rijksuniversiteit groningen / faculteit der letteren

Kun je volgers in een hokje plaatsen?

Voorspellen van politieke volgers op Twitter middels woordgebruik

Aileen Bus

Bachelor scriptie
Informatiekunde
Aileen Bus
S2546167
21 juni 2018

SAMENVATTING

In deze scriptie wordt onderzocht of er te voorspellen valt of een gebruiker op Twitter een bepaalde politieke leider volgt. In de programmeertaal Python met de Scikit-Learn module wordt door middel van 10-fold cross validatie en n-grammen gekeken naar de f1-score van zes groepen van in totaal 14.742 unieke volgers van zes verschillende politici.

Uit dit onderzoek komt naar voren dat met een baseline van 26% en een hoogste score van 56% met het LinearSVC algoritme en 49% met het SGD-Classifier algoritme, men met deze dataset niet kan voorspellen middels tweets of een gebruiker op Twitter een bepaalde politieke leider volgt. De groepen met meer volgers scoren veel hoger dan politici met minder volgers. Daarnaast scoren specifiekere groepen volgers ook veel hoger dan algemenere groepen volgers.

Een volgend onderzoek zou deze score kunnen verbeteren door een betere balans aan te brengen in de dataset, specifieker te selecteren op volgers (profielinformatie) of door betere preprocessing tools te gebruiken.

INHOUDSOPGAVE

VOORWOORD

Voor u ligt mijn scriptie: *"Kun je volgers in een hokje plaatsen? Voorspellen van politieke volgers op Twitter middels woordgebruik"*. Deze scriptie is geschreven in het kader van het afstudeerproject van de bachelor Informatiekunde. Eindelijk mag ik deze woorden opschrijven. In september 2013 begon ik met de opleiding Informatiekunde. De afgelopen vijf jaar heb ik veel geleerd. Met amper kennis van programmeren beklom ik de eerste dag de treden van het Academiegebouw. Niet alleen vakinhoudelijk heb ik veel kennis opgedaan, maar ook op persoonlijk vlak heb ik progressie gemaakt. In de de afgelopen vijf jaar is er veel gebeurd. Na een burn-out en een jaar er tussenuit zijn geweest heb ik het dagelijkse leven heel anders in moeten delen. Met vallen en opstaan is nu eindelijk bijna het einde in zicht. Als een berg keek ik op tegen het maken van de scriptie. Uiteindelijk viel het mij mee, alhoewel ik blij ben dat het er nu op zit. Ik heb er veel van geleerd en het is leuk om alle theorie uiteindelijk bij elkaar in de praktijk te kunnen brengen.

Graag wil ik mijn moeder, broer en grootouders bedanken voor alle steun en vertrouwen de afgelopen jaren. "Schouders er maar weer onder kind", zegt mijn Oma altijd.

Die eerste dag daar op de treden van het Academiegebouw ontmoette ik mijn studiegenoot en goede vriend Reinard van Dalen. Reinard, bedankt voor je vriendschap, je hulp, alle lachbuien, de tranen die ik met je kan delen en de mooie herinneringen aan mijn studietijd die ik samen met jou heb meegemaakt.

Daarnaast Stijn Eikelboom, ook erg bedankt voor al jouw hulp, de koffie en je gezelschap. Mijn studiegenoot van het eerste uur Joël Coster: zonder onze ontmoetingen de laatste weken was de scriptie mij misschien niet gelukt, de stok achter de deur! Heel erg bedankt voor je steun en uitleg.

Mijn scriptiebegeleider Martijn Wieling wil ik bedanken voor de kritische feedback en de vroege eerste deadlines. Dat heeft toch erg geholpen.

Nog een paar maanden, maar ik kan bijna niet wachten om dan op de trappen van het Academiegebouw te mogen staan, dit keer hopelijk met Bul!

1 INTRODUCTIE

1.1 INLEIDING

Anno 2018 maken 2,8 miljoen Nederlanders gebruik van Twitter.[1] Dagelijks openen ongeveer één miljoen Nederlanders hun Twitter-app. In de laatste jaren heeft Twitter zich ontwikkeld van een platform dat door jongeren werd gebruikt tot een interactiekanaal tussen journalisten, wetenschappers, politici en staatshoofden. Voor politici is sociale media tegenwoordig een noodzakelijk goed. Online bereiken de partijen een groter en breder publiek dan met flyeren op straat. In het boek *"Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public"*, concluderen Pamelee and Bichard (2011) dat politiek geïnteresseerden niet langer voorbij kunnen gaan aan de kansen die Twitter biedt om de uitkomst van campagnes en wetgeving te beïnvloeden. Ook hier in Nederland leidt meer dan eens ophef op sociale media tot een debat in de Tweede Kamer. Uit een onderzoek in 2014 van communicatiebureau Weber Shandwick blijkt dat Kamerleden Twitter een belangrijker medium vinden dan radio en televisie om hun dagelijks werk onder de aandacht te brengen.[2] Tegenwoordig gebruiken Tweede Kamerleden het sociale medium ook om discussiepunten voor te leggen, het kabinet te bekritiseren of aan te spreken en op te roepen tot actie. Zo blijkt uit het artikel "Twitter-motie rukt op als wapen" in de Volkskrant van 12 april 2018.[3]

