Jubel.be | Powered by KNOPSPublishing.com

NIEUWE TECHNOLOGIEËN EN HET RECHT: DE IMPACT VAN ARTIFICIËLE INTELLIGENTIE OP DE RECHTSPRAKTIJK

DEEL III: Natural language processing (NLP)

Rémy Bonnaffé [1]

Jubel.be, 7 november 2019.

[1] Advocaat bij Freshfields Bruckhaus Deringer.

Natural language processing (NLP): dé artificiële intelligentie voor de vennootschapsjurist

In onze vorige bijdragen bespraken we voornamelijk *expert systems* en het toepassen van *machine learning* op gestructureerde data. Gestructureerde data zijn data die bestaan uit duidelijk gedefinieerde datatypes waarbij de tendensen in de data makkelijk doorzoekbaar zijn. Dit klinkt abstract, maar uiteindelijk komt het neer op de vraag of we data al dan niet in een tabel kunnen plaatsen. Figuur 2 is een goed voorbeeld van gestructureerde data, waarbij de data nauwkeurig is georganiseerd in een tabel met rijen en kolommen. Ongestructureerde data zijn in essentie alle data dat niet gestructureerd is. Het belangrijkste voorbeeld van ongestructureerde data is tekst. Om artificiële intelligentie toe te passen op tekstuele data moet men toepassing maken van het verwerken van natuurlijke taal (*natural language processing* of NLP). NLP is het "*subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content*"[2].

Het gebruiken van *machine learning* in de context van gestructureerde data is bijzonder nuttig, zoals in Figuur 2 reeds werd geïllustreerd met uitkomsten van rechtszaken (de uitkomst van rechtszaken is gestructureerde data omdat elke rechtszaak een rij is in een tabel en waarbij elke rij een kolom heeft met de uitkomst van die rechtspraak). Niettegenstaande het voorgaande, is het belangrijk te benadrukken dat de echte kerngegevens wat betreft het juridisch beroep vaak tekstuele data zal zijn. Advocaten werken voortdurend met juridische documenten: contracten, rechtsuitspraken, wetsbepalingen, enzovoort. Om de kracht van artificiële intelligentie en meer in het bijzonder *machine learning* volledig te kunnen gebruiken in een juridische context is het daarom cruciaal om *machine learning* te kunnen toepassen op tekstuele data. In dit hoofdstuk bespreken we hoe *machine learning*-modellen gebruikt kunnen worden in het raamwerk van tekstuele data.

Toepassen van *machine learning* op tekstuele data

Om *machine learning* toe te passen op tekstuele data zijn meerdere stappen noodzakelijk[3]. Een eerste cruciale stap die niet over het hoofd gezien kan worden is het verzamelen van de tekstuele data op zich. Zoals eerder werd uiteengezet is de hoeveelheid data die gebruikt kan worden door een *machine learning*-model cruciaal om zulk model goed te kunnen trainen. In een vennootschapsrechtelijke context kan het evenwel niet altijd evident zijn om de nodige tekstuele data te verzamelen. In dit verband zijn wetteksten de meest eenvoudige tekstuele data die verzameld kunnen worden. Zodra men echter *machine learning* wil toepassen op andere zaken zoals rechtsuitspraken of zelfs private contracten als *share purchase agreements*, stoot men vaak op het probleem dat deze documenten niet publiek worden

[2] J. HIRSCHBERG en C. MANNING, "Advances in Natural Language Processing", *Science* 2015 261, 261.
[3] E. TALLEY, "Is The Future of Law a Driverless Car? Assessing How the Data Analytics Revolution Will Transform Legal Practice", https://ssrn.com/abstract=3064926, 2017, 11.

