Twitter in The Netherlands

Erik Tjong Kim Sang
University of Groningen

20 January 2012

Twitter

Twitter is a widely used microblog service in which the 300+ million users broadcast text messages of up to 140 characters. Example:

rihanna Rihanna
@iAmQueenLatifah just bust out @ @iamdiddy party singing #G4L !!! I cud cry!!! Ok I'm really crying!! My year started SO right!!! #MAJAH
1 Jan

The sheer volume of the text data produced every day makes Twitter an interesting resource for computational linguists

We use Dutch tweets as a window on the Dutch speaking community

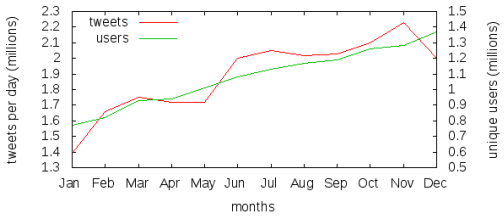
Collecting Dutch tweets

We use Twitter's filter API and retrieve tweets containing common Dutch words: een, het, ik, niet, maar, voor, ook, als, heb, naar, ...

An ngram language identifier by Thomas Mangin and language models developed by our students filter out the remaining non-Dutch tweets

The recall of this data collection method is about 37% while the precision is close to 100%

689,394,785 Dutch tweets in 2011



Visualizing Twitter data

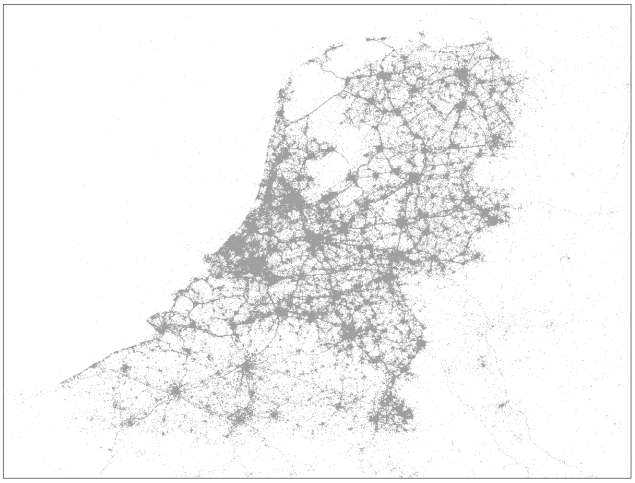
About one million people are broadcasting Dutch messages about their daily activities on Twitter

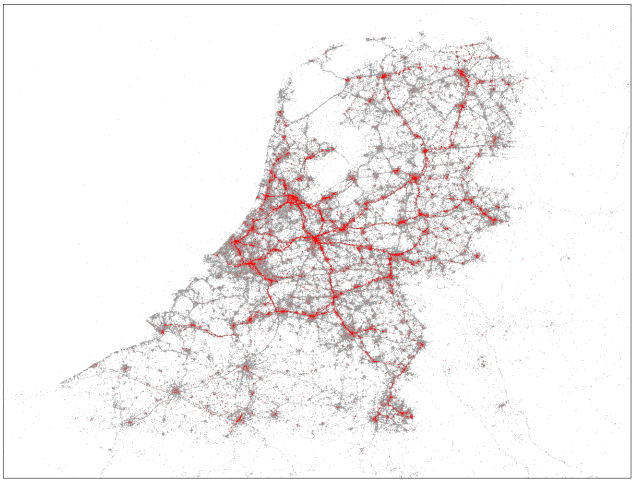
Can we use this data to see what the population of The Netherland and Flanders is occupied with?

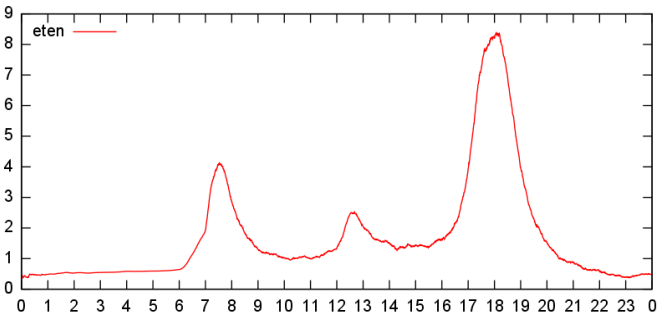
Are some words used more often in the morning than in the afternoon?

