Outomatiese genreklassifikasie vir Afrikaans

Authors:
Dirk Snyman[1]
Gerhard van Huyssteen[1]
Walter Daelemans[2]

Affiliations:
[1]Centre for Text Technology, North-West University, South Africa

[2]Computational Linguistics and Psycholinguistics Research Group, University of Antwerpen, Belgium

Correspondence to:
Dirk Snyman

Email:
dirk.snyman@nwu.ac.za

Postal address:
Private Bag X6001, Potchefstroom 2520, South Africa

Dates:
Received: 08 Aug. 2013
Accepted: 30 Apr. 2014
Published: 24 Nov. 2014

How to cite this article:
Snyman, D., Van Huyssteen, G. & Daelemans, W., 2014, 'Outomatiese genreklassifikasie vir Afrikaans', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 33(1), Art. #759, 12 pages. http://dx.doi.org/10.4102/satnt.v33i1.759

Op die terrein van teksverwerking speel die metadata oor 'n bepaalde teks in baie gevalle 'n belangrike rol. Sodanige metadata word dikwels toegevoeg met behulp van outomatiese teksklassifiseerders wat op grond van die inhoud van 'n teks een of meer vooraf bepaalde klasse of kategorieë outomaties aan 'n teks toeken. Een van die dimensies waarvolgens 'n teks geklassifiseer kan word, is die genre van 'n teks en in hierdie studie word die ontwikkeling van 'n outomatiese genreklassifikasiesisteem in 'n hulpbronskaars omgewing voorgehou. (Ander dimensies sluit in: outeur van 'n teks, domein van tekste, informele teenoor formele tekste, ensovoorts.) Die artikel het ten doel om 'n eksperimentele ondersoek te loods na bestaande genreklassifikasiesisteme, en om dan die tegnieke en benaderings te implementeer vir Afrikaans (as voorbeeld van 'n hulpbronskaars taal). Met die ontwikkeling van 'n outomatiese genreklassifikasiesisteem is daar 'n reeks veranderlikes wat in gedagte gehou moet word en wat 'n invloed op die prestasie van masjienleerbenaderings het (d.i. die algoritme wat gebruik word, die hoeveelheid afrigtingsdata, en die datavoorstelling as eienskappe). As dié veranderlikes reg hanteer word en 'n optimale versameling van hierdie veranderlikes geïdentifiseer kan word, kan die ontwikkeling van 'n genreklassifikasiesisteem suksesvol gedoen word. In die studie word daar 'n genreklassifikasiesisteem daargestel deur gebruik te maak van die volgende benadering wat eksperimenteel geïdentifiseer is: Die implementering van 'n MNB-algoritme, afgerig met woordversamelingbenadering as eienskapstel. Dié sisteem lewer 'n resulterende *f*-telling (prestasiesyfer) van 0.929.

Automatic genre classification for Afrikaans. When working in the terrain of text processing, metadata about a particular text plays an important role. Metadata is often generated, using automatic text classification systems which classify a text into one or more predefined classes or categories based on its contents. One of the dimensions by which a text can be can be classified, is its genre. In this study the development of an automatic genre classification system in a resource scarce environment is postulated. This study aimed to investigate the techniques and approaches that are generally used for automatic genre classification systems, and identify the best approach for Afrikaans (a resource scarce language). With the development of an automatic genre classification system, there is a set of variables that must be considered as they influence the performance of machine learning approaches (i.e. the algorithm used, the amount of training data, and data representation as features). If these variables are handled correctly, an optimal combination of them can be identified to successfully develop a genre classification system. In this article a genre classification system is being developed by using the following approach: The implementation of a MNB algorithm with a bag of words approach feature set. This system provides a resultant *f*-score (performance measure) of 0.929.

Inleiding

Hulpbronskaarste in die rekenaarlinguistiek is 'n onderwerp wat baie aandag geniet, beide plaaslike en internasionaal. Kongresse soos die *Association for Computing Machinery* se *Annual symposium on computing for development* fokus op die problematiek van ontwikkelende omgewings en die hulpbronskaarste van die tale in sulke omgewings, wat 'n gereelde onderwerp is van die navorsing wat daar bespreek word. Meta-net (http://www.meta-net.eu/whitepapers/overview) van die *Multilingual Europe Technology Alliance* het 'n reeks witskrifte wat die toekoms van die Europese tale aan die hand van die beskikbare tegnologiehulpbronne bespreek. In hierdie reeks word genoem dat sommige van dié tale in die digitale omgewing kan uitsterf weens 'n gebrek aan digitale hulpbronne. Ontbrekende digitale hulpbronne veroorsaak dat daar eerder uit die staanspoor van Engels gebruik gemaak word wat kleiner tale (soos Afrikaans) verdring. Hulpbronskaarste is dus 'n groot probleem waarvoor oplossings ernstig gesoek moet word. Hierdie probleem is weereens ter sprake by die ontwikkeling van 'n outomatiese genreklassifikasiesisteem vir 'n hulpbronskaars taal, wat 'n komplekse proses is waar 'n groot

aantal veranderlikes in gedagte gehou moet word. In dié navorsing word daar 'n benadering geïdentifiseer wat die veranderlikes saamvat en uiteindelik 'n suksesvolle, werkende genreklassifikasiesisteem daarstel vir Afrikaans, gegewe die beperkings van 'n hulpbronskaars omgewing. Volgens Grover, Van Huyssteen en Pretorius (2011) kan die hulpbronskaarste van 'n taal bepaal word deur 'n kritiese evaluering van die beskikbare digitale hulpbronne en verwante tegnologieë. Grover *et al*. (2011) vergelyk die Suid-Afrikaanse tale met 'n minimum verwagte versameling van hulpbronne om basiese navorsing oor mensetaaltegnologie te beoefen, en aan die hand daarvan kan gesê word dat Afrikaans as 'n hulpbronskaars taal beskou word.

Op die terrein van teksverwerking speel die metadata van 'n bepaalde teks in baie gevalle 'n belangrike rol. Cardinaels, Meire en Duval (2005) stel dat dit sonder toepaslike metadata moeilik of selfs onmoontlik sal wees om elektroniese leerinhoud outomaties te identifiseer en te onttrek. Wanneer daar byvoorbeeld korpusse saamgestel word vir toepassings vir natuurliketaalprosessering is dit dikwels nodig om te weet uit watter genres en domeine (as voorbeelde van metadata) die data afkomstig is, ten einde te verseker dat die korpus saamgestel word uit 'n wye verskeidenheid tekste. As 'n korpus slegs uit een of twee domeine saamgestel word, word die spreiding negatief beïnvloed en is die korpus nie meer verteenwoordigend nie. Sou die spreiding skeefgetrek wees, sal die sisteme en/of eksperimente wat gebaseer word op die korpus se veronderstelde graad van verteenwoordigendheid, onakkurate resultate lewer. In natuurliketaalprosessering wil 'n mens hierdie metadata ook dikwels outomaties toevoeg tot tekste – byvoorbeeld of 'n teks uit 'n bepaalde domein kom of nie, of dit gemorspos is of nie, of dit deur 'n bepaalde outeur geskryf is of nie, of dit tot 'n bepaalde genre behoort of nie, ensovoorts.

Dit is 'n algemene praktyk dat korpusse volgens 'n reeks genres gestratifiseer word. Voorbeelde hiervan is die Brown-korpus (Francis & Kucera 1979), asook die PAROLE-korpus (Instituut vir Nederlandse Leksikologie 2005). Deur 'n korpus te stratifiseer, word die graad van verteenwoordigendheid daarvan verseker. Wanneer 'n korpus saamgestel word, moet tekste geanaliseer word om die genre daarvan vas te stel voordat dit by die korpus gevoeg word. As die genre van 'n teks bekend is, kan daar na 'n opsomming van al die tekste gekyk word om 'n duidelike oorsig te kry van of die korpus verteenwoordigend genoeg is, of nie. Hierdie metadata oor 'n teks is egter nie altyd beskikbaar nie en omdat daar met groot hoeveelhede data gewerk word, is handmatige annotering van enige aard 'n arbeidsintensiewe en tydrowende aktiwiteit, wat dikwels omsit in hoë onkoste. As hierdie annotasie dus geoutomatiseer kan word, kan dit lei tot besparing van tyd en geld.

Sodanige metadata word dikwels toegevoeg met behulp van outomatiese teksklassifiseerders. 'n Teksklassifiseerder word gedefinieer as 'n sisteem wat op grond van die inhoud van 'n teks een of meer vooraf bepaalde klasse of kategorieë outomaties aan 'n teks toeken. Benaderings in statistiese patroonherkenning soos masjienleer en neurale netwerke word oor die algemeen gebruik om sulke klassifiseerders – byvoorbeeld genreklassifiseerders – af te rig (Goller *et al*. 2000).

Die primêre probleem wat in hierdie artikel bespreek word, is dat daar nie genreklassifikasiesisteme vir die Suid-Afrikaanse tale bestaan nie. Dit veroorsaak 'n dilemma vir die ontwikkeling van tekshulpbronne vir hierdie tale, veral waar die graad van verteenwoordigendheid belangrik is. Dié navorsing het dus die doel voor oë om 'n ondersoek te loods na bestaande genreklassifikasiesisteme en om dan die tegnieke en benaderings te implementeer vir Afrikaans (as voorbeeld van 'n hulpbronskaars taal).

