Domein-afhankelijke sentiment detectie op Twitter

door

Nathan CLAEYS

Scriptie ingediend tot het behalen van de academische graad van
burgerlijk ingenieur in de computerwetenschappen

Academiejaar 2013–2014

Promotor: prof. dr. ir. Bart DHOEDT
Scriptiebegeleider: ir. Steven VAN CANNEYT

Faculteit Ingenieurswetenschappen
Universiteit Gent

Vakgroep Informatietechnologie
Voorzitter: prof. dr. ir. Daniël DE ZUTTER

Samenvatting

In dit artikel beschrijven we een methodologie die probeert een betere methode te voorzien om negatieve tweets te detecteren. Om dit te doen classificeren we tweets in domeinen en voeren een domein-afhankelijke sentiment classificatie uit. Deze domein-afhankelijke sentiment classificatie werkt met een ensemble van classificeerders en schat voor elke tweet een kans in dat deze negatief is. Dit ensemble bestaat uit een algemene classificeerder getraind op alle tweets in de trainingsverzameling en domeinspecifieke classificeerders die enkel getraind zijn op tweets behorend tot hun domein. Voor een gegeven tweet berekent de domeinspecifieke classificeerder dan een probabiliteit dat de tweet negatief is, voor de tweets van zijn domein. Die probabiliteit wordt samengevoegd met de probabiliteit van de algemene classificeerder. Op deze manier kunnen de tweets gerangschikt worden volgens de kans dat ze negatief zijn. We hebben de domeinen van de tweets op verschillende manieren bepaald. Dit is gedaan door gebruik te maken van entiteiten, hashtag clusters en een domein classificeerder. De voorgestelde methode behaalt een gemiddelde precisie van 59,65% tegenover de gemiddelde precisie van 51,70% van de basismethode die enkel de algemene classificeerder gebruikt.

Trefwoorden

Twitter, sentiment classificatie, domein-afhankelijke classificatie

Hoofdstuk 3

Methodologie

3.1 Inleiding

In dit hoofdstuk wordt de methodologie van dit onderzoek uiteengezet. Dit hoofdstuk begint dan ook met het uiteenzetten van de methodologie van de sentimentdetectie van negatieve tweets in sectie 3.2. In sectie 3.3 wordt de domein bepaling uiteengezet, deze is noodzakelijk voor de sentimentdetectie.

3.2 Sentiment Classificatie

In deze sectie wordt uitgelegd hoe het onderzoek naar de sentiment classificatie is uitgevoerd. Hiervoor leggen we eerst uit wat de methodologie van de gebruikte basismethode (*baseline*) is in sectie 3.2.1. Vervolgens wordt uitgelegd wat het concept is achter ons onderzoek naar de verbeterde sentiment classificeerder in sectie 3.2.2. Ten slotte wordt er uitgelegd wat de gebruikte methodologie is van de sentiment classificeerder uit dit concept in sectie 3.2.3.

3.2.1 Basismethode

methode die hier beschreven staat, vormt de basismethode waartegen onze sentiment classificeerder vergeleken wordt. Deze basismethode wordt verder de standaard classificeerder genoemd. De standaard classificeerder heeft als trainingsinput een verzameling van tweets en kan daarna voor elke tweet de probabiliteit geven dat de tweet negatief is. Aangezien de probabiliteit wordt gegeven dat de tweet negatief is, is dit ook automatisch het complement van de probabiliteit

dat een tweet positief is. Dit is dus een classificatie in de twee klassen van positief sentiment en negatief sentiment in de tweet. Alle tweets in de trainingsinput hebben een label dat weergeeft of het een tweet is met een positief sentiment of een tweet met een negatief sentiment.

Om nu de classificeerder te doen werken met de tweets wordt elke tweet t in de trainingsinput K omgezet naar een feature vector. De feature vector waarop de classificeerder werkt is gebaseerd op de *bag-of-words* features. Hierin bevat de feature vector een dimensie voor elke token die voorkomt in de trainingsinput K. De verzameling van al deze tokens en dus alle dimensies is de verzameling W . Voor elke tweet moeten deze dimensies nu waardes krijgen. Hierbij is elke tweet een document dat wordt omgezet naar een *bag-of-words* door tokenisatie. Dit resulteert in de verzameling W_t van alle tokens in de tweet t. Het nadeel van de standaard *bag- of-words* is dat de volgorde van de woorden in de tweet verloren gaat maar dit vereenvoudigt de feature vector wel. De volgende stap is nu de omzetting van de *bag-of-words* naar een feature vector v_t. Deze feature vector v_t met component c_w met $w \in W$ krijgt de waarde 0 voor elke feature token w die niet in de tweet voorkomt en de waarde f_w voor elke token $w \in W_t$ die wel in de tweet voorkomt. Deze waarde, f_w, bestaat uit twee componenten. De eerste component is het genormaliseerde maximum van de absolute frequentie in positieve tweets en de frequentie in negatieve tweets in de verzameling K. Deze component zorgt ervoor dat tokens die veel voorkomen in e´´en klasse, positief of negatief, een hogere waarde krijgen. De tweede component is de ge¨ınverteerde documentfrequentie van de token w. Dit zorgt ervoor dat de tweets die algemeen veel voorkomen een lagere waarde krijgen. De symbolen p_w en n_w staan voor respectievelijk de frequentie waarmee de token w in positieve tweets of negatieve tweets voorkomt in de verzameling K.

$$f_w = \frac{\max(p_w, n_w)}{p_w + n_w} \times \frac{|K|}{p_w + n_w}$$

feature vector is gekozen na vergelijkingen met de binaire en token frequentie feature vectoren. Omdat uit deze initi¨ele experimenten bleek dat deze een mindere nauwkeurigheid gaven beschouwen we deze verder niet in ons onderzoek. Deze feature vector wordt uiteindelijk gebruikt in het classificeringsalgoritme om te trainen en te classificeren.

