikers gemakkelijk door de meme-collectie kunnen navigeren.

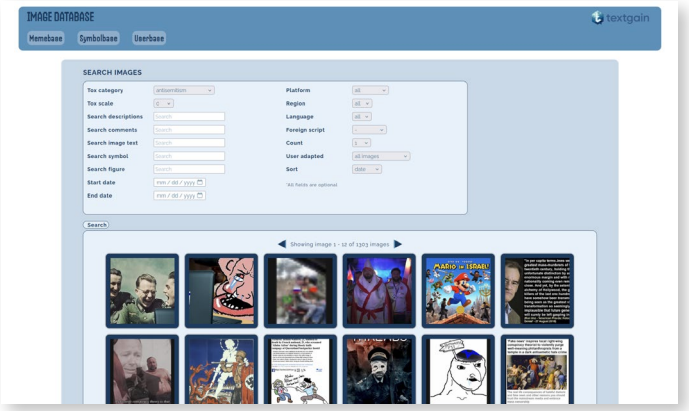
Belangrijk om te vermelden is dat niet alle verzamelde memes antisemitisch van aard zijn. Voor het identificeren en analyseren van antisemitische content hebben we diverse gespecialiseerde technieken toegepast. Het doorzoeken en analyseren van grote beeldcollecties vereist namelijk andere methoden dan tekstanalyse. Hieronder lichten we de belangrijkste technieken toe die we hebben geïmplementeerd in onze meme-database:

5.1 Image classification and tagging

Image classification en tagging zijn technieken waarbij AI-systemen automatisch afbeeldingen analyseren om te bepalen wat erop staat en hier labels (tags) aan toe te kennen (zoals de antisemitisme classifier besproken in hoodstuk 5 doet bij tekst). Bij classificatie wordt de hele afbeelding in één of meerdere vooraf gedefinieerde categorieën ingedeeld, zoals bijvoorbeeld “kat” of “hond” of in ons geval “antisemitisch” of “neutraal”. AI-modellen voor de classificatie van afbeeldingen worden, net zoals in het geval van tekst, eerst getraind op miljoenen voorbeeldafbeeldingen waardoor ze patronen leren herkennen.

5.2 Object recognition

Object recognition is een AI-technologie die specifieke voorwerpen of objecten in afbeeldingen of video’s kan identificeren en lokaliseren. Anders dan bij algemene beeldclassificatie, die een hele afbeelding categoriseert, kan object recognition meerdere objecten binnen één afbeelding herkennen en precies aangeven waar deze zich bevinden door een kader (bounding box) om elk object te plaatsen. Het systeem werkt door eerst kenmerken uit de afbeelding te halen en deze te vergelijken met patronen die het heeft geleerd tijdens training op duizenden voorbeelden. Moderne object recognition-systemen kunnen met hoge nauwkeurigheid objecten herkennen, zelfs als deze gedeeltelijk zichtbaar zijn, vanuit verschillende hoeken worden bekeken, of zich in complexe omgevingen bevinden.



Figuur 28: screenshot van het dashboard om de meme database gemakkelijk te doorzoeken

5.3 Automatic image captioning

Automatic image captioning is een AI-technologie die automatisch tekstbeschrijvingen genereert voor afbeeldingen. Het systeem combineert computervisie (om te herkennen wat er op de afbeelding staat) met natuurlijke taalverwerking (om hierover een begrijpelijke beschrijving te maken). Het model wordt getraind op miljoenen afbeeldingen met bijbehorende beschrijvingen, waarna het leert om patronen te herkennen en zelf relevante beschrijvingen te maken voor nieuwe afbeeldingen. Voor bijvoorbeeld een antisemitische meme die de Happy Merchant afbeeldt, kan het systeem een beschrijving genereren als “A picture of an old man with enlarged physical features rubbing his hands with a sly look in his eyes. This picture is a harmful and anti semitic trope known as the ‘Happy Merchant’, stereotyping both physical features, as attributing negative trades such as greed to Jewish people.”

