UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Van URL's naar Inzichten: AI-gedreven Klantensegmentatie en Duurzaamheidsrangschikking

Maxim Riebus

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

dr. Frank VANHOENSHOVEN

2023

2024

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Van URL's naar Inzichten: AI-gedreven Klantensegmentatie en Duurzaamheidsrangschikking

Maxim Riebus

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

dr. Frank VANHOENSHOVEN

Van URL's naar Inzichten: AI-gedreven Klantensegmentatie en Duurzaamheidsrangschikking

Maxim Riebus[a]

[a]*Universiteit Hasselt, Faculteit Bedrijfseconomische wetenschappen, Agoralaan Gebouw D, Diepenbeek, 3590, Limburg, België*

Abstract

Inzichten van bedrijven winnen op basis van website-informatie kan een essentieel aspect zijn in het begrijpen en benaderen van potentiële zakelijke relaties. Dit onderzoek focust op een gedetailleerde evaluatie van een AI-gedreven algoritme dat bedrijven automatisch classificeert in hun industrie en een score tussen 1 en 5 toekent op basis van hun duurzaamheidsprofilering. Tot slot wordt er ook een analyse uitgevoerd om de relatie tussen de industrie en de duurzaamheidsprofilering beter te begrijpen. Het startpunt is een lijst van 250 website-URL's van bedrijven die verdeeld zijn over vijf industrieën. Hier worden geavanceerde tekstherkenningstechnieken op toegepast om de meest relevante pagina's van bijhorende websites te identificeren, waarna de tekstuele inhoud verzameld wordt met behulp van *web scraping*. Die output wordt dan gebruikt voor het uitvoeren van drie verschillende analyses.

De eerste focus van deze studie betreft het toekennen van industrieën, waarbij het AI-model een *accuracy* van 92,37% laat zien. Dit toont aan dat het model effectief is in het toewijzen van bedrijven aan de correcte industrie op basis van gescrapete inhoud van de 'over-ons'-gerelateerde webpagina's. Daarnaast worden ook metrieken zoals de *confusion matrix*, *precision*, *recall* en *F1-score* gebruikt om de prestaties van het model binnen elke industrie afzonderlijk te meten.

Ten tweede geeft het model aan elk bedrijf een score tussen 1 en 5 op basis van duurzaamheidsprofilering. Deze resultaten werden vervolgens gespiegeld met de resultaten van menselijke input uit een enquête waar 34 respondenten aan hebben deelgenomen. In 88% van de gevallen is de afwijking van de AI-score ten opzichte van de mediaan uit de enquête gelijk aan nul, of maximaal één, ten opzichte van de mediaan uit de enquête. De gemiddelde absolute fout bedraagt 0,61 punten. Er werd ook een clustering uitgevoerd op de teksten afkomstig van de website om te achterhalen of er verschillen en/of gelijkenissen zijn in de duurzaamheidsprofileringen van bedrijven. Daarnaast werd het model getest zonder de input van gescrapete website-inhoud. Hierbij kreeg het model enkel de bedrijfsnaam als input en werd de score toegekend op basis van de trainingsdata dat het model ter beschikking heeft. Die trainingsdata bestaat uit data internetgegevens tot en met september 2023.

Tot slot werd de relatie tussen de industrie en de duurzaamheidsprofilering onderzocht. De bevindingen tonen aan dat er significante verschillen aanwezig zijn tussen de duurzaamheidsprofileringen van bedrijven uit verschillende industrieën. De energiesector profileerde zich volgens het onderzoek het sterkst op duurzaamheid met een gemiddelde score van 4.04, terwijl de autosector het laagste scoorde met een gemiddelde van 3.08.

De bevindingen benadrukken het potentieel van AI om complexe data te transformeren in waardevolle inzichten. Deze gedetailleerde segmentatie stelt *B2B-providers* in staat om hun benaderingsstrategieën te personaliseren en aan te passen aan de specifieke behoeften en interesses van elk bedrijf.

Keywords: Natural Language Processing, Web Scraping, Customer Segmentation, Sustainability Analysis

1. Introductie

We leven in een datagedreven wereld waar gegevens publiek beschikbaar zijn op het internet. Dat heeft ervoor gezorgd dat het gemakkelijk is om informatie te verzamelen dat kan worden omgezet in waardevolle inzichten (5). Dankzij de snelle vooruitgang in artificiële intelligentie (AI) en *machine learning* (ML) zijn er nieuwe mogelijkheden ontstaan om grote hoeveelheden data te analyseren (22). Bedrijven moeten meer focussen op het aspect van duurzaamheid in de brede zin. Concreet is het belangrijk dat sociale, ecologische, economische en culturele overwegingen geïntegreerd worden in de bedrijfsstrategie. Dit wordt gezien als essentieel om een concurrentieel voordeel te behouden op lange termijn (25). Andere voordelen die verbonden zijn aan het aspect van duurzaam ondernemen zijn: naleving van de wet, verbeterde klantrelaties en een goede merkreputatie (18). Het aspect van duurzaamheid is zeer breed. Onderzoek van Palmer et al. (44) geeft onder andere broeikasgasreductie, diversiteitstraining en het uitroeien van mensenhandel aan als voorbeelden van duurzaamheidspraktijken. Dezelfde studie geeft ook aan dat vooral grote bedrijven een rol spelen in duurzaamheid vanwege hun economische impact en beschikbare middelen. Steeds meer bedrijven worden dan

ook lid van wereldwijde duurzaamheidsinitiatieven zoals het *Global Compact* van de Verenigde Naties (44). Bedrijven maken gebruik van het internet om hun initiatieven rond duurzaam ondernemen bekend te maken aan hun *stakeholders*. Dit gebeurt bijvoorbeeld via sociale media, maar ook de website van het bedrijf speelt hier een belangrijke rol in (10). Recente technieken om inhoud van webpagina's te verzamelen, kunnen samen AI helpen om op grote schaal waardevolle inzichten van bedrijven te verkrijgen (46; 32; 47).

De bestaande literatuur biedt vooral inzichten over afzonderlijke technieken, genaamd *web scraping*, *Natural Language Processing* en *customer segmentation*. *Web scraping* is een methode voor het verzamelen van grote hoeveelheden data van websites (7). *Natural Language Processing* (NLP) wordt gebruikt voor het analyseren van tekstdata en biedt krachtige tools voor het begrijpen van menselijke taal (33) en *customer segmentation* is een techniek waarmee bedrijven hun klantgroepen kunnen identificeren en beter kunnen begrijpen (61). Ondanks de uitgebreide kennis over *web scraping*, *Natural Language Processing* en *Customer segmentation*, is er nog geen onderzoek gevoerd naar de toepassing om een bestaande klantenbestand te segmenteren, clusteren en profileren, startende van URL's. Nochtans kunnen die drie domeinen gecombineerd worden om een algoritme te ontwerpen dat op een geautomatiseerde manier bedrijven gaat toekennen aan een industrie en die bedrijven ook beoordeelt op basis van hun duurzaamheidsprofilering. Wanneer het mogelijk is om zo een algoritme toe te passen op je eigen klantenbestand, kan je die kennis gebruiken om geïnformeerde beslissingen te maken bij het benaderen van je klanten. Dit zorgt er vervolgens voor dat de klanttevredenheid toeneemt (12).

Deze thesis focust op het ontwikkelen en evalueren van een AI-gedreven algoritme dat bedrijven automatisch classificeert in hun industrie en een score geeft op basis van de duurzaamheidsprofilering die beschikbaar is op hun website. Het doel is om inzichten te verkrijgen in de prestaties van het voorgesteld model om vervolgens te onderzoeken wat de relatie is tussen de industrie en de duurzaamheidsprofilering.

Deze paper is opgebouwd als volgt. In sectie 2 wordt de gerelateerde literatuur omtrent *web scraping*, *natural language processing* en *customer segementation* beschreven. Sectie 3 gaat dieper in op de methodologie rond de literatuurstudie, het ontworpen algoritme, de data-analyse en de metrieken die gebruikt worden om het algoritme te valideren. In sectie 4 worden de resultaten besproken en sectie 5 behandelt de tekortkomingen en de aanbevelingen voor toekomstig onderzoek. Ten slotte wordt er een conclusie gegeven in sectie 6.

2. Gerelateerde literatuur

De literatuur rapporteert drie belangrijke domeinen: '*Web Scraping*', '*Natural Language Processing*' en '*Customer Segmentation*'. Dat zijn zeer brede domeinen, waarbij in deze studie telkens gekeken wordt naar welke toepassingen nuttig kunnen zijn voor deze specifieke case. In de drie volgende secties zullen de belangrijkste elementen die nuttig zijn binnen dit onderzoek, besproken worden.

2.1. Web scraping

Het startpunt van dit onderzoek is het extraheren van tekstuele data die op specifieke webpagina's van bedrijven staan. Het proces dat dit uitvoert, heeft de naam '*web scraping*' (7). Meestal wordt de verkregen tekst in een lokale database bijgehouden om er vervolgens analyses op te maken. Ook het manuele proces van teksten kopiëren en plakken kan gezien worden als *web scraping*, maar aangezien dat niet schaalbaar is, zal de focus van dit onderzoek liggen op de verschillende geautomatiseerde processen (7). Bovendien is er bewezen dat de data veel gedetailleerder, accurater en consistenter is wanneer je geautomatiseerde technieken gebruikt, vergeleken met de manuele methode (34).

Web crawling kan voorafgaan aan *web scraping*. Het doel van *web crawling* is om URL's te ontdekken die later gescraped kunnen worden (7). *Web crawling* is altijd een geautomatiseerd proces dat gebruik maakt van een bot om verschillende websites of webpagina's te scannen. Zulke bots worden ook wel *web-spiders* genoemd (34). Het doel van *web crawling* in dit onderzoek, is om verschillende relevante webpagina's omtrent duurzaamheid en industrie te verzamelen om ze later te gaan *scrapen* en analyseren.

Web scraping en *web crawling* zijn onderdelen binnen een groter proces, genaamd *web mining*. *Web mining* is de combinatie van het verkrijgen en analyseren van gegevens die op het web staan (34). Een toepassing van *web mining* die relevant is binnen dit onderzoek, is *web content mining*. *Web content* is de verzamelnaam voor webpagina's die verschillende datatypes bevatten, zoals tekst, foto's en videos (34). Een webpagina is een ongestructureerd datatype in formaat HTML. Na het scrapen van die webpagina is het de bedoeling dat die gestructureerd wordt opgeslagen in bijvoorbeeld csv, xls, SQL of XML-formaten (38). In dit onderzoek zal dat gebeuren in een csv-bestand.

Het is van groot belang dat de output van het *web scraping* proces kwalitatief is, aangezien dit een direct effect heeft op de prestatie van het *Natural Language Processing* proces (besproken in sectie 2.2). Dit staat ook wel bekend als het '*garbage in, garbage out*' principe. Uit een empirisch onderzoek van Dumitru et al. (23) blijkt dat de taal van een webpagina invloed heeft op welke tekst-extractie technologie het beste resultaat geeft bij het *Natural Language Processing* proces.

Er bestaan verschillende technieken om aan *web scraping* te doen (40). Het manuele kopiëren-plakken blijft ook een techniek om aan *web scraping* te doen. Die wordt enkel nog gebruikt als geautomatiseerde technieken niet werken op een specifieke website, aangezien de manuele techniek niet schaalbaar is bij grote datasets (45).

Volgens Shete et al. (55) zijn de meest relevante methoden: *HTML-parser*, *Tree-based technique* en *web wrapper*. De *HTML-parser* is snel en werkt in *real-time*. Je kan verschillende datatypes extraheren, waaronder ook tekst. De *HTML-parser* wordt zowel in Python als Java ondersteund (55).

Bij de *tree-based appraoch* worden er verschillende *tags* aan de HTML syntax toegevoegd waardoor er een hiërarchische structuur gevormd wordt. De data kan achteraf geëxtraheerd worden door gebruik te maken van XPath. Xpath is een taal om *nodes* aan te roepen in XML-documenten en kan ook gebruikt worden binnen HTML. Je kan met XPath verwijzen naar een bepaalde locatie in het HTML-document (28). De *tree-based approach* maakt ook gebruik van het *Document Object Model* (DOM). Dat is een standaard om HTML-elementen te verkrijgen, veranderen, toe te voegen of te verwijderen. Alle HTML-elementen kunnen daarbij gezien worden als objecten. Er kan dan toegang verkregen worden tot die objecten via programmeertalen (28).

De *web wrapper* (55) wordt gebruikt om de ongestructureerde HTML-data om te vormen naar gestructureerde data. Het kan gezien worden als een laag tussen de ruwe HTML-code en de gegevens die je wil extraheren. Het is ontworpen om door de HTML-structuur te navigeren en de relevante elementen te identificeren. Bij de *web wrapper* kan de gebruiker in een eenvoudige *interface* aangeven wat die zoekt. Vervolgens maakt het systeem automatisch die *wrapper*. Dat zorgt ervoor dat niet enkel mensen met specifieke kennis van programmeren aan *web scraping* kunnen doen, maar ook mensen die hier geen kennis van hebben (55).

De traditionele methoden die hierboven vermeld staan, hebben een echter beperkte *crawling*-strategie. Een grote tekortkoming is bijvoorbeeld dat ze niet in staat zijn om *focused-based* te gaan *crawlen*. Hierdoor verspillen ze veel tijd, aangezien ze bijvoorbeeld ook advertenties verzamelen (4). *WebCollectives* (4) is een in 2023 ontwikkelde content extractor in Java die hier een oplossing voor biedt. Het gaat effectief doorheen de webpagina's en het extractie-proces is gebaseerd op hiërarchie. Het grote voordeel is dat er minder regels code vereist zijn om toch een krachtige analyse uit te voeren vanuit het standaard DOM-model. *WebCollectives* werkt efficiënter dan het traditionele DOM-model, voornamelijk in termen van snelheid. Daarnaast is het ook minder gevoelig aan veranderingen van het domein en het is ook niet afhankelijk van het aantal HTML-elementen (in tegenstelling tot het standaard DOM-model) (4).

Er zijn reeds verschillende toepassingen van *web scraping* onderzocht, waaronder ook toepassingen binnen sociale media en marketing waarbij er achterhaald wordt wat de koopintenties van bepaalde *leads* zijn (40; 13; 67; 15). Er is echter geen onderzoek gevonden dat dezelfde *use-case* als deze thesis behandeld.

De methode voor *web scraping* die in dit onderzoek gebruikt zal worden, is de *HTML-parser*. De reden hiervoor is omdat die methode snel en in *real-time* werkt, maar ook omdat er voor Python werd gekozen als programmeertaal voor het algoritme (55). *WebCollectives* (4) is enkel beschikbaar in Java.

