Het gebruik van sentimentoriëntatie analyse bij het voorspellen van stemmingen in weblogberichten

Vincent Oord

25 juni 2006

Samenvatting

In dit onderzoek is gekeken naar de invloed van sentimentanalyse bij het voorspellen van de stemmingen van weweblogberichten. Hiervoor is gebruik gemaakt van een methode van Kamps et al. [6] om de sentimentele oriëntatie van adjectieven te bepalen en deze zijn daarna toegepast op een corpus van weblogberichten uit LiveJournal. Er is op een grafische manier gekeken naar de data, maar er zijn ook classificatie-experimenten uitgevoerd. Uiteindelijk blijkt het gebruik van sentimentfeatures alleen een niet afdoende methode om de stemming van een weblogbericht nauwkeurig te kunnen voorspellen.

Inhoudsopgave

1 Inleiding

In dit onderzoek staat het voorspellen van de stemmingen van weblogberichten door middel van sentimentanalyse centraal. LiveJournal[1] is een Engelstalige website die aan haar gebruikers de mogelijkheid biedt een weblog te beginnen: een website waarop berichten geplaatst kunnen worden en waar anderen vervolgens op kunnen reageren. Als extra optie biedt LiveJournal de schrijver van een bericht om diens stemming aan een bericht te koppelen, zodat anderen dit kunnen lezen. De gebruiker classificeert hiermee zijn eigen tekst als zijnde de representatie van diens huidige stemming. Dit biedt mogelijkheden voor tekstclassificatie, zoals het voorspellen van de stemming van een bericht.

Eerder onderzoek van Mishne [8], op hetzelfde corpus, leverde wisselende resultaten op. Hij gebruikte onder andere woordfrequenties en het gebruik van speciale tekens om de stemming van webblogberichten te voorspellen, maar dit bleek niet in alle gevallen een betrouwbare methode. In zijn artikel noemt hij al de mogelijkheid van sentimentanalyse als mogelijkheid om zijn resultaten te verbeteren en deze mogelijkheid wordt in dit onderzoek nader bekeken.

Sentimentanalyse is een methode van tekstclassificatie die kijkt naar de sentimentele lading van adjectieven om zodoende iets te kunnen zeggen over de sentimentele oriëntatie van een stuk tekst. Zo krijgen woorden met een negatieve lading als *bad* en *annoyed* een negatieve score en woorden met een positieve lading, zoals *happy* en *excited* een positieve score. Teksten met veel negatieve termen zullen dus in hoge mate negatief georiënteerd zijn, terwijl teksten met veel positieve woorden in hoge mate positief georiënteerd zullen zijn. Omdat de stemmingen die LiveJournal de gebruiker biedt om aan zijn of haar teksten te koppelen ook deze verdeling van positief en negatief kennen is het interessant om te kijken of er op basis van de sentimentele oriëntatie

[1] http://livejournal.com

van een bericht de corresponderende stemming kan worden voorspeld. Met andere woorden, zal een bericht met een erg negatieve sentimentele oriëntatie ook voorzien zijn van een erg negatieve stemming?

Om de relatie tussen verschillende stemmingen en de sentimentele waarde van het daaraan gekoppelde tekstbericht ook op een andere manier te benaderen, is gekeken naar de distributie van de sentimentscores binnen een bepaalde stemming. Dit wil zeggen dat er gekeken is of de berichten binnen een bepaalde stemming gelijkwaardig zijn verspreid binnen een bepaald domein van scores, of dat de scores juist totaal willekeurig zijn. Dit zou ook nieuwe inzichten kunnen bieden bij het voorspellen van stemmingen middels sentimentanalyse.

Doel van dit onderzoek is dan ook om te kijken of door het gebruiken van sentimentanalyse nauwkeurigere voorspellingen kunnen worden gedaan met betrekking tot de stemming van weblogberichten. Hiertoe worden eerst de sentimentwaardes van specifieke woorden in weblogberichten gemeten en wordt vervolgens geprobeerd deze te koppelen aan stemmingwaardes van die berichten. De resultaten daarvan worden gebruikt als input voor een stemming-voorspeller van weblogberichten, vergelijkbaar met de manier waarop stemmingen van weblogberichten worden voorspeld door Mishne in [8]. De onderzoeksvraag luidt dan ook als volgt:

> *In welke mate kan het voorspellen van stemmingen van weblogberichten worden verbeterd door het gebruik van sentimentoriëntatie?*

Deze vraag is op te delen in een aantal deelvragen:

1. Valt er iets op te maken uit de distributie van berichten bij een bepaalde stemming?
2. Is de keus voor een bepaald classificatiealgoritme van invloed op de classificatieresultaten?
3. Is de hoeveelheid trainingsdata van invloed op de classificatieresultaten?
4. Kunnen de resultaten nog verbeterd worden door gebruik te maken van andere features?

In het volgende hoofdstuk wordt verder in gegaan op de achtergronden van dit onderzoek en worden gerelateerde onderzoeken besproken. In het derde hoofdstuk gaat het over de opzet van de experimenten die zijn uitgevoerd en in het vierde hoofdstuk staan de resultaten hiervan. Daarna worden de resultaten van alle experimenten nog kort besproken in hoofdstuk 5. Uiteindelijk wordt dit verslag afgesloten met een algehele conclusie in hoofdstuk 6, waarin eerst de deelvragen één voor één worden beantwoord, om uiteindelijk antwoord te geven op de onderzoeksvraag.

2 Achtergrond

In dit hoofdstuk wordt achtergrondinformatie gegeven over het onderwerp van dit onderzoek. In de eerste sectie wordt uitgelegd wat tekstclassificatie precies is, waarbij de nadruk wordt gelegd op tekstclassificatie middels sentimentanalyse. Daarnaast worden er ook verwijzingen gegeven naar ander onderzoek in deze richting. De tweede sectie biedt een overzicht van de verschillende benaderingen en methoden binnen sentiment-analyse en geeft aan welke methode er in dit onderzoek is gebruikt, en waarom. Als laatste sectie is er een samenvatting van het onderzoek van Mishne [8], die met hetzelfde corpus onderzoek heeft gedaan naar het classificeren van stemmingen en waarvan de resultaten uiteindelijk worden vergeleken met de resultaten van dit onderzoek.

2.1 Tekstclassificatie

Onder tekstclassificatie wordt het proces waarbij aan een tekstdocument een bepaalde semantische klasse wordt toegekend verstaan. Dit kan zowel handmatig als automatisch gebeuren; in dit onderzoek kijken we echter naar het geautomatiseerde proces. Een toepassing van tekstclassificatie is bijvoorbeeld een spamfilter, dat op basis van, onder andere, de tekst in een e-mail beslist of een mailtje spam is of niet. Ook het tag-systeem dat tegenwoordig populair is op websites als Flickr[2] en Technorati[3], waarmee door een gebruiker bij een foto (Flickr) of een weblogbericht (Technorati) wordt aangegeven in welke categorie deze hoort, is een vorm van (tekst)classificatie, maar dan op een handmatige manier.

In dit onderzoek worden weblogberichten geclassificeerd, waarbij de klasses bestaan uit verschillende stemmingen die een persoon kan hebben. Dit is een verschil met het classificatiesysteem van systemen als Technorati, waarbij het gaat om objectieve zaken, zoals het onderwerp van een bericht (oftewel *topicality* - een ander veld in de tekstclassificatie). De focus in dit onderzoek ligt bij sentimentele classificatie, m.a.w. het bepalen of een tekst een positieve of een negatieve tendens heeft, waarna we dit proberen te koppelen aan een stemming.

De taak van tekstclassificatie is op te delen in een drietal subtaken. Ten eerste moeten er van de te classificeren teksten representaties gemaakt worden, zodat een classificatiealgoritme hiermee overweg kan. Deze subtaak wordt besproken in hoofdstuk 3. Daarna moet het algoritme geleerd worden naar bepaalde patronen in deze representaties te zoeken, op basis waarvan een classificatie gemaakt kan worden. Als laatste is er de evaluatiefase waarbij

[2] http://flickr.com
[3] http://technorati.com

het algoritme een aantal onbekende instanties dient te classificeren. Deze laatste twee subtaken worden uitvoerig besproken in hoofdstuk 4.

2.2 Onderzoek Mishne

Aanleiding tot dit onderzoek vormt het onderzoek van Gilad Mishne ([8]), waarin is geprobeerd individuele weblogberichten uit het LiveJournal-corpus (hierna: LJ-corpus) te classificeren middels een *bag-of-words* methode. Conclusie van dit onderzoek is dat het classificeren van individuele berichten een lastige taak is: hoewel op sommige stemmingen een redelijke nauwkeurigheid van tussen de 60% en 65% valt te behalen, komt het gros van de stemmingen niet veel boven de 50% nauwkeurigheid uit, waarmee het nauwelijks beter presteerde dan de op hetzelfde percentage gestelde *baseline*. Ander onderzoek van Mishne en De Rijke toont aan dat het voorspellen van een verzameling weblogberichten beter afgaat [9].

Ondanks deze conclusie van Mishne is toch geprobeerd te kijken of er niet een beter resultaat kan worden neergezet, ditmaal met een andere methode van sentimentclassificatie, namelijk sentimentoriëntatie.

2.3 Sentimentclassificatie

Bij het uitvoeren van sentimentele classificatie zijn er drie benaderingen te onderscheiden. Ten eerste is er de *bag-of-words* methode, zoals onder andere gehanteerd in [2, 6, 9, 14]. Hierbij worden alle woorden in een tekstdocument bijeengegooid en wordt er bijvoorbeeld op basis van woordfrequentie of woord n-grammen (samenstellingen van n-aantal woorden) een voorspelling over de juiste klasse gemaakt. Een tweede methode is die van de sentimentele oriëntatie, waarbij individuele woorden uit een bericht worden gescoord op een schaal van bijvoorbeeld goed tot slecht, waarna voor het hele bericht een totaal wordt berekend. Whitelaw et al. introduceren in [13] een derde, nieuwe benadering voor sentimentele classificatie: die van *appraisal-groups*. Dit is een complexe methode, waarbij op basis van *Appraisal Theory* gekeken wordt naar zinnen in teksten waarin iemand (de beoordelaar) een waardeoordeel (de attitude) geeft over iets (het beoordeelde). Er wordt hierbij ook gekeken naar context, dus bijvoorbeeld de woorden voor en na een specfiek woord. Vergelijk hierbij de frasen *happy*, *very happy* en *not very happy*, met elk een andere betekenis, die niet zou zijn opgemerkt als er alleen naar *happy* gekeken wordt. Dit rekening houden met de context waarin woorden in een text voorkomen is ook al eerder onderzocht, door Wilson et al. in [14].

Omdat Mishne in [8] gebruik maakt van de *bag-of-words* methode voor zijn sentimentclassificatie, is in dit onderzoek er voor gekozen om een andere benadering te kiezen, om te kijken of dit een positief effect heeft op de

classificatie. Hoewel er met de Appraisal-Group methode redelijk goede resultaten zijn geboekt, is er niet voor gekozen om met deze methode aan de slag te gaan. Het betreft een vrij nieuwe methode en kent dus (nog) weinig wetenschappelijke ondersteuning, waardoor er nog mogelijk kinderziekten in deze methode kunnen zitten. In plaats daarvan is gekozen voor de aanpak via sentimentele oriëntatie, omdat hier enkele veelbelovende methoden voor zijn gevonden.