Deze feature vector kan door verschillende classificatie algoritmes gebruikt worden om de negatieve tweets te detecteren. Het classificeringsalgoritme dat in dit onderzoek wordt gebruik is Naive Bayes met het Multinomiale model [3, 5]. Bij initiële experimenten is dit classificeringsalgoritme vergeleken met Support Vector Machines [3, 5, 9], hieruit bleek dat deze minder goed

werkt en ebben we deze verder niet beschouw komt overeen met de resultaten van vorig onderzoek [3]. De Naive Bayes Multinomial classificeerder berekent de negatieve probabiliteit $P_t(\text{negatief}|CL)$, waarbij CL het classificeerder model is dat het resultaat is van de training. De probabiliteit $P_t(\text{negatief}|CL)$ wordt berekend door gebruik te maken van de a priori probabiliteit dat een tweet negatief is $P(\text{negatief}|CL)$ en de probabiliteiten, $P(w|\text{negatief})\forall w \in W_t$, dat een token w voorkomt in een negatieve tweet t die de token verzameling W_t heeft.

$$P_t(\text{negatief}|CL) = P(\text{negatief}|CL) \cdot \prod_{w \in W_t} P(w|\text{negatief})$$

Deze probabiliteiten wo berekend door middel van het Naive Bayes Multinomial algoritme oor middel van de voorheen gedefini¨eerde feature vectoren.

3.2.2 Concept

Het doel van dit onderzoek is het verbeteren van de standaardmethode om negatieve tweets te vinden. De standaardmethode maakt gebruik van ´e´en classificeerder die een probabiliteit inschat de tweet ne f is onafhankelijk van andere informatie over de tweet. Om deze standaardme- thode te verbeteren wordt in dit onderzoek gebruik gemaakt van de semantiek in tweets, meer bepaald door rekening te houden met de domeinen waartoe een tweet behoort.

Tweets hebben vaak een onderwerp waarover een mening wordt geventileerd. De mening vormt een positief of negatief sentiment over het onderwerp. We definiëren een domein als een verzameling van onderwerpen die semantisch samen horen. De tweet '...ordered call of duty ghosts about a week ago and i didnt get it yet...' heeft bijvoorbeeld als onderwerp het computerspel 'call of duty' en de tweet 'My 3DS XL charging cradle shipped out...' heeft als onderwerp de spelconsole '3DS XL'. Beide onderwerpen maken deel uit van het domein 'Gaming', dus deze tweets behoren tot hetzelfde domein. Daartegenover gaat de tweet 'Can't believe how much my arm is hurting where I've had my flu jab :(' over het onderwerp 'griepvaccinaties' en dus past deze tweet in het domein 'Gezondheid'. Het is belangrijk om op te merken dat tweets ook tot meerdere domeinen kunnen behoren. Een voorbeeld hiervan is de tweet '...Want Facebook To Know That #Jesus Changed By Life...', waarin zowel religie een rol speelt, als sociale media. Deze tweet behoort dus tot zowel het domein 'Internet' als 'Religie'.

Deze domeinen kunnen nuttig zijn voor de sentimentdetectie omdat elk semantisch domein zijn eigen typische woordenschat heeft die in andere domeinen niet gebruikt wordt of daar een andere betekenis heeft. Deze woordenschat is dus domeinspecifiek. Een voorbeeld hiervan in

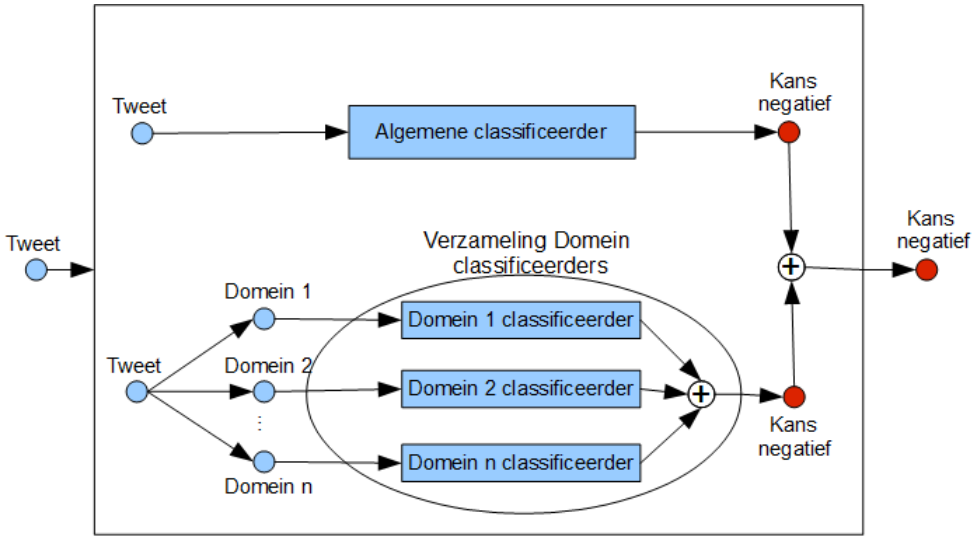
het 'Gaming' domein is de tweet '...next DLC in coming weeks will be a infected horde mode!...' waarin een positief sentiment over een uitbreiding op het spel wordt uitgedrukt. Het woord 'infected' in de tweet betekent echter ontstoken, wat in het domein 'Gezondheid' eerder als negatief beschouwd zal worden, maar hier in het domein 'Gaming' slaat dit op deze uitbreiding, waardoor het positief is. Dit is een voorbeeld van domeinspecifieke woorden die moeilijk op te pikken zijn door een classificeerder die geen kennis heeft over het domein van de tweet aangezien die tegenovergestelde sentimenten over de woorden ontvangt vanuit de verschillende domeinen.