2.2. Natural Language Processing

Natural language processing is een toepassing van AI die gebruikt wordt om menselijke taal te begrijpen en te analyseren (33). De laatste jaren is de hoeveelheid van online data en de beschikbaarheid van nieuwe IT-toepassingen exponentieel toegenomen. Dat zorgt ervoor dat er nieuwe onderzoeksdomeinen ontstonden, waaronder 'tekstclassificatie' (22). Het doel van tekstclassificatie is om vooraf gedefinieerde labels toe te kennen aan nieuwe tekstbronnen. Het wordt gebruikt in verschillende toepassingen, zoals: sentimentanalyse, websiteclassificatie en spamdetectie (57).

In dit onderzoek zullen *Natural Language Processing* technieken gebruikt worden op de tekstuele output van de *web scraping* resultaten. De resultaten daarvan zullen dan gebruikt kunnen worden als input voor het laatste subproces, namelijk het segmenteren van de verschillende klanten. Dat proces zal besproken worden in sectie 2.3.

Traditionele tekstclassificatie kan gezien worden als een *flow* die verschillende processen moet doorlopen, namelijk: *feature extraction*, datareductie, classificatie en evaluatie (65). Bij de recentere technieken zoals Large Language Models (LLMs), is dit proces niet meer van toepassing. Hierbij zijn de belangrijke aspecten namelijk: *pre-training*, *adaptation tuning*, *utilization*, and *capacity evaluation* (66). Het is echter wel interessant om te bespreken hoe de traditionele technieken achterliggend werken, om vervolgens de link te leggen met de meer *state-of-the-art* technieken. Daarom worden in de volgende subsecties de verschillende processen van de traditionele tekstclassificatietechnieken in meer detail toegelicht.

2.2.1. Feature extraction

Het eerste proces dat besproken zal worden, is *feature extraction*. *Feature extraction* houdt in dat relevante informatie geïdentificeerd en geëxtraheerd wordt uit de ruwe tekstuele gegevens. Het ultieme doel is om tekst om te zetten in een voor de computer interpreteerbaar formaat (57). *Word embedding* is een veelgebruikte techniek binnen dit subproces dat woorden gaat mappen op een X-dimensionele vector door ze om te zetten in getallen door middel van een neuraal netwerk. Deze techniek is vernieuwend omdat het ook de semantische informatie meeneemt. Het model neemt context uit de trainingdataset mee om de gewichten voor de *embeddings* te optimaliseren. Oudere technieken zoals het n-gram model kon geen betekenisvolle verbanden leggen tussen woorden onderling (65). *Word embedding* zal in dit onderzoek voornamelijk gebruikt worden om de meest relevante webpagina's van een website te vinden tijdens het proces van *web crawling*. Er zullen woordvectoren gemaakt worden van belangrijke *keywords* die een link hebben met duurzaamheid

en industrie. Aan de hand van die woordvectoren kunnen de verschillende webpagina's gesorteerd worden naarmate relevantie.

Binnen dat domein van *word embedding* zijn er twee belangrijke grondleggers: *Word2Vec* (42) en de opvolger, genaamd '*Global Vectors for Word Representation*' (GloVe) (48). Die technieken werden al succesvol toegepast binnen verschillende *deep learning* technieken en werden later geoptimaliseerd naarmate er nieuwe innovaties tevoorschijn kwamen (65).

Word2Vec is een model ontwikkeld door Tomas Mikolov (42) om woorden uit een grote dataset weer te geven in een vectorruimte, met als hoofddoel te bepalen hoe gelijkaardig de woorden zijn ten opzichte van elkaar. Mikolov stelt twee alternatieven voor om dit te doen: *Continuous Bag-Of-Words* (CBOW) en *Skip-gram*. CBOW gebruikt de n omringende woorden om het middelste woord te voorspellen. *Skip-gram* gebruikt het woord in het midden om de n omringende woorden te voorspellen. Beide alternatieven maken gebruik van *backward propagation*. Het model scoort heel hoog op vlak van syntactische en semantische gelijkenissen (42).

Word2Vec richt zich uitsluitend op lokale woorden en mist daardoor een globale context. Dit wordt opgelost door de opvolger, GloVe. GloVe maakt gebruik van globale informatie die zich uitstrekt over verschillende zinnen in een document. Het beginpunt van GloVe is een uitgebreide matrix die weergeeft hoe vaak ieder woord voorkomt met ieder ander woord. Vervolgens worden de vectoren voor ieder woord gevormd, gebaseerd op de kansen dat ze samen voorkomen. Uit onderzoek van Pennington et al. (48) blijkt dat GloVe een betere *accuracy* op *word analogy* taken heeft ten opzichte van CBOW en *Skip-gram*. Bij zulke *word analogy* taken gaat het over de onderlinge relatie tussen woorden. Neem bijvoorbeeld 'a behoort tot b, zoals c behoort tot ...', waarbij a, b en c gegeven zijn en '...' door het model wordt aangevuld. Daarnaast werd de *accuracy* ook vergeleken op basis van *Word similarity* en *Named entity recognition*. Ook hier presteert GloVe beter dan andere modellen (48).

Word2Vec en GloVe zijn voorbeelden van statische methoden. Ze genereren woordvectoren gebaseerd op het samen voorkomen van woorden binnen bepaalde grenzen en eenmaal getraind, blijven deze *embeddings* constant. Dynamische methoden daarentegen kunnen veranderen tijdens de uitvoering van een specifieke taak of training. De woordvectoren worden bijgewerkt op basis van de context van de huidige taak (57).

In 2023 presenteerden Sun et al. (57) een verbeterde versie van de klassieke *Word2Vec* waarmee tekortkomingen werden opgelost. Dat nieuwe model houdt rekening met de frequentie van woorden over verschillende documenten heen om te bepalen hoe belangrijk deze woorden zijn. Bovendien kijkt het, net als GloVe, naar een meer globale context. Ten slotte worden minder belangrijke woorden eruit gefilterd door de gewichten ervan op nul te zetten. Dit resulteert in een snellere training, vergeleken met de meeste *state-of-the-art* classificatie modellen zoals bijvoorbeeld BERT, die in sectie 2.2.3 besproken wordt.

2.2.2. Data reductie

Om de complexiteit van de geëxtraheerde *feautures* te beperken, wordt data reductie toegepast. Dit zorgt ervoor dat je de efficiëntie van je model kan verbeteren. Het vermindert niet alleen de kans op *overfitting*, maar zorgt er ook voor dat je model sneller getrained wordt. Twee veelgebruikte technieken zijn *Principal Component Analysis* (PCA) en *Linear Discriminant Analysis* (LDA). Die data reductie gebeurt op de vectoren die reeds gevormd zijn (36).

PCA is een *unsupervised* methode om de dimensionaliteit van de dataset te verminderen, zonder belangrijke data te verliezen. Concreet gebeurt dit door de oorspronkelijke n-dimensionale kenmerken om te zetten naar lagere k-dimensionale kenmerken. Dit resulteert in een nieuw coördinatensysteem dat de variabiliteit van de gegevens maximaliseert waardoor de belangrijkste kenmerken geïdentificeerd kunnen worden. Die gereduceerde dimensie wordt verkregen door de eigenwaarden en eigenvectoren van de covariantiematrix te berekenen en vervolgens te vermenigvuldigen met de oorspronkelijke gegevensmatrix (65).

In tegenstelling tot PCA is LDA een *supervised* methode met als doel laag-dimensionale data te projecteren waarbij de afstand tussen de verschillende klassen zo groot mogelijk is. LDA zoekt naar lineaire combinaties van kenmerken die ervoor zorgen dat de verschillende klassen zo ver mogelijk uit elkaar liggen, terwijl de spreiding binnen elke klasse geminimaliseerd wordt. Hierdoor ontstaan nieuwe kenmerken die de data efficiënter classificeren. Deze lineaire combinaties worden vervolgens gebruikt als input voor het trainen van je model (65).

2.2.3. Classificatie

Er is reeds veel onderzoek gedaan naar de toepassing van verschillende algemene *classificatie* technieken binnen tekstclassificatie. Zo werden in een literatuurstudie uit 2023 (43) negentien verschillende modellen met elkaar vergeleken op basis van het aantal vermeldingen in papers en hun *accuracy*. Een hoge *accuracy* geeft aan dat het model in staat is betere relaties te leggen tussen de inputwaardes, waarna het de toekomstige waarden beter kan classificeren. Belangrijk om te vermelden is dat dit geen *classifiers* zijn die specifiek ontworpen zijn voor tekst, dus de resultaten zijn afhankelijk van de kwaliteit van de vorige stappen in het proces (43). In tabel 1 worden de, volgens de studie, zes meest geciteerde technieken met elkaar vergeleken.

Wat opvalt, is dat het SVM-model het meest wordt aangehaald in papers en ook de hoogste *accuracy* heeft. *Logistic Regression* heeft de op twee na hoogste *accuracy*, maar wordt het minst vermeld in papers, vergeleken met de vijf andere modellen. Het model met de laagste *accuracy* is de *Random Forest*, met 92,60% (43).

Model	Aantal papers	Accuracy
Support Vector Machine (SVM)	118	98.88
Naive Bayes (NB)	92	97.89
k Nearest Neighbor (kNN)	66	97.89
Random Forest (RF)	42	92.60
Decision Tree (DT)	34	94.50
Logistic Regression (LR)	23	98.50

Tabel 1: tekstclassificatie technieken (43)

In de laatste jaren is het aantal complexe documenten toegenomen, wat ervoor zorgt dat de traditionele methodes van tekstclassificatie niet meer volstaan. Hoewel bijvoorbeeld de SVM ook werkt bij complexere non-lineaire verdelingen, kan het niet op tegen recentere *deep leaning* modellen die nog beter presteren bij complexe documenten (65). De meest voorkomende *deep learning* modellen binnen tekst classificatie zijn: *Long Short Term Memory* (LSTM), *Gated Recurrent Unit* (GRU) en *Transformers*. Andere recente modellen zijn: *Embeddings from Language Models* (ELMo), *Generative Pre-Training* (GPT), *Bi-directional Encoder Representation from Transformer* (BERT) en XL-Net (65). In wat volgt, zullen de verschillende modellen verder toegelicht worden.

LSTM (30) is een type neuraal netwerk dat lange tekstsequenties kan onthouden en dus ook gebruiken. LSTMs kunnen effectief zijn bij tekstclassificatie, aangezien de volgorde van woorden vaak cruciaal is voor de betekenis van die woorden. Zo een LSTM-netwerk heeft geheugencellen ter beschikking om informatie voor een langere periode op te slaan (63; 54).

GRU (16) is een opvolger van LSTM (17). GRUs hebben minder parameters en zijn daardoor minder complex, wat kan resulteren in een efficiëntere training en minder kans op *overfitting*, vooral bij beperkte datasets. Ze hebben een *reset gate* en een *update gate* en hebben daardoor maar één geheugen nodig. Deze *gates* stellen het model in staat om beter te bepalen welke informatie moet worden bijgewerkt en welke moet worden vergeten (65).

Een *transformer* is een specifiek ontworpen architectuur om sequentiële gegevens zoals tekst te verwerken. Voorbeelden hiervan zijn GPT en BERT. *Transformers* verschillend ten opzichte van LSTM en GRU dankzij hun zogenaamde '*Self-Attention Mechanism*' om relaties tussen verschillende delen van de invoer-sequenties te begrijpen. Hierdoor kan een *transformer* informatie vastleggen over lange afstanden en complexe relaties tussen woorden beter begrijpen dan de traditionele neurale netwerken. Het trainen van *transformers* gaat ook sneller vergeleken met LSTMs en GRUs, aangezien ze parallel kunnen werken (24; 39). Om die reden is er in dit onderzoek gekozen voor de *transformer* genaamd 'GPT'. In de volgende alineas worden de technieken EMLo, GPT, BERT en XL-net nog verder toegelicht zodat de verschillen, de gelijkenissen en de algemene evolutie duidelijk is.

ELMo is gebaseerd op LSTM, maar werkt in twee richtingen en heeft bijgevolg meer context ter beschikking. In plaats van een vaste vector te hebben voor elk woord, genereert ELMo verschillende *embeddings* voor hetzelfde woord op basis van de context waarin het gebruikt wordt. Ze zijn dus beter dan de traditionele LSTMs en GRUs, maar niet zo geavanceerd als de transformer-modellen, aangezien ze dat *Self-Attention Mechanism* niet hebben (8; 65; 24).

GPT is een vooraf getraind model dat gebruik maakt van de transformer-architectuur. In tegenstelling tot ELMo werkt GPT, maar in één richting, maar dankzij het *Self-Attention Mechanism* zijn de resultaten toch beter dan ELMo en andere modellen die gebaseerd zijn op de LSTM-architectuur. Het werkt beter voor langere teksten en generatieve taken zoals: 'Geef een samenvatting van deze tekst.' (26). BERT is net zoals GPT een transformer-architectuur, maar werkt wel in twee richtingen en er worden twee *pre-training* taken tegelijkertijd uitgevoerd: *Masked Language Model* (gemaskeerde woorden in een zin voorspellen, zodat context in twee richtingen kan worden begrepen) en *Next Sentence Prediction* (bepalen of twee zinnen elkaar opvolgen in de oorspronkelijke tekst) (21).

XL-Net combineert zowel de sterke punten van GPT als van BERT. Een belangrijke limitatie van GPT is dat het maar in één richting werkt en dat het dus geen beeld heeft van de woorden voor én na een woord (65). De grote limitatie van BERT daarentegen is het feit dat BERT tijdens de training willekeurig woorden maskeert en deze woorden vervolgens probeert te voorspellen. Maar tijdens de *fine-tuning* worden er geen woorden gemaskeerd. Dit verschil kan ervoor zorgen dat de prestatie van het model slechter is. Bovendien gaat BERT ervan uit dat de gemaskeerde woorden in een zin onafhankelijk zijn van elkaar, wat in werkelijkheid niet altijd klopt. Daardoor zijn er soms relaties tussen de gemaskeerde woorden die BERT niet kan herkennen (65).

