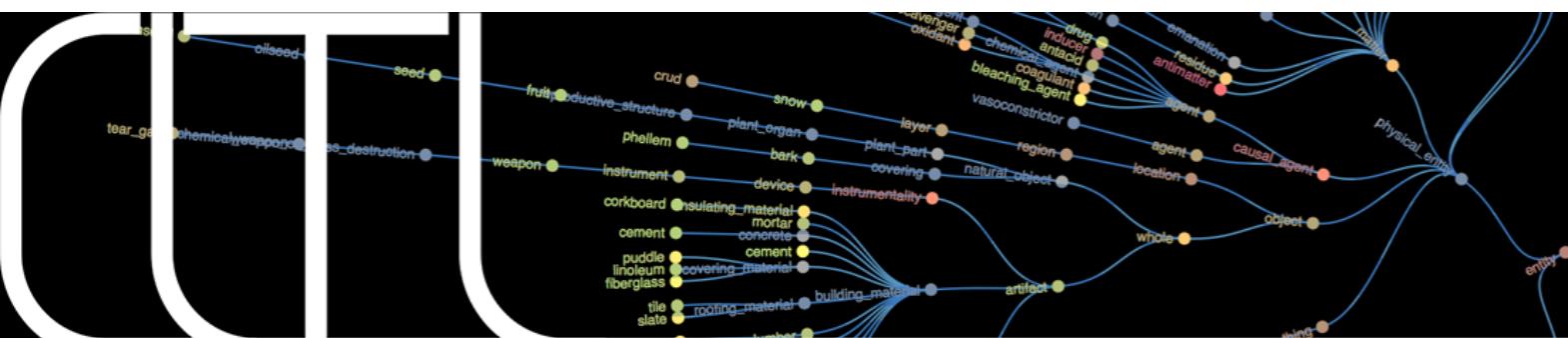


# Text Mining CBS



## Lecture 3: Subjectivity mining Piek Vossen



# Overview

- Part I: What is subjectivity? Many different things....
- Part II: Subjectivity mining
- Part III: Tools and resources

# Part I

# What is subjectivity in language?

- bella is the picture of health with boundless energy until a few days before she dies . this is absolutely and completely ridiculous and an insult to every family whose mother has suffered through the horrible pains of a death by cancer .

`nltk_data/corpora/sentence_polarity/rt-polarity.neg`

# Part I

# What is subjectivity in language?

- bella is the picture of health with **boundless energy** until a few days before **she dies** . this is absolutely and **completely ridiculous** and an **insult** to every **family** whose **mother** has **suffered** through the **horrible pains** of a **death** by **cancer** .

`nltk_data/corpora/sentence_polarity/rt-polarity.neg`

- **Explicit sentiment:** boundless energy, completely ridiculous, insult, horrible
- **Implicit sentiment:** dies, suffer, pains, death, cancer
- **Holders:**
  - **Author**
  - **Participants** of the text: she, family, mother about the cancer, the pain and the death
- But also what is and is not mentioned in the text is subjective (agenda setting)

# Part I

# What is subjectivity in language?

- A Colombia government trade official has urged the business community to aggressively diversify its activities and stop relying so heavily on coffee. Samuel Alberto Yohai, director of the Foreign Trade Institute, INCOMEX, said private businessmen should not become what he called "mental hostages" to coffee, traditionally Colombia's major export.

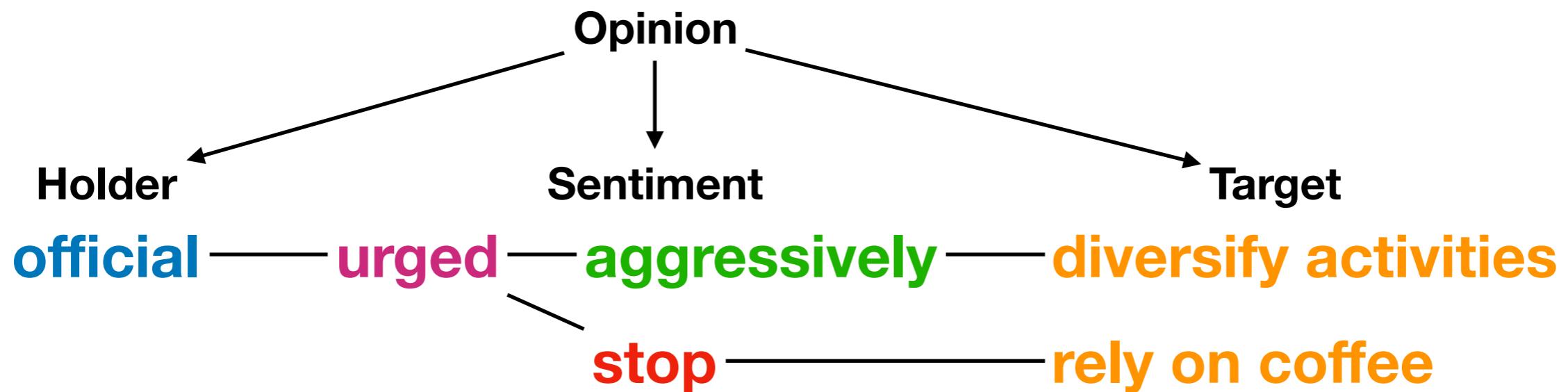
`nltk_data/corpora/reuters/test/15198`

# Part I

## What is subjectivity in language?

- A Colombia government trade **official** has **urged** the business community to **aggressively diversify** its **activities** and **stop relying** so heavily on **coffee**. **Samuel Alberto Yohai**, director of the Foreign Trade Institute, INCOMEX, said private businessmen should not become what he called "mental **hostages**" to coffee, traditionally Colombia's **major** export.

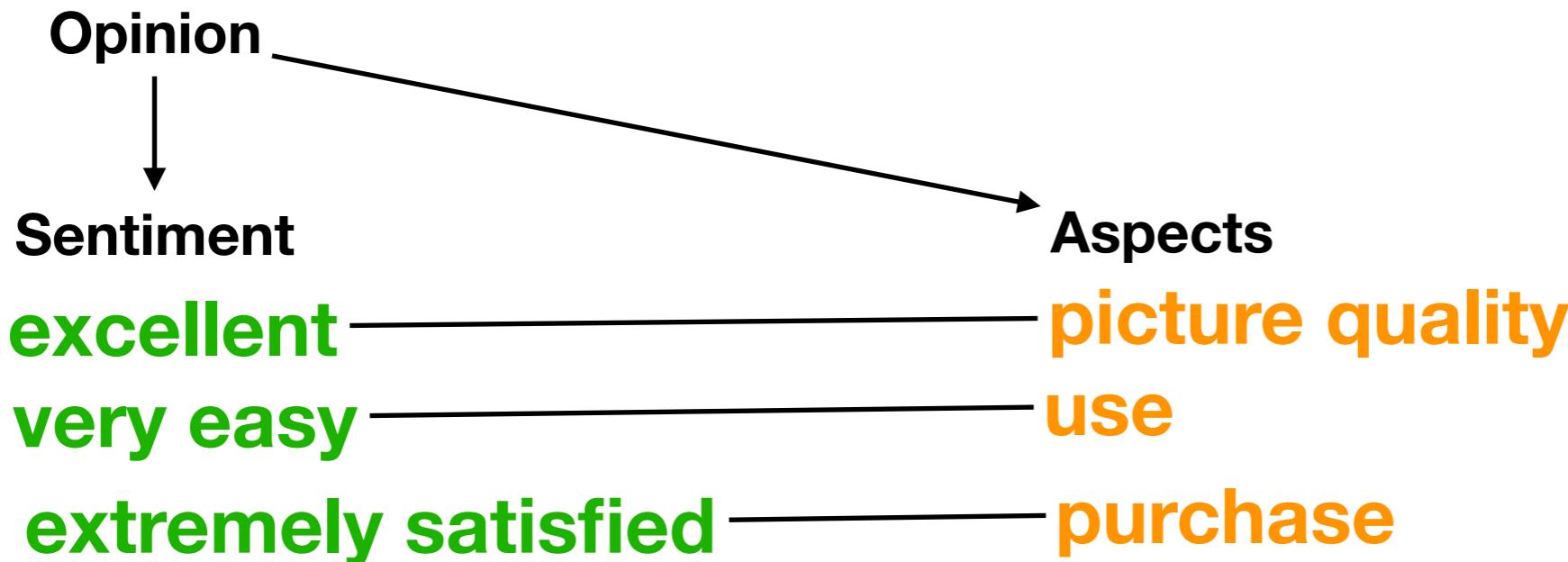
`nltk_data/corpora/reuters/test/15198`



# Product reviews

- [t]excellent picture quality / color canon powershot g3[+3]##i recently purchased the canon powershot g3 and am extremely satisfied with the purchase . use[+2]##the camera is very easy to use , in fact on a recent trip this past week i was asked to take a picture of a vacationing elderly group .

nltk\_data/corpora/product\_reviews\_1/Canon\_G3.txt



# Tweets & Irony?

- {"contributors": null, "coordinates": null, "text": "Everything in the **kids section** of IKEA is **so cute. Shame** I'm nearly 19 in 2 months :(,...}"}

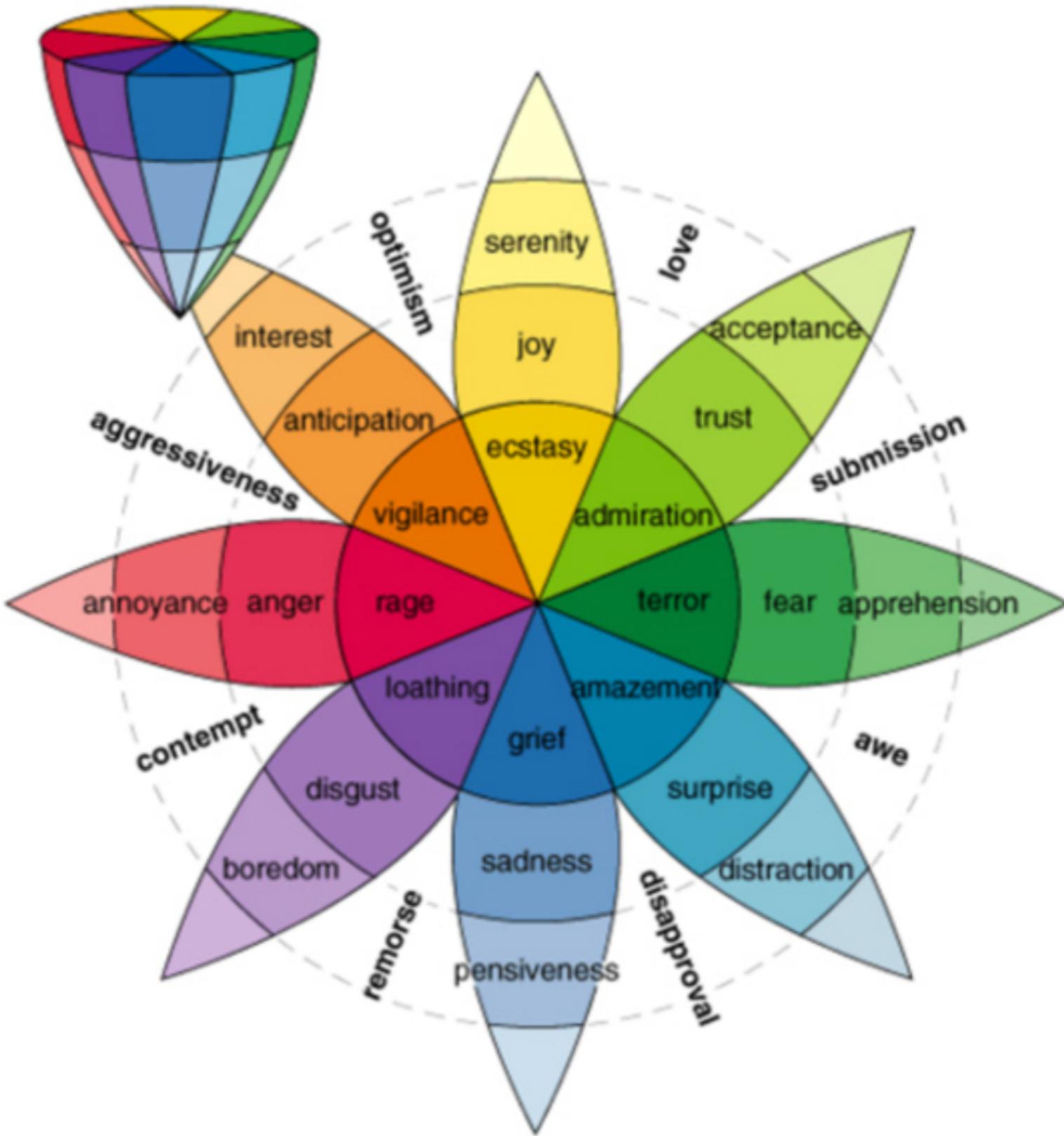
`nltk_data/corpora/twitter_samples/negative_tweets.json`

# Ekman (1972, ..)

- 6 Basic emotions in many different cultures from facial expressions related to described situations
  - ❖ Anger
  - ❖ Disgust
  - ❖ Fear
  - ❖ Happiness
  - ❖ Sadness
  - ❖ (Surprise)
- Extension but not all facially expressed:
  - ❖ Amusement
  - ❖ Contempt
  - ❖ Contentment
  - ❖ Embarrassment
  - ❖ Excitement
  - ❖ Guilt
  - ❖ Pride in achievement
  - ❖ Relief
  - ❖ Satisfaction
  - ❖ Sensory pleasure
  - ❖ Shame



# Plutchik's Wheel of Emotions (1980)



<http://wndomains.fbk.eu/wnaffect.html>

## A-Labels and corresponding example synsets

| A-Labels                    | Examples  |
|-----------------------------|---|
| EMOTION                     | <b>noun anger#1, verb fear#1</b>                      |
| MOOD                        | <b>noun animosity#1, adjective amiable#1</b>          |
| TRAIT                       | <b>noun aggressiveness#1, adjective competitive#1</b> |
| COGNITIVE STATE             | <b>noun confusion#2, adjective dazed#2</b>            |
| PHYSICAL STATE              | <b>noun illness#1, adjective all in#1</b>             |
| HEDONIC SIGNAL              | <b>noun hurt#3, noun suffering#4</b>                  |
| EMOTION-ELICITING SITUATION | <b>noun awkwardness#3, adjective out of danger#1</b>  |
| EMOTIONAL RESPONSE          | <b>noun cold sweat#1, verb tremble#2</b>              |
| BEHAVIOUR                   | <b>noun offense#1, adjective inhibited#1</b>          |
| ATTITUDE                    | <b>noun intolerance#1, noun defensive#1</b>           |
| SENSATION                   | <b>noun coldness#1, verb feel#3</b>                   |

# Relevance

<http://www.internetlivestats.com/one-second>

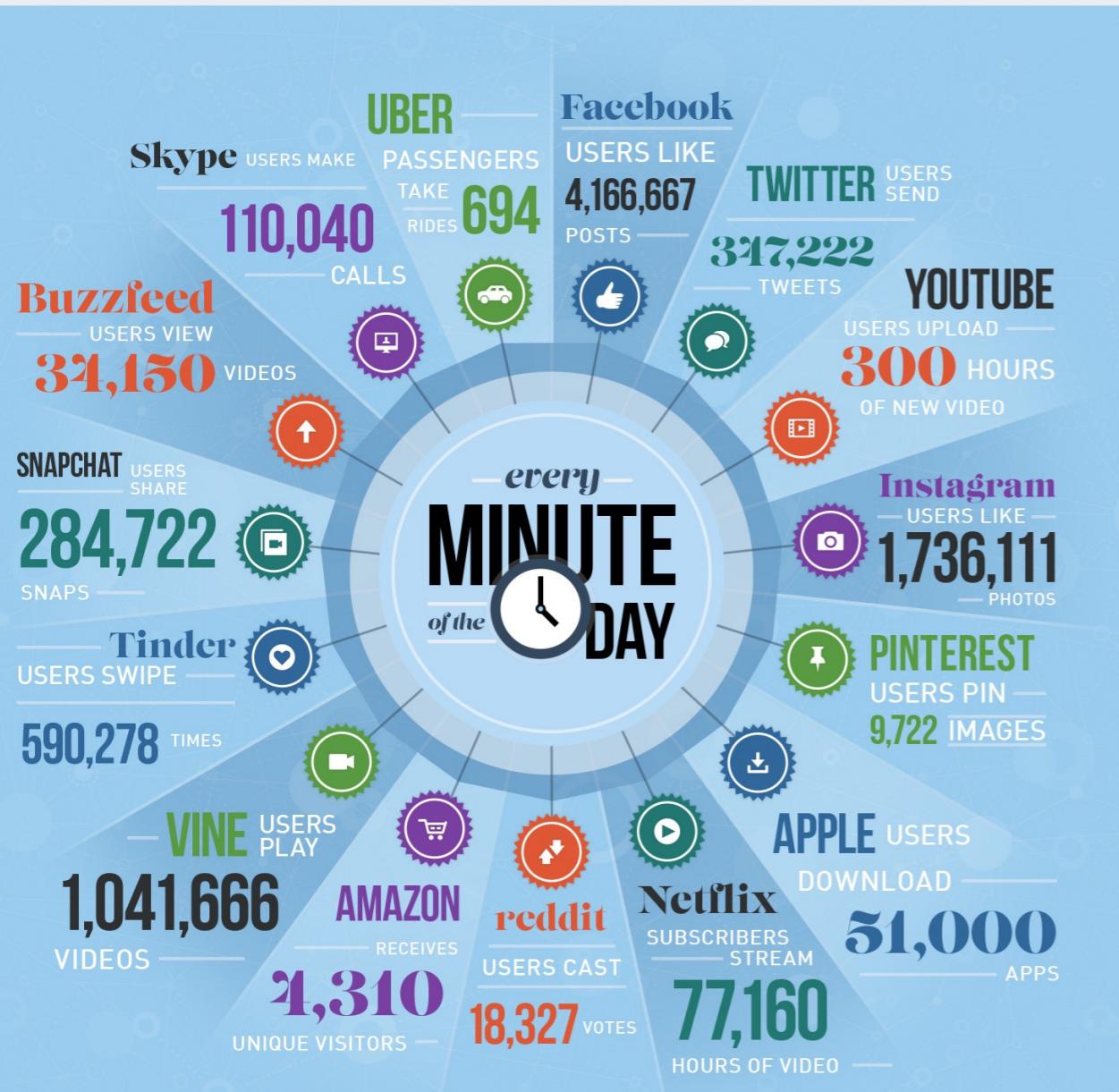




## DATA NEVER SLEEPS 3.0

How much data is generated **every minute**?

Data is being created all the time without us even noticing it. Much of what we do every day now happens in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. Just how much data do our tweets, likes and photo uploads really generate? For the third time, Domo has the answer—and the numbers are staggering.



THE GLOBAL INTERNET POPULATION GREW 18.5% FROM 2013–2015 AND NOW REPRESENTS

**3.2 BILLION PEOPLE.**

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. [Learn more at www.domo.com](#).



SOURCES:

<http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/>

FACEBOOK, TWITTER, YOUTUBE, INSTAGRAM, PINTEREST, APPLE, NETFLIX, REDDIT, AMAZON, TINDER, BUZZFEED, STATISTA, INTERNET LIVE STATS, STATISTICBRAIN.COM

# Social media 2015

- \* **Facebook:** 1.4 billion active monthly users, 4 million likes/minute
- \* **Instagram:** 300 million monthly users, 1.7 million likes/minute
- \* **Vine:** 1 million video's played/minute
- \* **Tinder:** 290K matches/minute
- \* **Twitter:** 347K tweets/minute
- \* **Youtube:** 300 hours upload/minute
- \* **Buzzfeed:** 34K video watches/minute

We do not live in an information society

...

We live in a **communication** society



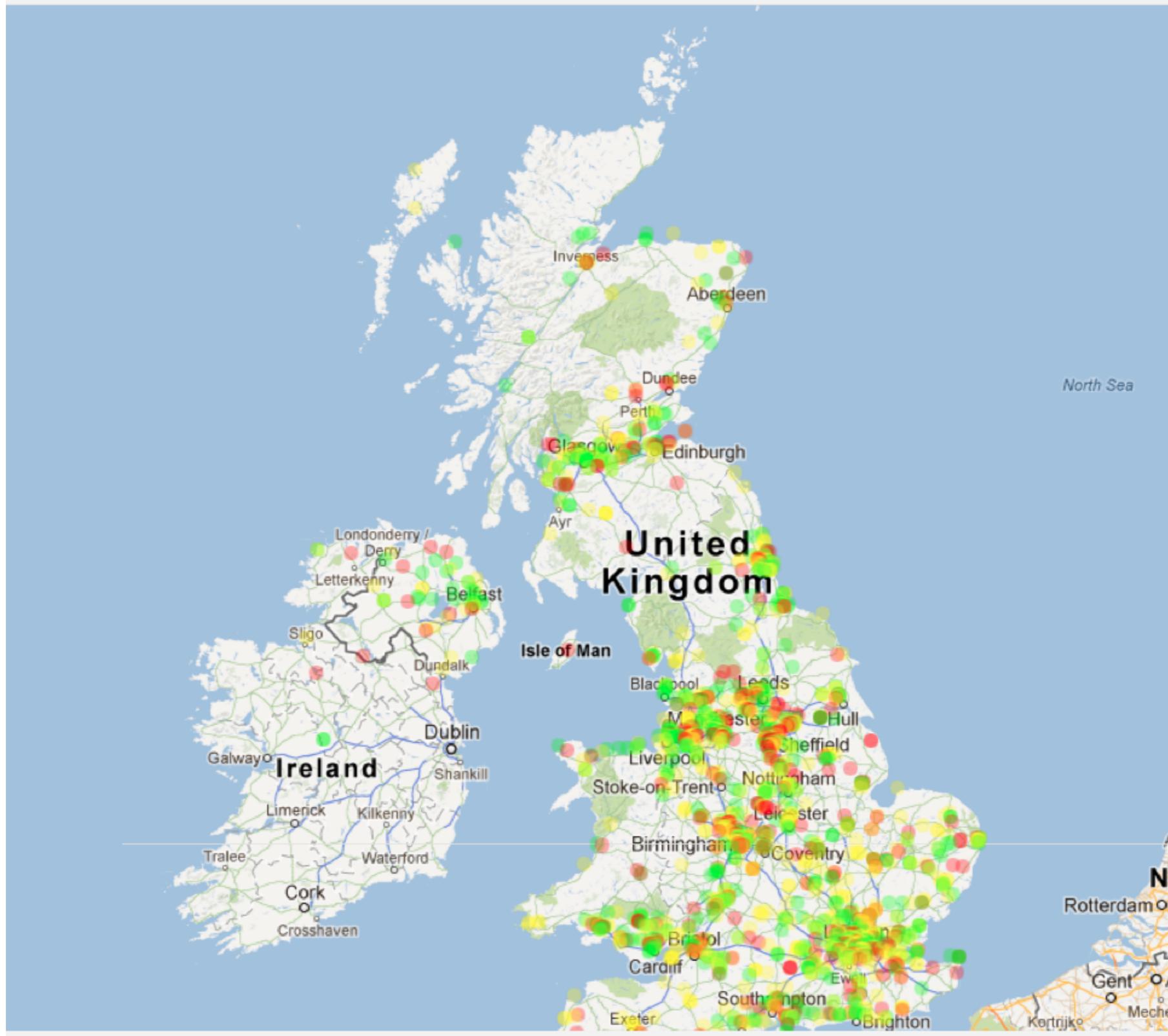
# Sentiment & Stock market



J. Sharma, 7/29 <http://www.csee.umbc.edu/2013/07/ms-defense-sentiment-analysis-on-tweets-and-their-relationship-with-stock-market-trends-j-sharma-729/>

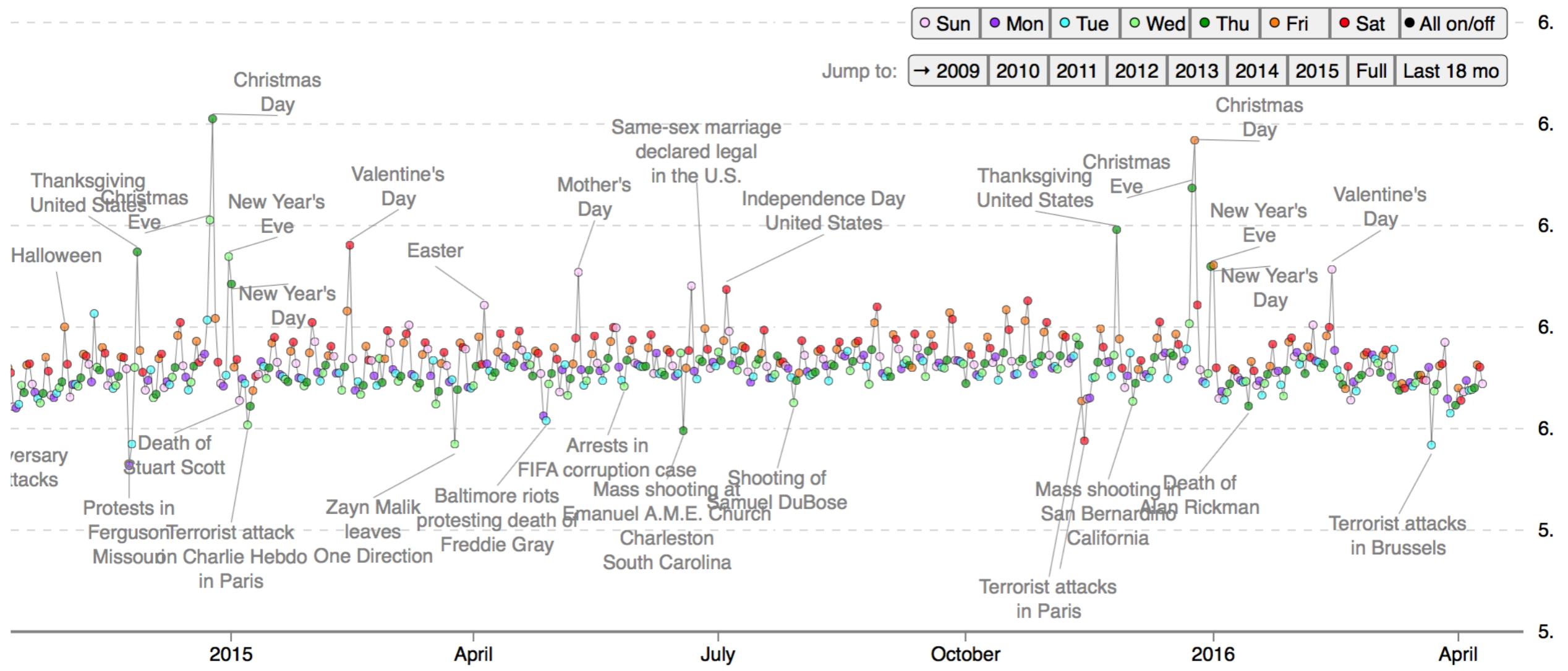
# moodmap

Created for Young Rewired State



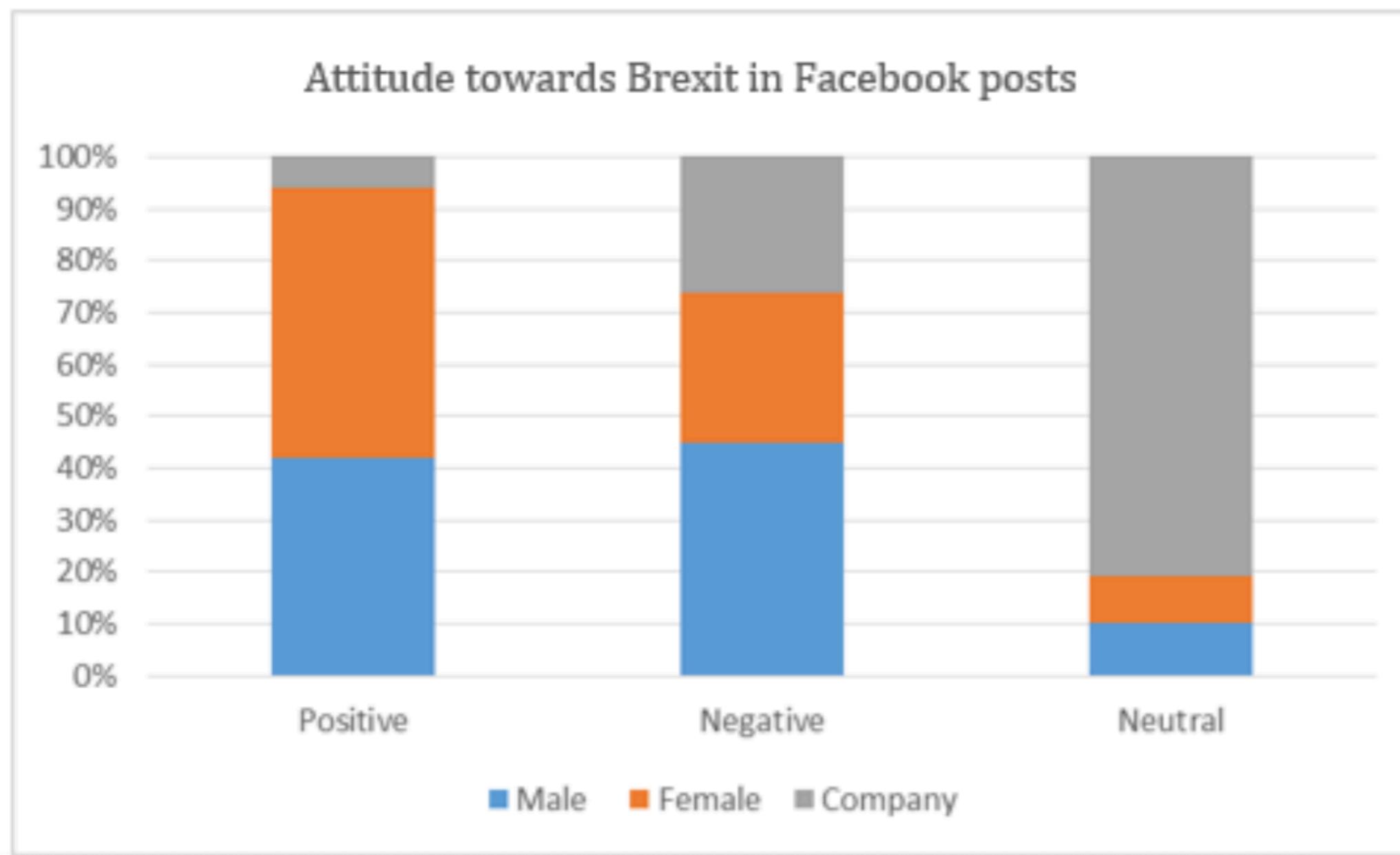
# hedonometer.org

## Average Happiness for Twitter



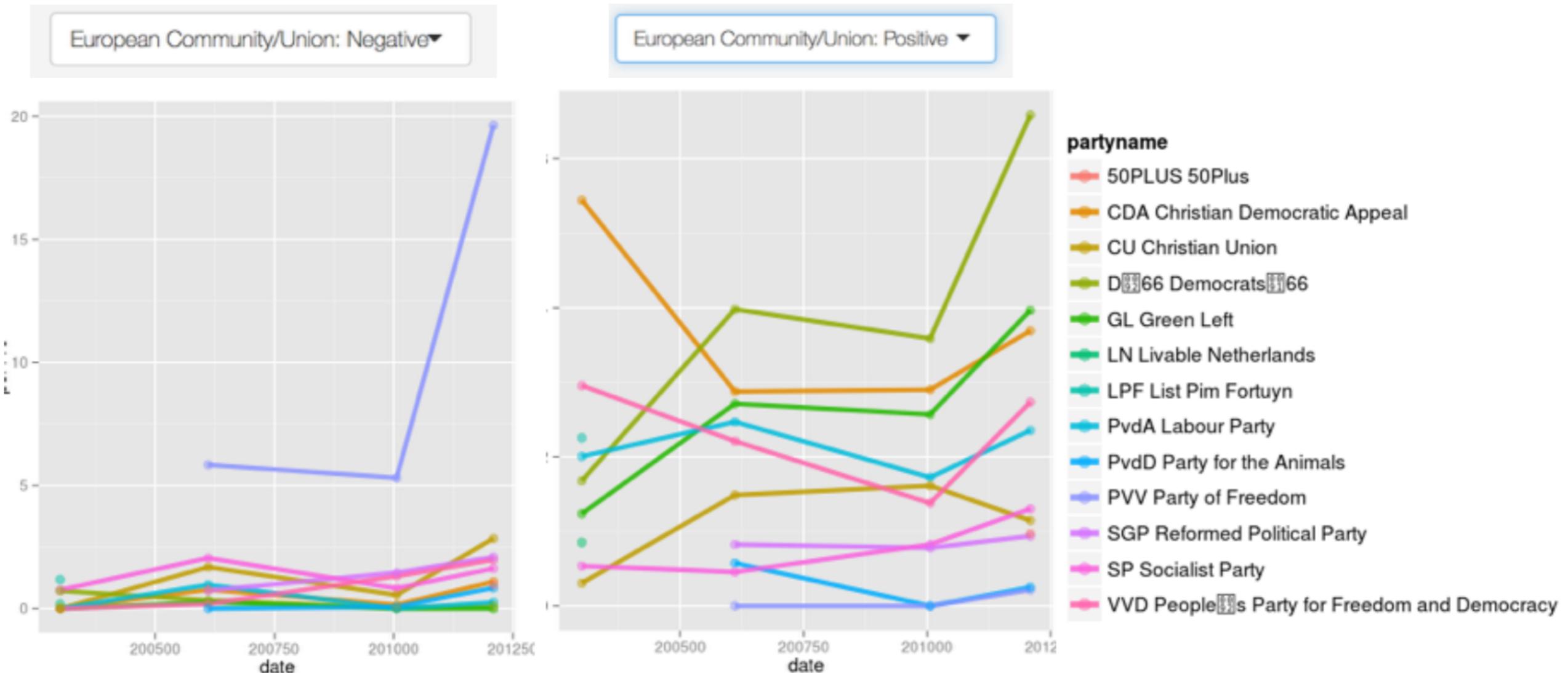
# pro/against Brexit (2016)

Interestingly, females were most likely to post **for** Brexit (68%) followed by men (55%), while companies posted mainly neutral content (71%). 36% of the posts made by men were actually **against** Brexit while only 23% of women voiced a negative opinion.



# The manifesto project:

The Manifesto Project provides the scientific community with parties' policy positions derived from a content analysis of parties' electoral manifestos. It covers over 1000 parties from **1945 until today** in over **50 countries** on five continents



Netherlands 2000-2015

<https://manifesto-project.wzb.eu>

# Love & Hate Letters

348 posts from: lovingyou.com

279 fragments from the millennium project:  
<https://ratbags.com/rssoles/categories.htm>

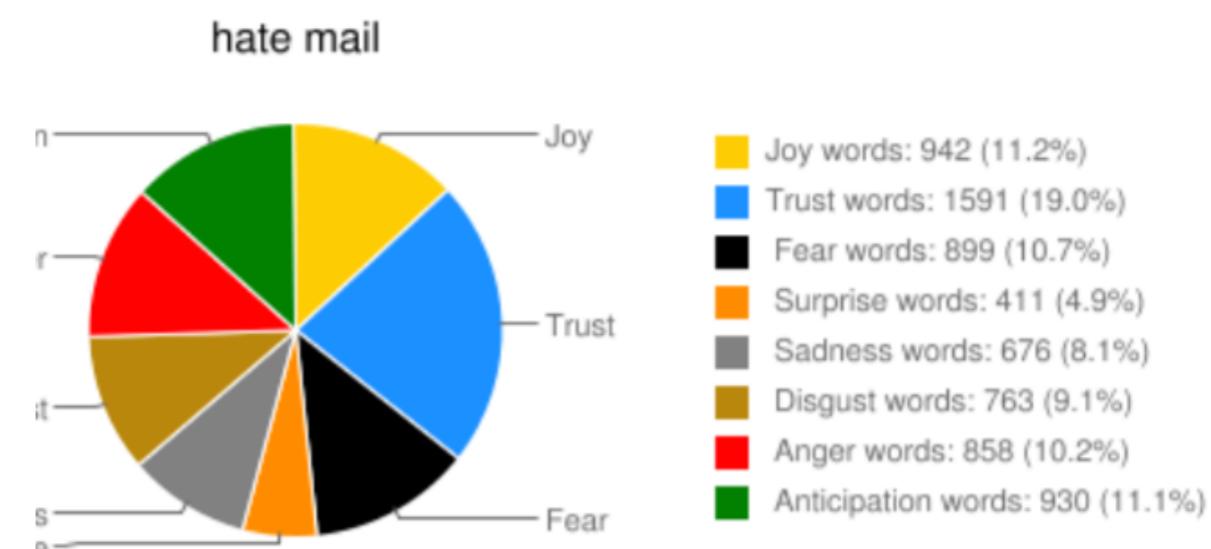
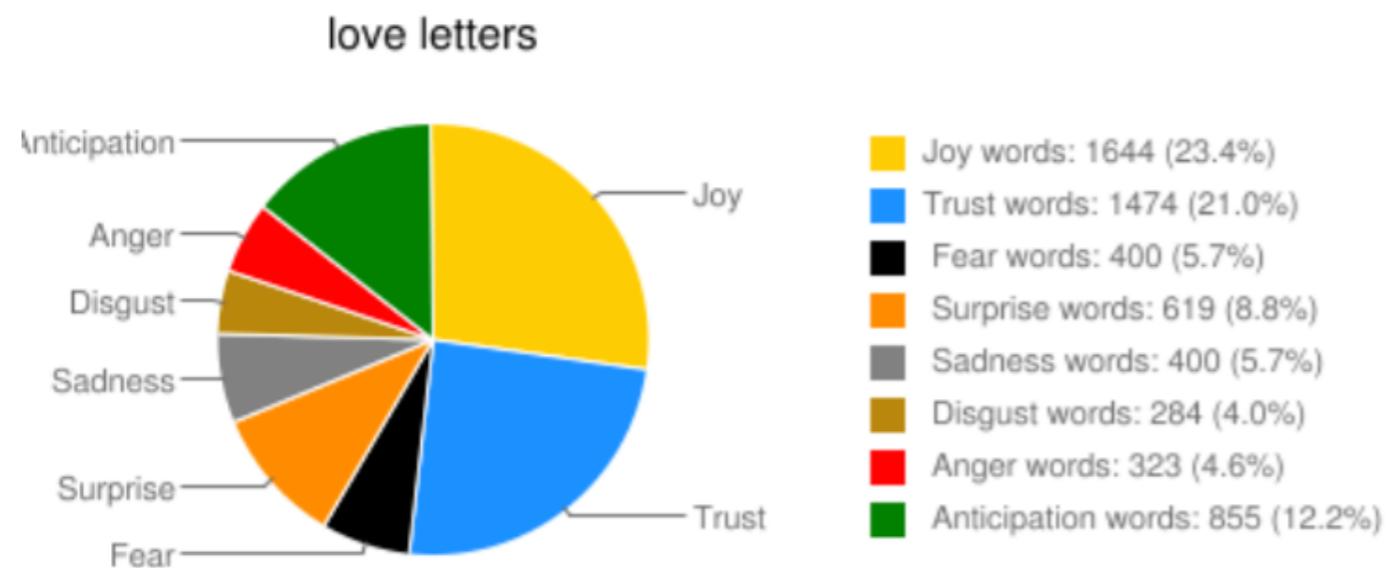


Figure 5: Percentage of emotion words in the love letters corpus.

Figure 6: Percentage of emotion words in the hate mail corpus.

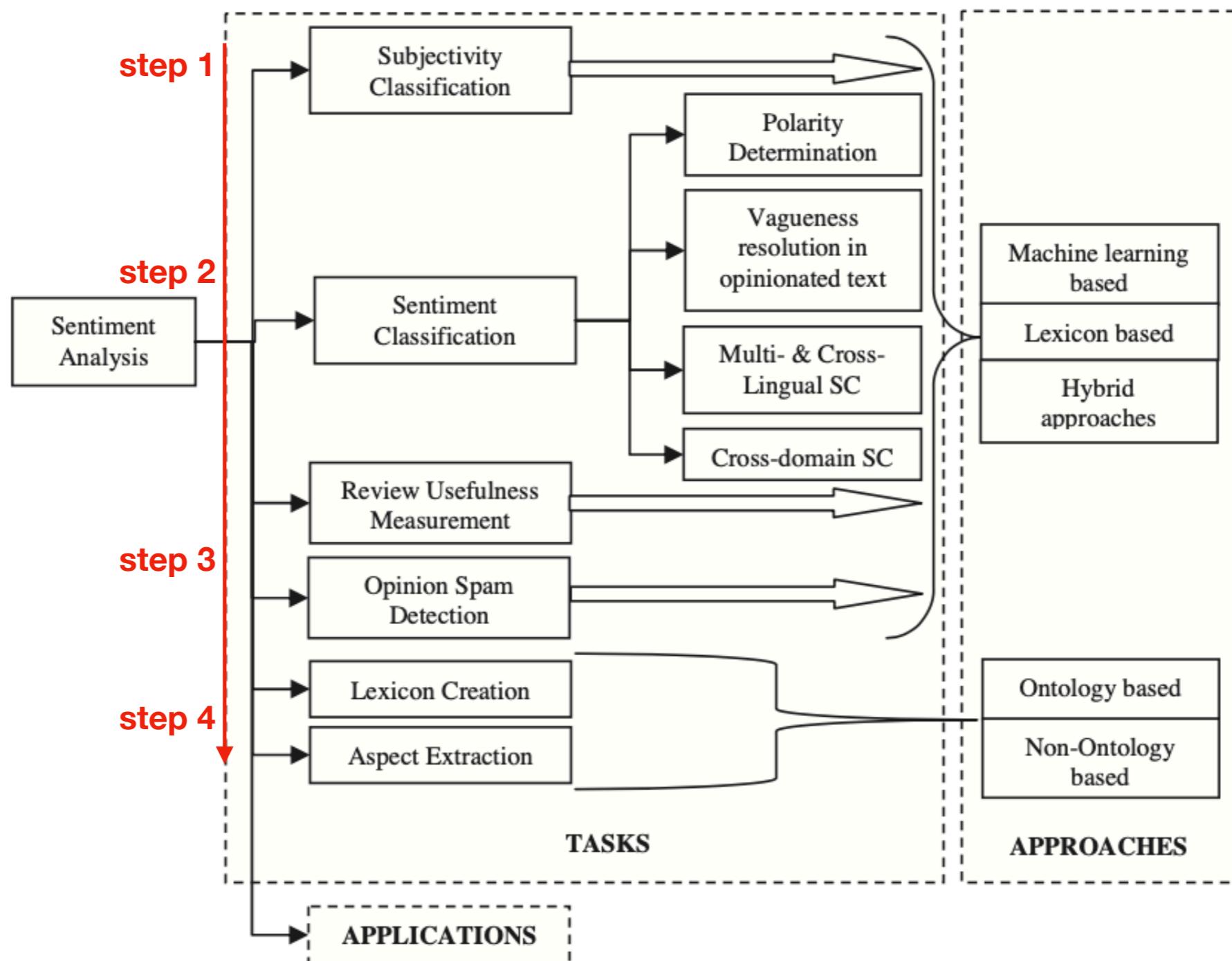
# Terminology

- **Subjectivity or Attitude:** broad term that covers all forms of opinion mining, sentiment analysis, stance but leaves open the lexical or structural realisation
- **Sentiment or Polarity:** explicit expression of being positive, negative or neutral about something/someone (**good, bad**)
- **Opinion:** a lexically or syntactically realised sentiment relation of a holder to a target (I **like** him, my **hero** failed)
- **Stance:** opinion in a debate, I **support** the anti-vaccination movement
- **Aspects-facets-features-properties:** opinion on an aspect of something such as products, (this iPhone has an **incredible good** battery)
- **Argumentation:** I like this iPhone **because** it has a very powerful battery
- **Emotion:** response to a situation that is deemed important: **anger, joy, sadness**
- **Attribution:** relation between a source, a cue (not differentiated) and some content, e.g. quotes
- **Perspective:** umbrella term for all the above as the values for the epistemic, emotional, deontic properties of a **source** with respect to an explicit or implicit **propositional content**
- **Other terms:** inner state, mental state, etc..