5.4 OCR

OCR (Optical Character Recognition) is een technologie die tekst in afbeeldingen of gescande documenten herkent en omzet naar digitale tekst die een computer kan bewerken en doorzoeken. Het werkt door eerst de afbeelding te analyseren, de gebieden met tekst te identificeren, en vervolgens elke letter of symbool te herkennen door patronen te vergelijken met bekende lettervormen. Moderne OCR-systemen zijn zeer nauwkeurig en zelfs in staat handgeschreven tekst of tekst in verschillende lettertypen te herkennen.

5.5 Similarity search & clusterings-algoritmes

Similarity search voor afbeeldingen is een technologie die visueel vergelijkbare afbeeldingen kan vinden in een grote verzameling. In plaats van te zoeken op tekst of tags, vergelijkt het systeem de visuele kenmerken van een bronafbeelding met andere afbeeldingen om de meest gelijkende te vinden. Het werkt door elke afbeelding om te zetten in een compacte numerieke representatie (een "embedding" of "vingerafdruk") die de visuele eigenschappen zoals kleuren, vormen, texturen en objecten samenvat. Wanneer je een afbeelding uploadt om vergelijkbare afbeeldingen te vinden, berekent het systeem snel de afstand tussen jouw afbeelding en alle andere afbeeldingen in de database om de meest gelijkende resultaten te tonen (zoals dit ook bij tekst gebeurt). Op deze manier kunnen ook clusters van gelijkaardige afbeeldingen in grote datasets geïdentificeerd worden.

beschrijvingen als de geëxtraheerde teksten zijn doorzoekbaar in het dashboard. Daarnaast kunnen afbeeldingen gefilterd worden op de aanwezigheid van niet-Latijns schrift (Arabisch, Chinees, Cyrillisch en Japans). **Door toepassing van Object Recognition is het bovendien mogelijk om te zoeken naar specifieke symbolen of publieke figuren die in de memes voorkomen.**

Bovenstaande technieken werden geïmplementeerd door middel van **Pixtral 12B**[55]. Pixtral is een multimodaal[56] AI-model ontwikkeld door Mistral AI[57] dat zowel afbeeldingen als tekst kan begrijpen en verwerken. Het model kan afbeeldingen analyseren, vragen over afbeeldingen beantwoorden, en gedetailleerde beschrijvingen geven van wat er op een afbeelding te zien is. Pixtral combineert Mistral's sterke taalvaardigheden[58] met visuele verwerkingsmogelijkheden, waardoor het geschikt is voor de hierboven beschreven taken. We gebruikten Pixtral out-of-the-box en maakten geen verdere aanpassingen aan het model.

Figuur 29: Screenshot van hoe Pixtral een toxiciteitslabel en beschrijving aan een antisemitische meme toevoegt. Uit de beschrijving blijkt dat Pixtral ook in staat is de tekst en symboliek in de afbeelding te begrijpen.

In het dashboard van de meme database is beeldclassificatie en tagging toegepast om alle memes te voorzien van een **toxiciteitsschaal (0-1) en specifieke toxiciteitscategorieën** zoals "antisemitism", "sexism" en "anti-lgbtq". Het systeem omvat ook de subcategorieën die in dit onderzoek zijn gehanteerd: "anti-zionism", "negationism", "conspiracy" en "TTL" (Threat To Life). Via het dashboard kunnen gebruikers filteren op zowel de toxiciteitsschaal als de categorieën. Dankzij Image Captioning is **elke meme verrijkt met een beknopte beschrijving,** terwijl OCR eventuele **tekst in de afbeeldingen heeft geëxtraheerd** en geïdentificeerd of het schrift Latijns is. **Zowel de**

[55] https://mistral.ai/news/pixtral-12b

[56] AI-modellen die verschillende soorten content kunnen verwerken, zoals tekst en ook beeld: https://en.wikipedia.org/wiki/Multimodal_learning

[57] Zie paragraaf 4.4.2 voor meer informatie over Mistral AI.

