Text Mining in Facebookberichten

Voorspellen van het geslacht van de schrijver

Corinca Zwijsen

332435

Bachelorscriptie Econometrie & Operationele Research

Marketing

Erasmus Universiteit Rotterdam

Abstract

In dit artikel zal ik door middel van een model voor logistische regressie onderzoeken of het mogelijk is om het geslacht van een persoon te voorspellen aan de hand van een bericht dat deze persoon geschreven heeft.. Ik zal de data op twee manieren gebruiken, en met verschillende methodes onderzoeken welke woorden of categorieën het nauwkeurigst het geslacht van de schrijver voorspellen. Verder zal ik onderzoeken welke woorden of categorieën significant vaker door een bepaald geslacht gebruikt worden.

1. Inhoudsopgave

2. Inleiding

Dat mannen en vrouwen verschillen, is een feit. De meeste mensen denken dan aan eigenschappen als uiterlijk en gedrag, maar ook verschil in DNA en biologische kenmerken. Echter, dat ook het taalgebruik verschilt is een verschijnsel waar minder snel aan gedacht wordt. Dit verschil is door vele onderzoeken aangetoond. Keune (2012) concludeerde dat mannen een meer 'informatieve' gespreksstijl hebben en vrouwen zich meer op een 'betrokken' manier uitdrukken. Deze betrokken stijl kenmerkt zich door het veelvuldig gebruik van werkwoorden. Mannen gebruiken daarentegen meer zelfstandig naamwoorden.

Volgens Trudgill (1972) is het verschil in taal deels psychologisch te verklaren. Vrouwen praten het liefst de - nette - standaardtaal om hun netwerk, dat wil zeggen, de mensen met wie ze sociale conversaties voeren, uit te breiden. Mannen daarentegen streven naar een speciale taal door niet-standaard vormen van woorden te gebruiken om te laten zien dat ze stoer zijn. Deuchar (1989) heeft een ander idee over het verschil van het taalgebruik. Zij is van mening dat vrouwen deze standaardvorm van woorden niet gebruiken om hun netwerk uit te breiden, zoals Trudgill beweerde, maar meer om het netwerk wat ze hebben te behouden. Het kon zo zijn, dat als een vrouw woorden gebruikte die niet standaard waren, en deze woorden werden in twijfel gebracht of niet herkend, dan zou de consequentie kunnen zijn dat ze contacten zou verliezen.

Al deze onderzoeken wijzen er dus op dat er duidelijk verschil is in het taalgebruik tussen mannen en vrouwen. Bamman et al (2012) onderzochten of dit verschil ook te vinden was in geschreven berichten. Zij maakten gebruik van het Social Media netwerk Twitter. Uit hun onderzoek bleek dat ook hier een verschil gevonden kon worden. Woorden werden opgedeeld in categorieën zoals bijvoorbeeld zelfstandig naamwoorden of afkortingen, en uit het onderzoek bleek dat mannen en vrouwen verschillen in het gebruik van deze categorieën. De categorieën kunnen als het ware worden geplaatst bij mannen en vrouwen, aan de hand van mate van gebruik.

Is het andersom ook zo? Als een bepaalde tekst wordt geanalyseerd, kan dan voorspeld worden of een man of vrouw deze tekst heeft geschreven?

Is het geslacht van de schrijver van een bericht te voorspellen aan de hand van zijn of haar schrijfstijl?

Wat als vraag, volgend op deze onderzoeksvraag zou kunnen zijn is of er met de mogelijke resultaten ook wat gedaan kan worden. Is het te gebruiken voor marketing doeleinden, die zich specifiek op bepaalde doelgroepen richten?

In dit onderzoek zal ik gebruik maken van een dataset met berichten van het veelgebruikte Social Media netwerk Facebook. Door deze bron te gebruiken is bij voorbaat al bekend of de schrijver man, dan wel vrouw is. De dataset bevat alleen maar Engelstalige berichten.

3. Literatuur

Om al bekende informatie over het onderwerp te vinden heb ik van enkele artikelen gebruik gemaakt.

Gender in Twitter – Styles, Stances and Social Networks (Bamman et al, 2012)

Het artikel van Bamman et al is vergelijkbaar met dit onderzoek. Omdat het onderzoek uit het artikel niet op dezelfde manier is uitgevoerd (hier is ook gekeken naar categorieën, maar slechts deels dezelfde categorieën als die in dit onderzoek gebruikt worden, en de bron van de gegevens is Twitter) maar wel erg lijkt op dit onderzoek, is het een goed naslagwerk. Uit het artikel kan opgemaakt worden dat er wel degelijk verschillende woordkeuzes zijn tussen man en vrouw. Een groot deel van de categorieën werd significant vaker door mannen dan wel vrouwen gebruikt. In dit onderzoek kunnen we eerdere resultaten wellicht reproduceren.

Explaining register and sociolinguistic variation in the lexicon: Corpus studies in Dutch (Keune, 2012)

Het artikel van Keune is compleet gericht op het verschil in taalgebruik. Het onderzoek is in Nederland gehouden, dus bepaalde taal-specifieke delen van het onderzoek (zoals welke woorden die eindigen op –lijk gebruiken mannen meer, en welke vormen gebruiken vrouwen meer) zijn niet bruikbaar. De achterliggende gedachte is daarentegen wel erg interessant, en is mogelijk een goede hulp bij dit onderzoek. Ook in het artikel is geconstateerd dat mannen en vrouwen een verschillende spreek- dan wel schrijfstijl hebben.