2.4 Sentimentele oriëntatie

Om sentimentanalyse middels sentimentele oriëntatie te kunnen toepassen, zijn er diverse methoden beschikbaar. In deze paragraaf zullen een viertal methodes die in de literatuur gevonden zijn, kort worden besproken, voor een uitgebreidere toelichting op de methodes zie [3, 4, 6, 12].

2.4.1 De methode Hatzivassiloglou & McKeown [4]

In deze methode wordt geprobeerd de oriëntatie van adjectieven te bepalen door te kijken naar adjectieven die in paren voorkomen in teksten, verbonden door voegwoorden als *and*, *or* en *but*. De onderliggende gedachte hierbij is dat *and* woorden van gelijke oriëntatie verbindt terwijl *but* juist woorden van een tegenovergestelde oriëntatie verbindt. Hatzivassiloglou & McKeown halen eerst alle samenvoegingen van adjectieven uit een corpus, waarna een supervised leeralgoritme verschillende karakteristieken van de paren gebruikt om te bepalen of de paren van gelijke of verschillende sentimentele oriëntatie zijn. Dit resulteert in een graaf, waarin de adjectieven gerepresenteerd worden als nodes en links tussen deze nodes laten zien of de oriëntatie tussen twee nodes gelijk of verschillend is. Deze graaf wordt door een ander algoritme, een clusteringsalgoritme, doorlopen om twee clusters te genereren met woorden van een gelijke (maar nog onbekende) oriëntatie. Op basis van het gegeven dat *semantically unmarked*[4] adjectieven vaker voorkomen in teksten en dat dit soort adjectieven in de meeste gevallen een positieve oriëntatie heeft[5], wordt het cluster met de grootste gemiddelde woordfrequentie gelabeld als zijnde de woorden met positieve sentimentele oriëntatie. Het nadeel van deze methode is dat er alleen gebruik gemaakt kan worden van adjectieven.

[4]semantically unmarked: als een adjectief geen oordeel geeft over de eigenschap waarna hij verwijst. Vergelijk 'Hoe lang is hij' en 'Hoe klein is hij'. Hierbij is 'lang' semantically unmarked en 'klein' niet.

[5]Dit lijkt taalafhankelijk te zijn. Recent onderzoek aan de UvA heeft aangetoond dat in het Nederlands negatieve termen vaker voorkomen

2.4.2 De methode Turney & Littman [12]

Turney en Littman hebben een methode ontwikkeld waarbij de sentimentele oriëntatie van een woord wordt berekend op basis van de semantische associatie van dat woord met woorden uit een *seed-set*. Dit houdt in dat er wordt gekeken naar de sterkte van de associatie van een woord met woorden uit een set positieve woorden minus de sterkte van de associatie van dat woord met woorden uit een set negatieve woorden. Als de resulterende SO-A (Semantic Orientation by Association) score positief is, betekent dit dat het betreffende woord een positieve oriëntatie heeft, als de score negatief is, is de oriëntatie van het woord juist negatief. Vervolgens worden twee statistische methodes gebruikt om deze methode te evalueren: Pointwise Mutual Information (PMI) en Latent Semantic Analysis (LSA). De laatste methode is helaas niet geschikt voor erg grote corpora, maar met beide methodes worden goede resultaten behaald. Het voordeel van de methode van Turney & Litmann is dat het toepasbaar is op verschillende soorten woorden en dus niet alleen op adjectieven.

2.4.3 De methode Esuli & Sebastiani [3]

Bij deze methode wordt ook gebruik gemaakt van een seed-set op basis van een positieve en negatieve categorie. Vervolgens wordt vanuit deze *seed-set* in een iteratief proces een thesaurus doorlopen en worden woorden die geassocieerd zijn met termen uit de seed-set toegevoegd aan de corresponderende categorie. Het uiteindelijke resultaat vormt de trainingsset waarmee de classifier getraind wordt. Daarna wordt met alle termen uit de twee uitgebreide *seed-sets* een *glossary* (definitie lijst) doorlopen waarbij van elke term uit de *seed-set* de definitie zoals die in de glossary staat wordt opgezocht en vertaald naar een vector. Een binaire tekstclassifier probeert vervolgens de termen in deze testset juist te classificeren.

2.4.4 De methode Kamps et al. [6]

In de methode van Kamps et al. wordt de oriëntatie van een woord bepaald door te kijken naar de relatieve afstand die synoniemen binnen WordNet[6] hebben ten opzichte van twee termen uit een *seed-set*, te weten *good* en *bad*. Als basis is hier de woordenlijst van Turney & Littman gebruikt en hier zijn vervolgens de woorden uit gehaald die volgens WordNet een synoniemrelatie hebben. Van deze termen is vervolgens de relatieve afstand tot de termen uit de *seed-set* berekend. Dit heeft als nadeel dat het aantal termen beperkt is tot de woorden die op de een of andere manier een synoniem relatie hebben tot *good* of *bad*.

[6] http://wordnet.princeton.edu/

2.5 Gekozen methode

Er is gekozen om de methode van Kamps et al. te gebruiken in dit onderzoek. De resultaten hiervan behoren niet tot de beste van de vier beschikbare methodes, maar zijn ook zeker niet slecht. Daarnaast heeft de methode als voordeel dat het onderzoek is uitgevoerd aan de UvA binnen dezelfde werkgroep als dit onderzoek, waardoor het eenvoudig was om de implementatie van deze methode te kunnen gebruiken. Hoewel er met deze methode alleen naar adjectieven wordt gekeken, hoeft dat niet als beperking gezien te worden: omdat adjectieven een specifieke eigenschap aan een zelfstandig naamwoord meegeven bepalen ze in hoge mate de sentimentele oriëntatie van een woord of tekst. Daarom zijn methodes die zich beperken tot adjectieven zeker bruikbaar.

De implementatie van Kamps et al. bestaat uit een woordenlijst met 5410 adjectieven en hun drie respectievelijke Osgood-scores. Deze scores zijn gebaseerd op het standaardwerk over het meten van emoties in teksten van Osgood et al. [10]. Hierin is aan mensen gevraagd woorden, frasen en teksten in te delen op verschillende schalen, bestaande uit verschillende paren van tegenovergestelde adjectieven als *active/passive* en *good/bad*. Hoewel er een redelijk aantal paren is onderzocht, blijkt dat er slechts drie paren het belangrijkst zijn, in de zin dat ze het best verschillen in oordelen van mensen kunnen uitleggen. Dit zijn *good/bad* (*evaluative* factor, kortweg EVA), *strong/weak* (*potency* factor, kortweg POT) en *active/passive* (*activity* factor, kortweg ACT). Dit zijn dan ook de drie scores die in de woordenlijst worden bijgehouden en op basis waarvan in dit onderzoek de sentimentele oriëntatie van een weblogbericht wordt bepaald.

Deze scores zijn waardes binnen het domein [-1,1], waarbij 1 een heel erg positieve oriëntatie aanduidt en -1 een heel erg negatieve oriëntatie. Deze scores zijn voor elk adjectief uitgerekend middels de volgende formules:

$$EVA(\omega) = \frac{d(\omega, bad)d(\omega, good)}{d(good, bad)}$$

$$POT(\omega) = \frac{d(\omega, weak)d(\omega, strong)}{d(weak, strong)}$$

$$ACT(\omega) = \frac{d(\omega, passive)d(\omega, active)}{d(passive, active)}$$

Waarbij de sentimentele oriëntatie van een woord ω wordt bepaald door het verschil in de afstand binnen WordNet van woord ω tot beide termen uit het paar, gedeeld door de afstand in WordNet tussen de twee termen zelf.

In hoofdstuk 3 wordt verder toegelicht hoe deze scores aan de berichten uit het LJ-corpus zijn toegekend.

```
<item>
  <guid>http://livejournal.com/users/psybstrd/141152.html</guid>
  <pubDate>Thu, 27 Jan 2005 20:08:17 GMT</pubDate>
  <title>Today is good...tomorrow will be better...</title>
  <author>cyric_316@hotmail.com</author>
  <link>http://livejournal.com/users/psybstrd/141152.html</link>
  <description>
    So today I am having a good day...
  </description>
  <comments>http://livejournal.com/users/psybstrd/141152.html</comments>
  <lj:music>Three Doors Down :: Let Me Go</lj:music>
  <lj:stemming>bouncy</lj:stemming>
</item>
```

Figuur 1: Een voorbeelditem uit het XML-bestand

3 Experimentele Opzet

Met de in hoofdstuk 2 gekozen methode wordt geprobeerd enkele onderzoeksvragen, zoals gesteld in hoofdstuk 1, te beantwoorden. Dit gebeurt aan de hand van een aantal visualisatie- en classificatie-experimenten. In dit hoofdstuk zal de opzet van de verschillende experimenten worden besproken. Een verslag van het uitvoeren van de feitelijke experimenten is te vinden in hoofdstuk 4.

3.1 Corpus

Het gebruikte corpus is hetzelfde als het corpus uit [8]: de LiveJournal-corpus. Dit is een corpus bestaande uit 815.494 weblogberichten van de site LiveJournal, een aanbieder van een gratis weblogdienst met een enorme gebruikersgemeenschap van enkele miljoenen mensen. LiveJournal biedt aan haar gebruikers de mogelijkheid om hun huidige stemming toe te voegen aan een bericht als ze dat op hun weblog plaatsen. Hiertoe kunnen ze kiezen uit een lijst van 132 vooraf gedefinieerde stemmingen, maar het is ook mogelijk om een eigen stemming in te voeren. Daarnaast kan de gebruiker er ook voor kiezen om helemaal geen stemming op te geven.

Het corpus bestaat uit al deze weblogberichten, bijbehorende stemmingen en andere gegevens in een groot XML-bestand van 1.7 gigabyte. Een voorbeeldbericht is te zien in figuur 1.

Tussen de `<item>`-tags staat de inhoud van een bericht. Dit bestaat uit een `<guid>`, waarin een URL is opgenomen die verwijst naar de locatie van het bericht op internet. Verder is er een `<pubdate>` te zien, waarin de datum van plaatsing van het bericht staat. Dit is niet per se de lokale tijd van de gebruiker, maar de tijd van de server waar LiveJournal op staat. Aangezien

de meeste gebruikers van LiveJournal echter Noord-Amerikaans zijn [9] komt dit vaak overeen. In de volgende velden, `<title>`, `<author>` en `<link>` staan respectievelijk de titel en auteur van het bericht en de permanente lokatie op het internet waar het bericht te vinden is. Het eigenlijke bericht staat tussen de `<description>` tags. Hier wordt later de sentimentanalyse op toegepast. Verder is alleen de `<lj:stemming>`-tag nog van belang in dit onderzoek: hier tussen staat de door de gebruiker aan het bericht toegekende stemming. De overige velden, `<comments>` en `<lj:music>` bevatten de reacties geplaatst op het bericht en het nummer waar de gebruiker tijdens het plaatsen van het bericht naar luistert. Deze twee velden zijn voor dit onderzoek echter niet van belang.