Het grote aantal berichten dat beschikbaar is, samen met meta-data als bijvoorbeeld datum, tijd en geografische locatie maakt Twitter ook voor taalkundig onderzoek een interessante bron. Vele onderzoeken zijn er al gedaan naar bijvoorbeeld de voorspellende gave van Twitter. Met Twitter kan men griepuitbraken in kaart brengen, de economische koers voorspellen of berekenen wat de kans is op voedselvergiftiging in een restaurant. Ook de uitslagen van verschillende politieke verkiezingen zijn geprobeerd te voorspellen met behulp van Twitter-data. Naast verkiezingen en referenda proberen onderzoekers ook dikwijls te voorspellen wat de politieke voorkeur van een gebruiker is. Dit is al gedaan op basis van profielinformatie, woordgebruik, vrienden (volgers) en hashtags meta-data (meer informatie over deze onderzoeken is te vinden in hoofdstuk 2).

In dit onderzoek ga ik kijken naar de volgers van politici in Nederland. Nederland heeft een uitgebreid politiek landschap met veel verschillende partijen. Als men het woordgebruik van de volgers van verschillende politieke leiders in kaart brengt, valt er dan te voorspellen wie ook bij deze groep hoort? Kun je volgers in een (politiek) hokje plaatsen?

1.2 ONDERZOEKSVRAAG

In deze scriptie ga ik onderzoek doen naar het taalgebruik van groepen volgers op Twitter. De hoofdvraag die ik hierbij stel is: *Valt er te voorspellen middels het woordgebruik in tweets of een gebruiker op Twitter een bepaalde politieke leider volgt?*

[1] van der Veer, N. et al. (2018, januari) Nationale Social Media Onderzoek 2018: het grootste trendonderzoek van Nederland naar het gebruik en verwachtingen van social media. *Newcom Research & Consultancy B.V.*

[2] de Boer, N. et al. (2014) Twitter en de Tweede Kamer. *Weber Shandwick*

[3] von Piekartz, H. (2018, april) De Twitter-motie: het nieuwe politieke wapen van Tweede Kamerleden. *Volkskrant*.

1.3 VERWACHTE UITKOMSTEN

Soort zoekt soort, is de veronderstelling. Volgens het principe van McPherson et al. (2001) zijn sociale netwerken homogeen. Mensen hebben de neiging om relaties te onderhouden met mensen die op hen lijken, bijvoorbeeld relaties tussen mensen met dezelfde leeftijd, hetzelfde ras, hetzelfde geslacht, dezelfde religie of hetzelfde beroep. De verwachting is dat woordgebruik hier dan ook onder zal vallen. De hypothese van mijn scriptie is dan ook dat er een overeenkomst is op basis van n-grammen in het woordgebruik tussen de tweets van een groep volgers op Twitter. De overeenkomst is aan te tonen door een hoge gemiddelde F1-score. Dit betekent dat op basis van de tweets van een gebruiker te voorspellen valt welke politicus deze gebruiker volgt.

1.4 OPBOUW

De opbouw van de scriptie ziet er als volgt uit. In hoofdstuk 2: Theoretisch kader is het literatuuronderzoek te vinden. In hoofdstuk 3: Data wordt beschreven hoe de data is verzameld en welke keuzes hierbij zijn gemaakt. Ook is hier te vinden welke stappen genomen zijn om de data voor te bereiden op het machine learning algoritme. In hoofdstuk 4: Methode staat vermeld welk classificatie algoritme is gebruikt en met welke kenmerken het onderzoek wordt uitgevoerd. De resultaten worden besproken in hoofdstuk 5: Resultaten en de conclusie vindt u in hoofdstuk 6: Conclusie.

1.5 CODE

Alle Python code en andere bestanden behorende bij dit onderzoek zijn te vinden met deze Github-link: `https://github.com/aileenbus/BachelorThesis.git`.

2 THEORETISCH KADER

In de laatste jaren zijn er veel en verschillende soorten onderzoeken gedaan naar (politieke) voorspellingen en taalgebruik op basis van internet gerelateerde data. Tumasjan et al. (2010) heeft onderzoek gedaan naar het voorspellen van de Duitse verkiezingen door meer dan 100.000 tweets te analyseren. De onderzoekers kwamen tot de ontdekking dat een derde van alle berichtgevingen op Twitter deel uitmaken van een conversatie. Dit indiceert dat niet alleen politieke ideeën publiekelijk gedeeld worden, maar dat hieruit ook discussies voortkomen. De conclusie van dit onderzoek is dat het aantal tweets waar politieke partijen genoemd worden het stemgedrag weerspiegelt. Met een mean absolute error (MAE) van 1,68% zitten de voorspellingen middels Twitter dicht in de buurt van de tradionele peilingbureau's. Het sentiment in de Twitter-berichten komt nauw overeen met de politieke programma's, de kandidaatprofielen en de berichtgeving in de media.

Tjong Kim Sang and Bos (2012) hebben ook onderzoek gedaan naar het voorspellen verkiezingen; zij deden dit voor de Nederlandse verkiezingen in 2011. Met het tellen van gefilterde tweets waar politieke partijen werden genoemd kwamen de onderzoekers met hun resultaten ook dicht in de buurt van de professionele peilingen.

Barberá (2013) heeft een model gemaakt dat de politieke ideologie van een actieve Twitter-gebruiker kan afleiden. Dit is uitsluitend gebaseerd op wie de gebruiker volgt op Twitter.

