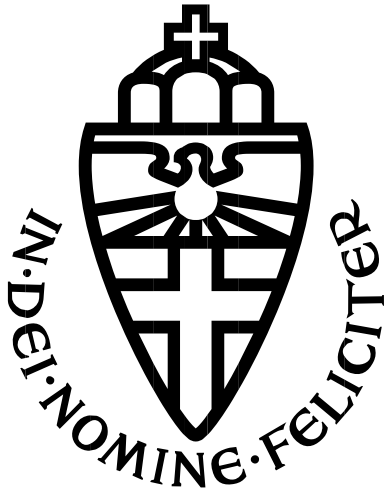
Radboud University Nijmegen



De automatische herkenning van drogredenen met GPT-3

Bachelor Thesis Taalwetenschap

Auteur:

Noa van Helleman, s1010462

Primair begeleider:

Dr. B.J.M. van Halteren[1,2]

[1] Centre for Language Studies

[2] Departement Taal en communicatie

Radboud University

22 Juli 2022

Abstract

In dit onderzoek is gekeken naar de mate waarin het taalmodel GPT-3 drogredenen kan herkennen en classificeren. Om dit te onderzoeken is een classificatietaak uitgevoerd waarbij dit geavanceerde taalmodel drogredenen in zinnen moest onderscheiden. De data bevatte 441 zinnen met 9 verschillende types drogredenen en 30 zonder drogreden. Deze zinnen zijn aangeboden aan GPT-3 met als taak classificatie van de zinnen op drogreden. Bij deze classificatie zijn de aantallen echt positieven, foutpositieven, echt negatieven en foutnegatieven geteld. Van deze waarden is een confusion matrix gemaakt en zijn de precision, recall en F1-score berekend. Hieruit bleek dat de herkenning en classificatie van geselecteerde drogredenen tot op zekere hoogte goed verloopt. De drogredenen ad hominem, appeal to authority, bandwagon en false dilemma behalen een F1-score van 0,74 of hoger. De prestaties van het taalmodel bij andere drogredenen vallen echter tegen. Dit wijst erop dat het mogelijk is met GPT-3 drogredenen te herkennen en classificeren, maar dat de mate waarmee dit gaat verschilt per drogreden.

Inhoudsopgave

1. Inleiding

Rede is een belangrijk onderdeel van menselijke communicatie. Het is de mogelijkheid om bewust logica toe te passen bij het trekken van conclusies uit informatie, met het doel om de waarheid te vinden (Proudfoot, 2010). In menselijke redeneringen komen echter fouten voor; drogredenen. Drogredenen zijn fouten in een argumentatie. Zo kan er een keuze tussen twee opties worden voorgelegd. Als het de ene optie niet is, dan moet het de andere optie wel zijn. In werkelijkheid zijn er echter meer keuzes. Dit wordt ook wel een false dilemma genoemd. Een andere mogelijkheid is dat een argument ingaat op de tegenstander in plaats van diens argument. Dit staat ook wel bekend als een ad hominem of persoonlijke aanval. De zin "*Je bent veel te jong om dit te begrijpen!*" is een voorbeeld van zo'n persoonlijke aanval. De persoon wordt in deze zin aangevallen, terwijl dit niets te maken heeft met een eventueel standpunt dat hij heeft gemaakt. Jong zijn betekent niet dat iemand niets weet over een discussiepunt.
Als mens is een fout in een argumentatie vrij gemakkelijk te herkennen als je weet waar je naar kijkt en je er bewust van bent, maar met de recente toename in de populariteit van taalmodellen komt de vraag naar boven: Kunnen taalmodellen dit ook?

1.1 Overzicht van de GPT modellen

De Generative Pre-Trained Transformer (GPT) modellen zijn natural language processing (NLP) taalmodellen ontwikkeld door OpenAI. De ontwikkeling van het eerste GPT model werd gedaan op een nieuwe manier (Radford et al., 2018). De meeste en beste NLP modellen tot 2018 werden getraind op specifieke taken door middel van supervised learning. Dit houdt in dat het systeem wordt getraind op gelabelde data totdat het de relatie tussen data en label heeft geleerd. Bij GPT-1 werd er echter gebruik gemaakt van een generatief taalmodel. Dit is getraind op niet gelabelde data, waarna het model werd gefinetuned door het aanleveren van voorbeelden. Op deze manier werden twee beperkingen van supervised models ontweken, namelijk de behoefte voor grote hoeveelheden specifieke trainingsdata en het feit dat supervised models niet voor andere taken kunnen worden ingezet dan de taak waar ze voor gebouwd zijn. De eerste versie van het GPT model kwam uit in 2018 en werd getraind op de dataset BooksCorpus met 7000 ongepubliceerde boeken. Dit leverde een model op met 117 miljoen parameters in totaal. De prestaties van GPT-1 waren beter dan de eerder gebruikte supervised modellen bij 9 van 12 tests. GPT-1 behaalde ook redelijke scores op zero-shot performance taken. Radford et al. beweren dat GPT-1 heeft bewezen dat deze alternatieve wijze van het opzetten van een model een positief effect had op de generaliseerbaarheid en prestaties van taalmodellen.

Een jaar later, in 2019, werd GPT-2 uitgebracht (Radford et al, 2019). Het idee hiervan was grotendeels hetzelfde als bij GPT-1, alleen werd het uitgevoerd op een grotere schaal. Het gebruik van een grotere dataset en meer parameters leidde tot een beter presterend taalmodel. Nadruk werd hier gelegd op zero-shot task transfer en zero-shot learning; dit houdt in dat de taak begrepen wordt, en dat de taak kan worden uitgevoerd zonder het aanleveren van voorbeelden. De schaal van training ging van 7000 boeken met 117 miljoen parameters, naar de WebText dataset met 40GB aan tekstdata uit meer dan 8 miljoen documenten met 1,5 miljard parameters. Een vergelijking met modellen met minder parameters is uitgevoerd, waaruit bleek dat meer parameters zorgden voor een beter begrip en betere prestaties op taken als tekstbegrip, samenvatten, vertalen en het beantwoorden van vragen. GPT-2 verbeterde 7 van de 8 beste modellen bij language modeling, ook wel het voorspellen van het volgende woord, woordherkenning en leesbegrip. Bij vertalen en samenvatten lieten de resultaten zien dat het model een redelijk niveau had, maar niet mee kon met de beste, specifiek getrainde modellen. Het belangrijkste gegeven uit de resultaten was echter het feit dat trainen op een grotere dataset en het hebben van meer parameters de prestaties van het taalmodel in zulke mate verbeterden dat deze beter waren dan de resultaten van verschillende taak specifieke modellen. Ook bleek uit de analyse dat grotere modellen nog betere resultaten konden opleveren.

In 2020 bracht OpenAI de nieuwste versie van het taalmodel uit, namelijk GPT-3 (Brown et al., 2020). Met het doel om een taalmodel te creëren dat geen fine-tuning nodig heeft en taken kan begrijpen zonder expliciete training of voorbeelden, werd de grootte van het model verder opgeschaald vergeleken met

GPT-2. Het trainingsmateriaal werd uitgebreid naar documenten uit vijf verschillende corpora, ieder met zijn eigen gewicht. Deze corpora waren Common Crawl, WebText2, Books1, Books2 en Wikipedia, waar Common Crawl het grootste van is en het zwaarste gewicht had. Hieruit volgde een model met 175 miljard parameters; meer dan 100 keer het aantal parameters van GPT-2. Naast het voorspellen van het volgende woord gebaseerd op context, werd ook patroonherkenning toegepast. De grootste verbetering van GPT-3 boven zijn voorgangers en concurrerende modellen, was de prestatie in een zero-shot setting. GPT-3 werd getest op meerdere language modeling datasets en gaf betere resultaten in de zero-shot setting dan de best presterende modellen tot dan. De resultaten van GPT-3 verbeterden in een one- en few-shot setting. Er zitten echter ook nadelen aan het GPT model. Zo heeft het systeem moeite met het produceren van begrijpelijke, lange zinnen en heeft het de neiging stukken tekst te herhalen. Ook loopt het soms tegen problemen aan als het aankomt op implicaties of tekstbegrip. Een meer technisch aspect wat kan zorgen voor problemen zijn de grootte en complexiteit van het model. Hierdoor kan het lastig zijn te interpreteren hoe de resultaten gegenereerd zijn. Daarbovenop komt het mogelijke gevaar van training bias en misbruik. Door de data waar het op is getraind zou het kunnen dat GPT-3 een bias heeft. Ook zou de tekstgeneratie gebruikt kunnen worden voor praktijken zoals phishing, spamming en het verspreiden van desinformatie (Brown et al., 2020).