Mining the Blogosphere: Age, gender, and the varieties of self-expression (Argamon et al, 2007)

In het artikel van Argamon et al is een voorspellingsmodel gebouwd met behulp van bijna 20.000 Engelse blogs (welke ruim 680.000 berichten, en daarmee 140 miljoen woorden bevatte). Hier werden de 1000 woorden die de meeste informatie bevatten in het model opgenomen, en het model had een nauwkeurigheid van ruim 80%, wat betekent dat van de groepen die voorspeld werden gemiddeld 80% goed werd voorspeld. Ze hebben de 1000 meest gebruikte woorden genomen, en deze bleken

ongeveer 80% van alle voorkomende woorden te bevatten. Het onderzoek maakt gebruik van veel dezelfde categorieën die ook in dit onderzoek gebruikt zullen worden. Het kan dus goed vergelijkingsmateriaal zijn. De resultaten bij het onderzoeken van de categorieën die ook in dit onderzoek gebruikt worden zijn dat voorzetsels significant meer door mannen worden gebruikt, en dat voegwoorden en hulpwerkwoorden (in dit onderzoek hebben wij dan wel alleen werkwoorden) significant meer door vrouwen worden gebruikt.

4. Data

4.1 Uitleg

De data heb ik ontvangen als een Excel bestand. Het is oorspronkelijk verworven van Facebook. Het Excel bestand bestaat uit twee gegevenssets.

Op de eerste pagina van het bestand staat de eerste kolom voor het unieke nummer wat aan een bericht is toegekend. Dit zijn 4400 berichten in totaal. De tweede kolom geeft aan wat het geslacht van de schrijver is, en de rest van de kolommen zijn woorden die in alle berichten bij elkaar het meest zijn voorgekomen. Dit zijn er 295 in totaal. Het is niet bekend waarom er voor dit aantal is gekozen.

De eerste pagina heb ik opgedeeld in een vector y en een matrix X.

De vector y is een kolom met 0/1 waarden. Als bericht *i* geschreven is door een man, dan staat op plek y_i de waarde 1. Is dit bericht door een vrouw geschreven, dan staat op plek y_i de waarde 0.

De matrix X is de matrix met de verklarende variabelen. De rijen stellen de berichten voor. In de kolommen staan getallen. Als een woord niet in het bericht is voorgekomen staat er een 0 op de plek. Als een woord 2 keer is voorgekomen staat er een 2 op die plek.

Als woord *j* in bericht *i* *k* keer is voorgekomen, dan staat op plek $X_{i,j}$ het getal *k*.

De tweede pagina van het bestand bevat de data zoals deze is opgehaald uit Facebook. Ook hier staat aan het begin over welk bericht het gaat. Het betreft de originele berichten, met daarbij de afzender, de informatie wat voor bericht het was (status, foto, video). Uit deze informatie kan opgemaakt worden uit hoeveel woorden elk bericht in totaal bestaat.

4.1.1 Korte statistieken

De verdeling mannen/vrouwen is redelijk gelijk verdeeld (2069 mannen en 2263 vrouwen, 47.7% om 52.3%. Deze percentages zijn van de berichten waarvan duidelijk is of het van een man of vrouw afkomstig is.) Van 68 berichten is het geslacht onbekend.

Voor het berekenen van het gemiddelde heb ik alleen de berichten meegenomen waar minimaal één veelgebruikt woord in voorkomt. Het is dus het gemiddelde aantal woorden in een bericht, gegeven dat er tekst in het bericht staat.

Een bericht bestaat uit gemiddeld 38.9 woorden, waarvan gemiddeld 16.2 woorden deel van de 295 meest gebruikte woorden zijn.

Mannen gebruiken gemiddeld meer woorden dan vrouwen (42.5 versus 35.7), en ook meer veelgebruikte woorden (gemiddeld 16.6 versus gemiddeld 15.8).

Er zijn een paar berichten die dit gemiddelde hoog hebben gemaakt. Zo zijn er berichten geplaatst waarin een complete Wikipedia pagina is weergegeven, of een uitgebreid verhaal of gedicht werd verteld. Om deze reden heb ik ook de mediaan van de berichten berekend. Ook hier heb ik alleen de berichten in de berekening meegenomen waarvan gegeven is dat er veelgebruikte woorden in voorkomen.

De mediaan van het aantal woorden in een bericht is 19. De mediaan van het aantal veelgebruikte woorden in een bericht is 8.

De mediaan van het aantal woorden in een bericht van een man is 19, de mediaan van het aantal veelgebruikte woorden in een bericht van een man is 7.

De mediaan van het aantal woorden in een bericht van een vrouw is 20, de mediaan van het aantal veelgebruikte woorden in een bericht van een vrouw is 10.

Van de 4332 overgebleven berichten waarvan duidelijk is of het van een man of vrouw afkomstig is, zijn 215 berichten leeg. Deze bevatten geen tekst. Daarnaast zijn er 73 berichten die wel tekst bevatten, maar geen enkel veelgebruikt woord.

4.2 Opmerkingen over de data:

Bij het ontvangen van de data waren er enkele punten die nog verbeterd moesten worden voordat er mee gewerkt kon worden. Allereerst waren er 68 berichten waarvan niet bekend was of de schrijver man of vrouw is. Omdat het onderzoek hoofdzakelijk over het geslacht gaat zijn deze berichten dus niet bruikbaar. Verder stonden er berichten in het bestand waar geen enkel woord gebruikt werd. Deze konden dus ook niet gebruikt worden.

De data zijn op één dag verworven. Dit kan ervoor zorgen dat het resultaat niet representatief is voor in het algemeen. Op de dag van de werving was er onder andere een voetbalwedstrijd tussen Manchester City en Manchester United. De woorden 'city' en 'united' zijn dus vaker voorgekomen dan misschien verwacht zou worden op een willekeurige andere dag.

Er is bij het aanleveren van de data niet duidelijk gebleken of het een willekeurige steekproef is van berichten op Facebook. Het kan zijn dat het allemaal berichten zijn binnen het netwerk van de persoon die de data heeft verworven. Als dit zo is, dan zijn de data ook niet representatief. Het is mogelijk dat het netwerk van de verwerver een groep is uit een bepaald milieu of van een bepaald intellectueel niveau. Deze factoren kunnen ook voor een ander soort berichten zorgen.