Conover et al. (2011) onderzochten of te voorspellen valt wat voor politieke ideologie een Twitter-gebruiker heeft. Hiervoor onderzochten zij verschillende methodes. De onderzoekers concluderen dat een Support Vector Machine (SVM) getraind met hashtag meta-data beter werkt dan een SVM welke getraind is op de teksten in tweets van gebruikers.

Park (2013) heeft uitgezocht of Twitter de betrokkenheid in politiek motiveert. Uit zijn onderzoek blijkt dat opinieleiders op Twitter een belangrijke bijdrage leveren aan de politieke betrokkenheid van individuen. Dit bleek niet per se het geval te zijn wanneer iemand alleen Twitter of een ander sociaal medium gebruikt zonder specifieke opinieleiders te volgen.

Bouma (2015) laat zien hoe de n-gram frequentie data is verkregen uit een grote steekproef van Nederlandse tweets uit een periode van vier jaar. Na het filteren op bepaalde kenmerken als retweets, duplicaten en niet-Nederlandse tweets, bleven er ruim 2.6 miljard tweets over. De webinterface van de Rijksuniversiteit Groningen stelt gebruikers in staat om interactieve queries op deze dataset uit te voeren.[1]

[1] http://www.let.rug.nl/gosse/Ngrams/ngrams.html

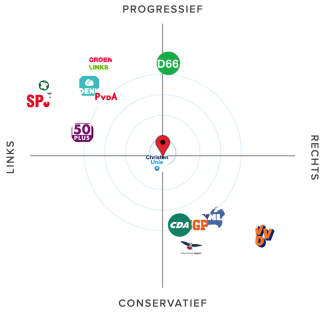
3 DATA

In dit hoofdstuk wordt beschreven welke data is verzameld en hoe deze gegevens zijn verwerkt.

3.1 SELECTIE VAN POLITICI

3.1.1 Politiek landschap

Voor een goede selectie voor dit onderzoek van politici in Nederland, is het belangrijk om ten eerste te kijken naar het politieke landschap. Nederland kent veel partijen en er is een onderscheid te maken in confessionele en niet-confessionele partijen, progressieve en conservatieve partijen of linkse, rechtse en middenpartijen. In figuur 1 is een overzicht te zien van het politieke landschap in 2017 tijdens de Tweede Kamerverkiezingen. Om alle richtingen te representeren is gekeken naar een extreemlinkse partij, linkse partij, midden partij, christelijke partij, rechtse partij en een extreemrechtse partij.



Figuur 1: Het politieke landschap in Nederland tijdens de Tweede Kamerverkiezingen van 2017 (bron: Kieskompas)

3.1.2 Representatie van politici op Twitter

In Nederland deden er 28 partijen mee tijdens de Tweede Kamerverkiezingen van 2017. De veelbesproken belangrijkste 18 partijen zijn meegenomen in het onderzoek. Om te bepalen welke politici geselecteerd werden is er gekeken naar de representatie op Twitter van de lijsttrekkers van deze 18 partijen. In tabel 1 zijn alle lijsttrekkers van de verkiezingen van 2017 opgenomen inclusief hun aantal tweets en volgers in april 2018.

3.1.3 Selectie

De volgende zes politici zijn uiteindelijk gekozen:

1. Sybrand van Haersma-Buma (Christen-Democratisch Appèl)
2. Klaas Dijkhoff (Volkspartij voor Vrijheid en Democratie)
3. Jesse Klaver (GroenLinks)

Tabel 1: Nederlandse lijsttrekkers tijdens de Tweede Kamerverkiezingen van 2017 op Twitter met aantal tweets en volgers (april 2018).

Naam	Partij	Tweets gestuurd	Volgers
Mark Rutte	VVD	16	112.644
Lodewijk Asscher	PvdA	11.889	325.168
Geert Wilders	PVV	10.656	950.032
Emile Roemer	SP	3.255	203.021
Sybrand van Haersma-Buma	CDA	2.013	77.270
Alexander Pechtold	D66	13.106	692.342
Jesse Klaver	GroenLinks	2.065	234.114
Kees van der Staaij	SGP	4.335	75.340
Marianne Thieme	Partij voor de Dieren	19.177	123.132
Henk Krol	50Plus	12.123	16.950
Norbert Klein	Vrijzinnige Partij	3.486	1.779
Ancilla van de Leest	Piratenpartij	61.352	53.980
Robert Valentine	Libertarische Partij	1.142	1.127
Tunahan Kuzu	DENK	4.074	31.322
Jan Roos	VoorNederland	93.814	137.908
Thierry Baudet	Forum voor Democratie	7.492	147.208
Jan Dijkgraaf	GeenPeil	9.999	32.799
Gert-Jan Segers	ChristenUnie	39.174	33.122

4. Lilian Marijnissen (Socialistische Partij)
5. Alexander Pechtold (Democraten 66)
6. Geert Wilders (Partij voor de Vrijheid)

Voor extreemlinks is gekozen voor de SP, omdat dit de meest linkse partij is van Nederland. Emile Roemer trad op 13 december 2017 af als fractievoorzitter, daarom is hier gekozen voor de huidige fractievoorzitter Lilian Marijnissen (sinds 23 maart 2017 lid van de Tweede Kamer). Van de extreemlinkse politici tweet Marijnissen veel en regelmatig. Haar collega van GroenLinks, Jesse Klaver, heeft minder tweets geplaatst, maar heeft wel meer volgers. De reden om voor GroenLinks te kiezen als linkse partij en niet voor bijvoorbeeld DENK (te weinig volgers) of PvdA, is dat GroenLinks een tikje linkser is en "groener" (staat voor klimaat, milieu en duurzaamheid). Ook is Jesse Klaver erg populair onder jongeren, waardoor er weer een andere doelgroep bereikt wordt.