1.2 Verwante literatuur

In 2020 hebben Samghabadi et al. onderzoek gedaan naar de herkenning van agressie en seksisme met een model gebaseerd op het taalmodel BERT (Devlin et al., 2018). De data voor dit onderzoek bestaat uit geclassificeerde teksten uit drie talen; het Engels, het Hindi en het Bengali. Voor taak A, het herkennen van agressie, waren de teksten verdeeld over drie labels: Not Agressive, Covertly Agressive en Overtly Agressive. Voor taak B, het herkennen van seksisme, waren de teksten verdeeld over twee labels: Gendered en Non-gendered. Eerst wordt door BERT contextinformatie opgehaald. Daarna wordt de belangrijkheid van elk woord bepaald. Een vector met deze informatie wordt dan geclassificeerd door het model om het agressie en seksisme gehalte te voorspellen. F1-scores (zie 2.5 voor uitleg) bij taak A lagen rond de 0,72, waar de scores bij taak B boven de 0.8 lagen. Samghabadi et al. (2020) geven aan dat het verschil in de scores bij taak A en B mogelijk kan komen door de binaire aard van de classificatie.

Ook Chiu et al. (2022) hebben gekeken naar het vraagstuk of taalmodellen haatdragende taal kunnen herkennen en teksten kunnen classificeren als seksistisch of racistisch. Om dit te onderzoeken hebben ze gebruik gemaakt van het taalmodel GPT-3. Een dataset van opmerkingen van YouTube en Reddit is beoordeeld door het systeem. Deze data bevatte seksistische, racistische en neutrale opmerkingen die zijn voorgelegd aan het taalmodel om te classificeren. Dit is uitgevoerd in drie verschillende settingen; een no-shot setting, een one-shot setting en een few-shot setting. Het verschil tussen deze settingen is de hoeveelheid voorbeelden die het model krijgt vóór het uitvoeren van de taak. Een no-shot setting krijgt alleen de taak aangeboden zonder voorbeeld van hoe de taak moet worden uitgevoerd. In een one-shot setting wordt de taak gegeven met één voorbeeld. Een few-shot setting geeft de taak en meerdere voorbeelden. Hieruit volgde dat de F1-scores van de no-shot, one-shot en few-shot varianten respectievelijk 0,70, 0,55 en 0,62 waren als de categorieën apart werden beoordeeld. De few-shot setting is ook uitgevoerd met gemengde categorieën. Dit bracht de F1-score naar een waarde van 0,78. Hieruit trekken zij de conclusie dat, met de juiste instellingen en voorbeelden, taalmodellen zoals GPT-3 seksisme en racisme kunnen herkennen. Ze vermelden hier echter ook bij dat er een paar mogelijke problemen kunnen optreden. Zo zouden spelling en snel veranderende taalgebruiken invloed kunnen hebben op de prestaties van zulke systemen.

Verder hebben Nakpih en Santini (2020) onderzocht of het mogelijk was om non sequitur, ook bekend als onjuist gevolg, drogredenen te herkennen in juridische argumentaties. Tot dit doel hebben zij een model opgezet wat de verschillende onderdelen van deze argumentaties doorloopt. Ze maken gebruik van de standaard structuur van de non sequitur fallacy en laten alternatieve en complexere vormen buiten beschouwing. Hieruit volgde dat het nagaan van vier onderdelen van een claim het probleem van drogredenen oploste. Deze onderdelen waren de validiteit, degelijkheid, noodzaak en de toereikendheid

van de argumentatie. Het model liep de geformaliseerde teksten door en keek naar de logische argumentatiestructuren die hierin werden gebruikt. Het bleek dat de drogreden kon worden herkend door het nalopen van de vier onderdelen. Wel zorgen de regels en gebruiken van het opstellen van juridische teksten voor een unieke situatie, waarin de teksten bepaalde standaarden volgden. Nakpih en Santini (2020) geven aan dat dit invloed heeft op de generaliseerbaarheid van het onderzoek.

Ook Jin et al. (2022) hebben onderzoek gedaan naar de herkenning van drogredenen. Dit hebben ze gedaan door middel van het opstellen van een dataset met verschillende types drogredenen. Deze dataset hebben ze voorgelegd aan een collectie van geavanceerde taalmodellen. De dataset bevatte een lijst met bijna 2500 zinnen, waarin 13 verschillende drogredenen voorkwamen. Er is gekeken naar de prestaties van verschillende types modellen in dit onderzoek: zero-shot modellen (zoals GPT-2 en RoBERTa-MNLI) en finetuned modellen (zoals BERT, DeBERTa en Electra). De resultaten lieten zien dat de classificatie van drogredenen door deze taalmodellen lastig is; de prestaties lagen tussen de 0.09 en 0.53 bij de F1 scores, waar de hoogste scores werden behaald door de gefinetunede modellen en de maximum score van de zero-shot modellen werd gehaald door GPT-2 met een score van nog geen 0.14. Daarnaast stellen ze een model voor dat kijkt naar de logische structuur van de tekst. Hieruit bleek dat dit model een waarde van 0.59 behaalde bij de F1 score, met de beste scores bij ad hominem en bandwagon. Dit was 5,46% hoger dan het best presterende model uit de eerste test. Jin et al. (2022) geven aan dat een mogelijke verklaring voor de scores bij deze twee drogredenen liggen aan het woordgebruik. Zinnen waar termen in zitten die verwijzen naar een meerderheid van mensen kunnen gemakkelijker worden herkend als zijnde een bandwagon. Hetzelfde geldt voor ad hominem met zinnen waar beledigingen of stukken die de betrouwbaarheid van de spreker aantasten in zitten.

1.3 Probleemstelling en onderzoeksopzet

Er wordt dus volop onderzoek gedaan naar de mogelijkheden van geavanceerde taalmodellen. De prestaties van deze modellen wisselen echter nogal per onderzoek. De resultaten van zowel Samghabadi et al. (2020) als Chiu et al. (2022) laten zien dat de herkenning van agressie, haat en seksisme tot op zekere hoogte lukt. Beide komen uit op maximum F1-scores van rond de 0,8. De onderzoeken van Nakpih en Santini (2020) en Jin et al. (2022) lijken nog een stap verder te gaan; ze testten de mogelijkheid van taalmodellen om de structuur achter taal mee te nemen in het maken van een classificatie en testten zo de herkenning van drogredenen door deze modellen. De prestaties van de modellen in het onderzoek van Jin et al. (2022) laten echter nogal wat te wensen over. De F1-score van het GPT-2 model bij de classificatie van drogredenen komt niet boven de 0.14 uit. Gefinetunede modellen presteerden beter, maar kwamen nog steeds slechts uit op een maximum F1-score van 0.53.

Sinds de vrijgave van GPT-3 is er veel aandacht voor de mogelijkheden van dit taalmodel. Het ene na het andere artikel verschijnt over het kunnen en de potentiële gevaren van dit model. Naar aanleiding van het onderzoek van Jin et al. (2022), waarin een varia aan taalmodellen is getest op de classificatie van drogredenen, is dit onderzoek specifiek gericht op de vraag *in welke mate het taalmodel GPT-3 drogredenen kan herkennen en classificeren.* Dit wordt gedaan aan de hand van een classificatie-taak, waarbij zinnen met verschillende drogredenen worden voorgelegd aan het model met de taak om de drogreden te herkennen en classificeren.

De verwachting aan de hand van het onderzoek van Jin et al. (2022) is dat de prestaties van het model verschillen per drogreden. Zo wordt er verwacht op basis van de resultaten van Samghabadi et al (2020) en Chiu et al. (2022), die lieten zien dat de herkenning van haatdragende tekst tot op zekere hoogte goed liep, dat de ad hominem drogreden met redelijke scores herkend kan worden. Ook wordt er verwacht dat de bandwagon drogreden relatief goed herkend kan worden op basis van Jin et al. (2022). De redeneringen van deze conclusies volgend wordt er verder verwacht dat appeal to authority en false dilemma ook een relatief goede score zullen halen bij de herkenning en classificatie, aangezien beide drogredenen mogelijk herkenbare delen bevatten. Bij appeal to authority zou dit de vorm van ‘volgens (autoriteit)’ kunnen aannemen. Bij false dilemma kan dit een ‘dan wel x, dan wel y’ constructie volgen.