Goller *et al*. (2000) onderskei tussen twee fases van outomatiese teksklassifikasie:

- Die eerste fase is die afrigtingsfase waar voorbeeldteks te van elkeen van die vooraf bepaalde klasse (handmatig of semi-outomaties) geklassifiseer word en dan as afrigtingsdata vir die sisteem gebruik word. Die sisteem lei dan die verskillende eienskappe van elke klas vanuit die afrigtingsvoorbeelde af deur statistiese inferensie, veralgemening, abstraksie, eksplisiete onttrekking, ensovoorts. Dit is dus belangrik dat die afrigtingsdata in die algemeen en die afrigtingstekste vir elke klas so verteenwoordigend moontlik moet wees. Afhangende van die benadering wat gevolg word, vereis sommige teksklassifikasiesisteme teenvoorbeelde vir elke klas wat dan as voorbeeld dien vir tekste wat definitief nié deel van die klas is nie. Elke teks wat aan die een klas toegeken word, dien outomaties as 'n teenvoorbeeld vir die ander klasse.
- Die tweede fase word die klassifikasiefase genoem waar voorheen onbekende tekste deur die masjienleeralgoritme geklassifiseer word. Die klassifikasiesisteem kan dan die klas van die invoerteks bepaal, of as die teks nie geklassifiseer kan word volgens die bepaalde klasse nie, kan dit as onbekend geklassifiseer word. Daar is egter dikwels heelwat voorverwerking van die invoerdata wat uitgevoer moet word voordat klassifikasie kan plaasvind, byvoorbeeld omskakeling van die invoertekste na die regte dataformaat.

Evaluasie word gedoen aan die hand van die presisie en herroeping, tesame met die resulterende *f*-telling van die sisteem. Hierdie evaluasiemetrieke word algemeen gebruik in inligtingherwinning en is die standaardmetode vir evaluasie van klassifiseerders (McCallum & Nigam 1998).

In die afdeling oor eksperimentele opstelling word die verskillende algoritmes wat algemeen vir genreklassifikasie gebruik word, voorgehou, tesame met datavoorstellings tegnieke vir eienskappe en klasse. Die resultate vir eksperimente met die verskillende algoritmes en die datavoorstellingstegnieke, asook verskillende kombinasies van die vooraf genoemde, word in die resultate beskryf.

Die optimering van die klassifiseerders en die uiteindelike optimale klassifiseerder word voorgehou. Die artikel sluit af met die samevatting van hierdie artikel.

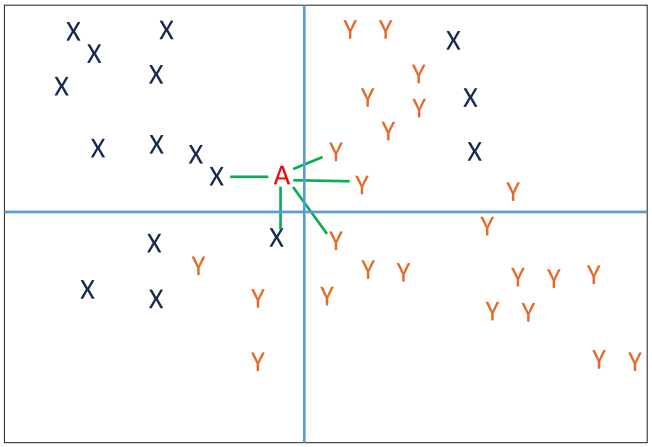
Eksperimentele opstelling

Algoritmes

In die literatuur is daar 'n paar verskillende masjienleerbenaderings wat algemeen gebruik word vir teksklassifikasie. Khan, Baharudin, Lee en Khan (2010) bied 'n oorsig van die mees vername benaderings, waarvan dié wat in ons eksperimente gebruik word hier kursories bespreek word, te wete *k*-naastebuurpuntklassifiseerders, steunvektorklassifiseerders, multinomiale naïewe Bayes-klassifiseerders, besluitnemingsbome en die RIPPER-algoritme. WEKA (Hall *et al*. 2009) is 'n geïntegreerde eksperimentele omgewing waar masjienleeralgoritmes en eienskappe maklik vergelyk kan word. Die onderstaande algoritmes wat in ons eksperimente gebruik word, dui telkens die implementering van die algoritme aan soos wat dit in WEKA beskikbaar is.

k-Naastebuurpuntklassifiseerders

Die *k*-naastebuurpuntbenadering (*k*-nn-benadering, Khan *et al*. 2010) is 'n geheuegebaseerde leermetode wat gebruik word om die ooreenkoms tussen onbekende tekste en die afrigtingsgevalle te bepaal. Die algoritme stel eienskappe wat uit die afrigtingsdata onttrek word in 'n multidimensionele ruimte voor. Die ruimte word dan ingedeel in verskillende areas wat deur die afrigtingsdata se klasse bepaal word. Wanneer 'n onbekende teks geklassifiseer moet word, word die teks ook as 'n punt in die ruimte gestip aan die hand van die betrokke teks se eienskappe. Die afstand tussen die onbekende teks en die *k*-naaste omliggende punte word dan bepaal (sien Figuur 1). Die mees frekwente klas onder die *k*-naastebuurpunte word vervolgens toegeken as die klas van die onbekende teks. Die *k*-nn-benadering is effektief en maklik implementeerbaar en vaar goed by klassifikasieprobleme met 'n wye reeks klasse.



Nota: Sien asb. Die volle literatuurverwysingslys van die artikel, Snyman, D., Van Huyssteen, G. & Daelemans, W., 2014, 'Outomatiese genreklassifikasie vir Afrikaans', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 33(1), Art. #759, 12 pages. http://dx.doi.org/10.4102/satnt.v33i1.759 , vir meer inligting.

FIGUUR 1: Grafiese voorstelling van die *k*-nn-benadering.

Die benadering is egter baie sensitief vir irrelevante eienskappe en uitskieters, en die teenwoordigheid hiervan in die afrigtingsdata kan die prestasie van die benadering ernstig benadeel (Khan *et al*. 2010).

Steunvektorklassifiseerders

Die steunvektorklassifiseerder (SVM) is gebaseer op die minimering van strukturele risiko (Khan *et al*. 2010). Struktureleriskominimering is 'n masjienleerbeginsel wat deur Vapnik (1995) soos volg verduidelik word: In masjienleer word 'n model saamgestel uit 'n eindigende datastel wat lei tot oormatige passing (d.i. die model word te spesifiek op die afrigtingstel gemodelleer en is nie meer algemeen genoeg om nuwe data te klassifiseer nie). Struktureleriskominimering minimeer die probleem van oormatige aanpassing deur die model se pasgemaakte kompleksiteit te balanseer met die model se vermoë om veralgemenings te hanteer.

Die hoofidee agter die benadering is om 'n hipotese te vind wat die laagste werklike fout waarborg, dit is die beste skeiding bepaal tussen die verskillende klasse van die klassifikasieprobleem. Die SVM bepaal dan 'n vlak in die ruimte wat die punte van die negatiewe en positiewe voorbeelde die beste verdeel deur 'n liniêre skeiding te maak. Hierdie vlak word die besluitnemingsvlak genoem. Die onbekende teks word ook in dié *n*-dimensionele ruimte voorgestel, en afhangend van die posisie van die onbekende tekste teenoor die besluitnemingsvlak, in die positiewe of negatiewe voorbeelde, word die klassifikasie bepaal. 'n Groot beperking van SVM's is dat so 'n klassifiseerder slegs 'n binêre klassifikasieprobleem kan hanteer. In 'n geval waar daar meer as twee klasse is waartussen die klassifiseerder moet onderskei, sal elke klas met elke ander klas vergelyk moet word, en aan die hand van 'n gewigtoekenningskema moet die beste klas vir die klassifikasie gekies word. Dit kan 'n geweldige toename tot gevolg hê in die aantal klassifiseerders wat uiteindelik in 'n multiklasomgewing gebruik word om 'n bepaalde klas aan 'n teks toe te ken. Dié toename in kompleksiteit veroorsaak 'n toename in die gebruik van fisieke geheue en verwerkingskrag. Die afrigtingsfase en klassifikasie neem ook derhalwe langer. Verder word daar verwarring opgemerk tydens klassifikasie, omdat 'n klomp verskillende klasse aan die teks toegeken word met elke iterasie van vereenduidiging tussen die klasse.

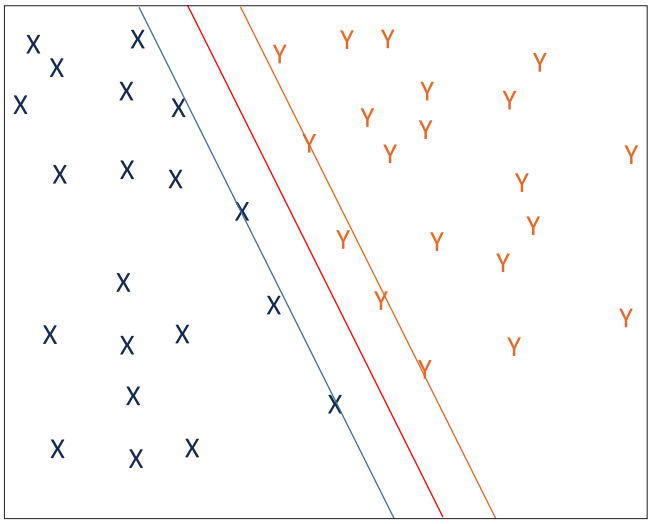
Figuur 2 toon die skeiding tussen die positiewe en negatiewe afrigtingsgevalle saam met die ondersteuningsvektore.