Tabel 2: Geselecteerde Nederlandse politici op Twitter met aantal tweets en volgers (april 2018).

Naam	Ideologie	Partij	Tweets	Volgers	Link
Lilian Marijnissen	Links	SP	11.967	16.932	https://twitter.com/MarijnissenL
Jesse Klaver	Links	GroenLinks	2.072	241.717	https://twitter.com/jesseklaver
Klaas Dijkhoff	Rechts	VVD	19.044	51.921	https://twitter.com/dijkhoff
Geert Wilders	Rechts	PVV	10.656	950.032	https://twitter.com/geertwilderspvv
Alexander Pechtold	Midden	D66	13.106	692.342	https://twitter.com/APechtold
Sybrand van Haersma-Buma	Midden	CDA	2.014	77.377	https://twitter.com/sybrandbuma

Voor de rechtse stroming is gekozen voor Geert Wilders (extreemrechts) en Klaas Dijkhoff (rechts). Mark Rutte, de minister-president van Nederland (VVD) is op persoonlijke titel amper aanwezig op Twitter, alhoewel hij wel veel volgers heeft. De kans is alleen dat mensen hem volgen omdat hij minister-president is en niet zijn ideologie delen. Zijn minister-presidentaccount is opgezet als promotie voor zijn werk en Nederland, hier wordt de redactie dan ook gedaan door de Rijksvoorlichtingsdienst. Daarom is hier gekozen voor Klaas Dijkhoff (fractievoorzitter van de VVD), die regelmatig tweet en genoeg volgers heeft. Geert Wilders representeert hier de extreemrechtse partij. Geert Wilders doet bijzondere uitspraken en heeft gigantisch veel volgers. Voor de politici in het midden zijn Alexander Pechtold en Sybrand Buma geselecteerd. D66 is een progressieve middenpartij. Alexander Pechtold tweet regelmatig en is populair. Buma tweet weer wat minder, maar zijn partij CDA is een christelijke middenpartij, waardoor het midden completer gerepresenteerd wordt. Gert-Jan Segers van de ChristenUnie heeft ongeveer de helft minder volgers dan Buma, waardoor de keus niet op Segers is gevallen.

3.2 TWEETS VERZAMELEN

3.2.1 Nederlands Twitter corpus

Om het onderzoek uit te kunnen voeren is er Twitterdata verzameld. Hier is gebruik gemaakt van het Nederlands Twitter corpus van de Rijksuniversiteit Groningen. Vanaf 2010 is de Rijksuniversiteit Groningen (RuG) bezig om Nederlandse Twitterdata te verzamelen. Hier wordt naast de tekst van de tweets ook metadata vergaard, waaronder de datum en de tijd van het versturen van de tweet, identiteitsnummer van de auteur, profielnaam van de auteur en geografische coördinaten indien aanwezig. Met behulp van de Twitter Application Programming Interface (API)[1] haalt de RuG de tweets op. De tweets worden geselecteerd op basis van een lijst met Nederlandse woorden (Tjong Kim Sang, 2011). In deze lijst staan woorden die niet of nauwelijks voorkomen in andere talen behalve het Nederlands. De tweets worden gecomprimeerd per uur opgeslagen in het corpus en geordend op datum. Aan dit corpus worden elke dag een half- tot anderhalf miljoen tweets toegevoegd. De laatste jaren wordt dit iets minder doordat het verkeer op Twitter afneemt.

Om aan het einde van de selectie genoeg data over te houden zijn voor dit onderzoek de tweets van de afgelopen twee jaren verzameld. Dit zijn alle tweets uit het Nederlandse corpus van mei 2016 tot en met april 2018. De meest recente tweets zijn nodig, omdat de huidige lijst van de volgers van een politicus wordt gebruikt en er geen datum van volgen bekend is.

3.2.2 Alle volgers verzamelen

De huidige lijst van de volgers van de politici worden opgehaald om uiteindelijk de tweets van deze gebruikers uit de verzameling tweets van de afgelopen twee jaar te kunnen filteren. Hiervoor is gebruik gemaakt van het programma Postman (een API Development Environment)[2] en de Twitter API. De uiteindelijk opgehaalde zes tekstbestanden bestaan uit een lijst met identiteitsnummers van alle volgers van een politicus. Met een programma zijn alle identiteitsnummers die in een van de andere lijsten voorkomen verwijderd, zodat elke politicus een lijst van gebruikers overhoudt die alleen hem of haar volgen. Het aantal unieke volgers per politicus is weergegeven in tabel 3. Zoals te zien in de tabel is er een onbalans in de aantallen unieke volgers. Geert Wilders heeft bijvoorbeeld veel volgers in vergelijking tot Lilian Marijnissen. Dit is de realiteit en om zo dicht mogelijk tegen de werkelijkheid aan te kunnen voorspellen of een gebruiker op Twitter een bepaalde politieke leider volgt, wordt hier geen gebalanceerde dataset gebruikt. Wel wordt dit opgevangen in het classificatie-algoritme met de "balanced-modus (zie hoofdstuk 4).

