university of groningen / faculty of arts

Online opinie over immigranten

Geografische sentimentanalyse van Nederlandse tweets over immigranten

Moniek Nieuwenhuis

Bachelor thesis
Informatiekunde
Moniek Nieuwenhuis
s2570785
20 juni 2017

SAMENVATTING

Immigratie was een groot onderwerp tijdens de verkiezingen in Nederland afgelopen maart. Om een beter beeld te krijgen van de voor- en tegenstanders van immigratie hebben we sentimentanalyse toegepast op Twitter met tweets over immigratie. Deze tweets bevatten een locatie waardoor voor elke gemeente kan worden bepaald of deze gemeente positiever tegenover immigranten staat of negatiever.

De data voor dit onderzoek is verzameld uit het Twittercorpus van de Rijksuniversiteit van Groningen. Voor dit onderzoek zijn alle tweets uit 2016 gebruikt die gaan over het onderwerp immigratie. Van deze tweets werden alleen de tweets gebruikt die een meegegeven gemeente bevatten. Van de 390 gemeenten in Nederland zijn er uiteindelijk 361 gemeenten in de data te vinden.

Voor de sentimentanalyse is een classifier gemaakt. Deze is getraind en getest met 1000 met de hand geannoteerde tweets. Alle data voor dit onderzoek is getokeniseerd. Voor dit onderzoek zijn er twee verschillende classifiers getest, naive Bayes en een linear-kernel Support Vector Machine(SVM). Beide classifiers worden geëvalueerd aan de hand van de f1-score die is afgeleid van de precisie en recall.

De beide classifiers zijn getest in combinatie met zowel letter n-grammen als woord n-grammen als feature. Tijdens de testfase bleek dat de SVM classifier de beste resultaten behaalt met als feature woord n-grammen in unigrammen, bigrammenen, trigrammen en daarnaast letter n-grammen in trigrammen, 4-grammen, 5-grammen en 6-grammen. Op de testset behaalde deze classifier uiteindelijk een f1-score van 0.53 en een accuratesse van 0.53.

Uit de resultaten van de gemeenten bleek dat 140 gemeenten negatief staan tegenover immigranten en 145 gemeenten positief. In de bijlage in Tabel 12 is het totale overzicht te zien per gemeente. Over het geheel van tweets waren er meer negatieve tweets dan positieve tweets. Ook is bepaald welke gemeenten het meest positief zijn en het meest negatief. De top-3 positieve gemeenten zijn: Bernheze, Papendrecht en Raalte. De top-3 negatieve gemeenten zijn Lingewaal, Noordoostpolder en Het Bildt.

INHOUDSOPGAVE

VOORWOORD

Toen de Nederlandse verkiezingen werden gehouden op 15 maart 2017 wist ik dat ik daar iets mee wilde gaan doen. Aangezien de verkiezingen al geweest waren voor het schrijven van mijn onderzoeksvoorstel, heb ik besloten om een groot onderwerp tijdens de debatten te onderzoeken.

Na maanden hard werken is deze scriptie nu afgerond. Ik ben zeer tevreden met het resultaat en hoe deze scriptie is gegaan. Deze scriptie is het slotstuk van alle Informatiekunde vakken die ik heb mogen volgen, waarbij ik veel verschillende dingen heb geleerd en deze technieken ook heb mogen toepassen in deze scriptie.

Ik wil graag een paar mensen bedanken. Allereerst bedank ik Malvina Nissim, mijn scriptiebegeleider. Toen ik gestrand was geraakt met de zoektocht naar een goed onderwerp hielp ze me op weg. Daarnaast wil ik graag Johan Bos bedanken, die de tijd heeft genomen om te kijken naar het Nederlands in deze scriptie en daarover zeer bruikbare feedback heeft gegeven.

1 INTRODUCTIE

Op 15 maart 2017 werden er verkiezingen in Nederland gehouden. Tijdens deze verkiezingen was immigratie een groot onderwerp. Dit komt mede door de grote vluchtelingenstroom die de afgelopen jaren naar Europa is getrokken. Hierdoor kende Nederland in 2015 43.093 nieuwe asielaanvragen.[1] Deze vluchtelingen worden in Nederland opgevangen in asielzoekerscentra (azc) gelegen op verschillende locaties in het land. Één van deze locaties is Ter Apel, hier is plek voor 2.000 vluchtelingen.[2] Voor dit onderzoek zijn wij benieuwd of inwoners van bijvoorbeeld Ter Apel juist negatiever of positiever staan tegenover de komst van vluchtelingen.

Daarom willen we sentimentanalyse toepassen op Twitter door middel van tweets met een locatie die gaan over immigranten, omdat we willen weten of twittergebruikers positief of negatief tegenover immigranten staan, teneinde te kunnen bepalen of er een meerderheid voor of tegen de komst van immigranten naar Nederland is en om te kunnen bepalen in welke gemeenten in Nederland mensen positiever of negatiever tegenover de komst van immigranten zijn. In deze scriptie staat de volgende vraag centraal:

In welke gemeenten in Nederland staan twittergebruikers positiever tegenover de komst van immigranten en in welke gemeenten negatiever tegenover de komst van immigranten?

Voor de beantwoording van deze vraag gaan we eerst kijken naar relevante literatuur over wat er is gedaan in onderzoek op Twitter (Hoofdstuk 2). Vervolgens zullen we data van Twitter verzamelen (Hoofdstuk 3). Om alle tweets te analyseren maken we een classificatiesysteem, deze methode bespreken we in hoofdstuk 4. De resultaten van het onderzoek worden besproken in Hoofstuk 5 en tot slot in Hoofdstuk 6 de conclusie.

[1] https://www.vluchtelingenwerk.nl/sites/public/u895/Vluchtelingeningetallen2016.pdf
[2] https://www.coa.nl/nl/zoek-locatie/ter-apel

2 RELEVANTE LITERATUUR

Voor dit onderzoek kijken we naar de methodes die zijn uitgevoerd door Pak and Paroubek (2010) en Sobhani et al. (2016). Beide onderzoeken richten zich op Twitter, waar wij voor dit onderzoek ook gebruik van maken. Daarnaast kijken we ook naar een onderzoek van Sang and Bos (2012), dit onderzoek richt zich speciaal op Nederlandse tweets.

Pak en Paroubek hebben onderzoek gedaan naar opinion mining en sentimentanalyse op Twitter (Pak and Paroubek, 2010). Hierbij werd geen gebruik gemaakt van één bepaald onderwerp, maar keken ze naar het algemene sentiment op Twitter. Voor dit onderzoek wordt gebruikt gemaakt van een naive Bayes classifier. Deze classifier behaalde voor dit onderzoek betere resultaten dan SVM, welke ook is getest. Hierdoor willen wij ook een naive Bayes classifier testen op de tweets die wij hebben verzameld.

Vorig jaar was er een shared task[1] over stance detection (Mohammad et al., 2016). Tijdens deze shared task is er gebruik gemaakt van tweets over vijf verschillende targets: 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton' en 'Legalization of Abortion'. De deelnemers aan de shared task moesten een classifier bouwen voor deze vijf targets met de classificatie of de tweet positief, negatief of neutraal was tegenover de target. Voor dit onderzoek maken we geen gebruik van meerdere targets, maar hebben we één onderwerp, namelijk immigratie. Opvallend aan deze shared task is dat het algoritme gebruikt voor de baseline, SVM met n-grammen, betere resultaten behaalde dan het beste team met hun eigen algoritme. Deze classifier wordt ook beter bescheven in Sobhani et al. (2016). Voor deze classifier gebruikten ze scikit-learn en en het linear suport vector machines algoritme. Daarnaast werd er gebruik gemaakt van n-grammen als feature, die naast woord n-grammen ook letter n-grammen maakt. Deze feature nemen we mee in ons onderzoek.