In dit onderzoek wordt dus gebruik gemaakt van het domein van de tweet om tot een betere detectie van de negatieve sentimenten te komen. Omdat de domeinspecifieke woorden moge- lijks niet correct ge¨ınterpreteerd worden wanneer een classificeerder tegenstrijdige sentimenten ontvangt, wordt in dit onderzoek per domein een classificeerder getraind. Dit betek dat de domeinspecifieke classificeerder enkel op tweets van zijn domein wordt getraind. Deze domeinspecifieke classificeerder zou dan in staat moeten zijn om beter gebruik te maken van deze woorden dan een algemene classificeerder die op alle tweets getraind is. Doordat de domeinspecifieke classificeerders slechts een fractie van de totale verzameling tweets kunnen gebruiken, hebben deze echter minder training tweets tot hun beschikking. Dit kan als gevolg hebben dat hun classificatie minder goed is bij woorden die niet specifiek bij het domein horen. Om dit nadeel te compenseren maken we ook gebruik van een algemene classificeerder die op alle tweets getraind is. Een combinatie van de algemene classificeerder en de domeinspecifieke classificeer- ders zou dan een betere inschatting kunnen geven of een tweet negatief is.

3.2.3 Ensemble van classificeerders

De sentiment classificatie heeft als doel om voor elke tweet de probabiliteit in te schatten dat de tweet negatief is. Op die manier kunnen we alle tweets sorteren in een gesorteerde lijst volgens deze probabiliteit. Bij een perfecte classificatie geeft deze lijst dan eerst de meest negatieve tweets ug, wat noodzakelijk is. Om deze sentiment classificatie op te stellen zoals in sectie 3.2.2 wordt er gebruik gemaakt van een ensemble van classificeerders, zie Figuur 3.1. Dit ensemble van classificeerders werkt met een trainingsset K van tweets, een validatieset V en een testset U. Deze sets zijn allemaal volledig gesplitst van elkaar en worden verder beschreven in hoofdstuk
4. Voor elk van deze sets wordt er verondersteld dat de domeinen van de tweets gekend zijn, de domein bepaling van deze tweets gebeurt volgens de methode beschreven in sectie 3.3.

Er zitten twee types van classificeerder in dit ensemble, het eerste type is de algemene classi-



Figuur 3.1: Ensemble van classificeerders

ficeerder die op elke tweet in de trainingsset is getraind. Vervolgens is er de verzameling van domeinspecifieke classificeerders, waarin er een classificeerder is per gegeven domein die enkel getraind is op tweets van dit domein. Hierbij is D de verzameling van alle mogelijke domeinen. De probabiliteit dat een gegeven tweet t negatief is wordt aangegeven door $P_t(\text{negatief})$. Deze probabiliteit wordt bepaald door een gewogen gemiddelde te nemen van twee probabiliteiten. De eerste probabiliteit $P_t(\text{negatief}|A)$ is de probabiliteit dat de tweet negatief is volgens de algemene classificeerder. De tweede probabiliteit $P_t(\text{negatief}|D_t)$ is de probabiliteit dat de tweet negatief is gegeven alle domeinen D_t van de tweet t. Het gewogen gemiddelde maakt gebruik van het gewicht van de domeinspecifieke classificeerder, α, dat tussen 0 en 1 ligt en het complement van dit gewicht, $1 - \alpha$:

$$P_t(\text{negatief}) = (1 - \alpha) \cdot P_t(\text{negatief}|A) + \alpha \cdot P_t(\text{negatief}|D_t)$$

De tweede probabiliteit, $P_t(\text{negatief}|D_t)$, is op zich een gecombineerde probabiliteit van de domeinspecifieke classificeerders die bij deze tweet gebruikt zijn. Om deze probabiliteit in te schatten worden de domeinen van de tweet bekeken. Deze vormen de verzameling D_t van alle domeinen waartoe de tweet behoort. Voor elk van de domeinen van de tweet wordt de tweet opnieuw geclassificeerd door de domeinspecifieke classificeerder. Elk van deze domein classificeerders geeft een probabiliteit $P_t(\text{negatief}|d)$; dit is de probabiliteit dat de tweet t negatief is gegeven het domein d van de verzameling D_t. Alle domein classificeerders $d \in D_t$ hebben ook een gewicht, β_d, tussen 0 en 1. Dit gewicht wordt geoptimaliseerd aan de hand van de validatieset V.

Dit gewicht heeft twee doelen. Ten eerste worden de gewichten van alle domein classificeerders $d \in D_t$ gebruikt om het gewicht α te berekenen. Dit gebeurt door het gemiddelde te nemen van alle gewichten $\beta_d \forall d \in D_t$:

$$\alpha = \frac{\sum_{i=1}^{|D_t|} \beta_i}{|D_t|}$$

Ten tweede worden deze gewichten gebruikt om alle domeinspecifieke probabiliteiten in de kans $P_t(\text{negatief}|D_t)$ te combineren. Om deze probabiliteit in te schatten gebruiken we de probabiliteiten $P_t(\text{negatief}|d) \forall d \in D_t$ en de gewichten $\beta_d \forall d \in D_t$ in een gewogen gemiddelde. Het probleem is nu nog dat de gewichten β_d niet op voorhand op elkaar afgestemd kunnen worden en dus geen totale som van een gewicht van 1 vormen zoals vereist. Dit komt doordat het niet mogelijk is op voorhand te weten welke domeinen D_t allemaal samen zullen voorkomen, zodanig dat het elke combinatie van gewichten anders genormaliseerd moet worden. Het is dus nodig om de gewichten eerst te normaliseren zodanig dat de som van alle gewichten van de domeinen D_t 1 is. Dit gebeurt door de volgende berekening. Hierbij staat β'_d voor het genormaliseerde gewicht van domein d.

