university of groningen / faculty of arts

Sarcasme in tweets: aanwijzingen voor de detectie van sarcasme

Thijs Horstman

Bachelorscriptie
Informatiekunde
Thijs Horstman
s2569205
12 juni 2016

SAMENVATTING

Bij sentimentanalyse wordt zelden rekening gehouden met sarcasme. De prestaties van sentimentsystemen kunnen verbeteren bij het toepassen van sarcasmedetectie. Aanwijzingen voor de detectie van sarcasme in tweets worden in dit onderzoek besproken. Tweeduizend tweets uit 2015 met de hashtag #sarcasme of #not worden gebruikt om een SVM-classifier te trainen. Features gericht op interpunctie en inhoud worden gebruikt. Met machine learning wordt een handmatig geannoteerde set van 230 sarcastische tweets (aangevuld met niet-sarcastische tweets) gebruikt om de prestaties te meten bij verschillende verhoudingen sarcastische en niet-sarcastische tweets. Het systeem haalt een nauwkeurigheid van 62 procent bij een set van 50 procent sarcastische en 50 procent niet-sarcastische tweets. Intensifiers, ellipsen, uitroeptekens en in mindere mate emoticons zijn de beste voorspellers van sarcasme. Het systeem presteert beter dan de baseline, maar het blijkt niet mogelijk om alle tweets goed te classificeren.

INHOUDSOPGAVE

VOORWOORD

Aan het begin van de scriptieperiode werd het thema *sentimentanalyse* voor de Bachelorscriptie Informatiekunde gepresenteerd. Al eerder dit jaar heb ik mij beziggehouden met het automatisch bepalen van de sentimentwaarde van tweets. Om me niet weer met ongeveer hetzelfde bezig te houden, besloot ik dat een onderzoek op het gebied van sarcasmedetectie interessant zou kunnen zijn. Omdat de detectie van sarcasme niet vanzelfsprekend is, leek mij dit juist een interessant en uitdagend onderwerp voor mijn scriptie.

Ik wil graag mijn familieleden die mij hebben geholpen bij het annoteren van tweets bedanken voor hun tijd. Daarnaast bedank ik mijn scriptiebegeleidster voor haar goede hulp en flexibiliteit.

1 INTRODUCTIE

De detectie van figuurlijk taalgebruik is in veel wetenschapsgebieden moeilijk. Toch kan een goede detectie erg belangrijk zijn. Met name op het gebied van sentimentanalyse is het herkennen van stijlfiguren, sarcasme en ironie belangrijk. Systemen op het gebied van sentimentanalyse beoordelen of een (deel van een) tekst als positief, negatief of neutraal opgevat kan worden. Er wordt gebruik gemaakt van natuurlijke taalverwerking en het heeft als doel de emotie die achter een tekst schuilgaat te extraheren. Zo kan sentimentanalyse worden ingezet voor het voorspellen van verkiezingsuitkomsten, het peilen van reacties op reclamespotjes of het voorspellen van aandelenprijzen. Wanneer hierbij figuurlijk taalgebruik buiten beschouwing wordt gelaten, geeft dit een vertekend beeld. Mensen gebruiken regelmatig positieve woorden voor een negatieve uiting. Ze bedoelen dan juist het tegenovergestelde van wat ze schrijven. Dit soort uitingen kunnen vervolgens eenvoudig verkeerd worden geïnterpreteerd door automatische systemen. Om de prestaties te verbeteren, is het van belang om sarcasme te kunnen detecteren. Sarcasme- en sentimentdetectie gaat het eenvoudigst bij korte teksten, zoals die in microblogs voorkomen. De onderzoeksvraag waarop antwoord wordt gegeven: kan sarcasme in tweets worden gedetecteerd door tekstpatronen te analyseren? Een subvraag die hierbij hoort, is welke eigenschappen van belang zijn voor de detectie van sarcasme.

Hoewel er een verschil is tussen sarcasme en ironie, wordt sarcasme in dit onderzoek gebruikt als verzamelterm voor alle vormen van sarcasme, ironie en cynisme. Traditioneel is sarcasme direct, terwijl ironie indirect is. Sarcasme is bijna altijd negatief en opzettelijk; ironie kan ook positief en onopzettelijk zijn. Bij ironie is de betekenis tegengesteld aan de letterlijke interpretatie. Ironie en sarcasme worden als sterk gerelateerd aan elkaar gezien (Attardo, 2007). Het is dan ook niet verwonderlijk dat sarcasme bij de meeste mensen het idee van een tegengestelde betekenis oproept.

Er worden tweets verzameld die de hashtag #sarcasme of #not bevatten. Ironie en sarcasme zullen door elkaar gebruikt worden onder deze hashtags, omdat Twitter-gebruikers niet altijd het verschil tussen beide zullen weten. Eigenschappen die mogelijk iets zeggen over sarcasme worden gebruikt om met behulp van *machine learning* te kunnen voorspellen of tweets sarcastisch of niet-sarcastisch zijn.

Eerst zal een overzicht worden gegeven van bestaand onderzoek en gerelateerde literatuur. Vervolgens wordt de collectie en annotatie van tweets besproken in hoofdstuk 3. In hoofdstuk 4 wordt de gehanteerde methode besproken. Er wordt afgesloten met een overzicht van de resultaten en de conclusie.