Dit onderzoek richt zich op Nederlandse tweets. Eerder onderzoek met Nederlandse tweets op het gebied van opinion mining en sentimentanalyse is gedaan in 2011 door Sang and Bos (2012). In 2011 hebben zij geprobeerd de Nederlandse Provinciale Statenverkiezingen te voorspellen aan de hand van Twitter. Hiervoor gebruikten ze sentimentanalyse om het sentimentgewicht te bepalen voor elke partij. Voor dit onderzoek berekenen we ook het sentimentgewicht om de gemeenten met elkaar te kunnen vergelijken.

[1] http://alt.qcri.org/semeval2016/task6/

3 DATA AND MATERIAAL

Dit hoofdstuk zal ingaan op de verzameling van tweets. Er zal worden uitgelegd hoe deze tweets zijn verzameld en op welke manier de gemeenten zijn bepaald. Daarnaast zal dit hoofdstuk ingaan op de annotatie van tweets en hoe de tweets in stukjes zijn verdeeld.

3.1 DATA COLLECTIE

De twitterdata die gebruikt is voor dit onderzoek is verzameld door de Rijksuniversiteit van Groningen. Deze tweets zijn gefilterd op de Nederlandse taal, waardoor de meeste tweets ook in Nederland zijn getwitterd. Voor dit onderzoek gebruiken we alle tweets uit 2016. Deze tweets hebben wij met een shellscript van de server gehaald. Daarbij zijn alleen tweets opgeslagen die een van de volgende relevante termen bevatten:

- 'immigrant'
- 'migrant'
- 'vluchteling'
- 'migratie'
- 'immigratie'

Het selecteren van deze tweets hebben we gedaan met het linux commando 'grep' met als argument 'i', waardoor er geen onderscheid wordt gemaakt tussen hoofd- en kleine letters. Argument 'w' is niet meegegeven, waardoor er ook wordt gezocht naar woorddelen en niet alleen naar woorden. Daardoor is het ook mogelijk dat de meervoudsvorm van de woorden wordt meegenomen en ook samenvoegingen zoals 'vluchtenlingencrisis'. Met de Tweet2Tab tool in het shellscript zijn de tweets opgehaald van de server. In deze dataset zitten hierna 1.168.487 tweets. Elke tweet gescheiden door een tab heeft de volgende structuur:

- ID
- gebruikersnaam
- tweet
- woorden
- hastags
- datum
- gemeente/plaats

3.1.1 gemeentenamen filteren

Niet elke tweet in de opgehaalde twitterdata bevat een gemeentenaam. Ondanks dat 'place' is meegegeven met Tweet2Tab worden tweets zonder gemeentenaam leeg gelaten. Voor dit onderzoek willen we alleen de tweets gebruiken die wel een

gemeentenaam hebben meegekregen. Hiervoor is een pythonscript geschreven om alleen de tweets te bewaren die wel een gemeentenaam bevatten. Als je de tweets split op de tab dan hebben de tweets zonder gemeentenaam zes elementen in de lijst en tweets met een gemeentenaam zeven. De overgebleven 19.791 tweets met gemeentenaam worden gebruikt voor het onderzoek.

3.1.2 Gemeente groepering

De 19.791 tweets in de dataset worden gegroepeerd per gemeente. Het databestand is door een pythonprogramma opgesplitst dat voor elke gemeente een apart bestand met tweets heeft gemaakt. Niet alle tweets hebben een gemeente meegekregen in Nederland. De reden hiervoor is dat de Rijksuniversiteit Groningen alle tweets die in het Nederlands geschreven zijn heeft verzameld. Daardoor zijn er ook veel tweets uit België of landen waar Nederlanders misschien op vakantie zijn of wonen.

Tweet2Tab geeft bij 'place' vaak twee of drie elementen mee. Er zit veel variatie in de structuur van de elementen. Het eerste is meestal de gemeente, het tweede element is vaak de provincie, maar soms ook het land en als het derde element aanwezig is, is dit het land. Nu komt het ook voor dat het land niet wordt aangeduid als 'Nederland', maar ook 'The Netherlands', 'Niederlande' en 'Pays-Bas' kwamen voor in combinatie met Amsterdam. Eerst wordt er in het python programma gekeken of het land Nederland is voordat het de tweets opslaat in een tekstbestand.

Aangezien dit onderzoek wordt uitgevoerd met alleen gemeenten hebben we een check toegevoegd of de gemeentenaam wel een gemeente is. Hiervoor hebben we gebruik gemaakt van het Excel bestand van het CBS met alle gemeenten in 2016 CBS (CBS).

Nederland telde in 2016 390 gemeenten[1]. In de dataset zitten 361 gemeenten met een totaal aantal van 13.201 tweets. In de bijlage is Tabel 11 opgenomen met het aantal tweets per gemeente. In Figuur 1 is een heatmap van Nederland te zien met de verdeling van tweets over Nederland.

3.2 ANNOTATIE

Voor dit onderzoek zijn er handmatig 1.000 tweets geannoteerd. Deze zijn met een random seed python programma verzameld uit de totale dataset, hierbij is nog geen rekening gehouden met gemeentenamen. Door dezelfde waarde voor random.seed te gebruiken blijft het onderzoek reproduceerbaar.

De tweets zijn geannoteerd door middel van scores. In Tabel 1 zijn de verschillende klassen te zien.

Tabel 1: Manier van de annotatie van de data.

klasse	Positieve score	Negatieve score
Positief over immigranten	1	0
Negatief over immigranten	0	1
Neutraal	0	0
Bevat zowel positieve als negatieve elementen	1	1

In Tabel 2 staan voorbeelden van tweets uit elke klasse.

In Tabel 3 is te zien hoeveel tweets van elke klasse de dataset bevat. Deze geannoteerde dataset wordt gebruikt voor het trainen en testen van de classifier.

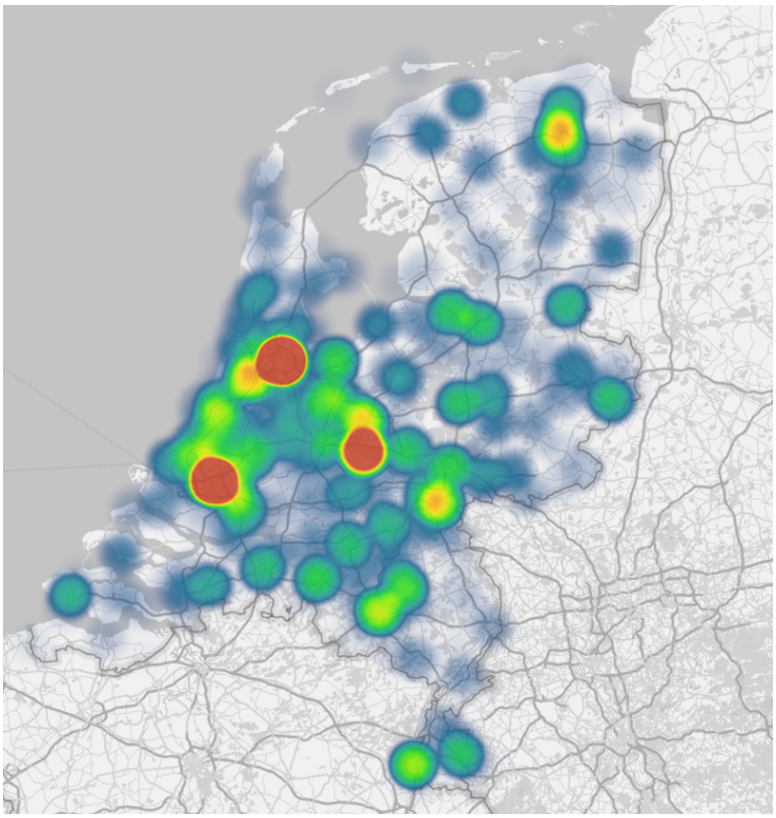
[1] cbs.nl

Tabel 2: Voorbeelden van tweets per klasse.

