Twiqs.nl: Searching in billions of Dutch tweets

Erik Tjong Kim Sang
Meertens Institute, Amsterdam, The Netherlands
23 November 2015

Twitter

Twitter is a microblog website started in 2006 by the American company Obvious (Jack Dorsey)

Users broadcast messages with a maximum length of 140 characters

Users can subscribe to messages of other users, in which case they follow the other users

Relations on Twitter are asymmetrical: you can follow somebody but that user does not have to follow you

The most followed Twitter users can be found on twitaholic.com/

Behind the screen (1)

Behind the screen (2)

Research based on tweets

An increasing number of students and researchers wants to study tweets

Tweets cover a wide variety of topics and are written in many different languages and dialects

Researchers and students often lack the technical knowledge to collect, store and analyze tweets

For this reason the Netherlands eScience Center has developed the website twiqs.nl for supporting research involving tweets written in Dutch

Erik Tjong Kim Sang

• MSc from Delft, PhD from Groningen
• computer linguist, specialized in machine learning
• working with social media data since 2010
• 2011: prediction of Dutch elections with tweets
• 2012-2013: built website twiqs.nl for researching Dutch tweets

What does a tweet look like?

Twiqs.nl: Searching in billions of Dutch tweets

Erik Tjong Kim Sang

• MSc from Delft, PhD from Groningen
• computer linguist, specialized in machine learning
• working with social media data since 2010
• 2011: prediction of Dutch elections with tweets
• 2012-2013: built website twiqs.nl for researching Dutch tweets
How do we collect tweets?
We use the Twitter streaming API for continuously searching for new tweets which contain at least one of a list of 229 frequent words. We collect about one million tweets per day of which about 800,000 are identified by our software as written in Dutch. The tweets are stored in compressed files, each of which contains one hour of tweets and their metadata encoded in the data format JSON. The tweet collection contains 4.3 billion tweets of which 2.9 billion are written in Dutch.

The challenge: searching in the collection of tweets
Question: how can we find tweets in this large collection?
Answer: use an information retrieval system
But...
• we have a really large collection of data
• each second new tweets are added to the collection
• we want to search in both tweets and in metadata

Solution: use a computer cluster (cloud)
We use a computer cluster to process search queries. Every search query is split into several smaller queries, each of which processes one file with tweets from the same one hour. These smaller queries can be processed in parallel. The search software of twiqs.nl runs on a Hadoop system, which is based on Google’s MapReduce algorithm. The search software runs on the experimental Hadoop cluster of SURFsara (170 nodes each with 8 cpus; 2.3 petabyte storage). This cluster is available for all researchers in The Netherlands.

Searching with MapReduce: Map task
The map task selects the relevant tweets per hour:

```java
for (each tweet in data file) {
    if (query matches tweet) {
        process tweet
        put results on output with data key (minute)
    }
}
```

By developing the (Java) search software ourselves, we have full control over the search process.

Example Reduce task
The Reduce task combines the results of the various Map tasks:
combine Map results of the current data key (minute) put results per data key on the output
The Reduce task sorts the output of the Map tasks based by the data key values.
Advantages of MapReduce

- Full control over search and data processing
- Searching in several hours of data is almost as fast as searching in one hour of data

Disadvantages of MapReduce

- It is not as fast as when using information retrieval
- Processing might have to wait on jobs of other users

TOPICS ON TWITTER

When are certain topics discussed on Twitter?

With twiqs.nl we can determine how popular a certain topic is on Dutch Twitter

We count how often tweets mention a certain topic in a time frame and put the counts in a graph

But during the night fewer tweets are written and we try to keep that effect out of the graphs

Therefore we do not make graphs with tweet totals but graphs with percentages, for example the percentage of Dutch tweets that contain the word Parijs

Graphs of daily word usage

With twiqs.nl you can make graphs of the frequencies of words used on Dutch Twitter

The graphs will present you the frequencies of words on Dutch Twitter almost live

You can also request graphs for other days than today, for example for Friday 13 November 2013 (Paris attacks)
Summary

We can draw graphs of word frequencies on Dutch Twitter.