Een probleem van de classificatie van deze berichten is dat het toekennen van een klasse niet het werk is van een professional, die enige consistentie in acht neemt bij het classificeren van de teksten, maar vam een grote groep verschillende mensen die elk hun eigen interpretatie van de verschillende stemmingen hebben. Neem daarbij ook nog het feit dat gebruikers eigen stemmingen in kunnen vullen, waarvan de precieze betekenis helemaal onduidelijk kan zijn en het is duidelijk dat er sprake is van een lastig corpus. Aan de andere kant, stelt Mishne [8] kan dit ook tot een voordeel worden omgekeerd, aangezien er direct toegang is tot de gemoedstoestand van de gebruiker op het moment van plaatsen van het bericht, in plaats van een externe annoteur die dit achteraf doet.

Om het probleem van de zelf ingevoerde stemmingen te omzeilen is er voor gekozen deze stemmingen te negeren en alleen berichten uit het corpus te halen die zijn geclassificeerd volgens een van de 132 bekende LiveJournal stemmingen[7]. Ook berichten zonder stemming worden dus genegeerd. Vervolgens is uit de overgebleven reeks berichten de tekst gehaald, alsmede het aantal woorden van het bericht, de datum geconverteerd naar een UNIX timestamp en het bericht-id. Verder werd ook nog alle in de berichten aanwezige HTML-code verwijderd. Deze berichten zijn per stemming in losse bestanden opgeslagen, zodat later de sentimentwaarde van elk bericht afzonderlijk kon worden berekend. Dit resulteerde uiteindelijk in totaal in 508.180 berichten, met een totale grootte van 555 megabyte. Per stemming betekent dit gemiddeld ongeveer 3850 berichten, wat niet erg veel is om een betrouwbare classificatie op uit te voeren.

Gelukkig zijn de stemmingen niet gelijkmatig verdeeld over de berichten. Besloten is, analoog aan [8], te kiezen om alleen van de top 40 stemmingen sentimentwaarden te gaan berekenen en deze mee te nemen in het classificatieproces. Dit resulteerde in een set van stemmingen, met een totaal van 488.251 berichten, zoals te zien in tabel 1 (het percentage is op basis van het aantal berichten met een van de 132 LJ-stemmingen).

[7] Zie http://www.livejournal.com/stemminglist.bml

Stemming	*Aantal berichten*		*Stemming*	*Aantal berichten*	
amused	22757	(4,7%)	exhausted	6532	(1,3%)
tired	19098	(3,9%)	depressed	6165	(1,3%)
happy	15534	(3,2%)	crazy	6123	(1,3%)
cheerful	12290	(2,5%)	drained	5993	(1,2%)
bored	12179	(2,5%)	curious	5907	(1,2%)
accomplished	11403	(2,3%)	sad	5838	(1,2%)
sleepy	10799	(2,2%)	aggravated	5686	(1,2%)
content	10669	(2,2%)	ecstatic	5658	(1,2%)
excited	10521	(2,2%)	blank	5617	(1,2%)
contemplative	10240	(2,1%)	okay	5440	(1,1%)
blah	10192	(2,1%)	hungry	5264	(1,1%)
calm	9535	(2,0%)	cold	5220	(1,1%)
bouncy	9487	(1,9%)	creative	5134	(1,1%)
awake	9485	(1,9%)	hopeful	5059	(1,0%)
chipper	9048	(1,9%)	pissed off	4738	(1,0%)
confused	7798	(1,6%)	good	4734	(1,0%)
annoyed	7755	(1,6%)	thoughtful	4210	(0,9%)
sick	7457	(1,5%)	frustrated	4128	(0,8%)
busy	7342	(1,5%)	cranky	3979	(0,8%)
anxious	6779	(1,4%)	loved	3877	(0,8%)

Tabel 1: Meest voorkomende stemmingen in het corpus

3.2 Datarepresentatie

Het voorspellen van stemmingen van weblogberichten middels sentimentanalyse is een vorm van tekstclassificatie. Aangezien een classificatiealgoritme niets kan met de letterlijke tekst zoals die in het bericht staat, is het zaak een representatie van de tekst te maken. Deze representatie bestaat uit een zogenaamde woordvector, waarin verschillende eigenschappen van een weblogbericht zijn opgenomen op een manier dat een classificatiealgoritme hier wel iets mee kan. Deze verzameling van verschillende eigenschappen wordt ook wel een featureset genoemd.

Het is belangrijk om vooraf te stellen hoe de te gebruiken featureset er uit moet zien en welke weging deze features meekrijgen. Een in het veld van tekstclassificatie veelgebruikte methode om een featureset te genereren is de bag-of-words methode. Hierbij is de aanname dat de woorden die in een bericht voorkomen in zekere zin iets vertellen over de inhoud van een bericht. Er wordt dan gekeken naar de woordfrequentie, onder de aanname dat hoe vaker een woord voor komt in een tekst, hoe belangrijker het zal zijn. Op basis van deze woordfrequentie wordt gekeken in welke klasse het de tekst

geplaatst moet worden, door die klasse toe te wijzen die andere teksten bevat waar dat soort woorden ook vaak in voorkomen. Toch is woordfrequentie niet altijd een goede indicator: hoe vaker een woord in meerdere berichten voorkomt, hoe minder sterk het als discriminerende factor tussen berichten uit verschillend klassen zal zijn.

Uiteraard zijn ook andere featuresets mogelijk. In het onderzoek van [8] wordt bijvoorbeeld als feature het voorkomen van speciale tekens of het voorkomen van benadrukte woorden gebruikt, allemaal onder de aanname dat de aanwezigheid van dergelijke woorden of tekenreeksen veelzeggend is over de inhoud van het bericht.

Bij dit onderzoek is als belangrijkste feature de sentimentwaarde van een specifiek bericht genomen, verdeeld over drie verschillende dimensies zoals herkend door Osgood et al. in [10], te weten de *evaluative* factor, de *potency* factor en de *activity* factor. Deze drie waarden, samen met het aantal woorden in een bericht en de tijd waarop het bericht geplaatst is, zijn gebruikt als featureset om de woordvector van elk bericht te construeren. In de volgende sectie wordt dieper ingegaan op de manier van berekenen van de sentimentwaarde.

3.3 Berekenen sentimentwaarde

Van elk bericht uit het corpus is een sentimentwaarde berekend aan de hand van de methode zoals gehanteerd door Kamps et al. in [6]. Uit dit onderzoek is een woordenlijst voortgekomen, waarop meer dan 5000 termen met hun respectievelijke sentimentwaarde binnen drie verschillende dimensies worden genoemd. Op deze lijst kunnen woorden een score hebben binnen het domein [-1,1], waarbij negatieve en positieve scores aangeven dat een woord een negatieve, dan wel positieve oriëntatie heeft. Hoe tot deze scores is gekomen is na te lezen in hoofdstuk 2. Bij elk bericht worden de scores van ieder woord dat in het bericht voorkomt opgezocht in de woordenlijst. De scores van alle individuele woorden in het bericht worden opgeteld en zo ontstaat voor elk bericht een totaalscore op elk van de drie dimensies. Omdat er ook op een grafische manier naar deze scores gekeken zou gaan worden, zijn vanwege de overzichtelijkheid deze scores vervolgens genormaliseerd, door de totale score van een bericht op een bepaalde dimensie te delen door het aantal woorden in het bericht dat ook daadwerkelijk op de woordenlijst voorkomt. Zo wordt elke score teruggerekend naar een score binnen het domein [-1,1], waarbij een score van -1 staat voor een zeer negatieve sentimentele oriëntatie en een score van 1 staat voor een zeer positieve oriëntatie van het bericht als geheel.

3.4 Woordvector

Deze drie sentimentwaarden worden vervolgens, samen met een UNIX timestamp van het bericht, het aantal woorden gebruikt in het bericht, de unieke combinatie van gebruikersnaam plus berichtindentificatiecode (*docid*) en de bijbehorende stemming opgeslagen in zogenaamde tuples, die er als volgt uitzien:

```
1112465416,carthia26884,83,0.19,0.12,-0,accomplished
```

Met van links naar rechts: timestamp, docid, aantal woorden, EVA-score, POT-score, ACT-score en de corresponderende stemming.

Omdat berichten waarin geen woorden uit de woordenlijst in voorkomen automatisch drie keer een 0 scoren op de sentimentwaarden en daardoor niet erg nuttig zouden zijn in het classificatieproces, zijn de berichten waarbij dit het geval is niet in de dataset meegenomen. Anders is het als door optellen en aftrekken de scores van alle drie de sentimentwaarden op 0 uitkomt. Hoewel de kans klein is dat dit daadwerkelijk gebeurt is dat toch een andere situatie dan een bericht zonder woorden uit de woordenlijst. Door deze beslissing blijven er uiteindelijk nog 488.251 berichten over om mee te experimenteren.

3.5 Testdata en trainingsdata

Nadat van alle berichten de sentimentwaarden zijn berekend en de bovenstaande tuples zijn gegenereerd kan het classificatiealgoritme worden ingezet om te kijken hoe goed deze scoort op de data. Belangrijk hierbij is dat de verkregen data wordt gescheiden in testdata en trainingsdata. Trainingsdata zijn de gegevens waarmee de classifier wordt getraind, d.w.z. op basis waarvan een model wordt gebouwd waarmee nieuwe, onbekende instanties zullen worden geclassificeerd. Deze nieuwe, onbekende instanties worden geleverd door de testdata, een van de oorspronkelijke data afgesplitste groep berichten. Ter controle is de correcte stemming bij elk bericht geleverd, maar de classifier negeert deze informatie. Door deze controle kan uiteindelijk worden bekeken hoeveel berichten de classifier correct heeft voorspeld. Hoe hoger dit percentage, hoe beter uiteraard.

Een andere methode is om met één dataset te werken en hierop n-fold cross-validation toe te passen, waarbij n een willekeurig natuurlijk getal kan zijn. Dit houdt in dat de dataset in n delen wordt gesplitst waarbij n-1 delen worden gebruikt als trainingsdata en het resterende deel als testdata. Dit gebeurt zo voor alle delen en uiteindelijk wordt de totale score gemiddeld. Het is belangrijk test- en trainingsdata strikt gescheiden te houden, omdat deze twee anders kunnen interfereren: het is logisch dat een classifier goed scoort op testdata die net als trainingsdata is gebruikt, omdat de data in de testset exact voldoet aan de regels die uit die trainingsdata zijn afgeleid.

4 Experimenten

In deze sectie worden alle uitgevoerde experimenten beschreven. Eerst de precieze opzet van een experiment, daarna de resultaten en ten slotte een korte discussie. De resultaten van alle experimenten samen worden gezamenlijk besproken in de discussiesectie in hoofdstuk 5.

4.1 Experiment 1

Doel van het eerste experiment is om te kijken of er overeenkomsten zijn in de sentimentwaarden van verschillende vergelijkbare stemmingen, zodat er kan worden gekeken of bepaalde stemmingen te groeperen zijn. Dit zou het classificatieproces kunnen vereenvoudigen, omdat berichten dan eerst in een abstracte subklasse zouden kunnen worden ingedeeld, waarna ze binnen deze subklasse specifieker te geclassificeerd kunnen worden.