2 ACHTERGROND

In vergelijking met andere gebieden op het vlak van sentimentanalyse, is er vrij weinig onderzoek gedaan naar het automatisch detecteren van sarcasme. Bestaand werk richt zich veelal op het Engels en het Spaans, al is er ook een onderzoek gedaan met Nederlandstalige tweets. In de meeste onderzoeken wordt gebruik gemaakt van korte `User Generated Content` (UGC). Blogs - met name microblogs - zijn door sociale media de afgelopen jaren sterk in opkomst. Twitter is (een van) de meestgebruikte microbloggingsdienst(en). In maximaal 140 tekens kunnen gebruikers aangeven wat ze aan het doen zijn of wat hun mening is over een bepaald onderwerp. Door hun geringe lengte lenen tweets zich goed voor sarcasmedetectie. Het blijkt daarentegen lastig om duidelijk aan te wijzen welke eigenschappen typerend zijn voor sarcastische uitingen. Daarbij helpt het ook niet dat sarcasme taal- en cultuurgebonden kan zijn.

2.1 SARCASME EN SENTIMENTANALYSE

In "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis" laten Maynard en Greenwood (2014) zien wat het belang van sarcasmedetectie is. Hun experimenten laten zien dat sentimentdetectie door het correct detecteren van sarcasme flink kan worden verbeterd. Zoals velen, maken ook Maynard en Greenwood geen onderscheid tussen ironie en sarcasme. Ze definiëren sarcasme als 'datgene waar de tegenovergestelde betekenis wordt bedoeld'. De zin "Wat houd ik ervan om door de regen naar het werk te lopen" krijgt met bovenstaande definitie een negatieve sentimentwaarde.

Maynard en Greenwood proberen vooral een probleem van de reikwijdte van sarcasme op te lossen. In bijvoorbeeld "Ik ben niet blij dat ik vanochtend om 5:15 uur wakker werd #goedestart #sarcasme" is sarcasme alleen van toepassing op #goedestart en niet op de rest. Bruikbare sentimentinformatie kan dus ook in hashtags voorkomen. Hashtags zijn echter moeilijk te identificeren, want ze worden normaliter als een enkel woord getokeniseerd. Hashtags bestaande uit meerdere woorden (#goedestart, #echtniet) worden dan als één woord gezien. Ze proberen individuele tokens uit de hashtag te halen en deze te matchen met bestaande woorden, locaties en organisaties uit een speciaal woordenboek. Vervolgens wordt met een Viterbi-algoritme de best mogelijke match opgeslagen. Voor deze hashtag-tokenisatie halen ze een F1-score van 0,97. De implementatie bouwt voort op hun eerdere onderzoek en is helaas alleen bruikbaar voor Engelse woorden. Om de prestaties te meten, hebben ze een corpus van 134 tweets met de hashtag #sarcasm verzameld en geannoteerd. Een sentimentanalysator die geen rekening houdt met sarcasme scoort een F1 van 0,27. De analysator die wel sarcasme detecteert met behulp van hun hashtag-tokenisatie scoort een F1 van 0,77.

Barbieri et al. (2014) pakken het iets anders aan. Zij gebruiken een corpus van zestigduizend tweets, gelijk verdeeld over de categorieën `sarcasme`, `educatie`, `humor`, `ironie`, `politiek` en `nieuwsberichten`. De berichten

zijn op Twitter verzameld door te zoeken naar de corresponderende hashtags. Ze gebruiken geen woordpatronen als *features*, omdat dat veelal taalgebonden is. In plaats daarvan analyseren ze elk woord in elke tweet onder andere op frequentie, onverwachtheid, intensiteit, aantal beschikbare synoniemen en amiguïteit. Onverwachtheid is bijvoorbeeld een kenmerk van ironie. Veelvoorkomende of juist zeldzame synoniemen van een woord kunnen iets zeggen over de mate van sarcasme. Voor de evaluatie wordt elke categorie vergeleken met de categorie sarcasme (sarcasme vs. educatie, sarcasme vs. humor etc.) Het blijkt dat sarcasme het best van nieuwsberichten kan worden onderscheiden. Ook in de categorieën educatie en politiek worden goede scores gehaald. Het is moeilijk om sarcasme van ironie te onderscheiden. Hieruit blijkt nogmaals de overlap tussen sarcasme en ironie. Een tekortkoming aan het onderzoek is dat ze geen rekening houden met de vaardigheden van de Twitter-gebruikers. Zij kunnen het verschil tussen sarcasme en ironie niet allemaal weten en #sarcasme en #ironie zullen waarschijnlijk door elkaar worden gebruikt.

2.2 AANWIJZINGEN VOOR SARCASMEDETECTIE

Carvalho et al. (2009) laten zien dat door *orale* en *gesturele* aanwijzingen te bestuderen, sarcasme relatief goed voorspeld kan worden. Ze gebruiken hiervoor Spaanstalige UGC. Er wordt onderscheid gemaakt tussen positieve en negatieve uitingen, maar ze richten zich vooral op de positieve ironische uitingen. Daarbij wordt gekeken naar verbale ironie als overkoepelende term voor sarcasme, ironie, hyperbolen, retorische vragen en dergelijke. Om de ironie te kunnen detecteren, maken Carvalho et al. gebruik van acht verschillende features. Sommige features zijn alleen bruikbaar voor het Spaans. Verkleinwoorden, aanwijzende voornaamwoorden, werkwoordmorfologie en zogenaamde cross-constructies zijn specifiek gericht op de Spaanse taal. Tussenwerpsels, overmatige interpunctie, aanhalingstekens en allerlei gelachsuitdrukkingen zijn ook bruikbaar in andere talen.

Na analyse van korte teksten (met een gemiddelde lengte van ongeveer vier zinnen) blijkt dat de productiefste features juist direct gerelateerd zijn aan leestekens, interpunctie en veelvoorkomende tekstuele tekens of emoticons (`"lol"`, `";-)"`, `":P"`). Van de teksten die matchen met de gelachsuitdrukkingen en de aanhalingstekens blijkt maar liefst respectievelijk 85 en 68 procent ironisch te zijn. Uit het onderzoek van Carvalho et al. (2009) blijkt dat de meer gestructureerde taalkundige patronen weinig bijdragen aan de herkenning van ironie.