The graphs show which topics are popular on Twitter at a certain time.

This provides insights in what topics an average user of Dutch Twitter is interested in.

Noise can be removed from the graphs by increasing the smooth factor.

---

Predicting election results

Can we use Twitter for predicting the results of an election?

Tumasjan et al (2010): yes, this is possible:

1. count the number of times that a tweet mentions a political party
2. convert the tweet counts to seat counts
3. done!

---

<table>
<thead>
<tr>
<th>Party</th>
<th>Short name</th>
<th>Long name</th>
<th>Seats</th>
<th>Seats PN</th>
<th>Seats SMH</th>
<th>Average polls</th>
</tr>
</thead>
<tbody>
<tr>
<td>PvdA</td>
<td>2226</td>
<td>12227</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>VVD</td>
<td>1562</td>
<td>1592</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>CDA</td>
<td>1964</td>
<td>1904</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>9.5</td>
</tr>
<tr>
<td>PvdA</td>
<td>1056</td>
<td>1037</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>9.5</td>
</tr>
<tr>
<td>SP</td>
<td>109</td>
<td>119</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>GL</td>
<td>243</td>
<td>505</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>D66</td>
<td>610</td>
<td>610</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5.5</td>
</tr>
<tr>
<td>CU</td>
<td>119</td>
<td>238</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>PvdD</td>
<td>105</td>
<td>119</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SGP</td>
<td>139</td>
<td>139</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SN</td>
<td>6</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DSP</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

PREDICTING ELECTION RESULTS
Twiqs.nl: Searching in billions on Dutch tweets 23 November 2015

The bad news
Nobody has been able to reproduce the excellent prediction results of Tumasjan (2010)

Problems of this approach:
- not all tweets mentioning a party are positive
- sometimes one person will mention a party many times on Twitter
- Twitter users are not a representative sample of the Dutch population

Examples of negative tweets
- De VVD en de PvdA heven slechts 1% centjes naar Brussel over te maken en voor de rest kunnen ze (het UD) doen en laten wat ze willen.
- VVD and PvdA only need to send your money to Brussels and otherwise they can do as they please
- Mijn pappie wordt nog steeds gepest door dat #pvv schorren op Twitter! Morgen ga ik angaf doen, munt toch pas ‘s middags winken.
- My dad is still being harassed by those PVV bastards on Twitter. I’m going to press charges tomorrow, I only working in the afternoon.
- SP heeft nog steeds niet door dat alienen dure woningen worden gebouwd, omdat dit maar geld in de Gemeentekas brengt http://t.co/0tvpeX8a
- SP still does not understand that only expensive houses are built because this earns the municipality more money

Ages of Twitter users

<table>
<thead>
<tr>
<th>Netherlands</th>
<th>Age</th>
<th>Twitter (Dutch)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5%</td>
<td>90+</td>
</tr>
<tr>
<td></td>
<td>3.4%</td>
<td>80-89</td>
</tr>
<tr>
<td></td>
<td>6.7%</td>
<td>70-79</td>
</tr>
<tr>
<td></td>
<td>11.1%</td>
<td>60-69</td>
</tr>
<tr>
<td></td>
<td>13.7%</td>
<td>50-59</td>
</tr>
<tr>
<td></td>
<td>15.6%</td>
<td>40-49</td>
</tr>
<tr>
<td></td>
<td>13.2%</td>
<td>30-39</td>
</tr>
<tr>
<td></td>
<td>12.1%</td>
<td>20-29</td>
</tr>
<tr>
<td></td>
<td>12.1%</td>
<td>10-19</td>
</tr>
<tr>
<td></td>
<td>11.6%</td>
<td>0-9</td>
</tr>
</tbody>
</table>

Method for predicting elections
1. count the number of tweets that mention a political party
2. subtract the negative tweets
3. link the counts to a reliable political poll of the same time
4. repeat steps 1 and 2 right before the elections
5. determine an election prediction based on steps 1–4

Example: PvdA was polled at 12 seats and at the same time was mentioned in 25% of the tweets. If the party is mentioned in 25% of the tweets right before the election, we estimate it to win 25/20 * 12 = 15 seats.