De twee sheets komen niet goed overeen met elkaar. Een aantal berichten die op de eerste pagina voorkomen komen niet voor op de tweede pagina en andersom. Dit maakt het lastig om de berichten aan elkaar te koppelen. Er kan dus niet van uit gegaan worden dat bericht *i* op pagina 1 hoort bij bericht *i* op pagina 2. Dit zou als consequentie hebben dat de data compleet handmatig aangepast moet worden. Dit is te veel werk, dus een paar onderzoeken kunnen niet worden gedaan - zoals met relatieve waarden werken - en andere onderzoeken kunnen resultaten geven die niet compleet betrouwbaar zijn. Voor verder onderzoek zou ik dus aanraden om de data of op een andere manier te verzamelen, of om ruim de tijd te nemen om de data kloppend te maken.

Ik heb geprobeerd uit te zoeken of alle woorden die volgens de X matrix in een bepaald bericht zouden voorkomen ook daadwerkelijk in de berichttekst terug te vinden zijn. Voor ongeveer 800 berichten geldt dat een aantal van de in X getelde woorden niet kunnen worden gevonden in het oorspronkelijke bericht. Het verschil per bericht tussen het beweerde aantal en het daadwerkelijk gevonden aantal varieert van 0 tot 66 woorden.
Door de wijze van zoeken van een woord in de berichttekst is hier een bepaalde onzuiverheid ontstaan. Bijvoorbeeld als het woord ''day'' is gezocht in de tekst 'from today we are ...', Dan is dit woord gevonden terwijl het er niet in stond. Deze onzuiverheid heeft tot gevolg dat het aantal van 800 in werkelijkheid hoger kan liggen. Een andere onzuiverheid is dat in X het woord '' can't '' is opgenomen als "cant". Indien gezocht wordt naar het laatste woord dan kan het zijn dat het niet

gevonden wordt. Deze onzuiverheid kan tot gevolg hebben dat aantal van 800 in werkelijkheid lager kan liggen.

4.3 Gebruik van de data

Er zijn meerdere manieren om de data te gebruiken in regressies. Het is interessant om naar al deze modellen te kijken, mogelijk zijn er nog verschillen in uitkomsten.

De meest voor de hand liggende regressie is die met de tellingen van de meest gebruikte woorden. Hiermee kan onderzocht worden of er specifieke woorden zijn die vaker door mannen of door vrouwen gebruikt worden. Deze tellingen zijn per woord in een bericht hoe vaak dit woord is voorgekomen.

De woorden kunnen ook gecategoriseerd worden. Er worden in de engelse taal 8 categorieën onderscheiden waarin een woord ingedeeld kan worden. Hiermee kan onderzocht worden of bepaalde categorieën vaker door mannen of vrouwen gebruikt worden. Woorden kunnen soms in twee categorieën horen. Omdat het per bericht bekijken waar een woord hoort erg veel werk is, neem ik aan dat de kans dat het in één van deze categorieën hoort 50% is, en deel deze woorden dus voor de helft bij de ene categorie, en voor de helft bij de andere categorie in. De mogelijke categorieën zijn:

- Zelfstandig Naamwoord (Noun)
- Lidwoord (Determiner)
- Voornaamwoord (Pronoun)
- Bijvoeglijk Naamwoord (Adjective)
- Werkwoord (Verb)
- Bijwoord (Adverb)
- Voegwoord (Conjunction)
- Voorzetsel (Preposition).

Deze lijst met categorieën is afkomstig van Wikipedia.

5. Model

Om voorspellingen te maken en tot conclusies te kunnen komen is er een keuze voor een model nodig. Er zijn verschillende modellen die gebruikt kunnen worden. Het ene model zal niet altijd even net en efficiënt werken als een ander. Naar aanleiding van het gebruik van het binaire logit model wat Bamman et al (2012) heeft gebruikt, is voor dit onderzoek de keuze gemaakt om op dezelfde wijze te onderzoeken of het model ook op deze dataset goede resultaten levert.

5.1 Logistische Regressie

Dit model richt zich specifiek op 0/1 data als te verklaren variabele, wat op de data van dit onderzoek van toepassing is. Het logistische model ziet er als volgt uit:

$$P(y = 1|X) = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \tag{5.1}$$

Hierbij is X de matrix met verklarende variabelen voor de uitkomst van y. In een logistisch model kan de waarde van y alleen een 0 of een 1 zijn. Voor X zijn geen specifieke eisen.

Het model is een functie om de kans op een waarde y=1 te vinden. De Griekse letter β is de vector van coëfficiënten die bij de kolommen van de matrix X horen. De eerste waarde van β hoort bij de eerste kolom van X, de tweede waarde van β bij de tweede kolom van X, etc.

De β wordt geschat door middel van Maximum Likelihood. Het berekenen van de Maximum Likelihood gaat in enkele stappen.

Als de kans op 'succes' gelijk is voor alle observaties, dan geldt

$$P(y_i = 1) = p \tag{5.2}$$

voor elke observatie. In dat geval is de kansverdeling van de *i*-de observatie gegeven door

$$p(y_i) = p^{y_i}(1 - p)^{1-y_i} \tag{5.3}$$

In dit onderzoek volgt de variabele y_i de Bernouilli verdeling met kans p_i, waarbij

$$p_i = P(y_i = 1) = F(x_i'\beta) \tag{5.4}$$

de kans is op de uitkomst y_i=1, en (1-p_i) de kans is op de uitkomst y_i=0.