Een ander onderzoek naar sarcasmedetectie is gedaan aan de hand van Nederlandstalige tweets (Liebrecht et al., 2013). Ze verzamelden een groot corpus (circa 78 duizend tweets) met de hashtag #sarcasme. Daarbij gaan ze ervan uit dat de tweets met een dergelijke hashtag sarcastisch zijn. Uit een steekproef blijkt dat 85 procent van de tweets met #sarcasme ook daadwerkelijk sarcastisch is. Een *machine learning*-classifier vergelijkt een testset van 135 sarcastische tweets met een dag aan twitterdata (ruim 3 miljoen tweets). De hashtags zijn verwijderd uit de testset. Ze halen een nauwkeurigheid van 75 procent. De genoemde nauwkeurigheid is echter het percentage correct voorspelde tweets van de 135 sarcastische tweets. De testset bestaat daarnaast uit ruim 3 miljoen niet-sarcastische tweets. Hiervan worden er ook bijna 490 duizend onterecht als sarcastisch geclassificeerd. Een analyse toont aan dat tweets die synoniemen van sarcasme (ironie, cynisme) bevatten als

goede aanwijzingen voor sarcasmedetectie gelden. Waar deze aanwijzingen in het onderzoek van Liebrecht et al. (2013) ook als features worden gebruikt, worden in dit onderzoek #sarcasme en #not juist alleen gebruikt om de tweets te verzamelen. Ze dan ook als features gebruiken geeft vertekende resultaten. Woorden als humor, LOL, en not (met of zonder hashtag) blijken goede voorspellers in het onderzoek van Liebrecht et al. (2013). Ook noemen ze sommige intensifiers (`geweldig, prachtig`) en positieve woorden (`fijn, gezellig, leuk`) als goede voorspellers. Verder wordt in een verkennend onderzoek van Van Hee et al. (2016) genoemd dat contrasterende sentimentwaarden goede indicators voor de detectie van sarcasme kunnen zijn. Ze observeerden dat in hun corpus in de meeste gevallen een positief sentiment werd gebruikt, terwijl een negatief sentiment werd beoogd.

Tot slot geven Joshi et al. (2016) in "Automatic Sarcasm Detection: A Survey" nog een overzicht van bestaand werk over sarcasmedetectie en sentimentanalyse. Tweets worden veel gebruikt en veelal verzameld via de Twitter API. Sommige onderzoekers annoteren alle tweets handmatig, terwijl het merendeel vertrouwd op hashtags. Voor relevante eigenschappen voor sarcasmedetectie gaan sommige onderzoekers zelfs zo ver dat ze informatie over de context halen uit de gebruikersnaam, de historische posts van die gebruiker of het onderwerp van de conversatie.

3 DATA EN MATERIAAL

3.1 COLLECTIE

Als datacollectie wordt bij het herkennen van sarcasme veelal gebruik gemaakt van korte teksten. Hoe langer de tekst, hoe meer mogelijke uitingen van sarcasme in dezelfde tekst te vinden zijn. Alsnog kan dan worden beoordeeld of de tekst in het geheel sarcastisch is of juist niet, maar kortere teksten zijn eenvoudiger te analyseren. Tweets zijn populair vanwege hun grote aantal en beperkte lengte. Bovendien zijn ze eenvoudig te verzamelen. Dat tweets af en toe wat *noisy* of ongrammaticaal zijn, is meestal niet zo belangrijk. Ik maak dan ook gebruik van (Nederlandstalige) tweets.

Voor het verzamelen van tweets is gebruik gemaakt van een corpus met Nederlandstalige tweets van de Rijksuniversiteit Groningen. Dit corpus is beschikbaar voor iedereen van de universiteit. Het bevat (bijna) alle Nederlandstalige tweets die de afgelopen jaren op Twitter zijn geplaatst. Het systeem verzamelt continu alle tweets die gebruikers op Twitter plaatsen. Een taaldetectiesysteem zorgt er aan de hand van een lijst met frequente Nederlandse woorden voor dat alleen Nederlandstalige tweets overblijven. Elk uur wordt een gecomprimeerd bestand opgeslagen in JSON-formaat. Het bevat voor elke tweet onder andere een uniek id, de gebruikersnaam, de tweet en een getokeniseerde versie van de tweet.

De datacollectie die gebruikt wordt in dit onderzoek is opgebouwd door in het corpus te zoeken naar tweets die #sarcasme of #not bevatten. Tweets van heel 2015 zijn hiervoor gebruikt. Van de bijna 170 miljoen tweets, bevat maar een klein gedeelte de hashtags #sarcasme en #not. Na het verwijderen van dubbele tweets (retweets, `RT's`), levert het zoeken ruim 1.500 tweets op met #sarcasme en ruim 15.000 tweets met #not.

Om een eventueel verschil tussen #sarcasme en #not uit te sluiten, heb ik gekozen voor een verhouding tussen beide van 1:1. Dit geeft een totale set van 3000 sarcastische tweets. Deze set is vervolgens opgedeeld in een trainset (2000 tweets), een testset (500 tweets) en een developmentset (500 tweets). Tweets zijn willekeurig over de verschillende sets verdeeld. Naast de sarcastische tweets, worden alle sets aangevuld met verschillende hoeveelheden niet-sarcastische tweets (achtergrondtweets). Dit zijn willekeurige tweets uit 2015 die geen #sarcasme en geen #not bevatten. Om te controleren of het systeem verschillende nauwkeurigheden haalt bij wisselende verhoudingen tussen sarcastische en niet-sarcastische tweets, zijn er steeds vijf versies van een set. Dit zijn sets met verhoudingen van 50/50, 60/40, 70/30, 80/20 en 90/10 tussen niet-sarcastische en sarcastische tweets.