Multinomiale naïewe Bayes-klassifiseerders

Naïewe Bayes-klassifiseerders is gebaseer op 'n eenvoudige toepassing van Bayes se wet vanuit die waarskynlikheidsleer. Bayes se wet word deur Vergelyking 1 voorgestel.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{[Vgl. 1]}$$

Bayes se wet verdeel die voorwaardelike waarskynlikheid van 'n onbekende gebeurtenis in 'n paar kleiner waarskynlikhede,



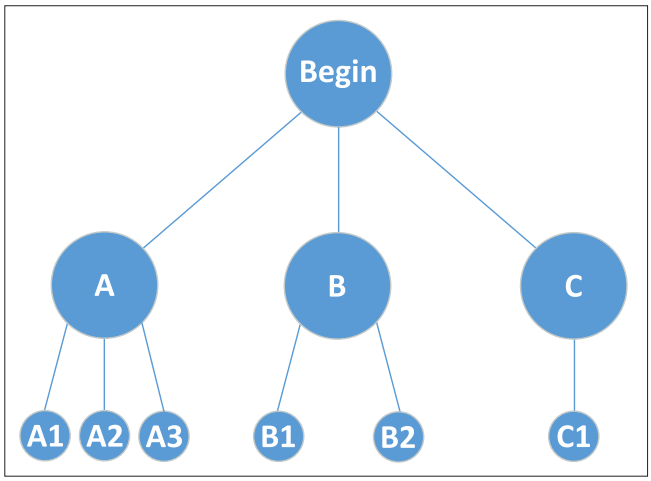
Nota: Sien asb. die volle literatuurverwysingslys van die artikel, Snyman, D., Van Huyssteen, G. & Daelemans, W., 2014, 'Outomatiese genreklassifikasie vir Afrikaans', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 33(1), Art. #759, 12 pages. http://dx.doi.org/10.4102/satnt.v33i1.759 , vir meer inligting.

FIGUUR 2: Skeiding van afrigtingsgevalle en ondersteuningsvektore in steunvektorklassifiseerder.

wat makliker is om te bereken. Hierdie ontbondeling van voorwaardelike waarskynlikheid vereenvoudig die taak van teksklassifikasie. Vir die gebruik van Bayes se wet vir teksklassifikasie word die waarskynlikheid van elke klas, gegewe die inhoud van die onbekende teks, bereken en word Vergelyking 1 dus soos in Vergelyking 2 herskryf (met onbekende teks W en die onbekende klas C).

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)} \qquad \text{[Vgl. 2]}$$

Naïewe Bayes-klassifiseerders is al suksesvol gebruik in vele domeine (Yi-Hsing & Hsiu-Yi 2008; Peng & Schuurmans 2003), ondanks die eenvoud van die model en die sterk onafhanklikheidsaannames wat dit maak. Peng en Schuurmans (2003) stel dat naïewe Bayes-klassifiseerders egter byna optimale prestasie kan bereik, al konformeer die domein onder bespreking glad nie tot die onafhanklikheidsaannames nie. McCallum en Nigam (1998) stel dat 'n variant van die klassieke, naïewe Bayes-klassifiseerder, naamlik die multinomiale naïewe Bayes-klassifiseerder (MNB) meer geskik is vir gebruik as 'n teksklassifiseerder as die klassieke naïewe Bayes-benadering. Die MNB-klassifiseerder is 'n aanpassing van standaard-naïewe Bayes, waar woordfrekwensies ook in ag geneem kan word (McCallum & Nigam 1998). Die belangrikste voordeel van 'n MNB-klassifiseerder is dat slegs 'n (relatief) klein hoeveelheid afrigtingsdata kompeterende resultate lewer (Khan *et al*. 2010). Dit kan dus in 'n hulpbronskaars omgewing van groot nut wees. MNB is ook al in talle ander hulpbronskaars, taaltegnologiese toepassings gebruik met goeie resultate. Cocks en Keegan (2011) gebruik dit vir die restourering van diakritiese Maori-tekens wat verlore gegaan het tydens teksverwerking. Mogadala en Varma (2012) gebruik MNB vir die onttrekking van opinies uit nuusartikels wat in Hindi geskryf is, en Peché, Davel en Barnard (2007) gebruik MNB vir gesproketaal-identifisering.



Nota: Sien asb. die volle literatuurverwysingslys van die artikel, Snyman, D., Van Huyssteen, G. & Daelemans, W., 2014, 'Outomatiese genreklassifikasie vir Afrikaans', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 33(1), Art. #759, 12 pages. http://dx.doi.org/10.4102/satnt.v33i1.759 , vir meer inligting.

FIGUUR 3: Eenvoudige voorstelling van 'n besluitnemingsboom.

Besluitnemingsbome

Besluitnemingsbome (Khan *et al*. 2010) kan gesien word as 'n versameling *if-then*-stellings wat in 'n hiërargiese boomstruktuur geïmplementeer word. Die eienskappe van die afrigtingsdata word in die takke van die boomstruktuur weergegee. Die pad vanaf die oorsprong van die boom na die blaarnode word bepaal deur 'n reeks waar-of-vals-vrae te volg. Die blaarnode van 'n besluitnemingsboom dui dan die klas vir die betrokke teks aan. Die pad wat deur die boom gevolg word van die wortelnode af tot by die spesifieke blaarnode, word bepaal deur die eienskappe van die teks. 'n Eenvoudige boomstruktuur word in Figuur 3 voorgestel.

Die hoofvoordeel van besluitnemingsbome is dat hulle eenvoudig is om te verstaan en te interpreteer. Die uitkomste van 'n spesifieke klassifikasie geval kan ook maklik bepaal word deur te verwys na die roete wat deur die boom gevolg word om by die uiteindelike klassifikasienode uit te kom. Besluitnemingsbome neig daartoe om so min moontlik eienskappe in ag te neem tydens klassifikasie om die uitvoertyd van die algoritme te minimeer. Dit kan egter lei tot laer akkuraatheid by komplekse klassifikasieprobleme. Die belangrikste probleem by besluitnemingsbome is dat oormatige passing by kleiner versamelings afrigtingsdata maklik voorkom (wat negatief kan wees in 'n hulpbronskaars konteks) en die sisteem dan faal wanneer dit op onbekende data en data uit ander domeine toegepas word.

RIPPER-algoritme

Die RIPPER-algoritme (*Repeated incremental pruning to produce error reduction*) is 'n reëlinduseringsmetode waar reëls vir klassifikasie outomaties uit die afrigtingsdata afgelei word. RIPPER bestaan uit twee fases, die groeifase en die snoeifase, waarna die resulterende reëls geoptimeer word (Cohen 1995). Die minimum beskrywende lengte vir elke reël in die versameling word bepaal en die reëls met die kortste beskrywende lengte, wat steeds die funksie maksimeer, word dan gebruik as die geoptimeerde stel reëls waarvolgens

klassifikasie gedoen word. Van die hoofvoordele van die RIPPER-algoritme is dat dit oor die algemeen vinniger is as ander algoritmes, omdat die reëls in liniêre tyd geïnduseer word. Cohen (1995) stel verder dat RIPPER ook beter resultate toon as ander algoritmes wanneer die afrigtingstel uitskieters of ander 'geraas' bevat, maar dat dié algoritme op sulke afrigtingstelle meer reëls genereer wat die algoritme se veralgemeningsvermoë kan inperk.

Data

Die data wat gebruik is om die algoritmes af te rig, is afkomstig van data wat saamgestel is vir die ontwikkeling van tekshulpbronne vir die Suid-Afrikaanse tale in die Nasionale Sentrum vir Mensetaaltegnologie (NCHLT). Dit bestaan grotendeels uit data wat beskikbaar is in die openbare domein, asook data van medewerkers wat dit beskikbaar stel. Volgens Vargas Sierra (2005) speel die outeur van 'n teks 'n belangrike rol by die samestelling van spesialiskorpusse (soos byvoorbeeld teksklassifikasie), omdat die outeurs outeurspesifieke invloed op die dimensies van die teks het en dat dit 'n faktor is wat in gedagte gehou behoort te word. Om die graad van verteenwoordigendheid van 'n korpus te bevorder, behoort die data van meerdere outeurs gebruik te word. Omdat die bovermelde data egter grotendeels uit ongestruktureerde bronne (soos die Internet) kom, is dit nie noodwendig moontlik om die outeurs van tekste vas te stel nie en word daar nie spesifieke aandag aan die outeurs gegee by die korpussamestelling nie.

Daar moet wel in gedagte gehou word dat die samestelling van die afrigtingsdata 'n belangrike rol speel by die uiteindelike prestasie van die genreklassifikasiesisteem en dat die prestasie moontlik negatief beïnvloed kan word as die afrigtingsdata se spreiding skeefgetrek is. Dit is belangrik om die graad van verteenwoordigendheid van hierdie afrigtingsdata te verseker ten einde 'n sisteem te hê wat goed kan veralgemeen by die klassifikasie van onbekende tekste. Dimensies wat 'n invloed op die uiteindelike prestasie kan hê, soos byvoorbeeld die outeur van 'n teks, domein van tekste, informele teenoor formele tekste, ensovoorts, behoort gediversifiseer te word. Die moontlikheid en invloed van oormatige passing behoort ook in gedagte gehou te word, veral tydens optimering, waar daar 'n uithoutoetsstel gebruik word om die effek van die verskillende parameters van die algoritme teenoor die uiteindelike prestasie daarvan te meet. Dit kan moontlik gebeur dat die parameters te nou volgens die datastel gepas word en dat die algoritme dan nie meer goed kan veralgemeen wanneer onbekende tekste geklassifiseer moet word nie, wat uiteindelik 'n laer prestasie van die genreklassifikasiesisteem tot gevolg sal hê wanneer onbekende tekste geklassifiseer moet word.