De veronderstelling in dit experiment is dat er in de distributie van verschillende stemmingen binnen de Osgood-dimensie patronen te herkennen zullen zijn, met bepaalde kenmerken die karakteristiek zijn voor een specifieke groep stemmingen. Van een geselecteerde groep stemmingen zijn deze scores in een 3D-scatter plot uiteengezet, om te kijken of er daadwerkelijk patronen te herkennen zijn.

4.1.1 Opzet

Allereerst zijn de 132 stemmingen handmatig ingedeeld in vier groepen, te weten duidelijk positieve stemmingen (zoals *excited* en *happy*), duidelijk negatieve stemmingen (zoals *enraged* en *sad*), ambigue stemmingen (stemmingen die zowel positief als negatief opgevat kunnen worden, zoals *drunk* en *tired*) en vage stemmingen (stemmingen waarvan de precieze oriëntatie onduidelijk is, zoals *awake* en *calm*). In de literatuur was niet een bepaalde methode te vinden om stemmingen te groeperen en daarom is gekozen voor een evaluatieve benadering, waarbij van elke stemming gekeken werd in welke groep hij het beste paste, op basis van het gevoel van de auteur bij elke stemming. Dit hele proces is erg subjectief en dus kan er gediscussieerd worden over de validiteit van deze indeling. Zo is het af en toe moeilijk een duidelijke scheidslijn tussen vage en ambigue stemmingen te maken en moet het ook zeker niet gezien worden als een definitieve lijst, het is slechts een hulpmiddel in dit experiment.

Nadat de stemmingen over de vier categorien waren verdeeld, zijn van elke categorie de tien sterkste stemmingen geselecteerd. Ook dit is weer middels een evaluatieve benadering gebeurd en is dus onderhevig aan subjectiviteit, maar komt voort uit gebrek aan een betere methode. Er is ook geen rekening

gehouden met het aantal berichten dat beschikbaar was voor de betreffende stemming, er is slechts gekeken naar de sterkte van de stemming.

In totaal zijn zo veertig stemmingen verkregen en voor elk van deze stemmingen is met het programma PSI-plot[8] een 3D-scatter plot gemaakt. De scores van de drie Osgood-waarden EVA, POT en ACT zijn op de x-,y- en z-as uitgezet, waarbij elke as een bereik van [-1,1] heeft. Berichten worden binnen dit driedimensionale vlak gerepresenteerd door een punt op het snijvlak van de drie coördinaten. Deze plots werden vervolgens een voor een bekeken waarbij notities werden gemaakt over de distributie van de data.

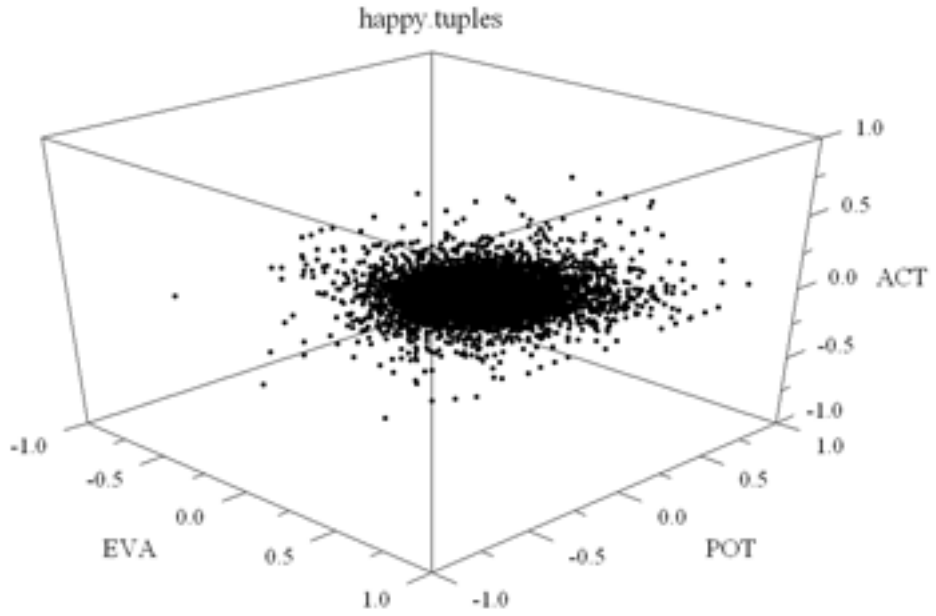
De hypothese in dit experiment is dat berichten met een verschillend karakter voor verschillende stemmingwolken zorgen in de plot. Voor berichten met een duidelijk positief karakter betekent dit dat wordt verwacht dat ze groeperen in het positieve vlak [(0,1);(0,1);(0,1)] van de plot en dat berichten met een duidelijk negatief karakter juist een stemmingwolk hebben in het negatieve vlak [(-1,0);(-1,0);(-1,0)]. Bij ambigue stemmingen is de verwachting dat er in beide vlakken een stemmingwolk ontstaat, doordat berichten met een dergelijke stemming zowel een positieve of negatieve lading kunnen hebben, en bij vage stemmingen wordt verwacht dat er helemaal geen wolk ontstaat, maar dat er sprake is van totale willekeurigheid van de positie van de verschillende berichten.

4.1.2 Resultaten

Positieve stemmingen
Er is sprake van een wolk van berichten die samenvalt rond het centrale punt van de plot, het nulpunt. Wel valt op dat er vanuit alle standpunten een neiging naar het positieve vlak is waar te nemen, alhoewel dit bij de EVA-scores sterker aanwezig is dan bij de ACT scores. POT-scores vallen er een beetje tussenin. Het gros van de berichten houdt zich echter op in het vlak [(-0,25;0,25);(-0,25;0,25);(-0,25;0,25)]. Ook zijn er extreme waarden (aan de rand van de plot, op -1 of 1 dus) die los van de wolk staan te zien. Deze bevinden zich zowel in het positieve als negatieve vlak van de plot, maar het merendeel is toch positief.

Negatieve stemmingen
Bij deze plots is een zelfde beeld te zien als bij de plots van de positieve stemmingen. De meeste waarden vallen in het vlak rond de 0, dus in het midden van de plot. Bij de POT en EVA waarden vallen de meeste berichten in het positieve vlak, alleen bij de ACT waarde is er sprake van een tendens naar het negatieve vlak, vooral door fragmentatie, niet in de vorm van een duidelijk wolk. Ook hier zijn weer extreme waarden te herkennen, aan beide zijden van het spectrum.

[8] http://www.polysoftware.com/psiplot.htm



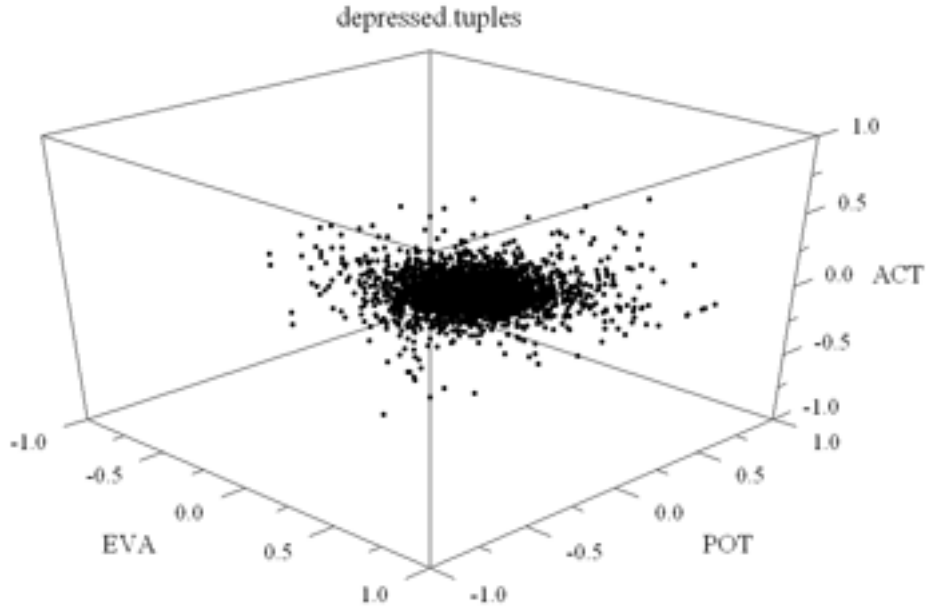
Figuur 2: 3D-scatterplot van berichten met de positieve stemming *happy*

Ambigue stemmingen
Deze berichten verzamelen zich rond de 0, maar er is ook sprake van fragmentatie, meer dan bij de positieve en negatieve plots. Waar de ACT en POT waarden vooral berichten hebben in het neutrale vlak rondom de 0, kenmerkt de EVA waarde zich door veel berichten in het positieve vlak, met name in het gebied [0;0,5]. Extreme waardes worden weer aan beide zijden van de plot aangetroffen, met een licht voordeel voor de positieve kant.

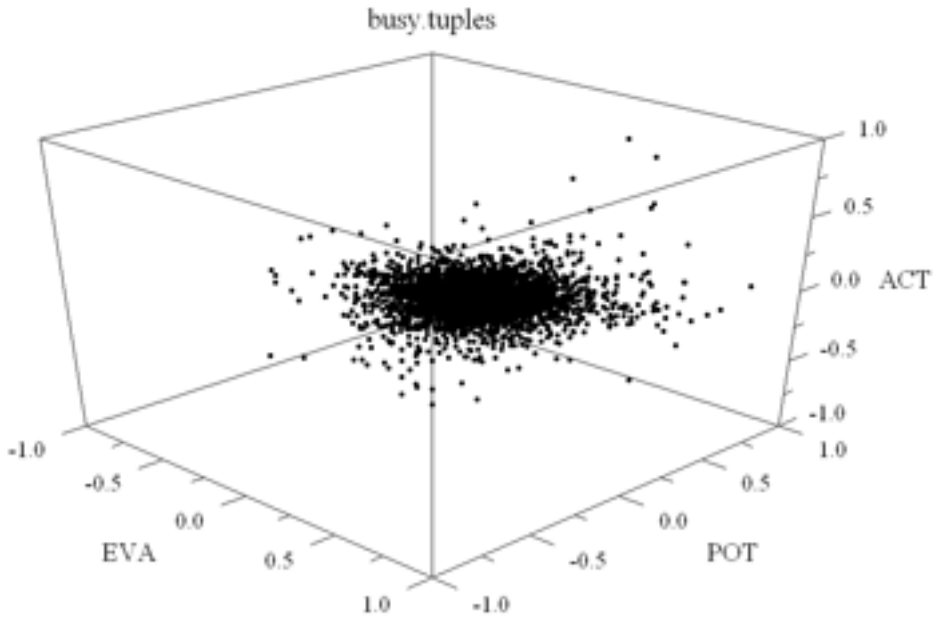
Vage stemmingen
Bij deze plots is meer fragmentatie waar te nemen dan bij de plots van de andere typen stemmingen. Nog steeds echter, verzamelt het gros van de berichten zich rondom de 0. De meeste berichten neigen naar het positieve vlak, gezien vanuit alle standpunten, hoewel er bij de ACT waarden redelijk wat fragmentatie in het negatieve vlak is. Ook in deze plots zijn weer extreme waardes te herkennen, zonder dat er een duidelijke voorkeur is voor het positieve of negatieve vlak.