Gebruiker	Tweet	Gemeente
	Positieve tweet	
BenjaminDThomas	Interessant discussie over het opnamesystem van vluchtelingen in Europa. Ik find alle landen in de EU mouten vluchtelingen opnemen. cdjanj	's-Hertogenbosch, Nederland, Niederlande
pvaznunes	'Vluchtelingenprobleem'?? De oorlog waar ze voor vluchten is het probleem toch?	Amsterdam, North Holland, Nederland
	Negatieve tweet	
sandraroeters	eliasvanhees @BarthCarlo @omroepwest @RSBaldewsingh fuck die zooi vluchtelingen in ons land!NL is van ons take itâ	Bedum, Nederland, Nederland
helderalsglas	MP Rutte sluit onze dakloze bevolkingsgroep uit tegen riante regelingen voor vluchtelingen. Stuur die naar het Leger des Heils. Ruil ze uit.	Amsterdam, The Netherlands, Nederland
StichtingStaan	Hoppa! Bus met meer als 80 vluchtelingen verongelukt. PVV GRENZENDICHT	Haren, Groningen, Nederland
	Neutrale tweet	
eenvandaag	Bijna zes op de tien (58%) vluchtelingen uit ons onderzoek geeft aan homoseksualiteit geen probleem te vinden	Hilversum, Nederland, Nederland
ErnstLissauer	VVD heeft een 'plan B' - instroom vluchtelingen. PvdA weet van niets.	Den Haag, Zuid-Holland, Nederland
	Positief en negatief	
PWEM	Ik ben voor 't opnemen van vluchtelingen, niet migranten of economische gelukszoekers met 'eisen' :(	Rotterdam, Nederland, Nederland

Tabel 3: Geannoteerde dataset.

Positief	Negatief	Neutraal	Tegenstrijdig	Totaal aantal tweets
244	274	470	12	1000



Figuur 1: Heatmap van Nederland met het totaal aantal tweets in het onderzoek.

3.3 DATA VERWERKING

Om de classifier zo goed mogelijk te kunnen trainen hebben we de data getokeniseerd en gefilterd. Eerst zijn de hoofdletters uit de tweet gehaald en veranderd in kleine letters, vervolgens is ook alle punctuatie uit de tekst verwijderd met het string.punctuation commando van python.

Hierna is de tekst getokeniseerd met de tweet tokenizer van de NLTK python libary Bird et al. (2009). Daarnaast zijn er nog een aantal dingen uit de tweet verwijderd, zoals de gebruiksersnamen, het hekje voor een gebruikte hashtag en linkjes. Vervolgens zijn ook stopwoorden verwijderd uit tweets, aangezien er dan alleen gekeken wordt naar belangrijke woorden in de tweets. Deze stopwoorden zijn verwijderd met behulp van het NLTK stopwoorden corpus en daarbij zijn de Nederlandse stopwoorden gebruikt.

4 METHODE

In dit hoofdstuk zal worden besproken hoe de classifier, die gebruikt is voor de sentimentanalyse is gemaakt. Daarnaast zal er aandacht besteed worden aan de evaluatie van deze classifier en de evaluatie van de onderzoeksresultaten van de gemeenten. Voor alle codering van dit onderzoek wordt verwezen naar: https://github.com/Moniekleonie/Ba-scriptie

4.1 FEATURES SELECTIE VOOR SENTIMENT CLASSIFIER

Voor de selectie van features voor de sentimentanalyse classifier kijken we vooral naar de methode die ze hebben gebruikt tijdens de shared task Mohammad et al. (2016). Daarbij werd gebruik gemaakt van n-grammen. Deze n-grammen werden gemaakt van de woorden, maar ook van letters.

Om de tekst van de tweets te verdelen in n-grammen hebben we gebruik gemaakt van scikit-learn [1]. Via de CountVectorizer is het mogelijk om n-grammen te maken. We hebben zowel woord n-grammen gebruikt als n-grammen die de woorden zelf opsplitst in stukjes. Dit hebben we gedaan met de optie 'char_wb', waarmee wordt aangegeven dat er niet buiten het woord n-grammen worden gemaakt. Voor dit onderzoek hebben we voor de woord n-grammen gebruik gemaakt van unigrammen, bigrammen en trigrammen en voor de letter n-grammen hebben we gebruik gemaakt van bigrammen, trigrammen, vier-grammen, vijf-grammen en zes-grammen.

4.2 SENTIMENT CLASSIFICATIE

Voor de classifier hebben we gebruik gemaakt van het script dat de scriptiebegeleider aanbood. Deze classifier is gemaakt met scikit-learn. In de flowchart die scikit-learn op de website heeft staan [2] komen zowel een linear-kernel Support Vector Machine als naive Bayes naar boven als beste classifier voor dit onderzoek. In de relevante literatuur waren beide methodes te vinden. Pak and Paroubek (2010) gebruikten een aanpak met naive Bayes en Sobhani et al. (2016) gebruikten een aanpak met een lineare SVM. Daarom zijn beide classifiers geïmplementeerd om te testen en evaluaren welke classifier de beste resultaten voor dit onderzoek behaalt.

4.3 EVALUATIE

4.3.1 Evaluatie van de sentiment classifier

Om de classifier te testen hebben we van de geannoteerde 1000 tweets de laatste 150 tweets apart gezet.

Tijdens het bouwen en trainen van de classifier hebben we gebruik gemaakt van een 5-fold cross-validation om tussendoor te testen. Dit moest voorkomen dat we de

[1] http://scikit-learn.org/stable/modules/feature_extraction.html
[2] http://scikit-learn.org/stable/tutorial/machine_learning_map/

classifier te specifiek maken op de data. Na elke fold worden de resultaten geprint in een classificatie-rapport van scikit-learn.

De resultaten worden uitgedrukt in accuratesse, precisie, recall en het harmonieuze gemiddelde van recall en precisie, de f1-score. Deze worden per categorie berekend en uitgeprint. Van deze scores per categorie wordt een gewogen gemiddelde genomen voor je totale score. Bij de 5-fold cross-validation in de trainingsfase wordt na de folds een gemiddelde van de 5 folds berekend over alle gewogen gemiddelden van de folds.

Om de classifier goed te evalueren hebben we een baseline berekend. De baseline is vastgesteld op een fictieve 'domme classifier' en deze classificeert alle tweets in de grootste klasse. In het geval van dit onderzoek is dit de neutrale klasse. De baseline is ook berekend met een gewogen gemiddelde over alle klassen. Deze staat genoemd in Tabel 4.

4.3.2 Evaluatie van de gemeente resultaten

Om de uitslag van de gemeenten te bepalen wordt elk gemeentetekstdocument ingeladen en elke tweet geclassificeerd. Dit gebeurt met de classifier die de beste resultaten heeft behaald. Van elke gemeente wordt opgeslagen hoeveel tweets er in welke klasse zijn geplaatst. Voor het beantwoorden van de onderzoeksvraag zijn we alleen geïnteresseerd in de uitslag van de positieve tweets en negatieve tweets, aangezien we willen bepalen welke gemeenten negatiever zijn en welke positiever. Om dit te visualiseren gebruiken we de 3D-map tool van Excel(2016). Deze visualiseert per gemeente de categorie met de meeste tweets. Hij kleurt rood bij meer negatieve tweets dan positieve en kleurt groen bij meer positieve dan negatieve.