Gebaseer op resultate van McCullum en Nigam (1998) word gepoog om tussen 30 000–60 000 woorde per klas te versamel; na gelang van die resultate sal die hoeveelhede afrigtingsdata egter aangepas word om beter resultate te lewer (sien die afdeling oor klasse). Soos reeds in die afdeling oor eienskappe hierbo genoem, moet eienskappe tydens voorverwerking op so 'n wyse uit die afrigtingsdata onttrek word dat hulle deur die bogenoemde algoritmes gebruik kan word. Die volgende afdelings bespreek die verskillende eienskappe wat algemeen in die literatuur genoem word.

By die vergelyking van die verskillende algoritmes, word daar van 10-voudige kruisvalidasie gebruik gemaak om die algoritmes telkens te evalueer. Die data word in 90% afrigtingsdata en 10% toetsdata verdeel wat tydens evaluering gebruik word. Die eksperiment word tien maal herhaal deur elke keer 'n ander 90/10-verdeling van die data te maak en die algoritme weer te evalueer. Dit gee 'n aanduiding van die algoritme se vermoë om effektief te kan veralgemeen. By die optimeringsfase (sien die afdeling oor algoritmes) word daar lukraak 'n eenmalige 90/10-verdeling van die data gedoen wat met elke iterasie van die optimeringsproses gebruik word.

Eienskappe

Eienskapseleksie verwys na die identifisering van 'n stel eienskappe vanuit die afrigtingsdata, wat die afrigtingsdata as't ware kan beskryf. Hierdie eienskappe moet op só 'n wyse geënkodeer word dat dit deur die masjienleeralgoritmes verstaan kan word. Die eienskappe dien dus as die vertrekpunt vir die masjienleeralgoritmes om die onderskeid tussen die klasse te leer en dan onbekende tekste daarvolgens te klassifiseer. Die eienskappe van die onbekende tekste moet op dieselfde wyse geënkodeer word om te verseker dat die masjienleeralgoritme die eienskapsinligting kan herken en dit kan analiseer om dit te vergelyk met die bestaande 'kennis' oor die betrokke klasse. In die afdelings oor woordversameling tot woordsoortinligting word die eienskappe wat in die eksperimente gebruik word, bespreek.

Woordversameling: 'n Woordversamelingbenadering is die eenvoudigste vorm waarin eienskappe van afrigtingsgevalle voorgestel kan word (Wurst 2007). Dit behels dat alle woorde in die afrigtingsteks net soos wat dit in die teks voorkom aan die masjienleeralgoritme gegee word. Die voorstelling van die woordversameling van 'n teks word dikwels as binêre vektor weergegee. Op hierdie wyse word slegs die teenwoordigheid of die afwesigheid van 'n woord in die afrigtingsgeval aangedui (Finn & Kushmerick 2006). Die aanwesigheid van 'n woord in 'n afrigtingteks word aangedui deur al die woorde in al die afrigtingtekste in 'n vektor te stoor. Dan word daar vir elke afrigtingsteks 'n ooreenstemmende vektor saamgestel, waarvan elke veld in die vektor verwys na 'n indeks van die woordvektor. As die woord in die betrokke indeks in die vektor ook in die afrigtingsteks voorkom, word daar 'n een (1) in die ooreenstemmende veld van die vektor gestoor, en as die woord nie voorkom nie, word daar 'n nul (0) gestoor.

Tabel 1 wys 'n voorstelling van die woordvoorkomste wat as binêre vektore gestoor word.

***tf-idf*-Tellings:** Die tweede stel eienskappe wat algemeen gebruik word, is *tf-idf*-tellings, waar *tf* die termfrekwensie en *idf* die inverse van die dokumentfrekwensie is. Om die frekwensie van 'n term (woord) in 'n afrigtingsgeval te bereken (d.i. *tf*), word die aantal voorkomste van die woord in die afrigtingsgeval getel en die produk geneem met die inverse van die hoeveelheid afrigtingsgevalle waarin die term voorkom (d.i. *idf*) (Manning, Prabhakar & Schütze 2009; Wurst, 2007). Die eenvoudigste formule vir die berekening van 'n *tf-idf*-telling word in Vergelyking voorgehou:

$$(tf \cdot idf)_{i,j} = tf_{i,j} \times idf_{i,j} \qquad \text{[Vgl. 3]}$$

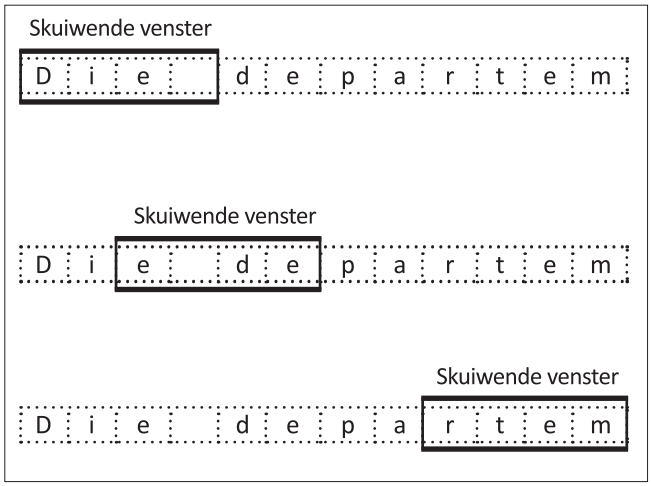
Die waarde van 'n *tf-idf*-telling kan vir die algoritme 'n aanduiding van die belangrikheid van 'n woord se bydrae tot die identifikasie van die klas gee. As 'n term herhaaldelik in 'n betrokke afrigtingsgeval voorkom, is dit waarskynlik dat die term verband hou met die klas van die afrigtingsgeval. Dit word egter genormaliseer deur die term se voorkomste in die versameling van afrigtingsgevalle, want as die term weer by ander klasse opgemerk word, word die uniekheid daarvan in die betrokke klas verflou. Die term dra daarom minder gewig by al die klasse waar dit moontlik in die afrigtingsgevalle kan voorkom.

Die *tf-idf*-tellings word dan op 'n soortgelyke wyse as die benadering tot woordvoorversameling in 'n vektor geënkodeer. Die verskil is egter dat daar nou gewigte toegeken word aan elkeen van die woorde wat wel voorkom in die betrokke afrigtingsgeval. Vergelyking 4 toon aan hoe die *tf-idf*-tellings in WEKA (Hall *et al*. 2009) bepaal word deur gebruik te maak van logaritmiese terme om *idf* te normaliseer.

$$(tf \cdot idf)_{ij} = tf_{ij} \log idf_{i,j} \qquad \text{[Vgl. 4]}$$

Karakter- en woord-*n*-gramme: Karakter-*n*-gramme word bepaal deur 'n skuiwende venster van karakterwydte *n* oor die data te beweeg en telkens die karakters wat in hierdie *n* posisies voorkom as 'n eienskap aan te teken (sien Figuur 4). Karakter-*n*-gramme het die voordeel dat dit morfologiese inligting van die woorde in die afrigtingstel kan vasvang. Morfeme (bv. *ge-* in *geloop*; of op 'n basale manier *-heid* in *belangrikheid*) bestaan selde uit 'n groot getal karakters, en daarom kan daar met 'n venster van 'n klein grootte gewerk word, wat weer die hoeveelheid eienskappe per afrigtingsgeval vergroot en die afrigtingsgeval volledige voorstel. Daar kan inligting oor leestekens, skryftekens en spasiegebruik ingewin word as die venster gekonfigureer word om die gebruik van sulke karakters in te sluit. Die hoeveelheid ongewone tekens en spasies kan ook 'n goeie identifiseerder van die aard van 'n teks wees by die klassifikasie van formele teenoor informele tekste. Dié benadering kan maklik aangepas word om woord-*n*-gramme te konstrueer deur die venster oor die *n*-hoeveelheid woorde op 'n slag te skuif. Sodoende kan die medevoorkomstes (d.i. woorde wat in dieselfde omgewing voorkom in 'n teks) wat prototipies van 'n klas is in die voorstelling ingesluit word.

Woordsoortinligting: Die gebruik van woordsoortinligting (byvoorbeeld dat *kat* 'n naamwoord en *hardloop* 'n werkwoord



n = 4.

FIGUUR 4: Voorbeeld van 'n skuiwende venster vir *n*-gram-onttrekking.

TABEL 1: Voorstelling van die woordversamelingvektore.

Afrigtings tekste	Woord	t1	t2	t3	t4	t5	†	t*n*
Woordskikking	w1	1	1	0	1	0	†	1
	w2	0	1	0	0	0	†	1
	w3	0	1	1	1	0	†	1
	w4	1	0	1	0	0	†	1
	†	†	†	†	†	†	†	†
	w*n*	1	0	0	1	0	†	1

w, woord; w*n*, *n*-de woord; t, term; t*n*, *n*-de term.
†, = tot (bv. 1...*n* [1 tot *n*]).