4.1.3 Discussie

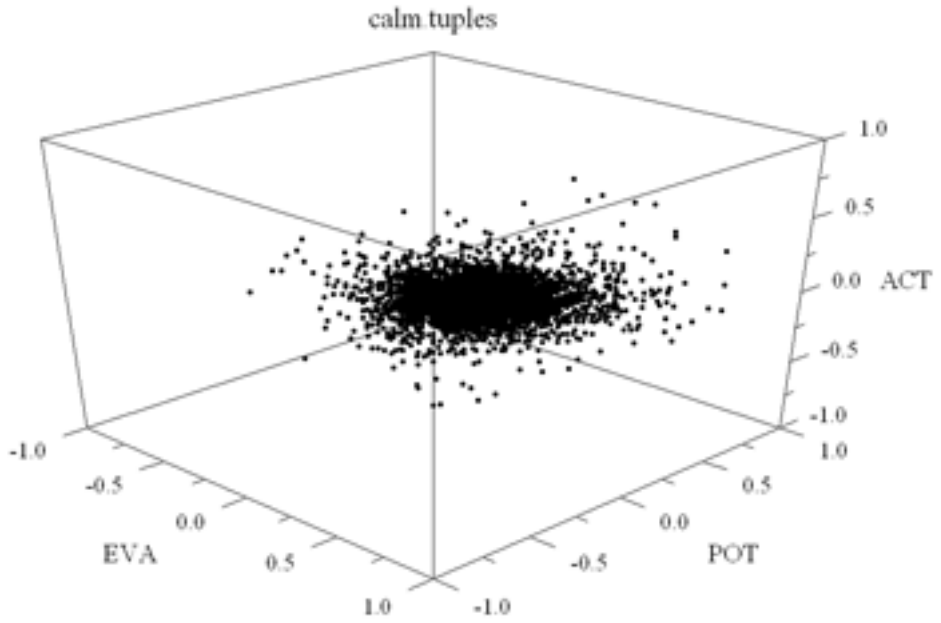
De vooraf gestelde hypothesen over de distributie van berichten van de verschillende soorten stemmingen blijken niet te kloppen. In zekere zin zijn er wel verschillen waar te nemen tussen de plots van verschillende stemmingen, maar dit zijn vaak nuanceverschillen. In grote lijnen komen de veertig



Figuur 3: 3D-scatterplot van berichten met de negatieve stemming *depressed*



Figuur 4: 3D-scatterplot van berichten met de ambigue stemming *busy*



Figuur 5: 3D-scatterplot van berichten met de vage stemming *calm*

gemaakte plots behoorlijk met elkaar overeen, zoals ook te zien is in de hierboven gegeven voorbeelden. Deze zijn willekeurig gekozen en geven een representatief beeld van hoe de plots er uit zien. Het is duidelijk te zien dat er geen sprake is van één of meerdere stemmingwolken in een bepaald vlak van de plot. Er ontstaat steeds een wolk rond de 0-lijn van de drie waardes, waarbij alleen de mate van fragmentatie licht fluctueert. Echt duidelijk onderscheidende patronen zijn er helaas niet te herkennen.

4.2 Experiment 2

Doel van het tweede experiment was om te kijken in welke mate een tekstclassifier in staat zou zijn om de stemming van een bericht te voorspellen op basis van de sentimentele oriëntatie van dat bericht. De resultaten van dit experiment worden vergeleken met resultaten uit vergelijkbare onderzoeken, om te kijken of deze resultaten nauwkeuriger zijn en hoe dat komt.

4.2.1 Opzet

Voor het classificeren van de berichten is WEKA [15] gebruikt. Dit betekent dat alle data moet worden omgezet naar het ARFF-bestandsformaat, zodat WEKA hiermee overweg kan. Gelukkig waren de bestanden met daarin de datarepresentaties (*tuples*) van de berichten compatibel met dit formaat,

wat in de praktijk betekent dat alleen de kolomnamen en het type gegevens in de kolommen gedefinieerd moeten worden.

Zoals in hoofdstuk 3 te lezen valt, zijn er verschillende manieren om test- en trainingsdata te genereren. In dit experiment zijn de test- en trainingsdata gegenereerd door voor elke stemming willekeurig een deel positieve berichten en negatieve berichten te pakken in een 50/50 verhouding. Dus de test- en trainingssets bestaan voor 50% uit berichten die met de betreffende stemming zijn geclassificeerd en voor 50% uit berichten met een andere stemming. Uiteraard is hierbij gelet op dat trainingsdata en testdata elkaar niet overlappen. Uiteindelijk wordt gebruik gemaakt van een testset van 390 items (met 195 berichten die bij de gekozen stemming horen en voor alle 39 andere stemmingen 5 berichten per stemming) en twee trainingssets: een van 1950 items (met 975 berichten die bij de gekozen stemming horen en voor alle 39 andere stemmingen 25 berichten per stemming) en een van 7020 items (met 3510 positieve items en eenzelfde aantal evenredig verspreid over de andere stemmingen). Dit werd gedaan om te zien of het aantal trainingsinstanties van invloed is op de nauwkeurigheid van de classificatie. In het derde experiment wordt hier nog gedetailleerder op ingegaan.

Om de berichten te classificeren zijn classificatie-algoritmes nodig. In dit experiment is er voor gekozen om analoog aan [8] gebruik te maken van support-vector machines (SVM's), in de implementatie van WEKA SMO geheten, en een baseline in de vorm van ZeroR. Daarnaast is er voor gekozen om ook een bayesiaanse classifier (Naive Bayes) en een classifier op basis van de nearest-neighbour methode (IBk) te gebruiken. Hierna volgt een beknopte uitleg van elk van deze classifiers, voor uitgebreidere toelichting zie [1, 5, 7, 11].

Van de genoemde classifiers is ZeroR de eenvoudigste: van een gegeven trainingsset wordt bij nominale klassen gekeken naar de modus en bij numerieke klassen naar het gemiddelde en vervolgens wordt in een testset elke instantie volgens die waarde geclassificeerd. Gezien de hier gebruikte test- en trainingsset betekent dat, dat de nauwkeurigheid van deze classifier steeds op 50% zal zitten, namelijk alle items worden of wel, of niet tot de betreffende stemming gerekend en daarbij zit de classifier automatisch in de helft van de gevallen goed. Voor deze classifier is gekozen om als *baseline* te dienen: met deze simpele methode is sowieso 50% goed geclassificeerd, een andere methode die minder dan dit scoort is dus niet erg nuttig.

Het SMO (sequential minimal optimization)-algoritme maakt gebruik van zogenaamde *support-vectors* die dienen om in een classificatieproces zonder numerieke waarden toch gebruik te maken van een lineair model. Dit gebeurt door van de originele vector middels non-lineaire mapping een nieuwe vector te maken, waarin classificatie categorien wel lineaire grenzen kennen. Voor elke classificatie categorie wordt op deze manier een vergelijking op-

gesteld die gebaseerd is op de informatie uit trainingsdata. Testdata wordt geclassificeerd door voor elke mogelijke klasse de vergelijking uit te rekenen waarna de testinstantie wordt toegekend aan de klasse waarvan de corresponderende vergelijking de hoogste uitkomst had. Deze classifier werd gekozen omdat hij, weliswaar in een iets andere vorm, ook in [8] is gebruikt, waardoor de uitkomsten van beide experimenten goed met elkaar kunnen worden vergeleken.

Als derde classifier wordt Naive Bayes gebruikt. Deze classifier maakt gebruik van Bayes theorema, wat inhoudt dat de kans dat een instantie in een bepaalde klasse hoort berekend wordt, door de kans dat een instantie gezien een bepaalde attribuutwaarde in een bepaalde klasse hoort uit te rekenen en deze kansen voor elke attribuutwaarde van de instantie op te tellen. Dit gebeurt voor alle klassen die er zijn en vervolgens wordt de instantie ingedeeld in de klasse met de grootste waarschijnlijkheid. Het algoritme gaat hierbij uit van onafhankelijkheid tussen de attribuutwaarden en wordt daarom naïef genoemd. Er is gekozen om deze classifier te gebruiken omdat er vaak goede resultaten mee te halen zijn.

De laatste gebruikte classifier is het IBk-algoritme, wat uitgaat van het zogenaamde 'n-nearest neighbour' principe. Dit houdt in dat van bij een onbekende instantie wordt gekeken naar het n-aantal klassen dat daarbij het dichtst in de buurt ligt. De onbekende instantie krijgt dan automatisch de klasse toegewezen van deze dichtstbijzijnde bekende instantie. Als er meerdere instanties van verschillende klassen in de buurt van de onbekende instantie liggen, wordt gekeken naar welke van die klassen de meerderheid heeft en wordt vervolgens die klasse aan de onbekende instantie toegekend. Voor deze classifier is gekozen vanwege het ruimtelijke aspect van de drie sentimentwaarden: zoals in het eerste experiment te zien is vallen de berichten goed te representeren door punten in een driedimensionaal vlak, waardoor het kijken naar buurklassen een interessante voorspeller kan zijn.

Als featureset worden de in hoofdstuk 3 al genoemde kenmerken van de berichten gebruikt: timestamp, aantal woorden en de drie sentimentwaarden. Alleen de kolom met daarin het bericht-id werd, gezien de aard van de data (tekststring) en de geringe voorspellende waarde, weggelaten uit de dataset. Vergeleken met andere onderzoeken is dit een kleine featureset, maar dat kan ook in het voordeel van deze methode werken, omdat er minder complexe regels nodig kunnen zijn, wat het risico op overfitting verkleint.

4.2.2 Resultaten

In tabellen 2 tot en met 4 staan de resultaten van experimenten, per classificatiealgoritme, met zowel trainingssets van 1950 instanties als 7020 instanties. De testset bestaat in alle gevallen uit 390 instanties. Alleen voor het

ZeroR-algoritme is geen tabel geplaatst; deze scoort zoals in de vorige sectie al gezegd steeds 50% correct.