[58] Zie paragraaf 4.4.2 waar wordt beschreven hoe Mistral's model Mixtral werd ingezet voor het genereren van synthetische tekstuele trainingsdata

Het dashboard biedt tevens de mogelijkheid om **naar vergelijkbare memes te zoeken en deze te clusteren via similarity search.** Hiervoor is niet Pixtral gebruikt, maar een combinatie van het multimodale model Blip59[60], de Python-bibliotheek *Sentence Transformers*[61] en Meta's Faiss[62], een techniek voor similarity search en contentclustering. De pipeline verloopt als volgt: Blip genereert beschrijvingen van de memes, Sentence Transformers zet deze om in numerieke representaties (vectoren), waarna Faiss berekent welke afbeeldingen gelijksoortig zijn.

Verder geeft het dashboard ook de mogelijkheid comments te doorzoeken (als deze aanwezig waren bij de memes) en te filteren op datum, platform, regio, taal en het aantal malen dat de meme voorkomt in de database[63]. Aan de hand van de laatste functionaliteit kan gezocht worden naar populaire of virale memes.

Figuur 30: Screenshot van similarity search voor antisemitische memes met de 'Happy Merchant' trope

[59] https://huggingface.co/docs/transformers/main/en/model_doc/blip

[60] We testten ook het Clip vision model en Meta's Dino vision model. Blip presteerde echter het best in het matchen van gelijkaardige afbeeldingen op zowel pixel als context niveau. Zie https://huggingface.co/docs/transformers/en/model_doc/clip en https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/

[61] https://sbert.net/examples/sentence_transformer/applications/image-search/README.html

[62] https://github.com/facebookresearch/faiss

[63] Duplicaten zijn verwijderd. Met deze functie kan gezocht worden naar memes die meer dan eens gepost zijn (populaire, virale of gespamde memes)

6. Topic Analyse

Topic analyse binnen datawetenschappen is een techniek die grote hoeveelheden ongestructureerde tekst analyseert om thematische patronen te ontdekken. Bij toepassing op social media data worden clusteringsalgoritmes ingezet die gerelateerde berichten groeperen op basis van woordfrequenties en -samenhang. Clusteringsalgoritmes voor tekst werken door tekstdocumenten eerst om te zetten in numerieke representaties (embeddings). Vervolgens groeperen algoritmes deze numerieke documenten op basis van hun onderlinge afstand of gelijkenis. **Het resultaat is een indeling waarbij tekstdocumenten met vergelijkbare inhoud in dezelfde cluster worden geplaatst, zonder dat menselijke annotatie nodig is**[64,65].

Vervolgens kunnen Large Language Models (LLMs) de inhoud van de berichten binnen deze clusters automatisch **samenvatten tot een kernboodschap per cluster.** Dit resulteert in beknopte inzichten zoals "alle berichten binnen deze cluster benoemen samenzweringen over George Soros" - zonder dat analisten duizenden individuele berichten hoeven te lezen en te labelen. Deze gecombineerde aanpak van clustering en LLM-summarisatie zorgt voor efficiënte extractie van waardevolle inzichten uit grote hoeveelheden social media data[66].

We pasten topic analyse toe om de verschillende antisemitische narratieven, tropen en onderwerpen binnen onze onderzoeksdata in kaart te brengen. Alle antisemitische berichten werden getransformeerd naar numerieke embeddings met het model multilingual-e5-large-instruct[67], dat we selecteerden vanwege zijn uitstekende prestaties op het MTEB (Massive Text Embedding Benchmark) leaderboard[68,69]. Voor de clustering gebruikten we de Python-bibliotheek BERTopic[70] met het HDBSCAN clusteringsalgoritme[71], waarbij we alleen clusters met minimaal 10 vergelijkbare berichten als relevant topic beschouwden.

Aangezien gangbare LLMs zoals OpenAI's ChatGPT en Google's Gemini toxische inhoud blokkeren vanwege moderatie-richtlijnen, kozen we het Mistral model Mistral-Nemo-Instruct-2407[72,73] om een korte samenvatting te schrijven van de berichten per cluster. Mistral-Nemo is compact genoeg voor lokaal gebruik, ondersteunt meerdere talen, kent minimale moderatiebeperkingen en levert toch nauwkeurige resultaten. We implementeerden dit model via Ollama[74], een open-source tool waarmee gebruikers LLMs lokaal kunnen draaien zonder externe API's of internetverbinding, wat extra privacy en controle biedt.