gemaakt. In grotere jurisdicties zoals de Verenigde Staten of het Verenigd Koninkrijk is er in dat verband veel meer tekstuele data beschikbaar. Rechtsuitspraken zijn veel meer publiek raadpleegbaar en ook private contracten, zoals *share purchase agreements*, worden vaker openbaargemaakt bv. omdat deze openbaar gemaakt moeten worden als gevolg van de notering van een bepaalde vennootschap (waarvan er in elk geval grotere aantallen zijn dan in België). Indien we in België aan de slag willen gaan met artificiële intelligentie in een juridische context en indien we willen dat deze artificiële intelligentie rekening houdt met de Belgische eigenheid, dan zullen alle actoren in het juridisch beroepsleven zich moeten afvragen hoe meer tekstuele data beschikbaar kunnen worden gesteld. Zo wordt in België minder dan twee procent van de rechtspraak gepubliceerd[4].

In het resterende deel van dit hoofdstuk bespreken we de twee belangrijkste stappen die genomen moeten worden eens voldoende tekstuele data verzameld werd. De eerste stap is de data 'opschonen' en, meer in het bijzonder, de data 'normaliseren' en de '*tokenisation*' toepassen. Het eindresultaat voor wat betreft het toepassen van *machine learning* op tekstuele data is in essentie dezelfde als haar toepassing voor gestructureerde data: we willen tendensen in de data herkennen. Voor tekstuele data betekent dat dat we pogen om statistische correlaties tussen woorden te vinden gepaard met de frequentie van dergelijke woorden.

Om op een degelijke manier tendensen in tekstuele data te vinden moeten we sommige woorden normaliseren. Het normaliseren houdt in dat alle hoofdletters worden weggenomen en dat alle woorden herleid worden tot hun stam (het zogenaamde *stemming*). Dit om oppervlakkige variatie in de woorden te minimaliseren[5]. Geen enkel persoon heeft het moeilijk om te identificeren dat 'Lopen', 'lopen', 'liep', 'gelopen' variaties zijn van hetzelfde woord. Zonder het normaliseren van deze woorden zal een computer deze woorden kwalificeren als woorden met een verschillende inhoud. Door het proces van normalisatie zullen alle variaties herleid worden tot 'lopen'. Daarenboven zal normalisatie er ook toe leiden dat stopwoorden worden verwijderd, met name woorden die met een hoge frequentie gebruikt worden maar die geen substantiële informatie inhouden. Voorbeelden zijn functiewoorden (dit, een, met) en voornaamwoorden (zij, het, de). Ten slotte zal *tokenisation* – het proces waarbij één of meerdere woorden als één eenheid worden beschouwd – één (*unigrams*) of meerdere (n-grams) woorden samennemen in een *token*: "De Verkoper verkoopt aandelen" wordt 'verkoper' 'verkoopt' 'aandelen' (op voorwaarde dat 'de' verwijderd wordt gezien het een stopwoord is).

[4] KU LEUVEN, "Opzoeken van Belgisch Recht", *KU Leuven* (25 oktober 2018), https://bib.kuleuven.be/rbib/collectie/zoekleidraad/belgisch-recht.

[5] K. ASHLEY, *Artificial Intelligence and Legal Analytics*, Cambridge, Cambridge University Press, 2017, 236.

Het voorgaande voorbeeld is een *unigram*, waarbij één woord één token vertegenwoordigd. N-grams laten toe dat twee (*bigrams*), drie (*trigrams*) of meerdere woorden in één token gestoken worden. N-grams zijn belangrijk omdat ze er beter toe in staat zijn om de complexiteit van onze natuurlijke taal vast te leggen[6]. Een eenvoudig voorbeeld is de volgende zin, waarvan we het sentiment willen achterhalen: "Ik ben niet goed in tennis, ik ben er zeer slecht in". Als we gebruik maken van *unigrams* ('ik' 'ben' 'niet' 'goed' 'in' 'tennis', 'ik' 'ben' 'er' 'zeer' 'slecht' 'in') dan hebben we zowel de termen 'goed' als 'slecht'. Daardoor elimineren beide woorden elkaars sentiment. Als we in de plaats daarvan gebruik maken van *bigrams* ('ik ben' 'ben niet' 'niet goed' 'goed in' 'in tennis' 'tennis ik' 'ik ben' 'ben er' 'er zeer' 'zeer slecht' 'slecht in') dan vangen we zowel de termen 'niet goed' als 'zeer slecht'. Gezien beiden een gelijkaardige betekenis hebben en gezien deze betekenis overweegt in deze zin, is het makkelijker voor een *machine learning* die toepassing maakt van *bigrams* om het sentiment van deze zin te achterhalen.