XL-Net gaat deze tekortkomingen aanpakken door te erkennen dat er wel relaties tussen woorden kunnen bestaan. Hierdoor kan het niet alleen individuele woorden in een zin voorspellen, maar ook de volgorde van woorden begrijpen (65). Om dat te doen, wordt er een context gecreëerd. Neem bijvoorbeeld een zin dat bestaat uit i=9 woorden (voorgesteld als 'x'): x1, x2, x3, x4, x5, x6, x7, x8, x9. Het doel kan bijvoorbeeld zijn om het vijfde woord te voorspellen door de eerste vier woorden te geven. Maar om ook de context na een woord mee te nemen, wordt er vervolgens een random permutatie uitgevoerd op de oorspronkelijke reeks en toegevoegd aan de training. De nieuwe volgorde is nu bijvoorbeeld x4, x2, x3, x1, x5, x6, x8, x9, x7. Door het model op deze manier te trainen met verschillende permutaties van woorden in een zin, kan het model de context rond een specifiek woord beter begrijpen. Het uiteindelijke doel is om het model in staat te stellen zowel de woorden

voor als na een doelwoord in een zin te begrijpen, wat nuttig is voor verschillende NLP-taken (65).

2.3. Customer Segmentation

Er is een sterk verband tussen de begrippen '*customer segmentation*', '*customer clustering*' en '*customer profiling*'. *Customer segmentation* is een methode om verschillende klanten die je hebt te groeperen aan de hand van verschillen en gelijkenissen (61). Het doel hiervan is om vervolgens de verschillende segmenten te benaderen op hun eigen manier, aangezien ze elk hun eigen behoeften hebben. Binnen elk klantensegment zitten dus individuele klanten met gemeenschappelijke waarden, normen en behoeften. Elk klantensegment heeft daarom ook een unieke marketingstrategie (61). Om die verschillende klantensegmenten vanuit data te ontdekken, kan je *customer clustering* toepassen (12). Dat is een datamininingtechniek waarmee je gelijkaardige datapunten kan groeperen op basis van hun eigenschappen en kenmerken. Het is een vorm van *unsupervised learning* waarbij het algoritme patronen in de data identificeert op basis van gelijkenissen en verschillen, zonder op voorhand de data te labelen. Je kan verschillende datatypes clusteren (12).

Websites van bedrijven zijn ontworpen om informatie te delen, dus deze informatie kan gebruikt worden als input om verschillende klanten via hun website te clusteren. Wanneer je die verschillende clusters geïdentificeerd hebt, kan je elke cluster een profiel gaan geven. Dat noemt men dan '*customer profiling*' (15). Deze verschillende 'profielen' zijn vervolgens nuttig om de beste klantensegmenten te identificeren.

Bestaande klanten opdelen in verschillende segmenten kan belangrijk zijn om te bepalen hoe verschillende soorten klanten het best benaderd worden. Het geeft een duidelijk beeld van de voorkeuren en de noden van de verschillende soorten klanten. Hierdoor kan je als bedrijf geïnformeerde beslissingen nemen die de klanttevredenheid zullen vergroten (12). Enerzijds houdt dat in dat je specifieker kan gaan *targetten* op nieuwe klanten waarvan de kernwaarden overeenkomen met jouw product of dienst. Stel dat je een bedrijf hebt dat focust op de verkoop van groene energie. Wanneer geweten is dat één van je klantensegmenten actief wil inzetten op het gebruik van duurzame grondstoffen en energiebronnen, zal het verkopen van jouw groene energie waarschijnlijk gemakkelijker zijn bij die groep bedrijven, vergeleken met bedrijven die hier geen belang aan hechten. Anderzijds kan je als bedrijf je bestaande product gaan aanpassen op de kernwaarden van je klanten. Het is ook belangrijk om te beseffen dat de kernwaarden van klanten evolueren doorheen de tijd. Dat zorgt ervoor dat het profileren en segmenteren van je klanten een continu proces is (11).

Paschen et al. (46) hebben onderzoek gevoerd naar hoe AI kan helpen bij het verkrijgen van kennis in de B2B markt. AI-systemen maken hierbij gebruik van zes fundamentele bouwstenen: gestructureerde gegevens, ongestructureerde gegevens, voorbewerkingen, hoofdprocessen, kennisbank en informatie. Deze bouwstenen gaan gegevens omzetten in informatie en kennis, die nuttig kunnen zijn voor B2B-marketingdoeleinden (46). Gestructureerde en ongestructureerde gegevens worden in dit proces gebruikt als input, waarna de voorbewerking plaatsvindt. In dat proces worden de gegevens opgeschoond, eventueel genormaliseerd en getransformeerd zodat een AI-algoritme deze gegevens kan gebruiken. Vervolgens kan er door AI geredeneerd worden door bijvoorbeeld gebruik te maken van *machine learning* om hier informatie uit te halen. Die informatie wordt vervolgens opgeslagen in een kennisbank, waarna die kunnen worden omgezet in waardevolle inzichten (46).

Omdat we leven in een datagedreven wereld waar gegevens publiek beschikbaar zijn op het internet, is het verkrijgen van informatie over bedrijven gemakkelijker geworden (5). De vraag naar toegang hiertoe stijgt ook. Echter kost het veel tijd en het is ook moeilijk om inzichten uit ongestructureerde gegevens manueel te verkrijgen. Dit is waar *machine learning* aan te pas komt om het proces te automatiseren en te versnellen (32). Door middel van *web mining* kunnen er taken zoals clustering en classificatie uitgevoerd worden op grote schaal (47). Volgens Al-Azmi et al. (5) zijn er vier belangrijke fasen in dit proces. Ten eerste moeten er gegevens worden opgehaald door op zoek te gaan op het internet naar welke gegevens relevant kunnen zijn. Vervolgens worden die gegevens geëxtraheerd om daarna via generalisatie bepaalde patronen te gaan ontdekken. Hier kunnen tot slot analyses worden op uitgevoerd (5). Binnen *web mining* worden er naast clustering en classificatie ook andere taken uitgevoerd, zoals het samenvatten van teksten, het maken van associatieregels en het visualiseren van informatie (37). Informatie-extractie en categorisatie helpt vooral bij het efficiënt ophalen van relevante informatie. Clustering en associatieregels zijn vooral nuttig bij het identificeren van klantsegmenten (37).

3. Methodologie

3.1. Literatuurstudie

Het startpunt van dit onderzoek is een uitgebreide literatuurstudie. Na een exploratieve fase werden nieuwe artikels gezocht en gelezen aan de hand van *citation chaining*. Zowel *forward reference searching* als *backward reference searching* werd toegepast om een beter beeld te krijgen van de beschikbare literatuur (51). De online tool 'Litmaps' (1) werd hiervoor gebruikt. Litmaps vergemakkelijkt het vinden van nieuwe artikels via *citation chaining* door het onmiddellijk tonen van de abstracten en geeft ook de onderliggende relaties van de artikels visueel weer.

Op het einde van de exploratieve fase werd het duidelijk dat dit onderzoek een combinatie zal zijn van drie grote domeinen:

- *Web scraping*: het verkrijgen van website-inhoud, startende van de URL.

- *Natural Language Processing*: Verschillende tekstherkenningstechnieken vergelijken.
- *Customer Segmentation*: Verschillende manieren entiteiten profileren en segmenteren op basis van verschillen en gelijkenissen.

Er werd gebruik gemaakt van de ScienceDirect en Google Scholar voor het zoeken van nieuwe bronnen. De zoektermen die gebruikt werden, staan opgelijst in tabel 2.

Domein	**Zoekterm**	**Aantal**
Web scraping	('Web scraping') AND (('web content scraping' OR 'web page scraping' OR 'extraction') AND ('scraping methods' OR 'extraction techniques') AND (('URL' OR 'HTML parsing') OR 'text extraction from websites'))	465
Natural Language Processing	('Natural Language Processing') AND ('text classification techniques' OR 'text labeling methods') AND ('latest advancements' OR 'state-of-the-art' OR 'recent developments') AND ("word embedding")	412
Customer Segmentation	(('Customer segmentation' OR 'market segmentation') AND ('KNN' OR 'decision tree' OR 'machine learning' OR 'clustering techniques')) AND ('web content mining' OR 'online profiling' OR 'website content')	483

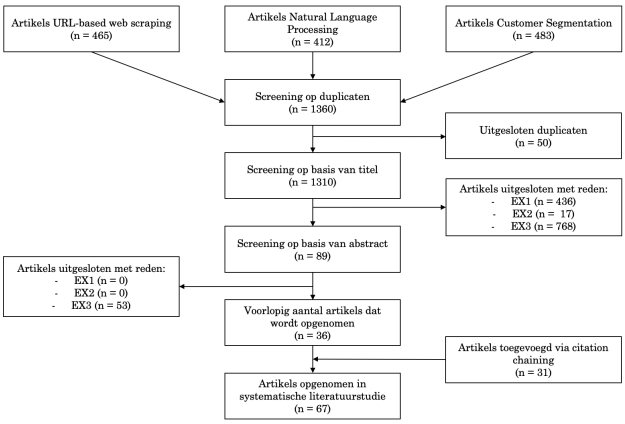
Tabel 2: Zoektermen per domein

Die gebruikte zoektermen leveren 1360 resultaten op. Na het filteren op duplicate artikels, bleven er nog 1310 artikels over. Vervolgens worden de artikels gescreend op basis van de vooropgestelde inclusie- en exclusiecriteria, te vinden in tabel 3.

Inclusiecriteria	**Exclusiecriteria**
IN 1: het artikel is peer-reviewed.	EX 1: Het is geen artikel in een wetenschappelijk tijdschrift.
IN 2: er is toegang tot de volledige tekst van het artikel.	EX 2: het artikel is niet Engelstalig.
	EX 3: Het artikel focust niet op één van de drie domeinen (URL- based web scraping, Natural Language Processing, Customer Segmentation).

Tabel 3: Inclusie- en exclusiecriteria

De artikels worden eerst gescreend op basis van de titel en indien er twijfel is op basis van het abstract. Na de selectie werden er 36 artikels opgenomen in deze literatuurstudie. Ten slotte zal *citation chaining* ook worden toegepast. Dat volledige proces leidt tot de uiteindelijke verzameling van 67 papers die in deze literatuurstudie zijn opgenomen zoals weergegeven in figuur 1.



Figuur 1: *Prisma Flow Diagram*

3.2. Algoritme

De kern van dit onderzoek omvat het bouwen van een geautomatiseerd algoritme dat, startende van een csv-bestand met 250 URL's, een analyse uitvoert en de resultaten teruggeeft in een nieuw csv-bestand. Het hoofddoel van dit algoritme is om data te verzamelen betreffende de industrie en de duurzaamheidsprofilering van ieder bedrijf. De 250 bedrijven zijn verdeeld over vijf industrieën: technology, gezondheidszorg, automotive, energie en retail. Elke industrie telt 50 bedrijven. Het is voor 249 van de 250 bedrijven gelukt om resultaten terug te krijgen. Er ontbreekt één bedrijf uit de healthcare sector in de output. Mogelijke verklaringen zijn netwerk *time-outs* of *request failures*.

Er werden drie analyses uitgevoerd:

- **4.1 Herkennen van de industrie**: De evaluatie van het vermogen van het AI-model om correct de industrieën te identificeren.
- **4.2 Ranking op duurzaamheid**: De evaluatie van het vermogen van het AI-model om de bedrijven een score te geven op basis van hun duurzaamheidsprofilering. In deze analyse zal ook clustering worden toegepast om te achterhalen of er verschillen en gelijkenissen in duurzaamheidsprofilering zijn.
- **4.3 Relatie tussen industrie en duurzaamheid**: De relatie tussen de industrie van een bedrijf en haar duurzaamheidsprofilering.

Het geautomatiseerde algoritme dat ontwikkeld werd, bestaat uit drie subalgoritmen. Het eerste algoritme 'scrape_websites.py' ontdekt relevante subpagina's van

websites en gaat die vervolgens *scrapen*. Algoritme 2, 'analyze_websites.py', gaat die output gebruiken om de bedrijven te classificeren per industrie en om een duurzaamheidsscore tussen 1 en 5 toe te kennen. Tot slot gaat algoritme 3, 'cluster_websites.py', een clustering op de duurzaamheidsteksten uitvoeren. De pseudocodes zijn terug te vinden in 'Algorithm 1', 'Algorithm 2', en 'Algorithm 3'. [1]

Algorithm 1 scrape_websites.py

Input: `websites_to_scrape.csv`
Output: `scraped_websites.csv`

Procedure ScrapeWebsites(file_path, output_path)
Definieer keywords en zet om naar woordvectoren
Lees input CSV-bestand uit
for elke rij in dataframe **do**
    Haal URL op uit rij
    Maak een HTTP-GET request naar de URL
    Vind URL's van alle subpagina's
    Vind drie meest relevante pagina's
    for elke gefilterde link **do**
        Maak een HTTP-GET request naar de link
        Scrape inhoud met BeautifulSoup
        Voeg pagina_tekst toe aan combined_text
    end for
    if combined_text leeg of onvoldoende is **then**
        Gebruik Selenium om extra inhoud te scrapen
        Voeg resultaat toe aan combined_text
    end if
    Voeg combined_text toe aan combined_texts lijst
end for
Schrijf combined_texts naar output CSV-bestand

Algorithm 2 analyze_websites.py

Input: `scraped_websites.csv`
Output: `analyzed_websites.csv`

Procedure AnalyzeWebsites(input_path, output_path)
Laad OpenAI API-key
Lees input CSV-bestand uit
for elke rij in dataframe **do**
    Haal URL en scraped content op uit rij
    Analyseer duurzaamheid met context
    Analyseer duurzaamheid zonder context
    Voorspel industrie met context
    Voorspel industrie zonder context
    Maak resultatenwoordenboek aan met URL, duurzaamheidsscores en industrievoorspellingen
    Voeg resultaat toe aan resultatenlijst
end for
Schrijf resultaten naar output CSV-bestand

[1] De originele broncode is beschikbaar op: https://github.com/MaximRiebus/thesis

Algorithm 3 cluster_websites.py

Input: `analyzed_websites.csv`
Output: `clustered_websites.csv`

Procedure ClusterWebsites(input_path, output_path)
Laad dataset
Laad sentence transformer model
Encodeer zinnen naar embeddings
Voer K-means clustering uit
Voeg clusterlabels toe aan dataset
Schrijf geclusterde data naar output CSV-bestand

De programmeertaal die gebruikt werd om het outputbestand te genereren, is Python (64). Om relevante pagina's met betrekking tot duurzaamheid en de industrie te ontdekken, startende van enkel de homepagina, werden de *libraries* 'Requests' (14), 'BeautifulSoup (bs4)' (50), 'Spacy' (31) en 'Numpy' (29) gebruikt. Requests werd gebruikt om HTTP-verzoeken te maken waarna BeautifulSoup de HTML van die webpagina's gaat doorzoeken. De output hiervan zijn alle beschikbare subpagina's van die website. Om vervolgens de drie belangrijkste subpagina's over te houden, werd Spacy gebruikt om vectorrepresentaties van de teksten achter de links te berekenen. Vervolgens worden die vector-representaties gebruikt om met Numpy (29) de *cosine similarity* te berekenen met vooraf gedefinieerde sleutelwoorden. Op basis van deze gelijkenis worden de drie relevante links overgehouden en opgeslagen in de variabelen 'sustainability_relevant_urls' en 'industry_relevant_URLs'.