3.2 ANNOTATIE

Tweets worden geclassificeerd als `sarcastisch` of `niet-sarcastisch`. Ik ga ervan uit dat de tweets met #sarcasme of #not als sarcastisch kunnen worden beschouwd. Eerder onderzoek laat zien dat uit een steekproef van tweets met #sarcasme, 85 procent daadwerkelijk als sarcastisch beschouwd kan

worden (Liebrecht et al., 2013). Hoewel ik tevens gebruik maak van #not, verwacht ik dat dit niet veel zal afwijken van de bevindingen van Liebrecht et al. Voor de trainset krijgen alle tweets die #sarcasme of #not bevatten direct het label sarcastisch. Alle aanvullende achtergrondtweets krijgen het label niet-sarcastisch. De tweets in de testset met #sarcasme en #not worden handmatig geannoteerd.

3.2.1 Annotators

Tweets met de hashtags #sarcasme en #not zullen grotendeels sarcastisch zijn, maar voor een hogere betrouwbaarheid dient de testset handmatig geannoteerd te worden, zodat een 'gouden standaard' ontstaat. Hiervoor is een webpagina (zie bijlage, Figuur 4) opgezet. De 500 tweets zijn in een database geladen. Op de webpagina wordt steeds één tweet getoond, waarna gebruikers kunnen aangeven of de tweet als sarcastisch danwel niet-sarcastisch kan worden opgevat. De hashtags zijn hiervoor niet eerst uit de tweet verwijderd, omdat de tweets in beginsel allemaal sarcastisch horen te zijn. Tweets worden in willekeurige volgorde getoond totdat iedere tweet twee keer is beoordeeld. In totaal hebben tien verschillende personen bijgedragen aan de beoordeling van de tweets. Gemiddeld hebben ze elk 100 tweets beoordeeld. Het is niet uit te sluiten dat sommige tweets tweemaal door dezelfde persoon zijn beoordeeld.

Tweets met #sarcasme en #not in de developmentset zijn door mijzelf eenmaal handmatig gecontroleerd. Net als in de testset, zijn tweets met een dergelijke hashtag die toch niet als sarcastisch te interpreteren zijn uit de set verwijderd.

3.2.2 Inter-annotator overeenstemming

Om de mate van overeenstemming tussen de verschillende annotators te berekenen, wordt gebruikt gemaakt van Cohen's kappa coëfficiënt (Cohen, 1968). De coëfficiënt geeft de mate van overeenstemming tussen verschillende annotators weer en levert een getal op tussen -1 en 1. Een waarde van 1 geeft een perfecte overeenstemming aan. Bij 0 is er geen overeenstemming en bij -1 is er sprake van tegenovergestelde overeenstemming (Manning et al., 2009). De formule luidt als volgt:

$$\kappa = \frac{P_A - P_E}{1 - P_E}$$

Hierbij is P_A de verhouding tussen hoe vaak de annotators het met elkaar eens zijn. P_E is verhouding tussen hoe vaak ze bij toeval hetzelfde zouden kiezen.

Tabel 1: Kappa-resultaten voor de annotatie van de testset

	Annotatie 2		
Annotatie 1	Sarcastisch	Niet-sarcastisch	Totaal
Sarcastisch	237	71	308
Niet-sarcastisch	95	97	192
Totaal	332	168	500

Uit Tabel 1 blijkt dat het voor de verschillende annotators vrij moeilijk is om aan te geven of een tweet daadwerkelijk als sarcastisch kan worden

beschouwd. Van 403 van de 500 tweets vinden ze dat de tweet mogelijk sarcastisch is. Slechts 237 van de 500 tweets worden door beide annotators duidelijk als sarcastisch beschouwd. Ze bereiken over 237 + 97 = 334 tweets overeenstemming (sarcastisch danwel niet-sarcastisch). Hieruit volg een P_A van 0,67, een P_E van 0,54 en $\kappa = 0.28$. Dit komt overeen met lichte tot redelijke overeenstemming.

3.3 VERWERKING

Na het toekennen van de labels `sarcastisch` en `niet-sarcastisch` zijn er twee datasets. De trainset zal gebruikt worden om een model te trainen dat op basis van informatie in de trainset nieuwe, ongeziene tweets uit de testset het label sarcastisch of niet-sarcastisch toe kan kennen. Een derde dataset wordt als developmentset gebruikt om het model goed af te stellen en de score op te krikken. Pas wanneer de prestaties van het model bij de developmentset niet meer verbeteren, wordt het model geëvalueerd aan de hand van de testset.

Zoals te zien in Tabel 1, zijn er in de testset in totaal 237 bruikbare sarcastische tweets. Beide annotators vinden dat deze tweets duidelijk als sarcastisch beschouwd kunnen worden. Hiervan zijn 230 tweets gebruikt voor de opbouw van de testsets. De 50/50-set wordt aangevuld met 230 willekeurige niet-sarcastische tweets. De 60/40-set heeft 276 niet-sarcastische en 184 sarcastische tweets, de 70/30-set 322 niet-sarcastische en 138 sarcastische tweets, etc. Er ontstaat zo steeds een testset van 460 tweets.

Voor de trainset geldt hetzelfde principe. Hiervoor zijn 2000 sarcastische tweets beschikbaar. Er worden weer sets gemaakt in de verhoudingen 50/50, 60/40, 70/30, 80/20 en 90/10. Iedere trainset bestaat in totaal steeds uit 4000 tweets.