Nu dat we het makkelijker gemaakt hebben om de frequentie en de onderlinge relatie tussen de woorden te achterhalen in tekstuele data, is de volgende stap om deze data te structureren op een wijze die het mogelijk maakt om *machine learning* toe te passen. Om de tekstuele data te structureren wordt veelal gebruik gemaakt van een *document-term matrix*. Dit is een spreadsheet-achtig document waarbij elke instantie georganiseerd wordt in rijen en waarbij de *unigrams* of *n-grams* georganiseerd worden in kolommen[7]. Belangrijk daarbij is dat *alle unigrams* of *n-grams* doorheen alle instanties in de kolommen wordt geplaatst. Een instantie kan eender wat zijn, en hangt af van de finaliteit van de oefening. Indien we bijvoorbeeld willen achterhalen welke paragrafen *change of control*-paragrafen zijn, dan is elke instantie een paragraaf. Naast paragrafen is het uiteraard ook mogelijk om volledige documenten te gebruiken als instanties, zoals contracten of rechterlijke uitspraken. Figuur 3 illustreert hoe een *document-term matrix* eruitziet (in Figuur 3 bestaat de *document-term matrix* uit vijf documenten, aangeduid als D1 tot D5).

[6] E. ALPAYDIN, *Machine Learning*, Cambridge, MIT Press, 2016, 69.

[7] K. ASHLEY, *Artificial Intelligence and Legal Analytics*, Cambridge, Cambridge University Press, 2017, 238.

- D1: *Hickory, dickory, dock;*
- D2: *The mouse ran up the clock;*
- D3: *The clock struck one;*
- D4: *The mouse ran down;*
- D5: *Hickory, dickory, dock.*

	hickory	dickory	dock	the	mouse	ran	up	clock	struck	one	down
D1	1	1	1	0	0	0	0	0	0	0	0
D2	0	0	0	2	1	1	1	1	0	0	0
D3	0	0	0	1	0	0	0	1	1	1	0
D4	0	0	0	1	1	1	0	0	0	0	1
D5	1	1	1	0	0	0	0	0	0	0	0

Figuur 3 Unigram document-term matrix[8]

Het gevolg van het opzetten van een *document-term matrix* is dat ongestructureerde data omgezet wordt tot gestructureerde data. Door deze gestructureerde data is het mogelijk om statistische modellen toe te passen om tot bepaalde inzichten te komen en die het computersysteem ertoe in staat stelt te 'leren' zoals eerder in dit artikel werd uiteengezet. Ook hier kan gebruik gemaakt worden van zowel gesuperviseerd als niet-gesuperviseerd leren.

Gebruik van NLP in het kader van fusies, overnames en andere juridische transacties

Nu we een overzicht hebben hoe artificiële intelligentie, en *machine learning* meer in het bijzonder, zich verhouden tot tekstuele data, kunnen we nagaan hoe deze technieken toegepast kunnen worden op het juridisch beroep. Momenteel is er een explosie van commerciële toepassingen die worden ontwikkeld en die gebruik maken van NLP om het werk van de advocaat te vereenvoudigen of zelfs, in sommige gevallen, de advocaat voor een bepaalde taak volledig vervangt en de dienstverlening onmiddellijk aan de eindgebruiker verleent. Een van de meest prominentie toepassingen van *machine learning* en NLP in een vennootschapsrechtelijke context zijn de computertoepassingen die de vennootschapsjurist helpen bij het *due diligence* onderzoek. Het *due diligence* onderzoek is het onderzoek waarbij "op meer of minder diepgaande wijze beschikbare informatie omtrent een vennootschap of groep van vennootschappen bestudeerd [wordt] om inzicht te krijgen in de waarde van de onderliggende onderneming en de eventuele aansprakelijkheden en risico's die er, in het licht van de voorgenomen transactie, aan verbonden zijn"[9].