Dit wordt onderzocht aan de hand van een classificatietaak. Hierbij worden zinnen met drogredenen voorgelegd aan GPT-3 via de OpenAI API. De taak voor GPT-3 is het toewijzen van een type drogreden, of aangeven dat er geen drogreden voorkomt. Dit wordt in meer detail beschreven in hoofdstuk 2. Hoofdstuk 3 gaat vervolgens in op de resultaten die de taak opleverde en de fouten die hierbij werden gemaakt. Als laatste worden in hoofdstuk 4 de conclusie en discussie gepresenteerd.

2. Methode

2.1 Materiaal

Om te bepalen of het taalmodel drogredenen kan herkennen is een classificatietaak voorgelegd aan GPT-3. Deze taak bestond uit 471 zinnen die een of een drogreden of geen drogreden bevatten. Elke zin was geclassificeerd met een specifieke drogreden. De drogredenen die zijn gebruikt zijn terug te vinden op de volgende pagina in Tabel 1 met een uitleg en een voorbeeld. De namen van de drogredenen en de voorbeelden zijn afkomstig uit de gebruikte dataset. Aangezien GPT-3 voornamelijk is getraind op Engelse teksten, is voor de data gekozen voor Engelse teksten.

De dataset is opgesteld uit handmatig verzamelde drogredenen van websites met een informerend en uitleggend doel over drogredenen, gecombineerd met geselecteerde voorbeelden die studenten gevonden hebben tijdens een opdracht in de cursussen Informationwetenschap en Information Science aan de Radboud Universiteit. Dit is aangevuld met drogredenen uit een externe dataset. Deze dataset was de LOGIC dataset, samengesteld door Jin et al. (2022). De dataselectie voor de zinnen van de studenten Informatiewetenschap en Information Science, en de LOGIC dataset was nodig om de kwaliteit van de dataset te waarborgen. Tijdens de selectie is het voorkomen van sterk op elkaar lijkende en identieke voorbeelden zoveel mogelijk beperkt en is er gelet op de classificatie van de drogreden. Dit laatste gold voornamelijk voor de data van de studenten; de data bevatte voorbeelden van drogredenen die door de studenten zelf waren geclassificeerd. Ook bij de LOGIC dataset is er gelet op deze punten. Met name het herhaaldelijk voorkomen van (varianten van) dezelfde zin was in deze dataset een punt van aandacht. De zinnen zonder drogreden zijn afkomstig uit de FEVER dataset van Thorne et al. (2018). Deze dataset is ontworpen voor feitenextractie en verificatie.

Hieruit volgde de volledige dataset met 471 zinnen (Van Helleman, 2022). Van de 471 zinnen bevatten er 87 een ad hominem, 50 een appeal to authority, 30 een appeal to emotion, 50 een bandwagon, 30 een circular reasoning, 50 een false cause, 65 een false dilemma, 49 een faulty generalization, 30 een straw man drogreden en 30 zinnen bevatten geen drogreden. De drogredenen zijn deels geselecteerd op herkenbaarheid. Vanuit de hypothese werd verwacht dat ad hominem, appeal to authority, bandwagon en false dilemma herkend zouden moeten worden. Dit is toen aangevuld met drogredenen waar genoeg voorbeelden bij gevonden konden worden. Uiteindelijk bestond de dataset uit zinnen met 9 verschillende drogredenen met ieder minimaal 30 voorbeelden. Het toevoegen van de zinnen zonder drogreden is gedaan om te testen of het model de neiging heeft om een drogreden toe te wijzen, zelfs als deze niet in de tekst zit.

Tabel 1
Drogredenen inclusief uitleg en voorbeeld

Naam	Uitleg	Voorbeeld
Ad hominem	Aanvallen van de spreker in plaats van het standpunt te beargumenteren.	Socrates' arguments about human excellence are rubbish. What could a man as ugly as he know about human excellence.
Appeal to authority	(Foutief) aannemen van een standpunt op basis van de positie of autoriteit van een persoon.	Richard Dawkins, an evolutionary biologist and perhaps the foremost expert in the field, says that evolution is true. Therefore, it's true.
Appeal to emotion	Manipuleren van emotie in plaats van een standpunt beargumenteren.	It is an outrage that the school wants to remove the vending machines. This is taking our freedom away!
Bandwagon	(Foutief) aannemen van een standpunt op basis van de mening van een grote groep/de meerderheid.	Marie notices that many of her friends have started eating a low-carb diet and drinking protein shakes. Marie decides that this must be the healthy way to eat, so she joins them.
Circular reasoning	Redenering beginnen met de conclusie; aannemen als waar zijnde van hetgeen bewezen moet worden.	The word of Zorbo the Great is flawless and perfect. We know this because it says so in The Great and Infallible Book of Zorbo's Best and Most Truest Things that are Definitely True and Should Not Ever Be Questioned.
False cause	Aannemen dat een (schijnbare) relatie tussen zaken betekent dat één de oorzaak van de ander is.	A rainbow appeared in the sky the day before basketball try-outs, and I made the team! If there's a rainbow when I try out in the spring, I'm sure I'll make the team again.
False dilemma	Beweren dat er slechts twee mogelijkheden zijn, waar er meer zijn.	Whilst rallying support for his plan to fundamentally undermine citizens' rights, the Supreme Leader told the people they were either on his side, or they were on the side of the enemy.
Faulty generalization	(Foutief) aannemen van een standpunt op basis van te weinig/niet-representatieve data.	A driver with a New York license plate cuts you off in traffic. You decide that all New York drivers are terrible drivers.
Straw man	Het argument van een ander verdraaien om het gemakkelijker te kunnen weerleggen.	Mayor Kerr wants to create more bicycle lanes in Greenpoint. Why is he forcing us to give up our cars and bike everywhere?

Noot. Uitleg gebaseerd op (The School of Thought, z.d.) en (Williamson, 2015)

2.2 GPT-3 online platform

Voor het onderzoeken van de mogelijkheden van GPT-3 wat betreft de herkenning van drogredenen is gebruik gemaakt van de OpenAI API (OpenAI, 2020a). Dit is een online platform wat, zoals OpenAI adverteert, voor vrijwel alle taken kan worden gebruikt die betrekking hebben tot het begrijpen of genereren van natuurlijke taal of code. Via de API is toegang mogelijk tot verschillende taal- en codemodellen, waaronder verschillende varianten van het GPT-3 model. Van deze varianten wordt het krachtigste model gebruikt; het Davinci model. Dit model wordt omschreven als de meest capabele als het gaat om complexe taken zoals samenvatten, creatieve tekstgeneratie en taken waarvoor begrip en logica een rol spelen. Ook wordt dit model aangeraden als het gaat om het uitproberen en verkennen van de mogelijkheden. De nieuwste versie hiervan is *text-davinci-002* en is getraind op data tot Juni 2021.

Toegang tot de API kan worden verkregen door het aanmaken van een account. Voor het gebruik van de API wordt aangeraden om *prompts* te gebruiken. Dit zijn aanwijzingen voor het model die aangeven wat verlangd wordt. De prompt is ontworpen om een tekst te geven en hierbij een drogreden in te vullen als er een is. Dit wordt gedaan door de aanduidingen 'Text:' en 'Name the logical fallacy if there is one:'. Het systeem herkent dit patroon en maakt hieruit op dat een drogreden moet worden gekoppeld aan de tekst.

Voorbeeld 1
Basis van de prompt

> Text: You didn't even finish high school. How could you possibly know about this?
> Name the logical fallacy if there is one: ad hominem attack
> Text:
> Name the logical fallacy if there is one:

Aangezien er wordt aangegeven in de API handleiding (OpenAI, 2020b) dat de resultaten verbeteren in een one- of few-shot setting, is er een voorbeeld inbegrepen in de prompt voor het model (zie Voorbeeld 1). Dit voorbeeld is voor elke zin die beoordeeld moet worden hetzelfde en dient alleen het doel om de taak te verduidelijken voor het model. Deze basis werd aangevuld met een zin uit de dataset en een afsluitende prompt voor de drogreden (zie Voorbeeld 2).

Voorbeeld 2
Voorbeeld van een volledige prompt met ad hominem voorbeeld

> Text: You didn't even finish high school. How could you possibly know about this?
> Name the logical fallacy if there is one: ad hominem attack
> Text: You cheated and lied to your wife, but you expect the jury to believe you now?
> Name the logical fallacy if there is one:

Ook is er een andere variant van de test uitgevoerd. Dit is gedaan met een ander voorbeeld in de prompt (zie Voorbeeld 3). Dit is gedaan om te kijken of er variatie zit in de classificaties en prestaties van GPT-3 als er een ander voorbeeld wordt gegeven.