4 METHODE

Voor de uitvoering wordt gebruik gemaakt van *supervised machine learning*: de te voorspellen categorieën (`sarcastisch` en `niet-sarcastisch`) zijn vooraf bekend. Er wordt gebruik gemaakt van Python in combinatie met scikit-learn voor machine learning. Alle tweets uit de trainsets worden geanalyseerd op features die mogelijk iets zeggen over de aanwezigheid van sarcasme. Iedere tweet wordt zodoende gerepresenteerd als een *featurevector*. Wanneer bijvoorbeeld zowel de aanwezigheid van een bepaald woord in een tweet en het totaal aantal woorden in een tweet iets zouden kunnen zeggen over de aanwezigheid van sarcasme, heeft de featurevector van elke tweet een lengte van twee. Een gecombineerde featurevector voor alle tweets in een set, zou er dan uit kunnen zien als [[0, 2], [0, 3], [1, 2], etc]. De getallen corresponderen in dit geval met een bepaald woord of het aantal woorden in de tweet. In tegenstelling tot wat Liebrecht et al. (2013) deden, is de punctuatie niet verwijderd uit de tweets. Punctuatie kan juist belangrijke aanwijzingen voor sarcasme bevatten. Alle tweets zijn getokeniseerd, zodat elk woord of leesteken afzonderlijk geanalyseerd kan worden. Daarnaast zijn de hashtags waarmee de tweets verzameld zijn (#sarcasme en #not) niet verwijderd. Zolang er geen feature geïmplementeerd wordt die de aanwezigheid van bovengenoemde hashtags als aanwijzing voor sarcasme beschouwt, maakt dit niet uit. De labels (`sarcastisch` of `niet-sarcastisch`) van elke tweet in de trainset zijn bekend. De featurevectors worden samen met de bijbehorende labels verwerkt door een machine learning-algoritme. Het algoritme probeert gezamenlijke patronen te vinden die relevant zijn voor de data. Er ontstaat een getraind model dat probeert te generaliseren, zodat er voorspellingen gemaakt kunnen worden voor nieuwe voorbeelden.

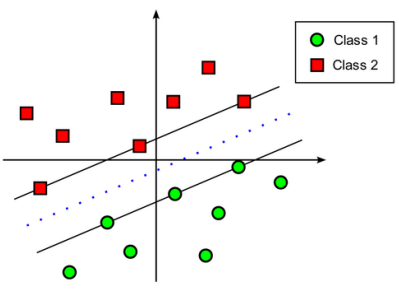
Ook alle tweets in de testsets worden geanalyseerd en omgezet in een featurevector. Deze featurevectors worden door het getrainde model geanalyseerd en vergeleken met de featurevectors uit de trainset. Aan de hand van de voorbeelden in de trainset, kan het model voor elke tweet in de testset voorspellen of het label `sarcastisch` of `niet-sarcastisch` van toepassing is.

4.1 CLASSIFIERS

Er zijn verschillende classifiers beschikbaar die kunnen dienen als machine learning-algoritme. Voor categorische supervised learning zijn SVM's (Support Vector Machines) en Naive Bayes de belangrijkste. Beide zijn geïmplementeerd in scikit-learn.

SVM's[1] worden gebruikt voor binaire classificatie en regressie-analyse. De SVM wijst aan de hand van eerder bepaalde features en het bijbehorende label een tweet toe aan de klasse `sarcastisch` of de klasse `niet-sarcastisch`. De SVM

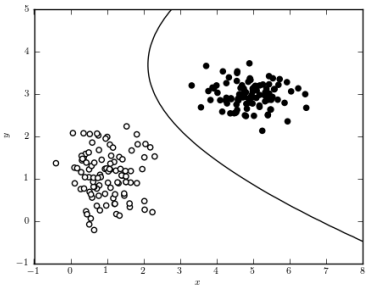
[1] http://scikit-learn.org/stable/modules/svm.html



Figuur 1: SVM met optimale scheidingslijn tussen twee klassen (lineaire kernel)

kan gebruik maken van een lineaire kernel of een soepelere, niet-lineaire kernel. De tweets worden ruimtelijk weergegeven waarna de SVM een scheiding tussen de klassen bepaalt. Om de optimale scheiding te bepalen, komt de scheiding in een 'hypervlak' te liggen (zie Figuur 1). Het hypervlak zorgt ervoor dat de marge tussen de scheiding maximaal is.

De Naive Bayes classifier[2] is gebasseerd op het theorema van Bayes. Het beschrijft de kans op een gebeurtenis aan de hand van voorwaarden die er mogelijk aan gerelateerd zijn. Belangrijk bij de Naive Bayes classifier is de voorwaarde dat features onafhankelijk van elkaar zijn. De classifier probeert op een 'naïeve' manier te classificeren. Het legt daarbij geen verbanden tussen verschillende features. Figuur 2 laat een fictieve, optimale scheiding tussen twee klassen zien.



Figuur 2: Naive Bayes met optimale grens tussen twee fictieve klassen

4.2 FEATURES

De prestatie van het model hangt voor een groot deel af van de gebruikte features. Hoe meer een feature specifiek gericht is op sarcastische tweets, hoe accurater het model nieuwe tweets juist kan classificeren. De zes onderstaande features zijn geïmplementeerd en zeggen mogelijk iets over de aanwezigheid van sarcasme in een tweet.

Uitroepteken (F_{excl})

Uitroeptekens komen vrij veel voor in sarcastische teksten. Wanneer in de tweet één of meerdere uitroeptekens voorkomen, krijgt de feature F_{excl} de waarde 'exclamation'. Als er geen uitroeptekens in de tweet voorkomen, krijgt F_{excl} de waarde 'no_exclamation'.