Stemming	*Correct*		*Stemming*	*Correct*	
	1950	*7020*		*1950*	*7020*
ecstatic	55,64%	55,90%	accomplished	51,28%	51,54%
drained	53,59%	55,38%	excited	52,56%	51,03%
amused	54,10%	54,10%	bored	50,26%	51,03%
hopeful	50,26%	53,85%	cranky	49,74%	51,03%
calm	50,51%	53,59%	pissed off	53,08%	50,77%
loved	55,38%	53,33%	cheerful	49,49%	50,77%
sad	52,31%	53,33%	creative	54,62%	50,51%
exhausted	51,54%	53,08%	aggravated	53,85%	50,51%
contemplative	55,38%	52,82%	curious	51,79%	50,51%
happy	54,87%	52,82%	bouncy	49,74%	50,51%
okay	50,77%	52,82%	good	56,15%	50,00%
crazy	54,36%	52,56%	chipper	48,97%	49,74%
sleepy	53,33%	52,56%	depressed	52,31%	48,97%
thoughtful	55,13%	52,31%	content	48,21%	48,72%
confused	52,31%	52,31%	blah	50,51%	48,46%
busy	51,54%	52,31%	hungry	49,49%	48,46%
annoyed	55,13%	52,05%	awake	48,21%	48,46%
sick	51,54%	52,05%	cold	47,69%	48,46%
frustrated	53,59%	51,79%	blank	51,28%	47,18%
tired	49,23%	51,79%	anxious	49,23%	46,41%

Tabel 2: Resultaten met SMO classifier (testset van 390 items)

4.2.3 Discussie

De resultaten van het classificeren vallen behoorlijk tegen. De beste resultaten scoren zo'n 5 tot 6 procent beter dan de baseline, een groot deel scoort slechter. Opvallend is ook dat een grotere trainingsset nauwelijks positief effect heeft op de resultaten. Met Naive Bayes wordt met 1950 trainingsinstanties gemiddeld 51,34% correct geclassificeerd tegenover 51,41% als er gebruik wordt gemaakt van 7020 trainingsinstanties: een marginale verbetering. Bij de andere twee classifiers gaat het gemiddelde zelfs omlaag bij meer trainingsinstanties, namelijk van 51,97% (1950 trainingsinstanties) naar 51,35% (7020 trainingsinstanties) correct geclassificeerd bij SMO, en van 51,49% correct naar 50,31% correct geclassificeerd bij IBk.

Ook is er geen trend waar te nemen wat betreft concrete stemmingen tegenover vage stemmingen. Een concrete stemming als *sad* scoort bij classificeren

Stemming	*Correct*		*Stemming*	*Correct*	
	1950	*7020*		*1950*	*7020*
contemplative	55,38%	56,92%	pissed off	50,00%	51,28%
thoughtful	53,59%	55,90%	okay	49,49%	51,28%
amused	56,41%	54,36%	bouncy	49,23%	51,28%
drained	54,62%	54,36%	accomplished	52,05%	51,03%
ecstatic	53,85%	54,10%	hopeful	51,28%	51,03%
creative	52,31%	53,33%	aggrevated	51,03%	51,03%
happy	52,56%	53,08%	loved	48,97%	51,03%
crazy	53,59%	52,82%	frustrated	52,05%	50,77%
content	52,05%	52,82%	cranky	50,51%	50,26%
sick	51,79%	52,82%	busy	49,74%	50,26%
good	52,31%	52,56%	anxious	50,51%	50,00%
calm	51,79%	52,31%	annoyed	49,23%	50,00%
sleepy	50,26%	52,31%	cold	51,54%	49,23%
excited	52,05%	52,05%	cheerful	50,26%	49,23%
bored	50,77%	52,05%	sad	49,74%	49,23%
tired	52,05%	51,79%	blah	47,44%	48,97%
curious	51,79%	51,79%	confused	51,54%	47,95%
hungry	51,28%	51,79%	chipper	50,26%	47,95%
depressed	51,03%	51,79%	blank	50,00%	47,95%
exhausted	50,77%	51,28%	awake	48,46%	46,41%

Tabel 3: Resultaten met Naive Bayes classifier (testset van 390 items)

met Naive Bayes onder de 50%, terwijl een vage stemming als *calm* bij dezelfde classifier rond de 52% scoort. Nu zijn dit geen enorme verschillen en hier valt dus weinig waarde aan te geven, hoewel een zelfde soort waarneming ook door Mishne in [8] werd gedaan.

Wat wel interessant is, is kijken naar precision en recall bij de verschillende classificatiemethoden. *Precision* is het aantal berichten dat als een bepaalde stemming is geclassificeerd, als percentage van het aantal berichten dat ook daadwerkelijk die stemming heeft. *Recall* is het aantal berichten dat als een bepaalde stemming geclassificeerd is, als percentage van het aantal berichten dat die stemming heeft. Als er bijvoorbeeld 100 berichten zijn, waarvan 50 van een bepaalde stemming en 50 van andere stemmingen en de classifier heeft classificeert 60 berichten als positief, terwijl er daarvan maar 40 echt positief zijn (*true positives*), dan resulteert dat in een *precision* van (40/60 = 66,7%) en een recall van (40/50 = 80%). De verschillende classifiers scoren hier afwisselend op.

Zo is Naive Bayes redelijk nauwkeurig wat betreft het herkennen van berichten die bij de gevraagde stemming horen. Het scoort hier een recall van

Stemming	*Correct*		*Stemming*	*Correct*	
	1950	*7020*		*1950*	*7020*
pissed off	51,28%	56,15%	awake	49,74%	49,74%
busy	50,00%	54,36%	chipper	49,23%	49,74%
cheerful	53,08%	54,10%	hopeful	50,26%	49,49%
sad	52,31%	53,59%	exhausted	48,97%	49,49%
drained	55,64%	52,82%	happy	52,31%	49,23%
calm	55,13%	52,56%	excited	52,05%	49,23%
tired	51,79%	52,31%	aggravated	49,49%	48,97%
amused	55,90%	52,05%	content	55,13%	48,72%
blah	52,31%	51,79%	ecstatic	54,36%	48,72%
good	52,05%	51,79%	curious	52,05%	48,72%
blank	50,51%	51,79%	hungry	48,97%	48,72%
depressed	54,87%	51,28%	frustrated	53,33%	48,46%
crazy	53,33%	51,03%	annoyed	51,54%	48,46%
anxious	52,05%	51,03%	sleepy	49,74%	48,46%
creative	51,28%	51,03%	accomplished	51,54%	48,21%
sick	51,03%	51,03%	cranky	48,97%	48,21%
cold	50,51%	50,51%	loved	51,28%	47,95%
bored	47,95%	50,51%	bouncy	48,72%	47,95%
contemplative	52,82%	50,26%	okay	50,00%	47,18%
confused	49,74%	50,26%	thoughtful	48,21%	46,67%

Tabel 4: Resultaten met IBk classifier (testset van 390 items)

69,3% en 68,4%. Bij de andere classifiers liggen deze percentages meer richting 50%, wat er op duidt dat er niet echt een duidelijk verschil is tussen berichten die wel en niet bij een specifieke stemming horen. Vooral de IBk-classifier scoort hierbij redelijk stabiel, met waarden tussen de 50% en 51% bij zowel precision als recall. SMO lijkt meer een neiging te hebben om berichten als 'negatief' te bestempelen, gezien een recall-waarde van 57,1% voor berichten die niet bij een stemming horen tegenover 45,6% voor berichten die wel bij de gezochte stemming horen, bij een trainingsset van 7020 instanties.

4.3 Experiment 3

Om te kijken of er een verband is tussen het aantal trainingsinstanties en de nauwkeurigheid van de classifiers wordt nog een experiment uitgevoerd. Doel van dit experiment is onderzoeken of de grootte van de trainingsset invloed heeft op de nauwkeurigheid van classificeren en of deze nauwkeurigheid te maximaliseren is, door een steeds grotere trainingsset te gaan gebruiken.

4.3.1 Opzet

Aangezien er van de meeste stemmingen uit de stemmingen top 40 niet genoeg berichten zijn om met heel grote trainingssets te gaan werken, is besloten om alleen met de elf stemmingen met de meeste berichten te gaan werken in dit experiment. Dit zijn allemaal stemmingen met meer dan 10.000 berichten. Omdat geen van de classifiers in experiment 2 een duidelijke stijging toonde bij meer traningsdata, worden alle drie de gebruikte classifiers in dit experiment weer ingezet. Bij dit experiment is niet gebruik gemaakt van een test-set, in plaats daarvan wordt er middels tenfold cross-validation gewerkt. Er worden trainingssets gemaakt van 2.500, 5.000, 10.000 en 20.000 trainingsinstanties, allemaal weer met dezelfde 50/50 verhouding als in experiment 2. Grotere trainingssets zijn met dit corpus niet mogelijk, aangezien de meest voorkomende stemmingen niet veel meer dan 10.000 berichten hebben. Bij de IBk-classifier waren er problemen met de trainingsset van 20.000 instanties, dus daarvan ontbreken de gegevens.

4.3.2 Resultaten

In tabellen 5 tot en met 7 staan de resultaten van experimenten, per classificatiealgoritme, met zowel toenemende trainingssets van 2.500, 5.000, 10.000 en 20.000 instanties.

Stemming	2500	5000	10000	20000
accomplished	51,76%	51,94%	52,38%	52,36%
amused	54,52%	55,86%	56,36%	55,86%
blah	51,44%	51,54%	50,70%	50,84%
bored	50,32%	50,94%	51,78%	51,07%
cheerful	49,12%	49,72%	50,00%	50,42%
contemplative	54,76%	56,04%	54,45%	55,23%
content	51,64%	53,16%	51,56%	52,82%
excited	53,72%	53,14%	52,99%	52,69%
happy	50,00%	50,48%	52,39%	52,00%
sleepy	50,12%	50,50%	50,97%	51,11%
tired	50,04%	52,54%	52,19%	51,59%

Tabel 5: Resultaten met Naive Bayes classifier en toenemende trainingsset

4.3.3 Discussie

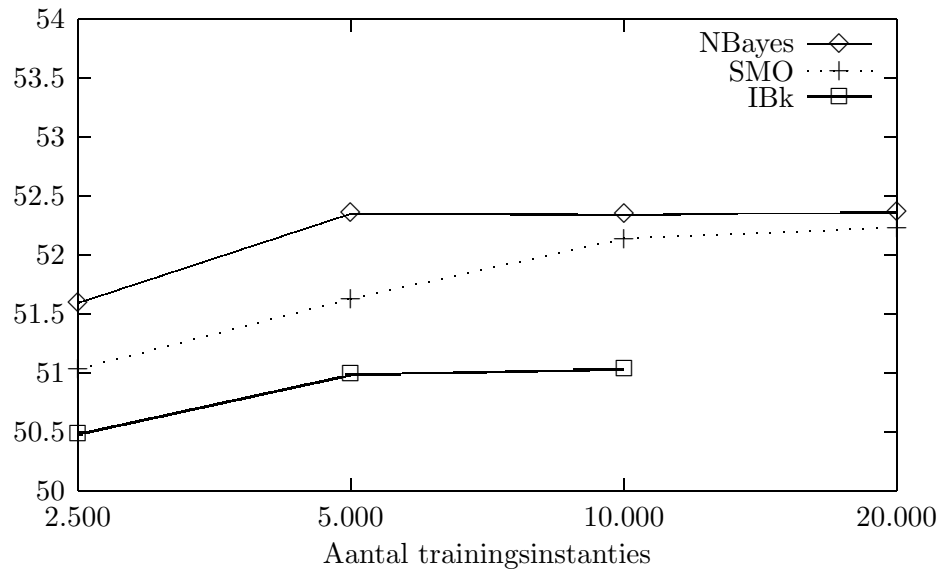
Er zijn weinig grote verschillen te zien tussen de nauwkeurigheid van de classifiers met trainingssets van verschillende grootte. Bij vrijwel alle stemmin-