Voorbeeld 3
Voorbeeld van een volledige prompt met false dilemma voorbeeld

> Text: You must be a Republican or Democrat. You are not a Democrat. Therefore, you must be a Republican.
> Name the logical fallacy if there is one: false dilemma
> Text: You cheated and lied to your wife, but you expect the jury to believe you now?
> Name the logical fallacy if there is one:

2.3 Modelinstellingen

Na het toegang krijgen tot het online platform moesten de instellingen van het model bepaald worden. In het model kunnen opties worden aangepast wat betreft het type model, de mate van willekeurigheid bij het selecteren van woorden, de frequentie- en aanwezigheids-penalty's, en start en restart teksten. De keuzes voor start en restart teksten draaien om wanneer de API herkent dat er een antwoord moet worden gegeven en wanneer er gewacht moet worden op input. Dit is zo geregeld dat er een antwoord werd verwacht na de prompt "Name the logical fallacy if there is one:" en een input na "Text:". De opties voor de frequentie- en aanwezigheids-penalty draaien om het voorkomen van eerder gebruikte antwoorden. De frequentie-penalty voorkomt het hergebruik van eerdere zinnen, waar de aanwezigheids-penalty het waarschijnlijker maakt om over te gaan op andere onderwerpen. Allebei deze opties zijn niet wenselijk voor het huidige doel. Dit is omdat een specifiek antwoord wordt verwacht op de vraag of er een drogreden wordt herkend, waarbij herhaling niet uitmaakt en een wisseling van onderwerp niet wenselijk is.
De belangrijkste instelling was de mate van willekeurigheid. Dit is een waarde tussen 0 en 1. Hoe hoger deze instelling, hoe waarschijnlijker het wordt dat de API een woord selecteert bij het vormen van een antwoord wat niet met de meeste waarschijnlijkheid volgt op de eerdere woorden. Een waarde van 0 bij deze instelling zorgt voor een deterministisch en repetitief model. Een zeer lage waarde bij deze instelling zorgde voor consistente antwoorden, waar een hoge waarde zorgde voor meer inconsistentie in de antwoorden. De waarde van 0 is uiteindelijk gekozen, aangezien met deze waarde de resultaten de opties zijn met de hoogste waarschijnlijkheid en bovendien te reproduceren zijn. Ook is deze waarde eerder gebruikt in een soortgelijk onderzoek van Chiu et al. (2022) naar de herkenning en classificatie van haatdragende tekst.

2.4 Dataverzameling

De classificaties van GPT-3 zijn verzameld met behulp van de dataset met drogredenen, de OpenAI API (OpenAI, 2020a) en Python (Van Rossum & Drake, 2009). De dataset is een CSV bestand met hierin een kolom met de teksten en een kolom met de bijbehorende drogreden. Deze dataset wordt met behulp van een Python script (zie Bijlage C) doorgegeven aan de OpenAI API. Dit kan door middel van het openai pakket (OpenAI, 2022). Het script leest de dataset in, gebruik makend van het pandas pakket (McKinney et al., 2010). Vervolgens wordt één voor één een drogreden gecombineerd met de basis van de prompt, resulterend in een complete prompt (zie Voorbeeld 2). Om de twee seconden wordt een prompt doorgespeeld aan de API, gebruikmakend van het openai pakket en de api key die bij het account komt. Deze key zorgt ervoor dat de prompt gekoppeld wordt aan het account en via het script kan worden verwerkt door de API. Het is nodig om gebruik te maken van het time pakket om de prompts te vertragen. De API staat slechts 60 verzoeken toe per minuut, en zonder de vertraging leidt dit tot foutmeldingen. De antwoorden die GPT-3 geeft worden opgeslagen in een dataframe. Na het doorlopen van de hele dataset wordt het bestand, inclusief tekst, drogreden en de door het model toegewezen drogreden weggeschreven naar een (nieuw) CSV bestand.
Dit CSV bestand met de resultaten is vervolgens doorlopen om de antwoorden naar vaste benamingen om te zetten. Dit is gedaan omdat de antwoorden die zijn gegeven door het model variëren qua benaming of specificiteit. Om de dataverwerking mogelijk te maken en te zorgen dat ieder voorkomen van een benaming zo goed mogelijk te classificeren is, is dit proces handmatig gebeurd. Zo zijn bijvoorbeeld de classificaties 'ad hominem attack' en 'poisoning the well' omgezet naar 'ad hominem'. Ook zijn classificaties door GPT-3 die niet duidelijk onder een bepaald type drogreden vielen omgezet naar 'miscellaneous' om aan te geven dat de classificatie niet onder een van de drogredenen viel. De hele lijst met omgezette benamingen is terug te vinden in Bijlage B.

2.5 Statistiek

Om te bepalen of GPT-3 drogredenen kan herkennen en correct kan classificeren, zijn de resultaten omgezet naar een confusion matrix. Dit is gedaan in R (R Core Team, 2019). Daarna zijn de kolommen herordend met behulp van het dplyr pakker (Wickham et al., 2022). Met de weergave van de confusion

matrix zijn per drogreden de precision, recall en F1-score berekend in Python (Van Rossum & Drake, 2009) met behulp van de pakketten pandas (McKinney et al., 2010) en sklearn (Pedregosa et al., 2011). Deze waarden zijn berekend om te kijken hoe goed GPT-3 drogredenen kan herkennen en onderscheiden. Precision, recall en de F1-score baseren zich allemaal op de concepten echt positief of true positive (TP), foutnegatief of false positive (FP), echt negatief of true negative (TN) en foutnegatief of false negative (FN). Een TP houdt in dat, in het geval van de classificering van een ad hominem drogreden, de drogreden correct is geclassificeerd. Een FP betekent dat een andere drogreden is geclassificeerd als een ad hominem. Ditzelfde gaat op voor TN en FN, maar dan voor een classificatie als iets anders dan een ad hominem. Een TN betekent dat een andere drogreden is herkend als een andere drogreden dan de ad hominem, waar een FN betekent dat een ad hominem is geclassificeerd als een andere drogreden. Om te bepalen waar de gegeven antwoorden van GPT-3 vallen wordt vastgehouden aan de classificaties gegeven door de datasets.

Voorbeeld 4
Formules van precision, recall en F1-score

Precision:	TP / (TP + FP)
Recall:	TP / (TP + FN)
F1-score:	(2 x Precision x Recall) / (Precision + Recall)

Met deze concepten kunnen de precision, recall en F1-score worden berekend (zie Voorbeeld 4). De precision is de verhouding true positives met alles wat geclassificeerd is als positief. De recall is de verhouding true positives met alle waarden die positief zouden moeten zijn. De F1-score is het harmonisch gemiddelde van de precision en recall. Dit geeft meer gewicht aan een lagere waarde; een hoge F1 betekent dus dat zowel precision als recall hoog zijn. Deze waarden geven ons een mogelijkheid om de classificatie van het model te beoordelen.
Verder is er een extra confusion matrix gecreëerd om te kijken of er verschillen zitten in de classificatie door het gebruik van een ander voorbeeld. In deze matrix zijn de classificaties van het model met een ad hominem als voorbeeld en de classificaties van het model met een false dilemma als voorbeeld tegen elkaar uitgezet.

3. Resultaten

3.1 Resultaten

In Tabel 1 zijn de classificaties van de drogredenen door het taalmodel te zien. Hierin is te zien dat bij een groot deel van de drogredenen de grootste groep van classificaties onder de juiste categorie valt. Dit is terug te zien aan de diagonale lijn met groene vakken, waarbij een donkerder groen staat voor een hoger getal. De diagonaal in deze tabel staat voor een match tussen werkelijke drogreden en de drogreden zoals hij is geclassificeerd door GPT-3. De rode vakken staan voor classificaties onder andere drogredenen die in de dataset voorkomen. De classificaties in het oranje staan voor classificaties onder drogredenen die niet in de data voorkomen. Ook hier staat een donkerdere kleur voor een hoger aantal. In Tabel 1 is te zien dat 80 van de 87 ad hominem drogredenen zijn geclassificeerd als een ad hominem. Ook is er te zien dat de drogreden straw man vaak wordt herkend als een andere drogreden; de donkerste kolom in de bijbehorende rijen is niet die van straw man, maar die van ad hominem. Dit houdt in dat een straw man drogreden het grootste deel van de tijd wordt geclassificeerd als een ad hominem. Verder valt het op dat circular reasoning in minder dan de helft van de gevallen goed wordt geclassificeerd.
Ook valt de classificatie van zinnen zonder drogreden op. Geen enkele drogreden wordt herkend door het model als none en geen enkele none wordt herkend als een van de drogredenen uit de dataset. Alle zinnen die worden herkend als een drogreden bevattende worden geclassificeerd onder non sequitur. De confusion matrix van de tweede test laat zien dat in veel gevallen de classificaties redelijk gelijk zijn (zie Bijlage A, Matrix A2).