$$\beta'_d = \frac{\beta_d}{\sum_{i=1}^{|D_t|} \beta_i}$$

Met deze nieuwe gewichten is het nu mogelijk om de domeinspecifieke probabiliteit $P_t(\text{negatief}|D_t)$ te bekomen dat de tweet negatief is door een gewogen gemiddelde van elke probabiliteit $P_t(\text{negatief}|d)$ met de gewichten β'_d.

$$P_t(\text{negatief}|D_t) = \sum_{i=1}^{|D_t|} \left(\beta'_i \cdot P_t\left(\text{negatief}|D_i\right)\right)$$

Voor de berekening van $P_t(\text{negatief}|A)$ en $P_t(\text{negatief}|d)$ van de respectievelijke algemene classificeerder en domeinspecifieke classificeerders wordt de basismethode gebruikt die beschreven staat in de vorige sectie 3.2.1. Hierbij zijn de gebruikte A en d de classificeerder modellen voor de basismethode van sectie 3.2.1. Deze classificeerders gebruiken dus allemaal dezelfde methode om tweets te classificeren en verschillen dan ook enkel in hun classificeerder model dat afhangt van de tweets die ze krijgen als training. Om te meten hoe goed dit ensemble van classificeerders een negatieve sentiment lijst kan opstellen vergelijken we deze met de basismethode (*baseline*), die eerder beschreven is (sectie 3.2.1).

3.3 Domein Bepaling

3.3.1 Inleiding

In het vorige deel hebben we uitgelegd hoe de detectie van negatieve tweets werkt. Deze methode vereist echter dat het domein van de tweets gekend is. In deze sectie wordt uitgelegd hoe de domein bepaling van tweets werkt. Eerst wordt een methode besproken op basis van entiteiten in de tweets in sectie 3.3.2. Vervolgens wordt een methode uitgelegd die gebaseerd is op de hashtags in tweets in sectie 3.3.3. Ten slotte wordt een domein classificatie uitgelegd die een domein bepaalt voor tweets zonder hashtag in sectie 3.3.4.

3.3.2 Entiteiten

Concepten

Om de methodologie van de domein bepaling via entiteiten duidelijk uit te kunnen leggen, worden hier eerst alle concepten en hun relatie met elkaar uitgelegd. Eerst zijn er de entiteiten: de entiteiten die in dit onderzoek gebruikt worden zijn zelfstandige naamwoorden die een semantisch begrip duidelijk voorstellen. Dit wil zeggen dat het een zelfstandig naamwoord is dat voor slechts ´e´en interpretatie open staat, zoals ‘christendom’ of ‘basketbal’. Vervolgens zijn er de onderwerpen en domeinen overeenkomstig met de uitleg in sectie 3.2.2. Een onderwerp is iets waarover op Twitter gediscussieerd wordt. Elke entiteit komt overeen met een onderwerp, een onderwerp in een tweet kan dus herkend worden aan de entiteit die in de tweet voorkomt. Zo zijn er de entiteiten ‘Playstation’ en ‘Xbox’ die elk een onderwerp voorstellen. De domeinen zijn de klassen waarin we een tweet willen classificeren. Een domein wordt gevormd door een verzameling van onderwerpen die semantisch samen horen. Een domein hoort dus bij meerdere onderwerpen. Zo bestaat het domein ‘Gaming’ uit alle onderwerpen die over computerspellen gaan. Aangezien een onderwerp gelijk is aan een entiteit bestaat een domein dus ook uit de groep entiteiten die overeenkomt met zijn onderwerpen. Ten slotte zijn in deze sectie alle domeinen statisch vastgelegd.

Entiteiten bepalen

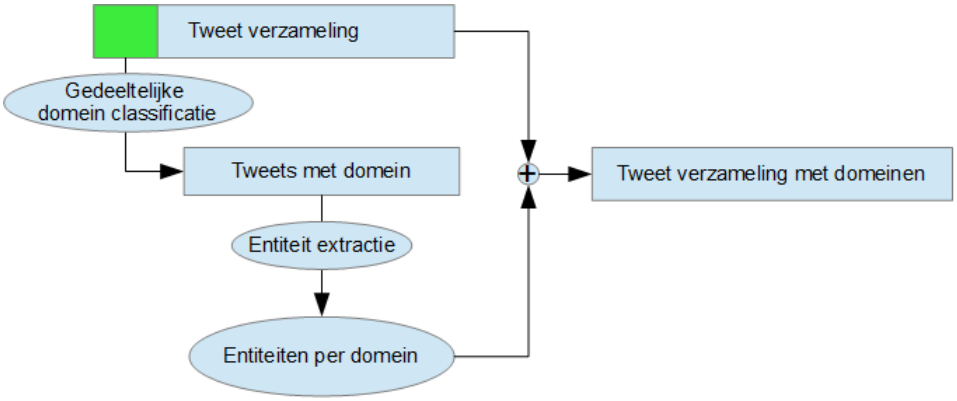
Het bepalen van de domeinen van de tweets gebeurt in verschillende stappen. Het doel is om een verzameling van entiteiten voor elk domein te verkrijgen zodanig dat tweets die een entiteit

Domeinen
Computer en Internet
Gaming
Gezondheid
Kunst en entertainment
Ondernemen
Ontspanning
Recht en Misdaad
Religie
Samenleving en Politiek
Sport
Weer

Tabel 3.1: Originele domeinen van de entiteiten

Domeinen
Beroemdheden
Computer
Gaming
Gezondheid
Internet
Kunst
Recht en Misdaad
Religie
Samenleving en Politiek
Sport

Tabel 3.2: Domeinen van de entiteiten



Figuur 3.2: Domein via entiteiten

bevatten in een domein geclassificeerd kunnen worden, zie Figuur 3.2.