Stemming	2500	5000	10000	20000
accomplished	51,28%	51,48%	51,88%	51,91%
amused	52,96%	52,62%	54,64%	54,65%
blah	51,04%	51,06%	51,25%	50,46%
bored	49,56%	50,08%	52,40%	51,92%
cheerful	51,44%	51,74%	49,73%	50,91%
contemplative	51,56%	53,56%	52,21%	52,58%
content	48,52%	51,60%	51,91%	51,54%
excited	52,08%	53,40%	53,68%	54,77%
happy	50,64%	51,60%	52,98%	52,95%
sleepy	50,56%	49,90%	51,19%	51,52%
tired	51,64%	50,84%	51,63%	51,37%

Tabel 6: Resultaten met SMO classifier en toenemende trainingsset

Stemming	2500	5000	10000
accomplished	50,32%	50,34%	51,71%
amused	52,16%	53,18%	51,58%
blah	51,44%	50,70%	50,95%
bored	48,52%	50,98%	50,31%
cheerful	50,52%	51,10%	49,71%
contemplative	51,76%	52,24%	51,35%
content	49,20%	50,30%	52,07%
excited	52,92%	51,70%	51,66%
happy	50,28%	50,04%	50,92%
sleepy	48,52%	50,36%	49,96%
tired	49,60%	49,84%	51,16%

Tabel 7: Resultaten met IBk classifier en toenemende trainingsset



Figuur 6: Verloop aantal correcte berichten bij toenemende trainingsdata

gen neemt de nauwkeurigheid bij een verdubbeling van het aantal instanties van 2.500 naar 5.000 iets toe; een groot deel daarvan kent ook toename als het aantal instanties wordt verdubbeld naar 10.000. Een trainingsset van 20.000 instanties zorgt er in de meeste gevallen echter voor dat de score iets verslechtert. Kijkend naar gemiddeldes is er duidelijk sprake van een afnemende stijging van de nauwkeurigheid, zoals ook blijkt uit de grafiek in figuur 6. De grens hierbij ligt rond de 10.000 trainingsinstanties: daarna is de groei van het aantal correct geclassificeerde berichten miniem. Een reden voor de afnemende nauwkeurigheid bij meer dan 20.000 trainingsinstanties kan zijn dat de aanwezigheid van zo veel trainingsinstanties het verschil tussen de klassen niet duidelijker maakt, door de aanwezigheid van een hoop ruis tussen die trainingsinstanties. In hoofdstuk 5 wordt hier verder op in gegaan.

4.4 Experiment 4

Het doel van dit experiment is om te kijken of het combineren van de featureset van Mishne [8] met de featureset uit dit onderzoek een verbetering van de resultaten uit experiment 2 tot gevolg heeft. De resultaten van dit experiment zullen dus zowel met de resultaten van experiment 2 als met de resultaten van het onderzoek van Mishne worden vergeleken.

4.4.1 Opzet

Hiertoe zijn beide featuresets samengevoegd en is van de berichten in het corpus een nieuwe woordvector berekend. Gezien het aantal features uit de featureset van Mishne is er in dit experiment voor gekozen om gebruik te maken van de SVMlight-classifier[9]. Dit is een methode die gebruik maakt van *support-vectors* en is vergelijkbaar met de SMO-classifier zoals gebruikt in experiment 2. Een ander voordeel van dit type classifier is dat ze een stuk beter presteren dan andere classifiers, zoals ook al door Mishne werd aangehaald in [8].

De verdere opzet van dit experiment is gelijk aan de opzet van experiment 2. Weer is er gebruik gemaakt van twee trainingsets, een met 1950 instanties en een met 7020 instanties, beide in een 50/50 verhouding van berichten (voor een uitgebreidere uitleg zie paragraaf 4.2) en een testset van 390 instanties. De resultaten van de classificatie staan in de volgende sectie.

4.4.2 Resultaten

In tabel 8 staan de resultaten van het classificeren van de berichten met de Mishne featureset, uitgebreid met de sentimentanalyse featureset. In de linkerkolom staan de resultaten zoals behaald in dit onderzoek met de featureset van Mishne, uitgebreid met sentimentfeatures. In de rechterkolom staan de originele gegevens integraal overgenomen uit [8].

4.4.3 Discussie

Vergeleken met de resultaten uit experiment 2 zijn de uitkomsten van dit experiment een stuk beter te noemen, maar ze blijven nog steeds achter bij de resultaten van Mishne in [8]. Enkele stemmingen springen er wel uit, namelijk *depressed*, *amused*, *contemplative* en *thoughtful*, met een nauwkeurigheidspercentage van 9-13% boven de baseline. Zulke hoge percentages werden in experiment 2 niet gehaald, met geen enkele van de classifiers. De verbetering uit zich ook in de gemiddelde score van 52,37% bij een trainingset met 1950 instanties en 53,06% met 7020 trainingsinstanties. Dit is een lichte toename van de gemiddelde nauwkeurigheid met minder dan 1%, ten opzichte van de resultaten uit experiment 2.

Het toevoegen van extra trainingsinstanties heeft een duidelijk positief effect. Dit is het meest duidelijk te zien bij de stemming *depressed*, met een stijging van het aantal correct voorspelde berichten van bijna 11%. Er zijn echter ook enkele stemmingen die slechter of gelijkwaardig scoren, waarvan *exhausted*, *cold* en *sick* het meest opvallen, met een daling van 2-3%. Het dalen van

[9] http://svmlight.joachims.org

Stemming	*Correct*		*Correct (Mishne)*	
	1950	*7020*	*1600*	*6400*
depressed	52,31%	63,08%	58,25%	62,50%
amused	58,72%	59,74%	57,75%	60,75%
contemplative	60,00%	59,23%	53,25%	57,00%
thoughtful	56,67%	58,97%	52,75%	59,00%
hopeful	55,64%	56,92%	51,50%	57,50%
bored	51,28%	55,64%	51,52%	55,25%
loved	52,82%	55,38%	57,00%	57,75%
curious	53,85%	55,13%	60,25%	63,25%
ecstatic	53,08%	54,87%	54,00%	59,75%
busy	50,51%	54,87%	50,75%	53,50%
accomplished	54,10%	54,62%	54,75%	55,75%
pissed off	53,59%	54,62%	no data	no data
drained	54,36%	54,36%	47,50%	52,25%
sad	51,79%	53,59%	53,00%	60,25%
annoyed	53,33%	53,33%	57,00%	59,00%
aggravated	50,51%	53,33%	52,75%	54,25%
frustrated	52,82%	53,08%	57,00%	60,25%
excited	52,56%	52,82%	55,50%	59,75%
confused	50,26%	52,82%	56,00%	65,75%
good	53,08%	52,05%	48,50%	50,50%
creative	51,54%	52,05%	47,75%	50,50%
chipper	51,28%	52,05%	no data	no data
hungry	50,26%	51,79%	51,50%	50,75%
exhausted	53,85%	51,54%	52,50%	52,50%
crazy	52,82%	51,54%	54,00%	55,00%
awake	50,26%	51,54%	51,50%	53,75%
okay	51,28%	51,03%	46,75%	49,00%
tired	50,00%	51,03%	no data	no data
calm	49,74%	51,03%	44,75%	49,00%
happy	51,28%	50,77%	54,50%	60,75%
sleepy	51,28%	50,77%	44,25%	55,00%
cold	53,08%	50,51%	50,25%	53,25%
sick	52,56%	50,26%	54,75%	60,25%
blah	51,03%	50,00%	53,75%	57,75%
cheerful	51,03%	50,00%	52,50%	54,25%
cranky	50,51%	50,00%	55,00%	57,25%
blank	51,03%	49,74%	56,00%	54,50%
content	49,23%	49,74%	50,75%	54,00%
anxious	50,26%	49,49%	51,75%	54,25%
bouncy	51,28%	49,23%	51,00%	59,50%

Tabel 8: Resultaten van uitgebreide Mishne featureset en SVM classifier (testset van 390 items)

de nauwkeurigheid valt misschien te verklaren door dat deze stemmingen geen duidelijk of een ambigue oriëntatie hebben. Door het toevoegen van meer trainingsinstanties ontstaat er dus meer ruis, waardoor het moeilijker is voor de classifier om duidelijke verschillen tussen deze stemmingen te zien. Opmerkelijk in die zin is het omlaag gaan van de nauwkeurigheid van de stemming *happy*, wat een duidelijk positieve stemming is. Kennelijk is hier nog iets anders aan de hand, maar wat is niet meteen duidelijk.

Opvallend is verder dat er een redelijke scheiding lijkt te zijn tussen concrete en vage stemmingen. Stemmings als *blah*, *blank*, *calm* en *okay* scoren allemaal niet erg hoog. Dit lijkt intuitief, aangezien er aan deze stemmingen moeilijk een bepaalde sentimentele oriëntatie te binden is. Ook is te zien dat stemmingen die wel een duidelijke sentimentele oriëntatie hebben als *depressed*, *amused* en *loved* aanzienlijk beter scoren. Toch is het niet mogelijk hier een wetmatigheid uit af te leiden, gezien de slechte prestaties van duidelijke stemmingen *happy* en *cheerful* en de redelijke prestaties van vage stemmingen als *busy* en *curious*. Verder is er ook geen opvallende trend tussen negatieve en positieve stemmingen te zien: beide zijn redelijk verspreid over de resultaten en er is dus niet sprake van dat een van beide stemmingen beter scoort dan de andere.

5 Discussie

De prestaties van de classificatietaak zijn niet erg goed te noemen. Daarom is het interessant te kijken of er bepaalde factoren aan te wijzen zijn die een negatief invloed hebben op de nauwkeurigheid van het classificeren. Op het eerste gezicht lijken de classificatieresultaten redelijk willekeurig. In de experimenten zonder uitgebreide featureset scoren vage en concrete stemmingen niet opvallend anders: vage stemmingen worden niet consequent slechter geclassificeerd dan concrete stemmingen en andersom is dit ook niet het geval. Als er wordt gekeken naar de tien best geclassificeerde stemmingen per classificatiealgoritme, valt op dat er in totaal 21 verschillende stemmingen in deze top 10 lijsten voorkomen. Een zelfde beeld valt te zien bij de tien slechts geclassificeerde stemmingen per classificatiealgoritme: hier zijn zelfs 22 verschillende stemmingen te vinden, waarvan zeven stemmingen die ook voorkomen op de best geclassificeerde lijst (zie figuur 2, 3 en 4). In beide lijsten is de verdeling van stemmingen vrijwel gelijk: iets meer dan de helft concrete stemmingen en voor de rest vage stemmingen met een enkele ambigue stemming als uitzondering. De nauwkeurigheid van de classificatie van een bepaalde stemming hangt dus af van het gebruikte algoritme en hoewel het niet vreemd is hier enig verschil in aan te treffen is het erg opvallend te noemen dat er zulke grote verschillen bestaan, wat het idee van willekeurigheid alleen maar versterkt.