De door Mistral-Nemo gegenereerde topic-samenvattingen werden vervolgens doorgestuurd naar Anthropic's Claude[75] en Google's Gemini[76] aan de hand van de LiteLLM[77], met de opdracht een beknopte analyse te schrijven die alle antisemitische topics overzichtelijk presenteert. Hierbij vroegen we specifiek aandacht voor taalgebonden verschillen en topics die kenmerkend zijn voor de Nederlandse context. **Claude en Gemini kunnen geen rauwe dataset van antisemitische berichten verwerken vanwege moderatierichtlijnen, maar het model is wel in staat om geanonimiseerde samenvattingen van antisemitische thema's te analyseren en te integreren in een samenhangend overzichtsartikel dat de patronen en narratieven in kaart brengt.**

[64] https://www.ibm.com/think/topics/topic-modeling

[65] https://www.sciencedirect.com/science/article/pii/S1877050922010158

[66] https://en.wikipedia.org/wiki/Automatic_summarization

[67] https://huggingface.co/intfloat/multilingual-e5-large-instruct

[68] Open lijsten die aangeven hoe goed AI modellen scoren op bepaalde taken

[69] https://huggingface.co/spaces/mteb/leaderboard

[70] https://maartengr.github.io/BERTopic/index.html

[71] https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

[72] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

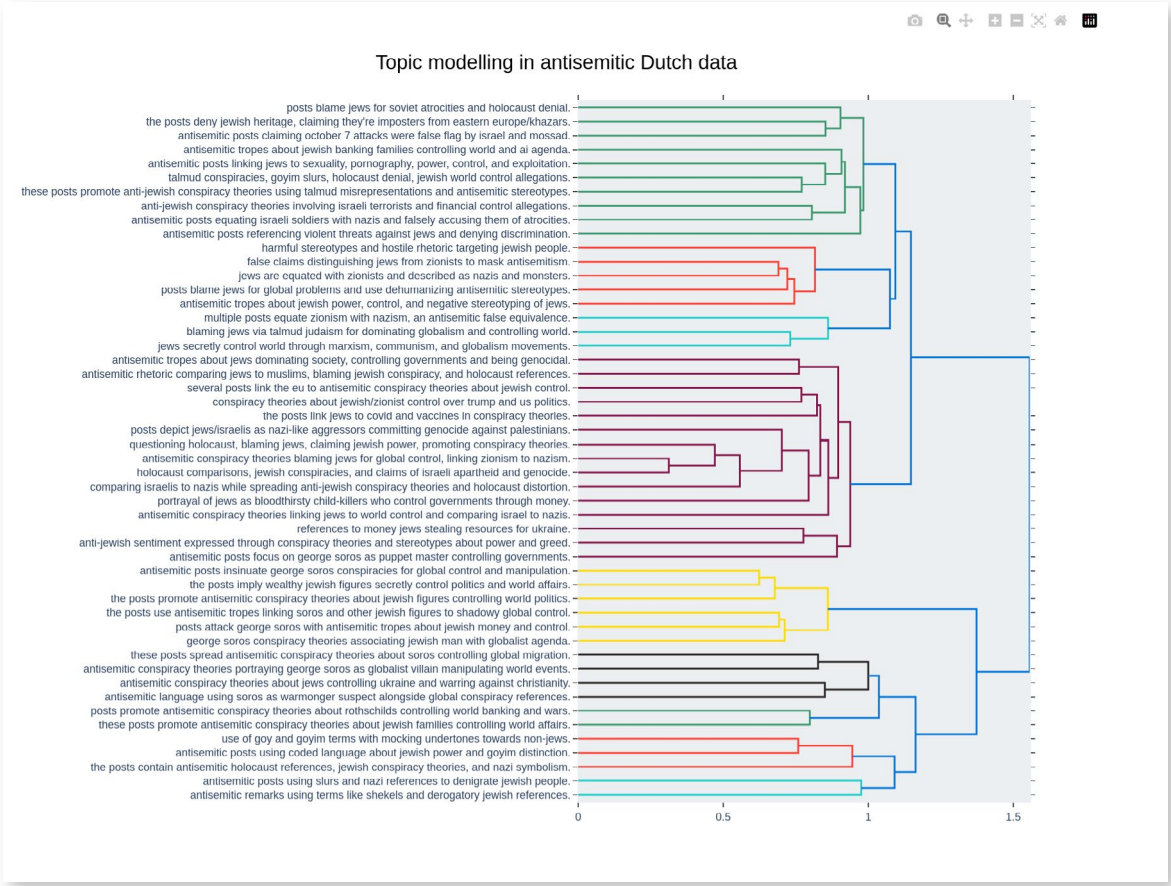
[73] Zie het tech report paragraaf 4.2.2 voor meer info over Mistral en Mistral modellen