TABEL 2: Voorstelling van die *tf-idf*-vektore.

Afrigtings tekste	Woord	t1	t2	t3	t4	t5	†	t*n*
Woordskikking	w1	1.614734	1.614734	0	1.614734	0	†	1.614734
	w2	0	0.522042	0	0	0	†	0.522042
	w3	0	2.032903	2.032903	2.032903	0	†	2.032903
	w4	1.580069	0	0.773130	0	0	†	1.580069
	†	†	†	†	†	†	†	†
	w*n*	0.773130	0	0	0.773130	0	†	0.773130

w, woord; w*n*, *n*-de woord; *t*, term; t*n*, *n*-de term.
†, = tot (bv. 1...*n* [1 tot *n*]).

is) as eienskappe vir genreklassifikasie is 'n benadering wat in die literatuur goeie vrugte afwerp (Finn & Kushmerick 2006). Die woorde in die teks moet eers geanaliseer word om die woordsoort van die woord in sy betrokke konteks te bepaal. Hierdie woordsoortetikettering kan óf handmatig óf outomaties met behulp van 'n woordsoortetiketteerder gedoen word. Die inligting wat van hierdie analise verkry word, word dan as eienskappe (gewoonlik gepaardgaande met ander eienskappe) vir die masjienleeralgoritme gebruik. Hierdie benadering is egter nie geskik vir hulpbronskaars tale nie, omdat die outomatiese annotasie van die woordsoorte afhanklik is van die beskikbaarheid van ondersteunende hulpbronne (d.i. woordsoortetiketteerders) wat nie vir die meerderheid van hulpbronskaars tale beskikbaar is nie. 'n Woordsoortetiketteerder vir Afrikaans is deur Pilon (2005) ontwikkel met 'n akkuraatheid van 85.87%. Dié etiketteerder maak gebruik van TnT, wat 'n gekontroleerde masjienleertegniek is. Só 'n etiketteerder kan die toekenning van woordsoortinligting vir gebruik in afrigtingsdata bespoedig, maar enige foute wat moontlik deur die etiketteerder gemaak word, sal 'n negatiewe uitwerking hê op die prestasie van die algoritme wat hiermee afgerig word. Derhalwe is die insette van kundiges steeds nodig wat finansiële en tydsimplikasies teweeg bring en word woordsoortinligting nie in hierdie artikel ondersoek nie.

Klasse

Wachsmuth en Bunja (2011) redeneer dat die funksies (intensies) van tekste konstante eienskappe van genre bied om dit mee te analiseer. Hulle definieer dan drie diskrete klasse wat die funksie van 'n teks kan beskryf: ekspressiewe teks (NF-EXP – waarvan die funksie van die teks nie kommersieël is nie en waar 'n persoonlike mening weergegee word), appellatiewe teks (NF-APP – die teks het 'n kommersiële funksie, soos om die leser te oortuig om 'n produk te koop) en informatiewe teks (NF-INF – inligting wat nie kommersieël georiënteer is nie, maar wat inligting op 'n joernalistieke wyse weergee). Hierdie drie klasse is dan handmatig aan die afrigtingtekste toegeken en gebruik in teksklassifikasie-eksperimente wat goeie resultate aangetoon het. Dié klasse van Wachsmuth en Bunja (2011) word gebruik in hierdie studie as die genreklasse waartussen die klassifiseerder moet kan differensieer. Die onderstaande tabel toon die uiteindelike getal afrigtingsgevalle wat vir elke klas gebruik word.

Resultate

Algoritme

In hierdie afdeling word die vyf algoritmes wat vroeër bespreek is (*k*-nn, SVM, MNB, besluitnemingsbome en RIPPER), vergelyk deur die eienskappe en die datagroottes konstant te hou. Elkeen van die algoritmes word met hul verstekinstellings, telkens met dieselfde data en eienskappe, afgerig. 'n Benadering tot woordversameling (sien die afdeling oor woordversameling) word as die aanvanklike eienskapvoorstelling gekies vir die eksperimente waar die bogenoemde algoritmes geëvalueer word. Die twee algoritmes wat die beste resultate lewer, word geïdentifiseer en word dan in verdere eksperimente gebruik.

TABEL 3: Genreklasse.

Genreklasse: Teks	Aantal Tekste
NF-EXP	229
NF-APP	439
NF-INF	536
Totaal	**1204**

NF-EXP, ekspressiewe teks; NF-APP, appellatiewe teks; NF-INF, informatiewe teks.

Om die sukses van die resultate van die algoritmes se prestasie te peil, moet die resultate teen 'n basislynsisteem gemeet kan word. Petrenz (2012) stel die gebruik van twee basislynsisteme vir die beoordeling van kruistalige genreklassifikasie voor. Die eerste is die toekenning van 'n lukrake klas tydens klassifikasie. Daar bestaan dus een kans uit drie dat die korrekte klas toegeken word, en die basislyn as 'n presisie van 0.333 bereken word. Hierdie basislyn neem egter nie die klasverspreiding in ag nie. Die klasse van die sisteem is nie gebalanseer nie en daarom is daar nie presies een uit drie kanse dat die regte klas lukraak toegeken gaan word nie.

As die waarskynlikhede van elke klas bygereken word (Vgl. 5), verhoog die bogenoemde basislyn na 0.367.

$$Basislyn = p(klas1)^2 + p(klas3)^2 + p(klas3)^2 \quad \text{[Vgl. 5]}$$

Petrenz (2012) stel 'n tweede basislyn voor, naamlik die toekenning van die mees frekwente klas. Hierdie basislyn sisteem kan afgelei word deur die frekwensie van die mees frekwente klas in die afrigtingsdata deur die hoeveelheid voorkomste in die volledige afrigtingstel te deel. Deur telkens die mees frekwente klas (NF-NEU) toe te ken (534/1202), kan die presisie bereken word as 0.444. Beide van hierdie lae basislynsisteme word aangeneem omdat die resultate van eksperimente uit die literatuur moeilik vergelykbaar is en die prestasie van die algoritmes hiermee vergelyk sal word.

Die resultate van die algoritmes se prestasie kan teen hierdie basislynsisteme gemeet word om 'n aanduiding te kry van die sukses van die algoritmes se prestasie. Die resultate word in Tabel 4 voorgehou.

Uit Tabel 4 word daar gesien dat die resultate van al die algoritmes die bogenoemde basislyn klop en dat MNB en SVM die beste resultate lewer. Uit die literatuur is dit te verwagte dat hierdie twee algoritmes beter resultate as die ander algoritmes sal lewer. Soos reeds in die multinomiale

TABEL 4: Resultate: Algoritmes en woordversameling.

Algoritme	Woordversameling: Drie klasse		
	Presisie	Herroeping	*f*-Telling
k-nn	0.860	0.855	0.856
SVM	0.902	0.901	0.901
MNB	0.931	0.930	0.929
Besluitnemingsbome	0.878	0.878	0.877
RIPPER	0.870	0.870	0.870

k-nn, *k*-naastebuurpunt-benadering; SVM, steunvektorklassifiseerder; MNB, multinomiale naïewe Bayes-klassifiseerder.

doi:10.4102/satnt.v33i1.759

naïewe Bayes-klassifiseerders afdeling genoem, het 'n MNB-klassifiseerder relatief min data nodig om kompeterende resultate te lewer, daarom sal dit verwag word dat MNB geskik sal wees vir 'n genreklassifikasietoepassing in 'n hulpbronskaars omgewing en in die algemeen beter as die ander algoritmes sal vaar. In 'n vergelykende studie van algoritmes vir teksklassifikasie, bevind Khan *et al*. (2010) ook dat SVM en MNB goeie keuses vir teksklassifikasie sal wees. Hoewel die onafhanklikheidsaanname (sien die afdeling oor multinomiale naïewe Bayes-klassifiseerders) die MNB-algoritme negatief beïnvloed as daar 'n hoë korrelasie tussen die eienskappe is, werk dit goed vir beide numeriese en tekstuele data en is dit maklik implementeerbaar in vergelyking met ander algoritmes.

Verder stel Khan *et al*. (2010) dat SVM aanvaar kan word as een van die effektiefste algoritmes vir teksklassifikasie. SVM kan die onderliggende karakteristieke (eienskappe) beter as die meeste ander algoritmes vasvang, danksy die implementering van die minimering van strukturele risiko (sien die afdeling oor steunvektorklassifiseerders) wat die veralgemening van die algoritme bevorder. SVM bied wel uitdagings by die optimering van die algoritmiese parameters (Khan *et al*. 2010), by name die keuse van die waarde vir die kompleksiteitsparameter wat die afrigtingsfase vertraag.