Om vervolgens relevante informatie over duurzaamheid en de industrie van bedrijven te verzamelen, werd *web scraping* toegepast. Dit gebeurt in eerste instantie met BeautifulSoup (50), wat goed werkt voor statische pagina's, maar wanneer er geen output gegenereerd werd door BeautifulSoup, werd 'Selenium' (9) gebruikt. Dat is nodig omdat het dan gaat om een dynamische webpagina, die op een andere manier gescrapet moet worden. Deze combinatie van tools stelt ons in staat om zowel statische als dynamische inhoud van de websites te verzamelen. De tekst die gescrapet is, werd opgeslagen in de variabelen 'scraped_text_for_sustainability' en 'scraped_text_for_industry'. Voor 229 bedrijven van de 249 bedrijven in het output-bestand is dit gelukt aan de hand van het algoritme. Met andere woorden, voor 20 bedrijven is het niet gelukt om waardevolle inhoud van de website te extraheren op de geautomatiseerde manier. Om deze bedrijven toch te kunnen opnemen in verdere analyses, werd die inhoud manueel gescrapet.

Op basis van de verkregen teksten werd vervolgens met behulp van de Open AI API (2) enerzijds de industrie voorspeld en anderzijds een duurzaamheidsscore toegekend op een schaal van 1 tot 5. Het gebruikte model is gpt-3.5-turbo-0125, het meest recente GPT-3.5 Turbo model dat beschikbaar was tijdens het uitvoeren van de studie. De training data is tot en met september 2021 en

het model heeft een context window van 16.385 tokens (2). De resultaten werden respectievelijk opgeslagen in de variabelen 'predicted_industry_with_context' en 'sustainability_score_with_context'. Om achteraf de robuustheid van het model te meten, werd er onderzocht of er afwijkingen ontstaan wanneer het model meerdere keren een score moet geven aan elk bedrijf, op basis van dezelfde input. Daarom worden er voor elk bedrijf vijf verschillende antwoorden gegenereerd en opgeslagen voor de 'predicted_industry_with_context'-variant. De resultaten hiervan staan in de variabelen 'predicted_industry_with_context_1', 'predicted_industry_with_context_2', 'predicted_industry_with_context_3', 'predicted_industry_with_context_4' en 'predicted_industry_with_context_5'. De score met de hoogste frequentie werd dan gekozen als waarde voor de 'sustainability_score_with_context'-variabele.

Aangezien de verkregen tekst beperkt kan zijn en niet alle relevante informatie van dat bedrijf kan omvatten, werd dezelfde oefening ook gedaan met een andere prompt waarin enkel de URL van de homepagina werd meegegeven als input. Omdat het model getraind is op een grote hoeveelheid data beschikbaar op het internet, is het mogelijk dat het meer context meekrijgt en dus een beter resultaat geeft. Het is wel een typisch voorbeeld van een '*black box*' waarbij het niet mogelijk is om te weten op basis van welke input de score werd gegenereerd. Mogelijke verbetervoorstellen worden later gedeeld in sectie 5. Deze aanpak werd zowel toegepast op de industrie-classificatie als op de toekenning van de duurzaamheidsscores. De resultaten werden respectievelijk opgeslagen in de variabelen 'predicted_industry_without_context' en 'sustainability_score_without_context'.

Vervolgens werden de teksten opgeslagen in de scraped_text_for_sustainability variabele ook geclusterd. Op die manier kan er achterhaald worden of er verschillen of gelijkenissen zijn in de manier waarop bedrijven zich profileren naar duurzaamheid toe. Om de semantische betekenis van de tekst te begrijpen, wat belangrijk is bij deze clustering taak, werd in Python de 'SentenceTransformer' (49) library gebruikt. Specifiek werd er gebruik gemaakt van het 'all-MiniLM-L6-v2'-model (3) dat de teksten omvormt in numerieke vectoren. Vervolgens werd K-means clustering toegepast uit de 'sklearn' library. Het aantal clusters werd vastgelegd op acht na een analyse op basis van de Elbow-methode (41).

Tot slot werd er ook een analyse uitgevoerd om de relatie tussen de effectieve industrie en de voorspelde AI-scores op basis van de gescrapete tekst te achterhalen. De resultaten van deze sectie representeren slechts 249 bedrijven, wat niet voldoende is om conclusies te trekken op globale schaal.

Tabel 4 geeft de variabelen van het output-bestand '`clustered_websites.csv`' weer, gegenereerd door het algoritme.

Kolomnaam	**Uitleg**
url	URL van de homepagina
sustainability_relevant_urls	3 relevante webpagina's met betrekking tot duurzaamheid
industry_relevant_URLs	3 relevante webpagina's met betrekking tot de industrie
scraped_text_for _sustainability	gescrapete tekst voor duurzaamheidsdoeleinden
scraped_text_for_industry	gescrapete tekst voor industriedoeleinden
sustainability_score_with_ context_1	eerste duurzaamheidsscore op basis van gescrapete tekst
sustainability_score_with_ context_2	tweede duurzaamheidsscore op basis van gescrapete tekst
sustainability_score_with_ context_3	derde duurzaamheidsscore op basis van gescrapete tekst
sustainability_score_with_ context_4	vierde duurzaamheidsscore op basis van gescrapete tekst
sustainability_score_with_ context_5	vijfde duurzaamheidsscore op basis van gescrapete tekst
sustainability_score_with_ context	meest frequente duurzaamheidsscore op basis van gescrapete tekst
sustainability_score_without_ context	duurzaamheidsscore op basis van bedrijfsnaam
predicted_industry_with_ context	voorspelde industrie op basis van gescrapete tekst
predicted_industry_without_ context	voorspelde industrie op basis van bedrijfsnaam
cluster	Een cijfer tussen 0 en 7 op basis van hoe gelijkaardig de teksten over duurzaamheid zijn

Tabel 4: Uitleg kolomnamen output-bestand

3.3. data-analyse

Om de resultaten van het algoritme te analyseren, werd gebruik gemaakt van het softwarepakket R (59). In tabel 5 wordt een overzicht gegeven van de gebruikte bibliotheken.

3.4. Metrieken voor validatie

Om de resultaten van deze studie te kunnen valideren, is het belangrijk om een set van metrieken ter beschikking te hebben. In deze sectie zal besproken worden welke metrieken van toepassing zijn in de komende resultatensecties.

Library	Gebruik
readr	inladen van csv-bestand
readxl	inladen van Excel-bestand
reshape2	herstructureren en aggregeren van gegevens
caret	statistische analyses, zoals de *confusion matrix*
ggplot2	grafieken en visualisaties
dplyr	data manipulatie
tidyr	wijzigen van de opmaak van *dataframes*

Tabel 5: Gebruikte bibliotheken RStudio

3.4.1. Herkennen van de industrie

In het eerste deel van de resultaten zal het vermogen van het AI-model om de correcte industrie te identificeren, getest worden. Om de industrie die door het AI-model geclassificeerd wordt te vergelijken met de vooraf manueel gelabelde industrie, wordt de *accuracy* onder andere gebruikt. De *accuracy* is een van de meest gebruikte metrieken voor modelvalidatie en is ook makkelijk te verifiëren (60; 58). Het geeft het percentage correct geclassificeerde gevallen weer over de hele dataset, dus over alle industrieën heen. Een hoge *accuracy* wijst op een goed presterend model. Ook *precision*, *recall* en *F1* zijn veelgebruikte metrieken (52). *Precision* geeft voor elke specifieke industrie weer hoeveel van de bedrijven die als die industrie zijn geclassificeerd, daadwerkelijk in die industrie zitten. *Recall* geeft daarentegen weer hoeveel van de bedrijven die daadwerkelijk in die industrie zitten, correct zijn geclassificeerd. De *F1-score* combineert *precision* en *recall* in één cijfer. Het is vooral nuttig wanneer je een balans wilt vinden tussen hoe nauwkeurig en hoe volledig je model is (27).

Grandini et al. (27) hebben een overzicht gepubliceerd van veelgebruikte metrieken voor *multi-class* classificatie. De berekeningen voor *accuracy, precision, recall en F1* worden respecitvelijk weergegeven in formules 1, 2, 3 en 4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Hierbij staan de volgende symbolen voor:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

Daarnaast werd ook de *confusion matrix* gebruikt als maatstaf om de prestatie van het model te meten. Dit is een kruistabel die het aantal voorvallen tussen de werkelijke classificatie en de door het model gegenereerde classificatie weergeeft (27). De getallen op de diagonaal zijn de gevallen die correct geclassificeerd zijn.

Cohen's Kappa is een metriek die ook gebruikt zal worden om het model te evalueren. Dat is een metriek die gebruikt wordt om de mate van overeenstemming tussen twee beoordelaars te evalueren (27). In dit geval zijn de beoordelaars enerzijds de manuele industrie-*labeling* en anderzijds de classificatie die door het model gegenereerd werd. Ook dit is een metriek die veel gebruikt wordt binnen *multi-class* classificatie.

3.4.2. Ranking op duurzaamheid

Om te onderzoeken wat de verschillen in scores zijn bij het model dat de gescrapete tekst als input heeft, en het model dat enkel de URL van het bedrijf krijgt en zelf beoordeelt op basis van de eigen trainingsdata, werd er gekeken naar de spreiding en naar de Spearman correlatie. De Spearman correlatie kan gebruikt worden voor het vergelijken van verschillende *rankings* (6). Het beoordeelt de sterkte en richting van de twee afzonderlijke beoordelingen. Een hoge positieve Spearman correlatie duidt op een sterke overeenkomst, terwijl een negatieve correlatie duidt op een sterke onenigheid. Deze metriek is specifiek ontworpen voor ordinale data (6), zoals hier het geval is.

Om de prestatie van de scores gegenereerd door het AI-model te meten, werd er gekozen om een *survey* uit te sturen waarbij de respondenten een beperkt aantal teksten van websites moeten rangschikken ten opzichte van elkaar op een schaal van 1 tot 5.[2] In deze schaal wil 1 zeggen dat het bedrijf duurzaamheid heel serieus neemt, terwijl een 5 wil zeggen dat er bijna niets over duurzaamheid vermeld wordt. Merk op dat het AI-model anders redeneert. Bij het AI-model wijst een score van 1 op de assumptie dat het bedrijf geen duurzaamheidsprofilering heeft, en een score van 5 op de assumptie dat het bedrijf haar uiterste best doet om duurzaam te zijn. Om deze punten later met elkaar te kunnen vergelijken, werden de punten die in de enquête gegeven werden, getransformeerd (een score van 1 werd 5, een score van 2 werd 4 ... een score van 5 werd 1). Op die manier is de interpretatie van de punten weer gelijk. De *survey* bestaat uit vijf vragen waarin telkens vijf korte teksten van drie zinnen worden weergegeven. Deze teksten zijn handmatig geselecteerd uit het outputbestand dat is gegenereerd door het 'scrape_websites.py'-algoritme. Hierbij is geen rekening gehouden met de industrie waarin de bedrijven zich bevinden, maar enkel met de scores die door AI zijn gegeven op basis van de gescrapete tekst. Voor elke vraag geldt dat de getoonde teksten een diverse reeks van

[2] De *survey* is terug te vinden in appendix A.1.

gegenereerde AI-scores vertegenwoordigen. Concreet betekent dit dat er bij elke vraag één tekst is met een AI-score van 1, één met een score van 2, één met een score van 3, één met een score van 4, en één met een score van 5. Dit zorgt ervoor dat de ranking van de respondenten beter vergeleken kan worden met de ranking van het algoritme.

Om de robuustheid van het model te meten, werd er onderzocht of er afwijkingen zijn wanneer het model meerdere keren op dezelfde vraag moet antwoorden. In dit onderzoek zal het model ieder bedrijf vijf keer een score geven, op basis van dezelfde input. Deze vijf scores kunnen dan vergeleken worden door te kijken naar de standaarddeviatie. Het model wordt als robuust beschouwd wanneer het meerdere keren hetzelfde antwoord geeft.

Tot slot gebeurt er ook een clustering op de duurzaamheidsteksten die afkomstig zijn van de website. Dit gebeurt op basis van K-Means clustering (56). Dat is over het algemeen de meest bekende en meestgebruikte methode om teksten te clusteren (53). Het is een *unsupervised* techniek dat gebruik maakt van *pattern recognition* (56). Het aantal clusters werd vastgelegd op acht, na het uitvoeren van de Elbow-methode (19).

3.4.3. Relatie tussen industrie en duurzaamheid

Voor de laatste resultatensectie werd de relatie tussen de industrie en de duurzaamheidsprofilering onderzocht. Hierbij ligt de focus vooral op het onderzoeken van de verschillen en gelijkenissen van toegekende scores aan bedrijven, per industrie. Dit werd gedaan door frequenties, spreidingen en gemiddeldes te vergelijken.

Daarnaast werd de ANOVA-test ook uitgevoerd om de variantie verder te onderzoeken. Die test is geschikt voor het vergelijken van resultaten over meerdere groepen, in dit geval de duurzaamheidsscores over verschillende industrieën (35). De F-statistiek binnen de ANOVA-test beoordeelt de grootte van de variantie tussen de gemiddelden. Wanneer de F-waarde die voortkomt uit de test groter is dan de kritische F-waarde, duidt dit op significante verschillen tussen de groepsgemiddelden. De p-waarde helpt bij het bepalen van de statistische significantie (35).

Tot slot werd er ook een regressieanalyse uitgevoerd om na te gaan in welke mate de industrie zou bijdragen aan de duurzaamheidsprofilering. Regressieanalyse is een statistische techniek om in te schatten wat het verband is tussen variabelen met een oorzaak-gevolgrelatie (62). Het analyseert de relatie tussen een afhankelijke variabele (in dit geval de duurzaamheidsscore) en een of meer onafhankelijke variabelen (in dit geval de industrie) (62).

4. Resultaten

In deze sectie van de thesis worden de resultaten van de analyses gepresenteerd en geïnterpreteerd. De analyse is opgedeeld in drie onderdelen, elk gericht op een specifiek aspect van dit onderzoek:

- **4.1 Herkennen van de industrie**: Deze subsectie behandelt de evaluatie van het vermogen van het AI-model om correct de industrieën te identificeren.
- **4.2 Ranking op duurzaamheid**: Deze sectie richt zich op hoe het AI-model de bedrijven rangschikt op basis van hun focus op duurzaamheid.
- **4.3 Relatie tussen industrie en duurzaamheid**: In dit onderdeel wordt de relatie tussen de industrie van een bedrijf en zijn focus op duurzaamheid geanalyseerd.