De eerste stap is het vinden van een groep tweets met een domein. Deze tweets worden gevonden door een externe dienst, die we verder bespreken in hoofdstuk 4. Deze dienst krijgt de verzameling van alle tweets T en neemt hieruit een sub verzameling van tweets: de verzameling T'. Deze dienst probeert voor elke tweet in de verzameling T' een domein te suggereren. Doordat deze dienst niet gespecialiseerd is om met een korte tekst als een tweet om te gaan is deze suggestie niet nauwkeurig en werkt deze ook maar op een klein deel van de tweets. Verder stelt de dienst ook strikte gebruikerslimieten waardoor we niet alle tweets hiermee kunnen classificeren in een domein. Deze domeinen zijn echter wel bruikbaar om als basis voor de domein bepaling te gebruiken. Uit deze verzameling tweets met domeinen, de verzameling D van Tabel 3.1, kunnen we entiteiten halen die bij dit domein horen.

Dit gebeurt in de tweede stap waarin we opnieuw een externe dienst gebruiken om entiteiten te extraheren uit tweets. We gebruiken deze extractie enkel op de verzameling T'. Samen met het voorkomen van de entiteiten wordt ook de frequentie van de entiteit bij elk domein opgeslagen. Dit geeft dan als resultaat een verzameling van domeinen D samen met voor elk gegeven domein een verzameling van entiteiten, met de frequentie van deze entiteiten per domein.

Vervolgens wordt de kwaliteit van de groep entiteiten per domein verbeterd, zodat de entiteiten die niet passen in het domein niet gebruikt worden. Hiervoor passen we eerst een frequentiefiltering toe zodat we enkel entiteiten overhouden die minstens twee keer in het domein voorkwamen. Dan is er manueel gekeken naar de entiteiten die overbleven en de domeinen waarbij ze geplaatst werden. Aan de hand van de entiteiten is echter duidelijk te zien wat er echt besproken wordt

in deze domeinen. Dit betekent dat het mogelijk is om te zien hoe goed de domeinen van de externe domeindienst zijn. Op basis hiervan werd besloten om de domeinen van Tabel 3.1 aan te passen naar de uiteindelijke domeinverzameling van Tabel 3.2. Dit gebeurde door sommige domeinen van naam te veranderen, op te splitsen of, samen te voegen met andere domeinen zodanig dat elk domein een goede samenhangende verzameling entiteiten had. Een voorbeeld hiervan is 'arts_entertainment' dat gesplitst werd in 'human_interest' en 'arts' omdat hier vaak tweets in voorkwamen waarin getweet werd over moderne beroemdheden, zoals de entiteit 'George Clooney' samen met de entiteit 'american music awards'. Deze twee entiteiten liggen semantisch ver van elkaar, hetgeen de reden is voor de splitsing.

Domein bepaling uitvoeren

Met de verzameling van entiteiten voor elk domein is het mogelijk om de domeinen van de tweets te bepalen. Dit wordt gedaan door alle tweets in de verzameling T af te lopen en de tweets die een entiteit bevatten worden geclassificeerd in het domein van de entiteit. De entiteiten zijn nu gekend en de aanwezigheid hiervan is te controleren door te kijken of de string van de entiteit in de tweet staat. Het is mogelijk dat een tweet meerdere entiteiten bevat, dit is perfect normaal aangezien het mogelijk is dat een tweet tot meerdere domeinen behoort. Het voordeel van de entiteiten bij deze domein bepaling is dat ze heel eenduidig zijn en dus zelden verward worden met een ander domein, hetgeen een goede nauwkeurigheid geeft voor de domein bepaling. Het nadeel hierbij is dat dit slechts op een heel kleine groep van de tweets bruikbaar is aangezien, een tweet een gevonden entiteit moet hebben voor een domein bepaling.

Uitbreiding entiteiten

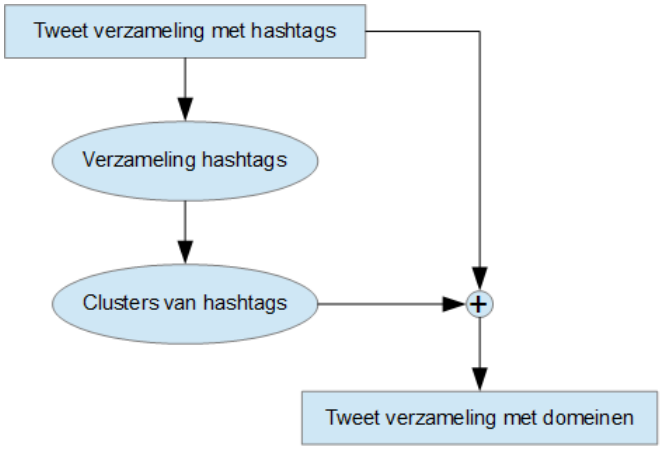
Om de entiteiten uit te breiden beschouwen we alle tweets in de trainingset K waarvan een domein bepaald is. De tokens van deze tweets kunnen mogelijk dienen als nieuwe woorden die het domein aangeven van de tweet. Om te bepalen welke tokens hiervoor het best gebruikt worden, wordt er een analyse gemaakt van hoe goed elk van de tokens als entiteit past in een domein. Voor deze analyse is gebruik gemaakt van de tokenisatie van sectie 3.2.1. Deze tokenisatie is gebruikt om alle tokens te verzamelen van elk domein en deze allemaal een waarde per domein te geven door middel van de statistieken Mutual Information (MI) en χ^2. Deze statistieken geven allebei een waarde die de afhankelijkheid van een token van het domein uitdrukken, waarbij de tokens die enkel in één domein voorkomen de hogere waardes krijgen en tokens die

