Analyse van Nederlandstalige tweets

Erik Tjong Kim Sang
Meertens Instituut, Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

25 augustus 2016

Tweets

Tweets zijn korte berichten (maximaal 140 tekens) die door gebruikers worden verspreid op het sociale netwerk Twitter

Tweets zijn interessant voor onderzoek: ze zijn actueel, beslaan diverse onderwerpen en zijn beschikbaar in vele talen

Voor onderzoek bevat `twiqs.nl` een archief van Nederlandstalige tweets dat teruggaat tot december 2010

Sentimentanalyse

Sentimentanalyse is het bepalen van het sentiment van een zin of een tekst

Hierbij wordt vaak onderscheid gemaakt tussen positief sentiment en negatief sentiment maar er kan ook naar andere sentimenten worden gezocht

Het sentiment dat tot uitdrukking wordt gebracht is een relatie tussen de schrijver van de tekst en een onderwerp

Toepassing van sentimentanalyse van tweets

Binnen ons onderzoek was sentimentanalyse van tweets nodig voor twee toepassingen:

1. voorspellen van verkiezingsresultaten
2. vinden van de gelukkigste provincie

Een probleem is dat er te veel data is om door mensen te laten analyseren en dat er geen systeem is om het automatisch te doen

Algemene aanpak

Tekstanalyseprojecten beginnen vaak met deze drie stappen:

1. Verzamelen van de relevante data
2. Manuele analyse van een deel van de data
3. Bouwen systeem voor automatische analyse

We laten nu zien hoe deze drie stappen verliepen voor de twee toepassingen

VERKIEZINGEN VOORSPELLEN

1. Verzamelen van de relevante data

We hebben op twiqs.nl gezocht naar tweets van de week voor de verkiezingen met daarin een van de 11 deelnemende partijen

Dit resulteerde in 64.395 tweets

Daarnaast verzamelden we op dezelfde manier tweets van de dag precies twee weken voor de verkiezingen

Dit leverde ongeveer 7000 tweets op

2. Manuele analyse van een deel van de data

Van de 7000 tweets zijn 1678 geannoteerd door twee annotatoren

De tweets kregen de classificatie: positief, neutraal of negatief

De annotatoren waren het eens over 1333 tweets (kappascore: 0,59)

Na het verwijderen van tweets met meerdere partijen en extra tweets van gebruikers bleven 227 negatieve en 534 niet-negatieve tweets over

3. Bouwen systeem voor automatische analyse

Een automatisch classificatiesysteem heeft veel trainingdata nodig en met bij voorkeur weinig klassen

We hadden te weinig tweets om een betrouwbaar automatisch analysesysteem te bouwen

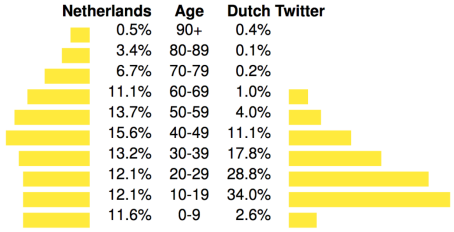
In plaats daarvan hebben we statistieken afgeleid om de latere data mee te kunnen analyseren

Bijvoorbeeld: 30% van de tweets over het CDA was negatief en daarom werd van de latere tellingen van CDA-tweets 30% afgetrokken

Resultaten: Provinciale Statenverkiezingen van 2011

Partij	Zetels	Zetels PB	Zetels MdH	Zetels Twitter
VVD	16	14	16	14
PvdA	14	12	11	16
CDA	11	9	9	8
PVV	10	11	12	10
SP	8	9	9	6
D66	5	7	5	8
GL	5	4	4	3
CU	2	3	3	3
50+	1	2	2	2
SGP	1	2	2	2
PvdD	1	1	2	2
OSF	1	1	0	1
	afwijking:	14	14	18

Twitter is niet hetzelfde als Nederland!



DE GELUKKIGSTE PROVINCIE

1. Verzamelen van de relevante data

Van `twiqs.nl` verzamelden we twee groepen tweets van dezelfde maand (januari 2013):

- positieve tweets, die :-) of :) bevatten (1.724.642 tweets)
- negatieve tweets, die :-(of :(bevatten (999.685 tweets)

2. Manuele analyse van een deel van de data

We hebben de frequenties van de woorden uit de twee verzamelingen vergeleken met de t-test (Church e.a., 1991)

Daaruit kwamen woorden en emoji naar voren die vaker in de positieve verzameling voorkwamen, of juist vaker in de negatieve

We hebben manueel 338 van deze woorden gekozen voor ons sentimentlexicon

Voorbeelden positieve woorden: :d blij gelukkig haha mooi veel winnen

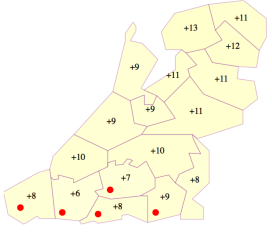
(automatisch de woorden kiezen, werkte niet zo goed)

3. Bouwen systeem voor automatische analyse

We gebruiken de volgende regels voor sentimentanalyse:

1. tweets zonder woorden uit het sentimentlexicon zijn neutraal
2. tweets met meer positieve dan negatieve woorden zijn positief
3. tweets met meer negatieve dan positieve woorden zijn negatief
4. bij een gelijk aantal positieve en negatieve woorden telt het laatste woord

Resultaat: gemiddelde sentimenten in januari 2014



Conclusies

Tekstanalyseprojecten beginnen met 1. data-verzameling, 2. manuele analyse en 3. het bouwen van een automatisch analysesysteem

Voor stap 3 zijn diverse gesuperviseerde machinale leerders beschikbaar maar deze hebben veel trainingdata nodig (duizenden) en bij voorkeur een klein aantal klassen

Een grondige analyse van de typen benodigde annotatie kan helpen om het te verrichten werk in stappen 2 en 3 te beperken

Technische kennis is zeer bruikbaar in dit soort projecten

Literatuur

Predicting the 2011 Dutch Senate Election Results with Twitter, by Erik Tjong Kim Sang and Johan Bos. Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, ACL, Avignon, France, 2012, pages 53-60. http://ifarm.nl/erikt/papers/sasn2012.pdf

Using Tweets for Assigning Sentiments to Regions, by Erik Tjong Kim Sang. In: Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data at LREC2014, Reykjavik, Iceland, 2014. http://ifarm.nl/erikt/papers/lrec2014b.pdf

EINDE