Tabel 1

Confusion matrix van de classificatie van drogredenen door GPT-3 met ad hominem voorbeeld

Drogreden \ GPT-3 classificatie	ad hominem	appeal to authority	appeal to emotion	bandwagon	circular reasoning	false cause	false dilemma	faulty generalization	straw man	none	false consensus	false equivalence	miscellaneous	non sequitur	red herring	slippery slope	tu quoque
ad hominem	80	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	6
appeal to authority	2	34	0	1	0	6	1	5	0	0	0	1	0	0	0	0	0
appeal to emotion	0	0	17	1	5	0	4	0	0	0	0	0	2	0	0	1	0
bandwagon	1	1	0	31	0	2	11	2	0	0	1	1	0	0	0	0	0
circular reasoning	2	0	3	1	13	1	1	6	0	0	0	0	3	0	0	0	0
false cause	0	0	2	0	0	28	0	12	0	0	0	0	6	2	0	0	0
false dilemma	0	0	3	0	3	0	59	0	0	0	0	0	0	0	0	0	0
faulty generalization	7	0	1	0	0	2	4	26	0	0	0	0	0	1	0	7	1
straw man	13	1	1	0	0	1	8	0	2	0	0	0	1	0	1	0	2
none	0	0	0	0	0	0	0	0	0	21	0	0	0	9	0	0	0

De bijbehorende waarden van precision, recall en F1-score, samengevat in Tabel 2, laten gemengde resultaten zien. De hoogste scores wat betreft precision worden gehaald bij straw man en none met een maximum van 1,00. Ook appeal to authority en bandwagon halen hier goede scores met 0,94 en 0,91. Dit betekent dat alles wat wordt geclassificeerd door het systeem als straw man in alle gevallen ook een straw man was in de data. Bij appeal to authority en bandwagon was dit in respectievelijk 94 en 91 procent van de gevallen. De hoogste scores bij recall worden gehaald bij ad hominem en false dilemma met een maximum van 0,92. Dit betekent dat bij de ad hominem en false dilemma drogredenen respectievelijk 92 en 91 procent van alle ad hominem en false dilemma gevallen correct werden geclassificeerd. De F1-score, het harmonisch gemiddelde van precision en recall, ziet de hoogste scores bij ad hominem, appeal to authority, false dilemma en none met een maximum van 0,83. Bij deze categorieën waren zowel de precision als recall van zo'n niveau dat het harmonisch gemiddelde hiervan uitkwam op een minimum van 0,77. De F1-scores bij de tweede test, met een false dilemma als voorbeeld, liggen hier rond dezelfde niveaus als bij de eerste test (zie Bijlage A, Tabel A2).

Tabel 2

Precision, recall en F1-score van de classificatie

	Precision	Recall	F1	Support
ad hominem	0,76	0,92	0,83	87
appeal to authority	0,94	0,68	0,79	50
appeal to emotion	0,63	0,57	0,60	30
bandwagon	0,91	0,62	0,74	50
circular reasoning	0,62	0,43	0,51	30
false cause	0,70	0,56	0,62	50
false dilemma	0,66	0,91	0,77	65
faulty generalization	0,51	0,53	0,52	49
straw man	1,00	0,07	0,12	30
none	1,00	0,70	0,82	30
gewogen gemiddelde	0,77	0,63	0,65	471

Tabel 3 laat de classificaties van het model met een ad hominem als voorbeeld zien uitgezet tegen de resultaten van het model met een false dilemma als voorbeeld. In de matrix is te zien dat een groot deel van de classificaties door beide varianten onder dezelfde categorie wordt geplaatst. Hierin valt op dat de categorie waarbij het voorbeeld hoort bij beide varianten vaker voorkomt. Het model met de ad hominem als voorbeeld classificeert dus vaker een zin als ad hominem en het model met een false dilemma classificeert vaker een zin als false dilemma. Verder is te zien dat er een afwijking zit bij de classificatie van een groep false cause/faulty generalization, waar 12 gevallen door de ene variant worden herkend als een false cause, waar ze bij de ander onder faulty generalization vallen.

Tabel 3
Confusion matrix van de classificaties van drogredenen door GPT-3

Classificatie met ad hominem voorbeeld	Classificatie met false dilemma voorbeeld																
	ad hominem	appeal to authority	appeal to emotion	bandwagon	circular reasoning	false cause	false consensus	false dilemma	false equivalence	faulty generalization	miscellaneous	non sequitur	none	red herring	slippery slope	straw man	tu quoque
ad hominem	91	1	0	0	0	0	0	9	0	4	0	0	0	0	0	0	0
appeal to authority	0	27	0	0	0	6	0	1	0	2	0	0	0	0	0	0	0
appeal to emotion	1	0	14	0	1	2	0	6	0	0	0	3	0	0	0	0	0
bandwagon	0	0	0	26	0	4	0	2	0	2	0	0	0	0	0	0	0
circular reasoning	0	0	0	0	16	0	0	3	0	1	0	1	0	0	0	0	0
false cause	0	0	0	0	0	38	0	0	0	1	1	0	0	0	0	0	0
false consensus	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
false dilemma	0	0	0	0	0	7	1	80	0	1	0	0	0	0	0	0	0
false equivalence	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
faulty generalization	0	2	0	0	0	12	0	1	0	36	0	0	0	0	0	0	0
miscellaneous	0	0	0	0	1	4	0	2	0	1	4	0	0	0	0	0	0
non sequitur	0	0	0	0	0	1	0	0	0	0	0	5	6	0	0	0	0
none	0	0	0	0	0	0	0	0	0	0	0	5	16	0	0	0	0
red herring	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
slippery slope	0	0	0	0	0	3	0	0	0	0	0	0	0	0	5	0	0
straw man	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
tu quoque	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	5

3.2 Foutenanalyse

De classificaties van het GPT-3 model met de ad hominem zin als voorbeeld zijn geanalyseerd. Deze classificaties zijn in te delen in drie groepen; juiste classificaties, onjuiste, maar begrijpelijke classificaties en onjuiste classificaties.

Voorbeeld 5
Voorbeelden van zinnen uit dataset + classificatie van dataset / juiste classificatie van GPT-3

1. You must be a Republican or Democrat. You are not a Democrat. Therefore, you must be a Republican.
   False dilemma / False dilemma
2. You're a fascist, so clearly we shouldn't listen to what you have to say about education.
   Ad hominem / Ad hominem
3. Channeling the spirits of the dead is real, because John Edward says he can do it and he is an expert.
   Appeal to authority / Appeal to authority

De categorie met juiste classificaties bevat zinnen waarin GPT-3 een drogreden classificeert onder de juiste drogreden. Voorbeeld 5.1 is correct geclassificeerd als een false dilemma drogreden. Er wordt een keuze voorgelegd tussen twee mogelijkheden; democraat of republikein zijn. Er zijn echter meerdere keuzes dan de twee die worden aangegeven. Voorbeeld 5.2 is bevat een ad hominem drogreden. De uitspraak van iemand wordt in twijfel gebracht door diens karakter te schaden. Voorbeeld 5.3 bevat een appeal to authority drogreden. Een standpunt wordt verdedigd door te refereren naar een zogenaamd expert die zegt dat het kan.