in alle domeinen voorkomen lagere waardes krijgen. Op deze manier is het mogelijk om nieuwe woorden te vinden die een goede indicatie zijn voor het bepalen van het domein. Aangezien de beste tokens automatisch een hoge waarde krijgen zou het mogelijk kunnen zijn om automatisch de beste tokens te nemen als deze woorden. Het probleem echter is dat deze top niet perfect is doordat de domeinen niet perfect zijn. Regelmatig komt er een tweet voor die niet echt over het domein gaat omdat het over een entiteit praat op een onverwachte manier. Bijvoorbeeld de tweet 'Ugh! Damn it! Lee Harvey Oswald shot President John F. Kennedy. That was an epic scene from the movie. :-O :-(#KillingKennedy #epic #JFK'. Deze tweet praat over een president in de Verenigde Staten en wordt dus geclassificeerd als een tweet over 'Samenleving en Politiek'. Maar deze tweet gaat niet over 'Samenleving en Politiek', maar over een film waarin de president voorkomt. Doordat Twitter sterk aan *trending* gevoelig is kan dit soort tweet overvloedig voorkomen in een bepaald domein. Dit heeft als gevolg dat er woorden een hoge score krijgen ook al passen ze niet in het domein. Indien dit soort foute woorden dan automatisch geselecteerd wordt zou dit de fouten in het domein vergroten, met als gevolg dat een nieuwe entiteit uitbreiding extra onnauwkeurigheid zou introduceren. Daarom hebben we de tokens met maximale MI/χ^2 manueel gefilterd. Deze uitbreiding heeft in dit onderzoek maar tot 16 woorden geleid die we verder hebben gebruikt om meer tweets met domeinen te vinden. Dit omdat de domein bepaling met deze uitbreiding nog altijd op slechts een beperkt deel van de tweets werkt. Dit betekent dat elke uitbreiding op de entiteiten in iteraties moet gebeuren die altijd arbeidsintensief zijn en slechts weinig nieuwe entiteiten toevoegen. Dit is dus geen bruikbare methode om tot een complete domein bepaling te komen waarbij alle tweets een domein krijgen.

3.3.3 Hashtag Clusters

Concepten

De methode die in deze sectie beschreven wordt, heeft als doel de problemen van domein bepaling via entiteiten (sectie 3.3.2) op te lossen. Het probleem met entiteiten is tweeledig. Aan de ene kant zijn de entiteiten een manuele selectie van de automatische selectie van een externe dienst. Deze vormen dus geen goed beeld van welke onderwerpen het meeste voorkomen in de tweets. Verder kan de domein bepaling via entiteiten slechts een beperkt deel van de tweets een domein geven.



Figuur 3.3: Domein bepaling via hashtags

De methode die in deze sectie beschreven wordt, is gebaseerd op het gebruik van hashtags, geïllustreerd in Figuur 3.3. Hashtags zijn een populair hulpmiddel op Twitter om aan te duiden waar een tweet over gaat. Hashtags op Twitter bestaan uit een hashtag '#' gevolgd door woorden zonder spaties, bijvoorbeeld '#callofdutytime'. Net als entiteiten kunnen hashtags dus gezien worden als een onderwerp van een tweet. Het voordeel van hashtags tegenover entiteiten is dat hashtags door de gebruikers worden toegevoegd aan tweets hetgeen als gevolg heeft dat de verzameling hashtags van een verzameling tweets T een betere representatie is van de onderwerpen in de verzameling T dan de verzameling entiteiten van dezelfde verzameling T. Dit omdat de hashtags heel gemakkelijk te vinden zijn en het dus niet mogelijk is om een hashtag te missen in een tweet. Daarentegen is het wel mogelijk om een entiteit over het hoofd te zien. Om nu het domein via hashtags te vinden, wordt gebruik gemaakt van het onderzoek van Antenucci et al. [11]. Net als in dat onderzoek wordt een domein hier gevormd door meerdere hashtags die semantisch samen passen. Om dus uit de hashtags domeinen te vormen, worden de hashtags geclusterd. Voor deze sectie wordt een domein dus beschouwd als een cluster van hashtags. Een voorbeeld hiervan is de cluster '... #healthyliving #instafood #nomnom #homemade #foodiechats #yummy #yelp #cooking ...' die gaat over onderwerpen als '#instafood' en '#cooking'. Een tweet hoort dan bij een domein als de tweet één van de hashtags van de cluster bevat. Dit betekent dus dat het nu mogelijk is om de verzameling domeinen automatisch op te stellen door middel van hashtag clusters. Dit geeft een goede weergave van de domeinen die dominant zijn in de verzameling tweets T. Een belangrijke opmerking echter is dat niet alle tweets een hashtag hebben. Deze maken dus ook geen volledige domein bepaling mogelijk.