Aangezien de gemeenten niet evenveel tweets hebben en we ze toch willen vergelijken normaliseren we de scores. Dit doen we om ten einde te kunnen bepalen welke gemeenten het meest negatief zijn en welke gemeenten het meest positief.

De formule die voor de normalisatie is gebruikt:

$$\frac{p+1}{p+n+2}$$

Hierbij staat 'p' voor het aantal positieve tweets en 'n' voor het aantal negatieve tweets. De +1 en +2 zijn toegevoegd, zodat er niet gedeeld hoeft te worden door 0 en de verhouding gelijk blijft.

5 RESULTATEN EN DISCUSSIE

In het eerste deel van dit hoofdstuk zullen alle resultaten van het onderzoek worden gepresenteerd. Hierbij zullen eerst de resultaten van de sentiment classifier worden besproken en vervolgens de resultaten van het onderzoek. Hierna zullen de resultaten bediscussieerd worden.

5.1 RESULTATEN VAN DE SENTIMENT CLASSIFIER

In Tabel 4 zijn de resultaten van de sentiment classifier tijdens trainingsfase te zien. We hebben twee verschillende classifiers getest, een linear-kernel Support Vector Machine(SVM) en naive Bayes. Bij beide classifiers hebben we geprobeerd de best mogelijke combinatie te vinden van features. De beste naive Bayes classifier is degene met alleen unigrammen, deze scoort beter dan de gestelde baseline. De beste SVM classifier is de classifier die gebruik maakt van woord n-grammen in unigrammen, bigrammen en trigrammen. Daarnaast maakt deze gebruik van letter n-grammen in trigrammen, 4-grammen, 5-grammen en 6-grammen. Dit verschilt met de methode van Sobhani et al. (2016), want daar gebruikten ze ook bigrammen en hadden ze geen 6-grammen toegevoegd. Bij dit onderzoek gaf het weglaten van bigrammen betere resultaten. Daarnaast verhoogde de score met het toevoegen van 6-grammen. Voor het onderzoek gaan we gebruik maken van de sentiment classifier met de beste resultaten en dat is de SVM classifier met de meeste woord n-grammen en letter n-grammen.

Tabel 4: De scores in deze Tabel zijn gewogen gemiddeldes van elke klasse waarvan een gemiddelde is gemaakt over alle 5 de folds.

Systeem	**Precisie**	**Recall**	**F1-score**
Baseline	0.22	0.47	0.31
SVM			
woord n-grammen (n=1)	0.50	0.50	0.49
woord n-grammen (n=1,2)	0.51	0.50	0.49
woord n-grammen (n=1,2,3)	0.51	0.50	0.49
woord n-grammen (n=1,2,3) + letter n-grammen (n=2)	0.50	0.50	0.49
woord n-grammen (n=1,2,3) + letter n-grammen (n=3)	0.51	0.52	0.50
woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4)	0.51	0.52	0.50
woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4,5)	0.52	0.52	0.50
woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4,5,6)	**0.53**	**0.53**	**0.51**
naive Bayes			
woord n-grammen (n=1)	**0.48**	**0.48**	**0.38**
woord n-grammen (n=1,2)	0.47	0.47	0.34
woord n-grammen (n=1,2,3)	0.46	0.47	0.33
woord n-grammen (n=1,2,3) + letter n-grammen (n=2)	0.21	0.46	0.29
woord n-grammen (n=1,2,3) + letter n-grammen (n=3)	0.21	0.46	0.29
woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4)	0.30	0.46	0.29
woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4,5)	0.35	0.46	0.30
woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4,5,6)	0.31	0.46	0.30

In Tabel 5 zijn de resultaten te zien over de folds van de gekozen SVM classifier. De slechts presterende fold heeft een f1-score van 0.42 en de best presterende fold 0.58. Gemiddeld over alle folds behaalt deze classifier een accuratesse van 0.53.

Tabel 5: Resultaten van de 5-fold cross-validation van de SVM classifier met woord n-grammen (n=1,2,3) + letter n-grammen (n=3,4,5,6). De scores zijn gewogen gemiddeldes van elke klasse.

	minimum	**maximum**	**gemiddelde**	**standaarddeviatie**
Precisie	0.46	0.58	0.5260	0.04669
Recall	0.45	0.58	0.5260	0.05030
F1-score	0.42	0.58	0.5140	0.05983

Tabel 6: Classificatie resultaten van de testset.

	Precisie	**Recall**	**F1-score**	**aantal**
Neutraal	0.64	0.63	0.64	82
Negatief	0.51	0.50	0.51	42
Positief	0.25	0.29	0.27	24
Tegenstrijdig	0.0	0.0	0.0	2
Avg/totaal	0.53	0.53	0.53	150

Tabel 7: Confusion matrix van de testset.

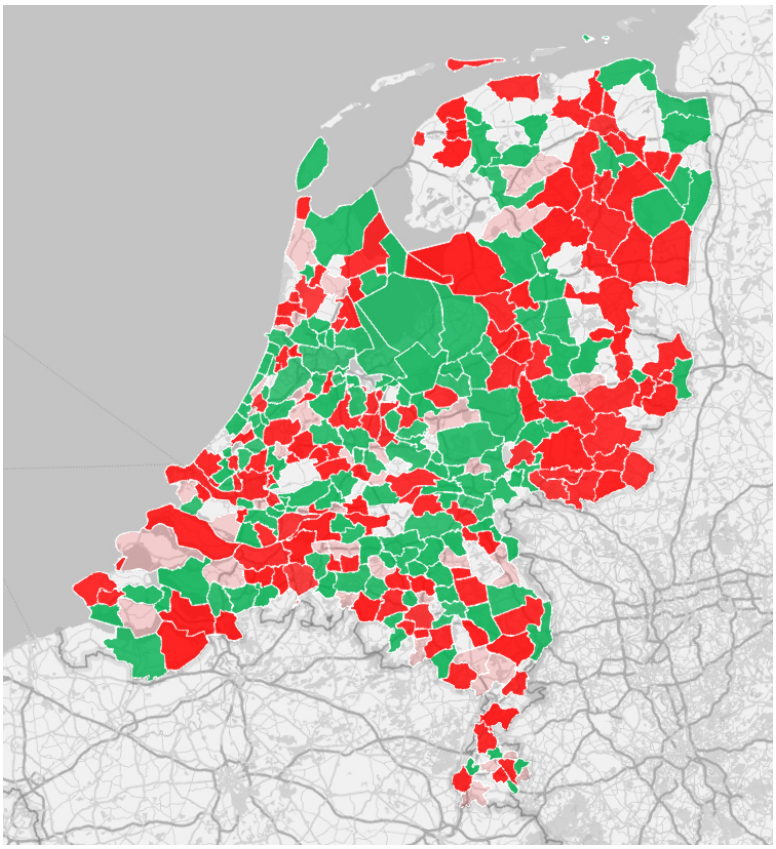
	Classifier voorspelt				
Document zit in	**Neutraal**	**Negatief**	**Positief**	**Tegenstrijdig**	***totaal***
Neutraal	51	13	18	0	82
Negatief	17	21	4	0	42
Positief	11	6	7	0	24
Tegenstrijdig	1	1	0	0	2
totaal	80	20	29	0	150

De sentiment classifier is uiteindelijk getest op de testset. De resultaten staan in Tabel 6. De resultaten die zijn behaald in de testset komen bijna overeen met de resultaten behaald tijdens het trainen van de sentiment classifier. De testset heeft een accuratesse van 0.53. In Tabel 7 is de confusion matrix van de testset te zien. Hieruit is op te maken dat van de neutrale klasse er 13 tweets als negatief zijn voorspelt en 18 tweets als positief.