Tabel 3: Aantal unieke volgers per politicus in de dataset.

Politici	Aantal volgers
Sybrand Buma	10.877
Klaas Dijkhoff	15.007
Jesse Klaver	91.642
Lilian Marijnissen	5.000
Alexander Pechtold	372.737
Geert Wilders	607.561
Totaal	1.102.824

[1] https://developer.twitter.com/content/developer-twitter/en.html
[2] https://www.getpostman.com/

3.2.3 Tweets selecteren

Om een dataset te maken met tweets van de volgers van de zes politici is eerst over de tweets uit het Nederlandse Twitter corpus geïtereerd om bij een vergelijkend identiteitsnummer van een volger deze tweets in een apart tekstbestand op te slaan. Uiteindelijk gaf dit een dataset van in totaal 38 miljoen tweets van alle volgers die te vinden zijn in het Nederlandse corpus van de afgelopen twee jaren.

3.2.4 Volgers selecteren

Omdat 38 miljoen tweets erg veel data is en niet alle 1 miljoen volgers een goede afspiegeling zijn van de volgersgroep, is hier ook een selectie gemaakt. De volgers zijn geselecteerd op de hoeveelheid tweets die zij verstuurd hebben in de afgelopen twee jaren. De bovengrens is hierbij gelegd op 2000 tweets. Dit, omdat een gemiddeld persoon maximaal 2 à 3 tweets op een dag stuurt, dit aantal keer 730 dagen komt op ongeveer 2000 tweets. Twitteraccounts die meer dan 3 tweets per dag sturen zijn hoogstwaarschijnlijk bots. Met deze bovengrens wordt geprobeerd om de accounts van (spam)bots, of bijvoorbeeld webcare-accounts en nieuwssites te filteren. De ondergrens is vastgezet op 300 tweets. De reden hiervoor is dat de dataset anders te groot wordt. Daarnaast, hoe meer data per gebruiker, hoe beter. Door deze selectie worden de personen die minder actief zijn op Twitter uit de dataset gefilterd. De uiteindelijke verdeling van de labels in de dataset ziet eruit als in tabel 4.

Tabel 4: Aantal volgers per politicus in de dataset.

Politici	Aantal volgers
Sybrand Buma	664
Klaas Dijkhoff	1.331
Jesse Klaver	2.326
Lilian Marijnissen	737
Alexander Pechtold	4.258
Geert Wilders	5.426
Totaal	14.742

3.3 DATA VOORBEREIDEN EN ANNOTEREN

Om de dataset bruikbaar te maken voor het onderzoek zijn de tweets eerst voorbewerkt. De metadata die niet wordt gebruikt is verwijderd. Hierdoor blijft alleen een identiteitsnummer van de volger en de tekst van de tweet over.

De volgende voorbewerkingsmethodes zijn gebruikt:

1. Tweet tekst omgezet in kleine letters
2. Verwijderen van retweets
3. URLs genormaliseerd naar een token
4. Mentions genormaliseerd naar een token
5. Begin en eind van een tweet aangeduid met een token
6. Tweets getokeniseerd met de NLTK TweetTokenizer (Bird et al. 2009)
7. Verwijderen van klanktekens

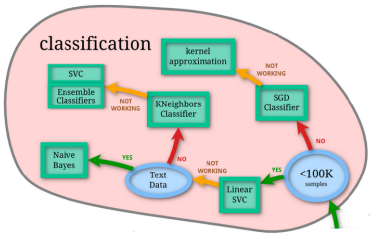
De klanktekens, zoals 'é' in 'hé', worden verwijderd tijdens het voorbereiden van de dataset. Hierdoor maakt het niet uit of iemand 'Hè', 'Hé' of 'He' schrijft. Alles wordt in kleine letters omgezet naar 'he'.

De tweets per gebruiker worden samengevoegd en zo als één stuk tekst verwerkt door het machine learning algoritme. Om ervoor te zorgen dat er geen bigrammen worden gegenereerd met een woord aan het einde van de ene tweet en een woord aan het begin van de volgende tweet, wordt hier een begin en een eind marker toegevoegd aan de tweets. Vervolgens zijn de labels van de politici bij de tweets gezocht en blijven er twee bestanden over: een tekstbestand met alle tweets, op elke regel één en een tekstbestand met op elke regel een label, waar elk regelnummer in de labels correspondeert met het regelnummer van de tweet.

4 METHODE

4.1 KEUZE VAN CLASSIFICATIE ALGORITME

Voor dit onderzoek is er een keus gemaakt voor twee classificatie algoritmes met behulp van de cheatsheet van de website Scikit-Learn.[1] Hier is de keus gevallen op de Stochastic Gradient Descent Classifier (SGD-Classifier)[2] en de LinearSVC.[3] Beide algoritmes zijn een lineaire Support Vector Machine (SVM), wat betekent dat ze gecontroleerd machinaal leren. Het verschil tussen deze twee algoritmes is dat LinearSVC altijd de volledige data gebruikt, terwijl de SGDClassifier de data in batches verwerkt. Op grote datasets is het beter om de SGD-classifier te gebruiken dan Logistische Regressie, want dit werkt veel sneller. De SGD-classifier is zo ingesteld dat het een lineaire SVM is. Deze keuze is gemaakt omdat SVM voor tekstdata wegens algemeen goede resultaten vaak aanbevolen wordt (Joachims, 1998).