Clustering

Het doel is nu om hashtags te clusteren in domeinen zodanig dat de domeinen bruikbaar zijn voor de sentiment classificatie. Het idee hierachter is dat elke hashtag een onderwerp voorstelt en dat de hashtags semantisch samen horen in een domein. De clustering van de hashtags moet ervoor zorgen dat elke cluster een domein voorstelt. Als input voor de clustering wordt de verzameling K van training tweets gebruikt samen met de verzameling T_H. De eerste stap is om uit deze verzameling van tweets K alle hashtags H te halen samen met de frequentie van elke hashtag. Om de clustering computationeel mogelijk te maken, wordt er gebruik gemaakt van een frequentiefiltering om de 2000 meest voorkomende hashtags H' hieruit te halen. Enkel deze hashtags H' worden gebruikt bij de clustering. Elk van deze hashtags wordt nu gebruikt in het clusteringalgoritme, dit werkt met Spectral Clustering [13] gebaseerd op het onderzoek van Antenucci et al. [11]. Dit clusteringalgoritme werkt met een in te stellen afstandssmetriek om de afstand tussen twee hashtags te kunnen bepalen tijdens de clustering. Er zijn twee verschillende afstandssmetrieken getest voor dit onderzoek. De eerste gebruikt een afstandssmetriek op basis van de co-occurence van de hashtags. De tweede maakt gebruik van de cosinusafstand tussen de tweets van de hashtags. Voor de informatie van deze afstandssmetrieken wordt een verzameling tweets T_H gebruikt. Deze verzameling T_H wordt uit de verzameling T gehaald, de verzameling T is de verzameling van alle verzamelde tweets van ons onderzoek (sectie 4.2). Deze verzameling T_H bestaat uit alle tweets van T die niet tot K, V of U behoren en één van de hashtags van H' bevat. K, V en U zijn de verzamelingen die gebruikt worden tijdens de sentiment classificatie, T_H wordt gescheiden gehouden van deze verzameling zodanig dat de informatie van de afstandssmetrieken niet beïnvloed is door de testverzamelingen. De methodologie van de afstandssmetrieken wordt verder uitgelegd in de volgende sectie.

Afstandssmetrieken

Om tot de clustering over te gaan moeten eerst de twee afstandssmetrieken gedefinieerd worden die getest zijn voor de clustering. Deze afstandssmetrieken maken gebruik van een verschillende voorbereiding en concept om de hashtags te clusteren. Beide clusteringtechnieken maken gebruik van de verzameling hashtags H' en de verzameling tweets T_H.

De co-occurence afstand is gebaseerd op het samen voorkomen van hashtags in een tweet. Het idee hierachter is dat hashtags die vaak samen voorkomen in een tweet ook semantisch gelijk-

aardig zijn. Voor de berekening van de co-occurence afstand wordt $n_{h1,h2}$ gedefinieerd als de frequentie waarmee de hashtags $h1$ en $h2$ samen voorkomen in tweets van T_H.

$$\text{co-occurence afstand}_{h1,h2} = 1 - \left(\frac{n_{h1,h2}}{\sum_{i=1}^{|H'|} n_{h1,hi}} + \frac{n_{h1,h2}}{\sum_{i=1}^{|H'|} n_{h2,hi}}\right) \times \frac{1}{2}$$

Om deze co-occurence afstand te gebruiken in de clustering is er ook nog een aparte filtering voor de clustering waarbij de hashtags met een te lage co-occurence eruit gefilterd worden.

De cosinusafstand tussen hashtags is gebaseerd op de cosinusgelijkaardigheid in een *bag-of-words* model. Hiervoor wordt voor elke hashtag een feature vector samengesteld. Deze vector gebruikt als dimensies alle tokens in de verzameling van tweets T_H, met als waarde voor een token in de vector hun tf-idf (*term frequency-inverted document frequency*) waarde. De $\text{tfidf}_{h,w}$ staat hier voor de waarde die wordt toegekend aan de token w die voorkomt als feature bij de hashtag h. De clustering is getest met meerdere definities voor tf-idf, waaruit is gebleken dat deze tf-idf definitie de beste clustering gaf voor de domein bepaling. De $\text{frequentie}_{h,w}$ is dan de frequentie van de token w in de tweets van verzameling T_H die hashtag h bevatten. Voor de idf is de $\text{hashtagfrequentie}_w$ het aantal hashtags waarbij de token w als feature voorkomt in de verzameling T_H.

$$\text{tfidf}_{h,w} = (1 + \log(\text{frequentie}_{h,w})) \times \frac{|H'|}{\text{hashtagfrequentie}_w}$$

Om nu voor de clustering de afstand tussen twee hashtags te kennen wordt de cosinusgelijkaardigheid berekend tussen de twee feature vectoren van de hashtags. Deze gelijkaardigheid wordt dan omgekeerd om de afstand te berekenen. Voor deze berekening is W de verzameling van alle tokens en is $h_{i,j}$ de waarde van de token j in de vector die bij hashtag h_i hoort.

$$\text{cosinusafstand}_{h_1,h_2} = 1 - \frac{\sum_{j=1}^{|W|} h_{1,j} \cdot h_{2,j}}{\sqrt{\sum_{j=1}^{|W|} (h_{1,j})^2} \cdot \sqrt{\sum_{j=1}^{|W|} (h_{2,j})^2}}$$