Predicting elections (Dutch Senate 2011)

<table>
<thead>
<tr>
<th>Party</th>
<th>Result</th>
<th>Seats</th>
<th>Pb</th>
<th>MdH</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>VVD</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>PvdA</td>
<td>14</td>
<td>12</td>
<td>11</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>CDA</td>
<td>11</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>PVV</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D66</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>GL</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>CU</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>SGP</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>PvdD</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>OSF</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>offset</td>
<td>-</td>
<td>14</td>
<td>14</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

Predicting elections without using tweets

<table>
<thead>
<tr>
<th>Party</th>
<th>Result</th>
<th>Seats</th>
<th>Pb</th>
<th>MdH</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>VVD</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>PvdA</td>
<td>14</td>
<td>12</td>
<td>11</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>CDA</td>
<td>11</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>PVV</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>D66</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>GL</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>CU</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>SGP</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>PvdD</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>OSF</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>offset</td>
<td>-</td>
<td>14</td>
<td>14</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

Summary
Tweets are not useful for predicting elections, at least not this time

The demographic features of Twitter users are different than those of the population of The Netherlands.

It is important to compare the performances of complex computer models with simple models, the so-called baselines (here: the prediction that did not use tweets)

WHAT IS THE HAPPIEST LOCATION IN THE NETHERLANDS?
What is the happiest location in The Netherlands?

We start with making this research question more specific:

We want to use tweets for answering the question so a more concrete version is: which location in The Netherlands broadcasts the highest percentage of positive tweets?

We need areas and therefore we examine provinces and municipalities

Sentiment analysis

Determining if a text is positive, negative or neutral is called sentiment analysis.

The best way to perform this task is to have people read the texts. They can assign a human emotion to the text: happy, unhappy, confused, angry, aggressive, ...

We want to process millions of texts (tweets) and therefore use an automatic solution, without determining topics

Automatic sentiment analysis

We use two word lists: one with positive and one with negative words.

When a tweet contains one positive word, we classify it as positive, and when it contains one negative word, we classify it as negative.

Whenever a tweet contains more than one word of the two word lists, we only consider the final relevant word.

The words geen (no) and niet (not) reverse the sentiment value of the following word.

Tweets without words from the sentiment word lists will be classified as neutral.

Determining the location of a tweet

A small percentage of tweets contains position information in the shape of a longitude and a latitude coordinate, for example 6.568027,53.219353.

Borders of regions are defined as sequences of such coordinate pairs (lines).

But how do we determine which region contains a coordinate point?

Solution: the algorithm point in polygon.

Computing a sentiment score of a region

1. find tweets that have been sent from the region
2. determine the sentiment score of each of the relevant tweet
3. subtract the percentage of negative tweets from the percentage of positive tweets

The result is a score between -100% (only negative tweets) and +100% (only positive tweets).

Next: display the sentiment scores of all regions on a map.

The most negative user on Dutch Twitter...

MeteoWestzaan.nl

Baro 1025,5mb-Stijgt langzaam. Temp 16,8°C (-0,4). Hum 44%. Rain last 24h 0,0mm. Wind 0,2kph WNW / gust 2,2kph. UV o. 

Next: display the sentiment scores of all regions on a map.

The most negative user on Dutch Twitter...
Solution

Now every user can only contribute one score to a region: his or her average score. This makes the region scores more fair: one person - one vote.

Results of December 2013:

1. +30 Ameland (FR)
2. +28 De Marne (GR)
3. +27 Ubbergen (GE)
4. +20 Reusel-De Mierden (NB)
5. +8 Uithoorn (NH)

Summary

A small percentage of tweets contains location information which can be used for mapping tweet data.

While determining the sentiment of a region it is important to maintain the one person - one vote principle.

On the average tweets written in Dutch are positive.

Literature


THE END

Bonus: Animation of Twitter data

Pulse of the Nation is an animation which combines tweets with time, tweet volume and sentiment.

You will see a map of the United States where:

- states become larger as the tweet volume increases
- states change color as the sentiment changes (green vs red)

Address: http://www.youtube.com/watch?v=uJcrJR6SGkg