Figuur 2: Classificatie algoritme cheatsheet van Scikit-Learn

4.2 SELECTIE N-GRAMMEN

Voor het voorspellen van de politici bij de volgers wordt gekeken naar woord n-grammen in de tweets. Dit zijn sets van woorden die samen voorkomen in een tekst. Gebruikt worden de volgende kenmerken:

1. unigrammen
2. bigrammen
3. trigrammen
4. combinaties van deze drie kenmerken

4.3 EVALUATIE

Om te kunnen trainen en testen met de dataset moet deze worden verdeeld in een trainingset en een testset. De dataset is daarom willekeurig verdeeld in 90% training en 10% test. De verdeling is te zien in tabel 5. De verdeelde dataset is opgeslagen

[1] http://scikit-learn.org/stable/
[2] http://scikit-learn.org/stable/modules/sgd.html
[3] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

in pickle-bestanden, omdat serialisatie (pickling) snel en makkelijk werkt. Serialisatie is het omzetten van een object naar een indeling die kan worden opgeslagen, verzonden of later kan worden gereconstrueerd.

Daarnaast wordt voor dit onderzoek gebruik gemaakt van 10-fold cross validation. Bij 10-fold cross validation wordt de dataset opgedeeld in 10 gelijke delen. Hiervan wordt op 9 delen getraind en op 1 deel getest. Dit wordt 10 keer gedaan met elke keer een ander deel waarop getest wordt, totdat elk deel een keer is gebruikt om op te testen.

Tabel 5: Verdeling van de dataset voor evaluatie.

Politici	trainingset	testset
Buma	606	58
Dijkhoff	1197	134
Klaver	2085	241
Marijnissen	666	71
Pechtold	3814	444
Wilders	4900	526
Totaal	13268	1474

4.4 OPTIMALISATIE

Voor dit onderzoek worden onder andere de functies SGDClassifier[4] en TFIDFVectorizer[5] en LinearSVC[6] gebruikt uit de Scikit-Learn module voor de programmeertaal Python.

Om de juiste parameters te vinden voor de dataset is hier getraind op een developmentset van 2000 volgers. Hier is voor gekozen, omdat voor het optimaliseren van de parameters de gehele trainingset erg groot is. Het duurt uren voordat het programma klaar is en er resultaten te zien zijn.

4.4.1 N-grammen

De functie TFIDFVectorizer heeft een parameter genaamd ngram_range waar onder andere de kenmerken unigram, bigram en trigram meegegeven kunnen worden. Om te bepalen wat de meest optimale instelling is voor de dataset is hier getraind en getest met de SGD-Classifier en een developmentset van de gehele dataset. Alhoewel de scores erg dicht tegen elkaar aanliggen, blijkt uit het optimalisatieproces dat unigrammen een betere F1-score geven (0.51), dan de andere kenmerken. De scores voor precision, recall en f1-score zijn een gemiddelde van alle labels in de dataset per kenmerk. (zie tabel 6).

Tabel 6: Resultaat van het optimaliseren van ngram_range.

Woord n-gram	precision	recall	f1-score
unigrammen	0.56	0.54	**0.51**
bigrammen	0.55	0.53	0.45
trigrammen	0.46	0.54	0.48
unigrammen en bigrammen	0.56	0.54	0.49
bigrammen en trigrammen	0.55	0.54	0.48
unigrammen, bigrammen en trigrammen	0.55	0.50	0.45

[4] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier

[5] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[6] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

4.4.2 Parameters classifiers

De SGD-Classifier en LinearSVC kennen verschillende parameters. Eén daarvan is class_weight. De standaardinstelling van deze parameter is None. Class_weight heeft echter ook een optie genaamd "balanced". Deze optie past automatisch het gewicht van elke klasse aan die omgekeerd evenredig is aan de klassefrequenties in de dataset. Omdat de dataset niet evenveel volgers per politicus bevat, wordt hierdoor toch een balans in de dataset aangebracht. De rest van de SGD-Classifier en de LinearSVC zijn ingesteld op de standaardwaarden zoals in versie 0.19 van Scikit-Learn.

5 RESULTATEN

Dit hoofdstuk laat de resultaten zien. De resultaten zijn gevonden door middel van de in hoofdstuk 4 beschreven methode.

5.1 BASELINE

Met de functie DummyClassifier van Scikit-Learn is de baseline bepaald. Op de gehele trainingsset en testset gaf deze classifier een f1-score van 0.26 zoals te zien in tabel 7. Onderaan de tabel is de gemiddelde score te zien van de precision, recall en f1-score van alle labels (politici) bij elkaar. Onder de kolom 'support' is het aantal volgers in totaal af te lezen voor deze politicus. Hier staat onderaan het totale aantal van de unieke volgers die in de dataset zitten.