4.1. Herkennen van de industrie

Het model presteert zeer goed wanneer het een industrie moet toekennen aan een bedrijf. In tabel 6 worden de algemene statistieken weergegeven. Het model heeft een *accuracy* van 92,37%, wat aantoont dat in 92,37% van de gevallen dezelfde industrie door AI wordt toegekend als door de manuele *labeling* die aan de start van het onderzoek gebeurde. Dat is een zeer hoge *accuracy*, wat wijst op een sterk model. Daarnaast kan er met 95% zekerheid gezegd worden dat de *accuracy* van het model ligt tussen de 88.34% en 95.34%. De *No Information Rate* (NIR) reflecteert de *accuracy* wanneer het model altijd de meest frequente industrie zou voorspellen. Die waarde is 20,08%. Aangezien er vijf gedefinieerde industrieën zijn die gelijk verdeeld zijn, wijst dit ook op een sterk model. De p-waarde bevestigt dit ook bij het testen van de hypothese dat de *accuracy* van het model gelijk is aan de NIR. Een zeer lage p-waarde (veel kleiner dan 0.05) wijst erop dat de *accuracy* van het model significant beter is dan de NIR, wat betekent dat het model aanzienlijk beter is dan een model dat willekeurige of constante voorspellingen doet. In dit geval is de p-waarde gelijk aan $< 2.2e-16$. Tot slot meet de Kappa-statistiek de overeenstemming tussen de voorspelde en werkelijke classificaties. De hoge waarde van 0,9046 toont aan dat er een grote overeenstemming is tussen de voorspellingen van het model en de werkelijke data.

Algemene statistieken industrie-toekenning	
Accuracy	0.9237
95% CI	(0.8834, 0.9534)
No Information Rate	0.2008
P-Value [Acc > NIR]	$< 2.2e-16$
Kappa	0.9046

Tabel 6: Algemene statistieken industrie-toekenning

Tabel 7 is de *confusion matrix* die per industrie in detail weergeeft hoe veel instanties correct en niet correct geclassificeerd werden. De kolommen geven de werkelijke industrieën weer, terwijl de rijen de voorspelde industrieën tonen. Hoge waardes op de diagonaal tonen aan dat het model sterk presteert.

Voorspeld / Werkelijk	Automotive	Energy	Healthcare	Retail	Technology
Automotive	45	0	0	1	0
Energy	0	48	0	0	0
Healthcare	0	0	46	1	1
Retail	1	0	1	43	1
Technology	4	2	2	5	48

Tabel 7: Confusion Matrix

Op basis van deze matrix, werden ook de *Precision*, *recall* en *F1-score* berekend voor elke industrie. Die zijn terug te vinden in tabel 8.

Industrie-statistieken			
Industrie	**Precision**	**Recall**	**F1-score**
Automotive	0.9783	0.9000	0.9375
Energy	1.0000	0.9600	0.9796
Healthcare and Pharmaceuticals	0.9583	0.9388	0.9485
Retail	0.9348	0.8600	0.8958
Technology and Information Technology	0.7869	0.9600	0.8649

Tabel 8: Precision, Recall en F1-score voor elke industrie

De *precision* reflecteert hoeveel van de bedrijven die tot een bepaalde industrie zijn geclassificeerd, effectief in die industrie zitten. Het model scoort het hoogst bij deze maatstaf binnen de energiesector. Alle bedrijven die volgens het model tot de energiesector behoren, zijn effectief energiebedrijven. Het model werkt ook goed voor bedrijven die behoren tot de automotive, healthcare en retailsector. De *precision* van technologiebedrijven is het laagste, namelijk 78,69%. Een mogelijke verklaring hiervoor is dat bedrijven uit andere sectoren ook veel technologie gebruiken, waardoor het mogelijk sneller geclassificeerd wordt als zijne een technologiebedrijf.

Recall geeft weer hoeveel van de bedrijven die effectief tot een bepaalde industrie behoren, ook correct zijn geclassificeerd. Bij deze maatstaf scoort het model hoog bij de bedrijven actief in automotive, energie, healthcare en technologie. Het valt op dat bedrijven uit de retail-sector het minst goed worden geclassificeerd tot de juiste industrie. Een verklaring hiervoor is dat sommige retailbedrijven zich meer focussen op een bepaalde niche, in plaats van een breed assortiment aan te bieden. Wanneer zo een niche geassocieerd kan worden met een andere sector, is er een grotere kans dat het model dat daaronder gaan classificeren.

De *F1-score* combineert *precision* en *recall* in één cijfer. Het is de balans tussen hoe nauwkeurig en hoe volledig het model is. Deze metriek is het hoogst voor bedrijven uit de energiesector.

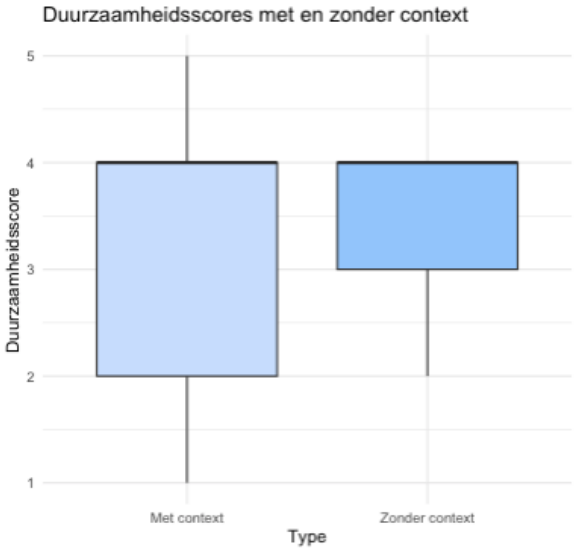
4.2. Ranking op duurzaamheid

De tweede analyse die werd uitgevoerd ging over het vermogen van het AI-model om bedrijven een score toe te kennen op basis van hun focus op duurzaamheid. De output van het model is een score van 1 tot en met 5, waarbij 1 wijst op geen focus op duurzaamheid en 5 op uitstekende focus op duurzaamheid. Aan het model werd voor elke score een set van criteria meegegeven om ervoor te zorgen dat alle scores consistent worden toegekend. [3]

Hierbij werden er twee verschillende methodes toegepast:

- **1. 'Met context':** bepalen van duurzaamheidscore op basis van de gescrapete tekst van relevante webpagina's als input.
- **2. 'Zonder context':** bepalen van duurzaamheidscore op basis van enkel de homepagina URL als input. Het GPT-model krijgt hierbij de opdracht om alle beschikbare informatie uit de training te gebruiken. Het model is getraind op internetdata tot en met september 2021.

Figuur 2 toont de respectievelijke boxplots van beide variaties. Hier is te zien dat de scores met context meer gespreid zijn dan de scores zonder context. Een mogelijke verklaring hiervoor is dat het model met context veel specifieker kan gaan zoeken op de aan- of afwezigheid van bepaalde zaken in de gescrapete tekst, waarnaar het model actief zoekt. Bij de input van de prompt is namelijk voor elke score tussen 1 en 5 een set van richtlijnen meegegeven zodat het geven van scores overheen alle bedrijven gelijkaardig verloopt, gebaseerd op de beschikbare teksten. Het model met context kan dus effectief gaan bekijken of het bedrijf volgens die teksten uitmuntend of juist totaal niet goed scoort op de duurzaamheidsprofilering. Het model zonder context is duidelijk voorzichtiger met de uiterste scores, wat impliceert dat het meer algemene input krijgt, die minder specifiek gefocust is op duurzaamheid. De mediaan ligt in beide gevallen op score 4.



Figuur 2: Boxplot van duurzaamheidsscores met en zonder context

Om het verschil in scores verder te onderzoeken, werd ook de Spearman correlatie berekend. Die bedraagt

[3] De criteria worden weergegeven in appendix A.2.

0,2587257. Dit is een positieve correlatie, wat betekent dat als de ene variabele toeneemt, de andere variabele ook de neiging heeft toe te nemen. De waarde is echter laag, wat wil zeggen dat de correlatie eerder zwak tot matig is. Ze komen dus niet sterk overeen, maar er is wel sprake van een samenhang.

De prestaties van dit model werden ook gemeten door de gegenereerde scores te spiegelen met scores die voortkomen uit een enquête die werd afgelegd door 34 respondenten. Hierbij werden 25 korte teksten van bedrijven geselecteerd om te laten rangschikken op dezelfde schaal van 1 tot en met 5. De algemene statistieken die voortkomen uit de enqûete, samen met de resultaten die door AI werden gegenereerd, worden weergegeven in tabel 9.

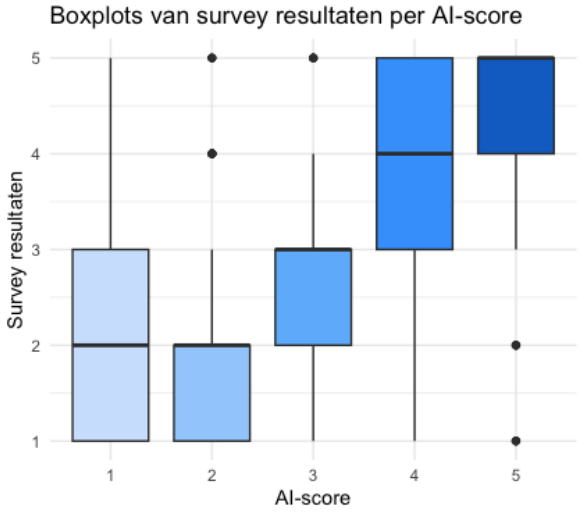
Bedrijf	AI Score	Gemiddelde	Mediaan	Variantie	Min	Max
1	5	4.2941	4	0.4563	3	5
2	5	3.2941	3	1.1836	1	5
3	5	4.3235	5	1.0740	1	5
4	5	4.6471	5	0.4171	2	5
5	5	4.6176	5	0.3645	3	5
6	4	3.2353	3	1.0339	1	5
7	4	4.2647	4.5	1.0490	1	5
8	4	4.5882	5	0.7344	2	5
9	4	4.0588	4	0.7237	1	5
10	4	3.8529	4	0.8565	1	5
11	3	3.3235	4	1.1346	1	5
12	3	3.0000	3	0.5455	1	5
13	3	2.9412	3	0.4813	1	5
14	3	2.2059	2	0.5927	1	4
15	3	2.0294	2	0.5143	1	4
16	2	1.4412	1	0.9813	1	5
17	2	2.2059	2	0.4109	1	4
18	2	1.8529	2	0.3717	1	4
19	2	2.3235	2	0.5891	1	4
20	2	1.6176	1	0.9706	1	5
21	1	1.4706	1	0.9840	1	5
22	1	3.4706	4	1.3476	1	5
23	1	3.1765	3	1.0588	1	5
24	1	1.1471	1	0.1898	1	3
25	1	1.5882	1	1.0980	1	5

Tabel 9: Algemene statistieken Survey

De rijen in tabel 9 zijn gesorteerd op basis van de AI-score (van groot naar klein). Op enkele uitzonderingen na, liggen de gemiddeldes en de medianen van de survey-resultaten in de buurt van de AI-scores. De variantie in deze tabel wijst op de spreiding van de scores rondom het gemiddelde. Een hogere variantie betekent dat de scores meer verspreid zijn, terwijl een lagere variantie betekent dat de scores dichter bij elkaar liggen. De hoogste variantie is die van bedrijf 22 en bedraagt 1.3476. Het valt ook op dat het verschil tussen de AI-score en de resultaten uit de survey daar het grootst is. Het gaat hier over een bedrijf uit de autosector. Wat opvalt aan de tekst, is dat de uitspraken over '*sustainability*' en '*reducing emissions and improving fuel efficiency*' wel vermeld worden, maar dat redelijk vaag blijft. Een mogelijke verklaring voor de lage AI-score en de hogere survey-score is mogelijk dat de ogen van de respondenten wel op die uitspraken gevallen zijn, terwijl er eigenlijk weinig diepgang of motivatie in die tekst verwerkt zit. Daarnaast is het ook belangrijk om te weten dat de teksten uit de survey maar een klein deel zijn van een groter geheel. De AI-score is gebaseerd op dat groter geheel waarin over het algemeen weinig over duurzaamheid vermeld wordt.

De variabelen 'Min' en 'Max' respectievelijk de minimale en maximale waardes die door respondenten gegeven werden in de survey. Hier valt het op dat er voor twee bedrijven die een AI-score van 5 hebben gekregen, ook een minimum score van 1 hebben. Andersom is er ook te zien dat voor vier bedrijven die een AI-score van 1 hebben gekregen, ook een maximum score van 5 hebben. Dit kan te maken hebben met het feit dat sommige respondenten weinig moeite hebben gedaan bij het invullen van de enquête. Dat werd verder onderzocht en het gaat hier echter om enkele uitzonderingen. Ook de antwoorden per respondent werden nagekeken, en daar is geen patroon terug te vinden dat wijst op het feit dat iemand weinig moeite gedaan heeft door *random* rangschikkingen door te sturen.

In 88% van de gevallen is de afwijking van de AI-score op basis van de gescrapete teksten gelijk aan nul, of maximaal één, ten opzichte van de mediaan uit de enquête. Dat is redelijk hoog, wat erop wijst dat het model goed presteert als het vergeleken wordt met menselijk input. De gemiddelde absolute fout (*Mean Absolute Error*) bedraagt 0,61. Dat betekent dat de voorspelde scores gemiddeld 0,61 punten afwijken van de menselijke scores. De correlatie tussen de AI-score en de mediaan uit de survey, bedraagt 0.7431975. Dat wijst op een sterk positieve relatie. Deze relatie wordt in meer detail toegelicht in figuur 3.