Typerend bij een dergelijk *due diligence* onderzoek is dat de advocaat in kwestie een grote hoeveelheid contracten of andere juridische documenten zal moeten doornemen, waarvan een reeks documenten gegroepeerd kunnen worden in documenten met een gelijkaardige

[8] E. TALLEY, "Is The Future of Law a Driverless Car? Assessing How the Data Analytics Revolution Will Transform Legal Practice", https://ssrn.com/abstract=3064926, 2017, 13.

[9] Y. VERLEISDONK, E. JANSSENS en M. WILKENHUYSEN, *Due Diligence*, Brussel, Larcier, 2011, 1-2.

inhoud. Als de vennootschap, bijvoorbeeld, goederen verkoopt en levert, is het gebruikelijk dat de commerciële overeenkomsten zullen onderzocht worden op bepaalde risico's. Deze commerciële overeenkomsten kunnen talrijk zijn en omvatten vaak dezelfde soort bepalingen, zoals de duurtijd van de overeenkomst en de al dan niet aanwezigheid van een *change of control*-bepaling, enzovoort. Vandaag de dag is het niet ongebruikelijk voor de advocaat om door deze documenten te gaan om (A) een overzicht op te stellen van deze overeenkomsten met de belangrijkste bepalingen, en/of (B) op zoek te gaan naar specifieke risico's (zoals de aanwezigheid van een *change of control* bepaling). Sommige (voornamelijk internationale) advocatenkantoren hebben innovaties ingevoerd (innovatie hoeft namelijk niet noodzakelijk gepaard te gaan met technologie) tijdens de laatste jaren, waarbij binnen het kantoor een afzonderlijke afdeling wordt opgericht en waarbij de werknemers binnen deze afdeling specifiek worden getraind om *due diligence* onderzoeken uit te voeren. Het bijzondere daarbij is dat de personen die het effectieve *due diligence* werk uitvoeren, of alleszins delen daarvan, geen advocaten zijn. Werk dat voordien door juristen werd uitgeoefend, wordt bijgevolg geoutsourcet naar niet-juristen.

Zoals in het voorgaande hoofdstuk echter werd uiteengezet, kan de hoeveelheid aan contracten en andere juridische documenten in een *due diligence* beschouwd worden als data. En door het gebruik van NLP kunnen we op deze data bijgevolg ook *machine learning* toepassen. Dit laat ons toe om, eens we een voldoende hoeveelheid data hebben, een computer aan te leren om deze data te herkennen (met name door het vinden van bepaalde patronen). Het herkennen van bepaalde contractuele bepalingen is net de taak die de jurist uitvoert tijdens een *due diligence* onderzoek. Gezien bepaalde contractuele bepalingen (denk aan de duurtijd van een overeenkomst of de *change of control*) zodanig veel voorkomen in overeenkomsten, is het relatief eenvoudig voor *machine learning*-toepassingen om deze bepalingen te herkennen.