5.2 RESULTATEN VAN HET GEMEENTEN ONDERZOEK

De classifier is toegepast op de gemeentebestanden. De resultaten hiervan zijn te zien in Tabel 12 in de bijlage. In Figuur 2 is een kaart te zien die een overzicht laat zien van de positieve gemeenten en negatieve gemeenten. Ter verduidelijking van figuur 2 is Tabel 8 toegevoegd. Hierin is te zien dat er 76 gemeenten neutraal zijn. Deze gemeenten hebben of helemaal geen positieve of negatieve tweets of een even aantal positieve tweets als negatieve tweets.

Voor dit onderzoek worden alleen de gemeenten meegenomen die positieve of negatieve tweets hebben. Nadat de tweets waren genormaliseerd, zoals beschreven in de methode, konden we de gemeenten met elkaar vergelijken. In Tabel 9 zijn de top 5 positieve en negatieve gemeenten te zien.



Figuur 2: Kaart van de positieve gemeenten en negatieve gemeenten. Hierbij zijn de gemeenten die lichtrood kleuren, gemeenten die een even aantal negatieve tweets als positieve tweets hebben.

Tabel 8: Tabel die weergeeft hoeveel gemeenten meer positieve tweets hebben dan negatieve en andersom. Daarnaast hoeveel gemeenten er geen positieve of negatieve tweets hebben of gelijk aan elkaar zijn, aangeduid als neutraal.

Klasse	Aantal gemeenten
Negatief	140
Positief	145
Neutraal	76
Totaal	361

Tabel 9: Tabel met de top 5 positieve gemeenten en de top 5 negatieve gemeenten.

	Meest positief	Meest Negatief
1.	Bernheze	Lingewaal
2.	Papendrecht	Noordoostpolder
3.	Raalte	Het Bildt
4.	Cranendonck	Oud-Beijerland
5.	Zeewolde	Aalten

Tabel 10 toont de totale resultaten van de 13.201 tweets. Hierin is te zien dat er meer negatieve tweets zijn dan positieve tweets.

Tabel 10: Tabel met totale sentimentindeling.

Klasse	Aantal
Negatief	3342
Positief	3010
Neutraal	6841
Tegenstrijdig	8
Totaal	13201

5.3 DISCUSSIE

De onderzoeksvraag voor dit onderzoek was of je met sentimentanalyse op tweets kon bepalen welke gemeenten positiever zijn tegenover de komst van immigranten en welke gemeenten negatiever zijn tegenover de komst van immigranten. Figuur 2 laat zien dat het mogelijk is om te bepalen of een bepaalde gemeente meer negatieve tweets heeft dan positieve tweets over immigranten. Daarnaast laat figuur 2 ook meteen zien welke gemeenten positief of negatief zijn tegenover de komst van immigranten. Ook is het mogelijk om te bepalen of een gemeente positiever of negatiever is dan een andere gemeente. Tabel 9 geeft een top 5 van positieve gemeenten en een top 5 van negatieve gemeenten.

Helaas zijn niet alle 390 gemeenten in Nederland zichtbaar. Dit komt omdat 29 gemeenten niet voorkomen in de dataset en omdat sommige gemeenten geen negatieve of positieve tweets hebben. Daardoor vallen 76 gemeenten in de neutrale categorie, waardoor je over deze gemeente niet kan zeggen of ze positief of negatief zijn. Hierdoor is er een uitval van 105 gemeenten in het onderzoek.
Daarnaast bevat de dataset voor enkele gemeenten minder dan 5 tweets. De vraag is of deze 5 tweets representatief zijn voor de gehele gemeente. Ook kan het voorkomen dat deze tweets door één persoon zijn geschreven.

Uit dit onderzoek is gebleken dat er meer gemeenten positief zijn dan negatief, terwijl over het geheel er meer negatieve tweets zijn dan positieve tweets.

In Tabel 6 is te zien dat de classifier een f1-score van 0.53 op de testset behaalt en daarnaast een accuratesse heeft van 0.53. Dit betekent dat het systeem maar net iets meer dan de helft van de tweets goed weet in te delen. De classifier presteerde beter dan de gestelde baseline. Één reden waarom de resultaten niet hoger zijn is dat de neutrale klasse heel erg groot is en de klasse tegenstrijdig erg klein. Tijdens het trainen en testen scoorde de klasse tegenstrijdig elke keer een f1-score van 0.0. Een accuratere classifier zou betrouwbaardere resultaten geven.

6 CONCLUSIE

6.1 CONCLUSIE

De gemeenten Bernheze, Papendrecht en Raalte komen als de top-3 uit dit onderzoek, die het meest positief zijn over immigranten op Twitter. Daartegenover staan Lingenwaal, Noordoostpolder en Het Bildt. Hiermee wordt op de inleidende vraag een antwoord gegeven. Deze onderzoeksresultaten zijn niet gebaseerd op alle gemeenten in Nederland. 29 gemeenten kwamen niet in de Twitterdata van 2016 voor en bij 76 gemeenten werden er geen negatieve of positieve tweets gevonden of hadden deze een gelijk aantal.

Om terug te komen op de gestelde case in de inleiding van Ter Apel, waar een azc is gevestigd, kan geconcludeerd worden dat Vlagtwedde, de gemeente waaronder Ter Apel valt, als positieve gemeente wordt gezien. Daarbij aangemerkt is dit bepaald aan de hand van slechts één positieve tweet in Vlagtwedde. Het omliggende gebied zoals Emmen heeft 44 tweets, waarvan het grote deel negatief is.

Deze resultaten zijn behaald met een linear-kernel Support Vector Machine(SVM) met als features een combinatie van woord n-grammen en letter n-grammen. Deze classifier presteerde het best voor deze opdracht, met een f1-score van 0.53 en een accuratesse van 0.53 op de testset.

Voor toekomstig onderzoek zijn er aantal aanbevelingen. Dit onderzoek heeft gebruikt gemaakt van één jaar aan tweets. Deze tweets moesten een plaats bevatten. Hierdoor zijn niet alle 390 gemeenten in de data te vinden en sommige gemeenten hebben niet meer dan 1 tweet. Voor toekomstig onderzoek zou er over meerdere jaren gekeken kunnen worden. Daarnaast is de neutrale klasse in dit onderzoek erg groot, deze zou verkleind kunnen worden door nieuwsaccounts uit te sluiten. Tot slot is het nu zo dat één twittergebruiker het sentiment kan bepalen van de hele gemeente. Om een beter beeld te krijgen van de gemeente zou je per gebruiker maar één tweet mee kunnen tellen, dit is ook gedaan bij het onderzoek van Sang and Bos (2012).

BIBLIOGRAFIE

Bird, S., E. Klein, and E. Loper (2009). *Natural Language Processing with Python*. O'Reilly Media.

CBS. Gemeenten alfabetisch 2016. Geraadpleegd op 29-05-2917.

Mohammad, S. M., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry (2016, June). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Pak, A. and P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, Volume 10, pp. 1320–1326.

Sang, E. T. K. and J. Bos (2012). Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the workshop on semantic analysis in social media*, pp. 53–60. Association for Computational Linguistics.

Sobhani, P., S. M. Mohammad, and S. Kiritchenko (2016). Detecting stance in tweets and analyzing its interaction with sentiment. *c2016 The* SEM 2016 Organizing Committee. All papers c2016 their respective authors. This proceedings volume and all papers therein are licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons. org/licenses/by/4.0*, 159.