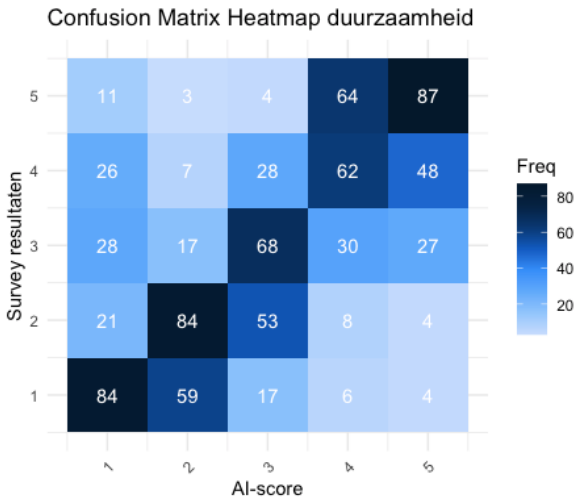


Figuur 3: Relatie tussen AI-scores en survey-resultaten

Uit figuur 3 valt af te leiden dat de medianen van de survey-resultaten voor AI-scores 2, 3, 4 en 5 gelijk zijn aan de scores die door het AI-model gegeven werden. Dit wijst ook op die positieve correlatie tussen de AI-score en de resultaten van de survey. Enkel bij de AI-score van 1 bedraagt de mediaan van de survey-resultaten 2, wat nog steeds niet slecht is aangezien er subjectieve invloeden kunnen voortkomen uit de antwoorden van respondenten en het algoritme zelf. Verder valt af te leiden dat de sprei-

ding van de resultaten voor AI-scores 2, 3 en 5 niet groot is. Bij een AI-score van 1 werden soms wel hogere cijfers toegekend in de survey, net zoals er voor de AI-score van 4 ook relatief meer lagere cijfers gegeven werden bij het invullen van de enquête. Een mogelijke verklaring hiervoor is dat enkele respondenten weinig moeite hebben gedaan bij het invullen van de enquête, maar dit kan na verder onderzoek van de antwoorden niet geconcludeerd worden omdat er geen opvallend patroon te vinden is in de data dat hierop kan wijzen.

Figuur 4 geeft in detail weer wat de frequenties zijn van de scores die de respondenten hebben ingevuld, vergeleken met de AI-scores die werden toegekend aan de bedrijven.



Figuur 4: Confusion Matrix Heatmap duurzaamheid

In het algemeen toont deze *heatmap* aan dat voor de AI-scores 1, 2, 3 en 5 de hoogste frequentie van antwoorden op dezelfde survey-score ligt. Een uitzondering op dit patroon wordt waargenomen bij de AI-score 4. Voor deze AI-score ligt de hoogste frequentie van antwoorden bij de survey-score op 5. Aangezien het AI-model effectief een score toekent, en de enquête een rangschikking is ten opzichte van elkaar, wordt dit niet gezien als een grote afwijking. Belangrijker om te vermelden is dat de frequentie van de lagere survey-scores van 1 en 2 in de grote minderheid zitten bij de door het model gegenereerde score van 4.

Er werd in deze analyse ook gemeten hoe robuust het AI-model is in het geven van de scores op basis van de gescrapete teksten. Dit werd gedaan door aan te geven in de code van de API dat het model telkens vijf keer moet antwoorden op de vraag. Hierdoor konden de vijf verschillende iteraties echter ook vergeleken worden ten opzichte van elkaar. De score met de hoogste frequentie werd dan gekozen als definitieve score. Het model geeft voor elk bedrijf telkens vijf keer dezelfde score. De standaarddeviatie is met andere woorden altijd gelijk aan nul, wat wil zeggen dat het model zeer robuust is in het toekennen van de scores.

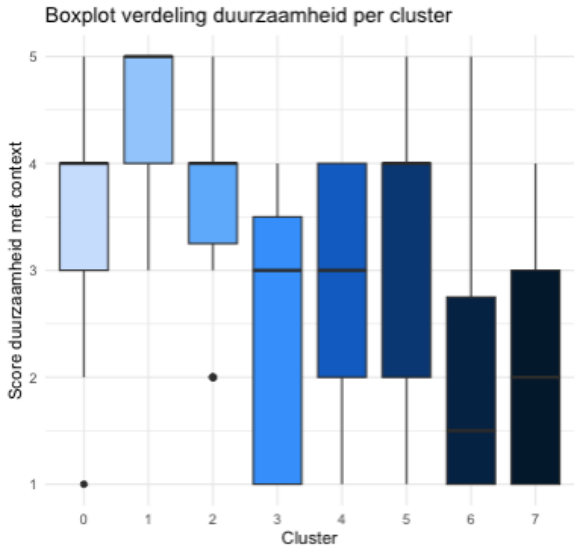
Tot slot werd er ook een clustering op de duurzaamheidsteksten uitgevoerd. Het doel van deze analyse, is om na te gaan of er verschillen of gelijkenissen te vinden zijn in de manier waarop bedrijven zich profileren naar duurzaamheid. Na het uitvoeren van de Elbow-methode, werd het optimale aantal clusters vastgesteld op acht. De gemiddelde duurzaamheidsscores en het aantal bedrijven per cluster zijn weergegeven in tabel 10.

Cluster	Average Score	Count
0	3.5	28
1	4.51	81
2	3.79	38
3	2.45	11
4	2.82	17
5	3.21	24
6	1.9	30
7	2.2	20

Tabel 10: Cluster statistieken

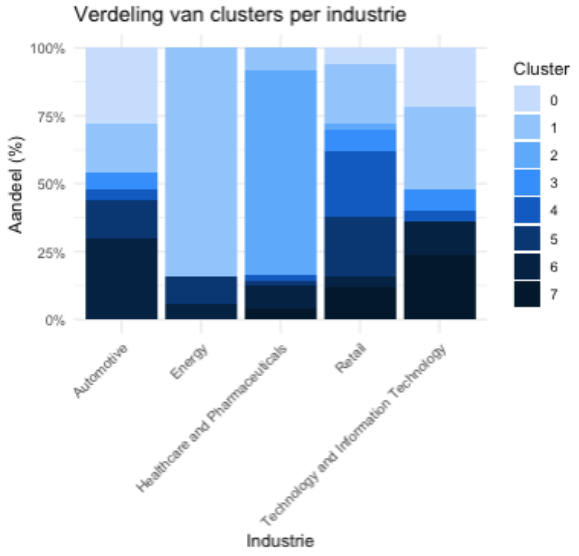
Opvallend is dat cluster 1 het meest aantal bedrijven bevat: 81 in totaal, maar ook de hoogste gemiddelde score van 4.51 heeft. Dit impliceert dat de teksten veel elementen over duurzaamheid bevatten. Meer dan drie vierde van de bedrijven in deze cluster zijn bedrijven uit de energiesector. Aan de andere kant staat cluster 6. In die cluster zitten 30 bedrijven en die cluster heeft de laagste gemiddelde score van 1.9 wat impliceert dat die teksten weinig tot geen informatie over duurzaamheid bevatten. De meerderheid van deze bedrijven bevinden zich in de autosector.

De verdeling van de scores per cluster is weergegeven in de boxplots in figuur 5. Die visualisatie bevestigt dat cluster 1 het hoogste scoort en toont ook aan dat er minder variabiliteit in die scores zit, aangezien de box eerder compact is, vergeleken met bijvoorbeeld cluster 3, 4, 5, 6 en 7. De box van cluster 2 is het meest compact en scoort ook redelijk hoog met een mediaan van vier. Dit kan erop wijzen dat bedrijven die onder cluster 2 vallen redelijk gelijkaardig communiceren over duurzaamheid, wat erop zou kunnen wijzen dat ze eerder een *template* gebruiken en dus minder oprecht zijn. Dit is echter een assumptie die niet verder wordt gevalideerd binnen dit onderzoek.



Figuur 5: Verdeling van de scores per cluster

Figuur 6 analyseert het verband tussen de industrie waarin het bedrijf zich bevindt en het resultaat van het clusteralgoritme. Deze analyse werd uitgevoerd om te kunnen valideren dat de clustering gebeurd is op basis van hoe bedrijven zich profileren naar duurzaamheid, onafhankelijk van de sector. Opvallend is dat de clusters mooi verdeeld zijn over de verschillende sectoren. Echter komt cluster 1 wel voor in meer dan 75% van de gevallen binnen de energiesector waar cluster 2 vooral domineert binnen de healthcare sector.



Figuur 6: Verdeling clusters per industrie

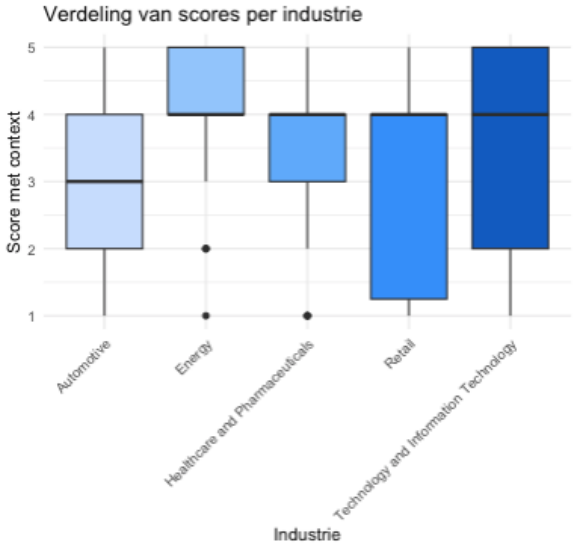
4.3. Relatie tussen industrie en duurzaamheid

In deze laatste subsectie wordt de relatie tussen de industrie en het duurzaamheidsaspect onderzocht. Tabel 11 toont de verdeling van scores per industrie. Over het algemeen valt het op dat de scores bij de energiesector het hoogst zijn en dat de spreiding daar het laagst is. Aan de andere kant van het spectrum staan de bedrijven uit de retail-, technologie-, en autosector. Het valt op dat die scores meer verspreid zijn en dat er ook relatief meer lagere duurzaamheidsscores toegekend werden.

Industrie/Score	**1**	**2**	**3**	**4**	**5**
Automotive	7	13	9	11	10
Energy	1	3	3	29	14
Healthcare and Pharmaceuticals	4	6	7	22	10
Retail	13	2	6	22	7
Technology and Information Technology	8	8	3	17	14
Aantal	**33**	**32**	**28**	**101**	**55**

Tabel 11: Verdeling van scores per industrie

Om meer context te krijgen, kunnen die gegevens ook gevisualiseerd worden in boxplots, zoals te zien in figuur 7. Wat hier opvalt, is dat de mediaan bij 'Energy', 'Healthcare and Pharmaceuticals', 'Retail' en 'Technology and Information Technology' het hoogste is, namelijk vier. De laagste mediaan van drie is terug te vinden bij de autosector.



Figuur 7: Verdeling van scores per industrie

De gemiddelde scores per industrie worden weergegeven in tabel 12. Die tabel bevestigt dat bedrijven uit de energiesector zich het meest profileren op duurzaamheid. Bedrijven uit de autosector presteren het minst goed volgens deze metriek.

Industrie	**Gemiddelde score**
Automotive	3.08
Energy	4.04
Healthcare and Pharmaceuticals	3.57
Retail	3.16
Technology and Information Technology	3.42

Tabel 12: Gemiddelde scores per industrie

In tabel 13 worden de resultaten van de ANOVA-test

weergegeven. Binnen deze tabel zijn vooral de *F-value* en de Pr(>F) van belang om te bepalen of er significante verschillen zijn tussen de duurzaamheidsscores per industrie. De hoge F-waarde van 4.406 geeft aan dat dat de variaties tussen de verschillende industrieën zeker aanwezig zijn in vergelijking met de variatie binnen de industrieën. Daarnaast zegt de zeer lage p-waarde van 0.00186 dat de nulhypothese (dat er geen verschillen zijn in de gemiddelde duurzaamheidsscores tussen de industrieën) verworpen kan worden. Dat impliceert dat er met 99% zekerheid kan worden vastgesteld dat er significante verschillen zijn tussen de industrieën wat betreft hun duurzaamheidsscores.

Source	**Df**	**Sum Sq**	**Mean Sq**	**F-value**	**Pr(>F)**
Industrie	4	29.2	7.305	4.406	0.00186 **
Residuals	244	404.5	1.658		
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Tabel 13: ANOVA tabel

Een laatste analyse die uitgevoerd is op de relatie tussen de industrie en het duurzaamheidsaspect, is een regressieanalyse.

Onder '*Coefficients*' zijn dit de belangrijkste opmerkingen:

- Als basisgroep werd 'Automotive' gekozen. De schatting van het intercept reflecteert de gemiddelde duurzaamheidsscore van die basisgroep. Aangezien dat het laagste gemiddelde was, gaan de coëfficiënt van de andere industrieën allemaal positief zijn.
- De gemiddeldes van de andere industrieën worden bekomen door de desbetreffende coëfficiënt op te tellen bij het intercept van 3.0800.
- Enkel de p-waarden van de energie- en healthcaresector zijn klein genoeg om te concluderen dat de verschillen in duurzaamheidsscores met de autosector significant zijn. De verschillen met de andere sectoren zijn niet significant.

Tot slot toont de *Multiple R-squared* van 0.06737 aan dat ongeveer 6.74% van de variatie in duurzaamheidsscores wordt verklaard door de verschillen in industrieën. Dat is niet bepaald hoog, wat wil zeggen dat er andere factoren zijn die die een belangrijke rol spelen. De overige statistieken uit deze tabel worden niet verder in detail besproken.

5. Tekortkomingen en aanbevelingen voor toekomstig onderzoek

Ondanks de waardevolle inzichten en resultaten die voortkomen uit dit onderzoek, zijn er enkele tekortkomingen die de interpretatie van de bevindingen beïnvloeden. Deze tekortkomingen kunnen aangepakt worden in toekomstig onderzoek.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	3.0800	0.1821	16.915	$< 2 \times 10^{-16}$ ***
factor(Industrie)Energy	0.9600	0.2575	3.728	0.00024 ***
factor(Industrie)Healthcare and Pharmaceuticals	0.4914	0.2588	1.899	0.05878 .
factor(Industrie)Retail	0.0800	0.2575	0.311	0.75632
factor(Industrie)Technology and Information Technology	0.3400	0.2575	1.320	0.18796
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residuals				
Min	-3.0400			
1Q	-1.0800			
Median	0.4286			
3Q	0.9200			
Max	1.9200			
Residual standard error:	1.288	on 244 degrees of freedom		
Multiple R-squared:	0.06737	Adjusted R-squared: 0.05208		
F-statistic:	4.406	on 4 and 244 DF, p-value: 0.001859		

Tabel 14: Regressieanalyse

Een eerste tekortkoming betreft de beperkte representativiteit van de onderzochte industrieën. Dit onderzoek focuste zich op vijf industrieën: technologie, gezondheidszorg, automotive, energie en retail. Hierdoor is de generaliseerbaarheid van de resultaten beperkt, aangezien er veel meer industrieën bestaan. Een uitbreiding naar meer industrieën zou de toepasbaarheid van de resultaten vergroten. Daarnaast was de steekproefgrootte van 250 websites relatief klein en dus niet representatief voor een globale context. Toekomstig onderzoek zou moeten streven naar een grotere steekproef om de betrouwbaarheid van de resultaten te verbeteren.