Vraagtekens ($F_{question}$)

In sarcastische tweets komen gemiddeld meer vraagtekens voor dan in niet-sarcastische tweets. De aanwezigheid van een enkel vraagteken blijkt echter niet kenmerkend voor sarcasme. Ook in niet-sarcastische tweets komen regelmatig vraagtekens voor. Dubbele vraagtekens ("??") lijken wel relatief vaker voor te komen in sarcastische tweets. Bij de aanwezigheid van twee opeenvolgende vraagtekens krijgt $F_{question}$ de waarde 'question'. Komen er geen vraagtekens voor of komt er slechts één vraagteken voor, krijgt $F_{question}$ de waarde 'no_question'.

[2] http://scikit-learn.org/stable/modules/naive_bayes.html

Ellips (F_{ellips})

Bij het doornemen van sarcastische tweets valt op dat er vrij vaak een ellips ('beletselteken') gebruikt wordt. Het wordt gebruikt om aan te geven dat een deel van de zin is weggelaten of om, in dit geval, de sarcasme extra kracht bij te zetten ("Wat een goede klantenservice..."). Op Twitter worden in plaats van drie punten achter elkaar ook regelmatig vier, vijf of nog meer punten gebruikt. F_{ellips} krijgt de waarde 'ellips' als er drie of meer punten achter elkaar in de tweet voorkomen. Zo niet, krijgt F_{ellips} de waarde 'no_ellips'.

Aanhalingstekens (F_{quote})

Aanhalingstekens kunnen een goede voorspeller zijn voor sarcasme in een tweet. Het komt regelmatig voor dat een gedeelte van een uitspraak van iemand wordt gequoot, om deze uitspraak vervolgens in twijfel te trekken. Zie bijvoorbeeld de tweet *"te vroeg om te zeggen of het om een afrekening gaat". Nee, 't was een verjaardagscadeau*. Als (een deel van een) tekst in een tweet tussen aanhalingstekens (' of ") wordt gedetecteerd, krijgt F_{quote} de waarde 'quote'. Anders krijgt F_{quote} de waarde 'no_quote'.

Emoticons ($F_{emoticon}$)

Volgens Carvalho et al. (2009) duidt de aanwezigheid van emoticons en gelachsuitdrukkingen vaak op sarcasme. Vooral emoticons die een positieve emotie uitbeelden zijn belangrijk. $F_{emoticon}$ krijgt daarom de waarde 'emoticon' wanneer er één of meerdere emoticons (":)", ";)", ":-)", ";-)", ":D", ":P", "=)") in de tweet aanwezig zijn. Als er geen positieve emoticons gevonden worden krijgt $F_{emoticon}$ de waarde 'no_emoticon'.

Intensifiers (F_{intens})

De laatste feature is meer inhoudelijk van aard. Woorden die de uiting versterken (*intensifiers*) komen vrij vaak voor in sarcastische tweets. Dit zijn woorden als *geweldig*, *fantastisch* en *ontzettend* (bijvoorbeeld in "De #rabobank heeft zon geweldig nieuwe app. #sarcasme"). Bij tweets die een van deze intensifiers bevatten krijgt F_{intens} de waarde 'intensifier'. Tweets zonder deze woorden krijgen de waarde 'no_intensifier'.

4.3 EVALUATIE

Voor iedere verhouding wordt steeds getraind en getest. Er wordt eerst getraind op alle tweets uit de set 50/50. Nadat alle featurevectors van iedere tweet in de traindata door de classifier zijn geanalyseerd, is het ontstane model in staat om nieuwe tweets (nieuwe featurevectors) het label `sarcastisch` of `niet-sarcastisch` toe te kennen. Daarna wordt op de testset 50/50 getest hoe goed de prestaties van het getrainde model zijn. Evaluatie kan plaatsvinden omdat de labels van de testset bekend zijn, maar het systeem heeft de labels uit de testset nooit te zien gekregen. De prestaties worden uitgedrukt in nauwkeurigheid (percentage goed geclassificeerde tweets) en F1-score (gewogen gemiddelde van precision en recall). De nauwkeurig-

heid wordt vergeleken met de *baseline* van de set. De baseline is 50 procent in de 50/50-set. Bij een willekeurige classificatie is theoretisch immers een score van 50 procent te halen. Hoe hoger de nauwkeurigheid is ten opzichte van de baseline, hoe beter het systeem presteert. Na het trainen, testen en evalueren van de set 50/50, wordt opnieuw getraind, getest en geëvalueerd op de overige verhoudingen. Per feature wordt vervolgens geanalyseerd in hoeverre deze heeft bijgedragen aan de totaalscore.

5 RESULTATEN EN DISCUSSIE

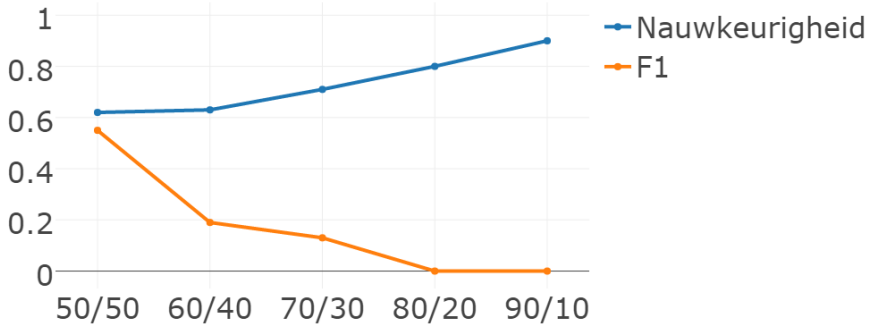
De verschillende features en classifiers blijken in sommige gevallen vrij grote effecten te hebben op de nauwkeurigheid. Op het gebied van nauwkeurigheid behaalt de SVM-classifier in alle gevallen betere resultaten dan de Naive Bayes-classifier. Op het gebied van precision en recall (en dus F1) scoort de Naive Bayes classifier lager, behalve bij de 70/30-set. Voor de resultaten is dan ook gebruikt gemaakt van de SVM-classifier. Het lijkt erop dat er geen duidelijke lineaire scheiding gemaakt kan worden tussen sarcastische en niet-sarcastische tweets. De SVM-classifier presteert beter zonder het forceren van een lineaire kernel.