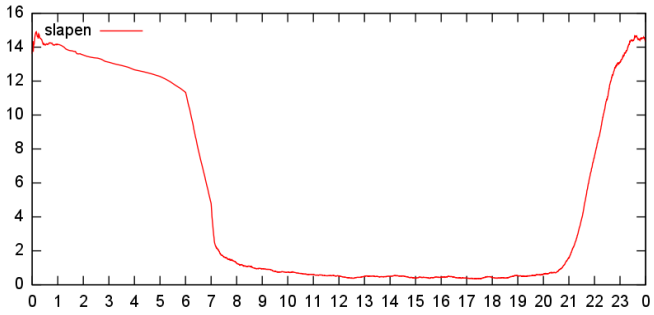
What is the impact of news events on Twitter?

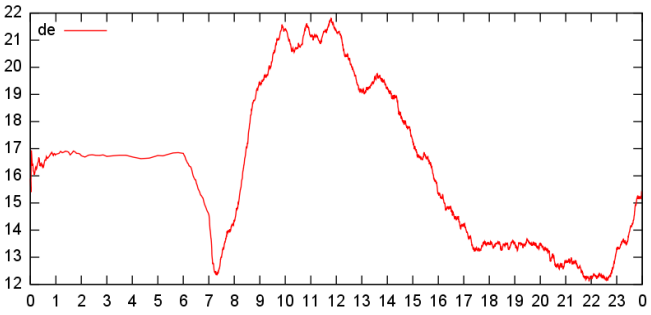
Is there a relation between word usage and location?

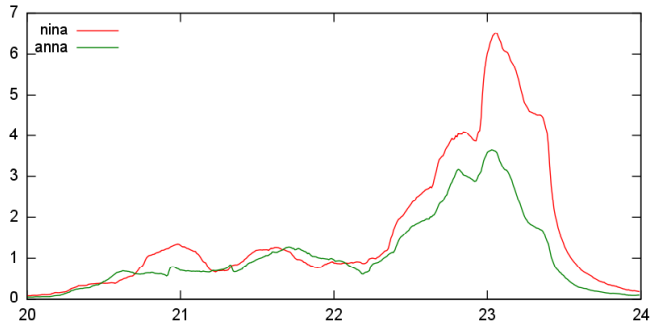












Dutch Provincial Elections 2011

Every four years the Dutch choose the boards of the twelve provinces

The boards contain between 39 and 55 seats and the number of all seats adds up to 564

After each election, the newly chosen board members select the members of the Dutch Senate

The 2011 provincial election was on Wednesday 2 March

Examples of political tweets

- Iedereen uit het Rivierengebied; stem op 2 maart. Stem verstandig, stem VVD, stem op Remco Dijkstra, nr 8 VVD lijst.
- Ik ben blij dat ik vandaag niet met de trein moet. :-) PvdA'ers staan koffie en rozen uit te delen op station. Lust geen koffie.
- Waar een WhatsAppje al niet goed voor is .. Net mijn PVV-neigende zus overgehaald D66 te stemmen #groenlinkswashelaaseenbrugtever
- NEE, Stemwijzer, ik stem geen 50PLUS #beledigd

Estimated sentiment scores per party

We estimated sentiment scores per party by manually analyzing 4000+ tweets containing Dutch party names:

Score	Party
0.952	50Plus
0.909	D66
0.846	PvdD
0.833	SP
0.789	GroenLinks
0.712	ChristenUnie
0.686	CDA
0.676	VVD
0.642	PvdA
0.521	PVV

Prediction results in Senate seats (total 75)

Party	Result	Pol.Bar.	De Hond	Twitter	
VVD	16	14	16	11	←
PvdA	14	12	11	10	←
CDA	11	9	9	10	
PVV	10	11	12	14	←
SP	8	9	9	7	
D66	5	7	5	8	
GL	5	4	4	9	←
CU	2	3	3	2	
SGP	1	2	2	1	
PvdD	1	1	2	2	
50Plus	1	2	2	0	
others	1	1	0	1	
deviation	-	14	14	21	

Prediction results high school elections

Party	Result	Twitter	
PVV	16	14	
PvdA	13	10	
VVD	11	11	
D66	8	8	
GL	6	9	
SP	6	7	
PvdD	5	2	
CDA	3	10	←
50Plus	3	0	
CU	2	2	
SGP	1	1	
others	1	1	
deviation	-	22	

Concluding remarks

We have discussed:

- how to collect text data from Twitter
- how to visualize this data
- how to predict event outcomes from Twitter data

Twitter is an interesting resource for computational linguists!

Demo available on: `http://www.let.rug.nl/erikt/twitter`

THE END