Een tweede beperking is de mogelijke subjectiviteit van respondenten bij het invullen van de enquête. Dit kan de *accuracy* van het AI-model beïnvloeden. Een aanbeveling is om een grotere steekproef te gebruiken en eventueel een expertenpanel samen te stellen om de subjectiviteit te verminderen. Daarnaast werden alleen de drie meest relevante pagina's van de websites geanalyseerd, wat kan leiden tot een onvolledig beeld van de duurzaamheidsprofilering van een bedrijf. Toekomstig onderzoek zou meer webpagina's kunnen analyseren, inclusief alle bijhorende publicaties, nieuwsartikelen, *social media* en rapporten, om een completer beeld te krijgen.

De derde tekortkoming betreft het 'black box'-karakter van het gebruikte AI-model, waardoor het niet duidelijk is hoe het model tot bepaalde resultaten komt. Toekomstig onderzoek zou gebruik kunnen maken van *Explainable AI* technieken, zoals *SHAP*, *LIME*, en *Saliency Maps*, die de transparantie van AI-modellen kunnen verbeteren (20).

Een vierde beperking betreft dat de trainingsdata van het 'gpt-3.5-turbo-0125'-model, gebaseerd is op gegevens tot en met september 2021. Voor duurzaamheidsscores op basis van de gescrapete tekst is dit geen probleem, maar voor analyses zonder die input is het beter om een recenter model zoals gpt-4o te gebruiken. Dit model heeft trainingsdata dat actueel is tot en met oktober 2023 (2). Toekomstig onderzoek zou het verschil in resultaten tussen meerdere AI-modellen kunnen onderzoeken om hun prestaties te vergelijken.

Tot slot is het ook interessant om te streven naar het ontwikkelen van een matuur duurzaamheidsmodel, zodat diverse studies hiervan uniform gebruik kunnen maken. Op die manier kan er op een meer objectieve manier vergeleken worden hoe bedrijven scoren op deze index.

Door deze aanbevelingen op te volgen, kan toekomstig onderzoek de tekortkomingen van het huidige onderzoek aanpakken en de betrouwbaarheid en toepasbaarheid van de resultaten verder verbeteren.

6. Conclusie

Dit onderzoek focust op het evalueren van een AI-model om enerzijds de juiste industrie toe te kennen aan bedrijven, om ze vervolgens te rangschikken op basis van hun duurzaamheidsprofilering. Het resultaat toont het potentieel van AI om complexe data te transformeren in waardevolle inzichten waarmee bedrijven hun benaderingsstrategieën kunnen personaliseren om zakelijke relaties te versterken. De resultaten van dit onderzoek worden opgedeeld in drie delen: het herkennen van de industrie, de ranking op duurzaamheid en de relatie tussen beide aspecten.

6.1. Herkennen van de industrie

De resultaten uit het eerste onderdeel van deze thesis tonen aan dat het AI-model zeer effectief is in het toekennen van industrieën op basis van de websiteinhoud. De *accuracy* van 92,37% toont aan dat in 92,37% van de gevallen de voorspelde industrie overeenkomt met de manuele labeling die aan het begin van het onderzoek werd uitgevoerd. Het betrouwbaarheidsinterval (95% CI) van 88,34% tot 95,34% bevestigt dat het model hoge prestaties levert. Deze statistieken worden verder onderbouwd door de zeer lage p-waarde (<2.2e-16) dat aantoont dat dat het model aanzienlijk beter is dan een model dat willekeurige of constante voorspellingen doet.

De *confusion matrix* (Tabel 7) toont in welke mate de industrieën correct worden geclassificeerd. Van de 50 bedrijven in de automotive sector, werden er 45 correct geclassificeerd als automotive. In de energiesector werden 48 van de 50 bedrijven correct geclassificeerd. De healthcare sector toonde ook hoge *accuracy* waar 46 van de 49 bedrijven correct geclassificeerd werden. De retail sector had een iets lagere *accuracy*, waarbij 43 van de 50 bedrijven correct geclassificeerd werden, en de technologiesector had 48 van de 50 bedrijven correct geclassificeerd.

Samenvattend kan worden geconcludeerd dat het AI-model goed is in het toekennen van industrieën aan bedrijven. Dit model kan een waardevol hulpmiddel zijn voor bedrijven en onderzoekers die op zoek zijn naar geautomatiseerde methoden voor bedrijfsclassificatie.

6.2. Ranking op duurzaamheid

Ook de analyse van het vermogen van het AI-model om bedrijven te rangschikken op basis van hun focus op duurzaamheid toont aan dat het model goed presteert. Er werden twee verschillende methoden toegepast voor de beoordeling:

- **1. 'Met context':** bepalen van duurzaamheidscore op basis van de gescrapete tekst van relevante webpagina's als input.
- **2. 'Zonder context':** bepalen van duurzaamheidscore op basis van enkel de homepagina URL als input. Het GPT-model krijgt hierbij de opdracht om alle beschikbare informatie uit de training te gebruiken. Het model is getraind op internetdata tot en met september 2021.

De scores met context zijn over het algemeen meer gespreid dan de scores zonder context. Dat impliceert dat de resultaten met context mogelijk meer informatie bieden, wat leidt tot een nauwkeurigere beoordeling. Beide methoden vertonen wel dat score 4 het meest frequent voorkomt.

De prestaties van het model met context werd verder geëvalueerd door de gegenereerde duurzaamheidsscores te vergelijken met de scores uit een enquête onder 34 respondenten. De algemene statistieken tonen aan dat in 88% van de gevallen de afwijking van de AI-score met context nul, of maximaal één is, ten opzichte van de mediaan uit de enquête. Dat wijst op een sterke correlatie tussen de modelvoorspellingen en menselijke beoordelingen.

De boxplots in figuur 3 tonen aan dat de medianen van de survey voor alle AI-scores, behalve die van 1, mooi in lijn liggen met de scores die werden toegekend door het model. Tabel 9 toont ook aan dat de gemiddeldes en medianen van de enquête in de meeste gevallen in de buurt van de AI-scores lagen, wat wijst op een consistente en betrouwbare prestatie van het model.

Samenvattend toont ook deze sectie aan dat het AI-model goed presteert bij het rangschikken van bedrijven op basis van hun duurzaamheid. De consistentie tussen de AI-scores en de menselijke beoordelingen, de gedetailleerde analyse van variantie en de positieve correlatie bevestigen dit.

6.3. Relatie tussen industrie en duurzaamheid

Tot slot werd de relatie tussen de industrie en het duurzaamheidsaspect ook nog onderzocht. De resultaten tonen aan dat er voor deze steekproef verschillen zijn in duurzaamheidsscores tussen verschillende industrieën, wat waardevolle inzichten biedt voor het begrijpen van hoe sectoren duurzaamheid benaderen en erover communiceren.

Tabel 11 toont aan dat de energiesector de hoogste duurzaamheidsscores behaalt, met een gemiddelde score van 4.04. Deze sector toont bovendien ook een consistente hoge betrokkenheid bij duurzaamheid, wat geïllustreerd wordt door de relatief kleine spreiding van scores. Aan de andere kant van het spectrum bevindt zich de autosector, die een grote variabiliteit en lagere duurzaamheidsscores laat zien, met een gemiddelde van 3.08.

De ANOVA-test resulteert in een hoge F-waarde van 4.406 en een zeer lage p-waarde van 0.00186 wat wil zeggen dat er significante variaties in duurzaamheidsscores aanwezig zijn tussen de verschillende industrieën. De Multiple R-squared geeft meer inzicht en zegt dat ongeveer 6.74% van de variatie in duurzaamheidsscores kan worden verklaard door verschillen tussen industrieën.

Referenties

[1] Litmaps | Your Literature Review Assistant.

[2] OpenAI GPT-3.5 Turbo API [gpt-3.5-turbo-0125], 2023.

[3] sentence-transformers/all-MiniLM-L6-v2 · Hugging Face, January 2024.

[4] Hayri Volkan Agun. WebCollectives: A light regular expression based web content extractor in Java. *SoftwareX*, 24:101569, 2023.

[5] A. A. R. Al-Azmi. Data, text and web mining for business intelligence: a survey. *arXiv preprint arXiv:1304.3563*, 2013. Publisher: arxiv.org.

[6] Ali Abd Al-Hameed and Khawla. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, 13(1):3249–3255, March 2022. Publisher: Semnan University.

[7] MS El Asikri, S. Knit, and H. Chaib. Using web scraping in a knowledge environment to build ontologies using python and scrapy. *European Journal of Molecular & ...*, 2020. Publisher: researchgate.net Type: PDF.

[8] D. S. Asudani, N. K. Nagwani, and P. Singh. Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review*, 2023. Publisher: Springer Type: HTML.

[9] Satya Avasarala. *Selenium WebDriver practical guide*. PACKT publishing, 2014.

[10] Luisa Bosetti. Web-Based Integrated CSR Reporting: An Empirical Analysis. *Symphonya. Emerging Issues in Management*, (1):18–38, July 2018. Number: 1.

[11] C. Bounsaythip and E. Rinta-Runsala. Overview of data mining for customer behavior modeling. *VTT Information Technology Research ...*, 2001. Publisher: inf.utfsm.cl Type: PDF.

[12] Z. Bošnjak, O. Grlević, and S. Bošnjak. Transforming Web data into knowledge: Implications for management. *Ekonomski horizonti*, 2019. Publisher: scindeks.ceon.rs.

[13] N. Chandra and K. Kumar. Designing Web Mining Technique for Customer Segmentation. *academia.edu*. Type: PDF.

[14] Rakesh Vidya Chandra and Bala Subrahmanyam Varanasi. *Python requests essentials*. Packt Publishing Ltd, 2015.

[15] L. S. Chen, C. C. Hsu, and M. C. Chen. Customer segmentation and classification from blogs by using data mining: an example of VOIP phone. *Cybernetics and Systems: An ...*, 2009. Publisher: Taylor &Francis.

[16] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, December 2014. arXiv:1412.3555 [cs].

[18] Carlo Cici and Diana D'Isanto. Integrating Sustainability into Core Business. In *Symphonya. Emerging Issues in Management*, pages 50–65, December 2017. ISSN: 1593-0319, 1593-0300 Issue: 1.

[19] Mengyao Cui. Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Accounting, Auditing and Finance*, 1(1):5–8, October 2020. Publisher: Clausius Scientific Press.

[20] R. S. Deshpande and P. V. Ambatkar. Interpretable Deep Learning Models: Enhancing Transparency and Trustworthiness in Explainable AI. *Proceeding International Conference on Science and Engineering*, 11(1):1352–1363, February 2023. Number: 1.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[22] V. Dogra, S. Verma, P. Chatterjee, J. Shafi, and ... A complete process of text classification system using state-of-the-art NLP models. *Computational ...*, 2022. Publisher: hindawi.com Type: HTML.

[23] V. Dumitru, D. Iorga, S. Ruseti, and ... Garbage in, garbage out: An analysis of HTML text extractors and their impact on NLP performance. *2023 24th International ...*, 2023. Publisher: ieeexplore.ieee.org.

[24] Armin Esmaeilzadeh and Kazem Taghva. Text Classification Using Neural Network Language Model (NNLM) and BERT: An Empirical Comparison. In Kohei Arai, editor, *Intelligent Systems and Applications*, Lecture Notes in Networks and Systems, pages 175–189, Cham, 2022. Springer International Publishing.

[25] Gabriel Eweje. A Shift in corporate practice? Facilitating sustainability strategy in companies. *Corporate Social Responsibility and Environmental Management*, 18(3):125–136, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/csr.268.

[26] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli. A survey on text classification algorithms: From text to predictions. *Information*, 2022. Publisher: mdpi.com Type: HTML.

[27] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for Multi-Class Classification: an Overview, August 2020. arXiv:2008.05756 [cs, stat].

[28] R. Gunawan, A. Rahmatulloh, and ... Comparison of web scraping techniques: regular expression, HTML DOM and Xpath. *2018 International ...*, 2019. Publisher: atlantis-press.com.

[29] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, pages 357–362, 2020.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[31] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[32] M. Kavitha and P. Prabhavathy. A review on machine learning techniques for text classification. *2021 4th International Conference ...*, 2021. Publisher: ieeexplore.ieee.org.

[33] M. Kavitha and P. Prabhavathy. Multi-Label Text Classification using Machine Learning Techniques. *Telematique*, 2022. Publisher: researchgate.net Type: PDF.

[34] M. A. Khder. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & ...*, 2021. Publisher: ijasca.zuj.edu.jo Type: PDF.

[35] Hae-Young Kim. Analysis of variance (ANOVA) comparing means of more than two groups. *Restorative Dentistry & Endodontics*, 39(1):74–77, February 2014.

[36] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text Classification Algorithms: A Survey. *Information*, 10(4):150, April 2019. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[37] S. N. Kumar. World towards advance web mining: A review. *American Journal of Systems and Software*, 2015. Publisher: academia.edu Type: PDF.

[38] T. B. Lalitha and P. S. Sreeja. Potential Web Content Identification and Classification System using NLP and Machine Learning Techniques. *International Journal of Engineering ...*, 2023. Publisher: researchgate.net Type: PDF.

[39] B. Liu, W. Guan, C. Yang, Z. Fang, and Z. Lu. Transformer and Graph Convolutional Network for Text Classification. *International Journal of ...*, 2023. Publisher: Springer Type: HTML.

[40] C. Lotfi, S. Srinivasan, M. Ertz, and I. Latrous. Web Scraping Techniques and Applications: A Literature. *researchgate.net*. Type: PDF.

[41] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. In *2018 International Seminar on Application for Technology of Infor-*

mation and Communication, pages 533–538, September 2018.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013. arXiv:1301.3781 [cs].

[43] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius. Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. *Algorithms*, 2023. Publisher: mdpi.com.

[44] Timothy B. Palmer and David J. Flanagan. The sustainable company: looking at goals for people, planet and profits. *Journal of Business Strategy*, 37(6):28–38, January 2016. Publisher: Emerald Group Publishing Limited.

[45] M. S. Parvez, K. S. A. Tasneem, and ... Analysis of different web data extraction techniques. ... *Conference on Smart* . . . , 2018. Publisher: ieeexplore.ieee.org.

[46] J. Paschen, J. Kietzmann, and T. C. Kietzmann. Artificial intelligence (AI) and its implications for market knowledge in B2B marketing. *Journal of business &* . . . , 2019. Publisher: emerald.com.

[47] S. K. Patnaik and C. Narendra Babu. Trends in web data extraction using machine learning. *Web Intelligence*, 2021. Publisher: content.iospress.com.

[48] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[49] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, October 2020. arXiv:2004.09813 [cs].

[50] Leonard Richardson. Beautiful soup documentation. *April*, 2007.

[51] Diana Ridley. *The Literature Review*. 2012.