De resultaten van het systeem bij alle verschillende verhoudingen staan in Tabel 2.

Tabel 2: Nauwkeurigheid, F1, precision en recall van getraind model per set

	set 50/50	set 60/40	set 70/30	set 80/20	set 90/10
Nauwk.	0,62	0,63	0,71	0,80	0,90
F1 (P/R)	0,55 (0,66/0,47)	0,19 (0,71/0,11)	0,13 (0,59/0,07)	N/A	N/A

Doordat de verhouding van sarcastische en niet-sarcastische tweets anders is in de verschillende sets, is de baseline ook voor iedere set verschillend. De 50/50-set bestaat voor 50 procent uit sarcastische en voor 50 procent uit niet-sarcastische tweets. Wanneer een classifier willekeurig labels zou toekennen, is theoretisch een score van 50 procent nauwkeurigheid te halen. Voor de overige verhoudingen geldt het percentage niet-sarcastische tweets als baseline. Het blijkt dat de prestaties het best zijn bij 50/50. De nauwkeurigheid ligt hier 24 procent (12 procentpunten) hoger dan de baseline.



Figuur 3: Nauwkeurigheid en F1-score van de verschillende verhoudingen

Zie Figuur 3 voor een plot van de nauwkeurigheid en F1-score. Hoewel het in eerste instantie lijkt alsof de prestaties beter zijn bij meer niet-sarcastische tweets in de set, is dit ten opzichte van de baseline niet het geval. Bij 60/40 en 70/30 wordt een nauwkeurigheid van slechts enkele

procenten boven de baseline gehaald. Bij 80/20 en 90/10 zijn er in verhouding zo weinig sarcastische tweets, dat alle tweets uit de testset als niet-sarcastisch worden geclassificeerd. Als gevolg hiervan kan er voor 80/20 en 90/10 geen F1-score worden berekend. De F1-score vertoont een dalende lijn. Dit heeft voornamelijk te maken met de slechtere recall van sarcastische tweets. Bij de 50/50-set is de recall van sarcastische tweets nog bijna 0,50. Bij 60/40 daalt deze naar 0,11 en bij 70/30 is de recall slechts 0,07. Er worden in de testset naar verhouding dus steeds minder tweets als sarcastisch geclassificeerd. De precision blijft redelijk constant, dus de tweets die wel als sarcastisch worden geclassificeerd, zijn dat in werkelijheid ook vrij vaak. De beste resultaten worden gehaald bij de 50/50-set. Bij naar verhouding meer sarcastische tweets in de set is de nauwkeurigheid groter. Het is echter onrealistisch om te trainen op een set met meer dan 50 procent sarcastische tweets. In werkelijkheid heeft slechts ongeveer één op de tienduizend tweets de hashtag #sarcasme of #not.

Tabel 3: Positieve featurematches in de trainset (4000 tweets) en de testset (460 tweets) bij 50/50

Feature	**Train**	**Test**
F_{excl}	797	102
$F_{question}$	25	3
F_{ellips}	797	55
F_{quote}	138	23
$F_{emoticon}$	125	18
F_{intens}	227	55
Totaal	**2109**	**256**

Zoals te zien in Tabel 3, levert de feature F_{excl} in zowel de trainset als de testset veel hits op. Ondanks dat uitroeptekens niet alleen in sarcastische tweets voorkomen, blijken ze toch een vrij goede voorspeller voor sarcasme. Ten opzichte van de behaalde nauwkeurigheid van 62 procent in de 50/50-set, daalt de nauwkeurigheid met 2 procentpunt als de feature wordt weggelaten. F_{intens} doet het nog iets beter (-3%-punt). F_{ellips} blijkt de beste voorspeller van sarcasme (-4%-punt). Emoticons zorgen voor een kleine verbetering bij de detectie (-0,5%-punt). De dubbele vraagtekens en de quotes zorgen niet voor een verbetering van de nauwkeurigheid. De slechts drie gevonden dubbele vraagtekens in de testset hebben dus niet geleid tot een verbetering van de sarcasmedetectie.

6 CONCLUSIE

In dit onderzoek is geprobeerd om met behulp van verschillende features sarcasme te detecteren. De detectie blijkt niet eenvoudig. De hoogste nauwkeurigheid wordt behaald wanneer het aantal sarcastische en niet-sarcastische tweets gelijk zijn verdeeld. De ellips en het uitroepteken blijken goede voorspellers van sarcasme. Andere interpunctie, zoals vraagtekens en aanhalingstekens, voegen niet veel toe. Woorden die de uiting versterken (zoals 'geweldig' en 'fantastisch'), zijn ook belangrijk voor een goede sarcasmedetectie. De behaalde nauwkeurigheid is iets lager dan de nauwkeurigheid van 75 procent die Liebrecht et al. (2013) haalden. Het systeem presteert beter dan wanneer willekeurig wordt geclassificeerd, maar een detectiegraad van 100 procent wordt bij lange na niet gehaald. Het is dus mogelijk om sarcasme in tweets te detecteren aan de hand van tekstpatronen, maar het systeem is niet in staat om met grote zekerheid te classificeren. Sarcasme blijft een moeilijk onderzoeksgebied en het halen van een dergelijke score blijft in dit wetenschapsgebied voorlopig een utopie.