Tabel 7: Resultaat baseline 10 fold Cross-Validation met DummyClassifier van Scikit-Learn.

	precision	recall	f1-score	support
Buma	0.04	0.04	0.04	606
Dijkhoff	0.08	0.07	0.07	1197
Klaver	0.17	0.18	0.17	2085
Marijnissen	0.04	0.04	0.04	666
Pechtold	0.29	0.29	0.29	3814
Wilders	0.36	0.36	0.36	4900
avg / total	0.26	0.26	0.26	13268

5.2 RESULTAAT SGD-CLASSIFIER

5.2.1 Unigrammen

Tabel 8 hieronder laat de resultaten zien van de SGD-Classifier met als parameter ngram_range=[1,1]. Onderaan de tabel is de gemiddelde score te vinden van de precision, recall en f1-score van alle labels (politici) bij elkaar. Onder de kolom 'support' is het aantal volgers af te lezen voor elke politicus en onderaan het totale aantal volgers in de testset van 10% van de data. Hieruit blijkt dat de gemiddelde f1-score van alle labels 0.46 is. Dit is beter dan de baseline.

Tabel 8: Resultaat unigrammen met de SGD-Classifier.

	precision	recall	f1-score	support
Buma	0.18	0.48	0.26	58
Dijkhoff	0.30	0.24	0.27	134
Klaver	0.54	0.35	0.42	241
Marijnissen	0.52	0.24	0.33	71
Pechtold	0.54	0.23	0.32	444
Wilders	0.55	0.87	0.68	526
avg / total	0.51	0.49	0.46	1474

5.2.2 Unigrammen en bigrammen

In tabel 9 zijn de resultaten van de SGD-Classifier te vinden met ngram_range=[1,2]. De f1-score is hier 0.47, wat 1% beter is dan de classifier met alleen unigrammen.

Tabel 9: Resultaat unigrammen en bigrammen met de SGD-Classifier.

	precision	recall	f1-score	support
Buma	0.15	0.55	0.24	58
Dijkhoff	0.30	0.22	0.25	134
Klaver	0.53	0.36	0.43	241
Marijnissen	0.50	0.24	0.32	71
Pechtold	0.53	0.26	0.35	444
Wilders	0.60	0.85	0.70	526
avg / total	0.52	0.49	0.47	1474

5.2.3 Unigrammen, bigrammen en trigrammen

Onderstaande tabel 10 laat het resultaat zien van ngram_range=[1,3]. Van alle resultaten voor de SGD-Classifier is deze combinatie van n-grammen het beste met een F1-score van 0.49.

Tabel 10: Resultaat unigrammen, bigrammen en trigrammen met de SGD-Classifier.

	precision	recall	f1-score	support
Buma	0.26	0.50	0.35	58
Dijkhoff	0.44	0.23	0.30	134
Klaver	0.39	0.57	0.46	241
Marijnissen	0.23	0.49	0.31	71
Pechtold	0.56	0.22	0.32	444
Wilders	0.69	0.81	0.75	526
avg / total	0.54	0.51	0.49	1474

5.3 RESULTAAT LINEARSVC

5.3.1 Unigrammen

Onderstaande tabel 11 laat het resultaat zien van ngram_range=[1,1] van de classifier LinearSVC. Deze classifier geeft een gemiddelde f1-score van 0.55. Dit is 6% hoger dan de SGD-Classifier.

Tabel 11: Resultaat unigrammen met de LinearSVC classifier.

	precision	recall	f1-score	support
Buma	0.39	0.40	0.39	58
Dijkhoff	0.32	0.22	0.26	134
Klaver	0.47	0.47	0.44	241
Marijnissen	0.41	0.37	0.39	71
Pechtold	0.52	0.49	0.51	444
Wilders	0.69	0.78	0.73	526
avg / total	0.54	0.56	0.55	1474

5.3.2 Unigrammen, bigrammen

De tabel 12 laat het resultaat zien van ngram_range=[1,2] van de classifier LinearSVC. Deze classifier geeft een gemiddelde f1-score van 0.56.

Tabel 12: Resultaat unigrammen en bigrammen met de LinearSVC classifier.

	precision	recall	f1-score	support
Buma	0.47	0.38	0.42	58
Dijkhoff	0.36	0.19	0.25	134
Klaver	0.48	0.45	0.46	241
Marijnissen	0.52	0.31	0.39	71
Pechtold	0.52	0.56	0.54	444
Wilders	0.68	0.79	0.73	526
avg / total	0.54	0.57	0.56	1474

5.3.3 Unigrammen, bigrammen en trigrammen

Het resultaat van de ngram_range=[1,3] van de LinearSVC classifier is te zien in tabel 13. De gemiddelde f1-score van alle labels is hier 0.55, wat iets slechter is dan in de combinatie unigrammen en bigrammen.

Tabel 13: Resultaat unigrammen, bigrammen en trigrammen met de LinearSVC classifier.

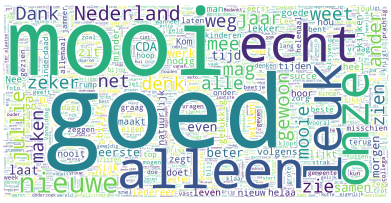
	precision	recall	f1-score	support
Buma	0.49	0.29	0.37	58
Dijkhoff	0.42	0.16	0.23	134
Klaver	0.49	0.44	0.46	241
Marijnissen	0.62	0.25	0.36	71
Pechtold	0.51	0.57	0.54	444
Wilders	0.67	0.81	0.73	526
avg / total	0.56	0.57	0.55	1474

Tabel 14: F1-scores van de kenmerken met de LinearSVC classifier en de SGD-Classifier.