Voorbeeld 6
Voorbeelden van zinnen uit dataset + classificatie van dataset / onjuiste, begrijpelijke classificatie van GPT-3

1. You're saying that I should clean up my room? But you never clean up yours!
   Ad hominem / Tu quoque
2. If you can't be loyal and support the way your government chooses to use taxes, then you should just leave the country and move somewhere else.
   Ad hominem / False dilemma
3. If you don't study, you'll fail your test. Then you will do poorly in the class and your GPA will fall. You won't get into a good college, so you'll never get a decent job and you'll end up poor and homeless.
   Faulty generalization / Slippery slope
4. Don't use iPhones. Teachers use iPhones and teachers are extremely dorky. If you use an iPhone, you will be dorky like a teacher!
   Faulty generalization / Ad hominem

De categorie met onjuiste, begrijpelijke classificaties gaat om voorbeelden waar meerdere drogredenen in herkend kunnen worden. Het systeem herkent in deze klasse drogredenen die wel in de data zitten, maar niet in de data geclassificeerd staan als deze drogreden. Voorbeeld 6.1 is door de data geclassificeerd als een ad hominem drogreden, waar het in werkelijkheid een tu quoque drogreden is. Volgens de aangehouden classificatie (The School of Thought, z.d.) zijn dit aparte drogredenen. Andere bronnen geven echter aan dat tu quoque een subcategorie van ad hominem is (Aikin, 2007). In dat geval zijn de classificaties wel terecht. Voorbeeld 6.2 staat als ad hominem in de data. Deze zin kan echter ook gezien worden als een false dilemma en wordt ook zo geclassificeerd. De false dilemma drogreden in deze zin is de keuze tussen loyaal zijn of weggaan. Voorbeeld 6.3 valt in de data onder een faulty generalization. De zin "If you don't study, you'll fail your test." kan gezien worden als een foute generalisatie, aangezien het niet per se het geval is dat je faalt als je niet leert. De rest van de zin zorgt er echter voor dat de drogreden slippery slope beter past, omdat het steeds een stap verder gaat. Voorbeeld 6.4 bevat een faulty generalization van leraren naar iPhone gebruikers. De zin bevat echter ook een persoonlijke aanval, namelijk "teachers are extremely dorky". Het systeem kan dit hebben gebruikt voor de classificatie als ad hominem.

Voorbeeld 7
Voorbeelden van zinnen uit dataset + classificatie van dataset / onjuiste classificatie van GPT-3

1. My opponent argues that we should abolish the soda tax. It's a shame that he wants to encourage people to eat and drink unhealthily. I say we keep it.
   Straw man / Ad hominem
2. Michael Giacchino composed the score for Doctor Strange.
   None / Non sequitur
3. It may be against the law to drink alcohol if you are under 18 years old, but almost everyone drinks anyway, so it must be fine.
   Bandwagon / False dilemma

De categorie met onjuiste classificaties bevat zinnen waarbij de classificatie van de dataset en GPT-3 niet overeenkomen. Ook is er bij deze zinnen geen sprake van de classificatie die wordt gegeven door het taalmodel. Voorbeeld 7.1 is een zin die een straw man bevat; het standpunt van de tegenstander wordt verdraaid. Het systeem classificeert deze zin, samen met andere straw man drogredenen, als een ad

hominem. Een mogelijke verklaring hiervoor is het feit dat de tweede zin een slecht beeld schetst van de tegenstander, wat ook gebeurt bij een ad hominem. Het systeem herkent de verdraaiing naar deze vorm wellicht als een ad hominem. Voorbeeld 7.2 bevat een zin zonder drogreden. GPT-3 koppelt echter aan meerdere van dit soort zinnen de drogreden non sequitur, die verder vrijwel niet voorkomt. Een mogelijke verklaring hiervoor is dat volgens GPT-3 er geen argumentatie te vinden is bij de claim. Dit kan er echter ook op wijzen dat het systeem een voorkeur heeft voor het toedelen van een drogreden. Voorbeeld 7.3 is een zin met een bandwagon drogreden; iedereen drinkt, dus dan zal het wel goed zijn. De classificatie die hierbij wordt gegeven is een false dilemma. Dit kan komen door de opzet van de zin. Het is of verboden, of het is oké. Iedereen doet het, dus het is oké. Dit volgt dezelfde opzet als een false dilemma kan hebben, zoals in 5.1; het is niet geval A, dus het is geval B. Dit kan de reden zijn dat het taalmodel dit classificeert onder false dilemma.

4. Conclusie en discussie

4.1 Conclusie en verklaringen

Dit onderzoek is uitgevoerd om te kijken naar de prestaties van het online-toegankelijke taalmodel GPT-3 bij het herkennen en classificeren van drogredenen. Om dit te onderzoeken is een classificatietaak opgezet. Toegang tot het model was mogelijk via de OpenAI API (OpenAI, 2020a). Een dataset bestaande uit 471 zinnen die verschillende types drogredenen bevatten is samengesteld. Deze dataset is gebruikt om drogredenen te laten classificeren door het taalmodel. De resultaten van deze classificatie zijn vervolgens geanalyseerd. Uit de data bleek dat de prestaties van GPT-3 ernstig wisselden per drogreden. Zo was de herkenning en classificatie van ad hominem drogredenen relatief goed, met een F1-score van 0,83. Ook de drogredenen appeal to authority, bandwagon, false dilemma en de zinnen zonder drogreden werden tot op zekere hoogte goed herkend. De herkenning van de andere drogredenen verliep moeizamer. De drogreden straw man werd niet goed herkend, met een F1-score van 0,12. Bij de overige drogredenen schommelden de prestaties tussen F1-scores van 0,51 en 0,62. De prestaties van het model zouden echter nog verbeteren als de onjuiste, begrijpelijke drogredenen geteld zouden worden als correct. Dit zou een optie zijn omdat de herkende drogredenen wel voorkomen in de data.
Deze resultaten komen overeen met de gestelde hypothese dat de prestaties van een dergelijk geavanceerd taalmodel bij de herkenning van drogredenen wisselend zijn. In overeenstemming met de gestelde hypothese worden de drogredenen ad hominem, appeal to authority, bandwagon en false dilemma het beste herkend, met F1-scores tussen de 0,74 en 0,83.

Deze resultaten bevestigen deels de conclusies getrokken door Jin et al. (2022) over het verschil in herkenbaarheid van drogredenen. In beide onderzoeken waren zowel ad hominem als bandwagon onder de best herkende drogredenen. Wel zit er een verschil in de herkenbaarheid van appeal to authority en false dilemma. In het onderzoek van Jin et al. (2022) lagen de F1-scores van deze drogredenen respectievelijk op 0,59 en 0,55, waar in dit onderzoek de scores op 0,79 en 0,77 uitkwamen. Ook is er nog onenigheid te vinden met het onderzoek van Jin et al. (2022); de prestaties van niet-gefinetunede modellen bereiken een maximum F1-score van 0.14 in hun onderzoek. De resultaten van dit onderzoek laten een stuk hogere waarden zien voor de classificatie van drogredenen door een niet-gefinetuned model, wat in dit geval GPT-3 is. Een mogelijke verklaring hiervoor is het verschil in het gebruikte model. Een andere verklaring hiervoor kan de aangepaste dataset zijn. Zoals eerder vermeld in 2.1 is een selectie van de LOGIC dataset (Jin et al., 2022) toegevoegd aan de huidige dataset. Deze selectie is gemaakt om sterk op elkaar lijkende zinnen uit de dataset te filteren. Het is mogelijk dat de gebruikte dataset gemakkelijker herkenbare drogredenen bevatte.

Een mogelijke verklaring voor de uitkomsten is dat van de drogredenen die het beste worden herkend simpelweg meer voorbeelden in de trainingsdata van GPT-3 zitten. Bij het samenstellen van de dataset viel al op dat het gemakkelijker was om voorbeelden te vinden bij bijvoorbeeld ad hominem dan bij andere types, zoals straw man. Het kan dus zijn dat het systeem meer voorbeelden in het trainingsmateriaal heeft gehad van bepaalde types drogredenen dan andere. Het model namelijk is

getraind op grote hoeveelheden data die voornamelijk online is terug te vinden (Brown et al., 2020). Dit in combinatie met de ervaringen bij het opstellen van de dataset maken dit tot een reële mogelijkheid.

Een andere verklaring voor de verkregen resultaten is dat de best herkende drogredenen gemakkelijker te herkennen en classificeren zijn door bepaald woordgebruik. Jin et al. (2022) geven aan dat het een mogelijkheid is dat bandwagon drogredenen worden herkend door het gebruikt van termen die een meerderheid aanduiden en dat ad hominem drogredenen herkend worden door het gebruik van haatdragende taal. De herkenning van haatdragende taal door taalmodellen tot op zekere hoogte is al eerder bevestigd door Samghabadi et al (2020) en Chiu et al. (2022). Het zou dus mogelijk zijn dat de drogredenen geclassificeerd worden op woordgebruik, of dat dit in ieder geval het proces vergemakkelijkt. Ook de analyse van goed geclassificeerde drogredenen sluit dit niet uit. Zo kan de drogreden ad hominem zijn toegewezen op het gebruik van het woord fascist of kan het gebruik van dit woord de classificatie hebben geholpen.