De kansverdeling is dan gegeven door

$$p(y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \tag{5.5}$$

Als de waarnemingen onderling onafhankelijk zijn, dan is de likelihood functie gegeven door:

$$L(p) = \prod_{i=1}^{n} p_i\,^{y_i}(1-p_i)^{1-y_i} \tag{5.6}$$

met y_i=0,1. De log-likelihood is daardoor gelijk aan:

$$\begin{aligned}\log\big(L(\beta)\big) &= \sum_{i=1}^{n} y_i \log(p_i) + \sum_{i=1}^{n}(1-y_i)\log(1-p_i)\\ &= \sum_{i=1}^{n} y_i \log\left(F\left(x_i'\beta\right)\right) + \sum_{i=1}^{n}(1-y_i)\log\left(1-F\left(x_i'\beta\right)\right)\\ &= \sum_{\{i;y_i=1\}} \log\left(F\left(x_i'\beta\right)\right) + \sum_{\{i;y_i=0\}} \log\left(1-F\left(x_i'\beta\right)\right)\end{aligned} \tag{5.7}$$

Om het maximum te vinden, stellen we de *k* eerste orde condities op

$$\begin{aligned}g(\beta) &= \frac{\partial \log(L)}{\partial \beta}\\ &= \sum_{i=1}^{n}\frac{y_i}{p_i}\frac{\partial p_i}{\partial \beta} + \sum_{i=1}^{n}\frac{1-y_i}{1-p_i}\frac{\partial(1-p_i)}{\partial \beta}\\ &= \sum_{i=1}^{n}\frac{y_i}{p_i}f_i x_i - \sum_{i=1}^{n}\frac{(1-y_i)}{1-p_i}f_i x_i\\ &= \sum_{i=1}^{n}\frac{y_i-p_i}{p_i(1-p_i)}f_i x_i = 0\end{aligned} \tag{5.8}$$

met f_i = f(x_i'β) de afgeleide van F(x_i'β). Deze *k* eerste orde condities kunnen numeriek worden opgelost, met bijvoorbeeld Newton-Raphson. Het resultaat van deze berekening is de geschatte β.

5.1.1 In dit onderzoek

In dit onderzoek is de vector y dus een 0/1 variabele. Bij de waarde 1 hoort het mannelijk geslacht, bij de waarde 0 het vrouwelijk geslacht. De matrix X verschilt in de onderzoeken. Hieronder per onderzoek de specificatie van X.

Tellingen van de woorden

Als er per woord naar de telling wordt gekeken, dan is X de matrix met het aantal keer dat een woord is voorgekomen.

> **Voor bericht *i*: als woord *j* in bericht *i* *k* keer is voorgekomen, dan staat op plek $X_{i,j}$ het getal *k*.**

Hierbij zal de β uit (5.1) een 296x1 vector zijn, met voor elk van de 295 woorden een coëfficiënt, en één voor de constante in het model.

Woordcategorieën

Als er naar de woord categorieën wordt gekeken, is X de matrix met het aantal keer dat een categorie is voorgekomen.

> **Voor bericht *i*: als categorie *j* in bericht *i* *k* keer is voorgekomen, dan staat op plek $X_{i,j}$ het getal *k*.**

Hierbij zal de β uit (5.1) een 9x1 vector zijn, met voor elk van de 8 categorieën een coëfficiënt, en één voor de constante in het model.

5.2 Voorspellen

Om de precisie te beoordelen gebruiken we 10-voudige kruisvalidatie. De data zullen worden opgedeeld in 10 willekeurige groepen, elk van deze groepen zal los van de rest voorspeld worden met de coëfficiënten die zijn geschat met de informatie over de overige 9 groepen. De groepen zijn dan wel willekeurig samengesteld, de verhouding man/vrouw moet wel redelijk gelijk zijn. Bij deze data is dat goed mogelijk, omdat de set zelf ook redelijk gelijk is verdeeld over man en vrouw.

Als de coëfficiënten van een groep geschat zijn, zal er een selectie plaatsvinden op basis van significantie van de coëfficiënten. Het model zal opnieuw geschat worden met een gereduceerd aantal variabelen, en op deze manier zal er hopelijk een beter model ontstaan waardoor de voorspellingen beter zijn.

Ook zal over de complete dataset worden geschat. Die bevat de meeste berichten, en vaak geldt dat als een dataset groter is, hij beter en nauwkeuriger geschat wordt. Het geschatte model zal over de complete set, en over delen ervan, toegepast worden om de kwaliteit te testen.

Doordat er verschillende groepen zijn en hierdoor verschillende schattingen gemaakt worden, kan de kwaliteit van zo'n model beter gezien worden. Het kan zo zijn dat als alleen de laatste berichten worden voorspeld dit geen representatieve uitkomst geeft. Mogelijk zijn deze berichten geen goede representatie van de werkelijke berichten die later met dit model geschat moeten worden.

Als de overgebleven coëfficiënten geschat zijn, dan hebben we de β gereduceerd tot een bepaalde grootte. Deze vector zullen we gebruiken in het model. Doordat niet alle woorden of categorieën in dit model worden gebruikt zal de X matrix ook kolommen verliezen. De nieuwe vector β en matrix X worden in (5.1) gebruikt om de kans op een man of vrouw te berekenen.

6. Resultaten

Omdat er op verschillende manieren geschat is met verschillende voorspellings-modellen, zijn er meerdere resultaten.

6. 1 Tellingen van de woorden

Hoewel er met 10-voudige kruisvalidatie is geschat, zijn de resultaten erg teleurstellend. De percentages van het aantal berichten die goed zijn voorspeld liggen tussen de 55% en 62%. Wel zijn er een aantal woorden waar de coëfficiënten significant verschillen van 0. In Tabel 1 zijn alle woorden die significant verschilden van 0 op een 5% significantieniveau opgenomen. Deze coëfficiënten zijn gevonden uit de regressie over de complete dataset.

Man	Coëfficiënt	Vrouw	Coëfficiënt	Vrouw	Coëfficient
After	0.55	Bed	-0.74	School	-0.62
Again	0.71	Bless	-0.49	Should	-0.58
Already	0.83	Boy	-0.53	Sleep	-0.64
Come	1.56	Bring	-0.76	Taking	-0.85
For	0.95	Church	-0.81	Than	-0.62
Fun	0.50	Could	-0.84	Their	-0.49
Girl	1.04	Does	-0.89	They	-0.72
Great	0.48	Face	-1.27	Think	-0.39
Hard	0.95	Feeling	-0.50	Wish	-0.90
Man	0.44	Friends	-1.10	Work	-0.44
Many	0.56	Gonna	-0.28	Worship	-0.84
Now	0.63	Got	-0.29		
Our	0.32	Have	-0.32		
Own	0.77	Home	-0.55		
Said	0.70	Lord	-0.37		
The	0.20	Love	-0.42		
Todays	0.51	Much	-0.45		
Tomorrow	0.49	Never	-0.55		
United	1.49	Off	-0.53		
When	0.39	Open	-0.65		
Why	0.91	Pray	-0.93		

Tabel 1 - Per geslacht staat aangegeven welke woorden significant een grotere kans geven dat het bericht door een man dan wel een vrouw geschreven is. In de tweede, vierde en zesde kolom staat de waarde van de coëfficiënt van het woord in de kolom links ervan. Deze coëfficiënten zijn de berekende waarden van een regressie over de complete data. Hoe groter de absolute waarde van een coëfficiënt wordt, des te meer invloed heeft het bijbehorende woord op de kans dat het bericht van een bepaald geslacht afkomstig is.