[52] Slamet Riyanto, Imas Sukaesih Sitanggang, Taufik Djatna, and Tika Dewi Atikah. Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(6), 2023. Number: 6 Publisher: The Science and Information (SAI) Organization Limited.

[53] J. Salminen, M. Mustak, M. Sufyan, and B. J. Jansen. How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. *Journal of Marketing* . . . , 2023. Publisher: Springer Type: HTML.

[54] W. K. Sari, D. P. Rini, and R. F. Malik. Text Classification Using Long Short-Term Memory with GloVe. *Jurnal Ilmiah Teknik Elektro* . . . , 2019. Publisher: pdfs.semanticscholar.org Type: PDF.

[55] D. Shete, S. Bojewar, and A. Sanghvi. Survey Paper on Web Content Extraction & Classification. *2021 6th International* . . . , 2021. Publisher: ieeexplore.ieee.org.

[56] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised K-Means Clustering Algorithm. *IEEE Access*, 8:80716–80727, 2020. Conference Name: IEEE Access.

[57] Guoying Sun, Yanan Cheng, Zhaoxin Zhang, Xiaojun Tong, and Tingting Chai. Text classification with improved word embedding and adaptive segmentation. *Expert Systems with Applications*, 238:121852, 2024.

[58] C M Suneera and Jay Prakash. Performance Analysis of Machine Learning and Deep Learning Models for Text Classification. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6, New Delhi, India, December 2020. IEEE.

[59] R Core Team. R: A Language and Environment for Statistical Computing, 2016.

[60] Violet Turri, Rachel Dzombak, Eric Heim, Nathan VanHoudnos, Jay Palat, and Anusha Sinha. Measuring AI Systems Beyond Accuracy, April 2022. arXiv:2204.04211 [cs].

[61] A. Ullah, M. I. Mohmand, H. Hussain, S. Johar, I. Khan, and ... Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time. *Sensors*, 2023. Publisher: mdpi.com Type: HTML.

[62] Gülden Kaya Uyanık and Neşe Güler. A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, 106:234–240, December 2013.

[63] A. K. Uysal and Y. L. Murphey. Sentiment classification: Feature selection based approaches versus deep learning. *2017 IEEE International Conference* . . . , 2017. Publisher: ieeexplore.ieee.org.

[64] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[65] Z. Wan. Text Classification: A Perspective of Deep Learning Methods. *arXiv preprint arXiv:2309.13761*, 2023. Publisher: arxiv.org.

[66] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, November 2023. arXiv:2303.18223 [cs].

[67] Jinfeng Zhou, Jinliang Wei, and Bugao Xu. Customer segmentation by web content mining. *Journal of Retailing and Consumer Services*, 61:102588, 2021.

A. Appendix

A.1. Survey

```
Thank you for participating in this survey.

This survey is a fundamental component of my master's thesis, which aims to perform profiling
    and segmentation of B2B customers based on their website content. After scraping data
    from company websites, an AI model will analyze the extent to which each company
    prioritizes sustainability, on a scale from 1 to 5. To assess the performance of this
    model, I have extracted several sentences related to sustainability from a limited number
    of websites. The goal is for participants to also rank these websites on a scale of 1 to
    5 based on sustainability, allowing for a comparison with the rankings generated by my AI
    model.

The purpose of this survey is to gather data to evaluate the accuracy of the AI model's
    sustainability rankings compared to human assessments. Your honest and thoughtful
    responses are greatly appreciated and will contribute significantly to the success of
    this research project.

Once again, thank you for taking the time to participate in this survey.

This survey exists of 5 questions.

Kind regards,
Maxim Riebus
Master in Business and Information Systems Engineering
Hasselt University

---

You'll encounter five questions, each asking you to rank five short texts on a scale from 1
    to 5, where 1 means the company is committed to sustainability and 5 means the company is
    not committed to sustainability.

To give you more context, the following bullet points can be examples of a committed company:
- Progressive sustainability policies, including innovative approaches and best practices.
- Excellent performance data in environmental management, social responsibility and
    transparent reporting.
- Sustainability leadership, with active involvement in cross-sector initiatives and positive
    influence on the industry.


Please rank the following text blocks from 1 to 5, where 1 means the company is committed to
    sustainability and 5 means the company is not committed to sustainability.

1. "We want to realize equality and all-inclusiveness by making technology accessible to all,
    develop reliable ICT infrastructure to power the digital world, balance social progress
    and environmental protection through innovation, and work with partners to build a
    harmonious ecosystem... Aligned with the UN SDGs and our vision and mission, XXX is our
    long-term digital inclusion initiative and action plan... Together with our partners, we
    are committed to innovating technologies and solutions that make the world a more
    inclusive and sustainable space for all."

2. "At XXX, climate and social responsibility are more than just talk.  W e re  working toward
    a sustainable future by focusing on three areas: rethinking consumption, value chain
    responsibility and a lower-impact production process...  W e re  changing the way we
    design and engineer our products to reduce waste and reuse resources by incorporating
    recycled and responsibly sourced materials, eliminating single-use plastics, and
    improving product reparability and recyclability... We take social and environmental
    considerations into account when forming and managing relationships with suppliers by
    working to build a fully traceable value chain, only working with suppliers committed to
    fair labour practices, and collaborating with industry stakeholders to be part of the
    solution."

3. "We understand the responsibility that comes with our role and are committed to operating
    sustainably for our investors, employees, communities and fellow citizens of the globe...
    Our integrated energy infrastructure network provides midstream energy services to
    producers and consumers of natural gas, natural gas liquids, crude oil, refined products
    and petrochemicals... The Enterprise model is woven into our culture - from the day a new
```

hire arrives at orientation, into our development programs, and through our weekly messages in staff meetings and other forums."

4. "Our strong ability to listen to the diverse needs of our many stakeholders is the key component in making high-quality products that provide value for patients... XXX work reflects its strategic vision to be an innovation-driven specialty pharmaceutical company focused on improving outcomes for patients with severe and critical conditions... From philanthropic efforts to advocacy initiatives, we are living our values throughout our communities."

5. "The path to 360 value starts h e r e featuring our most provocative thinking, extensive research and compelling stories of shared success... XXX seeks a diverse range of professionals to deliver specialized capabilities and solutions to clients across all industries... If you are an XXX employee with a question about company-related benefits, services or administrative issues, please use the resources available to you below."

Please rank the following text blocks from 1 to 5, where 1 means the company is committed to sustainability and 5 means the company is not committed to sustainability.

1. "Guided by multi-year goals, we are committed to reducing our environmental impact, and invests in hundreds of projects every year aimed at reducing energy, material and water consumption, and greenhouse gas (GHG) e m i s s i o n s We have a particular focus on using water responsibly and efficiently. In 2022, we implemented water reduction projects that saved an annualized amount of water equal to 3.2% of our 2021 u s a g e Our semiconductor products are powering innovation around the world, such as electric vehicles and renewable energy applications, and helping to make a positive impact in a growing number of ways."

2. "We're always striving to lead the way when it comes to sustainability and the c o m m u n i t y Take a look at how we're helping coffee farmers adapt to climate change and explore the roots of the 100% responsibly sourced cotton in our c l o t h i n g Through small changes, from carefully designing our packaging to supporting our local communities and working together to reduce our carbon footprint, we can take big steps to a more sustainable future."

3. "We are committed to conducting our business with integrity and the pursuit of excellence in all that we d o In particular, we are committed to acting responsibly, safely and with transparency in our interactions with patients, doctors and other stakeholders in the healthcare s y s t e m Everything we do at XXX is focused on three things: Putting patients first, being a great place to work and living our values: Integrity, Collaboration, Passion, Innovation, Pursuit of Excellence."

4. "We use every opportunity to think differently, make bold decisions and pioneer novel s o l u t i o n s At XXX, we work together in pursuit of one common goal: to dramatically improve and extend p a t i e n t s l i v e s We believe diversity is the key to unlocking our full potential. We push boundaries and take well-calculated risks, committing courageously and wholeheartedly to what we believe in."

5. "Charges vary depending on service provider and country/region... Free of charge from German landlines and mobile networks... Callers from abroad, please use the international number"

Please rank the following text blocks from 1 to 5, where 1 means the company is committed to sustainability and 5 means the company is not committed to sustainability.

1. "At XXX, we believe in acting with integrity. Through our environmental, social and governance (ESG) priorities, we aim to support long-term sustainability that benefits our people, communities and p l a n e t XXX's sustainability initiatives go beyond the design and manufacturing of our product. They encompass how we grow our business, support our people and communities and power a future that enriches life for a l l XXX takes a proactive approach to environmental stewardship. In all areas of operational sustainability, we make improvements and find ways to reduce our footprint as early as possible in a process."

2. "Together, with our customers, we are redefining the future of energy and developing innovative solutions that power communities and improve l i v e s W e re committed to supporting economic development and making smart infrastructure investments that power our communities and improve lives by attracting high-quality jobs, encouraging capital investment, and stimulating local e c o n o m i e s We believe effective corporate governance is an essential part of our sustained performance."

3. "XXX. is committed to the maintenance of high ethical standards, both internally and in its dealings with all those with whom it is involved XXX is relying on you to use good judgment and common sense in all of your business dealings XXX is committed to providing equal opportunity in employment to all employees and applicants."

4. "At XXX, we take the health and wellbeing of our team very seriously Remember, we re here to help you and we wont stand for abuse of any kind Read our energy guides Make a complaint Report a security concern Boost customer help Citizens advice star rating."

5. "The XXX plan offers coverage for services and component repairs, promoting sustainability by extending the life of vehicles and reducing the need for new ones The XXX technology, used with permission from XXX, helps to reduce emissions and improve fuel efficiency in XXX vehicles XXX is committed to sustainability, with trademarks such as 'Green Car Journal's Green SUV of the Year' awarded to the XXX Grand Cherokee EcoDiesel and 'Best Environmental Performance Brand' awarded to XXX."

Please rank the following text blocks from 1 to 5, where 1 means the company is committed to sustainability and 5 means the company is not committed to sustainability.

1. "To achieve sustainability and enrich communities together with our customers and partners, we are practicing sustainability management to fulfill our social responsibility and create shared social value We are working to realize our environmental vision by promoting decarbonization and closing the resource loop, developing environmental technologies, and providing products and services that reduce environmental impacts XXX emphasizes contributions involving the technologies and knowledge that underpin its business as a way to give something back to society. XXX will continue to engage in corporate citizenship activities, including contributions involving manpower."

2. "Our journey to develop as a broad energy company is founded on a strong commitment to sustainability, and our strategic pillars always safe, high value and low carbon are applied in everything we do A sustainable development path well below the two-degree target does not allow for further delays in policy, industry, and consumer action to reduce emissions We're one of the world's most CO2-efficient producers of oil and gas and we are proactively investing in renewables."

3. "XXX is our commitment to making a positive impact on our people, environment and our community A focus on combination therapies requires appropriate obesity-specific trial designs, long-term follow-up studies and diverse patient recruitment Our offering includes a range of globally scalable FSP solutions that are customised to provide support where and when you need it."

4. "Everything we do is focused on supporting you and helping you achieve your ambitions Intelligent vehicles that anticipate your every need, so you can go out there and live your best life Looking after our owners is deep within our DNA. So when you visit, you can expect award-winning customer service and absolutely no sales pressure."

5. "We examine metrics such as how you are shopping on our website, in our stores, and on our mobile application(s), the performance of our marketing efforts, and your response to those marketing efforts We also allow third-party companies (e.g., XXX) to place tags on our digital properties once approved through our tagging process. The tags may collect information from your interactions on XXX Our sites and mobile application(s) include social media features, such as the Facebook, Pinterest, and Twitter widgets. These features may collect information about you such as your IP add."

Please rank the following text blocks from 1 to 5, where 1 means the company is committed to sustainability and 5 means the company is not committed to sustainability.

1. "XXX meets global societal challenges such as climate protection, energy efficiency and resource management with innovative products XXX's climate target is to become carbon-neutral by 2030. Emissions are to be cut by 70 percent over the 2019 levels by 2025 We understand sustainability as the symbiosis between economy, ecology and social engagement, continuously respecting and recognizing the importance of cultural diversity."

2. "As we transform into a sustainable mobility tech company, we are guided by the clear vision of our Dare Forward 2030 strategic plan, which paves the way for XXX to achieve carbon net zero by 2038 We are leading the charge in electrification and software development, with cutting-edge technologies at the heart of our products and services With its ambitious decarbonization strategy, XXX is leading the industrys transition toward a carbon net zero future."

3. "At XXX, we are committed to Diversity, Equity & Inclusion. We recognize that an equitable and inclusive workplace culture where people are encouraged to bring their whole selves to work benefits everyone: our associates, our customers and our communities Find out how XXX is helping improve the lives of the people who live and work in our communities From the XXX Rewards program to our new Wellness Stores, XXX offers the tools and services you need to take a more active role in managing your familys well-being."

4. "We want to stay connected to our patients, their families and our communities, and to hear your voices We are seeking to create respectful and positive communities on social media XXX is not responsible for and does not assume any liability for any third-party content."

5. "The XXX brand is committed to sustainability, with features like the EcoDiesel design and Quadra-Lift technology that help reduce emissions and improve fuel efficiency In addition to its rugged capabilities, the XXX Wrangler Freedom Edition also supports a good cause, with a portion of its sales going towards the USO's Operation Enduring Care program The XXX Grand Cherokee Summit offers luxury and sustainability, with features like LaneSense and ParkView that help reduce accidents and improve overall efficiency."

A.2. Sustainability criteria

Sustainability criteria:

1. No importance to sustainability:
- No mention of sustainability initiatives, policies or objectives on the website.
- No information on environmental management, social responsibility or sustainability reporting.
- No involvement in sustainability programs, partnerships or industry initiatives.

2. Minimal commitment to sustainability:
- Limited mention of sustainability practices, possibly only superficial.
- Basic information about environmental management without specific goals or measurable indicators.
- Low involvement in social initiatives or limited transparency about social impact.

3. Average commitment to sustainability:
- Clear mention of sustainability initiatives, policies and objectives on the website.
- Moderate performance data on environmental management and social responsibility, possibly with some transparency.
- Some involvement in sustainability programs, although not leading within the sector.

4. Significant commitment to sustainability:
- Detailed description of sustainability strategies, including measurable goals and performance indicators.
- Strong performance data on environmental management and social responsibility, with clear reporting on progress.
- Active involvement in sustainability initiatives and partnerships, with impact within the sector.

5. Excellent commitment to sustainability:
- Progressive sustainability policies, including innovative approaches and best practices.
- Excellent performance data in environmental management, social responsibility and transparent reporting.
- Sustainability leadership, with active involvement in cross-sector initiatives and positive influence on the industry.