Ook in het experiment waarbij de featureset van Mishne in combinatie met de sentimentfeatures wordt gebruikt hebben de resultaten alle schijn van willekeur. Omdat er in de gegevens van Mishne van drie stemmingen de data ontbrak zijn de 37 overgebleven stemmingen en hun classificatieresultaten met elkaar vergeleken (zie figuur 8). Hieruit blijkt dat 17 van de 37 stemmingen met gebruikmaken van sentimentfeatures op enige wijze beter worden geclassificeerd dan zonder. Met op enige wijze wordt hier bedoeld dat niet alle stemmingen met beide trainingssets beter scoren: negen stemmingen scoren alleen beter met de kleine trainingsset, vier stemmingen scoren alleen beter met de grotere trainingsset en slechts vier stemmingen scoren in beide gevallen beter. Hierbij moet wel worden opgemerkt dat in sommige gevallen de verschillen miniem zijn en deze dus mogelijk kunnen worden toegewezen aan de onderlinge verschillen in grootte van de trainingssets. Van de vier stemmingen die met beide trainingssets beter scoren zijn er drie niet erg concreet (*good*, *creative* en *contemplative* en één (*drained*) juist wel. Ook in dit geval is een duidelijk patroon afwezig en lijken de resultaten vooral gebaseerd op willekeur.

Wat is nu precies de oorzaak van de matige classificatieresultaten? Dit lijkt twee mogelijke oorzaken te hebben. Ten eerste zijn er de vage verschillen tussen de klassen, die het moeilijk maken om nauwkeurig de weblogberichten te classificeren. Deze vage klassenverschillen kunnen twee oorzaken hebben: de data zelf en de gebruikte methode voor het berekenen van de sentimentwaardes van de berichten. Mishne noemt in [8] al dat het gebruikte corpus niet ideaal is: er is sprake van subjectieve annotatie, er zijn veel korte berichten met weinig woorden en de stemmingen die aan de berichten worden gekoppeld zijn ook niet altijd eenduidig te noemen. Dit kan hebben bijgedragen aan de resultaten, maar verklaart niet de uitkomst van experiment vier, waarbij de resultaten nog steeds achterblijven bij die van Mishne: hier zorgt het toevoegen van sentimentfeatures er misschien voor dat de klassen nog moeilijker te onderscheiden van elkaar zijn.

Het is daarom waarschijnlijker dat de methode gebruikt om sentimentanalyse toe te passen niet op een goede manier de sentimenten van de weblogberichten kan representeren. De gebruikte methode kijkt alleen naar adjectieven en dan ook nog alleen naar adjectieven die in WordNet een verbinding hebben met de adjectieven *good* of *bad*. Het kan dus betekenen dat de woorden uit de woordenlijst niet representatief zijn voor het woordgebruik binnen het corpus. Omdat er niet naar de context waarin de adjectieven voorkomen wordt gekeken, kan het ook nog gebeuren dat een bericht dat de frase *not happy* bevat, toch een positieve score toegekend krijgt. Dit soort ruis in de data die ontstaat en het feit dat de methode van Kamps et al. ook niet extreem nauwkeurig is, maakt het goed mogelijk dat het probleem met de onduidelijke klassenverschillen binnen dit corpus deels af te vangen is door het gebruikmaken van een andere methode voor sentimentanalyse. Dit

wordt nog verder ondersteund door de resultaten uit experiment 3, waaruit blijkt dat het gebruiken van meer trainingsdata nauwelijks effect heeft op de nauwkeurigheid van classificeren, wat aangeeft

Voor de tweede oorzaak van de matige classificatieresultaten is het nuttig om te kijken naar een andere factor die invloed heeft op de nauwkeurigheid van de classificatie, namelijk de keuze van de kernel bij het gebruikmaken van SVM's. In het vierde experiment is, net zoals door Mishne, gebruik gemaakt van een lineaire kernel. Bij het classificeren van documenten zoals weblogberichten is er echter geen sprake van numerieke waarden, zodat het stellen van lineaire grenzen tussen de verschillende klassen niet mogelijk is. Uit de visuele representaties van de sentimentsscores van de weblogberichten uit het eerste experiment blijkt al dat er sprake is van data die slecht lineair te scheiden is. Het zou daarom nuttig kunnen zijn eenzelfde experiment uit te voeren, gebruikmakend van non-lineaire kernels zoals een polynome of rbf-kernel. Omdat deze kernels beter in staat zijn klassenverschillen in een complexe ruimte te vangen, zou het gebruik van dergelijk kernels een positief effect kunnen hebben op de classificatietaak.

Uiteindelijk is het dus lastig om een precieze oorzaak voor de teleurstellende prestaties aan te wijzen. Voor de taak van het classificeren van individuele weblogberichten in het algemeen, lijkt het experimenteren met een andere kernel de meest interessante optie op uit te proberen, omdat dit de matige resultaten in beide onderzoeken zou kunnen verklaren. Als wordt gekeken naar het bepalen van de sentimentele oriëntatie van weblogberichten en hoe de vage verschillen tussen de verschillende stemmingen duidelijker gemaakt kunnen worden, is de optie om een andere methode voor sentimentanalyse te gebruiken meer voor de hand liggend.

6 Conclusie

In dit onderzoek is gekeken naar de invloed van sentimentanalyse bij het voorspellen van de stemmingen van weblogberichten. Hiervoor is gebruik gemaakt van een methode van Kamps et al. [6] om de sentimentele oriëntatie van adjectieven te bepalen en deze zijn daarna toegepast op een corpus van weblogberichten uit LiveJournal. Er is op drie verschillende manieren naar de data in het corpus gekeken: eerst door een grafische representatie middels zogenaamde stemmingwolken te maken, daarna door te kijken hoe enkele classificatiealgorithmen scoren met sentimentswaarden als featureset en vervolgens door te kijken of de nauwkeurigheid van de classificatie te maximaliseren is door geleidelijk meer trainingsinstanties te gebruiken. Ten slotte is het tweede experiment ook nog herhaald met een andere featureset, namelijk die van Mishne [8].

De in dit onderzoek uitgevoerde experimenten zijn één voor één terug te

koppelen aan de deelvragen uit de inleiding. In het eerste experiment is gekeken naar de distributie van berichten van verschillende stemmingen. Als uitgangspunt zijn vier verschillende type stemmingen herkend en van elk van deze vier typen zijn tien voorbeelden bekeken. Na afloop van het experiment blijken de vooraf gestelde hypothesen niet uit te komen. De verschillen in distributie tussen de verschillende typen stemmingen zijn niet erg duidelijk zichtbaar en in de meeste gevallen zijn de verschillen minimaal. Het blijkt totaal niet mogelijk om op basis van de grafiek van de distributie van een willekeurige stemming het type stemming te kunnen achterhalen. Het enige opvallende in de grafieken is dat de meeste berichten samenklonteren rond het nulpunt.

De onduidelijke visuele verschillen tussen verschillende type stemmingen geven een goede indruk over de moeilijkheid van de classificatietaak. Dit blijkt ook uit experiment 2, waarin de classificatieresultaten erg tegen vallen met als maximale score een verbetering van 5-6% ten opzichte van de baseline, en dit zelfs met slechts enkele stemmingen. Het gros van de stemmingen haalt een score van rond de 50% nauwkeurigheid, waarmee er dus nauwelijks beter wordt gepresteerd dan de baseline. Waar het de gemiddelde nauwkeurigheid betreft, is de keuze in classificatiealgoritme hierbij nauwelijks van invloed, aangezien deze bij zowel SMO, als bij Naive Bayes en IBk nagenoeg gelijk is. Voor specifieke stemmingen maakt de keuze van classificatiealgoritme echter wel uit: bij sommige stemmingen kan het verschil per algoritme zo'n 5% bedragen.

Verder valt op dat het toevoegen van trainingsdata nauwelijks voor verbetering zorgt. Dit is het duidelijkst zichtbaar in experiment 3, waar een afnemend stijgende curve valt waar te nemen wat betreft het percentage correct voorspelde berichten. Hoewel het te verwachten is dat de stijging van de nauwkeurigheid niet oneindig zou zijn, valt de stijging van ongeveer 1% bij een verdubbeling van het aantal trainingsinstanties erg tegen. Hier wegen de kosten van het toevoegen van meer trainingsinstanties vrijwel niet op tegen de opbrengsten, waardoor er gesteld kan worden dat het gebruiken van een groot aantal trainingsinstanties tamelijk zinloos is in de context van deze experimenten.

De resultaten van de bovengenoemde experimenten suggereren dat de gebruikte featureset waarschijnlijk niet goed in staat is de rijkheid van de weblogteksten te representeren. Daarom is ook nog in experiment 4 gekeken of de featureset wel van toegevoegde waarde is als deze wordt gebruikt als aanvulling op een andere featureset. De resultaten hiervan blijken in vergelijking met de resultaten dit de eerdere experimenten een stuk beter, maar geven de indruk dat dit vooral komt door het toevoegen van de andere features. Vergeleken met de resultaten van Mishne wordt er gemiddeld namelijk slechter gescoord. De bijdrage van sentimentanalyse aan de featureset lijkt dus eerder negatief te zijn.

Algeheel valt dus de zeggen dat het gebruik van alleen sentimentanalyse binnen dit corpus een te magere featureset oplevert om de teksten goed te kunnen representeren. Ook als toevoeging op een bestaande featureset bieden sentimentfeatures zo op het eerste gezicht weinig aanvulling. Het antwoord op de hoofdvraag uit dit onderzoek is dan ook dat het gebruik van sentimentele oriëntatie het voorspellen van stemmingen van weblogberichten niet verbetert, maar als toevoeging aan een featureset zou het nut kunnen hebben, hoewel dat hier niet direct is aangetoond. Dit dient dus nader onderzocht te worden, waarbij gekeken moet worden naar alternatieve methodes voor sentimentanalyse, aangezien het er met de hier gebruikte methode naar uitziet dat sentimentele oriëntatie vooral voor minder duidelijke klassen zorgt.

Referenties

[1] D Aha and D Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[2] E.S. Boese and A.E. Howe. Effects of web document evolution on genre classification. *CIKM 2005, Bremen, Germany*, 2005.

[3] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM 2005*, pages 617–624, 2005.

[4] V. Hatzivassiloglou and K.R. McKeown. Predicting the semantic orientation of adjectives. In *In Proccedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181, 1997.

[5] G.H. John and P. Langle. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[6] J. Kamps, M. Marx, R. Mokken, and M. De Rijke. Using wordnet to measure semantic orientations of adjectives. In *Proceedings LREC 2004*, volume IV, pages 1115–1118, 2004.

[7] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–649, 2001.

[8] G. Mishne. Experiments with mood classification in blog posts. *In Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access*, 2005. At Sigir 2005.

[9] G. Mishne and M. De Rijke. Capturing global mood levels using blog posts. *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[10] C.E. Osgood, G.J. Succi, and P.H. Tenenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana IL, 1957.

[11] J. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT-Press, 1998. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf, C. Burges and A. Smola (editors).

[12] P.D. Turney and M.L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346, 2003.

[13] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal taxonomies for sentiment analysis. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005)*, 2005.

[14] T. Wilson, J. Wiebe, and P. Hoffman. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

[15] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.