Wat opvallend is, is dat als naar de woordcategorieën van de woorden in

Man	Coëfficiënt	Vrouw	Coëfficiënt	Vrouw	Coëfficient
After	0.55	Bed	-0.74	School	-0.62
Again	0.71	Bless	-0.49	Should	-0.58
Already	0.83	Boy	-0.53	Sleep	-0.64
Come	1.56	Bring	-0.76	Taking	-0.85
For	0.95	Church	-0.81	Than	-0.62
Fun	0.50	Could	-0.84	Their	-0.49
Girl	1.04	Does	-0.89	They	-0.72
Great	0.48	Face	-1.27	Think	-0.39
Hard	0.95	Feeling	-0.50	Wish	-0.90
Man	0.44	Friends	-1.10	Work	-0.44
Many	0.56	Gonna	-0.28	Worship	-0.84
Now	0.63	Got	-0.29		
Our	0.32	Have	-0.32		
Own	0.77	Home	-0.55		
Said	0.70	Lord	-0.37		
The	0.20	Love	-0.42		
Todays	0.51	Much	-0.45		
Tomorrow	0.49	Never	-0.55		
United	1.49	Off	-0.53		
When	0.39	Open	-0.65		
Why	0.91	Pray	-0.93		

Tabel 1 wordt gekeken, er slechts 6 woorden zijn in het rijtje van de vrouwen welke niet (onder anderen) in te delen zijn als Zelfstandig Naamwoord of Werkwoord. Bij de mannen komt elke categorie minimaal één keer voor, en alle categorieën zijn redelijk gelijk verdeeld.

Niet in elke kruisvalidatie waren deze coëfficiënten significant verschillend van 0. Dit geeft aan dat er nog wel variatie zit binnen de groepen. Enkele coëfficiënten kwamen in elke regressie er uit, maar dit was voor lang niet elk woord het geval.

De coëfficiënt van de constante was niet in alle gevallen significant verschillend van 0. In de modellen waar dit wel zo was, had de coëfficiënt een positief teken. Dit betekent dat berichten zonder woorden vaker van mannen afkomstig zijn. Ook waar de constante niet significant verschilde van 0 was het teken positief. Deze conclusie kan worden afgeleid als de formule uit (5.1) wordt omgeschreven.

De formule die hier geldt is y=Xβ. Hierbij is Xβ een vector van de lengte 4021. De formule voor de *i*-de waarde van y kan als volgt worden weergegeven:

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_n X_{i,n} \quad (6.1)$$

met $X_{i,j}$ de het *j*-de woord van het *i*-de bericht van X, β_0 de coëfficiënt van de constante en β_1, β_2, … , β_n alle coëfficiënten die bij de woorden horen. Hierbij hoort β_1 bij de eerste kolom van X, β_2 bij de tweede kolom, etc.

Dit kan weer vereenvoudigd worden tot

$$y_i = \beta_0 + \beta_1 X_i \quad (6.2)$$

Hierbij is β_1 de vector met alle coëfficiënten die bij de woorden horen.

Als (6.2) in (5.1) wordt verwerkt, dan wordt de kans op een mannelijke schrijver van bericht *i*:

$$P(y_i = 1|X) = \frac{\exp(y_i)}{1 + \exp(y_i)} = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \quad (6.3)$$

Als X_i bestaat uit enkel nullen, wat betekent dat er geen enkel veelgebruikt woord voorkomt, dan neemt het tweede deel binnen de exponenten de waarde 0 aan.

Hierdoor wordt (6.3) gereduceerd tot

$$P(y_i = 1|X = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \quad (6.4)$$

Uit (6.4) valt af te lezen dat als β_0 een positief getal is, de waarde van de formule groter dan 0,5 zal worden. De kans is dus groter op een mannelijke schrijver dan op een schrijver van het vrouwelijke geslacht.

Voor de coëfficiënten van de woorden zelf is een andere interpretatie nodig. Dit wordt door middel van de odds-ratio bepaald. De odds-ratio is de mate van associatie tussen hoe vaak een bepaald iets voorkomt en een uitkomst. Hierbij is dus hoe vaak een bepaald iets voorkomt het aantal keer dat een bepaald woord voorkomt, en de uitkomst is of het bericht van een man of vrouw afkomstig is.

$$\frac{P(y_i = 1|X)}{P(y_i = 0|X)} \quad (6.5)$$

Dit is in formulevorm:

$$\frac{\Lambda(\beta_0 + \beta_1 X_i)}{1 - \Lambda(\beta_0 + \beta_1 X_i)} = \exp(\beta_0 + \beta_1 X_i)$$

$$= \exp\left(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_n X_{i,n}\right) \quad (6.6)$$

En, als dit wordt omgeschreven zodat het effect van het *j*-de woord apart staat, dan wordt dit:

$$= \exp\left(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{j-1} X_{i,j-1} + \beta_{j+1} X_{i,j+1} + \beta_n X_{i,n}\right) * \exp\left(\beta_j X_{i,j}\right) \quad (6.7)$$

Neem bijvoorbeeld "United" als het *j*-de woord. De coëfficiënt die bij dit woord hoort is 1.49. De mate waarin de odds (waarschijnlijkheid) dat het bericht afkomstig is van een man als er in het bericht één keer het woord "United" staat toeneemt is $e^{1.49}$=4.44. De kans dat het bericht afkomstig is van een man is dus ruim 4 keer zo groot als de kans dat het bericht afkomstig is van een vrouw.