De prestaties kunnen verder verbeterd worden door de trainset beter te annoteren. De aanwezigheid van alleen #sarcasme of #not betekent niet altijd dat een tweet ook daadwerkelijk als sarcastisch te interpreteren is. Van 403 van de 500 tweets (81 procent) in de testset werd door annotators aangegeven dat er mogelijk sprake was van sarcasme. Slechts 237 van de 500 tweets (47 procent) vonden beide annotators duidelijk sarcastisch. Hierbij moet wel worden opgemerkt dat de groep annotators uit tien personen bestond, in plaats van de gebruikelijke twee of drie. Omwille van de hoeveelheid tweets is toch gekozen voor meer dan twee annotators. Ze kregen een korte instructie, maar de betrouwbaarheid zou wellicht hoger zijn geweest wanneer slechts twee personen annoteerden. Daarnaast wordt de hashtag #sarcasme of #not door Twitter-gebruikers incidenteel verkeerd gebruikt of gebruikt terwijl de tweet helemaal niets met sarcasme te maken heeft. Het is tegelijkertijd een goede keuze geweest om geen onderscheid te maken tussen sarcasme en ironie. Twitter-gebruikers weten namelijk lang niet altijd het verschil tussen die twee. De trainset zal mogelijk ook gecontroleerd moeten worden voordat deze als 'zilveren standaard' als trainingsdata gebruikt kan worden.

Verder worden alleen tweets behandeld die de hashtag #sarcasme of #not bevatten. Tweets zonder een van deze hashtags kunnen ook sarcastisch zijn. Toekomstig werk moet uitwijzen of het mogelijk is om ook dat soort tweets te classificeren.

BIBLIOGRAFIE

Attardo, S. (2007). Irony as relevant inappropriateness. *Journal of pragmatics 32*(6), 793–826.

Barbieri, F., H. Saggion, and F. Ronzano (2014, April). Modelling sarcasm in twitter, a novel approach. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 50-58.

Carvalho, P., L. Sarmento, M. J. Silva, and E. de Oliveira (2009). Clues for detecting irony in user-generated contents: Oh...!! it's "so easy";-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, New York, NY, USA, pp. 53–56. ACM.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin 70*(4), 213.

Joshi, A., P. Bhattacharyya, and M. J. Carman (2016). Automatic sarcasm detection: A survey. *CoRR abs/1602.03426*.

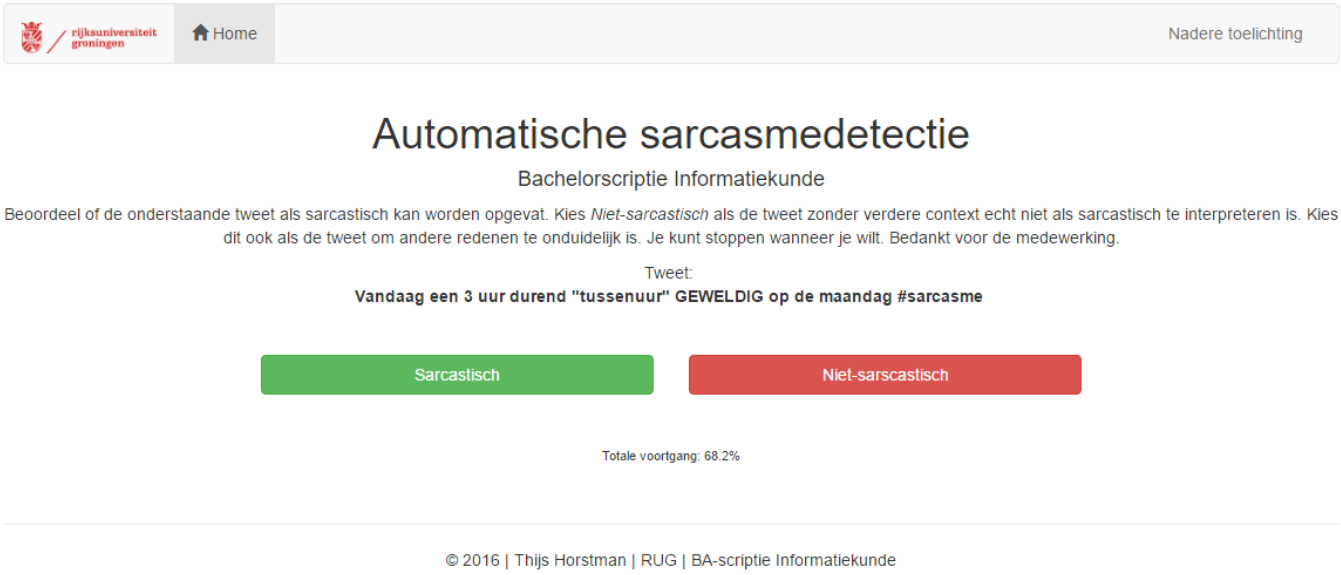
Liebrecht, C., F. Kunneman, and A. van den Bosch (2013, June). The perfect solution for detecting sarcasm in tweets #not. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 29-37.

Manning, C. D., P. Raghavan, and H. Schütze (2009). Evaluation in information retrieval. In C. U. Press (Ed.), *An Introduction to Information Retrieval*, Chapter 8, pp. 151–175. Camgridge, England: Cambridge University Press.

Maynard, D. en M. A. Greenwood (2014). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC*, pp. 4238–4243.

Van Hee, C., E. Lefever, and V. Hoste (2016). Exploring the realization of irony in twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*. European Language Resources Association (ELRA).

A | BIJLAGEN



Figuur 4: Webpagina gebruikt door verschillende annotators voor het beoordelen van tweets in de testset