De mogelijkheid van een *machine learning*-toepassing om bepaalde contractuele bepalingen te herkennen, zorgt ervoor dat een substantieel deel van het werk van de advocaat verlicht kan worden voor wat betreft de twee taken die eerder reeds werden aangehaald, met name documenten doornemen om (A) een overzicht op te stellen van deze overeenkomsten met de belangrijkste bepalingen, en/of (B) op zoek te gaan naar specifieke risico's. Voor wat betreft de eerste taak, zijn *machine learning*-toepassingen zoals Kira, Luminance en eBravia ertoe in staat om door een reeks gelijkaardige documenten te gaan (zoals commerciële overeenkomsten of huurovereenkomsten) en van dergelijke overeenkomsten een overzicht op te stellen in Word of eender welke andere uitvoermogelijkheid, met bijvoorbeeld daarin begrepen: de duurtijd van de overeenkomst, de (huur)prijs, opzegmogelijkheden, *change of control*, enzovoort. Dit gebeurt allemaal in fracties van seconden en dit voor honderden tot duizenden documenten. Voor wat betreft de tweede taak, is het ook mogelijk voor dezelfde

computertoepassingen om op zoek te gaan naar deze specifieke bepalingen in de overeenkomsten zonder noodzakelijkerwijze een overzicht op te stellen. Als de advocaat in kwestie bijvoorbeeld de overeenkomsten wil doornemen die een duurtijd hebben van minder dan een jaar, kan de advocaat deze meteen opvragen zonder manueel naar deze documenten te zoeken. Wederom gebeurt dit in enkele seconden, waardoor het werk van de advocaat in kwestie veel efficiënter wordt.

Het is belangrijk daarbij op te merken dat het momenteel nog niet aan de orde is dat deze computertoepassing de taak van de advocaat volledig overneemt. De toepassingen zijn momenteel nog niet geavanceerd genoeg dat deze toepassingen de taken volledig foutloos kan uitvoeren (hoewel de foutmarge van menselijke experten soms onderschat wordt, zoals eerder werd uiteengezet). Daarenboven is het zo dat ook slechts een beperkt aantal clausules voldoende gestandaardiseerd zijn (en dus voldoende data omvat) om *machine learning* te kunnen toepassen. Het gebruik van NLP in het *due diligence*-onderzoek moet daardoor eerder gezien worden als een hulpmiddel dat het mogelijk maakt voor de advocaat om een eerste schifting te maken. Op basis van deze eerste schifting kan de advocaat vervolgens verder met de doorlichting van de relevante documenten. Deze eerste schifting zal echter in ieder geval een tijdswinst en dus ook een verminderde kost met zich meebrengen.

Een laatste kanttekening moet nog gemaakt worden bij het gebruik van NLP in het kader van een *due diligence* onderzoek en dat in het bijzonder relevant is voor het meertalige België. Hierboven werd reeds vermeld dat overeenkomsten of andere juridische documenten gekwalificeerd kunnen worden als data. Hoe meer data, hoe beter *machine learning* en NLP er in staat toe zijn patronen te vinden in de data en dus bijgevolg accurater zijn. Een vereiste hierbij is evenwel dat de tekstuele data in een en dezelfde taal geschreven zijn. Indien eenzelfde document in een andere taal wordt opgesteld, dan wordt dit door een *machine learning*-model als een verschillend document beschouwd (daar waar een menselijke expert de gelijkenis van het document zou herkennen). Het gevolg is dat indien eenzelfde document of bepaling in bijvoorbeeld twee talen voorkomt, de data als het ware 'gehalveerd' wordt. Bijgevolg wordt de accuraatheid en de mogelijkheid van het *machine learning*-model om te leren ook gehalveerd.

Dit probleem speelt dan ook in het bijzonder in België, waarbij de Belgische jurisdictie getypeerd wordt door op z'n minst twee talen, Frans en Nederlands, en veelal ook door het Engels. Daardoor wordt het enerzijds moeilijker voor een *machine learning*-toepassing om eventuele 'typisch' Belgische bepalingen aan te leren. Anderzijds wordt het zo ook moeilijker om NLP toe te passen op documenten die in een bepaalde taal zijn geschreven, zoals in het Frans maar vooral in het Nederlands. Engels wordt immers veel meer gebruikt in een internationale juridische context dan het Frans, en zeker in vergelijking met het Nederlands.