Is het woord significant bij een vrouw ingedeeld, dan worden de odds juist lager. Nu is "Love" bijvoorbeeld het *j*-de woord. De mate waarin de odds dat het bericht afkomstig is van een man als er in het bericht één keer het woord "Love" staat toeneemt is $e^{-0.42}$=0.66. De kans dat het bericht afkomstig is van een man is dus 0.66 keer zo groot als de kans dat het bericht afkomstig is van een vrouw. Met andere woorden, de kans is groter dat het bericht afkomstig is van een vrouw.

De in-sample prediction error (de voorspelfout over de groep waar de regressie over is gedaan) is onder andere te bepalen door de Akaike Information Criterion (AIC) per model te bekijken. Het model met de laagste AIC wordt vaak aangenomen als het beste model. Dit betekent dat de voorspelfout binnen de 'sample groep' (de groep waar de coëfficiënten mee geschat worden om de rest te voorspellen), het kleinst zijn in dit model. Bij de kruisvalidatie is de laagste AIC die gevonden werd 1.296, en de hoogste AIC die gevonden werd 1.310. De AIC van de complete dataset is slechts één honderdste lager dan het minimum van de AIC bij de kruisvalidatie; de waarde hierbij is 1.295.

De Mean Squared Prediction Error (MSPE, gemiddelde gekwadratiseerde voorspelfout), is in dit model anders dan in een ander regressiemodel. Doordat de waarden van y alleen maar 0 of 1 kunnen aannemen, is de maximale afwijking 1. De MSPE is dus een waarde tussen 0 en 1. Wat de MSPE in dit model voorstelt is het percentage waar de

uitkomst van y niet goed voorspeld is. Dit is dus het tegenovergestelde van het voorspellingspercentage. Zoals eerder vermeld, verschilde het voorspellingspercentage tussen de 55% en 62%. De MSPE verschilt dus tussen 0.38 en 0.45. De standaarddeviatie van deze waarden is 0.02. Dit is zodanig klein dat aangenomen kan worden dat de groepen goed zijn verdeeld, en er dus geen groep was die een hele andere samenstelling had.

6.2 Woordcategorieën

Deze methode werkte niet veel beter dan de andere, hier werd een correct voorspellingspercentage van gemiddeld 58% bereikt. De modellen gaven bijna allemaal 7 coëfficiënten die significant verschilden van 0. Dit bevatte in elk van de gevallen de constante in het model. Twee woordcategorieën waren geen enkele keer significant; dit waren het voegwoord en het voorzetsel. Verder was er één categorie die in enkele gevallen wel en in enkele gevallen niet significant waren; dit was het bijwoord. De overige categorieën waren in elk van de gevallen significant verschillend en kunnen dus aangeven welk geslacht dit soort woorden vaker gebruikt. De coëfficiënten die in Tabel 2 staan zijn de uitkomsten van de regressie over de complete dataset.

Man	Waarde van de coëfficiënt bij regressie over complete data	Vrouw	Waarde van de coëfficiënt bij regressie over complete data
Lidwoord	0.15	Zelfstandig Naamwoord	-0.07
Voornaamwoord	0.04	Werkwoord	-0.12
Bijwoord	0.05	Bijvoeglijk Naamwoord	-0.10

Tabel 2 - Per geslacht staat aangegeven welke categorieën significant een grotere kans geven dat het bericht door een man dan wel een vrouw geschreven is. Op de categorie 'Bijwoord' na zijn alle categorieën in elk model significant verschillend van 0. Bijwoord is in enkele gevallen niet significant. De waarden in de tweede en vierde kolom geven de gemiddelde waarde aan die de coëfficiënten hadden.

Ook was de coëfficiënt van de constante in elk model positief, wat betekent dat berichten zonder woorden vaker van mannen afkomstig zijn. De coëfficiënten zijn erg klein vergeleken met de coëfficiënten die aan de woorden werden toegekend. De coëfficiënten zijn dus wel significant verschillend van 0, hoewel de kans dat het bericht afkomstig is

van een vrouw als er een zelfstandig naamwoord in het bericht voorkwam niet veel groter is dan de kans dat het bericht afkomstig is van een man. Wat wel als redenatie gebruikt kan worden, is dat de veelgebruikte woorden die significant aan een vrouw toegekend kunnen worden vrijwel allemaal zelfstandig naamwoord of werkwoord zijn, zie Tabel 1. Dit kan met Tabel 2 worden gecombineerd om het vermoeden te krijgen dat de woorden in de berichten van vrouwen zó veel zelfstandig naamwoorden en werkwoorden staan, dat deze samen zo vaak deze coëfficiënt hebben dat de kans op een vrouw als schrijver veel groter is. De waarden van de coëfficiënten kunnen op dezelfde manier worden geïnterpreteerd als de waarden van de coëfficiënten van de woorden.

7. Discussie en Conclusie

De nauwkeurigheid van de voorspellingen die is gevonden, is een stuk lager dan in eerste instantie was verwacht. In het onderzoek met berichten uit Twitter van Bamman et al (2012) is een voorspelling met 88% nauwkeurigheid van het voorspellen van geslacht. Zij hebben ook gebruik gemaakt van 10-voudige kruisvalidatie, maar in plaats van 9 groepen gebruiken om de tiende groep te voorspellen zoals in dit onderzoek is gedaan, hebben zij 8 groepen gebruikt om de coëfficiënten te schatten, één groep gebruikt om deze schattingen te controleren en eventueel aan te passen, en met dit model hebben ze de laatste groep geprobeerd te voorspellen. Het onderzoek van Bamman et al is dan wel voor een groot deel anders, zij hebben ook gebruik gemaakt van een logistische regressie. Een reden dat bij hun onderzoek een ander resultaat is, kan onder andere zijn dat hier veel specifiekere categorieën zijn gebruikt. De categorieën die in hun artikel werden gebruikt waren onder andere leestekens, scheldwoorden, getallen en afkortingen. Al deze soorten zijn in dit onderzoek niet meegenomen. Als er meer tijd was geweest had dit mogelijk een idee geweest om uit de dataset te filteren. Een andere reden kan zijn dat de dataset die in dit onderzoek is gebruikt een stuk beperkter is dan de dataset die in het onderzoek van Bamman et al is gebruikt. Hier bestond de dataset uit 4400 berichten, het bestand waar zij onderzoek mee hebben gedaan bestond uit 14000 gebruikers van Twitter. Het is niet duidelijk of per gebruiker ook meerdere berichten zijn gebruikt. Mocht dit niet zo zijn, dan is de dataset al ruim drie keer zo groot als de set die in dit onderzoek gebruikt is.

