Outomatiese genreklassifikasie vir hulpbronskaars tale

Dirk Snyman
20570856

Verhandeling voorgelê vir die graad
Magister Artium in Algemene Taal- en Literatuurwetenskap
aan die Potchefstroomkampus van die Noordwes-Universiteit

Studieleier: Prof. GB van Huyssteen
Medestudieleier: Prof. W Daelemans

November 2012

“Don't classify me, read me. I'm a writer, not a genre.”

Carlos Fuentes

VOORWOORD

Graag word die volgende persone en instansies bedank:

- Gerhard van Huyssteen, vir jou leiding, finansiële steun, dat jy my betrek het by jou projekte en vir alles wat ek by jou kon leer.
- Prof. Walter Daelemans, vir insigte oor teksklassifikasie en vir die konstruktiewe terugvoer oor my eksperimente.
- Die Navorsingseenheid: Tale en Literatuur in die Suid-Afrikaanse Konteks vir befondsing en ondersteuning.
- Die Sentrum vir Tekstegnologie (CTexT®) vir die gebruik van hulle data en toerusting en dat hulle my die tyd en ruimte gegee het om aan hierdie verhandeling te werk.
- Martin Puttkammer, vir al die raad en geselsies oor navorsing.
- Cindy McKellar en Marissa Griesel, vir die kere wat ek by julle kantoor kon inbars met vrae oor alles en nog wat.
- Jacques van Heerden en Wikus Pienaar, dat julle my gereeld kom wegsleep het om saam koffie te drink.
- My ouers en toekomstige skoonouers, vir hulle ondersteuning en gebede.
- My vriende wat ek ernstig afgeskeep het in die doodsnikke van my studies. Ek is nou weer beskikbaar.
- Annuschka, vir jou liefde, verstaan en dat jy altyd in my glo.

Die meetsnoere het vir my in lieflike plekke geval, ja, my erfenis is vir my mooi.

Psalm 16:6

ABSTRACT

AUTOMATIC GENRE CLASSIFICATION FOR RESOURCE SCARCE LANGUAGES

Dirk Snyman

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 classifies 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 the genre of a text. In this study the development of an automatic genre classification system in a resource scarce environment is postulated. This study aims to: i) 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), ii) transfer this approach to other indigenous South African resource scarce languages, and iii) investigate the effectiveness of technology recycling for closely related languages in a resource scarce environment.

To achieve the first goal, five machine learning approaches were identified from the literature that are generally used for text classification, together with five common approaches to feature extraction. Two different approaches to the identification of genre classes are presented. The machine learning-, feature extraction- and genre class identification approaches were used in a series of experiments to identify the best approach for genre classification for a resource scarce language. The best combination is identified as the multinomial naïve Bayes algorithm, using a bag of words approach as features to classify texts into three abstract classes. This results in an *f*-score (performance measure) of 0.929 and it was subsequently shown that this approach can be successfully applied to other indigenous South African languages.

To investigate the viability of technology recycling for genre classification systems for closely related languages, Dutch test data was classified using an Afrikaans genre classification system and it is shown that this approach works well. A pre-processing step was implemented by using a machine translation system to increase the compatibility between Afrikaans and Dutch by translating the Dutch texts before classification. This results in an *f*-score of 0.577, indicating that technology recycling

between closely related languages has merit. This approach can be used to promote and fast track the development of genre classification systems in a resource scarce environment.

Key Terms:

GENRE CLASSIFICATION, RESOURCE SCARCE LANGUAGES, MACHINE LEARNING, TECHNOLOGY RECYCLING, HUMAN LANGUAGE TECHNOLOGY, NATURAL LANGUAGE PROCESSING.

OPSOMMING

OUTOMATIESE GENREKLASSIFIKASIE VIR HULPBRONSKAARS TALE

Dirk Snyman

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 voorafbepaalde 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 hulpbronskaarsomgewing voorgehou. Hierdie studie het ten doel om: *i*) ’n ondersoek te loods na bestaande genreklassifikasiesisteme, en om dan die tegnieke en benaderings te implementeer vir Afrikaans (’n hulpbronskaars taal), *ii*) om die implementering vir Afrikaans toe te pas op die ander inheemse Suid-Afrikaanse hulpbronskaars tale, en *iii)* om die effektiwiteit van tegnologieherwinning van genreklassifikasiesisteme vir nabyverwante tale in ’n hulpbronskaars omgewing te ondersoek.

Om die eerste doelwit te bereik word daar vyf masjienleerbenaderings uit die literatuur geïdentifiseer wat in die algemeen gebruik word vir teksklassifikasie, tesame met die vyf algemeenste benaderings tot eienskaponttrekking. Twee verskillende benaderings tot die identifisering van genreklasse word voorgehou en vervolgens word die masjienleeralgoritmes, eienskappe en genreklasse in ’n reeks eksperimente gebruik om die beste benadering vir genreklassifikasie vir ’n hulpbronskaars taal te identifiseer, te wete ’n multinomiale naïewe Bayes-algoritme, met woordversameling eienskappe en ’n abstrakte drie-klas-benadering tot genreklasse. Hierdie kombinasie lewer ’n *f*-telling (prestasiesyfer) van 0.929 en daar word vervolgens aangetoon dat hierdie benadering suksesvol toegepas kan word op die ander inheemse Suid-Afrikaanse tale.

Om die lewensvatbaarheid van tegnologieherwinning vir nabyverwante tale te ondersoek, word Nederlandse toetsdata tot ’n suksesvolle mate met die Afrikaanse genreklassifikasiesisteem geklassifiseer. ’n Voorverwerkingtussenstap, deur die Nederlandse teks met ’n masjienvertaalsisteem te vertaal, is geïmplementeer om die Afrikaans en Nederlands beter versoenbaar te maak wat ’n *f*-telling van 0.577 tot gevolg

het. Die moontlikheid van tegnologieherwinning tussen nabyverwante tale blyk meriete te hê vir die bevordering en bespoediging van ontwikkeling van genreklassifikasie-sisteme in ’n hulpbronskaarsomgewing.

Sleutelterme:

GENREKLASSIFIKASIE, HULPBRONSKAARS TALE, MASJIENLEER, TEGNOLOGIEHERWINNING, MENSETAALTEGNOLOGIE, NATUURLIKETAALPROSESSERING.

INHOUDSOPGAWE

LYS VAN FIGURE

LYS VAN TABELLE

HOOFSTUK 1: INLEIDING

1.1. KONTEKSTUALISERING

Op die terrein van teksverwerking speel die metadata oor ŉ bepaalde teks in baie gevalle ŉ belangrike rol. Cardinaels *et al.* [5] stel dat sonder toepaslike metadata, dit moeilik of selfs onmoontlik sal wees om elektroniese leerinhoud outomaties te identifiseer en te onttrek. Voorbeelde van metadata sluit in: tekstuele statistiek (woord- en karaktertellings), onderwerpe, leesbaarheidsyfers, outeurnaam, titel, ensovoorts. Wanneer daar byvoorbeeld korpusse saamgestel word vir natuurliketaal-prosseseringstoepassings 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 tekste uit ŉ wye reeks domeine en genres; dit sal verseker dat die korpus ŉ goeie spreiding toon. As ŉ 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 verteenwoordigendheid, onakkurate resultate lewer. Dit is dus van groot belang dat dokumente gemonitor word voordat dit by ŉ korpus gevoeg word, of dat die korpus by nabaat geanaliseer kan word om die aard van die tekste vas te stel. Dit sal navorsers in staat stel om te bepaal hoe die korpus beïnvloed sal word as daar nuwe data bygevoeg word, as die herkoms van die nuwe data bekend is. Vir natuurliketaalprosessering wil ŉ mens hierdie metadata ook dikwels outomaties toevoeg tot tekste – byvoorbeeld of iets uit ŉ bepaalde domein kom of nie, of iets gemorspos is of nie, of iets deur ŉ bepaalde outeur geskryf is of nie, of iets tot ŉ bepaalde genre behoort of nie, ensovoorts.

Sodanige metadata word dikwels toegevoeg met behulp van outomatiese teksklassifiseerders. ŉ Teksklassifiseerder word gedefinieer as ŉ sisteem wat op grond van die inhoud van ŉ teks een of meer voorafbepaalde klasse of kategorieë outomaties aan ŉ teks toeken [14]. Statistiese patroonherkenningsbenaderings soos masjienleer en neurale netwerke word oor die algemeen gebruik om sulke klassifiseerders af te rig [14].

Die eerste en onmiddellike toepassing vir hierde navorsing sal in ŉ projek wees vir die Nasionale Sentrum vir Mensetaaltegnologie (NCHLT) wat handel oor die saamstel van elektroniese tekshulpbronne vir Suid-Afrikaanse tale. Onder andere word daar

gestratifiseerde korpusse vir hierdie tale ontwikkel. Dit is ŉ algemene praktyk dat korpusse volgens ŉ reeks genres gestratifiseer word. Voorbeelde hiervan is die Brown-korpus [13], asook die PAROLE-korpus [21]. Deur ŉ korpus te stratifiseer, word die verteenwoordigendheid daarvan verseker. Wanneer ŉ 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 ŉ teks bekend is, kan daar na ŉ opsomming van al die tekste gekyk word om ŉ duidelike oorsig te kry van of die korpus verteenwoordigend genoeg is, of nie. Hierdie metadata oor ŉ teks is egter nie altyd beskikbaar nie en omdat daar met groot hoeveelhede data gewerk word, is handmatige annotering, van enige aard, ŉ arbeidsintensiewe en tydrowende aktiwiteit, wat dikwels vertaal na hoë finansiële kostes. As hierdie annotasie dus geoutomatiseer kan word, kan dit lei tot besparings in beide tyd en geld. Vervolgens word daar in hierdie navorsing gekyk na die moontlikhede van die ontwikkeling van sisteme wat outomaties die genre van ŉ teks kan identifiseer.

1.2. PROBLEEMSTELLING

In die voorafgaande gedeelte word daar melding gemaak van die genre van ŉ teks. Yates en Orlikowski [58] definieer genre as ŉ karakteristieke tipe kommunikatiewe aksie, wat gekenmerk word deur ŉ sosiaal aanvaarde kommunikatiewe doel en gemeenskaplike aspekte van vorm (styl). Daar word dus verwys na eienskappe wat verder strek as die blote inhoud van ŉ teks; genre verwys na die kombinasie van die aard/karakter van ŉ teks en die doel waarvoor ŉ teks geskryf word. ŉ Genre kan iets soos briewe, advertensies, akademiese skrywes, en so meer, wees. *Genreklassifikasie* verwys dus na die outomatiese analise van hierdie eienskappe en die klassifikasie/toekenning van ŉ klas aan ŉ teks gebaseer op hierdie analise. Vir genreklassifikasie word dus ŉ sisteem benodig wat die inhoudelike aspekte van tekste kan analiseer en identifiseer as behorende tot ŉ voorafbepaalde klas (d.i. genre). Teksklassifikasie kan eenvoudig gedefinieer word as die toekenning van vooraf bepaalde klasse of kategorieë aan tekste gebaseer op die betrokke teks se inhoud, struktuur, stylkenmerke, ensovoorts [14]. Genreklassifikasie is ŉ spesiale geval van teksklassifikasie waar ŉ genre aan ŉ teks toegeken word aan die hand van die teksinhoud. Hierdie toekenning word dan deur ŉ teksklassifikasie-algoritme waargeneem.

Genreklassifikasie moet nie met onderwerpklassifikasie (wat ook ŉ toepassing van teksklassifikasie is) verwar word nie. Neem as voorbeeld ŉ koerantartikel: die artikel kan oor enige sport of politieke gebeure handel, maar die genre van die teks bly egter dié van ŉ koerantartikel. Hierdie verband is dieselfde vir enige genre, aangesien genre in essensie onderwerp-onafhanklik is, alhoewel sommige onderwerpe meermale by sekere genres as by ander teenwoordig is.

Genreklassifikasie is in verskeie alledaagse sisteme te sien:

- As deel van outomatiese data-indekseringsisteem [29];
- E-posklassifiseerders (byvoorbeeld gemorsposfiltreerders) [29]; en
- Outomatiese sentimentanalise uit tekste [29]

Die primêre probleem van hierdie navorsing is dat daar nie genreklassifikasiesisteme vir die Suid-Afrikaanse tale bestaan nie. Dit veroorsaak ŉ dilemma vir die ontwikkeling van tekshulpbronne vir hierdie tale, veral waar verteenwoordigendheid belangrik is. In ŉ projek[1] befonds deur die Departement van Kuns en Kultuur van die Suid-Afrikaanse regering, word daar onderneem om ŉ sisteem te ontwikkel wat hierdie genreklassifikasie outomatiseer. Dié genreklassifikasiesisteem is beskikbaar as ŉ webgebaseerde[2] demonstrasie. Die projek is ook volledig "oop bron" gelisensieerd en alle dokumente en lêers is ook beskikbaar[3].

Deur gebruik te maak van masjienleerbenaderings word die vereiste van ŉ voorkennis van die betrokke tale tot ŉ groot mate geminimeer. Tekstuele data word rekenaarmatig verwerk tot afrigtingsdata en kan tot ŉ groot mate taalonafhanklik wees. Afrigtingsdata is egter minder geredelik beskikbaar vir tale met meer beperkte hulpbronne, soos in die geval van Afrikaans, en soveel te meer vir die ander inheemse Suid-Afrikaanse landstale. ŉ Tekort aan hulpbronne kan ŉ groot struikelblok wees, veral in die geval waar groot hoeveelhede data gebruik word om ŉ masjienleeralgoritme af te rig. Die hoeveelheid data beskikbaar vir afrigting speel dikwels ŉ deurslaggewende rol met betrekking tot die uiteindelike resultate wat die sisteem lewer. Hoewel Afrikaans ook ŉ hulpbronskaars taal is, is daar heelwat meer teksdata beskikbaar as vir die ander tale en sal dit die aanvanklike navorsing oor genreklassifikasiemetodes vergemaklik, waarna die metodes oorgedra kan word na die ander tale.

[1] Projekwebblad: http://www.trifonius.co.za/projects/genre-classification/
[2] Demonstrasie beskikbaar by: http://196.33.156.2/genre_classification/
[3] Projekhulpbronne beskikbaar by: http://sourceforge.net/projects/gcsal/files/

Hulpbronskaarsheid is ŉ onderwerp wat baie aandag geniet, beide nasionaal en internasionaal by kongresse soos die *Association for Computing Machinery* se *Annual Symposium on Computing for Development* wat fokus op die problematiek van ontwikkelende omgewings en die hulpbronskaarsheid van die tale wat daarmee gepaardgaan, is ŉ gereelde onderwerp van die navorsing wat daar bespreek word. *Meta-net*[4] van die *Multilingual Europe Technology Alliance* het ŉ 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 kan uitsterf weens ŉ gebrek aan digitale hulpbronne. Hulpbronskaarsheid is dus ŉ groot probleem waarvoor oplossings ernstig benodig word.

Chan en Rosenfeld [7] definieer ŉ hulpbronskaars taal as ŉ taal met ŉ klein groepie gebruikers/sprekers wat (gewoonlik) uit ekonomies benadeelde omstandighede kom, en wat grotendeels deur die kommersiële wêreld geïgnoreer word. Dit is juis redes soos ekonomiese invloed wat veroorsaak dat ontwikkeling vir hulpbronskaars tale agterweë gelaat word. Hierdie agterstand wat die tale het ten opsigte van die beskikbare hulpbronne kniehalter die ontwikkeling van taaltegnologiese toepassings. Hierdie invloed blyk duidelik in die ontwikkeling van beide spraak- [33][7] sowel as teksgebaseerde [50][9] tegnologieë. Pilon *et al*. [50] stel voor dat die gebruik van tale wat ŉ verwantskap het, byvoorbeeld ŉ historiese verwantskap of tale wat uit dieselfde taalfamilies kom, die gapings wat gelaat word deur hulpbronskaarsheid kan oorbrug. Dit blyk duidelik dat dit van groot belang is om innoverende maniere te ondersoek om die invloed van hulpbronskaarsheid op tegnologie-ontwikkeling teen te werk of te mitigeer.

1.3. NAVORSINGSVRAE

Na aanleiding van die bogenoemde agtergrond en probleemstelling ontstaan die volgende navorsingsvrae:

- Wat is ŉ geskikte benadering tot genreklassifikasie vir ŉ hulpbronskaars taal?
- Op watter manier kan bestaande genreklassifikasiebenaderings herwin word vir nabyverwante tale?

[4] http://www.meta-net.eu/whitepapers/overview

1.4. DOELSTELLINGS

Ten einde die navorsingsvrae hierbo genoem te beantwoord, het hierdie navorsing die volgende doelstellings 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);
- om die implementering vir Afrikaans toe te pas op die ander inheemse Suid-Afrikaanse hulpbronskaars tale; en
- om die effektiwiteit van tegnologieherwinning van genreklassifikasiesisteme vir nabyverwante hulpbronskaars tale te ondersoek.

1.5. SENTRALE TEORETIESE STELLING

Die standpunt word ingeneem dat genreklassifikasie effektief uitgevoer kan word deur van algemeen bekende masjienleerbenaderings vir teksklassifikasie gebruik te maak. Hierdie benadering kan effektief gebruik word om dié taak vir ’n hulpbronskaars taal (Afrikaans) aan te pas en dan oor te dra na ander hulpbronskaars tale (res van die inheemse Suid-Afrikaanse tale). Daar bestaan tans geen genreklassifikasiesisteme vir die Suid-Afrikaanse tale nie en dit is daarom belangrik dat sulke kerntegnologieë ontwikkel word, ten einde dié tale se gebruik, sowel as die moontlikhede vir tegnologiese ontwikkeling in hierdie tale te bevorder. Tegnologieherwinning vir nabyverwante tale kan gebruik word om die ontwikkeling van sulke tegnologieë te bespoedig.

1.6. NAVORSINGSMETODE

Om die bogenoemde doelstellings te bereik, word die volgende navorsingsaktiwiteite in hierdie studie uitgevoer:

A. Literatuurstudie

Die konsep van “genre”, sowel as die klassifisering daarvan (en die verskillende aspekte wat dit behels) word bestudeer en omskryf aan die hand van die beskikbare literatuur. Daar word spesifiek aandag geskenk aan literatuur rakende:

- algoritmes wat algemeen gebruik word vir teksklassifikasie (spesifiek vir genreklassifikasie);
- eienskappe wat algemeen gebruik word by datavoorstelling vir masjienleeralgoritmes;
- klasse vir genreklassifikasie;
- die samestelling van datastelle; en
- evalueringsmetrieke en benaderings.

Die literatuurstudie word dan gebruik as die vertrekpunt vir die eksperimentering en ontwikkeling van genreklassifikasiesisteme.

B. Eksperimentering

Goller *et al.* [14] onderskei tussen twee fases van outomatiese teksklassifikasie:

(1) Die eerste fase is die afrigtingsfase waar voorbeeldtekste van elkeen van die voorafbepaalde klasse geklassifiseer word en dan as afrigtingsdata vir die sisteem gebruik word. Die sisteem lei dan die verskillende kenmerke van elke klas vanuit die afrigtingsvoorbeelde af deur statistiese inferensie, veralgemening, abstraksie, ensovoorts [14]. Dit is dus belangrik dat 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 as voorbeeld dien vir tekste wat definitief nie deel van die klas is nie [14].

Hierdie teenvoorbeelde is veral van waarde waar tekste volgens die versameling klasse geklassifiseer moet word, maar ook as onbekend geklassifiseer moet word as die teks nie eksplisiet as ’n klas geïdentifiseer kan word nie. Elke teks wat aan die een klas toegeken word, dien outomaties as ’n teenvoorbeeld vir die ander klasse.

(2) Die tweede fase van teksklassifikasie word die klassifikasiefase genoem waar voorheen onbekende tekste deur die masjienleeralgoritme geklassifiseer word [14]. Die klassifikasiesisteem kan dan die klas van die invoerteks bepaal, of as die teks nie geklassifiseer kan word volgens die bepaalde klasse nie, 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 invoer na die regte dataformaat.

Regdeur die literatuur is daar ’n reeks metrieke wat gebruik word om die prestasie van ’n klassifikasiesisteem te meet. 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 [30].

Presisie is ’n voorstelling van die mate van korrektheid [26] – hoeveel van die tekste wat geklassifiseer is, is reg geklassifiseer. Formeel word presisie bereken as die hoeveelheid *ware positiewe* (hoeveelheid positiewe elemente wat korrek as positief geklassifiseer is) gedeel deur die totale hoeveelheid elemente wat as positief geëtiketteer is. Vergelyking (1) word gebruik om die presisie te bepaal.

$$\boldsymbol{Presisie} = \frac{\boldsymbol{tp}}{\boldsymbol{tp}+\boldsymbol{fp}} \tag{1}$$

Herroeping (Vergelyking (2)) meet die mate van volledigheid van die sisteem [26] – hoeveel van die tekste wat geklassifiseer moet word, is wel geklassifiseer. Herroeping word formeel bereken as die hoeveelheid *ware positiewe* gedeel deur die totale hoeveelheid elemente wat werklik tot die positiewe klas behoort (met ander woorde die som van die *ware positiewe* en *vals negatiewe* – die elemente waaraan nie 'n klas toegeken is nie, maar in werklikheid aan die positiewe klas behoort).

$$\boldsymbol{Herroeping} = \frac{\boldsymbol{tp}}{\boldsymbol{tp}+\boldsymbol{fn}} \qquad (2)$$

Die presisie en herroeping word saamgevat in 'n *f*-telling wat 'n enkele waarde weergee as die harmoniese gemiddeld van presisie en herroeping. Die *f*-telling kan aangepas word om aan die presisie of die herroeping van die sisteem 'n swaarder gewig toe te ken na gelang van die belangrikheid wat die een bo die ander mag hê. Die *f-telling* word in hierdie geval gebalanseerd tussen presisie en herroeping bereken, aangedui deur f_1 in Vergelyking (3):

$$\boldsymbol{f_1} = \mathbf{2} \cdot \frac{\boldsymbol{Presisie} \cdot \boldsymbol{Herroeping}}{\boldsymbol{Presisie} + \boldsymbol{Herroeping}} \qquad (3)$$

Hierdie drie metrieke word voorgehou as die resultate vir beide die *n*-voudige kruisvalidasie, sowel as die uithoutoetsstelekserimente wat in afdeling 3.1 en 4.1 uitgevoer word.

Vervolgens word die benaderings uit die literatuur geïmplementeer en teen mekaar opgeweeg in 'n stel eksperimente, spesifiek só gestruktureer om te poog om die hoeveelheid eksperimente en die tyd wat die uitvoer daarvan in beslag neem te minimeer, maar terselfdertyd akkurate oorweging van die resultate te verseker.

Die uiteensetting van die eksperimente is soos volg:

1. Algoritme

   Hier word 5 algoritmes (*k*-naastebuurpuntklassifiseerders, steunvektorklassifiseerders, multinomiale naïewe Bayes-klassifiseerders, besluitnemingsbome en die RIPPER-algoritme) vergelyk deur die eienskappe en die datagroottes konstant te hou. Die algoritmes word op hul verstekinstellings afgerig, telkens met dieselfde data en eienskappe, maar met twee stelle klasse (drie klasse

teenoor dertien klasse) met 'n totaal van tien eksperimente vir die eerste rondte van die beplande eksperimente. Deur die resultate hiervan te analiseer, word daar dan twee algoritmes en een stel klasse wat die beste resultate lewer, geïdentifiseer. Hierdie twee algoritmes, gekombineer met die klasse waarop besluit word, word dan in die daaropvolgende eksperimente verder gebruik.

2. Data

2.1. Eienskappe (d.i. datavoorstelling)

Vervolgens word 'n reeks eienskappe (woordversameling, termfrekwensie en inversedokumentfrekwensie (*tf-idf)*, karakter *n*-gramme, woord *n*-gramme, teksstatistiek, en kombinasies hiervan) geëvalueer deur die twee algoritmes en klasse (in die vorige stap geïdentifiseer) telkens met 'n ander stel uit die reeks eienskappe af te rig en dan te identifiseer watter eienskappe die beste resultate lewer.

2.2. Datahoeveelhede

Gegewe die twee algoritmes, die klasse en die eienskappe (hierbo geïdentifiseer) word leerkurwes opgestel om die invloed van die hoeveelheid data wat vir die algoritme beskikbaar is, voor te stel en dan ook vas te stel wat die invloed sou wees as daar meer data vir die algoritme beskikbaar gestel word.

3. Optimering

Nadat die beste kombinasie van algoritmes, klasse en eienskappe, met die ideale hoeveelhede data bepaal is, gaan hierdie kombinasies geoptimeer word. Optimering word gedoen deur die beskikbare instellings vir die algoritmes iteratief te verander en die effek daarvan op die resultate van die algoritme te bepaal. Die proses word dan telkens vir elke kombinasie van die instellings wat vir die algoritme beskikbaar is, gedoen. Die optimale instellings word dan oorgedra na die ontwikkeling vir nege ander hulpbronskaars tale, te wete: isiNdebele, Sepedi, siSwati, Sesotho, Setswana, Xitsonga, Tshivenda, isiXhosa en isiZulu.

C. Ontwikkeling vir nabyverwante tale

As ’n uitbreiding van die eksperimente, word daar gekyk op watter manier die ontwikkeling van genreklassifikasiesisteme bevoordeel kan word deur gebruik te maak van die nabyverwantheid tussen tale. Afrikaans en Nederlands word gebruik in ’n reeks eksperimente in kombinasie met masjienvertaalsisteme waarvan die resultate voorgehou en bespreek word.

1.7. ONTPLOOIING

In Hoofstuk 2 word gefokus op die verskillende masjienleeralgoritmes wat algemeen in die literatuur gebruik word vir teksklassifikasie. Inleidend word teksklassifikasie in die breë sin omskryf, sowel as die onderliggende uitdagings en moontlike struikelblokke wat daarmee gepaardgaan. Elkeen van hierdie algoritmes word bespreek na aanleiding van die werking van die algoritme, die algoritme se sterkpunte en die toepassings waarvoor hierdie algoritmes in die algemeen gebruik word. Die uiteensetting van datastelle (d.i. die hoeveelhede data wat telkens gebruik word) vir die gebruik in genreklassifikasiesisteme word bespreek. ’n Oorsig word gegee van die verskillende wyses waarop die beskikbare data voorgestel/geënkodeer word voordat dit saam met die masjienleeralgoritmes gebruik word.

Hoofstuk 3 volg met ’n beskrywing van die ontwerp en die implementering van ’n genreklassifikasiesisteem vir Afrikaans, waaronder die eksperimente beskryf word waarvolgens die optimale kombinasie van algoritme, data, eienskappe en instellings vasgestel word. Soortgelyke sisteme uit die literatuur word voorgehou en daar word, sovêr moontlik, ’n vergelyking getref tussen die bestaande sisteme en die agtergrond waaruit hierdie sisteme ontwikkel is en die sisteem wat deur hierdie studie ontwikkel is. Die hoofstuk word afgesluit deur die beplande oordrag van die optimale instellings vir die ander hulpbronskaars tale van Suid-Afrika, asook die evaluering daarvan, te beskryf.