Ten einde 'n volledig ingeligte besluit oor die algoritmes te neem, is dit nodig om te bepaal of die verskil in resultate nie moontlik aan blote toeval te wyte is nie. Statistiese beduidendheid word gebruik om die verskil in resultate te analiseer aan die hand van die waarskynlikheid dat die verskil in resultate bloot toevallig is. Dié waarskynlikheid word teen die sogenaamde nul hipotese (H_0) getoets (Smucker, Allan & Carterette 2007; Morgan 2012). In die geval van statistiese beduidendheid by die verskil tussen masjienleer algoritmes se resultate, word H_0 uitgedruk as dat daar geen beduidende verskil tussen die resultate van algoritme A en algoritme B is nie. Hierdie hipotese word aanvaar of verwerp aan die hand van die *p*-waarde wat as resultaat vir 'n toets vir statistiese beduidendheid gelewer word (d.i. die waarskynlikheid dat die nul hipotese waar is). 'n *p*-Waarde wat kleiner as 0.05 is, word aanvaar as statisties beduidend en die nul hipotese word verwerp (met ander woorde die verskil in resultate is nie blote toeval nie). Uit die literatuur (Smucker *et al*. 2007; Morgan 2012; Yeh 2000) volg die argument dat die beste toets vir statistiese beduidendheid by masjienleeralgoritmes die *Approximate Randomisation Testing*-metode is. Hierdie toets word outomaties gedoen deur van die vrylik beskikbare sagteware art.py (by http://www.clips.ua.ac.be/scripts/art) gebruik te maak wat 'n implementering is van die tekentoets (*sign test* [Smucker *et al*. 2007]). Die tekentoets bepaal die afwyking van die verspreidingsmediaan, gegewe die verskil in klassifikasieresultaat tussen algoritme A en algoritme B; die *p*-waarde word dan daarvolgens bepaal. Dié toets word telkens uitgevoer om die beduidendheid in die verskil tussen die verskillende algoritmes se resultate vir dieselfde datastel te bepaal.

Die *p*-waardes vir die vergelyking tussen algoritmes in die afdeling oor algoritmes word in Tabel 5 voorgehou. Die waarde in vetdruk stel telkens 'n statisties beduidende waarde voor ($p < 0.05$). Slegs MNB en SVM toon 'n beduidende verskil in resultate in vergelyking met die ander algoritmes. Tussen MNB en SVM is daar egter nie 'n beduidende verskil nie. MNB teenoor RIPPER is die enigste ander algoritmiese kombinasiepaar waar 'n statisties beduidende verskil opgemerk word.

Vervolgens kan MNB en SVM, afgerig met drie klasse, as die optimale kombinasie van algoritmes en hoeveelheid klasse geïdentifiseer word om as die basis vir die hieropvolgende eksperimente te dien.

Ten slotte word enkele opmerkings voorgehou oor die foute wat tydens klassifikasie gemaak word. Die wyse waarop die optimale klassifiseerder met die verskillende klasse omgaan, word geanaliseer aan die hand van die klastoekennings wat deur die algoritme gemaak word. Die outomatiese klastoekennings word in 'n verwarrings matriks (Tabel 6) voorgehou.

Uit Tabel 6 word daar gemerk dat die optimale klassifiseerder slegs drie uit die moontlike 229 toetsgevalle vir NF-EXP verkeerd geklassifiseer het. Drie van hierdie wanklassifikasies is as NF-APP geklassifiseer en drie as NF-INF. Daar is egter heelwat meer verwarring tussen die NF-INF- en NF-APP-klasse, waar onderskeidelik 23 en 48 gevalle by NF-INF en NF-APP verkeerd as die ander geklassifiseer is. Dit is moontlik te wyte aan die ooreenkomste tussen die OFF- en NON-klas wat deel uitmaak van NF-INF- en NF-APP-klasse. Amptelike teks (OFF) en nie-fiksietekste (NON) stem ooreen na aanleiding van die tipe register wat gebruik word en die algemene woordkeuses wat tipies by sulke tekste opgemerk word.

Eksperimentele vergelyking van eienskappe

Die verskillende eienskappe, soos in die afdeling oor eienskappe uiteengesit (te wete woordversameling, *tf-idf*-

TABEL 5: *p*-Waardes vir algoritmiese vergelyking met art.py.

Drie klasse	*k*-nn	SVM	MNB	Besluitnemingsbome	RIPPER
k-nn	-	-	-	-	-
SVM	0.00500	-	-	-	-
MNB	0.00030	0.21875	-	-	-
Besluitnemingsbome	0.30817	0.20178	0.01200	-	-
RIPPER	0.29327	0.13989	0.00540	1.00000	-

k-nn, *k*-naastebuurpunt-benadering; SVM, steunvektorklassifiseerder; MNB, multinomiale naïewe Bayes-klassifiseerder.

tellings, karakter- en woordtrigramme, asook 'n kombinasie van al die bogenoemde eienskappe), word nou om die beurt gebruik om die afrigtingsdata voor te stel in 'n reeks eksperimente wat ten doel het om die optimale eienskappe vir genreklassifikasie te identifiseer. Deur die algoritmes, die hoeveelheid afrigtingsdata en die hoeveelheid klasse konstant te hou, kan die invloed van die eienskappe op die resultate van die algoritme vasgestel word. Die resultate vir hierdie reeks eksperimente word in Tabel 7 voorgehou.

Die resultate toon dat 'n woordversamelingbenadering tot eienskaponttrekking die beste resultate lewer. Met verskille so klein soos 0.005 vir die *f*-telling (0.929 vir woordversameling, teenoor 0.924 vir *tf-idf* [vgl. Tabel 7]), is die verskil in die resultate vir die verskillende eienskappe en die twee algoritmes egter nie statisties beduidend verskillend nie (telkens $p > 0.05$). Slegs by die woordtrigrambenadering word daar 'n statisties beduidende verskil ($p > 0.05$) in die resultate tussen dié benadering en die ander benaderings opgemerk vir beide algoritmes (die woordtrigrambenadering vaar telkens beduidend slegter as die ander benaderings). By karaktertrigramme, woordtrigramme en die kombinasiebenadering, word opgemerk dat SVM hier beter vaar as MNB. Die rede hiervoor, soos reeds genoem, is dat daar nie 'n statisties beduidende verskil opgemerk word in die prestasie van hierdie twee algoritmes nie. Dit is derhalwe te verwagte dat een algoritme soms beter en een soms slegter as die ander sal vaar. Buiten die feit dat die woordtrigrambenadering die enigste algoritme is wat sonder meer geëlimineer kan word, blyk dit 'n onbegonne taak te wees om hier uit die ander benaderings 'n beste benadering te identifiseer.

Optimering

Uit die bogenoemde eliminerende eksperimente word MNB- en SVM-klassifiseerders as die beste algoritmes geïdentifiseer, terwyl daar nie bewys kan word dat daar een stel eienskappe is wat statisties beduidend beter as die ander vaar nie. Daarom word die optimeringsfase vir elkeen van die eienskapstelle herhaal (behalwe die woordtrigrambenadering wat geëlimineer kon word). Vir hierdie eienskappe word die algoritme-instellings dan, volledigheidsonthalwe, geoptimeer om vas te stel of enige verbetering in die prestasie van die algoritmes teweeg gebring kan word.

Die WEKA-implementering van die MNB-klassifiseerder het geen instellings wat verander kan word om die algoritme te optimeer nie. Die bogenoemde versamelings (sien die eksperimentele vergelyking van eienskappe afdeling) dien dan as die volledig geoptimeerde resultate.

TABEL 6: Verwarringsmatriks.

Algoritme	Teks	Toegekende klas		
		NF-APP	NF-EXP	NF-INF
Werklike klas	NF-APP	382	9	48
	NF-EXP	3	223	3
	NF-INF	23	3	510

NF-APP, appellatiewe teks; NF-EXP, ekspressiewe teks; NF-INF, informatiewe teks.

Soos reeds genoem in die afdeling oor steunvektorklassifiseerders kan die kompleksiteitsparameter (C) van steunvektorklassifiseerders gestel word om die vertaling van die ruimte na hoër dimensies, vir die identifisering van die beste skeidingsvlak, te beheer. Die C-parameter bepaal die algoritme se geneigdheid tot afrigtingsfoute teenoor die model se kompleksiteit. Die verstekwaarde van hierdie parameter is C=1.0. Volgens Hsu, Chang en Lin (2003) is 'n eksponensieel groeiende reeks waarmee die positiewe en negatiewe eksponente telkens met 0.25 vermeerder word, 'n goeie reeks waardes om te ondersoek vir C om uiteindelik die optimale waarde van C te bepaal. 'n Hoër waarde vir C sal die kompleksiteit verhoog, terwyl 'n laer waarde afrigtingsakkuraatheid benadeel (Cherkassky en Yunqian, 2004).

Hsu *et al*. (2003) omskryf die afrigtingsakkuraatheid as die vermoë van die klassifiseerder om die afrigtingsdata, waar die klasse reeds bekend is, weer korrek te klassifiseer. Hulle stel verder dat dit egter belangriker is dat die algoritme goed kan veralgemeen as wat dit is vir die algoritme om die afrigtingsdata korrek te herklassifiseer.

Om die effek van die verandering van die kompleksiteitsparameter op die prestasie te bepaal, word die waarde van C telkens vir 'n reeks van drie eksperimente geïnkrementeer. Vervolgens word die reeks $C = \{2^{-5}, 2^{-4.75}, 2^{-4.5}, ..., 2^{4.5}, 2^{4.75}, 2^{5}\}$ gebruik vir die optimeringsfases vir SVM.