A BIJLAGE

A.1 AANTAL TWEETS PER GEMEENTE

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Aa en Hunze	15
Aalburg	6
Aalsmeer	8
Aalten	9
Achtkarspelen	4
Alblasserdam	10
Albrandswaard	6
Alkmaar	40
Almelo	47
Almere	153
Alphen aan den Rijn	59
Alphen-Chaam	2
Ameland	6
Amersfoort	231
Amstelveen	35
Amsterdam	2.126
Apeldoorn	129
Arnhem	141
Assen	35
Asten	3
Baarn	29
Barendrecht	12
Barneveld	12
Bedum	99
Beesel	1
Bellingwedde	1
Bergeijk	1
Bergen op Zoom	29
Berkelland	10
Bernheze	25
Best	14
Beuningen	100
Beverwijk	4
Binnenmaas	10
Blaricum	6
Bloemendaal	26

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Bodegraven-Reeuwijk	31
Boekel	1
Borger-Odoorn	3
Borne	7
Borsele	4
Boxmeer	6
Boxtel	10
Breda	102
Brielle	6
Bronckhorst	1
Brummen	5
Brunssum	17
Bunnik	19
Bunschoten	9
Buren	17
Capelle aan den Ijssel	105
Castricum	24
Coevorden	14
Cranendonck	8
Cromstrijen	4
Cuijk	15
Culemborg	9
Dalfsen	23
De Bilt	29
Delft	111
Delfzijl	6
De Marne	5
Den helder	20
De Ronde Venen	28
Deurne	8
Deventer	73
De Wolden	7
Diemen	57
Dinkelland	7
Doesburg	2
Doetinchem	40
Dongen	3
Dongeradeel	55
Dordrecht	82
Drechterland	13
Drimmelen	4
Dronten	30
Druten	4
Duiven	2

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Echt-Susteren	2
Edam-Volendam	9
Ede	110
Eemsmond	8
Eersel	5
Eijsden-Margraten	5
Eindhoven	218
Elburg	18
Emmen	44
Enkhuizen	10
Enschede	102
Epe	6
Ermelo	41
Etten-Leur	2
Franekeradeel	1
Geertruidenberg	9
Geldermalsen	44
Geldrop-Mierlo	19
Gemert-Bakel	7
Gennep	4
Giessenlanden	1
Gilze en Rijen	3
Goeree-Overflakkee	16
Goes	17
Goirle	13
Gorinchem	6
Gouda	113
Grave	27
Groningen	254
Haaksbergen	1
Haaren	3
Haarlem	3
Haarlemmerliede en Spaarnwoude	87
Haarlemmermeer	298
Halderberge	3
Hardenberg	111
Harderwijk	11
Hardinxveld-Giessendam	5
Haren	98
Harlingen	17
Hattem	8
Heemskerk	5
Heemstede	13
Heerde	4

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Heerenveen	12
Heerhugowaard	20
Heerlen	98
Heeze-Leende	7
Heiloo	22
Hellendoorn	12
Hellevoetsluis	27
Helmond	32
Hendrik-Ido-Ambacht	9
Hengelo	28
Het Bildt	8
Heumen	7
Heusden	8
Hillegom	2
Hilvarenbeek	2
Hilversum	154
Hof van Twente	23
Hollands Kroon	18
Hoogeveen	8
Hoogezand-Sappemeer	22
Hoorn	32
Horst aan de Maas	9
Houten	35
Huizen	20
Hulst	2
Ijsselstein	11
Kaag En Braassem	6
Kampen	135
Kapelle	5
Katwijk	19
Kerkrade	14
Koggenland	2
Kollumerland en Nieuwkruisland	2
Korendijk	5
Krimpen aan den Ijssel	16
Laarbeek	149
Landerd	3
Landgraaf	3
Landsmeer	17
Langedijk	1
Lansingerland	95
Laren	11
Leek	1
Leerdam	2

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Leeuwarden	46
Leeuwarderadeel	5
Leiden	97
Leiderdorp	21
Leidschendam-Voorburg	28
Lelystad	54
Leudal	3
Leusden	46
Lingewaal	21
Lingewaard	11
Lisse	11
Littenseradiel	1
Lochem	23
Loon op Zand	8
Lopik	16
Loppersum	1
Losser	6
Maasdriel	1
Maasgouw	1
Maassluis	11
Maastricht	218
Medemblik	7
Meerssen	8
Menameradiel	2
Menterwolde	2
Meppel	7
Middelburg	92
Midden-Delfland	1
Midden-Drenthe	28
Mill en Sint Hubert	3
Moerdijk	26
Molenwaard	7
Montferland	44
Montfoort	9
Mook en Middelaar	2
Neder-Betuwe	1
Nederweert	2
Neerijnen	1
Nieuwegein	22
Nieuwkoop	2
Nijkerk	2
Nijmegen	232
Noord-Beveland	1
Noordenveld	41

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Noordoostpolder	12
Noordwijk	6
Noordwijkerhout	10
Nunspeet	7
Nuth	6
Oegstgeest	15
Oirschot	13
Oisterwijk	12
Oldambt	25
Oldebroek	7
Oldenzaal	11
Olst-Wijhe	11
Ommen	4
Onderbanken	1
Oosterhout	23
Oost Gelre	8
Ooststellingwerf	6
Oostzaan	3
Opmeer	1
Opsterland	8
Oss	81
Oud-Beijerland	5
Oude Ijsselstreek	21
Ouder-Amstel	8
Oudewater	2
Overbetuwe	29
Papendrecht	7
Peel en Maas	4
Pekela	5
Pijnacker-Nootdorp	18
Purmerend	41
Putten	8
Raalte	15
Reimerswaal	12
Renkum	9
Renswoude	1
Rheden	20
Rhenen	8
Ridderkerk	101
Rijnwaarden	2
Rijssen-Holten	14
Rijswijk	42
Roerdalen	3
Roermond	21

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Roosendaal	81
Rotterdam	1.273
Rucphen	2
Schagen	15
Schiedam	32
Schijndel	11
Schinnen	1
Schouwen-Duiveland	40
S-Hertogenbosch	104
Simpelveld	1
Sint-Michielsgestel	10
Sint-Oedenrode	3
Sittard-Geleen	22
Sliedrecht	12
Slochteren	1
Sluis	5
Smallingerland	28
Soest	25
Son en Breugel	2
Stadskanaal	8
Staphorst	2
Stede Broec	9
Steenbergen	14
Steenwijkerland	16
Stein	3
Stichtse Vecht	62
Strijen	13
Ten Boer	1
Terneuzen	12
Terschelling	3
Texel	14
Teylingen	105
Tholen	12
Tiel	22
Tilburg	130
Tubbergen	1
Twenterand	5
Tynaarlo	26
Tytsjerksteradiel	2
Uden	20
Uitgeest	2
Uithoorn	16
Urk	7
Utrecht	43

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	aantal tweets
Utrechtse Heuvelrug	806
Vaals	1
Valkenburg aan de Geul	2
Valkenswaard	3
Veendam	5
Veenendaal	32
Veere	7
Veghel	22
Veldhoven	15
Velsen	21
Venlo	20
Venray	9
Vianen	4
Vlaardingen	15
Vlagtwedde	1
Vlissingen	9
Voerendaal	7
Voorschoten	5
Voorst	4
Vught	7
Waalre	8
Waalwijk	22
Waddinxveen	16
Wageningen	11
Wassenaar	13
Waterland	20
Weert	24
Weesp	14
Werkendam	3
Westerveld	5
Westervoort	1
Westland	64
West Maas en Waal	4
Weststellingwerf	6
Westvoorne	2
Wijchen	13
Wijdemeren	15
Wijk Bij Duurstede	2
Winsum	6
Winterswijk	14
Woensdrecht	7
Woerden	69
Wormerland	3
Woudenberg	6