N-grammen	LinearSVC	SGD-Classifier
unigrammen	0.55	0.46
unigrammen en bigrammen	**0.56**	0.47
unigrammen, bigrammen en trigrammen	0.55	0.49

5.4 DISCUSSIE

Alhoewel de resultaten beter zijn dan de baseline, komt de f1-score niet boven de 56% uit met de LinearSVC classifier op basis van unigrammen en bigrammen (zie tabel 14). Wat wel opvalt zijn de onderlinge verschillen tussen de politici. Het voorspellen van volgers bij Geert Wilders gaat de classifier aardig goed af. Bij de unigrammen, bigrammen en trigrammen combinatie met de SGD-Classifier is de f1-score het hoogst, namelijk 0,75. Het kan zijn dat in dit geval meer volgers in de dataset met meer tweets resulteert in een betere uitkomst. Daarnaast kan het ook zo zijn dat de doelgroep van Wilders veel specifiekere woorden gebruikt. Het specifiekere gebruik van woorden is duidelijker te zien bij bijvoorbeeld Buma, waar de minste volgers van in de dataset zitten, maar deze in vergelijking met de rest toch aardig scoort. De doelgroep van het CDA is christelijk, meestal wat ouder en gerichter op bijvoorbeeld gezin. Dit is goed te zien in de wordcloud gegeneerd met de belangrijkste woorden in de tweets van de volgers van Buma (zie figuur 3). Woorden als "CDA, gemeente, toekomst, werk, zorg, politiek, kabinet, stemmen, Nederland, Europa, actie, vrijwilliger, samen, liefde, aandacht, kinderen, vrouw, pa en gelukkig", zijn hier te vinden. In de wordcloud van Klaas Dijkhoff figuur 4, met meer volgers in de dataset, maar een veel slechtere score, zitten veel meer algemenere woorden. "Auto, Amsterdam, goed, mooi, alleen, leuk, echt, foto, haha, hahaha, VVD, weet, net, man en volgens", komen in deze wordcloud voor.



Figuur 3: Wordcloud van woorden die het meest voorkomen in de tweets van de volgers van Sybrand Buma

Figuur 4: Wordcloud van woorden die het meest voorkomen in de tweets van de volgers van Klaas Dijkhoff

Veel woorden uit de verschillende wordclouds komen overeen. Dit maakt het voor de classifier ook moeilijk om te kunnen bepalen welke volger bij welke politici hoort. Voor toekomstig onderzoek zou dit kunnen worden opgelost door bijvoorbeeld stopwoorden te filteren. Daarnaast zou een functie die uit werkwoorden de stam haalt de resultaten kunnen verbeteren.

Ook zouden de resultaten kunnen verbeteren wanneer data wordt geselecteerd van leiders met ongeveer evenveel volgers met ongeveer evenveel tweets. Hierdoor is er een betere balans in de data. Voor dit onderzoek is hier niet naar gekeken, omdat dit in werkelijkheid ook niet het geval is. De volgers zelf zouden ook nog beter geselecteerd kunnen worden. Er is hier bijvoorbeeld niet gekeken naar profielinformatie.

6 CONCLUSIE

In deze scriptie is onderzocht of er te voorspellen valt of een gebruiker op Twitter een bepaalde politieke leider volgt. Dit is gedaan door middel van een Stochastic Gradient Descent Classifier die is ingesteld als een lineaire Support Vector Machine en een LinearSVC classifier.

De hoofdvraag van dit onderzoek was: *Valt er te voorspellen middels het woordgebruik in tweets of een gebruiker op Twitter een bepaalde politieke leider volgt?*

In de hypothese stel ik dat ik verwacht dat er een overeenkomst is in het woordgebruik tussen de volgers van een groep, en dat dus te voorspellen valt of de gebruiker een bepaalde politieke leider volgt. Uit het onderzoek blijkt dat dit niet het geval is. Met een baseline van 0.26 en een f1-score van 0.56 is het algoritme weliswaar verbeterd, maar nog niet goed genoeg. In dit geval verkreeg de LinearSVC classifier betere resultaten dan de SGD-Classifier.

Van bepaalde groepen volgers is het label beter te voorspellen. Dit kan te maken hebben met een specifiekere doelgroep of met het feit dat er meer data beschikbaar is in sommige groepen.

BIBLIOGRAFIE

Barberá, P. (2013). Birds of the same feather tweet together. bayesian ideal point estimation using twitter data.

Bouma, G. (2015). N-grams frequencies for dutch twitter data. *CLIN Journal 5*.

Conover, M. D., B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer (2011). Predicting the political alignment of twitter users.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology 27*(1), 415–444.

Pamelee, J. H. and S. L. Bichard (2011). *Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public*. Lexington Books.

Park, C. S. (2013). Does twitter motivate involvement in politics? tweeting, opinion leadership, and political engagement. *Computers in Human Behavior 29*(4), 1641–1648.

Tjong Kim Sang, E. (2011). Het gebruik van twitter voor taalkundig onderzoek. *TABU: Bulletin voor Taalwetenschap 39*(1/2), 62–72.

Tjong Kim Sang, E. and J. Bos (2012). Predicting the 2011 dutch senate election results with twitter.

Tumasjan, A., T. Sprenger, P. Sandner, and I. Welpe (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment.