Real-time Language Independent Sentiment Analysis in Social Networks

Jürgen Nützel, Frank Zimmermann

4FriendsOnly.com Internet Technologies AG (4FO AG)
{jn, fz}@4fo.de

4FriendsOnly.com
Internet Technologies AG

Spin-off of Fraunhofer

Institut
Digitale
Medientechnologie

Outline

- Motivation
- Micro-blogging Service Twitter
- Sentiment Extraction
- Results – Different Types of Words
- Results – Pos. and Neg. Words
- Results – Weather and Mood In New York
- Results – High Standard Deviation
- Conclusion

Motivation

- Some groups (political parties, financial investors, companies) want know the current sentiment of the people about different topics.
- User of social networks like Twitter produce every day a huge amount of content (>340 mio tweets/day).
- They write almost about every topic. We have real-time access to this data stream for our sentiment research.

Micro-blogging Service Twitter

- 500mio users / 340mio tweets per day
- Per Tweet only 140 characters → user must use creative shortcuts like smileys ☺
- Access via streaming API (real-time access)
- We filter out Tweets from some areas (Germany, BosWash (incl. New York) and San Francisco bay area)

Region	Tweets per day	Total number of Tweets until now
Germany	10 800	3.1 mio
New York	72 000	7.1 mio
BosWash	236 000	15.6 mio
SF bay area	19 500	0.98 mio

Sentiment Extraction

- We use the language independent smiley detection in Tweets
- We calculate a normalized sentiment value (between -100 for a negative sentiment and +100 for a very good sentiment) for every word in every collected region for a certain time span

Sentiment	Normalized Sentiment Value	Smileys
Positive	+100	:-) ;) ;-) :P ;P ^.^ :D ;D :-D ;d -D ,-)
Negative	-100	:(:-(:| :-| :~(:'-(:,(

- Smiley dispensation over all tweets (40,3mio)
  - 1 843 000 pos. smileys
  - 367 620 neg. smileys

Sentiment Extraction for Words or Topics [1/3]

- First we create a text corpus *C*.
- The text corpus contains all Tweets of a selected region (e.g. New York) and of a certain time span (the last 7 days for BosWash).
- As newer Tweets will be added to the corpus, older Tweets leave the corpus. This will be done on a daily basis. The corpus length for each region was found by experiments.
- We calculate a sub corpus *C‘* of *C*. *C’* includes only Tweets which contain a smiley from the table before.

Sentiment Extraction for Words or Topics [2/3]

- For each word *w* we found in *C'* we calculate how many Tweets are in *C'* which contains *w*. We call it *count(w)*.
- If *count(w)* is smaller than a lower border (e.g. 10) the word *w* will be ignored. We have too less data to analyze sentiment of word *w*.
- We want to have comparable sentiment values in the same range: from +100 (positve) to -100 (negative)
- The summarized sentiment: *ss(w)*: For each word *w* we found in the Tweets of *C'* we add to *ss(w)* +100 for a positive smiley we found in the same Tweet. We add -100 for a negative smiley.

Sentiment Extraction for Words or Topics [3/3]

- In the last step we normalize the sentiment value

$$s(w) = \frac{ss(w)}{count(w)}$$

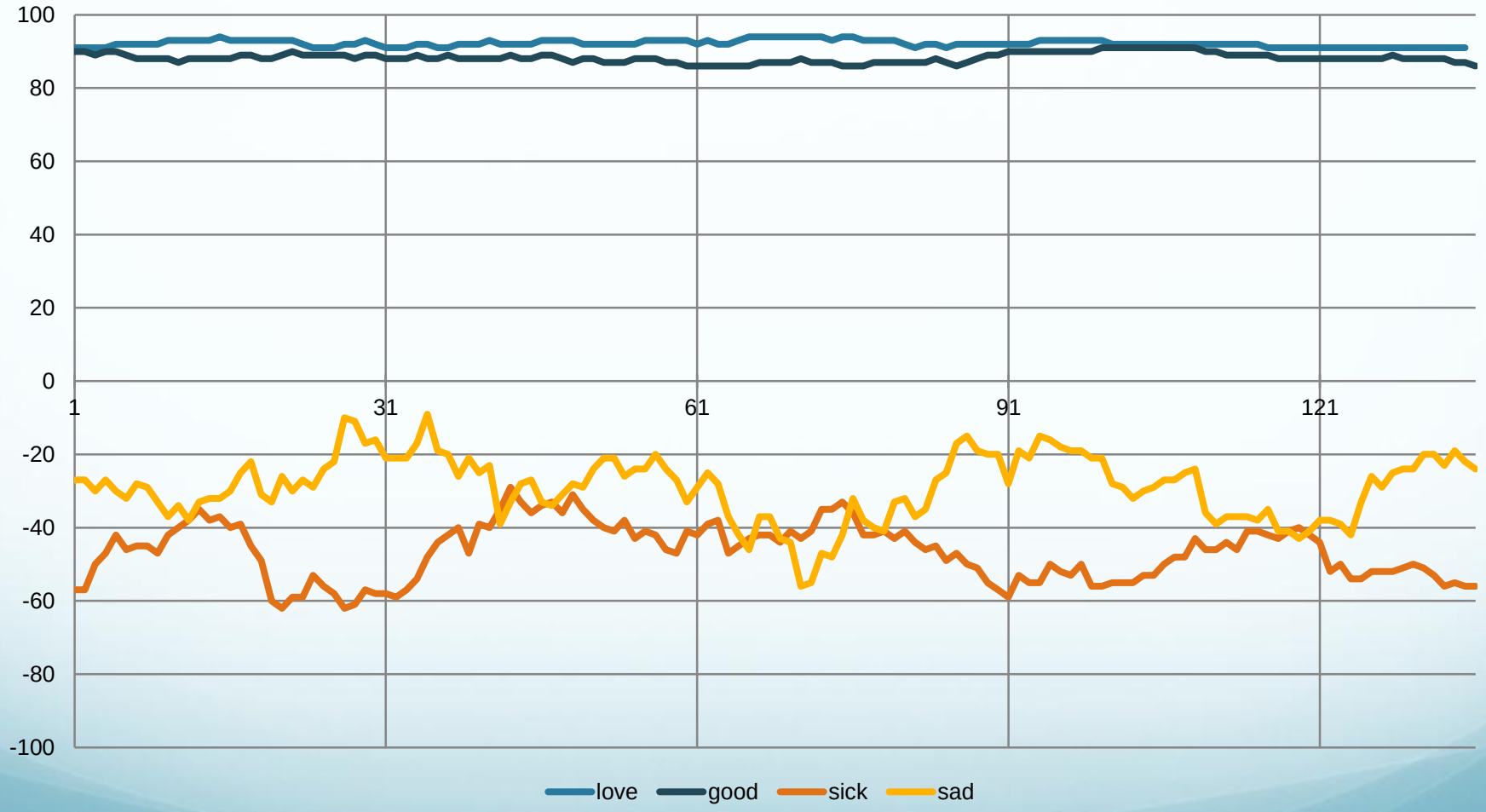
- nSENTI = s(w)

Results – Different Types of Words [1/2]

- We have found words with a constant high sentiment
- Normalized sentiment (nSENTI) over the time (*love*, *amazing*, *good*)
- Found words with „negative“ nSENTI like *sad* and *sick* over the time but with a high standard deviation
- Most interesting types of words have average nSENTI and a high standard deviation like *facebook* and *instagram*
- We saw interesting difference between the areas, the word weather (Wetter in German) has a more pos. sentiment in the Bay Area as in the other areas
  → more sunnier climate

Results – Pos. and Neg. Words

- time period between April to mid August



Results – Different Types of Words [2/2]

Word w	Area	Average of s(w)	Standard deviation of s(w)
Love	New York	92.18	0.91
Amazing	New York	92.49	2.68
Good	New York	88.29	1.33
Sad	New York	-46.61	7.69
Sick	New York	-28.95	8.79
Weather	New York	65.03	10.08
Weather	Bay Area	81.10	2.39
Weather	BosWash	58.29	5.62
Wetter	Germany	76.27	5.45
Vacation	New York	67.43	18.41
Sun	New York	73.01	12.26
Rain	New York	48.07	11.85
Tv	New York	71.29	12.41
Bahn	Germany	76.37	9.99
Mood	New York	72.02	9.44
money	New York	65.92	10.37
apple	New York	75.43	9.02
facebook	New York	75.06	11.70
instagram	New York	77.55	14.69
ipad	New York	71.21	19.84
iphone	New York	55.32	12.11
jfk	New York	54.90	8.28
macdonalds	New York	75.87	12.74

Results – Weather and Mood in New York

- April to mid August

Results – High Standard Deviation

Conclusion

- Summary
  - Social networks deliver us so much live content. This allows us to observe the sentiment of a large number of topics/words.
  - With position data we can calculate the sentiment of a word for every region
  - We are able to calculate the sentiment changes over time and can recognize big sentiment changes
  - We can also detect indirectly events with these sentiment analyse, a trivial example is the weather in a area
- Further Research Tasks
  - Some new open questions: Why is the standard deviation in neg. words so high.

Thank you

Jürgen Nützel, Frank Zimmermann

4FriendsOnly.com Internet Technologies AG (4FO AG)
{jn, fz}@4fo.de

4FriendsOnly.com
Internet Technologies AG

Spin-off of Fraunhofer

Institut
Digitale
Medientechnologie