4.2 Verbeterpunten

Er zijn meerdere potentiële verbeteringspunten voor dit onderzoek. Allereerst betreft dit de gebruikte dataset. Bij het samenstellen van de dataset kwamen verschillende problemen naar voren. Zo was de selectie van drogredenen een punt van aandacht. Van bepaalde drogredenen zijn meer voorbeelden te vinden. Dit is bijvoorbeeld te zien in de dataset bij het aantal ad hominem drogredenen. Van de 471 zinnen bevatten 87 een ad hominem. De dataset bevat 9 verschillende drogredenen in totaal. Ieder van deze types heeft minimaal 30 voorbeelden in de data. Het aanvullen van de dataset met meer verschillende types drogredenen zou een beter beeld kunnen geven van welke drogredenen wel en welke niet goed herkend kunnen worden. Ook zou er gelet kunnen worden op variatie in de zinsconstructies die binnen een drogreden voorkomen. Zo zou er gecontroleerd kunnen worden of alleen variaties op een standaardvorm van een drogreden herkend worden, of dat de drogreden in het algemeen herkend kan worden.

Een ander probleem wat het gebruik van drogredenen met zich meebrengt is het feit dat niet elk gebruik van een drogreden even duidelijk is. Sommige drogredenen zijn lastig te identificeren of liggen dichtbij andere drogredenen. Een voorbeeld hiervan is het verschil tussen de ad hominem en de tu quoque drogredenen. Waar de ad hominem een persoonlijke aanval is op de spreker, is de tu quoque gericht op een voorval. Een tu quoque kan echter ook een manier zijn om het karakter van de spreker te besmeuren; het kan worden gebruikt als een persoonlijke aanval. Sommige bronnen scharen de tu quoque zoals eerder gezegd ook onder ad hominem als een subcategorie (Aikin, 2007). Op deze manier wordt het lastig om de dataset te splitsen op drogreden en te garanderen dat de classificatie van drogredenen correct wordt beoordeeld. Dit kan invloed hebben op de uitkomsten als de classificatie wordt beoordeeld als fout, terwijl de zin uit de data dan wel meerdere drogredenen bevat, dan wel een twijfelgeval is.
Dit is bij classificatie van de dataset een punt van aandacht. Zoals eerder vermeld zitten er in de verzamelde data voorbeelden van drogredenen die ook onder andere drogredenen kunnen vallen.

4.3 Vervolgonderzoek

Bij eventueel vervolgonderzoek kan er worden gekeken naar het gebruiken van andere mogelijkheden van het GPT-3 model om drogredenen te herkennen. Zo zou er gekeken kunnen worden naar de prestaties van het model als er, in plaats van een one-shot setting, gebruik wordt gemaakt van een few-shot setting. In het onderzoek van Chiu et al. (2022) presteerde de few-shot setting beter dan de one-shot setting. Ook door Brown et al (2020) wordt aangegeven dat de one- en few-shot settingen betere resultaten geven dan de no-shot setting. Dit kan zorgen dat de prestaties verbeteren en de drogredenen beter herkend kunnen worden.
Bovendien geeft dit de mogelijkheid om verder te kijken naar de invloed van het voorbeeld in de prompt op de resultaten. De matrix die de twee classificaties van GPT-3 tegen elkaar uitzette gaf aan dat er een grote overlap zit in de classificatie, maar dat er ook afwijkingen zijn. Er lijkt een lichte voorkeur te zijn voor de drogreden gebruikt in het voorbeeld voor de classificatie. Ook Chiu et al. (2022) geven aan dat de

inhoud van voorbeelden mogelijk invloed heeft op de kwaliteit van de resultaten. Het gebruik van een few-shot setting kan deze invloed wellicht beperken door het aanbieden van verschillende drogredenen in het voorbeeld. Het is in ieder geval wenselijk verder te onderzoeken wat de invloed van de gegeven voorbeelden is.
Verder is het een mogelijkheid om de keuzemogelijkheden van het model te limiteren, zodat het alleen kan kiezen uit een selectie van antwoorden. Dit lost het probleem van verschillende benamingen voor dezelfde drogreden op, waardoor er niet meer gekeken hoeft te worden naar de namen van de classificaties. Dit kan ervoor zorgen dat de classificatie verbetert omdat het aantal mogelijkheden vermindert. Dit zorgt er echter ook voor dat de mogelijkheden van tevoren bekend moeten zijn, wat leidt tot minder generaliseerbaarheid van de herkenning van drogredenen in teksten.
Het was niet meer mogelijk deze opties te includeren in het huidige onderzoek. Dit komt omdat de OpenAI API een budget geeft voor een maximaal aantal tokens zonder extra te moeten betalen voor het verwerken van verzoeken. De limiet van dit budget was bereikt na het uitvoeren van het huidige onderzoek.

Een punt wat is gerelateerd aan het aanpassen van opties is het finetunen van het model. GPT-3 biedt de optie om een eigen versie van het model te laten creëren door het finetunen op trainingsdata. Op deze manier is het duidelijk voor het systeem wat ervan gevraagd wordt en kan je het model laten trainen op meer voorbeelden dan er in een prompt passen. Een nadeel hiervan is dat er wel nog meer data nodig is om een extra set te kunnen creëren voor de training. In de instructie van OpenAI (OpenAI, 2020b) voor het creëren van een finetuned model wordt aangeraden om minimaal 100 voorbeelden per classificatie te hebben. Het finetunen van het model kan er echter wel voor zorgen dat de kwaliteit van de resultaten hoger is dan bij het gebruik van het basismodel met prompts. De prestaties van een gefinetunede versie kunnen dan worden vergeleken met de resultaten van het basismodel.

Verder is er de mogelijkheid voor vervolgonderzoek om in te gaan op de herkenning van drogredenen in grotere context. In plaats van een taalmodel te vragen om de inhoud van één of een paar zinnen te classificeren, zou er gekeken kunnen worden naar de prestaties van een taalmodel als er wordt gevraagd een drogreden te halen uit een langer stuk tekst. Het gegeven dat GPT-3 moeite heeft met de productie van langere, begrijpelijke zinnen (Brown et al., 2020) kan wellicht ook betrekking hebben op begrip en verwerking. Dit is echter een punt wat onderzocht moet worden. De herkenning van drogredenen in langere teksten zou grote gevolgen kunnen hebben voor de toepassingen van dit model.

Dit onderzoek laat zien dat drogredenen tot op zekere hoogte herkend kunnen worden door het taalmodel GPT-3. De mate waarin dit kan wisselt echter sterk tussen verschillende drogredenen. De drogredenen ad hominem, appeal to authority, bandwagon en false dilemma laten de beste resultaten zien. Om een beter beeld te krijgen van de mogelijkheden van geavanceerde taalmodellen zoals GPT-3 zal echter meer onderzoek gedaan moeten worden naar dit vraagstuk.

Literatuurlijst

Aikin, S. F. (2007). Tu Quoque Arguments and the Siginificance of Hypocrisy. http://dx.doi.org/10.2139/ssrn.1012620

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Chiu, K., Collins, A. & Alexander, R. (2022). Detecting Hate Speech with GPT-3. https://doi.org/10.48550/arXiv.2103.12407

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R. & Schölkopf, B. (2022). Logical fallacy detection. https://doi.org/10.48550/arXiv.2202.13758

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 445, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a.

Nakpih, C. & Santini, S. (2020). Automated Discovery of Logical Fallacies in Legal Argumentation. *International Journal of Artificial Intelligence & Applications*, 11, 7-48. https://doi.org/10.5121/ijaia.2020.11203

OpenAI (2022). *Openai 0.20.0.* https://pypi.org/project/openai/

OpenAI. (2020a). *OpenAI API*. Geraadpleegd op 14 juli 2022, van https://beta.openai.com/playground

OpenAI. (2020b). *OpenAI API Documentation*. Geraadpleegd op 14 juli 2022, van https://beta.openai.com/docs/introduction/overview

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Proudfoot, M. (2010). The Routledge dictionary of philosophy. London: Routledge. 341. ISBN 978-0-203-42846-7.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Samghabadi, N. S., Patwa, P., Srinivas, P. Y. K. L. , Mukherjee, P., Das, A. & Solorio, T. (2020). Aggression and Misogyny Detection using BERT: A Multi-Task Approach. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 126–131, Marseille, France. European Language Resources Association (ELRA).

The School of Thought. (z.d.). *Thou shalt not commit logical fallacies*. Yourlogicalfallacyis.Com. Geraadpleegd op 15 juli 2022, van https://yourlogicalfallacyis.com/

Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 809–819. New Orleans, Louisiana. Association for Computational Linguistics.