Vir die doeleindes van hierdie optimeringseksperiment word daar van 'n uithoutoetsstel gebruik gemaak. Dié stel bestaan uit 10% van die oorspronklik beskikbare data wat nie by die afrigtingsdata van hierdie gevalle ingesluit is nie. Deur die toetsstel konstant te hou, word daar verseker dat die waarnemings ten opsigte van veranderinge in die resultate nie aan die toevallige insluiting van prototipiese voorbeelde by die outomatiese afrigting-/toetsstel-onttrekking van kruisvalidasie kan geskied nie. Vir die optimeringsfase word WEKA se *CVParameterSelection*-funksie gebruik, wat outomaties 'n reeks waardes vir 'n gegewe parameter van 'n algoritme toets en die optimale waarde vir die parameter, tesame met die uiteindelike resultate weergee. Die resultate van hierdie eksperimente word in Tabel 7 uiteengesit.

Tabel 8 toon die *p*-waardes vir die verskil tussen die prestasie van die geoptimeerde SVM-algoritmes en die verstek-SVM-

TABEL 7: Resultate vir steunvektorklassifiseerder en multinomiale naïewe Bayes-klassifiseerder met verskillende eienskappe.

Algoritme (drie klasse)	Presisie	Herroeping	*f*-Telling
MNB: Woordversameling	0.931	0.930	0.929
SVM: Woordversameling	0.902	0.901	0.901
MNB: *tf-idf*	0.925	0.924	0.924
SVM: *tf-idf*	0.901	0.900	0.900
MNB: Karaktertrigramme	0.902	0.889	0.888
SVM: Karaktertrigramme	0.891	0.890	0.891
MNB: Woordtrigtramme	0.822	0.786	0.769
SVM: Woordtrigtramme	0.861	0.853	0.854
MNB: Kombinasie	0.895	0.893	0.893
SVM: Kombinasie	0.896	0.895	0.895

MNB, multinomiale naïewe Bayes-klassifiseerder; SVM, steunvektorklassifiseerder.

doi:10.4102/satnt.v33i1.759

algoritme aan. Hieruit sien ons dat die optimering wel 'n verhoging in prestasie teweeg bring (soveel soos 0.024 in die geval van SVM met eienskapkombinasies), maar dat hierdie verskil nie statisties beduidend is nie (p is telkens > 0.05). Hierdie vergelyking word in Tabel 9 tussen die geoptimeerde SVM-algoritmes en die resultate van MNB getref. Ook hier word daar opgemerk dat daar geen statisties beduidende verskil tussen die resultate is nie (p is telkens > 0.05).

Vergelyking met ander genreklassifiseerders

Uit die bogenoemde elimineerende eksperimente word MNB geïdentifiseer as die beste algoritme, met 'n woordversamelingbenadering as die beste benadering tot eienskaponttrekking (geïdentifiseer met 'n resulterende *f*-telling van 0.929, sien Tabel 8). Daar kon egter nie bewys gelewer word dat die benadering statisties beduidend beter as die SVM-algoritme of die ander benaderings tot eienskap voorstelling is nie. Enige van dié benaderings kan dus gebruik word om ongeveer dieselfde resultate te verkry (in ag genome al die veranderlikes wat by die prestasie 'n rol speel) en dat die identifisering van die beste benadering dan gebaseer word op die uiteindelike *f*-telling.

Vervolgens wil ons bepaal of die resultate wat in hierdie artikel verkry is met soortgelyke genreklassifiseerders vir ander tale kan vergelyk. Die resultate vir die optimale klassifiseerder word hieronder in Tabel 11 saam met klassifiseerders vir ander tale uit die literatuur gelys. Hieruit kan gesien word dat die resultate in hierdie artikel goed vergelyk, ondanks die beperkings wat deur 'n hulpbronskaars omgewing gestel word. Daar moet egter gelet word dat die resultate van die klassifiseerders almal op verskillende hoeveelhede klasse, afrigtingsdata, toetsstelle en algoritmes gebaseer is en dat dit alles 'n invloed het op die uiteindelike resultate wat gelewer word. Klassifikasiesisteme kan nie volledig vergelyk word tensy hulle op dieselfde toetsstel geëvalueer word nie. Die voorbeelde uit die literatuur word daarom slegs daargestel om as verwysing te dien vir die klassifiseerder wat hierbo geïdentifiseer is, maar behoort steeds van waarde te wees ten einde 'n mate van vergelyking te verskaf. Wanneer die resultate uit die literatuur vergelyk word met die klassifiseerder, word daar opgemerk dat die prestasie vir die sisteme vergelykbaar is en dat die optimale klassifiseerder in al die gevalle beter presteer.

Slot

Die oorhoofse doel van hierdie studie was om 'n lewensvatbare oplossing te vind vir die outomatiese genreklassifikasie van tekste wat in Afrikaans geskryf is. Die daaropvolgende ondersoek na algemeen gebruikte metodes, die nodige hulpbronne en die voorstelling daarvan vir masjienleerbenaderings, asook die uiteindelike beoordeling daarvan, is die stappe wat gevolg is om hierdie doel te bereik. Sodoende word 'n bydrae gelewer tot die ontwikkeling van

TABEL 8: Algoritmiese optimering.

Algoritme	Eienskapstel	Optimale waarde vir parameter	Presisie	Herroeping	*f*-Telling
MNB	Woordversameling	N/A	0.931	0.930	0.929
	tf-idf	N/A	0.925	0.924	0.924
	Karaktertrigramme	N/A	0.902	0.889	0.888
	Kombinasie	N/A	0.895	0.893	0.893
SVM	Woordversameling	C = 0.03125 (2^{-5})	0.915	0.915	0.915
	tf-idf	C = 0.04420 ($2^{-4.5}$)	0.914	0.913	0.914
	Karaktertrigramme	C = 0.03125 (2^{-5})	0.919	0.919	0.919
	Kombinasie	C = 0.10511 ($2^{-3.5}$)	0.916	0.916	0.916

MNB, multinomiale naïewe Bayes-klassifiseerder; SVM, steunvektorklassifiseerder.

TABEL 9: p-Waardes vir eienskapvergelyking met art.py vir geoptimeerde steunvektorklassifiseerder teenoor standaard- steunvektorklassifiseerder.

SVM (drie klasse)	Woord-versameling C=1.0 (2^0)	*tf-idf*C=1.0 (2^0)	Karakter-trigramme C=1.0 (2^0)	Kombinasie C=1.0 (2^0)
Woordversameling C = 0.03125 (2^{-5})	0.21875	0.37500	1.00000	1.00000
tf-idf C = 0.04420 ($2^{-4.5}$)	0.12500	0.25000	1.00000	1.00000
Karaktertrigramme C = 0.03125 (2^{-5})	1.00000	1.00000	0.81702	0.82122
Kombinasie C = 0.10511 ($2^{-3.5}$)	0.28906	0.45312	0.45312	1.00000

TABEL 10: p-Waardes vir eienskapvergelyking met ART na optimering vir SVM en MNB (drie klasse).

Algoritme (drie klasse)	MNB: Woordversameling	MNB: *tf-idf*	MNB: Karaktertrigramme	MNB: Kombinasie
Woordversameling C = 0.03125 (2^{-5})	1.00000	1.00000	0.28906	0.68750
tf-idf C = 0.04420 ($2^{-4.5}$)	1.00000	1.00000	0.28906	0.72656
Karaktertrigramme C = 0.03125 (2^{-5})	0.42116	0.60034	1.00000	1.00000
Kombinasie C = 0.10511 ($2^{-3.5}$)	1.00000	1.00000	0.28906	0.62500

MNB, multinomiale naïewe Bayes-klassifiseerder.

TABEL 11: Beste kombinasie van klassifiseerder en eienskappe.

Aantal tekste	Korpus	Aantal klasse	Eienskappe	Taal	Algoritme	*f*-Telling	Bron
319	Internet: Tesisse en verhandelinge	11	*tf-idf*	Engels	NB†	0.890	Yi-Hsing and Hsiu-Yil (2008)
1224	Webdokumente	16	HTML-etikette; webadres inligting; leksikale eienskappe	Engels	SVM	0.757	Lim, Lee and Kim (2005)
800	Nuus webblaaie	2	Woordversameling; woordsoortinligting; teksstatistiek	Engels	Besluitnemingsbome	0.905	Finn and Kushmerick (2006)
1083	Koerantberigte	20	Letter-5-gramme; morfologiese inligting	Duits	SVM	0.540	(Goller *et al*. 2000)
499	Brown-korpus	6	Strukturele inligting; leksikale eienskappe, teksstatistiek	Engels	*k*-nn	0.870	(Kessler *et al*. 1997)
2172	NCHLT	3	Woordversameling	Afrikaans	MNB	0.929	-

†, Standaard-naïewe Bayes-klassifiseerder.
Nota: Sien asb. die volle literatuurverwysingslys van die artikel, Snyman, D., Van Huyssteen, G. & Daelemans, W., 2014, 'Outomatiese genreklassifikasie vir Afrikaans', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 33(1), Art. #759, 12 pages. http://dx.doi.org/10.4102/satnt.v33i1.759 , vir meer inligting.

Afrikaanse (as voorbeeld van 'n hulpbronskaars taal van Suid-Afrika) hulpbronne, sowel as om die gebruik daarvan in ander tegnologiese ontwikkeling te bevorder.

Ondanks die feit dat Afrikaans as 'n hulpbronskaars taal gesien word, is dit steeds moontlik om genreklassifikasiesisteme te ontwikkel, waarvan die prestasie vergelykbaar met sisteme uit die literatuur is. In hierdie studie word 'n genreklassifikasiesisteem voorgestel deur gebruik te maak van die volgende benadering: Die implementering van 'n MNB-algoritme, afgerig met woordversamelingbenadering vir eienskappe as voorstelling van drie genreklasse. Dié sisteem lewer 'n resulterende *f*-telling van 0.929 (Tabel 11).