[74] https://ollama.com/

[75] https://www.anthropic.com/claude

[76] Hiervoor gebruikten we gemini/gemini-2.5-pro: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-pro

[77] Aan de hand van de Python library LiteLLM: https://www.litellm.ai/



Figuur 31: Visualisatie van geclusterde berichten en terugkerende topics in de onderzoeksdata

7. Tijdslijn-analyse

Naast de inhoudelijke analyses van thema's, narratieven, netwerken en memes is het relevant om te onderzoeken hoe antisemitisme zich over de tijd heeft ontwikkeld. Dit hoofdstuk beschrijft de methode waarmee de tijdslijn-analyse is uitgevoerd en hoe de resultaten worden gepresenteerd.

7.1 Methode

Voor de tijdslijn-analyse wordt gebruik gemaakt van PELT (Pruned Exact Linear Time)[78] changepoint detection[79], een statistische methode die een tijdreeks opsplitst in perioden waarbinnen het signaal (het percentage antisemitische berichten in dit geval) relatief stabiel is, en de grenzen identificeert waar het signaal significant van karakter verandert. Concreet wordt per platform de dagelijkse verhouding van antisemitische berichten ten opzichte van het totale aantal verzamelde berichten berekend. PELT analyseert deze tijdreeks en deelt deze automatisch op in statistisch te onderscheiden tijdssegmenten. Op die manier kunnen onder andere korte, intense pieken van antisemitische activiteit worden onderscheiden van langere perioden met een structureel lager niveau[80].

7.1 Signaal constructie

Als invoer voor de tijdslijnanalyse wordt per platform voor elke dag een dagelijkse ratio berekend: het aantal berichten gemarkeerd als antisemitisch gedeeld door het totale aantal verzamelde berichten op die dag, uitgedrukt als percentage, oftewel het signaal. Door te werken met een verhouding in plaats van absolute aantallen wordt gecorrigeerd voor variaties in het dagelijkse verzamelde volume. Smoothing[81] werd toegepast op de ratiowaarden door voor elke dag het gemiddelde te nemen van die dag en de twee omliggende dagen (een schuivend venster van 3 dagen). Dit dempt incidentele eendagsschommelingen (bijvoorbeeld veroorzaakt door een tijdelijk lage dataverzameling) zonder structurele verschuivingen in het signaal te vertekenen.

7.2 Penalties

Een belangrijk onderdeel van PELT is de zogenoemde penalty-parameter. Deze parameter bepaalt hoe gevoelig het algoritme is voor het detecteren van nieuwe changepoints: een hoge penalty resulteert in minder, maar bredere tijdssegmenten, een lage penalty in meer en meer fijnmazige tijdssegmenten. De schatting combineert twee statistische maten in een gewogen verhouding: 75% van de waarde is gebaseerd op de Median Absolute Deviation (MAD)[82] en 25% op de logaritmisch getransformeerde standaarddeviatie[83]. De MAD is robuust tegen uitschieters en vormt daarmee de kern van de schatting; de standaarddeviatie voegt gevoeligheid toe voor bredere spreiding in het signaal.

Deze basis-penalty wordt vervolgens bijgesteld aan de hand van een volatiliteitscorrectie; platformen waarvan de dagelijkse ratio sterk schommelt, krijgen automatisch een hogere penalty. Dit voorkomt dat het algoritme bij inherent volatielere signalen te veel segmenten aanmaakt op basis van normale variatie in de data.