Tabel 11: Aantal tweets over immigranten in 2016 per gemeente

Gemeente	**aantal tweets**
Woudrichem	7
Zaanstad	57
Zaltbommel	17
Zandvoort	8
Zederik	3
Zeewolde	25
Zeist	35
Zevenaar	13
Zoetermeer	70
Zoeterwoude	3
Zuidhorn	11
Zuidplas	16
Zundert	1
Zutphen	35
Zwartewaterland	4
Zwijndrecht	12
Zwolle	123
totaal	13.201

A.2 RESULTATEN PER GEMEENTE

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Aa en Hunze	6	2	7	0
Aalburg	3	2	1	0
Aalsmeer	6	1	1	0
Aalten	5	0	4	0
Achtkarspelen	2	2	0	0
Alblasserdam	6	4	0	0
Albrandswaard	4	2	0	0
Alkmaar	17	11	12	0
Almelo	23	9	15	0
Almere	74	43	36	0
Alphen aan den Rijn	24	12	23	0
Alphen-Chaam	2	0	0	0
Ameland	0	2	4	0
Amersfoort	123	45	63	0
Amstelveen	19	10	6	0
Amsterdam	1.164	491	465	6
Apeldoorn	76	33	20	0
Arnhem	66	43	32	0
Assen	16	7	12	0
Asten	2	0	1	0
Baarn	19	4	6	0
Barendrecht	6	3	3	0
Barneveld	8	2	2	0
Bedum	32	13	54	0
Beesel	1	0	0	0
Bellingwedde	1	0	0	0
Bergeijk	1	0	0	0
Bergen op Zoom	15	8	6	0
Berkelland	5	2	3	0
Bernheze	6	18	1	0
Best	5	6	3	0
Beuningen	46	28	26	0
Beverwijk	2	1	1	0
Binnenmaas	3	5	2	0
Blaricum	2	3	1	0
Bloemendaal	13	9	4	0
Bodegraven-Reeuwijk	15	3	13	0
Boekel	0	0	1	0
Borger-Odoorn	2	1	0	0
Borne	6	1	0	0
Borsele	2	1	1	0
Boxmeer	4	1	1	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Boxtel	5	1	4	0
Breda	55	23	24	0
Brielle	5	0	1	0
Bronckhorst	0	0	1	0
Brummen	4	1	0	0
Brunssum	11	3	3	0
Bunnik	11	4	4	0
Bunschoten	5	3	1	0
Buren	4	5	8	0
Capelle aan den Ijssel	47	21	37	0
Castricum	16	2	6	0
Coevorden	4	2	8	0
Cranendonck	3	5	0	0
Cromstrijen	3	1	0	0
Cuijk	12	0	3	0
Culemborg	5	2	2	0
Dalfsen	17	4	2	0
De Bilt	16	5	8	0
Delft	54	22	35	0
Delfzijl	3	2	1	0
De Marne	5	0	0	0
Den helder	14	1	5	0
De Ronde Venen	15	11	2	0
Deurne	5	3	0	0
Deventer	38	26	9	0
De Wolden	7	0	0	0
Diemen	42	9	6	0
Dinkelland	4	1	2	0
Doesburg	1	1	0	0
Doetinchem	23	7	10	0
Dongen	2	0	1	0
Dongeradeel	10	17	28	0
Dordrecht	41	19	22	0
Drechterland	10	2	1	0
Drimmelen	2	0	2	0
Dronten	14	10	6	0
Druten	1	3	0	0
Duiven	1	1	0	0
Echt-Susteren	1	0	1	0
Edam-Volendam	5	1	3	0
Ede	59	30	21	0
Eemsmond	4	3	1	0
Eersel	2	2	1	0
Eijsden-Margraten	3	1	1	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Eindhoven	109	52	57	0
Elburg	11	4	3	0
Emmen	18	8	18	0
Enkhuizen	5	4	1	0
Enschede	50	13	39	0
Epe	1	2	3	0
Ermelo	24	9	8	0
Etten-Leur	1	0	1	0
Franekeradeel	1	0	0	0
Geertruidenberg	7	1	1	0
Geldermalsen	23	13	8	0
Geldrop-Mierlo	12	4	3	0
Gemert-Bakel	4	1	2	0
Gennep	2	2	0	0
Giessenlanden	0	1	0	0
Gilze en Rijen	2	1	0	0
Goeree-Overflakkee	6	2	8	0
Goes	8	7	2	0
Goirle	9	2	2	0
Gorinchem	4	1	1	0
Gouda	51	15	47	0
Grave	18	3	6	0
Groningen	96	73	85	0
Haaksbergen	1	0	0	0
Haaren	1	1	1	0
Haarlem	170	43	85	0
Haarlemmerliede en Spaarnwoude	1	0	2	0
Haarlemmermeer	52	18	17	0
Halderberge	2	0	1	0
Hardenberg	51	13	47	0
Harderwijk	5	4	2	0
Hardinxveld-Giessendam	4	0	1	0
Haren	81	5	12	0
Harlingen	8	4	5	0
Hattem	6	0	2	0
Heemskerk	2	1	2	0
Heemstede	8	1	4	0
Heerde	3	0	1	0
Heerenveen	5	4	3	0
Heerhugowaard	13	2	5	0
Heerlen	49	16	33	0
Heeze-Leende	4	1	2	0
Heiloo	17	2	3	0
Hellendoorn	8	4	0	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Hellevoetsluis	14	7	6	0
Helmond	13	10	9	0
Hendrik-Ido-Ambacht	5	4	0	0
Hengelo	15	4	9	0
Het Bildt	3	0	5	0
Heumen	5	2	0	0
Heusden	4	4	0	0
Hillegom	2	0	0	0
Hilvarenbeek	1	0	1	0
Hilversum	86	28	40	0
Hof van Twente	15	3	5	0
Hollands Kroon	8	7	3	0
Hoogeveen	4	1	3	0
Hoogezand-Sappemeer	11	2	9	0
Hoorn	16	3	13	0
Horst aan de Maas	6	1	2	0
Houten	10	14	11	0
Huizen	10	7	3	0
Hulst	1	0	1	0
Ijsselstein	7	2	2	0
Kaag En Braassem	3	2	1	0
Kampen	70	22	43	0
Kapelle	4	1	0	0
Katwijk	15	3	1	0
Kerkrade	10	2	2	0
Koggenland	0	1	1	0
Kollumerland en Nieuwkruisland	1	0	1	0
Korendijk	3	1	1	0
Krimpen aan den Ijssel	9	6	1	0
Laarbeek	56	72	21	0
Landerd	2	1	0	0
Landgraaf	2	1	0	0
Landsmeer	14	3	0	0
Langedijk	0	1	0	0
Lansingerland	37	16	42	0
Laren	2	4	5	0
Leek	1	0	0	0
Leerdam	2	0	0	0
Leeuwarden	29	12	5	0
Leeuwarderadeel	4	1	0	0
Leiden	61	23	13	0
Leiderdorp	8	10	3	0
Leidschendam-Voorburg	17	3	8	0
Lelystad	26	16	12	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Leudal	1	1	1	0
Leusden	19	15	12	0
Lingewaal	13	0	8	0
Lingewaard	8	2	1	0
Lisse	6	3	2	0
Littenseradiel	0	0	1	0
Lochem	14	3	6	0
Loon op Zand	5	1	2	0
Lopik	11	1	4	0
Loppersum	1	0	0	0
Losser	2	3	1	0
Maasdriel	1	0	0	0
Maasgouw	1	0	0	0
Maassluis	6	2	3	0
Maastricht	112	31	75	0
Medemblik	2	2	3	0
Meerssen	6	2	0	0
Menameradiel	1	0	1	0
Menterwolde	2	0	0	0
Meppel	2	3	2	0
Middelburg	37	19	36	0
Midden-Delfland	0	1	0	0
Midden-Drenthe	11	6	11	0
Mill en Sint Hubert	3	0	0	0
Moerdijk	11	6	9	0
Molenwaard	2	4	1	0
Montferland	16	13	15	0
Montfoort	5	1	3	0
Mook en Middelaar	1	1	0	0
Neder-Betuwe	0	1	0	0
Nederweert	0	1	1	0
Neerijnen	1	0	0	0
Nieuwegein	15	4	3	0
Nieuwkoop	0	1	1	0
Nijkerk	1	1	0	0
Nijmegen	111	66	55	0
Noord-Beveland	1	0	0	0
Noordenveld	8	7	26	0
Noordoostpolder	6	0	6	0
Noordwijk	4	4	2	0
Noordwijkerhout	4	1	1	0
Nunspeet	4	2	1	0
Nuth	2	2	2	0
Oegstgeest	11	2	2	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Oirschot	3	4	6	0
Oisterwijk	5	4	3	0
Oldambt	17	7	1	0
Oldebroek	2	1	4	0
Oldenzaal	6	1	4	0
Olst-Wijhe	6	2	3	0
Ommen	4	0	0	0
Onderbanken	1	0	0	0
Oosterhout	16	4	3	0
Oost Gelre	5	1	2	0
Ooststellingwerf	3	0	3	0
Oostzaan	2	1	0	0
Opmeer	1	0	0	0
Opsterland	4	2	2	0
Oss	52	16	13	0
Oud-Beijerland	1	0	4	0
Oude Ijsselstreek	8	5	8	0
Ouder-Amstel	5	1	2	0
Oudewater	1	0	1	0
Overbetuwe	10	12	7	0
Papendrecht	1	6	0	0
Peel en Maas	1	0	3	0
Pekela	2	1	2	0
Pijnacker-Nootdorp	9	5	4	0
Purmerend	19	10	12	0
Putten	3	2	3	0
Raalte	9	6	0	0
Reimerswaal	4	1	7	0
Renkum	7	1	1	0
Renswoude	1	0	0	0
Rheden	7	6	7	0
Rhenen	4	3	1	0
Ridderkerk	46	25	30	0
Rijnwaarden	2	0	0	0
Rijssen-Holten	10	2	2	0
Rijswijk	26	10	6	0
Roerdalen	3	0	0	0
Roermond	10	4	7	0
Roosendaal	38	28	15	0
Rotterdam	612	228	431	2
Rucphen	1	1	0	0
Schagen	7	4	4	0
Schiedam	22	4	6	0
Schijndel	9	2	0	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Schinnen	0	1	0	0
Schouwen-Duiveland	18	11	11	0
S-Hertogenbosch	51	37	16	0
Simpelveld	0	1	0	0
Sint-Michielsgestel	7	2	1	0
Sint-Oedenrode	1	0	2	0
Sittard-Geleen	13	3	6	0
Sliedrecht	8	2	2	0
Slochteren	1	0	0	0
Sluis	5	0	0	0
Smallingerland	14	11	3	0
Soest	12	8	5	0
Son en Breugel	0	1	1	0
Stadskanaal	4	3	1	0
Staphorst	1	1	0	0
Stede Broec	3	1	5	0
Steenbergen	8	3	3	0
Steenwijkerland	10	5	1	0
Stein	3	0	0	0
Stichtse Vecht	34	11	17	0
Strijen	8	2	3	0
Ten Boer	1	0	0	0
Terneuzen	7	3	2	0
Terschelling	3	0	0	0
Texel	7	5	2	0
Teylingen	47	27	31	0
Tholen	11	1	0	0
Tiel	19	1	2	0
Tilburg	69	36	25	0
Tubbergen	1	0	0	0
Twenterand	3	0	2	0
Tynaarlo	10	10	6	0
Tytsjerksteradiel	2	0	0	0
Uden	16	2	2	0
Uitgeest	0	1	1	0
Uithoorn	9	1	6	0
Urk	5	0	2	0
Utrecht	438	217	151	0
Utrechtse Heuvelrug	17	14	12	0
Vaals	1	0	0	0
Valkenburg aan de Geul	2	0	0	0
Valkenswaard	1	1	1	0
Veendam	0	1	4	0
Veenendaal	19	8	5	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Veere	4	1	2	0
Veghel	13	7	2	0
Veldhoven	7	3	5	0
Velsen	8	7	6	0
Venlo	9	7	4	0
Venray	4	3	2	0
Vianen	1	1	2	0
Vlaardingen	10	3	2	0
Vlagtwedde	0	1	0	0
Vlissingen	4	4	1	0
Voerendaal	2	1	4	0
Voorschoten	3	1	1	0
Voorst	3	0	1	0
Vught	5	2	0	0
Waalre	1	1	6	0
Waalwijk	14	4	4	0
Waddinxveen	12	2	2	0
Wageningen	6	4	1	0
Wassenaar	5	6	2	0
Waterland	10	6	4	0
Weert	8	5	11	0
Weesp	6	7	1	0
Werkendam	2	0	1	0
Westerveld	3	0	2	0
Westervoort	1	0	0	0
Westland	23	12	29	0
West Maas en Waal	1	1	2	0
Weststellingwerf	4	1	1	0
Westvoorne	0	1	1	0
Wijchen	5	3	5	0
Wijdemeren	6	5	4	0
Wijk Bij Duurstede	2	0	0	0
Winsum	4	0	2	0
Winterswijk	7	3	4	0
Woensdrecht	4	0	3	0
Woerden	48	12	9	0
Wormerland	0	2	1	0
Woudenberg	3	2	1	0
Woudrichem	3	4	0	0
Zaanstad	29	19	9	0
Zaltbommel	12	1	4	0
Zandvoort	4	3	1	0
Zederik	0	2	1	0
Zeewolde	8	15	2	0