Als wordt vergeleken met het onderzoek van Argamon et al (2007), dan is er niet veel te vergelijken. De categorieën die ook in hun onderzoek voorkwamen en die zij vooral aan mannen toewijzen zijn lidwoorden (determiner) en voorzetsels (preposition), en aan vrouwen wezen ze de voegwoorden (conjunction) toe. Dit is de enige categorie die zij toewezen aan vrouwen die ook in dit onderzoek verwerkt is. In de resultaten van dit onderzoek is het lidwoord de enige categorie die een coëfficiënt had die significant verschilde van 0. Deze werd hier ook toegewezen aan het mannelijke geslacht. De andere twee categorieën konden in dit onderzoek niet worden toegewezen aan een geslacht. Ook hier kan weer reden zijn dat de uitkomsten verschillen doordat de dataset

die in het onderzoek van Argamon et al werd gebruikt veel groter is (ruim 68000 berichten versus 4400 berichten)

Opvallend is ook, dat in het onderzoek van Keune (2012) geconcludeerd werd dat Zelfstandig Naamwoorden significant vaker werden gebruikt door mannen. In dit onderzoek is juist deze categorie toegewezen aan vrouwen. Het is niet duidelijk waardoor dit verschil wordt veroorzaakt.

De dataset is niet op elk vlak netjes en duidelijk, de berichten stonden niet net in een bepaalde kolom, waardoor het soms onduidelijk was welke woorden bij welk bericht hoorden. Ook was bepaalde informatie in verkeerde kolommen terecht gekomen. Hierdoor kan het zijn dat het aantal woorden in een bericht verkeerd berekend is. Het kan zijn dat er informatie over een bericht in de kolom met tekst stond, waardoor dit is mee gerekend. Op de pagina met veelgebruikte woorden waren er 4400 berichten waar dit voor gespecificeerd was. Op de pagina met de complete tekst waren dit er meer. De nummering van deze berichten loopt tot 4489, maar er ontbreken tussendoor enkele berichten. Het was niet duidelijk of deze berichten ook ontbreken op de eerste pagina. Ook is niet duidelijk welke berichten er niet op de eerste pagina zijn verschenen. Ook heb ik in het hoofdstuk Data aangegeven dat de sheet met het aantal keer dat een veelgebruikt woord voorkwam in een bericht niet compleet klopte. Als werd gecontroleerd of de woorden waarvan staat aangegeven dat deze voorkwamen in een bericht werkelijk in het bericht stonden, dan was dit niet altijd het geval. Hierdoor kunnen de resultaten niet helemaal vertrouwd worden.

Een aantal woorden die in de modellen significant aan een geslacht zijn toegewezen zijn voor een deel voor de hand liggend, deze zouden door de meeste personen aan een bepaald geslacht worden toegewezen (denk aan love, bless, man, united). Ook is het woord ‘the’ aan het mannelijk geslacht toegewezen, wat niet vreemd is gezien vanuit het artikel van Argamon et al (2007), waar de lidwoorden als categorie bij de man werden geplaatst. Dit werd ook in dit onderzoek geconcludeerd.

Niet alle categorieën konden significant aan een geslacht worden toegewezen. Dit is niet heel vreemd, omdat niet alles per se geslachtsgebonden moet zijn. In andere onderzoeken waren de persoonlijk voornaamwoorden meegenomen, en in dit onderzoek de veralgemeniseerde voornaamwoorden. In de vooraf gedane onderzoeken

kwam als resultaat dat deze bij vrouwen horen. De voornaamwoorden daarentegen worden in dit onderzoek aan mannen toegewezen. Dit is niet onmogelijk, omdat voornaamwoorden veel meer zijn dan alleen persoonlijk voornaamwoorden. Het kan zijn dat als deze categorie los zou zijn meegenomen in het onderzoek, deze dan wel bij de vrouw zou horen.

Dit onderzoek heeft mij tot de conclusie gebracht dat er wel degelijk woorden en categorieën toe te wijzen zijn aan het mannelijk dan wel vrouwelijk geslacht, maar dit is niet in elk van de methoden dezelfde. Ook is een grotere data set nodig om nauwkeurigere conclusies te kunnen trekken.

Verder onderzoek

Voor verder onderzoek zou ik aanraden om eerst ruim de tijd te nemen om ervoor te zorgen dat de data netjes en zorgvuldig is samengesteld. Data dat is verkregen van één dag is minder representatief dan data dat is verkregen over een week, maand of jaar. De data deels filteren in veelgebruikte woorden is een goed idee, maar zoals andere onderzoeken hebben uitgewezen is het opdelen in categorieën ook een doeltreffende wijze. Denk hierbij breed, en niet alleen aan lidwoorden, werkwoorden, bijwoorden, maar ook aan populaire tekens als een hashtag (#), een emoticon of smiley (:D), en namen van personen of steden.

8. Appendix

Veelgebruikte woorden plus categorieën

In Tabel 3 zijn de 295 veelgebruikte woorden weergegeven, met de categorieën waar deze zijn ingedeeld.