7.3 Filtering en classificatie van tijdssegmenten

Niet elk segment dat PELT produceert is inhoudelijk even relevant; sommige segmenten weerspiegelen slechts kleine, toevallige schommelingen. Daarom worden de ruwe segmenten in een *post-processing* stap gefilterd. Om te worden behouden moet een segment voldoen aan minimale drempelwaarden voor ten minste één van de volgende criteria: de **piekintensiteit** (de hoogste dagelijkse ratio binnen het segment), de **gemiddelde intensiteit** (het gemiddelde van alle dagelijkse ratio's over de duur van het segment), of het **berichtvolume** (het absolute aantal antisemitische berichten binnen het segment). Een segment kan ook worden behouden op basis van een combinatie van deze criteria wanneer het op geen enkel individueel criterium de drempel haalt, maar op meerdere criteria dicht bij de drempel zit.

[78] https://emmorts.github.io/SignalSharp/docs/detection/pelt.html

[79] https://en.wikipedia.org/wiki/Change_detection

[80] De methode werd geïmplementeerd aan de hand van Python library Ruptures: https://centre-borelli.github.io/ruptures-docs/

[81] https://en.wikipedia.org/wiki/Smoothing

[82] https://en.wikipedia.org/wiki/Median_absolute_deviation

[83] https://nl.wikipedia.org/wiki/Standaardafwijking

Elk segment dat dit filter doorstaat, krijgt een **significantie-score.** Deze score combineert intensiteit en volume: een segment scoort hoger naarmate het zowel een hoog aandeel antisemitische berichten als een groot absoluut aantal berichten kent, waarbij kortere perioden van hoge intensiteit zwaarder wegen dan lange perioden van lage activiteit met hetzelfde totale volume. **De score geeft daarmee aan hoe significant een tijdsperiode was,** oftewel hoeveel online commotie een specifieke periode of gebeurtenis teweegbracht. Aangrenzende segmenten met een lage score worden samengevoegd om fragmentatie te beperken.

Op basis van duur en intensiteit worden segmenten ingedeeld in vier typen:

- **Major event** (≤ 20 dagen, hoge intensiteit): een kortdurende, scherpe piek, doorgaans gerelateerd aan een concrete aanleiding of gebeurtenis.
- **Developing story** (21–60 dagen): een periode van geleidelijk oplopende of aanhoudende verhoogde activiteit rond een zich ontwikkelend thema.
- **Sustained coverage** (61–365 dagen): een langere periode met structureel verhoogde activiteit.
- **Long-term theme** (> 365 dagen): een breed, persistent thema dat de gehele onderzoeksperiode of een groot deel daarvan doorkruist.

7.4 Inhoud van tijdssegmenten

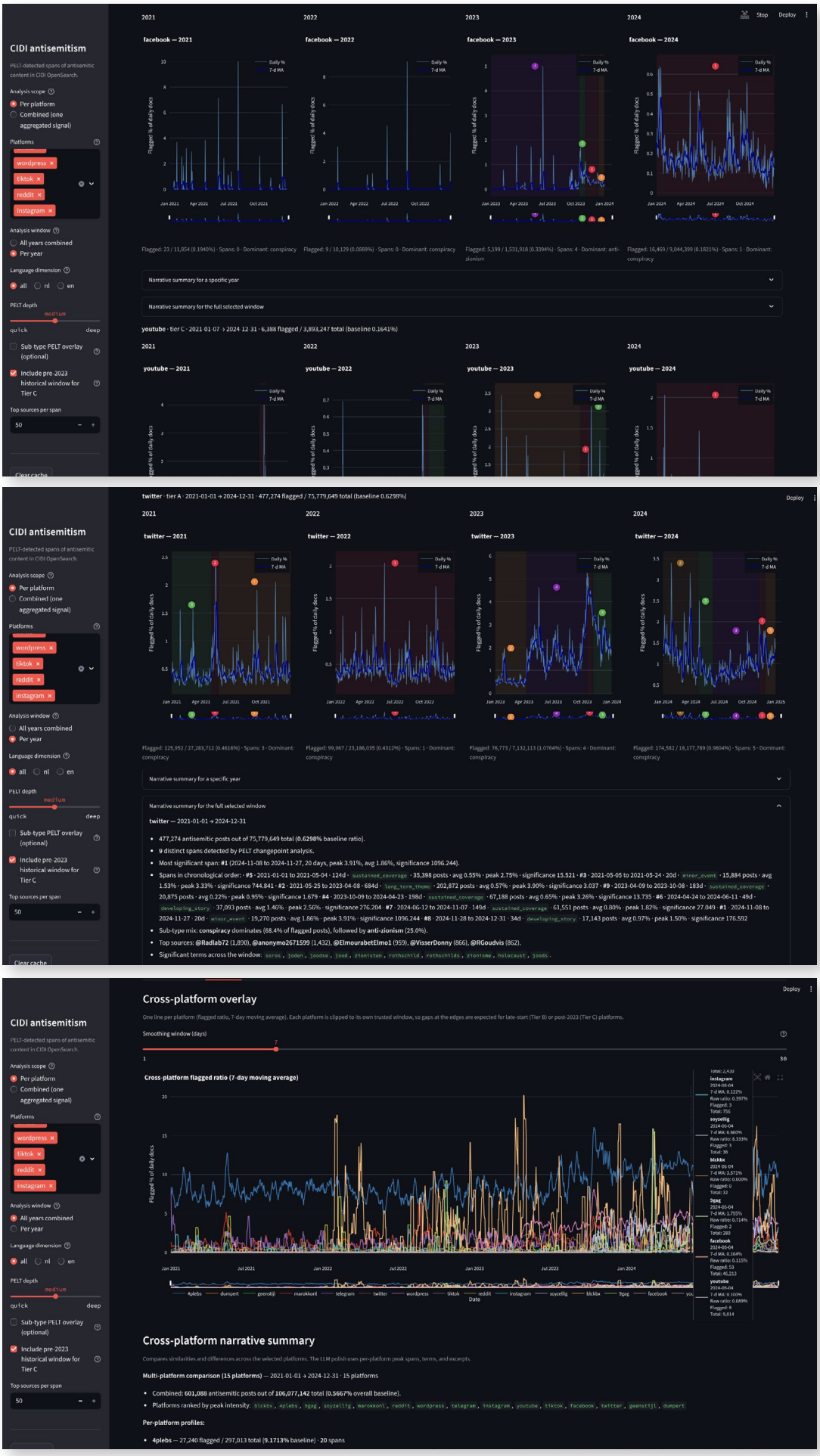
Voor elk gedetecteerd tijdssegment wordt bepaald welke antisemitische subtypes daarbinnen dominant zijn. Dit gebeurt door per sublabel (conspiracy, negationisme, TTL, anti-zionisme, VRWE en 4chan) de dagelijkse aantallen gemarkeerde berichten binnen het datumbereik van het tijdssegment op te tellen en dit te delen door het totale aantal gemarkeerde berichten in diezelfde periode. Omdat de sublabels niet-exclusief zijn (een bericht kan meerdere labels tegelijk dragen) kunnen de aandelen per segment samen meer dan 100% bedragen.

Elk tijdssegment wordt daarnaast automatisch voorzien van de voornaamste bijdragende bronnen en karakteristieke termen die statistisch oververtegenwoordigd zijn in de berichten binnen dat segment. Ook wordt een narratieve samenvatting gegenereerd door een taalmodel (LLM)[84], op basis van de meest representatieve berichten per segment. Dit transformeert de statistische segmentatie tot inhoudelijk interpreteerbare episodes: niet alleen wanneer antisemitische activiteit piekte, maar ook welk type antisemitisme domineerde, welke actoren een rol speelden en welke taal het discours kenmerkte.

7.5 Dashboard

We visualiseerden de resultaten in een interactief dashboard, waar tijdsdynamieken per jaar, per platform en geaggregeerd in kaart gebracht kon worden. Het dashboard gaf ook inzicht in de meest voorkomende antisemitische keywords, topics en narratieven en meest relevante comments per tijdssegment.

[84] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro



Figuur 32: Screenshots van dashboard voor tijdslijn-analyse

8. Conclusie

In dit rapport werd een technisch en methodologisch onderbouwd raamwerk gepresenteerd voor de detectie en analyse van antisemitische content binnen Nederlandse sociale media. Door een combinatie van lexicon-gebaseerde toxiciteitsdetectie, een specifiek voor antisemitisme getraind BERT-classificatiemodel en een geavanceerde annotatie-infrastructuur met multilinguale sublabels, werd een fijnmazig systeem ontwikkeld dat zowel de context als de variatie in antisemitische uitingen kan interpreteren.