Tabel 12: Tabel met de resultaten van de gemeente

Gemeente	Neutraal	positief	Negatief	Tegenstrijdig
Zeist	14	9	12	0
Zevenaar	5	5	3	0
Zoetermeer	39	16	15	0
Zoeterwoude	0	3	0	0
Zuidhorn	1	3	7	0
Zuidplas	8	7	1	0
Zundert	1	0	0	0
Zutphen	24	8	3	0
Zwartewaterland	3	0	1	0
Zwijndrecht	7	4	1	0
Zwolle	88	21	14	0
totaal	6.841	3.010	3.342	8

A.3 GEBRUIKTE SCRIPTS

De volgende lijst is een lijst met programma's die gebruikt zijn voor dit onderzoek deze zijn terug te zien in github. https://github.com/Moniekleonie/Ba-scriptie

A.3.1 Data en materiaal

Naam: Scriptiedata.sh
Input: Alle tweets van de Rijksuniversiteit Groningen van 2016
Output: Alle tweets in één tekstdocument met meegegeven metadata:

- ID
- gebruikersnaam
- tweet
- woorden
- hastags
- datum
- plaats

Functies: Geen

Naam: plaatsachterhalen.py
Input: Alle verzamelde tweets van scriptiedata.sh
Output: Alleen de tweets waarbij meegegeven plaats niet leeg is
Functies: main

Naam: maketraining.py
Input: Alle tweets met gemeentenaam
Output: 1000 random gekozen tweets met een vaste seed van 26
Functies: main

Naam: sortedonplacenames.py
Input: Alle tweets met gemeentenaam
Output: Per gemeente een tekstbestand met tweets uit die gemeente
Functies: main, gemeentes

A.3.2 Methode

Naam: classifier.py
Input: Genoteerde trainset, Geannoteerde testset, Gemeente tekstbestanden
Output: Per fold classifier rapport, Classifier rapport voor testset, Per gemeente aantal per klasse
Functies: main, getgoal, readTweets, tokenizeTweets, ngramschar, ngrams, cross-train, classifier, identity, evaluate, confusionMatrix