# What is Subjectivity Mining?

- Software for automatically extracting **opinions, emotions and sentiment** in text.
- It allows us to track attitudes and feelings on the web. People write **blogs, comments, reviews and tweets** about all sorts of different topics
- We can track products, brands and people for example and determine whether they are **positively or negatively** on the web
- It allows individuals to get **an opinion an a global scale**

# Sentiment analysis (2015)



**Fig. 1.** Organization of the review.

# Levels of Analysis

- Corpus or set of documents
- Document: one sentiment per document
- Sentence/ statement
  - Several opinions, on several topics by several opinion holders per document
  - Opinions on entities, topics or aspects

# subjective vs. non-subjective

- He was born in 1930 in Pennsylvania. He was an American actor best known for his portrayal in the HBO series the Sopranos. He died in New Jersey and was survived by his two sons.
- @USAirways please give Tara G a pat on the back and praise. She was very very helpful.

# Sentiment tagger with interpretation rules

Coen was a person, very modest in life style, of good character, not a drunkard, not haughty, a very proficient council and educated in bookkeeping and business.

# Intensification

- @USAirways please give Tara G a pat on the back and praise. She was very very helpful.
- Please go away!!!!!!
- He was really very helpful

# **Stance detection**

**pro or against or neutral**

**ideological discussions about:**

**legalisation of abortus**

**gay rights**

**brexit**

**vaccinations**

**black pete**

**PRO\_VACCINATION:**

**ANTI\_VACCINATION:**

**I hate anti-vaxxers (negative)**

**Children die from vaccinations!!! (negative)**

# Aspect based sentiment analysis

|                                       | Sentence3       | Sentence4          |
|---------------------------------------|-----------------|--------------------|
| Entity (= opinion target)             | Canon G12       | Canon G12          |
| Aspects (= related to opinion target) | Picture quality | Weight             |
| Sentiment (=opinion expression)       | Positive        | Negative           |
| Opinion holder                        | John Smith      | Wife of John Smith |
| Time                                  | September 2010  | September 2010     |

- (1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) However, my wife thinks it is too heavy for her. (John Smith, september 2010)

# Do we agree on opinions?

- **Hotel reviews**
  - Document level - 3 classes (positive, negative, neutral)
    - 0.87 Kappa
- **Black Pete debate**
  - Tweet level - 4 classes / 3 annotators : 0.65 (a1-a2);  
0.57 (a1-a3); 0.55 (a2-a3) Kappa
- **News annotations**
  - 0.70 Kappa on opinion expressions

| Review text   | User rating | Ann1 | Ann2 |
|---|-------------|------|------|
| The hotel is <b>fantastic</b> , but the area around the hotel is very very noisy.               | 4           | -2   | 3    |
| <b>Good</b> Hotel, in Down Town. Really for business. For holidays I had choose something else. | 4           | -2   | -2   |

# Opinion Annotation

Washington (AFP) - Obama said Thursday that he had included openly gay athletes in the US Olympic delegation to show the

32

United States would not abide discrimination in sport or anywhere else.

His comments, in an interview with NBC on the eve of the Sochi

20

Winter Olympics, came after a senior Russian official activists should not promote gay rights during the Olympics following the passage of controversial anti-gay legislation in Russia.

4

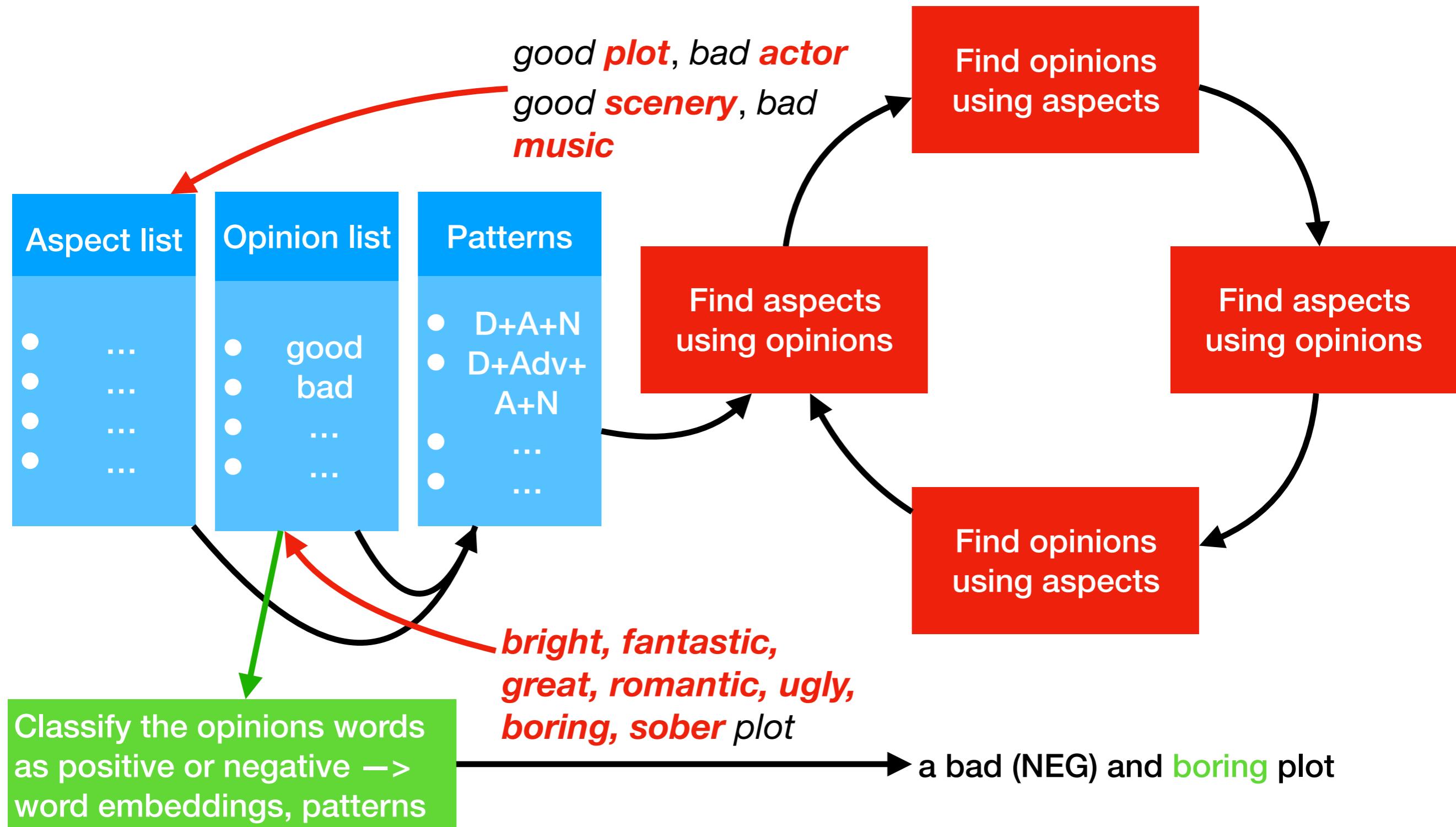
Obama picked several openly gay former athletes in the US delegation to the opening of the games and pointedly did not dispatch a cabinet-level official or a member of his family to Sochi

# Context dependence

- A **cold** person
- A **cold** soda
- A **cold** shower
  - to cool off
  - because the boiler broke (again!)
- Low prices, low ceilings, cheap fast food, cheap fast car

# Learning sentiments & aspects from targeted data sets

## *“Double Propagation”*



| <b>Top 15 ranked aspect terms for restaurants</b> | <b>Top 15 ranked aspect terms for laptops</b> | <b>Top 15 ranked aspect terms for hotels</b> |
|---|---|--|
| <b>1- food</b>                                    | <b>1- battery life</b>                        | <b>1- hotel</b>                              |
| <b>2- service</b>                                 | <b>2- keyboard</b>                            | <b>2- room</b>                               |
| <b>3- staff</b>                                   | <b>3- screen</b>                              | <b>3- staff</b>                              |
| <b>4- bar</b>                                     | <b>4- feature</b>                             | <b>4- service</b>                            |
| <b>5- drink</b>                                   | <b>5- price</b>                               | <b>5- food</b>                               |
| <b>6- table</b>                                   | <b>6- machine</b>                             | <b>6- view</b>                               |
| <b>7- menu</b>                                    | <b>7- toshiba laptop</b>                      | <b>7- stay</b>                               |
| <b>8- dish</b>                                    | <b>8- windows</b>                             | <b>8- breakfast</b>                          |
| <b>9- atmosphere</b>                              | <b>9- performance</b>                         | <b>9- pool</b>                               |
| <b>10- pizza</b>                                  | <b>10- use</b>                                | <b>10- floor</b>                             |
| <b>11- meal</b>                                   | <b>11- battery</b>                            | <b>11- area</b>                              |
| <b>12- bartender</b>                              | <b>12- program</b>                            | <b>12- location</b>                          |
| <b>13- price</b>                                  | <b>13- speaker</b>                            | <b>13- bed</b>                               |
| <b>14- server</b>                                 | <b>14- key</b>                                | <b>14- beach</b>                             |
| <b>15- dinner</b>                                 | <b>15- hard drive</b>                         | <b>15- bar</b>                               |

# Methods

- Rule-based approach, knowledge/lexicon based
  - Identify subjective sentences, exclude objective sentences
  - Count all positive and negative words (from a lexicon) per unit
  - Process negations
  - Process intensifiers: very, extremely, terribly
  - Aggregate sentiment at document level, corpus level, etc. (in order to generate opinions on a generic scale)
- Machine learning
  - need a lot of training data
  - classification problem
  - Decide on the features to use (feature engineering)

# Document level sentiment classification with and without preprocessing

- Ex. (2) De kamers zijn niet schoon en hebben geen eigen badkamer. (*The rooms are not clean and do not have an own bathroom*)
- Rule and lexicon-based: negation flips vote 78.3
- Machine Learning Naive Bayes, 1171 hotel reviews
  - Bag-of-WORDS 82.3
  - Bag-of-words + LEMMA 83.6
  - Bag-of-Words + sentiment lexicon and rules:
    - de kamer zijn niet schoon **neg\_tag** en hebben geen eigen badkamer **negator\_tag**84.7

# Machine Learning methods

**Table 6**

Distribution of articles based on intelligent techniques applied.

| Applied techniques                | #Articles | Articles' references  |
|-----------------------------------|-----------|---|
| SVM                               | 55        | [21,26,29,33,44–46,50,51,53,54,57,58,66,67,73,76,77,86,88,90,91,94,95,97,101,108,109,111,114,116,118,125,131,148,157,160,163,165,167,169,172,176,177,183,195,197,200,209,210,212,214,225,228,240] |
| Dictionary based approaches (DBA) | 41        | [13,18,23–25,35,36,47,55,64,67–69,85,96,110,112,117,126,127,158,170,171,175,183,193–196,202,203,206,207,209,210,213,216,218,220,229,241]  |
| NB                                | 28        | [33,46,50,54,56,73,80,86,90,94,98,101,111,114,116,118,121,124,125,131,148,156,163,167,197,209,217,228]  |
| NN                                | 11        | [48,50–53,57,76,101,116,213,226]  |
| DT                                | 9         | [53,66,73,76,86,94,116,118,209]   |
| Maximum entropy                   | 8         | [33,46,54,60,63,66,148,156,174]   |
| Logistic regression               | 9         | [53,77,88,99,116,118,163,197,220]   |
| Linear regression                 | 8         | [8,18,70,75,222–224,231]  |
| Ontology                          | 8         | [30,41,62,98,182,194–196]   |
| LDA                               | 8         | [61,92,107,185,189,191,182,240]   |
| Random forest                     | 4         | [77,81,228,210,228]   |
| SVR                               | 5         | [118,123,130,155,227]   |
| CRF and rCRP                      | 5         | [37,88,93,186,190]  |
| Boosting                          | 4         | [12,118,179,197]  |
| SVM-SMO                           | 4         | [76,97,118,169]   |
| Fuzzy logic                       | 3         | [23,41,62,213]  |
| Rule miner                        | 4         | [37,100,112,217]  |
| EM                                | 3         | [56,59,155]   |
| K-medoids                         | 1         | [52]  |
| RBF NN                            | 1         | [130]   |

**Table 5**

Sentiment classification accuracy reported on common datasets.

| S# | Dataset              | Articles | Obtained result                                       |
|----|----------------------|----------|---|
| 1  | Pang and Lee [167]   | [156]    | 92.70% accuracy                                       |
| 2  |                      | [112]    | 90.45% F <sub>1</sub>                                 |
| 3  |                      | [169]    | 90.2% accuracy  |
| 4  |                      | [35]     | 89.6% accuracy  |
| 5  |                      | [54]     | 87.70% accuracy                                       |
| 6  |                      | [46]     | 87.4% accuracy  |
| 7  |                      | [50]     | 86.5% accuracy  |
| 8  |                      | [26]     | 85.35% accuracy                                       |
| 9  |                      | [162]    | 81% F <sub>1</sub>                                    |
| 10 |                      | [124]    | 79% accuracy & 86% F <sub>1</sub>                     |
| 11 |                      | [61]     | 76.6% accuracy  |
| 12 |                      | [69]     | 76.37% accuracy                                       |
| 13 |                      | [48]     | 75% precision   |
| 14 |                      | [98]     | 79% precision   |
| 15 | Pang et al. [33]     | [109]    | Approx. 90% accuracy                                  |
| 16 |                      | [165]    | 88.5% accuracy  |
| 17 |                      | [172]    | 87% accuracy  |
| 18 |                      | [33]     | 82.9% accuracy  |
| 19 |                      | [156]    | 78.08% accuracy                                       |
| 20 |                      | [180]    | 75% accuracy  |
| 21 |                      | [48]     | 60% precision   |
| 22 |                      | [195]    | 86.04%  |
| 23 | Blitzer et al. [149] | [45]     | 84.15% accuracy                                       |
| 24 |                      | [99]     | 80.9% (Avg.) accuracy                                 |
| 25 |                      | [54]     | 85.15% (Avg.) Max. 88.65% accuracy on Kitchen reviews |
| 28 |                      | [165]    | 88.7% accuracy  |
| 29 |                      | [61]     | 71.92% accuracy                                       |

# Features for sentiment mining

- **Features:**

- Words (bag-of-words)
- N-grams
- Parts-of-speech (e.g. Adjectives and adjective-adverb combinations)
- Opinion words (lexicon-based: dictionary or corpus)
- Valence intensifiers and shifters (for negation); modal verbs; ...
- Syntactic dependency

- **Feature selection based on**

- frequency
- information gain
- Odds ratio (for binary-class models)
- mutual information

- **Feature weighting**

- Term presence or term frequency
- Inverse document frequency (TF.IDF)
- Term position : e.g. title, first and last sentence(s)

# (Some) Tools

- OpinionFinder: subjective sentences , source (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments: <http://mpqa.cs.pitt.edu>
- Basic sentiment tokenizer plus some tools, by Christopher Potts: <http://sentiment.christopherpotts.net>
- Sentiment for Tweets: VADER (for tweets); <https://github.com/cjhutto/vaderSentiment>
- Linguistic Inquiry and Word Count, Psychological features: <http://liwc.wpengine.com>

# (Some) Tools

- SentiStrength ([sentistrength.wlv.ac.uk](http://sentistrength.wlv.ac.uk))
- TheySay ([apidemo.theysay.io](http://apidemo.theysay.io))
- Sentic ([sentic.net/demo](http://sentic.net/demo))
- Sentdex ([sentdex.com](http://sentdex.com))
- Lexalytics ([lexalytics.com](http://lexalytics.com))
- Sentilo ([wit.istc.cnr.it/stlab-tools/sentilo](http://wit.istc.cnr.it/stlab-tools/sentilo))
- [nlp.stanford.edu/sentiment](http://nlp.stanford.edu/sentiment)

# Lexicons

- Bing Liu's opinion lexicon: <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>
- MPQA subjectivity lexicon: <http://www.cs.pitt.edu/mpqa/>
- SentiWordNet
  - Project homepage: <http://sentiwordnet.isti.cnr.it>
  - Python/NLTK interface: <http://compprag.christopherpotts.net/wordnet.html>
- WordNet Affect: <http://wndomains.fbk.eu/wnaffect.html>
- Harvard General Inquirer: <http://www.wjh.harvard.edu/~inquirer/>
- SenticNet: <http://sentic.net>
- NRC (emotions expressed by words): <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

# Overview

# sentiment datasets (2015)

**Table 9**

List of publicly available datasets.

| S# | Data set                      | Type                         | Lang.     | Web resource  | Details   |
|----|-------------------------------|------------------------------|-----------|---|---|
| 1  | Stanford large movie data set | Movie Reviews                | English   | <a href="http://ai.stanford.edu/~amaas/data/sentiment/">http://ai.stanford.edu/~amaas/data/sentiment/</a>   | Movie Reviews   |
| 2  | COAE2008                      | Product Reviews              | Chinese   | <a href="http://ir-china.org.cn/coae2008.html">http://ir-china.org.cn/coae2008.html</a>   | 2739 documents for movie, education, finance, economics, house, computer, mobile phones, etc. 1525 +ve, 1214 –ve    |
| 3  | Boacar                        | Car Reviews                  | Chinese   | <a href="http://www.riche.com.cn/boacar/">http://www.riche.com.cn/boacar/</a>   | 11 type of car TradeMarks and total review 1000 words, having 578 POS, 428 –ve reviews                              |
| 4  | [187]                         | Reviews, forums              | English   | <a href="http://sifaka.cs.uiuc.edu/~wang296/Data/">http://sifaka.cs.uiuc.edu/~wang296/Data/</a>   | Accessed: 27 August, 2014   |
| 5  | [188]                         | Reviews                      | English   | <a href="http://uilab.kaist.ac.kr/research/WSDM11">http://uilab.kaist.ac.kr/research/WSDM11</a>   | Aspect oriented dataset. Accessed: 18 December, 2014  |
| 6  | Movie-v2.0                    | Movie Reviews                | English   | <a href="http://www.cs.cornell.edu/people/pabo/movie-review-data/">http://www.cs.cornell.edu/people/pabo/movie-review-data/</a>   | Data size: 2000 Positive: 1000 Negative: 1000   |
| 7  | Multi-domain                  | Multi-domain                 | English   | <a href="http://www.cs.jhu.edu/~mdredze/datasets/sentiment">http://www.cs.jhu.edu/~mdredze/datasets/sentiment</a>   |   |
| 8  | SkyDrive de Hermit Dave       | Spanish Word Lists           | Spanish   | <a href="https://skydrive.live.com/?cid=3732e80b128d016f&amp;id=3732E80B128D016F%213584">https://skydrive.live.com/?cid=3732e80b128d016f&amp;id=3732E80B128D016F%213584</a>                                     |   |
| 9  | TripAdvisor                   | Reviews                      | Spanish   | <a href="http://clic.ub.edu/corpus/es/node/106">http://clic.ub.edu/corpus/es/node/106</a>   | 18,000 customer reviews on hotels and restaurants from Hopinion   |
| 10 | [38]                          | Multi-Domain                 | English   | <a href="http://www2.cs.uic.edu/~liub/FBS/sentiment-analysis.html">www2.cs.uic.edu/~liub/FBS/sentiment-analysis.html</a>  | 6800 opinion words on 10 different products   |
| 11 | TBOD [144]                    | Reviews                      | English   |   | Product Review on Cars, Headphones, Hotels  |
| 12 | [68]                          | Product Reviews              | English   | <a href="http://www.lsi.us.es/_fermin/index.php/Datasets">http://www.lsi.us.es/_fermin/index.php/Datasets</a>   | Product Reviews from <a href="#">Epinion.com</a> on headphones 587 reviews, hotels 988 reviews and cars 972 reviews |
| 13 | [148]                         | Movie Reviews                | Turkish   | <a href="http://www.win.tue.nl/~mpechen/projects/smm/#Datasets">http://www.win.tue.nl/~mpechen/projects/smm/#Datasets</a>   | 5331 positive and 5331 negative reviews on movie  |
| 14 | [148]                         | Product Reviews              | Turkish   | <a href="http://www.win.tue.nl/~mpechen/projects/smm/#Datasets">http://www.win.tue.nl/~mpechen/projects/smm/#Datasets</a>   | 700 +ve & 700 –ve reviews on books, DVD, electronics, kitchen appliances  |
| 15 | ISEAR                         | English sentences            | English   | <a href="http://www.affective-sciences.org/system/files/page/2636/ISEAR.zip">www.affective-sciences.org/system/files/page/2636/ISEAR.zip</a>  | The dataset contains 7666 such statements, which include 18,146 sentences, 449,060 running words.                   |
| 16 | [149]                         | Product Reviews              | English   | <a href="http://www.cs.jhu.edu/~mdredze/datasets/sentiment/">http://www.cs.jhu.edu/~mdredze/datasets/sentiment/</a>   | Amazon reviews on 4 domain (books, DVDs, electronics, kitchen appliances)   |
| 17 | DUC data, NIST                | Texts                        | English   | <a href="http://www-nplir.nist.gov/projects/duc/data.html">http://www-nplir.nist.gov/projects/duc/data.html</a> , <a href="http://www.nist.gov/tac/data/index.html">http://www.nist.gov/tac/data/index.html</a> | Text summarization data   |
| 18 | [70]                          | Restaurant and Hotel Reviews | English   | <a href="http://uilab.kaist.ac.kr/research/WSDM11">http://uilab.kaist.ac.kr/research/WSDM11</a>   | Restaurant and Hotel Reviews from Amazon and Yelp   |
| 19 | [114]                         | Restaurant Reviews           | Cantonese | <a href="http://www.openrice.com">http://www.openrice.com</a>   | Reviews on restaurant   |
| 20 | [125]                         | Biographical Articles        | Dutch     | <a href="http://www.iisg.nl/bwsa">http://www.iisg.nl/bwsa</a>   | 574 Biographical articles   |
| 21 | Spinn3r dataset               | Multi-Domain                 | English   | <a href="http://www.icwsm.org/2011/data.php">http://www.icwsm.org/2011/data.php</a>   |   |
| 22 | [86]                          | Ironic Dataset               | English   | <a href="http://users.dsic.upv.es/grupos/nle/">http://users.dsic.upv.es/grupos/nle/</a>   | 3163 ironic reviews on five products  |
| 23 | HASH [179]                    | Tweets                       | English   | <a href="http://demeter.inf.ed.ac.uk">http://demeter.inf.ed.ac.uk</a>   | 31,861 Pos tweets, 64,850 Neg tweets, 125,859 Neu tweets  |
| 24 | EMOT [179]                    | Tweets and Emoticons         | English   | <a href="http://twittersentiment.appspot.com">http://twittersentiment.appspot.com</a>   | 230,811 Pos & 150,570 Neg tweets  |
| 25 | ISIEVE [179]                  | Tweets                       | English   | <a href="http://www.i-sieve.com">www.i-sieve.com</a>  | 1520 Pos tweets, 200 Neg tweets, 2295 Neu tweets  |
| 26 | [177]                         | Tweets                       | English   | e-mail: apoovr@cs.columbia.edu  | 11,875 tweets   |
| 27 | [52]                          | Opinions                     | English   | <a href="http://patientopinion.org.uk">http://patientopinion.org.uk</a>   | 2000 patient opinions   |
| 28 | [96]                          | Tweets                       | English   | <a href="http://goo.gl/UQvdX">http://goo.gl/UQvdX</a>   | 667 tweets  |
| 29 | [39]                          | Movie Reviews                | English   | <a href="http://ai.stanford.edu/~amaas/data/sentiment/">http://ai.stanford.edu/~amaas/data/sentiment/</a>   | 50,000 movie reviews  |
| 30 | [164]                         | Tweets                       | English   | <a href="http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip">http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip</a>   |   |
| 31 | [210]                         | Spam Reviews                 | English   | <a href="http://myleott.com/op_spam">http://myleott.com/op_spam</a>   | 400 deceptive and 400 truthful reviews in positive and negative category. Last Accessed by: 12 April, 2015          |
| 32 | [230]                         | Sarcasm and nasty reviews    | English   | <a href="https://nlds.soe.ucsc.edu/iac">https://nlds.soe.ucsc.edu/iac</a>   | 1000 discussions, ~390,000 posts, and some ~73,000,000 words  |

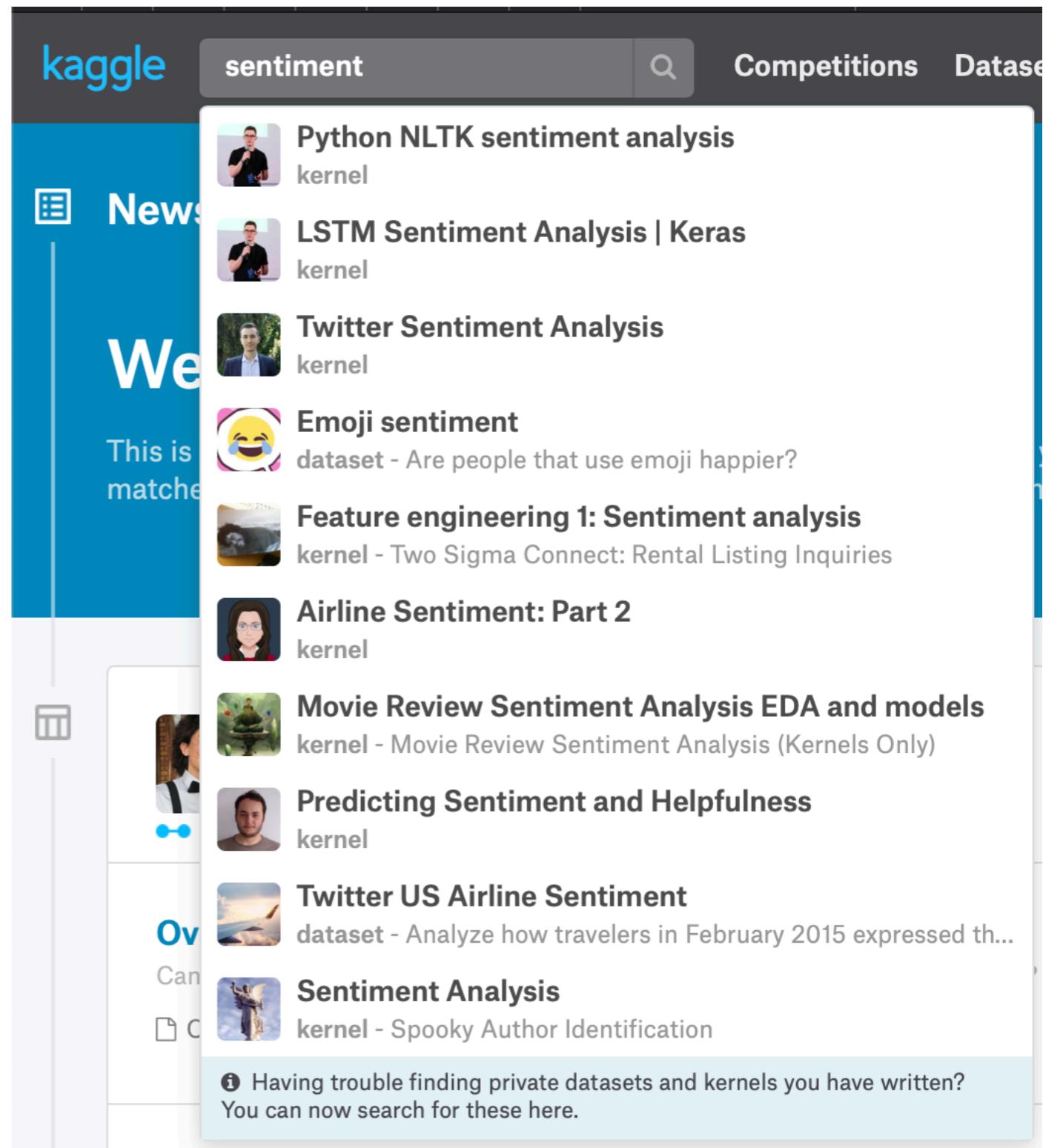
# (Some) datasets

- Data from Lillian Lee's group: <http://www.cs.cornell.edu/home/llee/data/>
- Data from Bing Liu: <http://www.cs.uic.edu/~liub/>
- Large movie review dataset: <http://ai.stanford.edu/~amaas/data/sentiment/>
- Pranav Anand & co. (<http://people.ucsc.edu/~panand/data.php>)
  - Internet Argument Corpus
  - Annotated political TV ads
  - Focus of negation corpus
  - Persuasion corpus (blogs)
- Data on AFS:
  - `/afs/ir/data/linguistic-data/mnt/mnt4/PottsCorpora`  
`README.txt`, `Twitter.tgz`, `imdb-english-combined.tgz`,  
`opentable-english-processed.zip`
  - `/afs/ir/data/linguistic-data/mnt/mnt9/PottsCorpora`  
`opposingviews`, `product-reviews`, `weblogs`
- Twitter data collected and organized by Moritz!  
[/afs.ir.stanford.edu/data/linguistic-data/mnt/mnt3/TwitterTopics/](http://afs.ir.stanford.edu/data/linguistic-data/mnt/mnt3/TwitterTopics/)

From Potts (2013), p.5

# More datasets

- SNAP review datasets: <http://snap.stanford.edu/data/>
- Yelp dataset: [http://www.yelp.com/dataset\\_challenge/](http://www.yelp.com/dataset_challenge/)
- More on Twitter datasets, including critical appraisal:  
Saif et al. (2013)
- Kaggle: <https://www.kaggle.com/>



# [https://www.kaggle.com/ search?q=sentiment](https://www.kaggle.com/search?q=sentiment)

← [sentiment](#)

---

Searching for sentiment within

Comments 1,888 Notebooks 1,420 Topics 519 Datasets 323 Competitions 160 Blogs 3

---

Filter by

4,313 Results Sort by: Relevancy ▾

---

**Date**

- Last 90 days 451
- Last week 39
- Today 5

**Dataset Size**

- small 160
- medium 156
- large 7

**Dataset File Types**

- csv 225
- txt 55
- tsv 32

---

 < Notebook  
**Python NLTK sentiment analysis**  
by Peter Nagy  
a year ago • 9m to run • Python • ^ 159  
Python NLTK [sentiment](#) analysis

---

 < Notebook  
**LSTM Sentiment Analysis | Keras**  
by Peter Nagy  
a year ago • 2m to run • Python • ^ 117  
LSTM [Sentiment](#) Analysis | Keras

---

 < Notebook  
**Twitter Sentiment Analysis**  
by Dario Dianomenti

# [https://www.kaggle.com/ search?q=opinion](https://www.kaggle.com/search?q=opinion)

← [opinion](#)

Searching for opinion within

Comments 2,915

Topics 480

Notebooks 267

Datasets 73

Competitions 4

Filter by

3,739 Results

Sort by: Relevancy ▾

#### Date

- Last 90 days 454
- Last week 28
- Today 5

#### Dataset Size

- small 51
- medium 21
- large 1

#### Dataset File Types

- csv 49
- pdf 10
- json 9



Notebook

### Political Opinion with Machine Learning

by Yunus Emre Gündoğmuş

2 years ago • 27s to run • Python • ^ 49

Political [Opinion](#) with Machine Learning



Notebook

### Comprehensive data exploration with Python

by Pedro Marcelino

3 months ago • 28s to run • Python • ^ 5920

In my [opinion](#), this heatmap is the best way to get a quick overview of our 'plasma soup' and its relationships



Dataset

### Deceptive Opinion Spam Corpus

# Google dataset search

<https://toolbox.google.com/datasetsearch>

The screenshot shows the Google Dataset Search interface. At the top, there's a search bar with the query "sentiment". Below the search bar, it says "100+ results found". On the left, there's a list of datasets:

- Sentiment Analysis Single Word (data.world, updated May 22, 2018)
- Sentiment Analysis in Text (data.world, updated Oct 24, 2019)
- Weather sentiment (data.world, updated Oct 28, 2019)
- Economic sentiment indicator (data.europa.eu, db.nomics.world, updated Nov 15, 2019)
- VnEmoLex: A Vietnamese emotion lexicon for sentiment... (zenodo.org, search.datacite.org, Published Jun 1, 2017)

On the right, a detailed view of the "Emoji Sentiment Ranking" dataset is shown. It includes:

- Emoji Sentiment Ranking**
- [Explore at figshare.com](#)
- 63 scholarly articles cite this dataset ([View in Google Scholar](#))
- Unique identifier**: <https://doi.org/10.6084/m9.figshare.1600931.v1>
- Dataset created**: Nov 12, 2015
- Dataset published**: Nov 12, 2015
- Dataset updated**: Jan 20, 2016
- Dataset provided by**: figshare
- Authors**: Igor Mozetic; Petra Kralj Novak; Jasmina Smailović; Borut Sluban
- License**: [Attribution 4.0 \(CC BY 4.0\)](#)
- License information was derived automatically
- Available download formats from providers**: txt, s, gif, html
- Description**: A lexicon of 751 emoji characters with automatically assigned sentiment. The sentiment is computed from 70,000 tweets, labeled by 83 human annotators in 13 European languages. The Emoji Sentiment Ranking web page at [http://kt.ijs.si/data/Emoji\\_sentiment\\_ranking/](http://kt.ijs.si/data/Emoji_sentiment_ranking/) is automatically generated from the data provided in this repository. The process and analysis of emoji sentiment ranking is described in the paper: P. Kralj Novak, J. Smailović, B. Sluban, I. Mozetič, Sentiment of Emojis, submitted; arXiv preprint, <http://arxiv.org/abs/1509.07761>, 2015.