De testresultaten van het BERT-classificatiemodel tonen aan dat het systeem in staat is om met hoge precisie (0,90–0,95) en recall (0,80–0,97) onderscheid te maken tussen neutrale en antisemitische content en subcategorieën van antisemitisme. Ook presteert het model accuraat in complexe contexten zoals ironie en quotes.

De analyse werd verder ondersteund door aanvullende componenten zoals een doorzoekbare meme-database met OCR, image captioning en object recognition, alsook een thematische topic analyse aan de hand van een clusteringsalgoritme en LLM-samenvattingen.

Deze infrastructuur maakt het mogelijk om met hoge nauwkeurigheid onderscheid te maken tussen neutrale, toxische en antisemitische uitingen, inclusief complexere vormen van antisemitisme zoals negationisme, VRWE-jargon en verhuld antisemitisme onder de noemer van anti-zionisme. De combinatie van explainable AI-methodes en privacybewuste lokale uitvoering onderstreept de inzet op transparantie, ethiek en bruikbaarheid voor maatschappelijke actoren in een Nederlandse context.

Colofon

Titel
Antisemitisme in het Nederlandse sociale media landscap:
Technisch rapport

Publicatiedatum
Mei 2026

Opdrachtgever
Centrum Informatie
en Documentatie Israël (CIDI)
in casu Stichting Informatie
en Meldpunt Antisemitisme
(meldantisemitisme.nl)

Uitvoering en onderzoek
Textgain

Financiële ondersteuning
De onderzoeken zijn mogelijk gemaakt door financiële ondersteuning van de Nationaal Coördinator Antisemitismebestrijding (NCAB).

Methodologie
Dit rapport is gebaseerd op de analyse van publiek toegankelijke online bronnen, waaronder sociale mediaplatformen, discussiefora en andere openbare digitale omgevingen. Voor de detectie en analyse van online antisemitische uitingen werd een combinatie ingezet van AI- en NLP-technieken, meertalige lexicons, contextuele analyse en menselijke validatie, gericht op het in kaart brengen van trends, narratieven, verspreidingsdynamieken en taalgebruik binnen online discussies.

Juridische en ethische disclaimer
De analyses in dit rapport zijn gebaseerd op geautomatiseerde detectie en interpretatie van publieke online content. Hoewel maximale zorgvuldigheid is betracht, kunnen classificaties en interpretaties onderhevig zijn aan contextuele beperkingen. Het rapport heeft als doel maatschappelijke trends en risico's inzichtelijk te maken en vormt geen juridische kwalificatie van individuele uitingen of personen. Persoonsgegevens worden waar mogelijk geaggregeerd, gepseudonimiseerd of geanonimiseerd verwerkt conform de AVG/GDPR en relevante Europese regelgeving.

Vormgeving
Luna 3

Contact
cidi.nl of meldantisemitisme.nl
textgain.com

TECHNISCH RAPPORT

Online antisemitisme: hoe staat het ervoor?

Sociale media zijn uitgegroeid tot een belangrijke plek waar antisemitische denkbeelden, complottheorieën en haat zich verspreiden. Om een volledig beeld te krijgen van antisemitisme in Nederland liet CIDI, samen met het Meldpunt Antisemitisme en de NCAB, onderzoek uitvoeren door Textgain naar het Nederlandse online landschap.

Deze rapportages brengen de belangrijkste trends, narratieven en ontwikkelingen op Nederlandse sociale media van 2021 tot 2024 in kaart.

Het onderzoek vormt onderdeel van een monitor die de ontwikkeling van online antisemitisme door de jaren heen inzichtelijk moet maken.

Alleen door online en offline ontwikkelingen structureel te volgen, kunnen we antisemitisme beter begrijpen, herkennen en bestrijden.

CIDI meld. antisemitisme

textgain

Nationaal Coördinator Antisemitismebestrijding
Ministerie van Justitie en Veiligheid