De NLP-toepassingen zoals Kira, Luminance en eBravia worden eveneens ontwikkeld in landen (de VS en het VK) waar voornamelijk Engels wordt gesproken. Hoewel deze toepassingen nu ook beginnen met het ondersteunen van het Frans, zal het nog enige tijd duren vooraleer Nederlands eveneens ondersteund wordt. Nederlandse en Franse documenten moeten vandaag de dag bijgevolg nog steeds manueel doorzocht worden. Wel is het mogelijk voor de NLP-toepassingen om alle documenten te identificeren op taal, en deze documenten automatisch in een map te plaatsen. Indien bijgevolg een internationaal *due diligence*-onderzoek wordt uitgevoerd, is het eenvoudig om alle documenten te groeperen per taal en deze documenten vervolgens toe te wijzen aan de relevante juristen (of niet-juristen) die de taal machtig zijn.

Conclusie: heb ik als vennootschapsjurist nog een toekomst in een wereld met artificiële intelligentie?

Artificiële intelligentie is zich momenteel razendsnel aan het ontwikkelen. Twee technologieën binnen het ruime onderzoeksveld rond artificiële intelligentie die van bijzonder belang zijn voor de rechtspraktijk zijn *machine learning* en *natural language processing* of NLP. De laatste overblijvende vraag is of deze technologieën echter voldoende geavanceerd zijn om de jobs van juristen en advocaten over te nemen. Het antwoord luidt: het hangt ervan af. Of artificiële intelligentie in staat is om de jobs van juristen en advocaten over te nemen hangt af van de specifieke taak die de jurist of advocaat uitoefent. Het is bijgevolg cruciaal om te begrijpen welke soort taken een jurist of advocaat op een dagdagelijkse basis uitvoert om de impact van artificiële intelligentie te kunnen inschatten. Belangrijk daarbij is de realisatie dat het werk van een jurist of een advocaat opgedeeld kan worden.

Een vennootschapsjurist kan bijvoorbeeld een van de volgende taken uitvoeren in een typische M&A-transactie: *due diligence*, opzoekingswerk, beheren van de transactie, selecteren van standaarddocumenten en modellen, onderhandelen, documenten herwerken op basis van de behoeften van de cliënt, beheer van documenten, juridisch advies en risicoanalyse[10]. Zoals we in het vorige hoofdstuk hebben bekeken, kan het *due diligence*-proces een pak efficiënter en minder tijdrovend gemaakt worden met behulp van artificiële intelligentie. Dat wil natuurlijk niet zeggen dat de job van de vennootschapsjurist binnen aanzienbare tijd zal verdwijnen. De advocaat zal nog steeds een cruciale rol spelen in een transactie voor wat betreft de taken die nog niet vervangen kunnen worden door artificiële intelligentie, zoals het onderhandelen van belangrijke bepalingen van een aandelenverkoopovereenkomst. Het is wel belangrijk te beseffen dat hoe meer geavanceerd

[10] R. SUSSKIND, *Tomorrow's Lawyers*, Oxford, Oxford University Press, 2013, 30-32.

artificiële intelligentie wordt, hoe meer van deze taken (gedeeltelijk) door artificiële intelligentie gedaan kunnen worden.

We kunnen concluderen dat de vraag of artificiële intelligentie de vennootschapsjurist kan vervangen nu nog niet aan de orde is. De vraag die we als vennootschapsjurist momenteel moeten stellen is eerder hoe we de efficiëntie kunnen maximaliseren die resulteren uit het gebruikmaken van artificiële intelligentie, en dit ten bate van de cliënt. Zoals Daniel Katz schrijft: "the equation is simple: Humans + Machines > Humans or Machines"[11].

[11] D. KATZ, "Quantitative Legal Prediction – or – How I learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry", *Emory Law Journal* 2013, 909, 929.