Domein bepaling uitvoeren

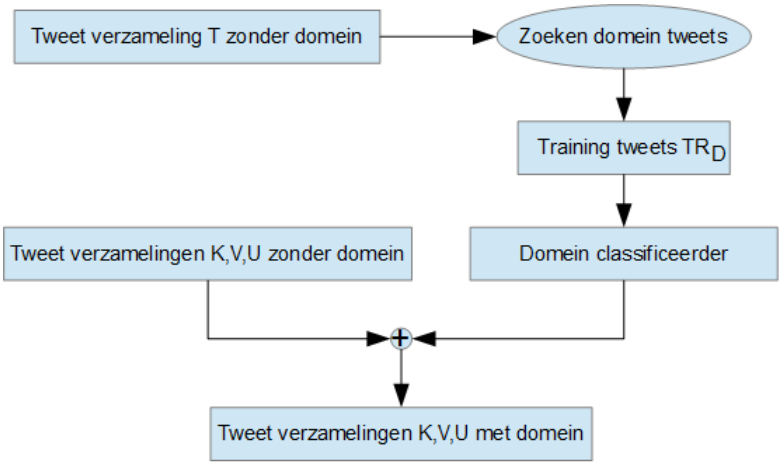
Nu de hashtag clusters beschikbaar zijn is het mogelijk om het domein van de tweets te bepalen. Dit wordt gedaan door de tweets in K, V en U af te lopen en te kijken naar de hashtags die aanwezig zijn in de tweets. Zodra een tweet een hashtag heeft, wordt de hashtag cluster opgezocht. Deze hashtag cluster is dan het domein van de tweet. In tegenstelling tot de vorige

methode is hier geen manuele controle van de domeinen en entiteiten nodig. Dit zorgt ervoor dat de hashtag clusteringmethode volledig automatisch is en dus heel goed schaalt naar het gebruik van meer domeinen. Verder zijn de domeinen nu aangepast aan de verzameling K. Dit zorgt ervoor dat de domeinverzameling zo goed mogelijk past bij de tweets van V en U waarvan men het sentiment wil bepalen. Dit aangezien de hashtags sterk veranderen in de tijd, doordat de verzameling K dicht in tijd bij V en U ligt, zorgt dit voor de beste domeinverzameling zonder kennis te nemen van de validatie- en testset. De hashtag clusteringmethode heeft echter ook het nadeel dat dit niet op elke tweet werkt. Niet alle tweets hebben immers een hashtag, waardoor niet elke tweet bij een cluster geplaatst kan worden. Om dit probleem op te lossen moet overgegaan worden naar de domein classificatie van sectie 3.3.4.

3.3.4 Domein Classificatie

Concepten

De methode die in deze sectie beschreven wordt, heeft als doel om de problemen van de domein bepaling van de vorige secties op te lossen. Het probleem met zowel de hashtag clusters als de entiteiten is dat slechts een beperkt deel van de tweets hiermee een domein kan krijgen. Deze methode probeert de ideale oplossing te vinden waarin alle tweets een domein krijgen dat zo goed mogelijk past. Om dit te bereiken wordt gebruik gemaakt van het onderzoek van [11]. In dit onderzoek wordt een classificeerder gebruikt die tweets zonder domein classificeert in een domein. Deze classificeerder werkt op basis van een verzameling tweets die wel een domein hebben om hieruit te leren bij welk domein een tweet het beste past. Dit betekent dus dat deze domein classificatie gebaseerd wordt op één van de twee beperkte domein bepalingen. We hebben ervoor gekozen om te werken met de hashtag clusters. Dit omdat de hashtags een beter beeld kunnen geven van de onderwerpen in de tweets, aangezien de hashtags in tweets een annotatie zijn van het onderwerp door de auteur van de tweet. Dit betekent dat de hashtags een betrouwbaar beeld geven van waar de tweets over gaan. Verder maken de hashtags het mogelijk om automatisch een verzameling domeinen samen te stellen. Dit maakt de hashtags makkelijker te gebruiken dan de entiteiten methode die altijd een manuele filtering vereist. Met deze methode is het dus mogelijk om elke tweet een domein te geven, dit lost de nadelen van de vorige methodes op. Dit concept wordt verduidelijkt in Figuur 3.4.



Figuur 3.4: Domein bepaling via classificeerder

Domein classificatie uitvoeren

Zoals gezegd maakt deze methode gebruik van een classificeerder die een verzameling tweets met domein nodig heeft. Als input voor deze methode wordt de gegeven verzameling TR_D gebruikt. Deze verzameling tweets is gesplitst van de verzamelingen K, V en U. Dit is noodzakelijk omdat de verzameling TR_D gebruikt zal worden om de domein classificeerder te trainen die de domeinen van de tweets in K, V en U moet bepalen. Verder bevatten alle tweets in TR_D een domein door middel van de methode van sectie 3.3.3, tweets die geen domein krijgen van deze methode zijn uit de verzameling gefilterd. Dit zorgt ervoor dat de verzameling TR_D gebruikt kan worden voor de training van de domein classificeerder.

De domein classificatie krijgt nu als input de verzamelingen K, V, U en TR_D van tweets. Om nu de tweets in K, V en U te classificeren wordt, in navolging van [11], een classificeerder getraind door middel van de tweets in TR_D. Op deze manier moet de classificeerder in staat zijn om tweets zonder hashtag te classificeren in een domein. De tweets waarop de classificeerder traint worden hiervoor omgezet naar binaire feature vectoren op basis van het *bag-of-words* model. Hier is voor gekozen omdat uit het onderzoek van Antenucci et al. [11] blijkt dat dit beter werkt dan andere feature vectoren die ze getest hebben. Bij deze feature voorstellingen zijn de feature dimensies alle tokens in de verzameling met de waarde van de vector: een binaire waarde afhankelijk van de aanwezigheid van een token in de tweet. Hierbij worden de hashtag tokens verwijderd om ervoor te zorgen dat de classificeerder de hashtags aanleert zonder de aanwezigheid van de hashtags. De classificeerder zelf maakt gebruik van het Naive Bayes classificatiealgoritme

met het Multinomiale model. Deze classificeerder wordt vervolgens gebruikt op alle tweets die nog niet in een domein zitten. Op deze manier krijgen alle tweets een domein.