Gegewe die uitslae van hierdie navorsing is daar moontlikhede vir verdere navorsing, wat in die toekoms uitgevoer kan word. Onderwerpe wat in die studie geïdentifiseer is wat aandag sou kon geniet, sluit die volgende in:

- Uit die resultate vir die ontwikkeling van genreklassifikasiesisteme vir Afrikaans, blyk die sisteme goed te presteer in 'n gekontroleerde eksperimentele omgewing. Hierdie sisteme kan egter aan verdere eksperimentering onderwerp word om die moontlikhede van dié sisteme in die regtewêreld te evalueer, sy dit gedoen word deur intrinsieke evaluering (waar die sisteem as 'n alleenstaande entiteit geëvalueer word) of ekstrinsieke evaluering (die sisteem se bydrae tot die prestasie van ander bestaande sisteme word geëvalueer, byvoorbeeld as deel van 'n dokumentbestuurstelsel).
- Deurlopende ontwikkeling van tekshulpbronne vir die hulpbronskaars, inheemse Suid-Afrikaanse tale beteken dat daar moontlike ruimte vir die uitbreiding van die afrigtingsdata bestaan. Die Afrikaanse sisteem kan verbeter word deur die hoeveelheid beskikbare data, sowel as die kwaliteit daarvan uit te brei en te verbeter. Die graad van verteenwoordigendheid van 'n afrigtingstel is van kardinale belang vir die uiteindelike prestasie van die sisteem wat daarop gebaseer is. Die klasverspreidings moet hier spesifiek in gedagte gehou word om te verseker dat sommige van die klasse nie onderverteenwoordig is nie.
- Die moontlikheid van oormatige passing by kleiner afrigtingstelle en klasse, sowel as benaderings vir die hantering daarvan, kan ondersoek word.
- Die effek van domeinoordrag op die klassifiseerders kan ondersoek word deur die sisteme eksplisiet hiervoor te evalueer en moontlike oplossings vir dié probleem te ondersoek. Dit kan gedoen word deur die domein waaruit die afrigtingsdata saamgestel word streng te beheer en dan spesifiek toetstekste uit 'n ander domein te gebruik om so die robuustheid van die sisteem te toets.

Erkenning

Die genreklassifikasiesisteem wat in hierdie studie beskryf word, is geïmplementeer in 'n projek wat befonds is deur die Departement van Kuns en Kultuur van die Suid-Afrikaanse regering en wat onderneem is deur Trifonius, met CTexT® (Noordwes Universiteit) en die Universiteit van Antwerpen as medewerkers. Die projek het ten doel gehad om die daarstelling van 'n genreklassifikasiesisteem vir Afrikaans te bewerkstellig en internasionale samewerking, met betrekking tot natuurliketaalprosesserings navorsing, te bevorder. 'n Webdemonstrasie en verwante navorsingsuitsette is beskikbaar by www.trifonius.co.za en die bogenoemde hulpbronne is beskikbaar by http://sourceforge.net/projects/gcsal/.

Opinies in hierdie artikel gelug is dié van die outeurs en kan nie aan die Departement Kuns en Kultuur toegedig word nie.

Mededingende belange

Die outeurs verklaar hiermee dat hulle geen finansiële of persoonlike verbintenis het met enige party wat hulle nadelig of voordelig kon beïnvloed het in die skryf van hierdie artikel nie.

Outeursbydrae

D.S. (Noordwes-Universiteit) was verantwoordelik vir die uitvoer van alle eksperimente. G.B.v.H. (Noordwes -Universiteit) was die projekleier en begeleier van die navorsing. W.D. (Universiteit van Antwerpen) het betekenisvolle konseptuele bydraes gelewer.

Literatuurverwysings

Cardinaels, K., Meire, M. & Duval, E., 2005, 'Automating metadata generation: The simple indexing interface', *WWW 2005: 14th International conference on World Wide Web*, pp. 548–556, ACM Press.

Cherkassky, V. & Yunqian, M., 2004, 'Practical selection of SVM parameters and noise estimation for SVM regression', *Neural networks* 17(1), 113–126.

Cocks, J. & Keegan, T., 2011, 'A word-based approach for diacritic restoration in Maori', *The Australasian Language Technology Association Workshop*, pp. 126–130, Canberra.

Cohen, W.W., 1995, 'Fast effective rule induction', *The 12th International conference on machine learning*, vol. 95, pp. 115–123, Lake Tahoe, California.

Finn, A. & Kushmerick, N., 2006, 'Learning to classify documents according to genre', *Journal of the American Society for Information Science and Technology* 57(7), 1506–1518.

Francis, W.N. & Kucera, H., 1979, 'Brown corpus manual – Revised and amplified', Dept. of Linguistics, Brown University, Providence.

Goller, C., Löning, J., Will, T. & Wolff, W., 2000, 'Automatic document classification: A thorough evaluation of various methods', *The Internationales Symposium für Informationswissenschaft, Informationskompetenz – Basiskompetenz in der Informationsgesellschaft*, Proceedings 7, pp. 145–162.

Grover, A.S., Van Huyssteen, G.B. & Pretorius, M.W., 2011, 'The South African human language technology audit', *Language resources and evaluation*, 45, 271–288.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H., 2009, *The WEKA data mining software: An update*, ACM SIGKDD Explorations Newsletter 11(1), 10–18.

Hsu, C.W., Chang, C.C. & Lin, C.J., 2003, 'A practical guide to support vector classification', Technical Report, Dept. of Computer Science, National Taiwan University', viewed 27 October 2012, from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Instituut vir Nederlandse Leksikologie, 2005, 'Parole Corpus', besigtig op 01 September 2010, by http://parole.inl.nl/html/main_info_dutch.html

Kessler, B., Nunberg, G. & Schütze, H., 1997, 'Automatic Detection of Text Genre', in P.R. Cohen & W. Wahlster (eds.), ACL-97, *35th Annual Meeting of the Association for Computational Linguistics proceedings*, Madrid, Spain, pp. 32–38.

Khan, A., Baharudin, B., Lee, L.H. & Khan, K., 2010, 'A review of machine learning algorithms for text-documents classification', *Journal of Advances in Information Technology* 1(1), 4–20.

Lim, C.S., Lee, K.J. & Kim, G.C., 2005, 'Multiple Sets of Features for Automatic Genre Classification of Web Documents', *Information Processing and Management* 41(5), 1263–1276.

Manning, C.D., Prabhakar, R. & Schütze, H., 2009, *An introduction to information retrieval*, pp. 117–119; 253–285, Cambridge University Press, Cambridge.

McCallum, A. & Nigam, K., 1998, 'A comparison of event models for naive Bayes text classification', *The AAAI-98 workshop on learning for text categorization, Tech. Rep*. WS-98-05, pp. 41– 48, AAAI Press.

Mogadala, A. & Varma, V., 2012, 'Retrieval approach to extract opinions about people from resource scarce language news articles', *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ACM, pp. 1–4.

Morgan, W., '*Statistical hypothesis tests for NLP or: Approximate randomization for fun and profit*', viewed 24 October 2012, from http://masanjin.net/sigtest.pdf

Peché, M., Davel, M. & Barnard, E., 2007, 'Phonotactic spoken language identification with limited training data', *The annual conference of the International Speech Communication Association*, pp. 1537–1540, Antwerp, Belgium.

Peng, F. & Schuurmans, D., 2003, 'Combining naive Bayes and *n*-Gram language models for text classification', *The 25th European conference on information retrieval research* (ECIR-03), pp. 335–350.

Petrenz, P., 2012, 'Cross-lingual genre classification', *The EACL 2012 Student research workshop*, pp. 11–21, Avignon, France.

Pilon, S., 2005, 'Outomatiese Afrikaanse woordsoortetikettering', Magister verhandeling, Noordwes-Universiteit, Suid-Afrika.

Smucker, M.D., Allan, J. & Carterette, B., 2007, 'A comparison of statistical significance tests for information retrieval evaluation', *The 16th ACM conference on information and knowledge management*, pp. 623–632.

Vapnik, N.V., 1995, *The nature of statistical learning theory*, pp. 159– 164, Springer-Verlag, New York.

Vargas Sierra, C., 2005, 'A pragmatic model of text classification for the compilation of special-purpose corpora', in J. Mateo & F.Yus (eds.), *Thistles, a homage to Brian Hughes, Essays in Memoriam*, pp. 295–315.

Wachsmuth, B. & Bunja, K., 2011, 'Back to the roots of genres: Text classification by language function', *The 5th international joint conference on natural language processing*, pp. 632–640, Chiang Mai, Thailand.

Wurst, M., 2007, '*The word vector tool: User guide, operator reference, developer tutorial*', viewed 26 August 2012, from http://www-ai.cs.uni-dortmund.de/SOFTWARE/WVTOOL/doc/wvtool-1.0.pdf

Yeh, A., 2000, 'More accurate tests for the statistical significance of result differences', *The 18th international conference on computational linguistics*, n.d., pp. 947– 953.

Yi-Hsing, C & Hsiu-Yi, H., 2008, 'An automatic document classifier system based on naïve Bayes classifier and ontology', *The seventh international conference on machine learning and cybernetics*, Kunming, China, July 12–15, 2008, pp. 3144–3149.