Van Rossum, G. & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Van Helleman, N. (2022). Fallacies. https://github.com/Aonnav/fallacies

Wickham, H., François, R., Henry, L. & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Williamson, O. M. (2015). *Master List of Logical Fallacies*. utminers.utep.edu. Geraadpleegd op 15 juli 2022, van http://utminers.utep.edu/omwilliamson/engl1311/fallacies.htm

Bijlagen

Bijlage A: Matrices en tabellen

Matrix A1

Confusion matrix van de classificatie van drogredenen door GPT-3 met ad hominem voorbeeld

Drogreden \ GPT-3 classificatie	ad hominem	appeal to authority	appeal to emotion	bandwagon	circular reasoning	false cause	false dilemma	faulty generalization	straw man	none	false consensus	false equivalence	miscellaneous	non sequitur	red herring	slippery slope	tu quoque
ad hominem	80	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	6
appeal to authority	2	34	0	1	0	6	1	5	0	0	0	1	0	0	0	0	0
appeal to emotion	0	0	17	1	5	0	4	0	0	0	0	0	2	0	0	1	0
bandwagon	1	1	0	31	0	2	11	2	0	0	1	1	0	0	0	0	0
circular reasoning	2	0	3	1	13	1	1	6	0	0	0	0	3	0	0	0	0
false cause	0	0	2	0	0	28	0	12	0	0	0	0	6	2	0	0	0
false dilemma	0	0	3	0	3	0	59	0	0	0	0	0	0	0	0	0	0
faulty generalization	7	0	1	0	0	2	4	26	0	0	0	0	0	1	0	7	1
straw man	13	1	1	0	0	1	8	0	2	0	0	0	1	0	1	0	2
none	0	0	0	0	0	0	0	0	0	21	0	0	0	9	0	0	0

Tabel A1

Precision, recall en F1-score van de classificatie met ad hominem voorbeeld

	Precision	Recall	F1	Support
ad hominem	0,76	0,92	0,83	87
appeal to authority	0,94	0,68	0,79	50
appeal to emotion	0,63	0,57	0,60	30
bandwagon	0,91	0,62	0,74	50
circular reasoning	0,62	0,43	0,51	30
false cause	0,70	0,56	0,62	50
false dilemma	0,66	0,91	0,77	65
faulty generalization	0,51	0,53	0,52	49
straw man	1,00	0,07	0,12	30
none	1,00	0,70	0,82	30
weighted avg	0,76	0,66	0,67	471

Matrix A2

Confusion matrix van de classificatie van drogredenen door GPT-3 met false dilemma voorbeeld

Drogreden \ GPT-3 classificatie	ad hominem	appeal to authority	appeal to emotion	bandwagon	circular reasoning	false cause	false dilemma	faulty generalization	straw man	none	false consensus	false equivalence	miscellaneous	non sequitur	red herring	slippery slope	tu quoque
ad hominem	79	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	3
appeal to authority	1	29	0	0	0	13	2	3	0	0	0	1	1	0	0	0	0
appeal to emotion	0	0	13	0	5	2	7	0	0	0	0	0	0	2	0	1	0
bandwagon	1	1	0	26	0	6	10	3	0	0	2	1	0	0	0	0	0
circular reasoning	1	0	0	0	12	5	3	6	0	0	0	0	1	2	0	0	0
false cause	0	0	0	0	0	38	1	7	0	0	0	0	3	1	0	0	0
false dilemma	1	0	1	0	0	1	62	0	0	0	0	0	0	0	0	0	0
faulty generalization	4	0	0	0	1	11	2	25	0	0	0	0	0	1	0	4	1
straw man	8	0	0	0	0	1	13	4	2	0	0	0	0	0	1	0	1
none	0	0	0	0	0	0	0	0	0	22	0	0	0	8	0	0	0

Tabel A2

Precision, recall en F1-score van de classificatie met false dilemma voorbeeld

	Precision	Recall	F1	Support
ad hominem	0,83	0,91	0,87	87
appeal to authority	0,97	0,58	0,72	50
appeal to emotion	0,93	0,43	0,59	30
bandwagon	1,00	0,52	0,62	50
circular reasoning	0,67	0,40	0,50	30
false cause	0,49	0,76	0,60	50
false dilemma	0,59	0,95	0,73	65
faulty generalization	0,52	0,51	0,52	49
straw man	1,00	0,07	0,12	30
none	1,00	0,73	0,85	30
weighted avg	0,78	0,65	0,66	471

Bijlage B: Benamingen

Tabel B1

Tabel met verschillende termen voor drogredenen gegeven door GPT-3

ad hominem	ad hominem ad hominem attack poisoning the well
appeal to authority	appeal to authority argument from authority false authority celebrity endorsement
appeal to emotion	appeal to emotion emotional manipulation fearmongering emotional appeal appeal to pity appeal to fear emotional blackmail
bandwagon	bandwagon peer pressure argument from popularity appeal to popularity
circular reasoning	circular reasoning begging the question
false cause	false cause post hoc ergo propter hoc
false dilemma	false dilemma black and white thinking
faulty generalization	faulty generalization hasty generalization generalization stereotype stereotyping
straw man	straw man
miscellaneous	self-blame survivorship bias superstition conspiracy theory blaming patriotism fallacy of composition personal experience special pleading blackmail appeal to force

Bijlage C: Python Code

Bijlage C1: GPT

```
# beschikbaar via https://github.com/Aonnav/fallacies

# importeren van de nodige pakketten

import openai

import pandas

import time


# key voor connectie met de API (koppeling aan account)

openai.api_key = ""


# inlezen van de dataset

df =
pandas.read_csv(r'D:\Documents\Uni\TW\B3\Scriptie\Data\Fallacies.csv',
sep = ';')


# aanmaken array voor classificaties

gpt_fallacy = []


def gpt3(prompt, engine='text-davinci-002', response_length=64,

         temperature=1, top_p=1, frequency_penalty=0,
presence_penalty=0,

         start_text='', restart_text='', stop_seq=[]):

    response = openai.Completion.create(

        prompt=prompt + start_text,

        engine=engine,

        max_tokens=response_length,

        temperature=temperature,

        top_p=top_p,

        frequency_penalty=frequency_penalty,

        presence_penalty=presence_penalty,

        stop=stop_seq,

    )
```

```
    answer = response.choices[0]['text']
    new_prompt = prompt + start_text + answer + restart_text
    return answer, new_prompt


# doorlopen van de regels (drogredenen) uit dataset (df), combineren
met prompt, doorspelen aan de API en opslaan van het antwoord.
def fallacies():
    for line in range(len(df)):
        prompt = """Text: You didn't even finish high school. How
could you possibly know about this?
        Logical fallacy: ad hominem attack
        Text:""" + df.at[line,'text']
        print(prompt)
        answer, prompt = gpt3(prompt,
                              temperature=0,
                              frequency_penalty=0,
                              presence_penalty=0,
                              start_text='\nLogical fallacy:',
                              restart_text='\nText: ',
                              stop_seq=['\nText:', '\n'])
        print('text:' + df.at[line,'text'])
        print('GPT-3:' + answer)
        gpt_fallacy.append(answer)
        time.sleep(2)



if __name__ == '__main__':
    fallacies()
    # toevoegen van de opgeslagen classificaties aan de dataset
    df['gpt_fallacy'] = gpt_fallacy
    print(df)
   # wegschrijven van de dataset inclusief classificaties
df.to_csv(r'D:\Documents\Uni\TW\B3\Scriptie\Data\FallaciesGPT.csv')
```

Bijlage C2: dataverwerking

```
# beschikbaar via https://github.com/Aonnav/fallacies

# importeren van de nodige pakketten

from sklearn.metrics import classification_report


if __name__ == '__main__':

    # inlezen van het bestand (met handmatig aangepaste)
classificaties

    df =
pandas.read_csv(r'D:\Documents\Uni\TW\B3\Scriptie\Stats\Scriptie\full.
csv', sep=';')

    print(df)

    # berekenen van de precision, recall en F1-scores

    print(classification_report(df['fallacy'], df['gpt_adjusted']))
```

Bijlage D: R Code

```
# beschikbaar via https://github.com/Aonnav/fallacies

# importeren van de nodige pakketten

library(dplyr)


# laden van de data

data <- read.csv2('full.csv')


# creëren confusion matrix

tab <- table(data$fallacy, data$gpt_adjusted)

tab2 <- as.data.frame.matrix(tab)


# herordenen kolommen

tab2 <- tab2 %>% relocate(c(`circular reasoning`, `false cause`,
`false dilemma`, `faulty generalization`, `straw man`), .after =
bandwagon)


# wegschrijven van de data

write.csv2(tab2, file = 'matrix.csv')


# creëren matrix GPTxGPT

tab3 <- table(data$gpt_adjusted, data$gpt_adjusted2)


# wegschrijven van de data

write.csv2(tab3, file = 'matrix3.csv')
```