Woord	Categorie 1	Categorie 2
"2012"	noun	
about	adjective	
after	conjunction	preposition
again	adverb	
ago	adverb	
all	determiner	
already	adverb	
also	adverb	
always	adverb	
and	conjunction	
another	adjective	
any	determiner	
anyone	pronoun	
anything	pronoun	
are	verb	
around	adverb	
ask	verb	
away	adverb	
awesome	adjective	
baby	noun	
back	noun	adverb
bad	adjective	
beautiful	adjective	
because	conjunction	
bed	noun	
been	verb	

Woord	Categorie 1	Categorie 2
make	verb	
making	verb	
man	noun	
many	determiner	
may	verb	
maybe	adverb	
might	verb	
mind	noun	verb
miss	verb	
mom	noun	
money	noun	
more	determiner	
morning	noun	
most	determiner	
much	determiner	
must	verb	
name	noun	
need	verb	
never	adverb	
new	adjective	
next	adjective	
nice	adjective	
night	noun	
not	adverb	
nothing	pronoun	
now	adverb	

before	adverb	conjunction
being	verb	
believe	verb	
best	adjective	
better	adjective	
big	adjective	
birthday	noun	
bless	verb	
blessed	adjective	verb
boy	noun	
bring	verb	
but	conjunction	
call	noun	verb
can	verb	
cant	verb	
care	verb	
check	verb	
church	noun	
city	noun	
come	verb	
coming	verb	
could	verb	
day	noun	
days	noun	
did	verb	
didnt	verb	
dinner	noun	
does	verb	
doing	verb	
done	verb	
dont	verb	
down	adverb	
early	adjective	

off	adjective	
old	adjective	
one	pronoun	
only	adjective	
open	adjective	verb
other	adjective	
our	pronoun	
out	adverb	
over	preposition	
own	verb	
party	noun	
people	noun	
person	noun	
place	noun	
please	verb	
power	noun	
praise	verb	
pray	verb	
profile	noun	
put	verb	
quite	adjective	
ready	adjective	
really	adverb	
remember	verb	
rest	noun	
right	adjective	
sad	adjective	
said	verb	
same	pronoun	
say	verb	
school	noun	
see	verb	
service	noun	

enjoy	verb	
even	adverb	
ever	adverb	
every	determiner	
everyone	pronoun	
everything	pronoun	
excited	adjective	
face	noun	
facebook	noun	
family	noun	
father	noun	
feel	verb	
feeling	noun	verb
few	determiner	
finally	adverb	
find	verb	
first	determiner	adverb
football	noun	
for	preposition	
forward	verb	
free	adjective	verb
friend	noun	
friends	noun	
from	preposition	
fun	noun	
game	noun	
gave	verb	
get	verb	
getting	verb	
giants	noun	
girl	noun	
give	verb	
giving	verb	

she	pronoun	
should	verb	
show	verb	
since	conjunction	
sleep	noun	
some	determiner	
someone	pronoun	
something	pronoun	
son	noun	
special	adjective	
start	noun	verb
stay	verb	
steelers	noun	
still	adverb	
sunday	noun	
sure	adjective	
take	verb	
taking	verb	
team	noun	
tell	verb	
than	conjunction	
thank	verb	
thanks	verb	
that	conjunction	
thats	verb	conjunction
the	determiner	
their	pronoun	
them	pronoun	
then	adverb	
there	adverb	
these	determiner	
they	pronoun	
thing	noun	

god	noun	
gods	noun	
goes	verb	
going	verb	
gonna	verb	
good	adjective	
got	verb	
gotta	verb	
great	adjective	
guess	verb	
had	verb	
happy	adjective	
hard	adjective	
has	verb	
have	verb	
having	verb	
head	noun	verb
hear	verb	
heart	noun	
help	noun	verb
her	determiner	pronoun
here	adverb	
him	pronoun	
his	determiner	pronoun
home	noun	
hope	noun	verb
house	noun	
how	adverb	
ill	adjective	
into	preposition	
its	determiner	
jesus	noun	
join	verb	

things	noun	
think	verb	
thinking	verb	
this	determiner	
those	determiner	
thought	verb	
through	preposition	
time	noun	
times	adverb	
tired	adjective	
todays	determiner	
together	adjective	
tomorrow	adverb	
too	adverb	
two	noun	
united	adjective	
until	adverb	
very	adjective	
viewed	verb	
wait	verb	
walk	noun	verb
want	verb	
was	verb	
watch	noun	verb
way	noun	
week	noun	
well	noun	adjective
went	verb	
were	verb	
what	determiner	pronoun
when	conjunction	
where	adverb	
which	determiner	pronoun

just	adverb	
keep	verb	
kids	noun	
know	verb	
last	adjective	verb
later	adjective	
left	verb	
let	verb	
lets	verb	
life	noun	
like	verb	
little	adjective	
live	verb	
lol	noun	
long	adjective	verb
look	verb	
looking	verb	
lord	noun	
lost	verb	
lot	pronoun	
love	noun	verb
luck	noun	
made	verb	

while	conjunction	
who	pronoun	
why	determiner	
will	verb	
win	verb	
wish	verb	
with	preposition	
without	preposition	
woke	verb	
wonderful	adjective	
wont	verb	
word	noun	
work	verb	
working	verb	
world	noun	
worship	noun	
would	verb	
year	noun	
years	noun	
yesterday	noun	
you	pronoun	
your	pronoun	

Tabel 3 – alle 295 veelgebruikte woorden, deze staan in de linkerkolom. In de andere twee kolommen staan de categorieën waar deze woorden zijn ingedeeld voor het onderzoek met gecategoriseerde woorden.

9. Bronvermelding

Argamon, Shlomo, Koppel, Moshe, Pennebaker, James & Schler, Jonathan. (2007). Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday* 12(9).

Bamman, David, Eisenstein, Jacob & Schnoebelen, Tyler. (2012). Gender in Twitter – Styles, Stances and Social Networks

Deuchar Margaret. (1989). Pragmatic account of women's use of standard speech. In Jennifer Coates & Deborah Cameron (eds.), *Women in their speech communities*, London: Longman. 27-32

Keune, Karen. (2012). Explaining register and sociolinguistic variation in the lexicon: Corpus studies in Dutch

Trudgill, Peter. (1972). Sex, covert prestige and linguistic change in the urban British English of the Norwich. *Language in Society* 1(2):179-195.

http://en.wikipedia.org/wiki/English_grammar. beschrijving van de categorieën (word classes)