Hoofstuk 4 beskryf die ontwikkeling van genreklassifikasiesisteme vir nabyverwante tale. Hier word die moontlikheid ondersoek om nabyverwante tale te gebruik om die ontwikkeling van genreklassifikasiesisteme te bespoedig, deur gebruik te maak van ’n taal wat oor meer hulpbronne beskik, waarin ’n genreklassifikasiesisteem reeds bestaan. ’n Reeks eksperimente word vervolgens uitgevoer om te bepaal wat die resultate sou wees as ’n genreklassifikasiesisteem, wat in een van die nabyverwante

tale nog nie bestaan nie, maar wel in die ander taal bestaan, toegepas word op die taal waarin dit nie bestaan nie. Hierdie eksperiment word dan herhaal deur die taal waarvan die genre geklassifiseer moet word, outomaties te vertaal na die taal waarin die genre-klassifikasiesisteem bestaan. Hierdie outomatiese vertaling word gedoen deur masjienvertaalsisteme. Die twee nabyverwante tale onder bespreking is Afrikaans en Nederlands.

In Hoofstuk 5 word ’n samevatting van die studie gegee en gevolgtrekkings word gemaak oor die ontwikkeling van genreklassifikasiesisteme vir hulpbronskaars tale. Toekomstige navorsingsmoontlikhede en -onderwerpe word dan ter afsluiting van die studie genoem.

HOOFSTUK 2: VERWANTE NAVORSING

2.1. INLEIDING

Wanneer daar op die terrein van masjienleer, of meer spesifiek, teks-/(genre) klassifikasie gewerk word, is daar sekere onderwerpe wat in gedagte gehou moet word, ten einde die beste moontlike resultate te verkry. Eerstens moet daar besluit word op die algoritme wat gebruik gaan word om die klassifikasie te doen. Hierby moet daar in ag geneem word hoe die algoritme se interne implementering werk: bou dit besluitnemingsbome, word daar outomaties klassifikasiereëls afgelei, word daar statistiese berekeninge gedoen, word daar gebruik gemaak van vektoralgebra, ensovoorts. Weens die intrinsieke verskille tussen die beskikbare masjienleeralgoritmes is ŉ omvattende begrip van die werking daarvan nodig om te verseker dat die regte algoritme vir die bepaalde toepassing gekies word. Verder het ŉ masjienleeralgoritme gewoonlik ŉ stel veranderlikes/parameters waaraan gestel kan word om te verseker dat die algoritme vir die betrokke implementering pasgemaak word. Dit is daarom ook nodig om die algoritme se werking te maksimeer deur die optimale kombinasie van hierdie instellings te identifiseer.

Tweedens is masjienleeralgoritmes direk afhanklik van die data waarmee dit afgerig word. ŉ Algemene opvatting is dat die resultate van ŉ masjienleeralgoritme slegs goed kan wees as dit met ŉ groot hoeveelheid data van hoë gehalte afgerig word. Data kan egter nie in ŉ rou formaat gebruik word om ŉ masjienleeralgoritme mee af te rig nie. Dit is daarom nodig om uit die data ŉ stel eienskappe te identifiseer wat die prototipiese gevalle van die klasse wat geklassifiseer moet word, kan voorstel. Hierdie eienskappe moet dan onttrek word vanuit die data en geënkodeer word op so ŉ wyse dat dit deur die masjienleeralgoritme verstaan kan word. Alvorens die eienskappe geïdentifiseer kan word, moet daar eers besluit word hoe onbekende tekste geklassifiseer moet word.

Laastens, soos reeds genoem in afdeling 1.2, gaan hierdie navorsing daaroor om outomaties ŉ spesifieke genre aan ŉ onbekende teks toe te ken. Daarom is dit nodig om te weet wat genre is voordat dit aan ŉ teks toegeken kan word. Die genres wat geïdentifiseer word, word dan as klasse vir die masjienleeralgoritmes gebruik.

Samevattend: Die volgende benaderings en metodes uit die literatuur word in die hieropvolgende gedeeltes uiteengesit:

- Algoritmes;
- Eienskappe;
- Klasse; en
- Data (korpusse en datahoeveelhede).

2.2. ALGORITMES

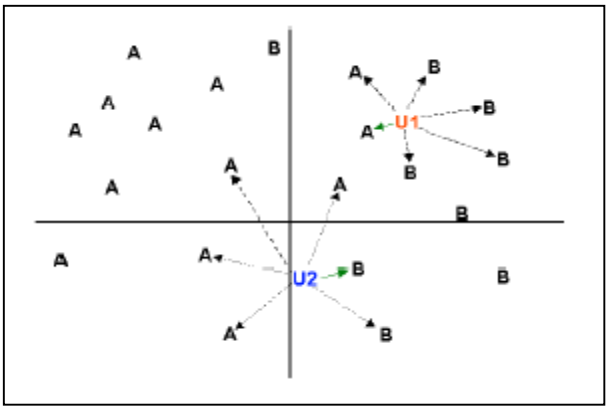
In die literatuur is daar 'n paar verskillende masjienleerbenaderings wat algemeen gebruik word as die kern vir teksklassifikasie. Khan *et al.* [24], bied 'n oorsig van die mees vername benaderings, waarvan enkele hier kursories bespreek word, te wete *k*-naastebuurpuntklassifiseerders, steunvektorklassifiseerders, multinomiale naïewe Bayes-klassifiseerders, besluitnemingsbome en die RIPPER-algoritme.

2.2.1. *k*-Naastebuurpuntklassifiseerders

Die *k*-naastebuurpunt (*k*-nn) benadering [24] 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 dan 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. Die mees frekwente klas onder die *k*-naastebuurpunte word dan 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. 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 [24]. Die benadering maak ook staat daarop dat die hele afrigtingstel beskikbaar is tydens klassifikasie en kan dus as "lui" beskryf word omdat daar nie baie prosessering gedoen word tydens die afrigtingsfase nie. Die *k*-nn benadering is 'n arbeidsintensiewe benadering omdat al die beskikbare eienskappe gebruik word om die afstande tussen die afrigtingsdata en die onbekende teks te bepaal. Hierdie afstandsmeting word

ingewikkelder namate die afrigtingsdata groei. Figuur 1 is ŉ voorstelling van onbekende tekste U1 en U2 wat in die multidimensionele eienskapruimte gestip word waarvan die vyf naaste buurpunte uitgewys word.



Figuur 1. Grafiese voorstelling van die *k*-nn benadering [24]

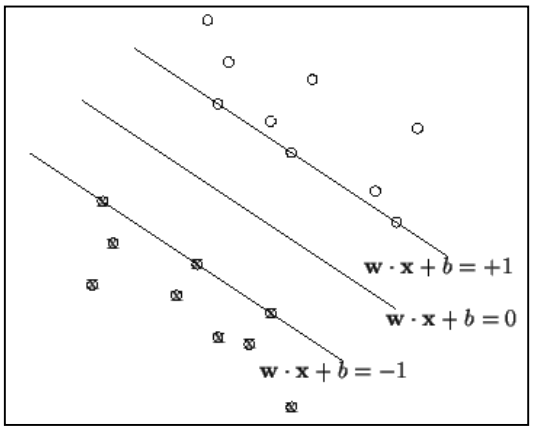
2.2.2. Steunvektorklassifiseerders

Die Steunvektorklassifiseerdermetode (SVM) is gebaseer op die minimering van strukturele risiko [24]. Strukturele risikominimering is ŉ masjienleerbeginsel wat deur Vapnik [53] soos volg verduidelik word: In masjienleer word ŉ model saamgestel uit ŉ eindigende datastel wat lei tot oormatige passing (d.i. die model word té spesifiek op die afrigtingstel gemodelleer en is nie meer algemeen genoeg om nuwe data te klassifiseer nie). Hierdie beginsel minimeer die oormatige passingsprobleem deur die model se pasgemaakte kompleksiteit te balanseer met die model se vermoë om veralgemenings te hanteer.

Die hoofidee agter die benadering is om ŉ hipotese te vind wat die laagste werklike fout waarborg, d.i. die beste skeiding bepaal tussen die verskillende klasse van die klassifikasieprobleem. Die afrigtingsdata vir die SVM word ook in die eienskapruimte gestip soos met die *k*-naastebuurpuntalgoritme, maar die verskil is dat die SVM beide positiewe en negatiewe afrigtingsdata vereis. Die SVM bepaal dan ŉ vlak in die ruimte wat die punte van die negatiewe en positiewe voorbeelde die beste verdeel. Hierdie vlak word dan die besluitnemingsvlak genoem. Dit word bepaal deur die ondersteuningsvektore te genereer aan die hand van die tekste wat naaste aan die natuurlike skeiding van die klasse lê. Hierdie parallelle vektore word dan die ondersteuningsvektore genoem en is in essensie die verteenwoordiging van die betrokke klas. Die besluitnemingsvlak word dan tussen die ondersteuningsvektore van elke klas vasgestel. Dit gebeur dat die verskillende klasse nie met ŉ liniêre vlak geskei

kan word nie, maar wel deur ŉ nie-liniêre een. Die eienskapsruimte word dan vertaal na ŉ ruimte met ŉ hoër dimensionaliteit, totdat die skeiding tussen klasse in die hoër dimensies wel liniêr kan wees Die hoeveelheid dimensies waardeur gesoek word na die optimale skeiding, word bepaal deur die algoritme se kompleksiteitsveranderlike (C). Hoe groter die waarde van hierdie veranderlike, hoe hoër die orde van die dimensies waardeur daar gesoek word na hierdie optimale skeiding. Die onbekende teks word ook in die n-dimensionele ruimte voorgestel, en afhangend van die posisie van die onbekende tekste teenoor die besluitnemingsvlak, word die klassifikasie bepaal.

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



Figuur 2. Skeiding van afrigtingsgevalle en ondersteuningsvektore in SVM [24]

ŉ Groot beperking van SVM's is dat so ŉ klassifiseerder slegs ŉ binêre klassifikasieprobleem kan hanteer. In ŉ 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 ŉ gewigtoekenningskema moet die beste klas vir die klassifikasie gekies word. Dit kan ŉ geweldige vergroting van die hoeveelheid klassifiseerders wat uiteindelik gebruik word om ŉ klas aan ŉ teks toe te ken in ŉ multiklasomgewing tot gevolg hê. Hierdie toename in kompleksiteit veroorsaak ŉ toename in die gebruik van fisiese geheue en verwerkingskrag. Die afrigtingsfase en klassifikasie neem ook derhalwe langer. Verder word daar verwarring opgemerk tydens klassifikasie, omdat daar ŉ klomp verskillende klasse aan die teks toegeken word met elke iterasie van vereenduidiging tussen die klasse.

2.2.3. Multinomiale naïewe Bayes-klassifiseerders

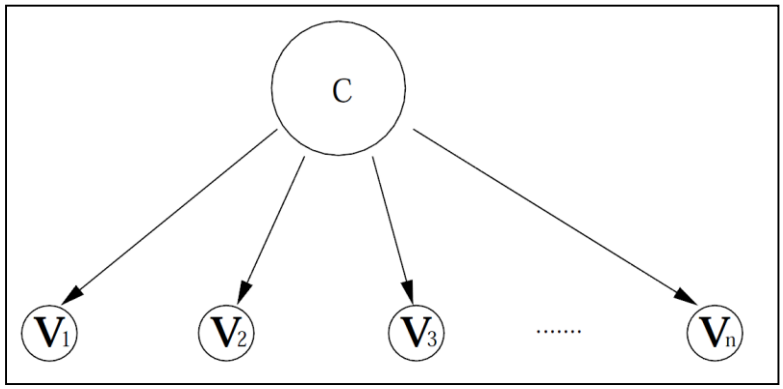
Naïewe Bayes-klassifiseerders is al suksesvol gebruik in vele domeine [60][34], ten spyte van die eenvoud van die model en die sterk onafhanklikheidsaannames wat dit maak. Peng en Schuurman [34] stel dat naïewe Bayes-klassifiseerders byna optimale prestasie kan bereik, al konformeer die domein onder bespreking glad nie tot die onafhanklikheidsaannames nie. Naïewe Bayes-klassifiseerders is gebaseer op 'n eenvoudige toepassing van Bayes se Wet vanuit die waarskynlikheidsleer. Bayes se Wet word deur Vergelyking (4) voorgestel.

$$\boldsymbol{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}} \tag{4}$$

Bayes se Wet breek die voorwaardelike waarskynlikheid van 'n onbekende gebeurtenis in 'n paar kleiner waarskynlikhede op, wat makliker is om te bereken. Hierdie ontbinding 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 (4) dus soos in Vergelyking (5) herskryf (met onbekende teks *W* en die onbekende klas *C*).

$$\boldsymbol{P(C|W) = \frac{P(W|C)\,P(C)}{P(W)}} \tag{5}$$

Die onafhanklikheidsaanname wat deur Bayes-klassifiseerders gemaak word, stel dat alle eienskapwaardes V_j (soos gebruik in Vergelyking (6)) onafhanklik is, gegewe die klas etiket *C.* [34]. Die onafhanklikheid van die eienskapsnode van elke ander eienskapsnode, gegewe die klasetiket C, word in Figuur 4 uitgebeeld en kan word die onafhanklikheidsaanname hier uitgebeeld.



Figuur 3. Voorstelling van die naïewe Bayes-benadering [34]

As gevolg van die naïewe aanname (naïef omdat die aanname selde waar is), kan die vergelyking vereenvoudig word na Vergelyking (6).

$$\boldsymbol{P(C|W) = P(C) \times \frac{\prod_j P(v_j|C)}{P(W)}} \quad (6)$$

Vergelyking (6) kan nog verder vereenvoudig word na Vergelyking (7), omdat *P(W)* nie verander oor die reeks kategorieë nie en dus uit die vergelyking gelaat kan word.

$$\boldsymbol{P(C|W) = P(C) \times \prod_j P(v_j|C)} \quad (7)$$

Uit die vergelykings hierbo genoem, word die vergelyking vir die berekening van die waarskynlikste klas vir ’n onbekende teks voorgestel in Vergelyking (8).

$$\boldsymbol{c^* = \arg max \left\{P(c) \times \prod_j P(v_j|C)\right\} \quad c \in C} \quad (8)$$

Die logiese uiteensetting van ’n naïewe Bayes-klassifiseerder is baie eenvoudig en maklik verstaanbaar en maklik aanpasbaar vir nuwe probleme. Sterk aannames word gemaak met die gebruik van naïewe Bayes, naamlik dat elke woord in ’n teks onafhanklik is van elke ander woord in die teks en dat ’n woord se posisie in die teks irrelevant is. Ten spyte hiervan lewer dit steeds kompeterende resultate [1]. Die uitvoerkompleksiteit van naïewe Bayes-klassifikasie is liniêr, wat beteken dat die klassifikasie vinnig gedoen kan word.

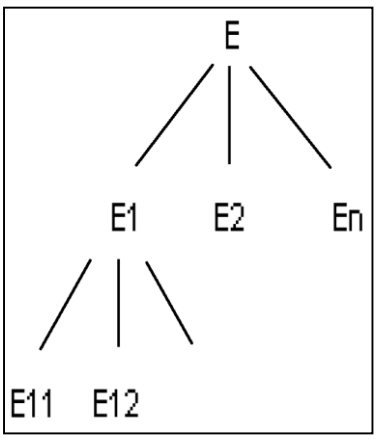
McCallum en Nigam [30] stel dat ’n variant van die klassieke naïewe Bayes-klassifiseerder, naamlik die multinomiale naïewe Bayes-klassifiseerder (MNB) meer geskik is vir gebruik as ’n teksklassifiseerder as die klassieke naïewe Bayes-benadering. Die MNB-klassifiseerder is ’n aanpassing van standaard naïewe Bayes, waar woordfrekwensies ook in ag geneem kan word [30]. Standaard naïewe Bayes neem die veronderstelling dat die voorkomste van ’n woord nie van belang is vir ’n betrokke klas nie en word alle woorde herlei na ’n enkel voorkoms in die waarskynlikheidsfunksie. Woordfrekwensie kan egter van groot belang wees tydens genreklassifikasie. Neem ’n advertensie as voorbeeld: ’n Advertensie sal heel waarskynlik baie meer voorkomste van woorde soos “spesiale aanbieding” en “uitverkoping” hê as wat ’n akademiese teks sal hê. As hierdie voorkomste in ag geneem word, kan dit ’n positiewe bydrae lewer tot die korrektheid van klassifikasie.

Die belangrikste voordeel van ’n MNB-klassifiseerder is dat dit ’n (relatiewe) klein hoeveelheid afrigtingsdata nodig het om kompeterende resultate te lewer [24]. 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 [9] gebruik dit vir die restourering van Maori diakritiese tekens wat verlore gegaan het tydens teksverwerking. Mogadala en Varma [31] gebruik MNB vir die onttrekking van opinies uit nuusartikels wat in Hindi geskryf is en Peché *et al*. [33] gebruik MNB vir gesproke taalidentifisering.

2.2.4. Besluitnemingsbome

Besluitnemingsbome [24] 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.



Figuur 4. Eenvoudige voorstelling van 'n besluitnemingsboom [24]

Die hoofvoordeel van besluitnemingsbome is dat dit eenvoudig is om te verstaan en te interpreteer. Die uitkomste van 'n spesifieke klassifikasiegeval 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 as 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 met besluitnemingsbome is dat oormatige passing met die afrigtingsdata maklik gebeur en die sisteem dan faal wanneer dit op onbekende data en data uit ander domeine toegepas word.

2.2.5. 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 [10].

Die afrigtingsdata word lukraak verdeel in 'n groeistel en 'n snoeistel in 'n 2:1-verhouding. 'n Reeks reëls word deur die algoritme gegenereer vir elke moontlike klaswaarde. Die groeifase het tot gevolg dat reëls bygevoeg word in die volgorde van minder frekwente reëls tot die mees frekwente reël. 'n Reël word saamgestel deur veronderstellings tot die reël by te voeg totdat die maksimum beskrywende lengte van die reël bereik word. 'n Veronderstelling word gekies om bygevoeg te word op grond van die veronderstelling se bydrae tot die inligtingswins van die groeiende stel en wat die deskriptiewe lengte van die reël maksimeer. 'n Reël sal aanhou groei totdat daar geen verdere inligtingswins moontlik is nie. Die reël word dan met die snoeistel gesnoei deur veronderstellings weg te laat sodat die funksie soos in Vergelyking (9) gemaksimeer word:

$$\boldsymbol{v} \equiv \frac{\boldsymbol{p}+(\boldsymbol{N}-\boldsymbol{n})}{\boldsymbol{P}+\boldsymbol{N}} \tag{9}$$

Waar *P* = die aantal positiewe voorbeelde in die snoeistel, *N* = die aantal negatiewe voorbeelde in die snoeistel, *p* = die aantal positiewe voorbeelde wat deur die reël gedek word en *n* = die aantal negatiewe voorbeelde wat deur die reël gedek word. Die minimum deskriptiewe lengte vir elke reël in die versameling word dan bepaal en die reëls met die kortste deskriptiewe lengte wat steeds die funksie maksimeer, word dan gebruik as die geoptimeerde stel reëls waarvolgens klassifikasie dan gedoen word.

Van die hoof voordele van die RIPPER-algoritme is dat RIPPER oor die algemeen vinniger is as ander algoritmes omdat die reëls in liniêre tyd geïnduseer word. Cohen [10] stel verder dat RIPPER ook beter resultate toon as ander algoritmes wanneer die afrigtingstel uitskieters of ander "geraas" bevat.

2.3. EIENSKAPPE

Eienskapseleksie verwys na die identifisering van ŉ stel eienskappe vanuit die afrigtingsdata wat die afrigtingsdata as't ware kan beskryf. Hierdie eienskappe moet dan geënkodeer word op só ŉ wyse 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 daarvolgens onbekende tekste 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.

2.3.1. Woordversameling

ŉ Woordversamelingbenadering is die eenvoudigste vorm waarin eienskappe van afrigtingsgevalle voorgestel kan word [56]. Dit behels dat alle woorde in die afrigtingsteks net soos wat dit in die teks voorkom aan die masjienleeralgoritme gegee word. Die woordversamelingvoorstelling van ŉ teks word dikwels as binêre vektor weergegee. Op hierdie wyse word slegs die teenwoordigheid of die afwesigheid van ŉ woord in die afrigtingsgeval aangedui [12]. Die aanwesigheid van ŉ woord in ŉ afrigtingsteks word aangedui deur al die woorde, in al die afrigtingstekste in ŉ skikking te stoor. Daar word vir elke afrigtingsteks dan ŉ vektor saamgestel waarvan elke veld in die vektor verwys na ŉ indeks van die skikking. As die woord in die betrokke indeks in die skikking ook in die afrigtingsteks voorkom, word daar ŉ een in die ooreenstemmende veld van die vektor gestoor, en as die woord nie voorkom nie, word daar ŉ nul gestoor. Tabel 1 wys ŉ voorstelling van die woordvoorkomste wat as binêre vektore gestoor word.

Afrigtingstekste		**t1**	**t2**	**t3**	**t4**	**t5**	**...**	***tn***
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

Tabel 1. Voorstelling van die woordversamelingvektore

2.3.2. *tf-idf*-tellings

Die tweede stel eienskappe wat algemeen gebruik word, is *tf-idf*-tellings waar *tf* die termfrekwensie is en *idf* die inverse van die dokumentfrekwensie. Om die frekwensie van 'n term (woord) in 'n afrigtingsgeval te bereken (d.i. *tf*), word die hoeveelheid voorkomste van die woord in die afrigtingsgeval getel en die produk daarvan word geneem met die inverse van die hoeveelheid afrigtingsgevalle waarin die term voorkom (d.i. *idf*) [29][56]. Die eenvoudigste formule vir die berekening van 'n *tf-idf*-telling word in Vergelyking (10) voorgehou.

$$(\boldsymbol{tf} \cdot \boldsymbol{idf})_{i,j} = \boldsymbol{tf}_{i,j} \times \boldsymbol{idf}_{i,j} \tag{10}$$

Die waarde van 'n *tf-idf*-telling kan vir die algoritme 'n aanduiding gee van die belangrikheid van 'n woord se bydrae tot die identifikasie van die klas. As 'n term herhaaldelik voorkom in 'n betrokke afrigtingsgeval, 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 in die afrigtingsgevalle mag voorkom. Die *tf-idf*-tellings word dan op 'n soortgelyke wyse as die woordvoorversamelingbenadering in 'n vektor geënkodeer. Die groot verskil is egter dat daar nou gewigte toegeken word aan elkeen van die woorde wat wel voorkom in die betrokke afrigtingsgeval. Vergelyking (11), toon aan hoe die *tf-idf*-tellings in WEKA [18] bepaal word deur gebruik te maak van logaritmiese terme om *idf* te normaliseer.

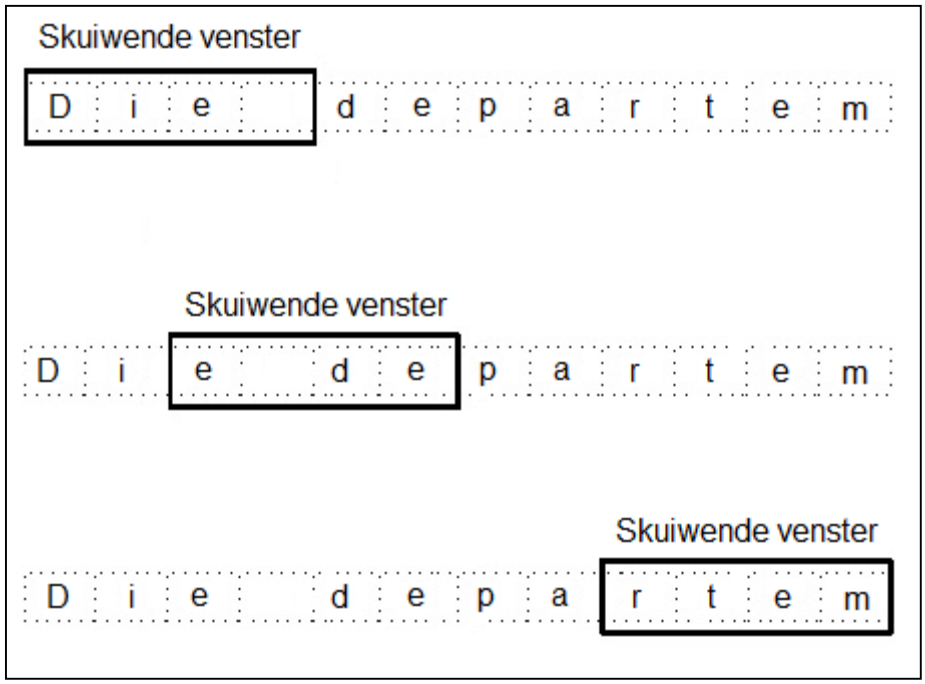
$$(\boldsymbol{tf} \cdot \boldsymbol{idf})_{ij} = \boldsymbol{tf}_{ij} \log \boldsymbol{idf}_{i,j} \tag{11}$$

Afrigtingstekste		**t1**	**t2**	**t3**	**t4**	**t5**	**...**	***tn***
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

Tabel 2. Voorstelling van die *tf-idf*-vektore

2.3.3. 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 5). Karakter-*n*-gramme het die voordeel dat dit die moontlikheid het om morfologiese inligting van die woorde in die afrigtingstel vas te vang. Morfeme bestaan selde uit ’n groot hoeveelheid karakters, en daarom kan daar met ’n venster van ’n klein grootte gewerk word, wat weer die hoeveelheid eienskappe per afrigtingsgeval vergroot. Dit het die moontlikheid om die afrigtingsgeval vollediger voor te stel. Daar kan ook inligting oor leestekens, skryftekens en spasiegebruik ingewin word as die venster gekonfigureer word om die voorkomste daarvan waar te neem. Die hoeveelheid ongewone tekens en spasies kan ook ’n goeie identifiseerder wees by die klassifikasie van formele teenoor informele tekste. Dié benadering kan maklik aangepas word om woord-*n*-gramme ook te konstrueer deur die venster oor die *n*-hoeveelheid woorde op ’n slag te skuif. Sodoende kan die mede-voorkomstes (d.i. woorde wat in dieselfde omgewing voorkom in ’n teks) wat prototipies van ’n klas is, vasgevang word in die voorstelling.



Figuur 5. Voorbeeld van ’n skuiwende venster vir *n*-gram-onttrekking

2.3.4. Woordsoortinligting

Die gebruik van woordsoortinligting as eienskappe vir genreklassifikasie is ŉ benadering wat in die literatuur baie vrugte afwerp [12]. Die woorde in die teks moet geanaliseer word om die woordsoort van die woord in sy betrokke konteks te bepaal. Hierdie woordsoortetikettering kan óf handmatig óf outomaties gedoen word. Om die woordsoorte outomaties toe te ken, sal die ideaal wees, omdat dit baie tydrowend sal wees om die woordsoortetikette handmatig toe te ken. Die inligting wat van hierdie analise verkry word, word dan gebruik as eienskappe (gewoonlik gepaardgaande met ander eienskappe) vir die masjienleeralgoritme. Hierdie benadering is egter nie geskik vir hulpbronskaars tale nie, omdat die outomatiese annotasie van die woordsoorte staat maak op die beskikbaarheid van ondersteunende hulpbronne (d.i. woordsoort-etiketteerders) wat nie beskikbaar is vir die meerderheid van hulpbronskaars tale nie.

2.3.5. Teksstatistiek

Teksstatistiek is van die eienskappe wat die maklikste vir ŉ afrigtingstel bepaal kan word. Dit sluit die volgende eienskappe in:

- Woordlengtes (die gemiddelde hoeveelheid karakters in ŉ woord);
- Sinlengtes (die gemiddelde hoeveelheid woorde in ŉ sin);
- Karaktertellings (die relatiewe spreiding van karakters in ŉ teks, gebaseer op frekwensies);
- Lettergreeptellings (die gemiddelde hoeveelheid lettergrepe in ŉ woord);
- Woordtellings (die frekwensies van woorde in ŉ teks);
- Sintellings (die hoeveelheid sinne in ŉ teks);
- Paragraaftellings (die hoeveelheid paragrawe in ŉ teks); en
- Leesbaarheidsmetrieke (d.i. metrieke wat gebruik word om die moeilikheidsvlak, ten opsigte van die lees van ŉ teks, te kwantifiseer), ens.

Hierdie eienskappe (met die uitsondering van leesbaarheidsmetrieke) is ook taalonafhanklik en is dus ideale eienskappe om vir hulpbronskaars tale te gebruik omdat die eienskappe nie staatmaak op bestaande kerntegnologieë soos woordsoortetiketteerders, morfologiese analiseerders, ensovoorts nie.

2.4. KLASSE

Daar is in die literatuur ’n mate van verdeeldheid oor wat die term “genre” regtig behels. Talle definisies word gegee om te poog om genre te omskryf. Vervolgens word ’n aantal van hierdie definisies voorgehou:

“We will use the term genre here to refer to any widely recognized class of text defined by some common communicative purpose or other functional traits, provided the function is connected to some formal cues or commonalities and that the class is extensible.” [23]

“They [genres] *have often been characterized in terms of purpose, form. This means that documents belonging to the same genre share the same purpose and the same form, either for the language and/or the layout.”* [44]

“...genres cover different properties of both documents and texts, such as their form, function, purpose, and target audience.” [55]

“[Genre definitions] *rely on a combination of two notions: one of structure and one of function. Structure is defined by factors which are reflected in the visual layout of the document while function is defined by the intended purpose of the document.”* [25]

“Genre has a range of definitions, but for Language Technology, a good one is a class of documents that share a communicative purpose.” [37]

’n Algemene tendens wanneer genre gedefinieer word (veral in die literatuur rakende genreklassifikasie) is dus om te verwys na die doel (funksie), teikengehoor of struktuur (uitleg) van ’n teks. Daar is ook verwysing na genre as kommunikatiewe gebeure, met ’n gedeelde kommunikatiewe doel [49] of karakteristieke tipe kommunikatiewe aksie, wat gekenmerk word deur ’n sosiaal aanvaarde kommunikatiewe doel en gemeenskaplike aspekte van styl [58][12]. Die veronderstelling uit die bostaande literatuur is dat die genre van ’n teks verder strek as die ooglopende eienskappe van die teks. Die genre van ’n teks word bepaal deur “eksterne” eienskappe waarna daar nie (noodwendig)

direkte verwysing in die teks gevind kan word nie, maar waar die eienskappe geïmpliseer word [3][37]. Die definisies vir genreklassifikasie is samevoegings en aanpassings van definisies van genre uit klassieke linguistiekliteratuur [49][3][47][58][12], waarvan die invloedrykste waarskynlik dié van Biber is [47]. Die definisie van Biber is van die vroeë definisies van genre en daarom word baie van die nuwer definisies daarop gebaseer.

Biber definieer genre soos volg:

"Genre categories are determined on the basis of external criteria relating to the speakers' purpose and topic; they are assigned on the basis of use, rather than on the basis of form. It is also possible to consider groupings of texts that are derived on the basis of linguistic form." [3]

Nog 'n tendens is om bloot nie melding te maak van 'n definisie vir genre nie [44]. Genre word dan bloot, vir die meerderheid van die navorsing wat oor genreklassifikasie handel, herdefinieer (of vae definisies van ander navorsing) wanneer dit gebruik word [44][55][25][37], maar ten spyte van die "nuwe" definisies, word ooreenkomste met die bogenoemde algemene/vae definisies steeds opgemerk. Dit sorg vir verwarring van wat genre eintlik is en hoe daar daarmee omgegaan moet word by die klassifikasie daarvan [44][55]. Selfs wanneer daar nie van outomatiese klassifikasiemetodes gebruik gemaak word nie, is daar verdeeldheid oor die toekenning van genre aan 'n teks. Dit gebeur baie dat menslike annoteerders van mekaar sal verskil oor die betrokke genre van 'n teks omdat daar bykans by alle tekste 'n toekenning van meer as een genre kan geskied. Hierdie omstredenheid rondom 'n definisie van genre is dan die beperking van 'n masjienleerbenadering tot genreklassifikasie, wat vereis dat genre (as dit in hierdie konteks gebruik word) as diskrete entiteite gesien moet word en dat dit nie moontlik is om meer as een genre aan 'n klas toe te ken nie [12][45]. Die meerderheid van die navorsing oor genreklassifikasie word dan ook gebaseer op hierdie "diskrete, enkel-klas, *supervised* klassifikasie" [44]. Daarom word navorsers "gedwing" om te besluit op genreklasse wat nie die omvang van genre regtig kan bevat nie, maar wat binne hul beperkte domein tog tot 'n mate van waarde kan wees [44][55][35].

Tipiese genreklasse uit die literatuur sluit onder andere die volgende in:

- Advertensie;
- Amptelike teks;
- Bespreking;
- Biografie;
- Dialoog;
- Drama;
- Fiksie;
- Glansartikel;
- Hulpgids;
- Inligting;
- Instruksie;
- Nie-fiksie;
- Nuus;
- Opstel;
- Poësie;
- Brief;
- Produkaanwysing;
- Toesprake;
- Tuisblaaie;
- Verslag;
- Dagboekinskrywing; en
- Notule.

Die tipe klasse wat gekies word, sal bepaal hoe die data voorgestel word om die maksimum hoeveelheid inligting vanuit die data vir ’n klas te verseker. Verder is dit ook belangrik dat die data ’n goeie verteenwoordigendheid toon. Die verteenwoordigendheid van die data is belangrik, sodat die afrigtingsdata ’n akkurate weerspieëling is van die algemene voorkomste van die betrokke klasse en dat die spreiding van die klasse ongeveer gebalanseerd is. Die hoeveelheid klasse waartussen die masjienleeralgoritme moet onderskei het ’n invloed op die kompleksiteit van die algoritme, en daarom moet daar ook aandag gegee word daaraan om die hoeveelheid klasse so omvattend moontlik te maak, maar terselfdertyd die hoeveelheid klasse te probeer beperk.

2.5. DATA

Soos reeds genoem, maak masjienleeralgoritmes staat op die beskikbare afrigtingsdata, en die uiteindelike prestasie van die algoritme hou direk verband met die kwaliteit, kwantiteit en tipe afrigtingsdata. Regdeur die literatuur word eksperimente in 'n wye reeks domeine, vir 'n wye reeks toepassings gedoen. Dit is dus vanselfsprekend dat die samestellings van die afrigtingstelle (data kwaliteit, kwantiteit en bron), sowel as die eienskappe en klasse waarmee die data voorgestel word, vir elke eksperiment anders sal wees. Telkens word daar ook verskillende algoritmes gebruik en die eksperimente word ook in verskillende tale gedoen. In Tabel 3, word 'n paar voorbeelde van die bogenoemde eksperimente uit die literatuur weergegee. Die uiteindelike *f*-telling vir elke eksperiment word ook genoem. Hierdie tabel bied 'n oorsig wat met een oogopslag die verskille tussen die eksperimente uitwys en word ook later (sien Tabel 16) gebruik om as verwysingsraamwerk te dien vir die resultate vir die eksperimente wat in hierdie studie uitgevoer word. Eksperimente met verskillende tale, eienskappe en klasse kan egter nie direk vergelyk word nie, maar behoort tenminste 'n aanduiding gee oor die verwagte prestasie vir soortgelyke sisteme.

# Tekste	Korpus	# Klasse	Eienskappe	Taal	Algoritme	*f*-Telling	
319	Internet: Tesisse en verhandelinge	11	*tf-idf*	Engels	NB[5]	0.890	[60]
1224	Webdokumente	16	HTML-etikette Webadresinligting Leksikale eienskappe	Engels	SVM	0.757	[27]
800	Nuus webblaaie	2	Woordversameling Woordsoortinligting Teksstatistiek	Engels	Besluitnemings-bome	0.905	[12]
1083	Koerantberigte	20	Letter 5-gramme Morfologiese inligting	Duits	SVM	0.540	[14]
499	Brown Korpus	6	Strukturele inligting Leksikale eienskappe Teksstatistiek	Engels	*k*-NN	0.870	[23]

Tabel 3. Opsomming van data uit die literatuur

[5] Standaard Naïewe Bayes-klassifiseerder

2.6. VOORPUNTNAVORSING

Yi-Hsing en Hsiu-Yi [60] hou die ontwikkeling van ŉ genreklassifikasiesisteem vir die klassifikasie van elf webgenres voor. ŉ Naïewe Bayes-benadering word gevolg vir die klassifikasie-algoritme. Ondersteunend tot die klassifikasie-algoritme word ŉ domeinontologietabel gebruik. Die tabel stel verhoudings tussen die verskillende woorde in die afrigtingsdata vas en herlei alle woorde met dieselfde betekenis na enkele sinonieme toe. Daar word dan gewigte toegeken aan die verhoudings tussen die sinonieme en die verhouding daarvan met die genreklas wat daardeur verteenwoordig word. Hierdie genormaliseerde afrigtingsdata word dan aan die klassifikasie-algoritme oorgedra. Hierdie genreklassifikasiesisteem word dan geïmplementeer as deel van ŉ groter dokumentbestuurstelsel. Yi-Hsing en Hsiu-Yi [60] rapporteer ŉ gemiddelde *f*-telling van 0.890, gemeet oor al elf klasse. Om afrigtingsdata te verkry, gebruik Hsing en Hsiu-Yi [60] ŉ "Collection and Split"-module wat outomaties Engelse tesisse en verhandelings uit ŉ beperkte domein vanaf die Internet verkry en wat dan die klassifikasie-algoritme afrig met 70% van die data. Die ander 30% van die data word dan as toetsdata gebruik. In totaal word daar 319 tekste gebruik in die finale afrigtingsdatastel.

In ŉ poging om die beste moontlike stel eienskappe vir genreklassifikasie van webdokumente te bepaal, onttrek Lim *et al*. [27] HTML-etikette, URL-inligting, leksikale eienskappe en strukturele inligting uit 1 224 tekste. Lim *et al*. [27] meld ŉ gemiddelde presisie van 0.757 oor 16 klasse by die gebruik van die beste kombinasie van die bogenoemde eienskappe. Timbl [11] se k-Naastebuurpuntalgoritme word gebruik om die klassifikasie mee te doen.

Fin and Kushmerick [12] weeg die impak van drie stelle eienskappe op genreklassifikasie van tekste. Besluitnemingsbome word gebruik om tussen twee klasse te onderskei, te wete opinies- en feitlike tekste afkomstig uit drie verskillende domeine. Vir afrigtingsdata gebruik Fin en Kushmerick [12] ŉ korpus van ongeveer 800 tekste waaruit eienskappe soos woordsoortinligting, teksstatistiek en woordversameling onttrek word. Hulle noem presisiesyfers van tussen 0.824 en 0.905, afhangend van die kombinasie van die bogenoemde eienskappe.

2.7. SAMEVATTING

In hierdie hoofstuk word die belangrikste masjienleerbenaderings tot genreklassifikasie uit die literatuur voorgehou in afdeling 2.2, naamlik:

- *k*-Naastebuurpuntalgoritme;
- Steunvektorklassifiseerders;
- Multinomiale naïewe Bayes-klassifiseerder;
- Besluitnemingsbome; en die
- RIPPER-algoritme.

Die benaderings word bespreek aan die hand van die werking van die algoritme, die geskiktheid daarvan vir genreklassifikasie en die voor- en nadele van die gebruik van die algoritme.

Die verskillende maniere waarop afrigtingsdata vir die masjienleer-benaderings as eienskappe geënkodeer word, word in afdeling 2.3 bespreek. Hierdie eienskappe is:

- Woordversameling;
- *tf-idf*-tellings;
- Karakter en woord *n*-gramme;
- Woordsoortinligting; en
- Teksstatistiek.

Ter afsluiting van die hoofstuk word die klasse (d.i. die kategorieë waarin die masjienleeralgoritmes onbekende tekste klassifiseer) bespreek wat algemeen gebruik word by genreklassifikasie. Verder word ’n opsomming van die samestelling van afrigtingstelle vir genreklassifikasiesisteme uit die literatuur voorgehou en ’n uittreksel van die voorpuntnavorsing word kortliks bespreek.

In die volgende hoofstuk word hierdie masjienleerbenaderings, eienskappe, klasse en data gebruik as basis vir ’n reeks eksperimente om die beste kombinasie daarvan vir die gebruik by genreklassifikasie van hulpbronskaars tale te identifiseer.

HOOFSTUK 3: EKSPERIMENTERING

3.1. INLEIDING

In hierdie gedeelte word daar 'n aantal eksperimente bespreek wat uitgevoer is, ten einde die beste algoritme, eienskappe, datagrootte en die hoeveelheid klasse te bepaal. Soos reeds genoem in Hoofstuk 2 speel hierdie veranderlikes almal 'n rol in die uiteindelike effektiwiteit van 'n sisteem wat op masjienleeralgoritmes gebaseer is. Elkeen van die veranderlikes word nou geëvalueer deur die ander veranderlikes konstant te hou om sodoende die effek van elkeen van die veranderlikes vas te stel. Hierdie proses van eliminasie word spesifiek so saamgestel om die uiteindelike aantal eksperimente tot 'n mate te probeer beperk om sodoende die hoeveelheid tyd wat dit in beslag neem, te minimeer. Sou 'n eindigende stel eksperimente uitgevoer moes word, sou die hoeveelheid eksperimente te veel raak. Die hoeveelheid eksperimente van 'n eindigende soektog deur die eksperimente sou soos volg bereken kon word:

$$\left(5Algoritmes \times \left(2Klasstelle \times \left(6Eienskapstelle \times (10Datainkremente)\right)\right)\right)$$

$$= 600\ Eksperimente \tag{9}$$

Daar sou dan vir elkeen van die 600 eksperimente 'n algoritme-optimeringsfase gedoen moes word waar daar vir sommige van die algoritmes meerdere parameters is wat geoptimeer kan word. Slegs net met een optimeringsfase sou die totale hoeveelheid eksperimente wat uitgevoer moet word, meer as 1 200 eksperimente wees. Dit is daarom van groot nut indien hierdie hoeveelheid beperk kan word.

3.2. EKSPERIMENTELE OPSTELLING

3.2.1. Klasse

In afdeling 2.4 word daar ŉ oorsig gegee uit die literatuur van die klasse wat algemeen in outomatiese genreklassifikasie-eksperimente gebruik word. Vervolgens word daar vir die eksperimentele doeleindes van hierdie studie twee stelle klasse geïdentifiseer, te wete konkrete en abstrakte klasse. Hierdie twee stelle klasse word in die volgende afdelings bespreek.

3.2.1.1. Konkrete klasse

Wachsmuth en Bujna [55] verwys na die benadering tot genre uit die literatuur (afdeling 2.4) as konkrete klasse, omdat hierdie klasse teruggeneem kan word na konkrete entiteite waar die klas sy oorsprong vandaan het (byvoorbeeld ŉ advertensie). Soos reeds genoem in afdeling 2.4 is die navorsing oor genreklassifikasie tot op hede tot ŉ groot mate gefragmenteer en kan dit nie maklik met mekaar vergelyk word nie. Die rede hiervoor is dat daar ŉ gebrek is aan duidelike “voorskrifte” vir die definisie van en omgang met genre en genreklasse [55][45]. Verskillende studies dek ook ŉ verskillende bestek van klasse, omdat die bronne van die data waarop die studies gebaseer is, verskil en omdat die navorsing sigself met verskillende dimensionaliteite van tekste bemoei. Verder word hierdie studies ook in verskillende domeine gedoen. Die data wat in hierdie studies gebruik word, word gewoonlik uit een spesifieke domein geneem, omdat domein-spesifieke klassifikasie gewoonlik beter resultate lewer [12][55]. Genreklassifikasie faal ook gewoonlik wanneer dit oor domeingrense heen gebruik word (d.i. domeinoordrag), juis oor hierdie vaagheid van die genreklasse en die gebrek aan duidelikheid oor die konsep wat versoenbaarheid belemmer. Wat in die een domein ŉ sinvolle klas kan wees, is nie noodwendig in die volgende sinvol nie [55][45]. Finn en Kushmerick [12] dui ŉ afname van tussen 13.2% en 29.7% in die presisie van ŉ genreklassifikasiesisteem aan wanneer die sisteem wat afgerig is in die een domein (byvoorbeeld sport), toegepas word op ŉ ander domein (byvoorbeeld politiek of finansies).

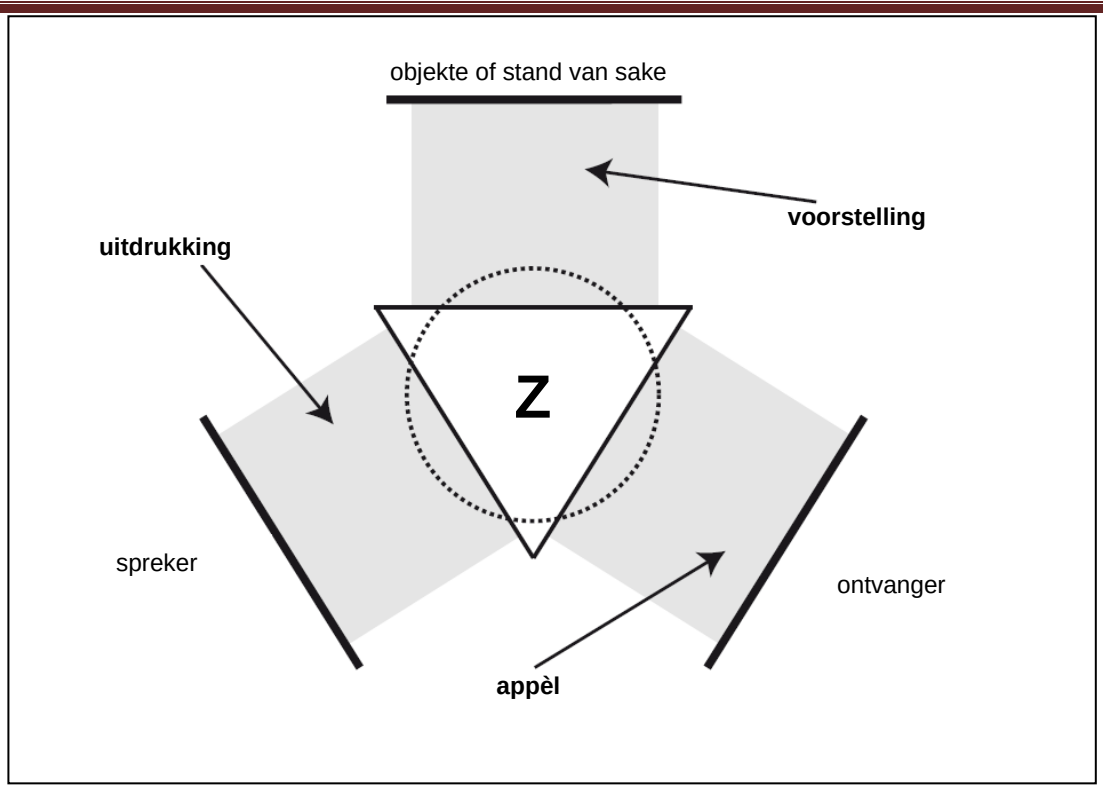
Hoewel hierdie beperktedomeinklassifikasie van waarde kan wees, is dit veral in die konteks van implementeerbaarheid belangrik om domeinoordrag te kan hanteer, omdat regtewêreldtoepassings selde in een domein gebeur. Ook met verwysing na die beperkte hulpbronne vir die Suid-Afrikaanse tale kan dit van waarde wees om klasse

meer generies te maak na aanleiding van die domein waarin dit gebruik kan word en om daarom dalk weg te beweeg van die eksplisiete verdeling in streng diskrete klasse wat veroorsaak dat sommige klasse in hulpbronskaars omgewings ernstig onderverteenwoordig word. In die volgende gedeelte word 'n moontlike oplossing vir die onderverteenwoordigendheid van klasse ondersoek.

3.2.1.2. Abstrakte klasse

Wachsmuth en Bujna [55] argumenteer dat om op slegs een van die eienskappe van genre te fokus, die geval van onvergelykbare genreklasse oorkom kan word. Hulle kyk in [55] dus slegs na die funksie van 'n teks. Soos reeds genoem in afdeling 2.4 is genre 'n konsep wat meerdere fasette omvat, en om 'n genre werklik te kan identifiseer, behoort al hierdie fasette in ag geneem te word. Deur net op die funksie van 'n teks te fokus, word daar in 'n mate 'n skuif gemaak na funksieanalise in plaas van volledige genreklassifikasie, maar dit hou egter geen nadeel vir die prestasie van 'n genreklassifikasiesisteem in nie. Wachsmuth en Bujna [55] redeneer verder dat dit 'n konstante eienskap bied om te analiseer, en sou dit deurlopend in navorsing gebruik word, sal die resultate wat in verskillende navorsing verkry word, meer vergelykbaar maak. Hulle [55] definieer dan drie diskrete klasse wat die funksie van 'n teks kan beskryf: ekspressiewe teks (waarvan die funksie van die teks nie kommersieel is nie en waar 'n persoonlike mening weergegee word), appellatiewe teks (die teks het 'n kommersiële funksie, soos om die leser te oortuig om 'n produk te koop) en informatiewe teks (inligting wat nie kommersieel georiënteer is nie, maar wat inligting op 'n joernalistieke wyse weergee) [55].

Wachsmuth en Bujna [55] het hierdie klasse afgelei uit Karl Bühler [4] se "Organon"-model vir die funksie van kommunikasie. Die verhouding tussen die funksie van 'n teks (kommunikatiewe gebeurtenis), gegewe 'n linguistiese uitdrukking Z, word in Figuur 6 uitgebeeld.



Figuur 6. “Organon”-model van Karl Bühler [55]

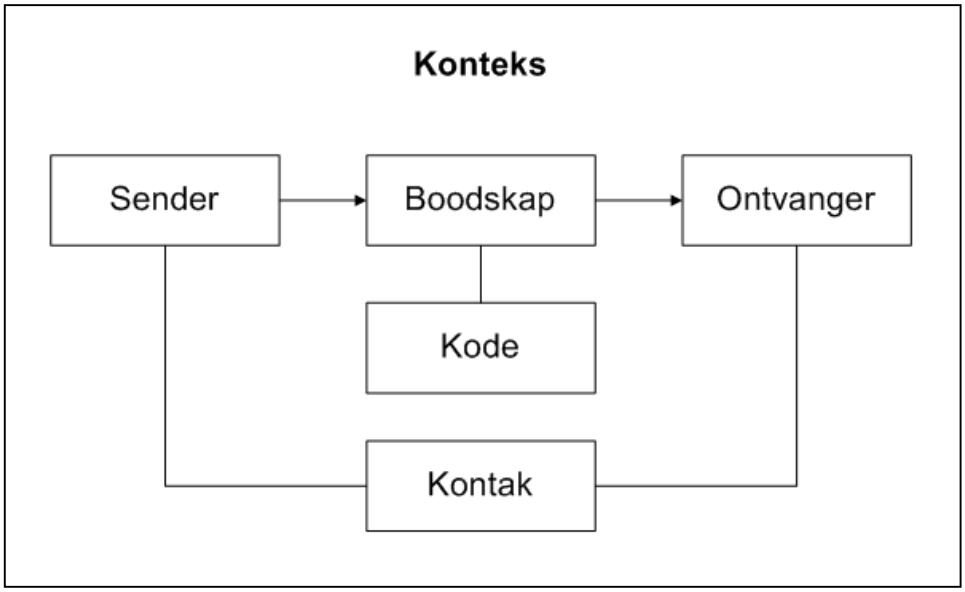
Bühler se model kan soos volg beskryf word [42]: Bühler se model beskryf die kommunikasie tussen ’n sender en ’n ontvanger deur ’n derde party (d.i. objekte of die stand van sake) by die kommunikasie te betrek. ’n Kommunikatiewe funksie word dan toegeken aan die kommunikasie gebaseer op die deelnemer (sender, ontvanger of objek) wat die meeste steun geniet. As die fokus op die sender is, word daar na ’n ekspressiewe funksie verwys; as die fokus op die objek is, word daar na ’n voorstellende (vgl. informatiewe teks van [55]) funksie verwys; en wanneer die fokus op die ontvanger is, word daar na ’n appellatiewe funksie verwys.

Daar word verwys na hierdie drie klasse as abstrakte klasse, omdat die oorsprong van die klasse nie op konkrete entiteite (soos hierbo in Tabel 4 genoem) gebaseer is nie, maar op ’n abstrakte eienskap (funksie) van ’n teks gebaseer is [55]. Eksperimente word ook voorgehou waarin die effek van domeinoordrag op hierdie genreklasse getoets word. Volgens die resultate, blyk die oordragpotensiaal tussen domeine goed te wees.

Die meer konkrete tipe klasse, soos hierbo in Tabel 4 genoem, kan gepas word na die funksie van die klasse en kan die klassifikasie van genre op hierdie meer abstrakte vlak gedoen word. Wachsmuth en Bujna [55] stel as voorbeeld ’n resensie wat na die ekspressiewe klas gepas kan word, ’n voorlegging na appellatief en ’n verslag na informatief. Gegewe die bostaande argument oor die wye reeks genreklasse wat tans in die literatuur te sien is en die problematiek daarrondom, word daar dus nou gepoog om

die genreklasse uit Tabel 4 te pas na die meer abstrakte vlak soos in [55] genoem. Hierdie passing kan egter nie sonder meer geskied nie. Om die eenvormigheid van die klassamestellings te verseker, moet die afrigtingsgevalle aan die hand van een of ander kriteria gemeet kan word. Die kriteria vir die klasse appellatief, ekspressief en informatief word vervolgens geïdentifiseer.

Hierdie klassebeskrywings word gedeeltelik gebaseer op die model vir die funksie van kommunikasie van Roman Jakobson, soos uiteengesit in [52]. Jakobson stel dat alle wyses van kommunikasie, verbaal of andersins, uit ses elemente bestaan. Hierdie ses elemente word in Figuur 7 voorgestel.



Figuur 7. Jakobson se kommunikasiemodel [52]

Jakobson se kommunikasiemodel is gedeeltelik gebaseer op die “Organon”-model van Bühler [4]. Jakobson brei egter die “Organon”-model uit deur die objekte of stand van sake (die informatiewe aspek) van die kommunikasie te vervaag na ’n abstrakte konteks, eerder as konkrete objekte of eienskappe. Rafferty en Salffner [41] definieer hierdie konteks soos volg:

“In order to have an effect, the message requires a context to which it refers (a reference in a different, rather vague terminology). This context has to be comprehensible for the recipient and it has to be a verbal context or must be expressible verbally.”

’n Vereenvoudigde verduideliking vir die idee agter hierdie model, word soos volg deur Van Zyl [52] uiteengesit: ’n Boodskap word gestuur deur ’n sender na ’n ontvanger (byvoorbeeld wanneer twee persone gesels): Persoon 1 kommunikeer met persoon 2.

Vir hierdie kommunikasie om te slaag, moet die sender (persoon 1) en die ontvanger (persoon 2) ŉ gemeenskaplike kode gebruik (byvoorbeeld dieselfde taal), sowel as dieselfde konteks vir die ontsyfering van hierdie kode vir die boodskap. Laastens word daar vereis dat die sender en die ontvanger kontak moet hê deur middel van ŉ fisiese kanaal (hetsy ŉ gesproke of geskrewe modus), wat die sender en die ontvanger toelaat om in verbinding te tree en te kommunikeer.

Hierdie konteks skep ŉ veronderstelde werklikheid. Die sender en die ontvanger het beide ŉ verwagting van wat hierdie werklikheid behels, maar slegs as beide die werklikheid van die sender en die ontvanger dieselfde werklikheid weerspieël, kan die kommunikasie tussen hulle geskied (anders gestel, die konteks waarin die kommunikasie geskied, moet vir beide partye dieselfde wees).

As dan verwys word na die genre van ŉ teks kan ons hier die parallelle trek tussen die verwagte/veronderstelde werklikheid en die funksie van die genre. ŉ Stel riglyne vir die interpretasie van ŉ genre se funksie kan vasgestel word; die denkbeeldige afwyking (al dan nie) van die verwagte werklikheid en die uitwerking daarvan kan gebruik word om te bepaal hoe ŉ konkrete genreklas na ŉ abstrakte klas gepas kan word.

ŉ Teks wat tot die ekspressiewe genreklas (EXP) gepas kan word, skep die verwagting dat die teks ŉ ware en neutrale voorstelling van die werklikheid bied. Hierdie voorstelling van die werklikheid word nie deur persoonlike voorkeur van die sender beïnvloed nie. Sou die ekspressiewe voorstelling egter afwyk van die veronderstelde werklikheid, word die voorstelling steeds aanvaar omdat daar geen nadelige uitwerking op die interpretasie van die ontvanger oor die kommunikasie sal wees nie. Die volgende konkrete klasse kan as ekspressief geklassifiseer word: glansartikel; nuus; nie-fiksie; en biografie.

ŉ Teks wat tot die appellatiewe genreklas (APP) gepas kan word, skep die verwagting dat die voorstelling van die werklikheid beïnvloed word deur die sender se sieninge en is as't ware juis ŉ weerspieëling van die sender se weergawe van die werklikheid. Die verwagting is dat die werklikheid verdraai word om die sender se denke te pas. Die ontvanger verwag dat die kommunikasie die werklikheid anders weerspieël as wat hy/sy die werklikheid ervaar. Die volgende konkrete klasse kan as appellatief geklassifiseer word: advertensie; private teks; en bespreking.

ŉ Teks wat tot die informatiewe (INF) genreklas gepas kan word, skep die verwagting dat dit altyd ŉ absoluut neutrale weerspieëling is van die werklikheid. Daar is geen

ruimte vir interpretasie/verwagting of bedoeling/opset wat afwyk van die werklikheid nie. ŉ Neutrale teks wat afwyk van die voorstelling van die werklikheid sal tot gevolg hê dat die kontak tussen die sender en die ontvanger verbreek word en die kommunikasie summier ongeldig word.

Die twee benaderings tot genreklasse uit afdeling 3.2.1.1 en afdeling 3.2.1.2 se implementering word in die volgende afdeling bespreek aan die hand van die beskikbare databronne sowel as die voor- en nadele wat die twee benaderings in ŉ hulpbronskaarsomgewing inhou.

3.2.2. Data

Tabel 4 wys die uiteensetting van beide die konkrete en die abstrakte genreklasse. Die konkrete genreklasse word geneem vanuit die voorgestelde genre-annotasie van die PAROLE-korpus [21]. PAROLE is ŉ Nederlandse korpus wat deur die Instituut vir Nederlandse Leksikologie[6] saamgestel is. Deur die genreklasse vir hierdie navorsing te modelleer op bestaande, gevestigde korpusse, word die platform daargestel vir wisselwerking tussen die korpusse, sodat eksperimente wat op beide korpusse uitgevoer word, vergelykbaar is. Dit verseker ook dat die genreklasse aan internasionale standaarde voldoen.

Gebaseer op resultate van Mcullum en Nigam [30] word gepoog om tussen 30 000 tot 60 000 woorde per klas te versamel; na gelang van die resultate sal die hoeveelhede data egter aangepas word om beter resultate te lewer. Die afrigtingsdata is afkomstig van data wat saamgestel is vir die ontwikkeling van tekshulpbronne vir die Suid-Afrikaanse tale vir die Nasionale Sentrum vir Mensetaaltegnologie (NCHLT). Dit bestaan grotendeels uit data wat beskikbaar is in die openbare domein, met spesifieke verwysing na die webblaaie van die Suid-Afrikaanse regering, asook data van medewerkers wat dit beskikbaar stel. Volgens Vargas Sierra [54] speel die outeur van ŉ teks ŉ belangrike rol by die samestelling van spesialiskorpusse (soos byvoorbeeld teksklassifikasie) omdat die outeurs, outeurspesifieke invloede op die dimensies van die teks het en dat dit ŉ faktor is wat ingedagte gehou behoort te word. Om die verteenwoordigendheid van ŉ korpus te bevorder, behoort die data van meerdere outeurs gebruik te word. Omdat die data egter grotendeels uit ongestruktureerde

[6] Sedert die verwysing na PAROLE se webblad is die PAROLE-korpus by die TST-Centrale se databasis ingesluit, en die oorspronklike verwysing na die genreklasse is verwyder van die webblad. Hierdie oorspronklike webblad is bekom deur ŉ versoek te rig aan die INL en word as bylaag (sien Bylaag A) ingesluit.

bronne soos die internet kom, is dit nie noodwendig moontlik om die outeurs van tekste vas te stel nie en word daar spesifieke aandag aan die outeurs gegee by die korpussamestelling nie. Die data word outomaties omgeskakel na ŉ platteksformaat en daarna handmatig geklassifiseer volgens genre deur ŉ meestersgraadstudent van die vakgebied Rekenaarlinguistiek. ŉ Annotasie-omgewing is geskep om annotasie te vergemaklik, en kwaliteitskontrole (deur lukrake tekste ŉ tweede maal te analiseer en verskille in die klastoekennings te identifiseer) word toegepas om te verseker dat die annotasies korrek is.

Die uiteindelike datahoeveelhede vir die afrigtingsdata word in Tabel 6 weergegee. Die hoeveelhede data vir elke klas is inkonsekwent. Die verskil in die syfers per klas is te wyte aan die beskikbaarheid uit die oorspronklike databronne. Daar is aanvanklik besluit om die verskillende klasse se afrigtingsdatasyfers gelyk te hou, sodat die een klas nie bo die ander bevoordeel word nie, maar na ŉ vinnige vergelyking tussen die prestasie van ŉ sisteem met gelyke hoeveelhede data en ŉ sisteem wat al die beskikbare data gebruik, sodat sommige klasse meer en ander minder data het, word daar opgemerk dat dit beter is om alle beskikbare data te gebruik. Die sisteem wat met al die beskikbare data afgerig is, het oor die algemeen ŉ verbetering van 3% in prestasies getoon, bo die sisteem waarvan die klasse gebalanseer is. Dit sou egter die teenoorgestelde effek gehad het as die klasse waarby data ontbreek, aangevul kon word om dieselfde hoeveelheid data as die grootste klasse te hê. He en Garcia [19] stel dat ŉ datastel eers regtig skeefgetrek is, of dat ŉ klas eers onderverteenwoordig is, as dit ŉ 1:100-verhouding of groter tot die ander klasse aanneem.

Die volledige afrigtingstel het ŉ uiteindelike totaal van 2 173 afrigtingsgevalle/tekste oor al die klasse. Hierdie totaal word dan outomaties verdeel in ŉ 90%:10%-verhouding van afrigtings- en toetsdata vir die gebruik in 10-voudige kruisvalidasie in die hierop-volgende eksperimente.

Klas	Afkorting	Beskrywing
Inligting	INF	Inligtingspamflette oor enige onderwerp
Nuus	NEW	Verslaggewing in koerante en ander nuusbronne
Nie-fiksie	NON	Studiegidse, akademiese skrywes
Biografie	BIO	Biografieë, outobiografieë en geskiedkundige skrywes
Private teks	PRI	Private korrespondensie, soos briewe en dagboekinskrywings
Bespreking	DIS	Gesprekke, debatte en toesprake van enige aard
Instruksie	INS	Handleidings
Amptelike teks	OFF	Wetgewing, beleide
Glansartikel	FEA	Langer inligtingstukke, soos in tydskrifte
Advertensie	ADV	Spesiale aanbiedinge en koerantadvertensies
Poësie	POE	Gedigte en lirieke
Drama	DRA	Dramas
Fiksie	FIC	Roman

Tabel 4. Klasse vir die genreklassifikasiesisteem

Die twee uiterste gevalle van die datasyfers is die POE- en FEA-klasse. Die POE-klas (“poetry”) se samestelling is van so ’n aard dat daar baie, maar kort afrigtingsgevalle is, omdat die gedigte wat ingesluit is by die afrigtingsdata oor die algemeen nie meer as ’n paar versreëls bevat nie. Ander klasse het weer minder, maar langer afrigtingsgevalle, byvoorbeeld wette, waar ’n enkele wet etlike bladsye kan beslaan. Die FEA-klas (“features”) bevat ooreenkomste met die INF-klas (“information”), omdat FEA ook ’n inligtingstuk is, maar in essensie langer en meer gedetailleerd is, soos byvoorbeeld in tydskrifartikels. Die meeste van die data wat beskikbaar was, is as INF geklassifiseer omdat dit meestal strooibiljette of soortgelyke data was.

Gebaseer op die veronderstellings vir die abstrakte genreklasse uit afdeling 3.2.1.2, kan die konkrete genreklasse dan soos in Tabel 5 ingedeel word, sodat klassifikasie op hierdie abstrakte vlak kan plaasvind. Uit die dokumentasie vir die LASSY-korpus [51] (sien afdeling 4.2.1) kon daar nie enige fiksietekste (d.i. die poësie, drama en fiksieklasse uit Tabel 4) geïdentifiseer word nie. By die samestelling van die

afrigtingsdata vir die abstrakte genreklasse word hierdie klasse dus uitgelaat om te verseker dat die eksperimente in afdeling 4.1 versoenbaar is. Die klasetikette word dus voorafgegaan met “NF-” om aan te dui dat die klasse slegs nie-fiksie bevat. Die resulterende klasetikette is dus NF-EXP; NF-APP; en NF-INF. Tabel 5 wys die passing van die konkrete klasse op die abstrakte klasse. Die weglating van die bogenoemde drie klasse het egter tot gevolg dat daar heelwat afrigtingsgevalle verlore gaan.

Tabel 6 toon die afrigtingsdatasyfers nadat die passing van die oorspronklike dertien klasse op die drie abstrakte klasse gedoen is. Die oorspronklike dertien klasse word weer weergegee ter wille van vergelyking.

Die volgende afdelings beskryf eksperimente wat op beide die abstrakte en die konkrete genreklasse uitgevoer is om die effektiwiteit van beide benaderings tot genreklasse vas te stel.

Abstrakte superklas	**Konkrete subklasse**
NF-EXP	FEA
	NEW
	NON
	BIO
NF-APP	ADV
	PRI
	DIS
NF-INF	INS
	OFF
	INF

Tabel 5. Uiteensetting van konkrete en abstrakte genreklasse

Konkrete genreklasse	**# Tekste**
ADV	343
BIO	96
DIS	148
DRA	61
FEA	52
FIC	87
INF	334
INS	92
NEW	85
NON	58
OFF	109
POE	528
PRI	180
Totaal	**2 173**
Abstrakte genreklasse	**# Tekste**
NF-EXP	229
NF-APP	439
NF-INF	536
Totaal	**1204**

Tabel 6. Vergelyking van konkrete en abstrakte genreklasse

3.2.3. Algoritme

In hierdie afdeling word die vyf algoritmes wat vroeër in afdeling 2.2 bespreek is (*k*-NN, SVM, MNB, besluitnemingsbome en RIPPER) vergelyk deur die eienskappe en die datagroottes konstant te hou. Elkeen van die algoritmes word afgerig met hul verstekinstellings, telkens met dieselfde data en eienskappe. Hier word die twee stelle klasse, soos bespreek in afdeling 3.2.1, met elkeen van die vyf algoritmes geëvalueer in ŉ totaal van tien eksperimente. ŉ Woordversamelingbenadering word gevolg met betrekking tot die eienskappe vir hierdie eksperiment. Woordversameling is die eenvoudigste stel eienskappe wat gebruik kan word en is inherent ŉ woord-vir-woord-voorstelling van die data. Die twee algoritmes en een stel klasse wat die beste resultate lewer, word geïdentifiseer en word dan in verdere eksperimente gebruik.

ŉ 10-Voudige kruisvalidasiemetode word in die vergelykende eksperimente gebruik. Dit behels dat die beskikbare data lukraak verdeel word in 90% afrigtingsdata en 10% toetsdata. Hierdie verdeling word dan tien maal herhaal telkens met ŉ ander lukrake 90%/10% verdeling van die data. Deur hierdie metode te gebruik word die algoritme se

vermoë om te veralgemeen ook gemeet. Die WEKA implementering van 10-voudige kruisvalidasie word gebruik.

Om die sukses van die resultate van die algoritmes se prestasie te peil, moet die resultate gemeet kan word teen 'n basislynsisteem. Petrenz [36] stel die gebruik van twee basislynsisteme voor vir die evaluasie van kruistalige genreklassifikasie (sien afdeling 4.1). Die eerste is die toekenning van 'n lukrake klas tydens klassifikasie. Daar bestaan dus, in die geval van die dertien klasse eksperimente, 'n een uit dertien kans dat die korrekte klas toegeken word en kan die basislyn bereken word as 'n presisie van 0.077. Hierdie basislyn neem egter nie die klasverspreidings in ag nie. Die klasse van die sisteem is nie gebalanseer nie en daarom is daar nie presies 'n een uit dertien kans dat die regte klas lukraak toegeken gaan word nie. As die waarskynlikhede van elke klas bygereken word, verhoog die bogenoemde basislyn na 0.148. Vir die eksperimente met drie klasse kan hierdie lukrakeklasbasislyn (met die klasverspreidings in ag geneem) gestel word op 'n presisie van 0.367. Hierdie twee basislynsisteme is egter baie laag, veral met verwysing na die eksperimente met dertien klasse.

Petrenz [36] stel 'n tweede basislyn voor, naamlik die toekenning van die mees frekwente klas. Hierdie basislynsisteem kan afgelei word deur die hoeveelheid voorkomste van die mees frekwente klas in die afrigtingsdata te deel deur die hoeveelheid voorkomste in die totale afrigtingstel. Vir die eksperimente met dertien klasse kan hierdie mees frekwente klas basislyn gestel word op 'n presisie van 0.243, deur telkens die mees frekwente klas (POE) toe te ken (528/2172). Die mees frekwente klas basislyn vir die eksperimente met drie klasse kan gestel word op 'n presisie van 0.444, deur telkens die mees frekwente klas (NF-NEU) toe te ken (534/1202). Hierdie lae basislynsisteme word aangeneem omdat die resultate van eksperimente uit die literatuur moeilik vergelykbaar is (sien afdeling 2.6). 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. Deur hier reeds van die algoritmes te elimineer, word die hoeveelheid eksperimente tot 'n minimum beperk (vgl. afdeling 3.1). Die resultate word in Tabel 7 en Tabel 8 voorgehou.

Algoritme	Woordversameling: 3 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

Tabel 7. Resultate: Algoritmes, woordversameling (3 Klasse)

Algoritme	Woordversameling: 13 Klasse		
	Presisie	**Herroeping**	***f*-Telling**
***k*-NN**	0.635	0.590	0.579
SVM	**0.844**	**0.849**	**0.845**
MNB	**0.859**	**0.866**	**0.856**
Besluitnemingsbome	0.770	0.780	0.773
RIPPER	0.748	0.769	0.750

Tabel 8. Resultate: Algoritmes, woordversameling (13 klasse)

Die resultate uit Tabel 7 en Tabel 8 toon dat MNB en SVM telkens die beste resultate lewer vir beide drie, sowel as dertien klasse. Al die bogenoemde resultate klop die basislynsisteme, hoewel die basislynsisteme laag gestel is. Uit die literatuur is dit te verwagte dat hierdie twee algoritmes beter resultate sal lewer as die ander algoritmes. Soos reeds genoem in afdeling 2.2.3, het ’n MNB-klassifiseerder relatief min data nodig om kompeterende resultate te lewer, daarom sal dit verwag word dat MNB geskik sal wees vir hierdie toepassing en in die algemeen beter sal vaar as die ander algoritmes. In ’n vergelykende studie vir algoritmes vir teksklassifikasie, bevind Khan *et al*. [24] ook dat SVM en MNB goeie keuses sal wees vir teksklassifikasie, aangesien MNB baie goed werk vir beide numeriese en tekstuele data en maklik implementeerbaar is in vergelyking met ander algoritmes, hoewel die onafhanklikheidsaanname (sien afdeling 2.2.3) die algoritme negatief beïnvloed as daar ’n hoë korrelasie tussen die eienskappe is.

Verder stel Khan *et al*. [24] dat SVM aanvaar kan word as een van die effektiefste algoritmes vir teksklassifikasie. SVM kan die onderliggende karakteristieke (eienskappe) beter vasvang as meeste ander algoritmes danksy die implementering van strukturele risikominimering (sien afdeling 2.2.2) wat die veralgemening van die algoritme bevorder. SVM bied wel uitdagings by die optimering van die algoritme-parameters [24].

Ten einde 'n volledig ingeligte besluit te neem oor die algoritmes 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 as gevolg van kans is. Hierdie waarskynlikheid word getoets teen die sogenaamde nulhipotese (H_0) [48][32]. In die geval van statistiese beduidendheid by die verskil tussen masjienleeralgoritmes se resultate, word H_0 uitgedruk as dat daar geen beduidende verskil tussen die resultate van algoritme A en algoritme B is nie [32]. 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 nulhipotese waar is). 'n p-Waarde wat kleiner as 0.05 is, word aanvaar as statisties beduidend en die nulhipotese word verwerp (met ander woorde die verskil in resultate is nie blote toeval nie). Uit die literatuur [48][32][59] volg die argument dat die beste toets vir statistiese beduidendheid by masjienleeralgoritmes die "Approximate Randomisation Testing"-metode is. Hierdie toets word outomaties gedoen deur gebruik te maak van die vrylik beskikbare sagteware: art.py[7] wat met verstekinstellings gebruik word. Hierdie 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 algoritmepare word in Tabel 9 en Tabel 10 voorgehou. Die waarde in vetdruk stel telkens 'n statisties beduidende waarde voor ($p > 0.05$). Vir die eksperimente met drie klasse toon slegs MNB en SVM 'n beduidende verskil in resultate in vergelyking met die ander algoritmes. Tussen MNB en SVM is daar egter nie 'n beduidende verskil nie. Vir die eksperimente met dertien klasse word gemerk dat daar 'n statisties beduidende verskil tussen die resultate van *k*-NN en al die ander algoritmes is (met die uitsondering van RIPPER). MNB teenoor RIPPER is die enigste ander algoritmekombinasiepaar waar 'n statisties beduidende verskil opgemerk word.

[7]Beskikbaar by http://www.clips.ua.ac.be/scripts/art

3 Klasse	k-NN	SVM	MNB	Besluitnemings-bome	RIPPER
k-NN					
SVM	**0.00500**				
MNB	**0.00030**	0.21875			
Besluitnemings-bome	0.30817	0.20178	**0.01200**		
RIPPER	0.29327	0.13989	**0.00540**	1.00000	

Tabel 9. p-Waardes vir algoritme vergelyking met ART (3 klasse)

13 Klasse	k-NN	SVM	MNB	Besluitnemings-bome	RIPPER
k-NN					
SVM	**0.00009**				
MNB	**0.00009**	0.41446			
Besluitnemings-bome	**0.00009**	0.34417	0.05969		
RIPPER	0.29267	0.05110	**0.00210**	0.29267	

Tabel 10. p-Waardes vir algoritme vergelyking met ART (13 klasse)

Om te bepaal watter hoeveelheid klasse uiteindelik beter is, word daar vervolgens gekyk of daar ŉ statisties beduidende verskil tussen die resultate van die benaderings bestaan. Om die statistiese beduidendheid hiervan te bepaal, word die gratis weergawe van "Statistics Calculator" van StatPac [8] gebruik. Die drie-klas- en dertien-klas-eksperiment word hier as twee afsonderlike steekproewe vir Student se *t*-toets [48] hanteer ten einde ŉ aanduiding van die statisties beduidendheid van die verskil in resultate te verkry.

k-NN lewer statisties beter resultate vir drie klasse, teenoor dertien klasse. Die *f*-telling is 0.277 hoër, wat statisties beduidend is ($p = 0.00001 < 0.05$). Die resultate vir die eksperimente met drie klasse is met ongeveer 0.056 (*f*-telling) hoër as die eksperimente met dertien klasse vir MNB ($p = 0.04750 < 0.05$), wat weereens ŉ statisties beduidende verskil is. Die resultate vir SVM is met 0.073 (*f*-telling) hoër vir die drie-klas eksperiment as vir die dertien-klaseksperiment. Hierdie verskil is egter nie statisties beduidend nie omdat $p = 0.15070 > 0.05$. Vir besluitnemingsbome word daar weer ŉ statisties

[8]Beskikbaar by http://www.statpac.com/statistics-calculator/free-version.htm

beduidende toename in *f*-telling (0.104) waargeneem ($p = 0.02040 < 0.05$) en die teenoorgestelde geld weer vir die 0.120 verhoging in *f*-telling vir die RIPPER-algoritme ($p = 0.07910 > 0.05$; daarom nie statisties beduidend). Drie uit die vyf algoritmes toon dus ŉ statisties beduidende verskil in die verhoging van die *f*-telling wanneer drie klasse teenoor dertien gebruik word.

Die verskil in gemiddelde *f*-telling vir al die algoritmes vir die drie-klas- (0.887) en die dertien-klas- (0.761) eksperimente is wel statisties beduidend ($p = 0.00540 < 0.05$). Hieruit en na aanleiding van die feit dat MNB en SVM oor die algemeen statisties beter resultate lewer as die ander algoritmes op dieselfde hoeveelheid klasse, kan dus afgelei word dat die drie-klasbenadering met MNB en SVM die beste benadering sal wees om te volg. 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 gebruik.

3.2.4. Eienskappe

In hierdie gedeelte word die verskillende eienskappe wat algemeen in die literatuur gebruik word vir genreklassifikasie mee geëksperimenteer en geëvalueer. Hierdie eienskappe is woordversameling, *tf-idf*-tellings, karakter- en woord-n-gramme en woordsoortinligting. Die onttrekking van hierdie eienskappe uit die afrigtings- en toetsdata word ook kortliks beskryf.

3.2.4.1. Eienskaponttrekking

Stap 1: Omskakeling na die ARFF-formaat

WEKA vereis dat alle toevoerdata, sonder uitsondering, moet voldoen aan ŉ formele struktuur wat bekend staan as die “Attribute-Relation File Format” (ARFF-formaat) [18] (sien Figuur 8 as voorbeeld). Die afrigtingsdata word na hierdie formaat vanuit die platteks omgeskakel deur teksmanipulering met ŉ Perl-skrip. Die ARFF-lêer bestaan uit ŉ reeks opskrifte wat die datatipes van die invoerdata, sowel as die verskillende klasse definieer. WEKA lei dan hieruit af hoe die verskillende velde van elke afrigtingsgeval in die lêer geïnterpreteer moet word. Hierdie opskrifte word telkens met ŉ “@” teken aangedui. Direk na die “@Data”-opskrif, word die afrigtings-/toetstekste (die proses word apart vir dié twee tipes tekste gedoen) gelys as een teks per lyn, waarvan slegs

die woorde van die teks (sonder enige leestekens, en so meer) weergegee word, gevolg deur die voorafbekende klas van elke teks). Al die tekste in die afrigting- en toetsstelle word gelyktydig na hierdie formaat verwerk. Nadat die tekste na die ARFF-formaat omgeskakel is, word hierdie lêer gebruik vir eienskapsonttrekking met die voorverwerkingsalgoritmes van WEKA.

```
@RELATION DocClass
@ATTRIBUTE words string
@ATTRIBUTE class{PRI,POE,ADV,BIO,DIS,DRA,FEA,FIC,INF,INS,NEW,NON,OFF}
@DATA
"....",ADV
"....",PRI
"....",POE
"....",DRA
.
.
.
```

Figuur 8. Die ARFF-formaat

Stap 2: Toepassing van StringToWordVectorFilter

Die masjienleeralgoritmes in WEKA kan nie eienskappe in 'n platteksformaat gebruik tydens die afrigtingsfase nie. Die eienskappe moet eers na 'n numeriese voorstelling omgeskakel word voordat dit bruikbaar is. Hierdie omskakeling na 'n numeriese voorstelling lewer eienskapvektore wat deur die StringToWordVectorFilter van WEKA vir elke teks in die datastel saamgestel word (sien Figuur 9).

Die woorde van al die betrokke tekste word gebruik om 'n woordelys te maak waarvan slegs unieke voorkomste van elke woord gebruik word. Die woordelys word dan gebruik om 'n nuwe opskrifstruktuur in die ARFF-lêer te genereer, waarvan elke woord 'n eienskap met 'n numeriese waarde definieer. Hierdie woorde in die opskrifstruktuur word as 'n skikkingstruktuur gestoor en kan daar dan daarna terugverwys word in die @Data-gedeelte. Die woorde van elke teks word vervang deur 'n groepie van twee numeriese waardes wat telkens met 'n komma geskei word. Die eerste van die twee waardes verwys na die *n*-de woord (aangedui deur die syferwaarde) in die opskrifte (of die *n*-de indeks van die skikking) van die ARFF-struktuur en die tweede waarde is 'n gewig wat aan daardie woord toegeken word deur die StringToWordVectorFilter. Hierdie gewig word gebaseer op die instellings van die filter wat gemanipuleer kan word om 'n hele reeks eienskappe van die woord in ag te neem en as 'n numeriese waarde te enkodeer. Byvoorbeeld, vir binêre vektore (woordversameling) word slegs die aanwesigheid van die woord in ag geneem en word daar slegs 'n 1 toegeken vir 'n spesifieke indeks van die skikking as die woord in die afrigtingsgeval voorgekom het. Vir *tf-idf*-tellings word die

verhouding van die woord se aantal voorkomste in ’n spesifieke teks se verhouding met die woord se voorkomste in al die ander tekste in die vektor as die gewig geënkodeer (sien afdeling 2.3.2.). Ander eienskappe, soos teksstatistiek, kan ook in hierdie tipe struktuur voorgestel word deur numeriese waardes (byvoorbeeld gemiddelde vir die teksstatistiek) eksplisiet aan elkeen van die teksstatistiekeienskappe toe te ken en dit so in die ARFF-formaat te enkodeer.

```
@RELATION  DocClass
@ATTRIBUTE class{PRI,POE,ADV,BIO,DIS,DRA,FEA,FIC,INF,INS,NEW,NON,OFF}
@ATTRIBUTE appel numeric
@ATTRIBUTE ander numeric
@ATTRIBUTE amper numeric
@ATTRIBUTE beste numeric
      .
      .
      .
@DATA
{0 ADV,2 0.184600 ,39 0.444778 ,43  0.889556...}
{0 PRI,2 0.331844 ,21 1.599103 ,112 1.599103...}
{0 POE,2 0.090516 ,4  0.436181 ,5   0.436184...}
{0 DRA,2 0.444778 ,73 0.889556 ,74  0.889556...}
      .
      .
      .
```

Figuur 9. ARFF-formaat van die woordvektore

3.2.4.2. Eksperimentele vergelyking van eienskappe

Die verskillende eienskappe, soos in afdeling 2.3 uiteengesit (te wete woordversameling, *tf-idf*-tellings, karakter- en woordtrigramme), word nou om die beurt gebruik as afrigtingsdata vir die twee algoritmes hierbo in afdeling 3.2.2 geïdentifiseer (SVM en MNB) 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 uitwerking van die eienskappe op die resultate van die algoritme vasgestel word. Die resultate vir hierdie reeks eksperimente word in Tabel 11 voorgehou.

Algoritme (3 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**

Tabel 11. Resultate vir SVM en MNB met verskillende eienskappe

Die resultate toon dat ŉ woordversamelingbenadering tot eienskaponttrekking die beste resultate lewer. Met verskille so klein as 0.005 vir die *f*-telling (0.929 vir woordversameling, teenoor 0.924 vir *tf-idf*) (vgl. Tabel 11) is die verskil in die resultate vir die verskillende eienskappe en die twee algoritmes egter nie statisties beduidend verskillend nie (telkens $p > 0.05$; sien Tabel 12). Slegs by die woordtrigrambenadering word daar ŉ statisties beduidende verskil ($p < 0.05$) in die resultate tussen dié benadering en die ander benaderings opgemerk vir beide algoritmes (die woordtrigram-benadering 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 is, soos genoem in afdeling 3.2.2, dat daar nie ŉ 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 ŉ onbegonne taak te wees om hier uit die ander benaderings ŉ beste benadering te identifiseer. ŉ Woordversamelingbenadering (soos genoem in afdeling 2.3.1) is die eenvoudigste voorstelling van eienskappe en maak nie staat op enige bestaande hulpbronne nie, wat dit besonder geskik maak vir ŉ hulpbronskaars omgewing. Verder het die woordversamelingbenadering beter resultate as die ander algoritmes gelewer (hoewel dit nie statisties beduidend beter is nie) en word daarom in die hieropvolgende datahoeveelheideksperimente gebruik.

Algoritme (3 Klasse)	MNB: Woordver-sameling	MNB: *tf-idf*	MNB: Karakter-trigramme	MNB: Woord-trigtramme	MNB: Kombinasie	SVM: Woordver-sameling	SVM: *tf-idf*	SVM: Karakter-trigramme	SVM: Woord-trigtramme	SVM: Kombinasie
SVM: Woord-versameling	0.12500	0.28906	1.00000	**0.00040**	0.75391		1.00000	0.25000	0.06059	0.50781
SVM: *tf-idf*	0.21875	0.45312	1.00000	**0.00020**	1.00000	1.00000		0.50000	**0.02280**	0.72656
SVM: Karakter-trigramme	0.62500	1.00000	0.50781	**0.00009**	1.00000	0.25000	0.50000		**0.00750**	1.00000
SVM: Woord-trigtramme	**0.00100**	**0.00360**	0.07719	0.13289	**0.02950**	0.06009	0.02370	**0.00750**		**0.00690**
SVM: Kombinasie	0.68750	1.00000	0.50781	**0.00009**	1.00000	0.50781	0.72656	1.00000	**0.00690**	
MNB: Woord-versameling		1.00000	0.06250	**0.00009**	0.45312	0.12500	0.21875	0.62500	**0.00100**	0.68750
MNB: *tf-idf*	1.00000		0.12500	**0.00009**	0.68750	0.28906	0.45312	1.00000	**0.00360**	1.00000
MNB: Karakter-trigramme	0.06250	0.12500		**0.00090**	0.68750	1.00000	1.00000	0.50781	0.07719	0.50781
MNB: Woord-trigtramme	**0.00009**	**0.00009**	**0.00090**		**0.00020**	**0.00040**	**0.00020**	**0.00009**	0.13289	**0.00009**
MNB: Kombinasie	0.45312	0.68750	0.68750	0.00009		0.75391	1.00000	1.00000	0.02950	1.00000

Tabel 12. p-Waardes vir eienskapvergelyking met ART (3 klasse)

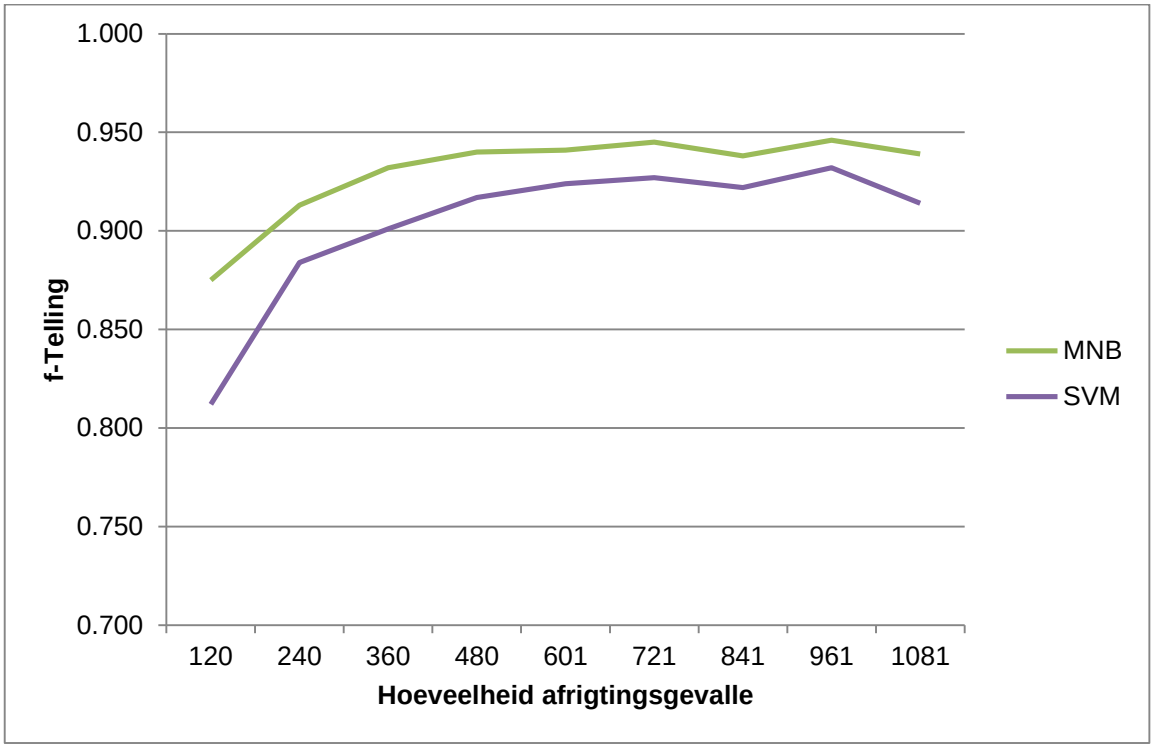
3.2.5. Data: hoeveelheid

In die vorige gedeeltes is die optimale kombinasie van hoeveelheid klasse, algoritmes, en eienskappe geïdentifiseer. Die vraag ontstaan nou: Watter hoeveelheid data moet gebruik word om die optimale prestasie vir die algoritme te bereik? Geld die algemene veronderstelling dat meer data altyd beter is, of word daar ŉ afname in die prestasie waargeneem wanneer meer data gebruik word as wat werklik benodig word? Ten einde hierdie optimale hoeveelheid van die afrigtingsdata te bepaal, word daar gekyk na die effek op die prestasie van die algoritme, namate die hoeveelheid data wat vir die algoritme beskikbaar gestel word, toeneem [17].

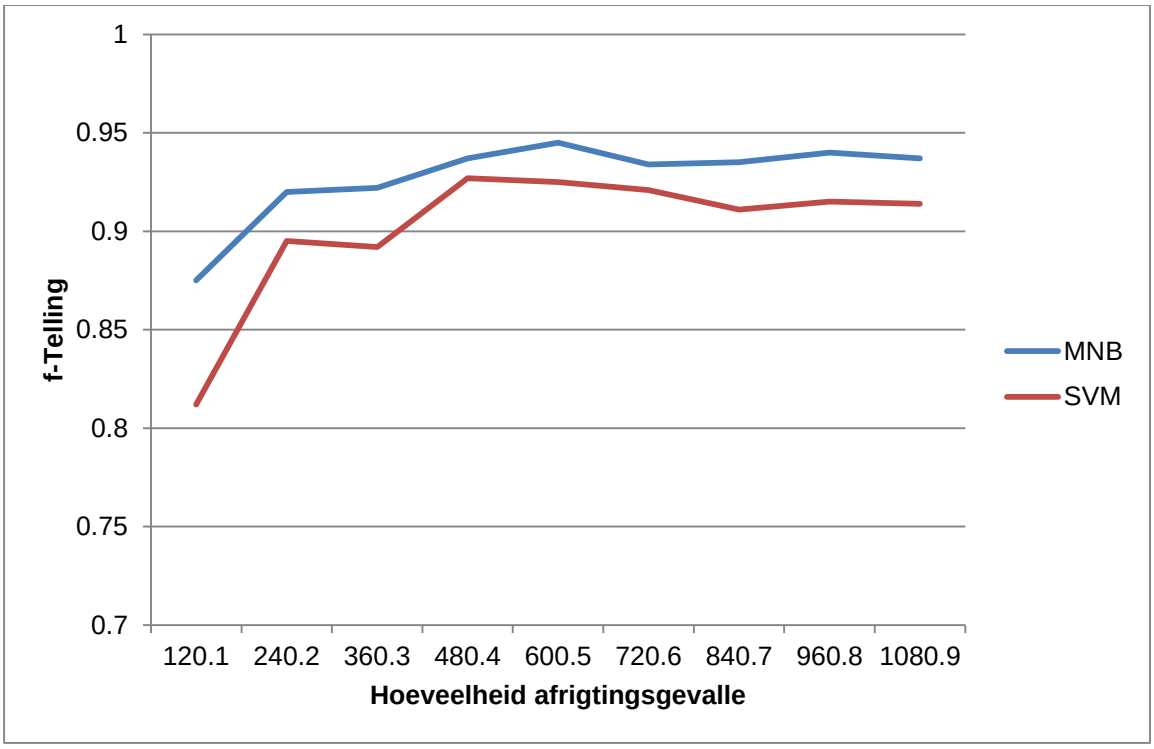
ŉ Uithoutoetstel van 10% van die afrigtingsgevalle word onttrek en bly konstant deur die loop van die datahoeveelheideksperimente Die oorblywende stel data word verdeel in kleiner gedeeltes wat elk 10% van die afrigtingsgevalle bevat. Hierdie gedeeltes word dan lukraak gekies, sonder terugplasing (d.i. die nuwe 10% wat gekies word, bou voort op die 10% inkremente wat dit voorafgegaan het deur bloot die nuwe 10% by te voeg), om met elke herhaling van die eksperiment 10% meer van die oorspronklike hoeveelheid data te gebruik vir afrigtingsdata.

Deur die resultate vir elke herhaling van hierdie eksperiment te gebruik, kan ŉ leerkurwe saamgestel word, wat ŉ aanduiding gee van die verwagte impak van die toevoeging of weglating van data wat vir die algoritme beskikbaar gestel word. Die leerkurwe word saamgestel deur ŉ reeks punte van hoeveelheid afrigtingsgevalle, teenoor ŉ metriek (byvoorbeeld *f*-telling) wat die prestasie van die algoritme voorstel te stip en die tendens van groei tussen hierdie punte te trek. Figuur 10 wys die groei in die resultate van die algoritmes, namate die beskikbare data groei. Om hierdie proses te illustreer word ŉ woordversamelingbenadering vir die eienskappe gebruik.

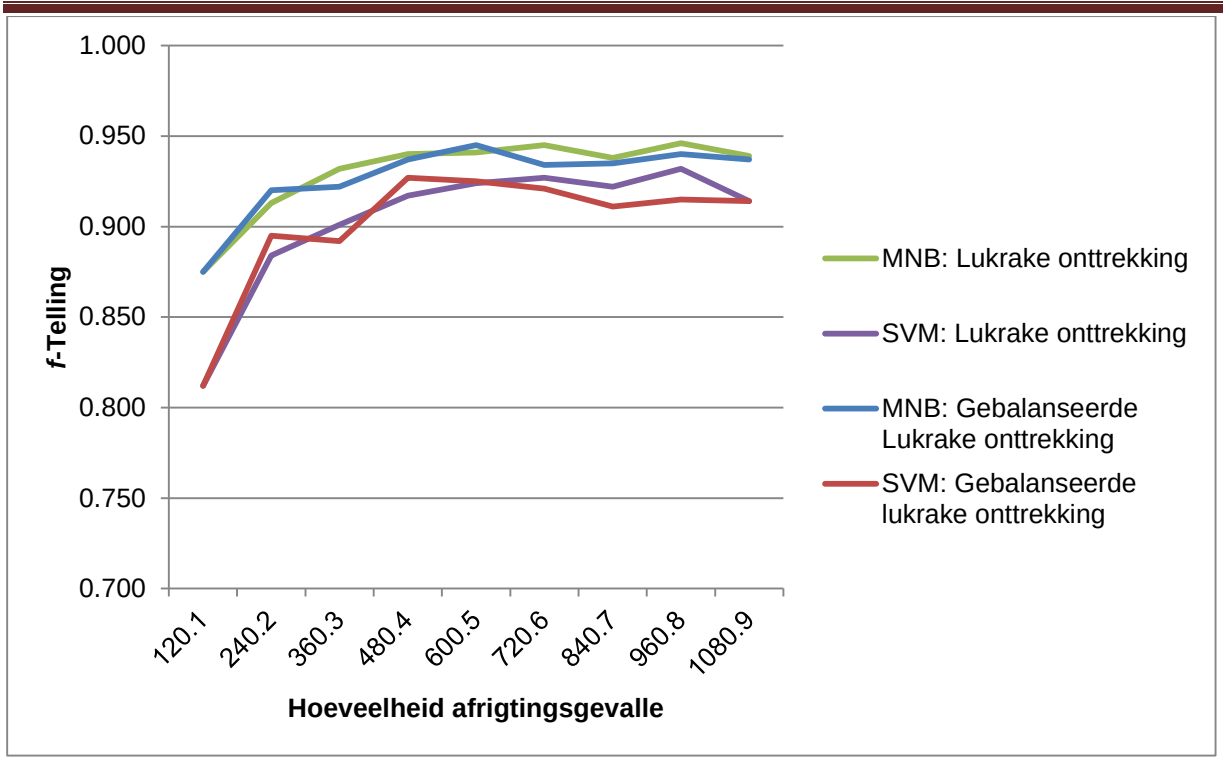
Figuur 10 toon dat daar ŉ skerp styging in die prestasie van die twee algoritmes is wanneer 10%-20% van die data aan die algoritmes beskikbaar gestel word, maar die gradiënt van die kurwe begin afneem (hoewel dit steeds positief bly) totdat daar slegs ŉ baie klein hoeveelheid groei waargeneem word namate die datahoeveelhede toeneem. Dit is die verwagte patroon van ŉ masjienleeralgoritme om by die aanvang van die leerproses vinnige groei te toon en dan gevolg te word deur ŉ gedeelte waar die gradiënt van die kurwe stadig begin afneem totdat ŉ plato bereik word, waar die toevoeging van data die prestasie tot ŉ mindere mate beïnvloed [17].



Figuur 10. Leerkurwes vir SVM en MNB algoritmes (lukrake afrigtingsgevalle uit volledige afrigtingstel)



Figuur 11. Leerkurwes vir SVM en MNB algoritmes (lukrake afrigtingsgevalle met gebalanseerde klasse uit volledige afrigtingstel)



Figuur 12. Vergelyking van lukrake afrigtingsgevalle met en sonder gebalanseerde klasse uit volledige afrigtingstel

Daar word egter gemerk dat die gradiënt van die leerkurwe aan die einde van die data-toevoeging negatief word en dat die toevoeging van nog afrigtingsgevalle die prestasie van die algoritme negatief beïnvloed. Hierdie patroon is egter teenstrydig met dit wat deur Gu *et al.* [17] beskryf word as die tipiese leerkurwe, omdat hulle geen melding maak word van negatiewe groei ná die plato bereik word nie.

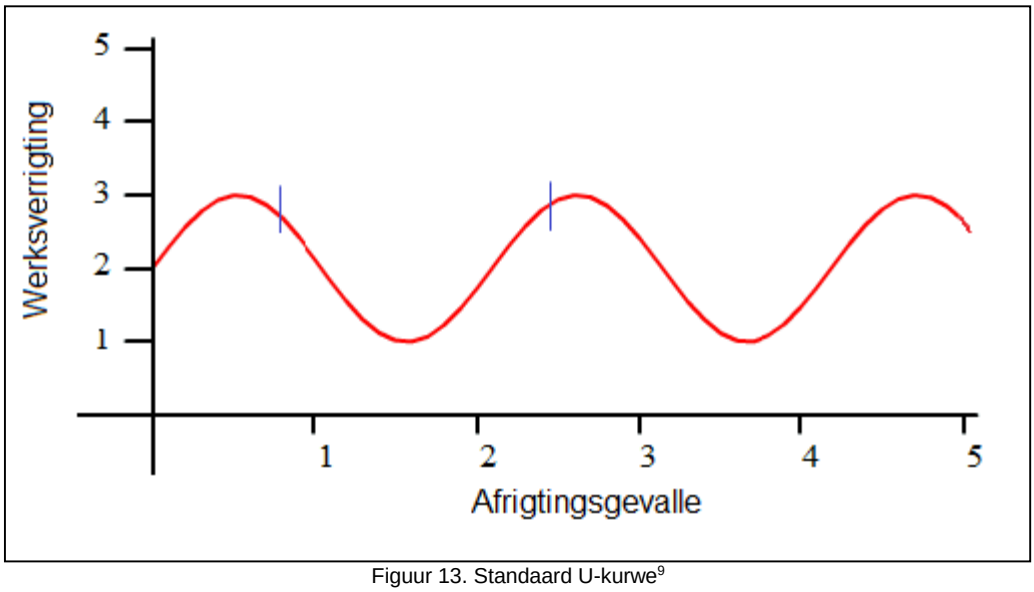
Ten einde vas te stel of die datasamestellings vir elkeen van die bogenoemde 10% inkremente moontlik die rede hiervoor kan wees, word die leerkurwes weer soos voorheen saamgestel. Vir die tweede stel leerkurwes word daar egter gelet op die balans tussen die klasse vir elkeen van die 10% inkremente. In plaas daarvan om 10% van die totale datastel telkens lukraak te onttrek en by te voeg, word daar nou telkens 10% van die totale afrigtingsgevalle op ŉ per klas basis lukraak onttrek. Deur bloot ŉ lukrake 10% van die data te onttrek, veroorsaak dit soms dat sekere klasse se verteenwoordiging vir sommige iterasies nie verteenwoordigend van die totale datastel is nie. Die tweede stel leerkurwes word in Figuur 11 voorgehou.

Figuur 11 toon leerkurwes wat min of meer dieselfde tendens volg as die leerkurwes in Figuur 10. Die afname in prestasie aan die einde van die leerkurwe word steeds opgemerk, maar die negatiewe groei is hier laer, ten spyte van die nuwe lukrake onttrekking van afrigtingsgevalle soos hierbo uiteengesit. Figuur 12 toon die leerkurwes op dieselfde grafiek gestip. Die begin- en eindresultate vir die algoritmes is dieselfde,

maar toon deurlopend verskillende resultate vir die verskillende hoeveelheid afrigtingsdata. 'n Moontlike verduideliking hiervoor is dat dit bloot te wyte is aan die standaard fluktuasies in leerkurwes, of sogenaamde U-kurwes, soos wat in [6] deur Carlucci en Case beskryf word:

"A U-shaped curve in a cognitive-developmental trajectory refers to a three-step process: good performance followed by bad performance followed by good performance once again. In learning contexts, U-shaped learning is a behaviour in which the learner first learns the correct behaviour, then abandons the correct behaviour and finally returns to the correct behaviour once again." [6]

Hierdie verskynsel van die U-kurwe uit die kognitiewe wetenskappe word in [6] getoets aan die hand van abstrakte wiskundige stellings en bewyse om die noodsaaklikheid daarvan in rekenaarmatige leerteorie vas te stel. Gegewe die relatiewe klein hoeveelheid afrigtingsdata wat beskikbaar is, is dit moontlik dat slegs die begin van die standaard U-kurwe (sien Figuur 13) deur hierdie hoeveelheid data voorgestel word en dat die patroon van groei, al is dit teen 'n lae tempo, voortgesit sal word met die toevoeging van nog afrigtingsdata. Daarom word die volledige afrigtingstel gebruik in die volgende gedeelte vir algoritme-optimering om die optimale klassifiseerder te identifiseer.



Figuur 13. Standaard U-kurwe[9]

[9] Aanpassing van http://upload.wikimedia.org/wikipedia/commons/d/dc/U-Shaped_development_graph.png

3.2.6. Optimering

3.2.6.1. Optimale versameling

Uit die bogenoemde stappe van die eliminerende eksperimente word MNB- en SVM-klassifiseerders geïdentifiseer as die beste algoritmes en drie klasse as die optimale stel. Soos genoem in afdeling 3.2.4.2, kan daar nie bewys word dat daar een stel eienskappe is wat statisties beduidend beter vaar as die ander nie. Daarom word die optimeringsfase vir elkeen van die eienskapstelle (behalwe die woordtrigrambenadering wat geëlimineer kon word) herhaal. Die volledige afrigtingsdatastel word gebruik in hierdie optimeringsfases (sien afdeling 3.2.5). Vir hierdie eienskappe word die algoritme instellings dan, om volledigheidsonthalwe, geoptimeer om te kyk of enige verbetering in die prestasie van die algoritmes te weeg 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 afdeling 3.2.4.2) dien dan as die volledig geoptimeerde resultate.

Soos reeds genoem in 2.2.2. 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 verstekwaarde van hierdie parameter is C=1.0. Om die effek van die verandering van die kompleksiteitsparameter op die prestasie te bepaal, word die waarde van C telkens geïnkrementeer vir ’n reeks van drie eksperimente. Volgens Hsu *et al*. [20] is ’n goeie reeks waardes om te ondersoek vir C ’n eksponensiële groeiende reeks waarmee die eksponent telkens met 0.25 vermeerder word en stel beide negatiewe en positiewe eksponente voor. 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. Die C-parameter bepaal die algoritme se geneigdheid tot afrigtingsfoute teenoor die model se kompleksiteit. ’n Hoër waarde vir C sal die kompleksiteit verhoog, terwyl ’n laer waarde afrigtingsakkuraatheid benadeel [8]. Hsu *et al.* [20] omskryf die afrigtingsakkuraatheid as die vermoë van die klassifiseerder om die afrigtingsdata, waarvoor 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 belangrik is vir die algoritme om die afrigtingsdata korrek te herklassifiseer.

Vir die doeleindes van hierdie optimeringseksperiment word daar van ’n uithoutoetsstel gebruik gemaak. Hierdie 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 optimerings-fase 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 13 uiteengesit.

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

Tabel 13. Algoritme-optimering (3 klasse)

Om vas te stel of hierdie verbetering in prestasie van die SVM-algoritme na optimering werklik beter resultate lewer, word daar weer gebruik gemaak van statistiese beduidendheid om die moontlikheid van kans in die verskil tussen die algoritmes vas te stel (vgl. afdeling 3.2.4.2). Tabel 14 toon die p-waardes vir die verskil tussen die prestasie van die geoptimeerde SVM-algoritmes en die verstek SVM-algoritme aan. Hieruit sien ons dat die optimering wel ’n verhoging in prestasie te weeg bring (soveel as 0.024 in die geval van SVM met eienskapkombinasies), maar dat hierdie verskil nie ’n statisties beduidende een is nie (p is telkens > 0.05). Hierdie vergelyking word in Tabel 15 getref tussen die geoptimeerde SVM-algoritmes en die resultate van MNB. Ook hier word daar opgemerk dat daar geen statisties beduidende verskil tussen die resultate is nie (p is telkens > 0.05).

SVM (3 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 14. p-Waardes vir eienskapvergelyking met ART vir geoptimeerde SVM teenoor standaard SVM (3 klasse)

Algoritme (3 Klasse)	**MNB: Woordver-sameling**	**MNB: *tf-idf***	**MNB: Karakter-trigramme**	**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

Tabel 15. p-Waardes vir eienskapvergelyking met ART na optimering vir SVM en MNB (3 klasse)

3.2.7. Optimale klassifiseerder

Uit die bogenoemde stappe van die eliminerende eksperimente word MNB geïdentifiseer as die beste algoritme, drie klasse as die optimale stel klasse, ŉ woordversamelingbenadering as die beste benadering tot eienskaponttrekking (geïdentifiseer met ŉ resulterende *f*-telling van 0.929). Daar kon egter nie bewys gelewer word dat die benadering statisties beduidend beter is as die ander benaderings wat in Hoofstuk 3 voorgehou word nie. Dit word daarom geargumenteer dat enige van dié benaderings dus gebruik kan word om ongeveer dieselfde resultate te verkry (inaggenome al die veranderlikes wat by die prestasie ŉ rol speel) en dat die identifisering van die beste benadering dan gebaseer word op die uiteindelike *f*-telling. Die resultate vir die optimale klassifiseerder word hieronder in Tabel 16 gelys saam met klassifiseerders uit die literatuur wat in Tabel 3 voorgehou is. Kompeterende resultate, vir die hoeveelheid afrigtingsdata en klasse, teenoor die sisteme uit die literatuur word waargeneem.

Daar moet egter daarop gelet word dat die resultate van die klassifiseerders almal gebaseer is op verskillende hoeveelhede klasse, afrigtingsdata, toetsstelle en algoritmes, en dat dit alles 'n invloed het op die uiteindelike resultate wat gelewer word. Klassifikasiesisteme kan eintlik nie volledig vergelyk word tensy dit op dieselfde toetsstel geëvalueer word en die voorbeelde uit die literatuur word slegs daargestel om as 'n verwysing te dien vir die optimale klassifiseerder wat hierbo geïdentifiseer is. Wanneer die resultate uit die literatuur vergelyk word met die optimale klassifiseerder word daar opgemerk dat die prestasie vir die sisteme vergelykbaar is en dat die optimale klassifiseerder in al die gevalle beter prestasie toon.

# Tekste	Korpus	# Klasse	Eienskappe	Taal	Algoritme	*f*-Telling	
319	Internet: Tesisse en verhandelinge	11	*tf-idf*	Engels	NB[10]	0.890	[60]
1224	Webdokumente	16	HTML-etikette Webadresinligting Leksikale eienskappe	Engels	SVM	0.757	[27]
800	Nuus webblaaie	2	Woordversameling Woordsoortinligting Teksstatistiek	Engels	Besluitnemings-bome	0.905	[12]
1083	Koerantberigte	20	Letter 5-gramme Morfologiese inligting	Duits	SVM	0.540	[14]
499	Brown-korpus	6	Strukturele inligting Leksikale eienskappe Teksstatistiek	Engels	*k*-NN	0.870	[23]
2172	**NCHLT**	**3**	**Woordversameling**	**Afrikaans**	**MNB**	**0.929**	

Tabel 16. Beste kombinasie van klassifiseerder en eienskappe

[10] Standaard Naïewe Bayes-klassifiseerder

Ten slotte word enkele opmerkings oor die foute wat tydens klassifikasie gemaak word voorgehou. Die wyse waarop die optimale klassifiseerder met die verskillende klasse omgaan, word geanaliseer aan die hand van die klastoekennings tydens evaluering en word in 'n verwarringsmatriks (Tabel 17) voorgehou.

Genreklas		**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**

Tabel 17. Verwarringsmatriks vir optimale klassifiseerder

Uit Tabel 17 word daar gemerk dat die optimale klassifiseerder slegs drie uit die moontlike 229 toetsgevalle vir NF-EXP verkeerd geklassifiseer het. Drie van hierdie misklassifikasies 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-fiksie tekste (NON) stem ooreen na aanleiding van die tipe register wat gebruik word en die algemene woordkeuses wat tipies by sulke tekste opgemerk word.

3.3. UITBREIDING VIR ANDER HULPBRONSKAARS TALE

In afdeling 31 word die optimale kombinasie van algoritme, die hoeveelheid klasse, eienskappe en die hoeveelheid data wat gebruik word vir outomatiese genreklassifikasie geïdentifiseer. Hierdie optimale instellings is gebaseer op data wat vir Afrikaans beskikbaar is. Afrikaans, hoewel dit steeds as hulpbronskaars beskou word, beskik oor meer en beter hulpbronne as enige van die ander inheemse tale van Suid-Afrika [16]. Die Suid-Afrikaanse regering het in die laaste dekade erns gemaak met die bevordering van veeltaligheid en die ontwikkeling van onderverteenwoordigende tale in die land [15]. Taalbeleide van die regering vereis dat staatsdienste voorsiening moet maak om dienste te lewer in die elf amptelike landstale, sovêr dit prakties uitvoerbaar is.

Dit sluit in die daarstelling van inligting in al die amptelike landstale oor die verskillende dienste wat deur die verskeie staatsdepartemente gebied word en die amptelike vorms wat deur die departemente gebruik word. Amptelike kommunikasie soos, onder andere, die staatsrede van die President en ander belangrike toesprake, soos die begrotingsrede wat deur die Minister van Finansies gelewer word, word ook in verskillende tale vertaal.

Die toenemende digitalisering van dokumentasie en die beleide wat geïmplementeer is, het aanleiding daartoe gegee dat die staat hierdie veeltalige tekste deur middel van nasionale[11] en provinsiale[12] webblaaie van die beskikbaar stel. Volgens Resnik en Smith [43] en Keller *et al.* [22] het hierdie tipe aanlyn bronne van parallelle tekste 'n essensiële deel van veeltalige natuurliketaalprosesseringswerk geword. Hierdie bronne kan ontgin word deur gebruik te maak van geoutomatiseerde snuffelprogramme ("web crawlers" of "web spiders") wat algemeen gebruik word om volledige webblaaie (insluitend die HTML bladsy en enige ander dokumente van die webblad) af te laai.

Die optimale benadering wat vir Afrikaans geïdentifiseer word nou vervolgens gebruik om genreklassifikasiesisteme af te rig en te evalueer vir nege[13] ander inheemse Suid-Afrikaanse tale, om vas te stel of die tendens wat in die vorige gedeeltes geïdentifiseer word steeds geld wanneer die taak oorgedra word na ander tale toe. Die afrigtingsdata vir elkeen van die tale word saamgestel uit parallelle tekste (d.i. tekste waarvan daar weergawes in die ander tale beskikbaar is) [43]. Soos reeds genoem, word die Afrikaanse afrigtingsdata geanaliseer en handmatig geannoteer met die betrokke genreklas. Die genre wat dan aan die Afrikaanse weergawe toegeken is, kan dan outomaties oorgedra word na die weergawes in die inheemse tale en sodoende kan die annotering van die afrigtingsdata en die samestelling daarvan bespoedig word. Hierdie benadering het egter baie min geannoteerde afrigtingsdata tot gevolg gehad omdat tekshulpbronne soos parallelle dokumente tussen Afrikaans en die ander tale skaars is [16]. Dit het tot gevolg gehad dat van die tale baie min tot geen afrigtingsgevalle vir van die klasse beskikbaar gehad het nie. Hierdie benadering sou die tydrowende proses van handmatige annotasie vir al die tale vermy, maar die gebrek aan afrigtingsgevalle noodsaak egter dat daar wel menslike intervensie sou moes plaasvind.

[11] Byvoorbeeld http://www.services.gov.za/
[12] Byvoorbeeld http://www.westerncape.gov.za/
[13] Geen eksperimente word vir Engels gedoen nie; Engels word nie as 'n hulpbronskaars taal gereken nie.

Om die dataskaarsheid wat opgemerk word wanneer slegs dokumente, wat parallel met die bestaande Afrikaanse afrigtingsdata is, gebruik word te verlig, word die bogenoemde parallelle tekste van die staatswebblaaie onttrek. Die volledige webblaaie, tesame met alle ander dokumente, word afgelaai en dan verwerk na ŉ platteksformaat. Die taal waarin ŉ teks geskryf is, word dan bepaal deur gebruik te maak van ŉ taalidentifiseringsprogram [38] wat die taal van ŉ teks kan vasstel met ŉ presisie van ongeveer 90%. Die program word gestel dat dit die taal van ŉ teks met ŉ 80% sekerheid moet kan identifiseer (d.i. dat die teks ten minste uit 80% van ŉ betrokke taal moet bestaan) alvorens die taal dienooreenkomstig geïdentifiseer word. Parallelle tekste word onttrek aan die hand van die lêername van die tekste soos wat dit op die oorspronklike webblad verskyn. Nadat die taal van die tekste vasgestel is, word die tekste deur moedertaalsprekers (een per taal) geanaliseer en geannoteer volgens genre.

Daar word uit die data wat deur CTexT® beskikbaar gestel is nog parallelle tekste geïdentifiseer waarvoor daar nie Afrikaanse weergawes uit die afrigtingsdata beskikbaar is nie. Hierdie tekste is afkomstig uit tydskrifte waarvoor daar parallelle weergawes in Engels, isiZulu, Sesotho en isiXhosa beskikbaar is. Hierdie tekste word weereens handmatig geannoteer.

Die hoeveelheid tekste wat handmatig geanaliseer moet word, kan beperk word deur die tekste in die taal met die meeste parallelle tekste eerste te annoteer en die annotasies oor te dra na die ander tale toe. Hierdie proses word dan herhaal vir die taal met die volgende meeste oorblywende parallelle tekste en hou aan herhaal totdat slegs die dokumente oorbly waarvoor daar nie parallelle weergawes in die ander tale is nie. Sodoende word die hoeveelheid tekste waarvoor handmatige annotasie nodig is, beperk. Die uiteindelike afrigtingsdatasyfers word in Tabel 18 voorgehou. Afrikaans word weer hier weergegee vir verwysingsdoeleindes.

	Genreklas			
Taal	**NF-EXP**	**NF-APP**	**NF-INF**	**Totaal**
Afrikaans	229	439	536	1204
isiNdebele	9	33	854	896
isiXhosa	740	574	1091	2405
isiZulu	475	437	1416	2328
Sepedi	27	15	1051	1093
Sesotho	208	376	1232	1816
Setswana	41	109	1199	1349
siSwati	9	41	849	899
Tshivenda	35	39	820	894
Xitsonga	147	26	719	892

Tabel 18. Afrigtingsdatasyfers vir die elf inheemse Suid-Afrikaanse tale

In Tabel 18 word daar groot verskille in die hoeveelheid beskikbare afrigtingsdata per klas en per taal opgemerk. Hierdie verskille is te wyte aan die parallelle bronne waaruit hierdie data onttrek is waarvan die balans van beskikbare tekste nie noodwendig versprei is oor al die klasse en tale nie. Die invloed van die tydskrifte op die datasyfers van isiXhosa, isiZulu en Sesotho kan duidelik gemerk word. Hierdie drie tale het die hoogste totale hoeveelheid afrigtingsgevalle. Wat die syfers vir isiXhosa verder bevoordeel is dat die provinsiale regering van die Wes-Kaap deur middel van hulle webblad Engelse, Afrikaanse en IsiXhosa tekste beskikbaarstel. Dit is te wyte aan die groot konsentrasie Xhosasprekers in hierdie provinsie. Deur op slegs twee tale te fokus vir die vertaling van hulle tekste, word daar meer tekste vir elkeen van die tale gebied, eerder as ’n klein hoeveelheid tekste vir ’n groot hoeveelheid tale.

Die voordeel wat isiZulu het met betrekking tot die hoeveelheid beskikbare afrigtingsgevalle kan moontlik te wyte wees aan die groot hoeveelheid sprekers van dié taal. Die nuutste nasionale sensusresultate[14] toon aan dat isiZulu deur 22.7% van Suid-Afrika se bevolking as huistaal beskou word, gevolg deur isiXhosa (16%) en Afrikaans (13.5%). Volgens Grover *et al.* [15] is isiZulu ook die derde grootste taal ten opsigte van die beskikbare elektroniese hulpbronne.

[14] http://www.statssa.gov.za/Census2011/Products/Census_2011_Key_results.pdf

MNB	Woordversameling: 3 Klasse		
Taal	**Presisie**	**Herroeping**	***f*-Telling**
Afrikaans	0.931	0.93	0.929
isiNdebele	0.971	0.950	0.955
isiXhosa	0.801	0.781	0.788
isiZulu	0.823	0.775	0.765
Sepedi	0.959	0.870	0.907
Sesotho	0.868	0.844	0.840
Setswana	0.908	0.707	0.768
SiSwati	0.970	0.960	0.964
Tshivenda	0.951	0.954	0.952
Xitsonga	0.889	0.802	0.823

Tabel 19. Resultate vir nege tale, optimale instellings

Tabel 19 wys die resultate verkry vir die oordrag van die optimale versameling (te wete MNB-algoritme, drie klasse met ’n woordversamelingbenadering) na die ander inheemse Suid-Afrikaanse tale. Afrikaans word weereens hier ingesluit vir verwysingsdoeleindes. Die resultate vir die oordrag van die optimale versameling blyk positief te wees. Die resultate verkry is soortgelyk aan dié vir Afrikaans, maar daar is egter onreëlmatigheid in die uitslae van die verskillende tale. Daar word uit Tabel 19 wel gemerk dat die tale wat die grootse hoeveelheid afrigtingsgevalle beskikbaar het slegter resultate lewer as die gevalle waar daar relatief min afrigtingsgevalle is.

Dit is moontlik te wyte aan die bogenoemde tydskrifte waaruit tekste bygevoeg is. Dit is moontlik dat die tydskriftekste nie noodwendig prototipiese gevalle van die genreklas voorstel nie. Stilistiese elemente van die tydskrif is moontlik hiervoor verantwoordelik. Hierdie ekstra data wat moontlik nie prototipies is nie, dra nie noodwendig by tot die uiteindelike prestasie nie en beïnvloed dit moontlik negatief. Dit is ook moontlik dat hierdie voorkoms nie geïsoleerd is tot die tydskrifte nie en dat daar moontlik verkeerdelik tekspare as parallel geïdentifiseer word (op grond van die lêername), wat in werklikheid nie parallel is nie, wat geraas in die klasse veroorsaak.

Verder word daar gemerk uit Tabel 19 dat die tale waarvoor die minste data beskikbaar

is, die hoogste resultate lewer (vgl. isiNdebele 0.955, Tshivenda 0.952 en siSwati 0.964). Liu *et al*. [28] stel dat oormatige passing algemeen is by kleiner afrigtingsstelle en klasse. Hierdie oormatige passing het tot gevolg dat die klassifiseerder té gefokus is op die afrigtingsdata en nie goed veralgemeen nie wat hierdie uitermatige hoë resultate tot gevolg het.

Beide van hierdie voorkomste dui derhalwe daarop dat die samestelling van die afrigtingsdatastelle 'n kritiese faktor is wat baie aandag behoort te geniet. Vir die onderverteenwoordigde klasse sou dit nodig wees om meer afrigtingsdata te verskaf en waar die klasse reeds redelike verteenwoordigendheid toon, moet daar omgesien word daarna dat slegs prototipiese gevalle, wat die klas beskryf, toegevoeg word. Dit is verder belangrik om alle outomatiese prosesse deurgaans te kontroleer.

3.4. SAMEVATTING

In afdeling 3.2 word 'n reeks eksperimente met die belangrikste masjienleer-benaderings tot genreklassifikasie uit die literatuur bespreek, te wete *k*-NN-klassifiseerders, SVM-klassifiseerders, MNB-klassifiseerders, besluitnemingsbome en die RIPPER-algoritme. Hierdie masjienleeralgoritmes is geëvalueer in kombinasie met die algemeenste maniere uit die literatuur waarop afrigtingsdata as eienskappe geënkodeer word. Hierdie eienskappe is woordversameling, *tf-idf*-tellings, karakter- en woord-*n*-gramme en 'n kombinasiebenadering. Die beskikbaarheid van verskillende hoeveelhede data vir die masjienleerbenaderings word ook ondersoek en die resultate word hier weergegee. Twee benaderings tot klasetikette vir genreklassifikasie word bespreek en 'n drie-klas-benadering word saam met die bogenoemde benaderings gebruik om uiteindelik 'n optimale klassifiseerder te identifiseer. Dié optimale klassifiseerder (MNB, drie klasse, met 'n woordversamelingbenadering) lewer 'n resulterende *f*-telling van 0.929. Hierdie benadering is reeds in afdeling 3.3 toegepas op die nege ander inheemse Suid-Afrikaanse tale. Die resultate wat op hierdie benadering gebaseer is, blyk positief te wees, en die resultate is ook vergelykbaar met dié van Afrikaans, wat aantoon dat die oordrag van hierdie benadering meriete het. Daar word egter gemerk dat die prestasie van die genreklassifikasiesisteme wat afgerig is met meer data, swakker vaar as die sisteme met minder beskikbare data. 'n Moontlike rede wat hiervoor verskaf word, is dat daar 'n moontlikheid is vir oormatige passing van die masjienleeralgoritmes op die kleiner datastelle. Verder bestaan die moontlikheid dat die tekste wat by die groter datastelle ingesluit is, dalk nie prototipies is van die genreklasse

wat deur die spesifieke tekste voorgestel moet word nie. Dit het tot gevolg dat onbekende tekste dan verkeerd geklassifiseer word tydens klassifikasie wat weer lei tot laer resultate.

Dié optimale klassifiseerder word derhalwe in die volgende hoofstuk gebruik om die moontlikheid van tegnologieherwinning tussen nabyverwante tale vir genreklassifikasie te toets. Die tale hier onder bespreking is Afrikaans en Nederlands.

HOOFSTUK 4: ONTWIKKELING VIR NABYVERWANTE TALE

4.1. INLEIDING

In afdeling 3.3 word daar ’n reeks probleme geopper oor die oordrag van genreklassifikasiesisteme van een taal na ’n ander in ’n hulpbronskaars omgewing. Een van die hoofprobleme wat hier geïdentifiseer word, is ’n gebrek aan beskikbare data om te gebruik as afrigtingsdata vir al die betrokke genreklasse. Die vraag ontstaan nou: Wat kan gedoen word om hierdie probleem van onder-verteenwoordigendheid van afrigtingsdata te oorkom? Die eenvoudige antwoord sou wees om bloot meer afrigtingsdata te annoteer en toe te voeg. In ’n hulpbronskaars omgewing sal daar egter meer “kreatiewe” maniere ondersoek moet word omdat dit gewoonlik die geval is dat daar nie nog data beskikbaar is om te annoteer en toe te voeg nie. ’n Moontlike oplossing hiervoor is die implementering van tegnologie-herwinning (“technology recycling”).

Volgens Pilon *et al*. [39] sou dit goedkoper en vinniger wees om bestaande tegnologieë van ’n nabyverwante taal te hergebruik om die ontwikkeling vir ’n hulpbronskaars taal te bespoedig, eerder as om hierdie tegnologieë van voor af te ontwikkel. Die bestaande tegnologieë kan hergebruik word deur verskeie aanpassings te maak aan die taal waarvoor die tegnologieë nog nie bestaan nie, as die taal waarin ’n tegnologie reeds bestaan, nabyverwant genoeg is [39]. Die taal waarvoor die tegnologie herwin word, word dus verander sodat dit meer soos die nabyverwante taal “lyk” deur metodes soos sintaktiese herordening, aanpassing van morfeme, of volledige of gedeeltelike masjienvertaling [39]. In die konteks van die Suid-Afrikaanse hulpbronskaars tale, sou hierdie benadering moontlik van waarde kon wees in gevalle waar daar bestaande hulpbronne vir een taal in ’n taalfamilie is, maar nie vir die ander nie, of waar die een taal oor meer hulpbronne beskik as die ander en ontwikkeling dan eers vir dié taal gedoen word, met die idee om oordrag na die ander tale te vergemaklik.

Om die moontlikheid hiervan te ondersoek vir genreklassifikasie, is daar ŉ reeks vereistes waaraan voldoen moet word voordat eksperimente hieroor uitgevoer kan word. Twee nabyverwante tale (L1 – taal waarin tegnologie reeds bestaan; L2 – taal waarvoor tegnologie oorgedra moet word deur tegnologieherwinning) waarvoor die volgende hulpbronne beskikbaar is, moet geïdentifiseer word:

- ŉ L1-genreklassifikasiesisteem (d.i. ŉ klassifikasiemodel, afgerig met genre-spesifieke inligting);
- ŉ Versoenbare L2-toetskorpus (d.i. ŉ korpus, geannoteer met genreklasse wat versoen kan word met die L1 klassifikasiemodel); en
- ŉ Masjienvertaalsisteem wat (gedeeltelike) vertaling van L2 na L1 kan uitvoer.

Omdat die bogenoemde hulpbronne ontbreek vir die meerderheid van die Suid-Afrikaanse tale, is eksperimentering oor die uitvoerbaarheid van tegnologieoordrag dus nog nie tans moontlik nie. Van Huyssteen en Pilon stel in [50] dat Afrikaans en Nederlands soortgelyk genoeg is om te gebruik vir tegnologieherwinning. Die taalpaarkombinasie van Afrikaans en Nederlands voldoen aan die bogenoemde vereistes van hulpbronne wat noodsaaklik is vir effektiewe tegnologieherwinning. Ten einde ŉ ondersoek te loods na tegnologieoordrag vir genreklassifikasie vir Afrikaans en Nederlands, sal daar ŉ aantal aannames gemaak moet word.

Omdat daar in hierdie studie reeds ŉ geoptimeerde genreklassifikasiesisteem vir Afrikaans daargestel is, kan hier argumentsonthalwe na Afrikaans as die hulpbronryke taal verwys word (Afrikaans word as die L1-taal geïdentifiseer), en Nederlands as die hulpbronskaars taal (L2). Sodoende kan die Afrikaanse genreklassifiseerder se effektiwiteit op onbekende, Nederlandse tekste getoets word. Hierdie toepassing behoort verder ŉ aanduiding te gee van die robuustheid van die genreklassifiseerder en die moontlikheid van tegnologieherwinning vir genreklassifikasie.

Vervolgens word daar in hierdie hoofstuk ŉ reeks eksperimente voorgehou waar onbekende Nederlandse tekste met ŉ Afrikaanse genreklassifiseerder geklassifiseer word, waarna die Nederlandse tekste met beide ŉ leksikale oordragbenadering (d.i. reëlgebaseerde masjienvertaling) en ŉ volwaardige masjienvertaalsisteem vertaal word na Afrikaans. Die verskille tussen dié twee benaderings en die moontlike voordele wat dit vir tegnologieherwinning inhou, word toegelig.

4.2. KRUISTALIGE GENREKLASSIFIKASIE

Relatief min navorsing is beskikbaar oor die toepassing van ’n genreklassifikasiesisteem wat in een taal afgerig is, op ’n ander taal. Die eerste navorsing oor kruistalige genreklassifikasie word in [36] voorgehou. Petrenz [36] stel dat baie navorsing poog om taalonafhanklike benaderings daar te stel vir teksklassifikasietoepassings, maar dat daar baie selde empiriese bewyse gelewer kan word dat die taalonafhanklike benaderings werklik werk. ’n Voorbeeld hiervan, volgens Petrenz [36], is Sharoff [46], wat voorgestel het dat deur woordsoortinligting van die mees frekwente woorde te onttrek en as eienskappe te gebruik om Engelse en Russiese tekste volgens genre te klassifiseer, ’n benadering is wat taalonafhanklik kan wees. Sharoff se eksperimente is egter nie tussen die twee tale uitgevoer nie, maar op ’n per taal basis [46]. Met ander woorde, die eksperimente is slegs uitgevoer deur Engelse tekste met ’n sisteem te klassifiseer wat op die Engelse woordsoortinligting gebaseer is, en die Russiese tekste te klassifiseer met ’n sisteem wat op die Russiese woordsoortinligting gebaseer is. Die doel was om die woordsoorteienskappe as taalonafhanklik te bewys (deur die eienskappe vir elke taal te onttrek), eerder as om genreklassifikasie buite die taalbeperking van die klassifiseerders te toets [46]. Taalonafhanklik is egter nie die regte beskrywing vir hierdie eienskappe nie, omdat die eienskappe per taal onttrek en getoets word. Volgens Petrenz [36] is taalonafhanklikheid in hierdie geval baie moeilik om te bewys omdat daar altyd taalspesifieke inligting inherent in die eienskappe is. Werklike taalonafhanklikheid sou dui op ’n model wat vir enige taal gebruik kan word om tekste in daardie taal te klassifiseer. Hierdie “taalonafhanklike” eienskappe van Sharoff [46] het by die klassifikasie van Engels resultate gelewer wat goed vergelyk het met soortgelyke sisteme, maar die eksperimente met Russies het getoon dat die sisteem nie die basislynsisteem (toekenning van die mees frekwente klas) kon klop nie [36]. Petrenz [36] stel voor dat die fokus vir eers sou moes wees op kruistalige klassifikasie. Deur hierop te fokus, sou ’n beter oorsig bied oor hoe klassifiseerders met ander tale omgaan en sou moontlike ooreenkomste en verskille uitlig wat as eienskappe gebruik kan word.

Hoewel genreklassifikasie oor taalgrense heen tot op hede min nagevors is, bestaan daar egter navorsing vir ander, meer algemene teksklassifikasietake [36]. Bel *et al*. [2] het in 2003 reeds van die eerste navorsing voorgehou oor onderwerpklassifikasie tussen Spaanse en Engelse tekste, wat volgens Petrenz [36] groot invloed gehad het op die rigting wat hierdie tipe navorsing ingeslaan het, en die metodes wat Bel *et al*. [2] beskryf, word dus in die algemeen gebruik vir kruistalige teksklassifikasie [36]. Die

benadering wat in [2] gevolg word, behels die gebruik van ’n onderwerp-klassifikasiesisteem wat in die een taal (Engels) afgerig word en dan gebruik word om tekste in die ander taal (Spaans) volgens onderwerp te klassifiseer. Daar ontstaan ’n diskrepansie tussen die verwagte woordeskat in die brontaal (d.i. die woordeskat van die afrigtingstel/klassifikasiemodel) en die uiteindelike woordeskat van die onbekende teks (in die teikentaal) wat geklassifiseer moet word. Daar word grotendeels gesteun op vertaling tussen die bron- en die teikentaal as ’n tussenstap om die afrigting en die toetsstelle te versoen. Bel *et al*. [2] stel voor dat hierdie vertaling gedoen kan word deur een van die volgende drie vertalingstrategieë te volg:

- Terminologievertaling [2]: Terminologielyste word op ’n per klas basis saamgestel en slegs die terme relevant vir klassifikasie in die spesifieke klas word dan vertaal in die teikentaaltekste. Dit word veronderstel dat al die woorde wat belangrik is vir klassifikasie hierby ingesluit word.
- Profielgebaseerde vertaling [2]: Slegs die terme wat uiteindelik in die afrigtingsdata vir elke klas voorkom, word vertaal.
- Volteksvertaling: Die volledige vertaling van die teikentaalteks na die brontaal. Hierdie benadering word deur Bel *et al*. [2] veroordeel weens die hoë kostes verbonde aan handmatige vertaling van tekste en die swak gehalte van masjienvertaalsisteme, en derhalwe word hierdie benadering nie deur Bel *et al*. [2] geëvalueer nie.

Soos genoem in afdeling 4.1 is tegnologieherwinning ’n manier om vinnig tegnologieë te ontwikkel deur gebruik te maak van ’n hulpbronskaars- en hulpbronryke taalpaar. Tegnologieherwinning word juis gebruik om die ontwikkelingsproses vir die hulpbronskaarstaal te bespoedig en om die afhanklikheid van menslike intervensie te mitigeer. Hoewel Bel *et al*. [2] die volteksvertalingsbenadering veroordeel, bewys Pilon *et al.* [39] die teendeel; dat selfs net gedeeltelike masjienvertaling reeds groot voordele inhou vir tegnologieherwinning en word die ontwikkeling van nuwe hulpbronne bespoedig en die behoefte vir menslike insette verminder. Daarom word daar in die volgende gedeelte tegnologieherwinning vir genreklassifikasie, deur gebruik te maak van ’n masjienvertalingtussenstap, bespreek.

4.2.1. Afrikaans en Nederlands

Soos in afdeling 4.1 genoem, word die effektiwiteit van die Afrikaanse genre-klassifikasiesisteem op Nederlandse toetsdata gemeet. Hiervoor word daar egter 'n toetskorpus benodig en vir die Nederlandse toetskorpus word 'n uittreksel uit die LASSY-korpus [51] gebruik. Hierdie amptelike uittreksel staan bekend as LASSY-Klein en is 'n miljoenwoordkorpus, geannoteer met sintaktiese inligting, sowel as woordsoortetikette en lemmas. Genre-inligting is ook teenwoordig, maar is ietwat moeiliker bekombaar; die genre-inligting word nie eksplisiet in die korpus genoem nie, maar daar word wel in die projekdokumentasie[15] melding gemaak hiervan. Daar kan uit die dokumentasie afgelei word watter genre sommige van die .xml-korpustekste bevat en sodoende kan daar 'n genre-geannoteerde subkorpus onttrek word. Die toekenning van die genre word gebaseer op die .xml-lêernaam wat in die dokumentasie gelys word. Die dokumentasie maak melding van die oorsprong van die teks wat moontlik 'n aanduiding gee van die genre van die betrokke teks. Nadat die genreklasse vir die tekste onttrek is, word die genreklasse van die LASSY-korpus gepas op die genreklasse van hierdie studie, deur 'n passing te maak na die oorspronklike 13 klasse en dan die passing van hierdie dertien klasse te maak na die drie meer abstrakte klasse (soos uiteengesit in 0). Die uiteindelike Nederlandse toetskorpussamestelling word in Tabel 20 voorgehou.

Abstrakte genreklasse	**# Tekste**
NF-EXP	75
NF-APP	546
NF-INF	107
Totaal	**728**

Tabel 20. Nederlandse toetskorpussamestelling

Vir die masjienvertaalkomponent word die "Dutch to Afrikaans Converter" oftewel D2AC van Van Huyssteen en Pilon [50] gebruik. D2AC is 'n reëlgebaseerde masjienvertaalsisteem wat gebaseer is op die ortografiese, morfosintaktiese en leksikale verskille tussen Afrikaans en Nederlands [50]. D2AC is dus nie 'n volledige masjienvertaalsisteem nie, maar doen slegs leksikale oordrag en is ontwikkel met die doel van tegnologieherwinning as motivering (d.i. om Nederlands só te verander dat dit meer na Afrikaans lyk). Hulle rapporteer 'n presisie van 71% vir woordvlakevaluasie en 'n BLEU-telling van 0.2519 [50].

[15] http://www.let.rug.nl/~vannoord/Lassy/deliverable1-1.pdf

Google se masjienvertaalsisteem, “Google Translate” (GT), word ook gebruik vir die masjienvertalingtussenstap en dien ŉ tweeledige doel. Eerstens word die resultate wat vir D2AC geverifieer is, vergelyk met GT; tweedens word die verskil in die bydrae tot die prestasie van die genreklassifikasiesisteem tussen die volledige masjienvertalingsbenadering van GT en die leksikale oordragbenadering van D2AC getoets. Van Huyssteen en Pilon [50] rapporteer die prestasie van GT op dieselfde toetsstel as waarmee D2AC geëvalueer is met ŉ BLEU-telling van 0.3162 [50].

Vervolgens word daar dan ŉ reeks eksperimente uitgevoer:

- Onbekende Nederlandse tekste word sonder enige aanpassing met ŉ Afrikaanse genreklassifikasiesisteem geklassifiseer. Daar word gelet op die resultate wat hierdie klassifikasie tot gevolg het.
- Onbekende Nederlandse tekste word met ŉ masjienvertaalsisteem (gedeeltelik) vertaal en die vertaalde weergawes van die tekste word dan deur die Afrikaanse genreklassifikasiesisteem geklassifiseer en die resultate word weer waargeneem.
- Die resultate word vergelyk om die effektiwiteit van hierdie benadering vas te stel.

4.2.1.1. Afrikaanse klassifiseerder – Nederlandse toetsdata

Bel *et al*. [2] verwys na ŉ informele eksperiment met Spaanse toetsdata wat deur ŉ Engelse teksklassifikasiesisteem (afgerig met 2 164 dokumente) geklassifiseer word. Die Engelse klassifiseerder kan die klas van ŉ Spaanse teks met ŉ presisie (resultate vir ander metrieke word nie genoem nie) van 10.75% identifiseer. Daar word egter nie verdere inligting verskaf oor die aard of omvang van die eksperiment nie. Ten spyte van die verskil in woordeskat tussen die Engelse afrigtingsdata en die Spaanse toetsdata is daar steeds elemente wat deur die woordeskatte gedeel word, bv. eiename, akronieme en moontlik syfers wat in beide tale dieselfde voorgestel word. Dít is volgens Bel *et al.* [2] ŉ moontlike rede hoekom só ŉ hoë presisie gesien word. Tussen Afrikaans en Nederlands word daar egter ŉ heelwat groter oorvleueling tussen die woordeskatte verwag omdat die tale nabyverwant is. Hierdie 10.75% presisie kan gebruik word as ŉ verwysingspunt vir die eksperiment met skoon Nederlandse data en die Afrikaanse klassifiseerder. Die verwagte resultate behoort egter hoër te wees gegewe die hoër oorvleueling tussen woordeskatte.

Die eerste stel resultate van die Nederlandse LASSY-toetsdata (waarvoor die genreklasse outomaties toegeken is aan die hand van die korpusdokumentasie) wat met ŉ Afrikaanse genreklassifikasiesisteem geklassifiseer word, word in Tabel 21 voorgehou. Vir eenvoudige eienskapsonttrekking word ŉ woordversamelingbenadering gebruik.

Algoritme	**Woordversameling: 3 Klasse (NL)**		
	Presisie	**Herroeping**	***f*-Telling**
MNB	**0.392**	0.281	0.277
SVM (C=0.03125 (2^{-5}))	0.312	**0.488**	**0.363**

Tabel 21. Resultate Nederlandse toetsdata, Afrikaanse klassifiseerder

As die resultate vir klassifikasie van Nederlands met ŉ Afrikaanse klassifiseerder met dié van Engels en Spaans vergelyk word, word daar gemerk dat die Afrikaanse klassifiseerder vir beide MNB en SVM beter resultate lewer. Wanneer die resultate egter met die basislynsisteme (sien afdeling 3.2.2) vergelyk word, kan slegs die SVM klassifiseerder die lukrake klasbasislyn klop. Die mees frekwente klasbasislyn is egter steeds hoër. Vir nabyverwante tale wat groter gedeeltes van hulle woordeskatte deel, is dit te verwagte dat die klassifiseerders beter resultate sou lewer as die Engels/Spaans-kombinasie, maar gebaseer op die resultate wat Pilon *et al.* [39] rapporteer vir Afrikaans en Nederlands, sou hoër resultate verwag word.

Die toetsstel word nou vervolgens aangepas deur dit te vertaal met D2AC en GT. Hierdie vertaalde toetsstelle word telkens weer met die Afrikaanse klassifiseerder geklassifiseer. Die verwagting uit die literatuur is dat die vertaalde weergawe van ŉ toetsstel baie beter resultate tot gevolg sal hê tydens evaluering as wat die toetsstel in die oorspronklike taal lewer [36][2][40][39][50]. Die resultate vir die vertaalde toetsstel word in Tabel 22 aangetoon.

Woordversameling: 3 Klasse (NL)			
Algoritme	**Presisie**	**Herroeping**	***f*-Telling**
MNB	**0.392**	0.281	0.277
SVM (C=0.03125 (2^{-5}))	0.312	**0.488**	**0.363**
Woordversameling: 3 Klasse (D2AC: NL)			
MNB	**0.423**	0.284	0.316
SVM (C=0.03125 (2^{-5}))	0.412	**0.404**	**0.407**
Woordversameling: 3 Klasse (GT: NL)			
MNB	**0.481**	0.282	0.381
SVM (C=0.03125 (2^{-5}))	0.462	**0.432**	**0.446**

Tabel 22. Resultate vir vertaalde Nederlandse toetsdata, Afrikaanse klassifiseerder

Uit Tabel 22 word daar ŉ verhoging van hoogstens 0.104 in die *f*-telling waargeneem (MNB met GT) wanneer die toetsstel vertaal word na Afrikaans. Die resultate hier blyk dan ietwat laag te wees, veral as die nabyverwante status van die tale in ag geneem word. Bel *et al*. [2] rapporteer akkuraathede van tussen 0.538 en 0.845 vir vertaalde tekste vir nie-verwante tale. Om ŉ moontlike verduideliking te kry van hoekom die verwagte resultate nie bereik word nie, word daar vervolgens gekyk na die veranderlikes wat by hierdie eksperiment ŉ rol speel. Die eerste is die genreklassifikasiesisteem wat geïmplementeer word. Dié genreklassifikasiesisteem is reeds in afdeling 3 bewys as ŉ sisteem waarvoor daar geen onreëlmatige resultate opgemerk is nie, wat aantoon dat die benadering wat gevolg word, sowel as die afrigtingsdata (in beide samestelling en voorstelling) wat gebruik word, korrek is. Die tweede veranderlike is die masjienvertaalkomponent van hierdie eksperiment. D2AC en GT is deur Van Huyssteen en Pilon [50] geëvalueer en suksesvol in soortgelyke tegnologieherwinningseksperimente gebruik. Die derde, en laaste veranderlike is die toetskorpus. Soos genoem in afdeling 4.2.1 is daar gepoog om outomaties die genreklasse van die LASSY-korpus se tekste te identifiseer aan die hand van die korpusdokumentasie. Ten einde vas te stel of daar moontlik foute by die toekenning van die genre voorgekom het, word die handmatige genreannotasie in die volgende gedeelte bespreek.

4.2.1.2. Versoening van genre-geannoteerde korpusse

Soos reeds genoem in 0 is daar ŉ wye reeks kwessies met betrekking tot die identifisering van genre. Die verdeeldheid oor wat die term genre behels wat in die literatuur sigbaar is en die tendens om genre telkens te herdefinieer wanneer dit in navorsing gebruik word, laat navorsers met ŉ twyfelagtige vertrekpunt wanneer daar met genres gewerk word. Die effek hiervan kan gevolglik waargeneem word by die poging om genre-geannoteerde korpusse te versoen. By nadere inspeksie van die afrigtingsdata is daar gemerk dat die genreklasse wat uit LASSY afgelei kon word, nie outomaties gepas kon word op die genreklasse van hierdie studie nie. Die outomatiese annotering van die genreklasse het tot gevolg gehad dat sommige tekste verkeerd ingedeel is, byvoorbeeld ŉ inligtingstuk oor ŉ geografiese area wat as ŉ geskiedkundige teks ingedeel word. Dit het tot gevolg gehad dat die klasse nie ŉ gepaste voorstelling was nie, weens die geraas wat die verkeerd geklassifiseerde tekste tot gevolg gehad het.

Wanneer genreklasse geïdentifiseer word vir ŉ korpus/studie, word daar ŉ spesifieke protokol vir elke klas saamgestel van watter tipe tekste by elkeen van die klasse ingesluit behoort te word. Dit kan dan gebeur dat klasse wat ongeveer dieselfde naam het in verskillende korpussamestellings op die oog af versoenbaar lyk, maar in die werklikheid nie noodwendig dieselfde inhoud het nie, byvoorbeeld tekste wat as “periodicals” aangedui word wat nie slegs nuus en tydskrifartikels bevat nie, maar tekste met ŉ akademiese strekking ook bevat. ŉ Rede hiervoor is dat verskillende riglyne by verskillende korpusse geld vir die tekste van ŉ betrokke klas. Na aanleiding van die onverwagte resultate in afdeling 4.2.1.1 is daar besluit om die tekste van die LASSY-korpus handmatig te annoteer met die oorspronklike dertien genreklasse (sien Tabel 4). Nadat hierdie handmatige genreannotasie voltooi is, kon die passing van die oorspronklike dertien klasse op die drie klasse (soos uiteengesit in Tabel 5) gedoen word. Daarna kon die eksperiment met die Afrikaans en Nederlands herhaal word om vas te stel wat die invloed van die nuwe passing van die LASSY-genreklasse op die resultate van die eksperiment is. Tabel 23 toon die nuwe uiteensetting van die LASSY-toetsstel.

Abstrakte genreklasse	**# Tekste**
NF-EXP	321
NF-APP	391
NF-INF	16
Totaal	**728**

Tabel 23. Nederlandse toetskorpussamestelling na handmatige genre-annotasie

Woordversameling: 3 Nuwe klasse (NL)			
Algoritme	**Presisie**	**Herroeping**	***f*-Telling**
MNB	**0.660**	**0.385**	**0.438**
SVM (C=0.03125 (2^{-5}))	0.597	0.352	0.472
Woordversameling: 3 Nuwe klasse (D2AC: NL)			
MNB	0.631	0.384	0.544
SVM (C=0.03125 (2^{-5}))	**0.644**	**0.516**	**0.558**
Woordversameling: 3 Nuwe klasse (GT: NL)			
MNB	**0.672**	0.429	0.485
SVM (C=0.03125 (2^{-5}))	0.669	**0.484**	**0.577**

Tabel 24. Resultate vir handmatig geklassifiseerde Nederlandse toetsdata, Afrikaanse klassifiseerder

Die resultate vir die handmatig geklassifiseerde Nederlandse toetsstel word in Tabel 24 voorgehou. Hier word ŉ aansienlike verhoging in die resultate teenoor die oorspronklike weergawe van die eksperiment waargeneem. Wanneer die Nederlandse data sonder vertaling geklassifiseer word, is daar ŉ 0.109 verhoging in die *f*-telling waargeneem vir SVM. Die maatreëls wat toegepas is om die toetsstel se versoenbaarheid te verseker, blyk positief by te dra en kan die ooreenkomste tussen Afrikaans en Nederlands beter benut word tydens klassifikasie. Wanneer die Nederlandse toetskorpus vertaal word met D2AC en GT word daar weereens verbetering opgemerk. Die *f*-telling groei met 0.106 vir MNB en D2AC en groei met 0.086 vir SVM en GT, die hoogste tellings onderskeidelik. Hoewel die resultate nou binne die verwagte resultate van 0.538 tot 0.845 lê [2], is die verskil tussen die oorspronklike Nederlandse teks en die vertaalde weergawe daarvan steeds nie so groot soos wat in die literatuur te sien is nie [2].

Bel *et al.* [2] stel dat die verskil en/of ooreenkomste in woordeskat tussen die betrokke tale met kruistalige teksklassifikasie ŉ baie belangrike rol speel in die uiteindelike klas wat deur die masjienleeralgoritme toegeken word. Vervolgens is ŉ analise van die woordeskat van die afrigtingsdata vir Afrikaans en Nederlands gedoen. Om die mees informatiewe woorde (d.i. die woorde wat die meeste bydrae lewer tot die klassifikasie) te identifiseer, word die woorde se inligtingswins ten opsigte van die klassifikasiefunksie bepaal. Inligtingswins word gebruik om die informatiwiteit/bydrae van ŉ spesifieke woord, relatief tot die klassifikasiefunksie, te bepaal [57]. Die informatiwiteit word bepaal deur die hoeveelheid "kennis" wat verkry word deur die klassifiseerder as die woord voorkom in ŉ teks. Die informatiwiteit word as ŉ gemiddelde bydrae tot al die klasse bereken [57]. WEKA bied die funksionaliteit om die eienskappe te identifiseer volgens die bydrae wat die eienskap lewer by klassifikasie. Onder die funksionaliteit is die opsie om die eienskappe te evalueer aan die hand van die eienskap se inligtingswins en ŉ gerangskikte lys van die eienskappe weer te gee [18]. ŉ Uiteensetting van die woorde wat die meeste bydra tot klassifikasie (met die hoogste inligtingswins) vir die toetsstel in die oorspronklike sowel as die vertaalde Nederlands, word in Tabel 25 gegee.

Wanneer die woorde met die hoogste inligtingswins (die 20 hoogste tellings en woorde word weergegee) voor en na vertaling vergelyk word uit Tabel 25, word daar waargeneem dat daar ŉ groot oorvleueling is. Die woorde met ŉ hoë inligtingswins, wat uiteraard die meeste bydra tot die klassifikasie, bestaan reeds in die skoon Nederlandse teks as woorde waarvan die Afrikaanse weergawes presies dieselfde lyk. Wanneer die toetsstel dus vertaal word, word die woorde net so weergegee, maar woorde met ŉ laer inligtingswins se vertalings is wel anders in Afrikaans as wat dit in Nederlands is. Die lae inligtingswins sal beteken dat die woorde nie so ŉ groot invloed sal hê op die klassifikasie nie, al is die woorde dan Afrikaans. Dit sal daarom verduidelik hoekom die verskil tussen die klassifikasie van die oorspronklike en die vertaalde toetsstel laer is as wat verwag sou word. ŉ Moontlike oplossing hiervoor is om stoplyste (d.i. lyste van woorde wat nie by die finale afrigtingsdata ingesluit word nie) van die gemeenskaplike woorde wat nie noodwendig ŉ bydrae lewer tot die klassifikasie nie, maar steeds ŉ hoë inligtingswins het, saam te stel. Sodoende bly slegs die woorde oor wat regtig belangrik is ten opsigte van die betrokke genre en sal die inligtingswins van die oorblywende terme verhoog. So kan die uiteindelike prestasie moontlik verhoog word.

Nederlands	Inligtingswins		Vertaling	Inligtingswins
nog	0.302910		ik	0.379400
is	0.295780		ons	0.327000
maar	0.291490		nog	0.302910
dit	0.283370		so	0.302790
die	0.280110		is	0.295780
van	0.244970		maar	0.291490
al	0.239600		dit	0.283370
dat	0.239070		die	0.280110
was	0.235000		weer	0.277060
wat	0.228150		’n	0.267350
het	0.223320		van	0.244970
alles	0.212040		dink	0.242140
moet	0.192760		al	0.239600
daar	0.189680		sy	0.238080
dan	0.189630		was	0.235000
Frank	0.187300		baie	0.234390
toe	0.185300		wat	0.228150
om	0.176210		weet	0.218060
in	0.172250		tog	0.212930
goed	0.165540		hy	0.201740

Tabel 25. Inligtingswins vir oorspronklike en vertaalde Nederlands

Nog ’n moontlike rede vir die lae prestasie van die Afrikaanse sisteem op die Nederlandse toetsdata is die invloed van die domeinoordrag. Die hoeveelheid wat die sisteem se prestasie afneem wanneer dit met ’n toetsstel uit ’n ander domein geëvalueer word, word deur Finn en Kushmerick [12] ondersoek vir ’n onderwerpklassifikasie-sisteem. Daar word ’n afname van ongeveer 0.100 (gemiddeld oor al die eksperimente van verskillende domeine) in die presisie opgemerk vir domeinoordrag. Hierdie syfer is onbekend vir die Afrikaanse genreklassifikasiesisteem en daarom ook die moontlike invloed hiervan. As die prestasieverlies van die domeinoordrag gekombineer word met die onbekende taal, kan die verlies in prestasie moontlik hoog wees.

Ten spyte daarvan dat die resultate nie noodwendig so hoog is as wat verwag word vir nabyverwante tale nie, kan tegnologieherwinning steeds ’n belangrike rol speel, veral in die Suid-Afrikaanse hulpbronskaars konteks. Hierdie tegnologieherwinningsmetodes kan aangewend word in die ontwikkeling van nuwe hulpbronne vir onderontwikkelde

tale, deur tale wat oor meer hulpbronne beskik as 'n steunpilaar te gebruik. Die inheemse Suid-Afrikaanse tale kan gegroepeer word in taalfamilies: isiZulu, isiXhosa, isiNdebele en siSwati is deel van die Nguni taalgroep en die Sotho taalgroep bestaan uit Sepedi, Setswana en Sesotho. Xitsonga en Tshivenda maak egter nie deel uit van enige taalgroepe nie [15]. Hierdie groeperings word gemaak op grond van die verwantskappe tussen die tale en tegnologieherwinning kan tot 'n groot mate baie voordele inhou.

4.3. SAMEVATTING

In hierdie hoofstuk is die moontlikheid van tegnologieherwinning vir genreklassifikasie ondersoek. In afdeling 4.1 is 'n oorsig gegee van die konsep van tegnologieherwinning en die voordele wat die implementering daarvan vir nabyverwante tale inhou.

In afdeling 4.2.1 is die nabyverwantheid tussen Afrikaans en Nederlands bespreek, sowel as die korpussamestelling vir die Nederlandse toetsstel. Die onttrekking van genreannotasies uit die korpus is ook bespreek.

In afdeling 4.2.1.1 word die eksperimente wat uitgevoer is vir Afrikaans en Nederlands beskryf. Hier is aangetoon dat Nederlandse tekste relatief goed deur 'n Afrikaanse genreklassifikasiesisteem geklassifiseer kan word. Vir hierdie eksperiment is daar 'n *f*-telling van 0.472 waargeneem. Om die prestasie van die Afrikaanse klassifiseerder op die Nederlandse tekste te verbeter, is 'n masjienvertalingtussenstap geïmplementeer. Dit behels die gedeeltelike of volledige vertaling van die tekste deur 'n masjienvertaler voordat dit deur die genreklassifikasiesisteem geklassifiseer word. Daar is 'n aansienlike toename van 0.106 in *f*-telling opgemerk.

In afdeling 4.2.1.2 word die belangrikheid van die versoenbaarheid tussen 'n toetskorpus en die klassifikasiemodel uitgelig. Die toetsstel is handmatig geannoteer en die kruistalige genreklassifikasie-eksperiment is weer uitgevoer om die effek daarvan op die prestasie van die genreklassifikasiesisteem te toets. Deur te verseker dat die genreklasse volledig versoenbaar is, word daar 'n aansienlike verhoging opgemerk met die hoogste *f*-telling van 0.577 wat bereik word. Dié resultate is verder geanaliseer aan die hand van die inligtingswins relatief tot die klassifikasiefunksie van die oorspronklike, sowel as die vertaalde Nederlandse woorde.

HOOFSTUK 5: SLOT

5. SAMEVATTING

Die oorhoofse doel van hierdie studie was om ŉ lewensvatbare oplossing te vind vir die outomatiese genreklassifikasie van tekste wat in een van die tien inheemse amptelike tale van Suid-Afrika geskryf is en die daaropvolgende ondersoek na algemeen gebruikte metodes, die nodige hulpbronne en die voorstelling daarvan vir masjienleerbenaderings en die uiteindelike evaluering daarvan. Moontlikhede vir die oordrag van metodes tussen tale, sowel as tegnologieherwinning tussen nabyverwante tale sluit aan by die tema. Sodoende word ŉ bydrae gelewer tot die ontwikkeling van die hulpbronskaars tale van Suid-Afrika om die gebruik daarvan in tegnologiese ontwikkeling te bevorder.

In Hoofstuk 1 word daar melding gemaak van die rasionaal van die ontwikkeling van outomatiese genreklassifikasiesisteme (d.i. vir die gebruik by korpusontwikkeling). Hulpbronskaarsheid het egter ŉ invloed by die ontwikkeling van natuurliketaalprosesseringstegnologieë (spesifiek by ontwikkeling vir Suid-Afrikaanse tale) en word die volgende doelstellings geïdentifiseer om dié navorsing te rig:

- om ŉ ondersoek te loods na die bestaande genreklassifikasiesisteme, die tegnieke en benaderings te implementeer vir Afrikaans en die implementering volledig te evalueer;
- om die implementering toe te pas vir die ander Suid-Afrikaanse hulpbronskaars tale; en
- om die moontlikheid van tegnologieherwinning van genreklassifikasiesisteme vir hulpbronskaars tale te ondersoek en die effektiwiteit daarvan vas te stel.

Hierdie doelstellings verskaf dan die onderbou vir die daaropvolgende hoofstukke se ondersoek na die literatuur en die eksperimente wat uitgevoer is. Ten einde die effektiwiteit van die eksperimente te toets, word die standaardevalueringsmetrieke vir inligtingherwinning en die evaluering van klassifiseerders beskryf. Hierdie drie metrieke is presisie, herroeping en *f*-telling. Deur hierdie drie metrieke te gebruik, kan die sisteme wat in die eksperimenteringsfase daargestel word met mekaar vergelyk word.

In Hoofstuk 2 word die belangrikste benaderings tot genreklassifikasie uit die literatuur ondersoek en beskryf aan die hand van die onderliggende verskille tot die benaderings. Hierdie benaderings word verdeel in (1) algoritmes; (2) eienskappe; en (3) data.

(1) Vyf algoritmes word uiteengesit: *k*-NN-algoritme, SVM-klassifiseerders, besluitnemingsbome, MNB-klassifiseerders en die RIPPER-algoritme. Die werking van die algoritmes word kortliks bespreek en die sterk en swakpunte van die algoritmes, sowel as die geskiktheid daarvan vir genreklassifikasie word uitgelig.

(2) Die eienskappe (datavoorstellingsmetodes) wat ondersoek word, is ŉ woordversamelingbenadering, *tf-idf*-tellings karakter en woord *n*-gramme en ŉ kombinasiebenadering. Die onttrekking van hierdie eienskappe uit die teks word voorgehou.

(3) Dataversamelings wat gebruik word vir die afrig van genreklassifikasiesisteme word gelys. Hierdie versamelings uit die literatuur kan gebruik word as verwysingsraamwerk vir die omvang van hierdie sisteem, sowel as die verwagte prestasie van die algoritmes wat op soortgelyke datastelle gebaseer is. Daar moet wel daarop gelet word dat sisteme wat op verskillende datastelle afgerig is nie direk vergelykbaar is nie.

Hoofstuk 3 sit ŉ reeks eksperimente uiteen om die benaderings wat in Hoofstuk 2 bespreek word te toets vir die geskiktheid daarvan vir genreklassifikasie vir hulpbronskaars tale. Die problematiek rakende die identifisering van genreklasse word bespreek aan die hand van die tendens om telkens die konsep van genre te herontwerp wanneer daar navorsing oor genreverwante onderwerpe gedoen word, veral waar genre outomaties geannoteer of geïdentifiseer moet word. Daar word vervolgens eksperimente voorgehou waarin die algemene benaderings tot genre, sowel as ŉ meer abstrakte benadering getoets word, tesame met die algoritmes uit Hoofstuk 2. Hier word twee algoritmes geïdentifiseer wat die beste resultate lewer vir klassifikasie, by name SVM-klassifiseerder (*f*-telling = 0.901) en MNB-klassifiseerder (*f*-telling = 0.929).

Vervolgens word dié twee algoritmes gebruik om kombinasies van die verskillende eienskappe (te wete woordversameling, *tf-idf* tellings, karakter en woord *n*-gramme en ŉ kombinasiebenadering) uit Hoofstuk 2 saam met die algoritmes te toets. Daar word aangetoon dat die woordversamelingbenadering tot eienskappe, gekombineer met die MNB-klassifiseerder, die beste resultate lewer (*f*-telling = 0.929) gevolg deur die SVM-

klassifiseerder met dieselfde eienskappe (*f*-telling = 0.901). Leerkurwes word saamgestel om die effek van die hoeveelheid beskikbare data op die prestasie van die algoritme vas te stel, waarna algoritme-optimering gedoen word.

Hoofstuk 4 beskryf eksperimente wat die moontlikheid van tegnologieherwinning vir nabyverwante tale ondersoek. Nederlandse toetsdata word met 'n Afrikaanse klassifiseerder geklassifiseer om vas te stel wat die uitslae vir nabyverwante tale sou wees. Om die uitslae te probeer verbeter, word die Nederlandse toetsdata met 'n masjienvertaalsisteem [50] na Afrikaans vertaal om die Nederlandse toetsdata meer na Afrikaans te laat lyk en dan die eksperiment te herhaal.

5.1. GEVOLGTREKKING

Die navorsingsvrae wat gestel word in Hoofstuk 1, word weer voorgehou en kortliks beantwoord:

- Wat is 'n geskikte benadering tot genreklassifikasie vir 'n hulpbronskaars taal (d.i. 'n benadering geskik vir Afrikaans)?

Ten spyte van die hulpbronskaarsheid van die Suid-Afrikaanse tale is dit steeds moontlik om genreklassifikasiesisteme te ontwikkel waarvan die prestasie vergelykbaar met sisteme uit die literatuur is (Hoofstuk 3). Met die ontwikkeling van 'n outomatiese genreklassifikasiesisteem is daar 'n reeks veranderlikes wat in gedagte gehou moet word wat 'n invloed op die prestasie van masjienleerbenaderings het (d.i. die algoritme wat gebruik word, die hoeveelheid klasse, die hoeveelheid afrigtingsdata, en die datavoorstelling as eienskappe). As hierdie veranderlikes reg hanteer word en 'n optimale versameling van hierdie veranderlikes geïdentifiseer kan word, kan die ontwikkeling van 'n genreklassifikasiesisteem suksesvol gedoen word. In hierdie studie word daar 'n genreklassifikasie sisteem daargestel deur gebruik te maak van die volgende benadering: Die implementering van 'n MNB-algoritme, afgerig met woordversamelingbenadering vir eienskappe as voorstelling van drie genreklasse. Hierdie sisteem lewer 'n resulterende *f*-telling van 0.929. Daar moet wel in gedagte gehou word dat die samestelling van die afrigtingsdata 'n belangrike rol speel by die uiteindelike prestasie van die genreklassifikasiesisteem. Dit is belangrik om die 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

(sien afdeling 3.2.1.1), behoort gediversifiseer te word. Die moontlikheid en invloed van oormatige passing behoort ook in gedagte gehou te word, veral tydens optimering, waar daar ŉ 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.

- Is hierdie benadering oordraagbaar na ander hulpbronskaars tale?

Die optimale versameling, soos hierbo genoem, kan suksesvol oorgedra word na ander tale om die ontwikkeling van genreklassifikasiesisteme te bewerkstellig. Die prestasie van die sisteme, waarvoor die benadering wat gevolg is vir Afrikaans, oorgedra word, lewer resultate wat vergelykbaar is met beide die Afrikaanse, sowel as die sisteme uit die literatuur. Die resultate verkry is soortgelyk aan dié vir Afrikaans, maar daar is egter onreëlmatigheid in die uitslae van die verskillende tale. Daar word wel gemerk dat die tale wat die grootse hoeveelheid afrigtingsgevalle beskikbaar het slegter resultate lewer as die gevalle waar daar relatief min afrigtingsgevalle is. Dit is moontlik te wyte aan geraas in die afrigtingsdatastel. Verder word daar gemerk dat die tale waarvoor die minste data beskikbaar is, die hoogste resultate lewer. Liu *et al*. [28] stel dat oormatige passing algemeen is by kleiner afrigtingsstelle en klasse. Beide van hierdie voorkomste dui derhalwe daarop dat die samestelling van die afrigtingsdatastelle ŉ kritiese faktor is wat baie aandag behoort te geniet. Vir die onderverteenwoordigde klasse sou dit nodig wees om meer afrigtingsdata te verskaf en waar die klasse reeds redelike verteenwoordigendheid toon, moet daar omgesien word daarna dat slegs prototipiese gevalle, wat die klas beskryf, toegevoeg word. Dit is verder belangrik om alle outomatiese prosesse deurgaans te kontroleer.

- Op watter manier kan bestaande genreklassifikasiebenaderings herwin word vir nabyverwante tale?

Eksperimente met nabyverwante tale (Hoofstuk 4) toon aan dat ŉ Afrikaanse klassifiseerder al tot ŉ redelik suksesvolle mate Nederlandse data kan klassifiseer, sonder dat daar enige aanpassings aangebring hoef te word. Uit die literatuur word daar gesien dat kruistalige genreklassifikasie meestal gedoen word deur die toetsstel te vertaal na die taal waarin die klassifiseerder afgerig is. Hiervoor kan beide handmatige of outomatiese metodes gebruik word. Die vertalingtussenstap verseker dat die afrigting en toetsstel oor dieselfde woordeskatte beskik wat dan die klassifikasie vergemaklik. Deur ŉ masjienleerbenadering te gebruik om die tale eenders te laat lyk, lei wel tot ŉ verbetering in die prestasie van die Afrikaanse klassifiseerder met die klassifikasie van Nederlandse data, maar die toename is nie so groot soos wat uit die literatuur verwag sou word nie. Dit is moontlik te wyte aan ŉ prestasieverlies as gevolg van domeinoordrag wat plaasvind wanneer die Nederlandse toetsdata nie volledig ooreenstem met die Afrikaanse afrigtingsdata nie. Dit is daarom belangrik om die invloed van domeinoordrag (d.i. die toepassing van die genreklassifikasiesisteem buite die omvang waarvoor die afrigtingsdata voorsiening maak) vas te stel.

Kruistalige genreklassifikasie blyk meriete te hê, veral in ŉ omgewing waar daar tale in dieselfde taalfamilies is (met ander woorde, waar die tale nabyverwant is). Ontwikkeling wat reeds in een van die tale gedoen is kan gebruik word om ontwikkeling vir die nabyverwante taal te bespoedig. ŉ Voorbeeld hiervan is die outomatiese annotering van dokumente vir afrigtingsdata. Die kwaliteit van die annotasie is nie goed genoeg om direk gebruik te word as afrigtingsdata nie. Die foute sal egter deur menslike insette reggemaak moet word, maar die proses sal vinniger verloop. Hierdie benadering kan ook gebruik word in skoenlussteekproefneming (“bootstrapping”) benaderings om vinnig ŉ werkende prototipe daar te stel en dan die prototipe iteratief te verbeter.

Ten slotte, die genreklassifikasiesisteme wat in hierdie studie uiteengesit is, is geïmplementeer in ŉ projek vir die Departement van Kuns en Kultuur van die Suid-Afrikaanse regering wat onderneem is deur Trifonius, met CTexT® en die Universiteit van Antwerpen as medewerkers. Die projek het ten doel gehad om die daarstelling van genreklassifikasiesisteme vir tien van die inheemse Suid-Afrikaanse tale te bewerkstellig en internasionale samewerking, met betrekking tot natuurliketaal-prosesseringsnavorsing, te bevorder. Die projekuitsette bestaan uit die volgende:

- ’n ondersoek na masjienleerbenaderings en ontologieë vir die ontwikkeling van genreklassifikasiesisteme vir hulpbronskaars tale;
- die ontwikkeling van die nodige hulpbronne, sowel as die genre-klassifikasiesisteme as ’n kerntegnologie;
- die implementering van die kerntegnologieë (genreklassifikasiesisteme) in ’n webgebaseerde demonstrasie waar gebruikers ’n lêer of ’n url na ’n weblêer kan verskaf om volgens genre geklassifiseer te word; en
- ’n opleidingsessie deur ’n internasionale spreker te reël waar daar deur plaaslike navorsers uit internasionale kennis gebaat kan word.

Hierdie hulpbronne, webdemonstrasie en verwante navorsingsuitsette is beskikbaar by www.trifonius.co.za.

5.2. VOORUITSKOUING

Gegewe die uitslae van hierdie navorsing is daar moontlikhede vir verdere navorsing, wat in die toekoms uitgevoer kan word. Onderwerpe wat in hierdie studie geïdentifiseer is wat aandag behoort te geniet is soos volg:

- Uit die resultate vir die ontwikkeling van genreklassifikasiesisteme vir hulpbronskaars tale, blyk die sisteme goed te funksioneer in ’n gekontroleerde eksperimentele omgewing. Hierdie sisteme kan egter aan verdere eksperimentering onderwerp word om die regtewêreldmoontlikhede vir hierdie sisteme te evalueer, sy dit met 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, bv. 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 sisteme kan verbeter word deur die hoeveelheid beskikbare data, sowel as die kwaliteit daarvan te bevorder. Aandag behoort hier spesifiek gegee te word aan die samestelling van die verskillende afrigtingstelle. Die 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.
- Verdere navorsing oor die moontlikheid van tegnologieherwinning vir nabyverwante tale kan uitgevoer word. Moontlikhede vir die oplossing van die woordeskatoorvleueling (sien 4.2.1.2) kan ondersoek word deur selektief om te gaan met die terme wat ingesluit word by die afrigtingsdata. Dit kan moontlik gedoen word deur funksiewoorde, ensovoorts, uit die afrigtingsdata te verwyder en dan net woorde wat eie is tot 'n spesifieke genreklas in te sluit. Die verskillende vertalingstrategieë wat in afdeling 4.2 genoem word kan ook ondersoek word om vas te stel watter benadering beter werk.
- 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.

Ten slotte: Die ontwikkeling van 'n outomatiese genreklassifikasiesisteem vir 'n hulpbronskaars taal is 'n komplekse proses waar daar 'n groot hoeveelheid veranderlikes, wat verder bemoeilik word deur 'n gebrek aan hulpbronne, in gedagte gehou moet word. In hierdie navorsing word daar 'n benadering geïdentifiseer wat hierdie veranderlikes saamvat en uiteindelik 'n suksesvolle, werkende genreklassifikasiesisteem daarstel vir Afrikaans. Hierdie studie is van nut vir ander hulpbronskaars tale en ten spyte van die moontlike probleme wat geopper word kan hierdie benadering tot 'n redelik suksesvolle mate oorgedra word na ander hulpbronskaars tale, spesifiek na die ander inheemse Suid-Afrikaanse tale. Die gebruik van nabyverwante tale (soos in die verskillende taalfamilies in Suid-Afrika) as 'n manier om die ontwikkeling van genreklassifikasiesisteme te bespoedig, blyk ook suksesvol te wees en kan gebruik word om onderontwikkelde tale te bemagtig deur dit te verryk met die ontwikkeling van nuwe tegnologieë.

BIBLIOGRAFIE

[1] Amor, N.B., Benferhat, S., Elouedi, Z., "Naive Bayes vs. Decision trees in Intrusion Detection Systems," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 420–424, 2004.

[2] Bel, N., Koster, C., Villegas, M., "Cross-Lingual Text Categorization," Research and Advanced Technology for Digital Libraries, 2769, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, 2003, pp. 126-139.

[3] Biber, D., "Variation Across Speech and Writing," Cambridge University Press: Cambridge, UK, 1988.

[4] Bühler, K., "Sprachtheorie: Die Darstellungsfunktion der Sprache," Verlag von Gustav Fischer, Jena, Germany, 1934.

[5] Cardinaels, K., Meire, M., Duval, E., "Automating Metadata Generation: The Simple Indexing Interface," in *Proceedings of WWW 2005: 14th international conference on World Wide Web*, ACM Press, pp. 548-556, 2000.

[6] Carlucci, L., Case, J. "On the Necessity of U-Shaped Learning," URL:"http://www.eecis.udel.edu/~case/papers/pr2.pdf". [Date of access: 2012-10-27]

[7] Chan, H., Y., Rosenfeld, R., "Discriminative Pronunciation Learning for Speech Recognition for Resource Scarce Languages," in *Proceedings of the Association for Computing Macinery Annual Symposium on Computing for Development*, Atlanta, USA, 2012.

[8] Cherkassky, V., Yunqian, M., "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression," Neural networks 17.1, 2004, pp. 113-126.

[9] Cocks, J., Keegan, T., "A Word-Based Approach for Diacritic Restoration in Maori," in *Proceedings of the Australasian Language Technology Association Workshop*, Canaberra, Australia, pp. 126-130, 2011.

[10] Cohen, W. W., "Fast Effective Rule Induction," in *Proceedings of the Twelfth International Conference on Machine Learning*, Lake Tahoe, California, 1995.

[11] Daelemans, W., Zavrel, J., Van Der Sloot, K., Van Den Bosch, A., "TiMBL: Tilburg Memory Based Learner, version 5.1, reference guide," ILK ResearchGroup Technical Report Series no. 04-02, 2004.

[12] Finn, A., Kushmerick, N., "Learning to Classify Documents According to Genre," Journal of the American Society for Information Science and Technology (JASIST), 7, 2006.

[13] Francis, W.N., Kucera, H., "Brown Corpus Manual - Revised and Amplified," Deptartment of Linguistics, Brown University, Providence, RI, USA, 1979.

[14] Goller, C., Löning, J., Will, T., Wolff, W., "Automatic Document Classification: A Thorough Evaluation of Various Methods," in *Proceedings of the Internationales Symposium für Informationswissenschaft*, Informationskompetenz - Basiskompetenz in der Informationsgesellschaft, Proceedings 7, pp. 145-162, 2000.

[15] Grover, A.S., Calteaux, K., van Huyssteen, G.B., Pretorius, M., "An overview of HLTs for South African Bantu languages," in *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, ACM, pp. 370-375, 2010.

[16] Grover, A.S., van Huyssteen, G.B., Pretorius, M.W., "The South African Human Language Technology Audit," Language Resources and Evaluation, Springer, 2011, pp. 271-288.

[17] Gu, B., Hu, F., Liu, H., "Modelling Classification Performance for Large Data Sets: An Empirical Study," in *Proceedings of Advances in web-age information management: Second International Conference*, Xi'an, China, pp. 317-328, 2001.

[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., "The WEKA Data Mining Software: An Update," SIGKDD Explorations, 11,1, 2009.

[19] He, H., Garcia, E., "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, 21, 9, 2009, pp. 1263-1284.

[20] Hsu, C.W., Chang, C.C., Lin, C.J., "A Practical Guide to Support Vector Classification," Technical Report, Department of Computer Science, National Taiwan University, 2003. URL:"http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf" [Date of access: 2012-10-27]

[21] Instituut vir Nederlandse Leksikologie (INL), "Parole Corpus," URL: http://parole.inl.nl/html/main_info_dutch.html, 2005. [Date of access: 2010-09-01]

[22] Keller, F., Lapata, M., Ourioupina, O., "Using the Web to Overcome Data Sparseness," in *Proceedings of the ACL-02 conference on Empirical methods in*

natural language processing, Association for Computational Linguistics, pp. 230-237, 2002.

[23] Kessler, B., Nunberg, G., And Schütze, H., "Automatic Detection of Text Genre," in *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics,* Madrid, Spain, pp. 32–38, 1997.

[24] Khan, A., Baharudin, B., Lee, L.H., Khan, K., "A Review of Machine Learning Algorithms for Text-Documents Classification," Journal of Advances in Information Technology, North America, 2010, pp. 10-11.

[25] Kim, Y. And Ross, S., "Detecting Family Resemblance: Automated Genre Classification," in *Proceedings of the 20th International CODATA Conference*, Beijing, CODATA Data Science Journal, 6, pp. 172-183, 2006.

[26] Liang, A., "Rotten Tomatoes: Sentiment Classification in Movie Reviews," Unpublished, 2006.

[27] Lim, C. S., Lee, K. J., Kim, G. C., "Multiple Sets of Features for Automatic Genre Classification of Web Documents," Information processing and management, 41, 5, 2005, pp. 1263-1276.

[28] Liu, H., Dougherty, E.R., Dy, J.G., Torkkola, K., Tuv, E., Peng, H., Ding, C., Long, F., Berens, M., Parsons, L., Zhao, Z., Yu, L., Forman, G. "Evolving feature selection," Intelligent Systems, IEEE, 20, 6, 2005, pp. 64-76.

[29] Manning, C.D., Prabhakar, R., Schütze, H., "An Introduction to Information Retrieval," Cambridge University Press, Cambridge, England, 2009, pp. 117-119; 253-285.

[30] Mccallum, A., Nigam, K., "A Comparison of Event Models for Naive Bayes Text Classification," in *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization,* Tech. Rep. WS-98-05, AAAI Press, pp. 41-48, 1998.

[31] Mogadala, A., Varma, V., "Retrieval Approach to Extract Opinions about People from Resource Scarce Language News Articles," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, Beijing, China, pp. 4:1-8, 2012.

[32] Morgan, W., "Statistical Hypothesis Tests for NLP or: Approximate Randomization for Fun and Profit," URL: "http://masanjin.net/sigtest.pdf". [Date of access: 2012-10-24]

[33] Peché, M., Davel, M., Barnard, E., "Phonotactic Spoken Language Identification with Limited Training Data," in *Proceedings of the Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, pp. 1537-1540, 2007.

[34] Peng, F., Schuurmans, D., "Combining Naive Bayes and n-Gram Language Models for Text Classification," in *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR-03)*, pp. 335–350, 2003.

[35] Petrenz, P., "Assessing Approaches to Genre Classification," M.Sc. Dissertation, School of Informatics, University of Edinburgh, Edinburgh, 2009.

[36] Petrenz, P., "Cross-Lingual Genre Classification," in *Proceedings of the EACL 2012 Student Research Workshop*, Avignon, France, pp. 11-21, 2012.

[37] Petrenz, P., Webber, B., "Stable Classification of Text Genres," Computational Linguistics, 37, 2011, pp. 385–393.

[38] Pienaar, W., Snyman, D.P., "Spelling Checker-based Language Identification for the Eleven Official South African Languages," in *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, pp. 213-217, 2010.

[39] Pilon, S., Van Huyssteen, G.B., Augustinus, L., "Converting Afrikaans to Dutch for Technology Recycling," in *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, pp. 219-224 , 2010.

[40] Prettenhofer, P., Stein, B., "Crosslanguage Text Classification Using Structural Correspondence Learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, pp. 1118-1127, 2010.

[41] Rafferty, W., Salffner, S., "Once Upon A Time in a Country Far, Far Away. Ritualisation and Ritualised Communication in African Orature," Unpublished term paper, Universität Bielefeld, Fachbereich Anglistik, 2002, p. 8.

[42] Reşceanu, A., "Notes for the English Linguistics Seminar," To appear in *Introduction to General Linguistics*, p. 155. URL: "http://www.scribd.com/doc /13384840/36/Buhler-s-organon-model" [Date of access: 2012-10-16]

[43] Resnik, P., Smith, N. A., "The Web as a Parallel Corpus," Computational Linguistics, 29, 3, pp. 349-380, 2003.

[44] Santini, M., "Automatic Genre Identification: Towards A Flexible Classification Scheme," Paper presented at the BCS IRSG Symposium: Future Directions in Information Access, Glasgow, Scotland, 2007.

[45] Santini, M., "Cross-Testing a Genre Classification Model," in *Proceedings of the 2nd Swedish Language Technology Conference (SLTC*), pp. 7-8, 2008.

[46] Sharoff, S., "Classifying Web Corpora into Domain and Genre Using Automatic Feature Identification," in *Proceedings of Web as Corpus Workshop*, Louvain-la-Neuve, 2007.

[47] Sharoff, S., Wu, Z., Markert, K., "The Web library of Babel: Evaluating Genre Collections," in *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*, Malta, 2010.

[48] Smucker, M.D., Allan, J., Carterette, B., "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation," in *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp. 623-632, 2007.

[49] Swales, J.M., "Genre Analysis: English in Academic and Research Settings," Cambridge: Cambridge University Press, Cambridge, England, 1990, p. 260.

[50] Van Huyssteen, G.B. & Pilon, S., "Rule-Based Conversion of Closely-Related Languages: A Dutch to Afrikaans Convertor," in *Proceedings of the 20th Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, pp. 23-28, 2009.

[51] Van Noord, G., "Huge Parsed Corpora in LASSY," in *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, Utrecht, The Netherlands, pp. 115-126, 2009.

[52] Van Zyl, C., "A Script Development Model for the Creation of Computer Games," M.A. dissertation, North-West University, Potchefstroom, 2008, pp.13-14.

[53] Vapnik, N, V., "The Nature of Statistical Learning Theory," Springer-Verlag, NewYork, 1995, pp. 159-164.

[54] Vargas Sierra, C., "A Pragmatic Model of Text Classification for the Compilation of Special-purpose Corpora," Thistles, A homage to Brian Hughes, Essays in Memoriam, 2, 2005, pp. 295-315.

[55] Wachsmuth, B., Bunja, K., "Back to the Roots of Genres: Text Classification by Language Function," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, pp. 632-640, 2011.

[56] Wurst, M., "The Word Vector Tool: User Guide, Operator Reference, Developer Tutorial," URL:"http://www-ai.cs.uni-dortmund.de/SOFTWARE/WVTOOL/doc/wvtool-1.0.pdf", 2007. [Date of access: 2012-08-26]

[57] Yang, Y., Pedersen, J., "A Comparative Study on Feature Selection in Text Categorization," in *Proceedings of the International Conference on Machine Learning (ICML)*, Nashville, TN, USA, pp. 412-420, 1997.

[58] Yates, J. And Orlikowski,W. J., "Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media," Acad. Manage. Rev. 17, pp. 299–326, 1992.

[59] Yeh, A., "More Accurate Tests for the Statistical Significance of Result Differences," in *Proceedings of the 18th International Conference on Computational Linguistics*, 2, pp. 947-953, 2000.

[60] Yi-Hsing, C., Hsiu-Yi, H., "An Automatic Document Classifier System Based on Naïve Bayes Classifier and Ontology," in *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics,* Kunming, 2008.

BYLAAG A: Genreklasse van PAROLE-korpus

PAROLE BROAD GENRE CLASSIFICATION

Deze classificatie is gebaseerd op de gedetailleerde genreclassificatie van het Deense woordenboek, zoals gepubliceerd in: Norling-Christensen, O. (1995) *Design and composition of reusable harmonised written language reference corpora for European languages.* Draft report for the PAROLE project.

N/A: not applicable/mixed/unknown/not identified

ADVertising

DIScussion: absorbs former CON, DEB, SPE

CONversation: written as well as spoken, including interviews

DEBate: including i.a. "letters to the editor"

SPEech including i.a. sermons, lectures and parliamentary speeches. NB: this is genre, not medium. A book of sermons, or the final, edited and printed official version of the parliamentary proceedings, go into this category.

FEAture: article in newspaper etc. which does not belong to news or another, more specific genre; reviews, radio/tv magazines, etc.

FICtion: including faction and e.g. comic strips, groups DRAma, ENTertainment and POEtry.

DRAma: including film manuscripts and e.g. tv-series

ENTertainment: including i.a. childrens'and youth pages, jokes, and games

POEtry: including poems and song lyrics

INFormation: including folders and leaflets from e.g. the authorities; posters; signs. (NB: this is not the category E.3.2.1 of Sinclair/Ball (1995:14) defined as "mainly reference compendia"; like other refence works these would belong to the category instruction); includes NEWs

NEWs: the main genre for newspaper texts, as well as similar programs in radio and television

INStruction: including reference and text books, but also that kind of correspondence column in e.g. magazines, where readers' questions are answered by specialists in tax, gardening, health etc.

NON-fiction: absorbs the former category BIOgraphy, as well as the following categories removed from FEAture: aphorism, calendar, cultural debate, report, school essay, student essay

BIOgraphy: including obituaries and autobiographies

OFFicial text: including laws, government circulars, official announcements, business correspondence

PRIvate text: like diaries and private letter

GEDETAILLEERDE GENRECLASSIFICATIE

Dit is de gedetailleerde genrelijst die gebruikt is door het Deense corpus en waar de PAROLE Broad Genre Classification op gebaseerd is. Deze lijst geeft een idee van welke subcategorieën onder de verschillende hoofdcategorieën zouden vallen en kan gebruikt worden als model voor andere corpusbouwers.

Broad genre	**Genre (detailed), with Danish abbreviations**	
N/A	n/a:	not applicable/unknown
ADV	ann:	advertisement
ADV	rekl:	advertising
ADV	til:	bargain paper
ADV	bro:	brochure
ADV	kat:	catalogue
ADV	rekt:	printed advertisement
ADV	foro:	prior publicity
DIS	sam:	conversation
DIS	kul:	cultural debate
DIS	deb:	debate
DIS	dia:	dialogue
DIS	grui:	group interview
DIS	grus:	group conversation
DIS	int:	interview
DIS	laes:	letter to the editor
DIS	fol:	parliamentary speech
DIS	telp:	phone-in programme
DIS	oplg:	presentation
FEA	art:	article
FEA	cau:	causerie
FEA	klu:	column
FEA	ess:	essay
FEA	fea:	feature (radio/tv)
FEA	kro:	feature article
FEA	mont:	montage
FEA	mus:	music programme
FEA	pro:	programme
FEA	cit:	quotation
FEA	mag:	regular programme

FEA	anm:	review
FIC	boer:	children's page
FIC	teg:	comic strip, comic book
FIC	undh:	entertainment
FIC	eve:	fairy tale
FIC	film:	film
FIC	spil:	game
FIC	hor:	horoscope
FIC	hum:	humour
FIC	vit:	joke
FIC	rom:	novel
FIC	lej:	occasional song
FIC	dra:	play
FIC	digt:	poem
FIC	ser:	series (esp. television)
FIC (cont.)	nov:	short story
FIC (cont'd)	san:	song lyrics
FIC	fort:	story
FIC	ung:	youth section
INF	rep:	(on-the-spot) report
INF	ref:	account, minutes
INF	forn:	association publication
INF	bag:	background article
INF	bil:	caption
INF	fori:	consumer tips
INF	akt:	current events programme
INF	dokp:	documentary
INF	reda:	editorial matters
INF	led:	editorial
INF	eng:	enquiry, poll
INF	fold:	folder
INF	bes:	informal message
INF	inf:	information leaflet
INF	infh:	information booklet
INF	ford:	lecture
INF	lis:	list
INF	med:	message
INF	nyha:	news article
INF	nyhu:	news broadcast

INF	not:	notice
INF	pje:	pamphlet
INF	pla:	poster
INF	ski:	sign
INF	pet:	small news item
INF	tra:	sports broadcast (radio/tv)
INS	bib:	bibliography
INS	brsp:	corr. column questions
INS	brsv:	corr. column answers
INS	brek:	correspondence column
INS	vej:	guide
INS	haan:	handbook
INS	mono:	monograph
INS	opsk:	recipe
INS	fag:	reference book
INS	skob:	school textbook
INS	undv:	teaching
INS	laer:	textbook
INS	rej:	travel book
INS	opsl:	work of reference
NON	nav:	'names'
NON	afo:	aphorism
NON	sel:	autobiography
NON	bio:	biography
NON	kal:	calendar
NON	nek:	obituary
NON	fes:	principal speech
NON	por:	profile
NON	rap:	report
NON	prae:	sermon
NON	skos:	school essay
NON	lev:	short biography
NON	opg:	student essay
OFF	bek:	announcement
OFF	forb:	business letter
OFF	kon:	contract
OFF	doku:	document
OFF	bla:	form
OFF	cir:	government circular

OFF	lov:	law
OFF	ret:	legal record
OFF	skr:	official letter
OFF	bet:	parl. report, white paper
PRI	dag:	diary
PRI	brev:	private letter