# Google dataset search

<https://toolbox.google.com/datasetsearch>

Google Dataset Search

 opinion



100+ results found



The Opinions of the Committee of the Regions

[data.europa.eu](http://data.europa.eu)  
[data.wu.ac.at](http://data.wu.ac.at)

Updated Jul 25, 2019

Consumer Opinion Surveys: Consumer Prices: Future Tendency of Inflation: European Commission and National Indicators for Spain  
CSINFT02ESM460S

[Explore at FRED](#)

Dataset updated Nov 13, 2019

License

[https://research.stlouisfed.org/fred\\_terms.html#copyright-citation-required](https://research.stlouisfed.org/fred_terms.html#copyright-citation-required)

Description

Graph and download economic data for Consumer Opinion Surveys: Consumer Prices: Future Tendency of Inflation: European Commission and National Indicators for Spain (CSINFT02ESM460S) from Jun 1986 to Oct 2019 about consumer sentiment, consumer prices, Spain, consumer, and inflation.



Opinion of the French on the relevance of their research onlin...

[www.statista.com](http://www.statista.com)



European Data Protection Supervisor opinions

[data.europa.eu](http://data.europa.eu)

Updated Jan 25, 2018

**Table 9**

List of publicly available datasets.

| S# | Data set                      | Type                         | Lang.     | Web resource  | Details   |
|----|-------------------------------|------------------------------|-----------|---|---|
| 1  | Stanford large movie data set | Movie Reviews                | English   | <a href="http://ai.stanford.edu/~amaas/data/sentiment/">http://ai.stanford.edu/~amaas/data/sentiment/</a>   | Movie Reviews   |
| 2  | COAE2008                      | Product Reviews              | Chinese   | <a href="http://ir-china.org.cn/coae2008.html">http://ir-china.org.cn/coae2008.html</a>   | 2739 documents for movie, education, finance, economics, house, computer, mobile phones, etc. 1525 +ve, 1214 -ve    |
| 3  | Boacar                        | Car Reviews                  | Chinese   | <a href="http://www.riche.com.cn/boacar/">http://www.riche.com.cn/boacar/</a>   | 11 type of car TradeMarks and total review 1000 words, having 578 POS, 428 reviews                                  |
| 4  | [187]                         | Reviews, forums              | English   | <a href="http://sifaka.cs.uiuc.edu/~wang296/Data/">http://sifaka.cs.uiuc.edu/~wang296/Data/</a>   | Accessed: 27 August, 2014   |
| 5  | [188]                         | Reviews                      | English   | <a href="http://uilab.kaist.ac.kr/research/WSDM11">http://uilab.kaist.ac.kr/research/WSDM11</a>   | Aspect oriented dataset Accessed: 18 December, 2014   |
| 6  | Movie-v2.0                    | Movie Reviews                | English   | <a href="http://www.cs.cornell.edu/people/pabo/movie-review-data/">http://www.cs.cornell.edu/people/pabo/movie-review-data/</a>   | Data size: 2000 Positive: 1000 Negative: 1000   |
| 7  | Multi-domain                  | Multi-domain                 | English   | <a href="http://www.cs.jhu.edu/~mdredze/datasets/sentiment">http://www.cs.jhu.edu/~mdredze/datasets/sentiment</a>   |   |
| 8  | SkyDrive de Hermit Dave       | Spanish Word Lists           | Spanish   | <a href="https://skydrive.live.com/?cid=3732e80b128d016f&amp;id=3732E80B128D016F%213584">https://skydrive.live.com/?cid=3732e80b128d016f&amp;id=3732E80B128D016F%213584</a>                                     |   |
| 9  | TripAdvisor                   | Reviews                      | Spanish   | <a href="http://clic.ub.edu/corpus/es/node/106">http://clic.ub.edu/corpus/es/node/106</a>   | 18,000 customer reviews on hotels and restaurants from Hopinion   |
| 10 | [38]                          | Multi-Domain                 | English   | <a href="http://www2.cs.uic.edu/~liub/FBS/sentiment-analysis.html">www2.cs.uic.edu/~liub/FBS/sentiment-analysis.html</a>  | 6800 opinion words on 10 different products   |
| 11 | TBOD [144]                    | Reviews                      | English   |   | Product Review on Cars, Headphones, Hotels  |
| 12 | [68]                          | Product Reviews              | English   | <a href="http://www.lsi.us.es/_fermin/index.php/Datasets">http://www.lsi.us.es/_fermin/index.php/Datasets</a>   | Product Reviews from <a href="#">Epinion.com</a> on headphones 587 reviews, hotels 988 reviews and cars 972 reviews |
| 13 | [148]                         | Movie Reviews                | Turkish   | <a href="http://www.win.tue.nl/~mpechen/projects/smm/#Datasets">http://www.win.tue.nl/~mpechen/projects/smm/#Datasets</a>   | 5331 positive and 5331 negative reviews on movie  |
| 14 | [148]                         | Product Reviews              | Turkish   | <a href="http://www.win.tue.nl/~mpechen/projects/smm/#Datasets">http://www.win.tue.nl/~mpechen/projects/smm/#Datasets</a>   | 700 +ve & 700 -ve reviews on books, DVD, electronics, kitchen appliances  |
| 15 | ISEAR                         | English sentences            | English   | <a href="http://www.affective-sciences.org/system/files/page/2636/ISEAR.zip">www.affective-sciences.org/system/files/page/2636/ISEAR.zip</a>  | The dataset contains 7666 such statements, which include 18,146 sentences, 449,060 running words.                   |
| 16 | [149]                         | Product Reviews              | English   | <a href="http://www.cs.jhu.edu/~mdredze/datasets/sentiment/">http://www.cs.jhu.edu/~mdredze/datasets/sentiment/</a>   | Amazon reviews on 4 domain (books, DVDs, electronics, kitchen appliances)   |
| 17 | DUC data, NIST                | Texts                        | English   | <a href="http://www-nlpri.nist.gov/projects/duc/data.html">http://www-nlpri.nist.gov/projects/duc/data.html</a> , <a href="http://www.nist.gov/tac/data/index.html">http://www.nist.gov/tac/data/index.html</a> | Text summarization data   |
| 18 | [70]                          | Restaurant and Hotel Reviews | English   | <a href="http://uilab.kaist.ac.kr/research/WSDM11">http://uilab.kaist.ac.kr/research/WSDM11</a>   | Restaurant and Hotel Reviews from Amazon and Yelp   |
| 19 | [114]                         | Restaurant Reviews           | Cantonese | <a href="http://www.openrice.com">http://www.openrice.com</a>   | Reviews on restaurant   |
| 20 | [125]                         | Biographical Articles        | Dutch     | <a href="http://www.iisg.nl/bwsa">http://www.iisg.nl/bwsa</a>   | 574 Biographical articles   |
| 21 | Spinn3r dataset               | Multi-Domain                 | English   | <a href="http://www.icwsm.org/2011/data.php">http://www.icwsm.org/2011/data.php</a>   |   |
| 22 | [86]                          | Ironic Dataset               | English   | <a href="http://users.dsic.upv.es/grupos/nle/">http://users.dsic.upv.es/grupos/nle/</a>   | 3163 ironic reviews on five products  |
| 23 | HASH [179]                    | Tweets                       | English   | <a href="http://demeter.inf.ed.ac.uk">http://demeter.inf.ed.ac.uk</a>   | 31,861 Pos tweets, 64,850 Neg tweets, 125,859 Neu tweets  |
| 24 | EMOT [179]                    | Tweets and Emoticons         | English   | <a href="http://twittersentiment.appspot.com">http://twittersentiment.appspot.com</a>   | 230,811 Pos & 150,570 Neg tweets  |
| 25 | ISIEVE [179]                  | Tweets                       | English   | <a href="http://www.i-sieve.com">www.i-sieve.com</a>  | 1520 Pos tweets, 200 Neg tweets, 2295 Neu tweets  |
| 26 | [177]                         | Tweets                       | English   | e-mail: <a href="mailto:apoory@cs.columbia.edu">apoory@cs.columbia.edu</a>  | 11,875 tweets   |
| 27 | [52]                          | Opinions                     | English   | <a href="http://patientopinion.org.uk">http://patientopinion.org.uk</a>   | 2000 patient opinions   |
| 28 | [96]                          | Tweets                       | English   | <a href="http://goo.gl/UQvdX">http://goo.gl/UQvdX</a>   | 667 tweets  |
| 29 | [39]                          | Movie Reviews                | English   | <a href="http://ai.stanford.edu/~amaas/data/sentiment/">http://ai.stanford.edu/~amaas/data/sentiment/</a>   | 50,000 movie reviews  |
| 30 | [164]                         | Tweets                       | English   | <a href="http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip">http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip</a>   |   |
| 31 | [210]                         | Spam Reviews                 | English   | <a href="http://myleott.com/op_spam">http://myleott.com/op_spam</a>   | 400 deceptive and 400 truthful reviews in positive and negative category. Last Accessed by: 12 April, 2015          |
| 32 | [230]                         | Sarcasm and nasty reviews    | English   | <a href="https://nlds.soe.ucsc.edu/iac">https://nlds.soe.ucsc.edu/iac</a>   | 1000 discussions, ~390,000 posts, and some ~73,000,000 words  |

# Survey Literature

- Ronen Feldman: Techniques and applications for sentiment analysis. Commun. ACM 56(4): 82-89 (2013).
- Bing Liu, Lei Zhang: A Survey of Opinion Mining and Sentiment Analysis. Mining Text Data 2012: 415-463.
- Bo Pang, Lillian Lee: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2): 1-135 (2007).
- Potts (2013). Introduction to Sentiment Analysis. <http://www.stanford.edu/class/cs224u/slides/2013/cs224u-slides-02-26.pdf>
- Mikalai Tsytsarau, Themis Palpanas: Survey on mining subjective data on the web. Data Min. Knowl. Discov. 24(3): 478-514 (2012)
- Saif M. Mohammad, (2012) From once upon a time to happily ever after: Tracking emotions in mail and books, Decision Support Systems 53 (2012) 730–741
- Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.

# The OpeneR project

- <https://www.opener-project.eu>
- Opinion detection and Named Entity recognition for 6 European languages
- Sentiment lexicons for 6 European languages
- Demonstrator: <http://tour-pedia.org/about/>

# A hotel review

Nothing special really. Comfortable and clean but very boring decor in comparison to other NH hotels. I stayed in NH in Brussels and Zurich and I really liked them because of their modern and stylish design and big rooms. This one was just like any other hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and 20 euros for breakfast!!! It was good but way overpriced! The best thing about the hotel was the location - city centre, 2min from a metro stop.



# A hotel review

Nothing special really. Comfortable and clean but very boring decor in comparison to **other NH hotels**. I stayed in **NH** in **Brussels** and **Zurich** and I really liked **them** because of **their** modern and stylish design and big rooms. **This one** was just like any **other** hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and **20 euros** for breakfast!!! It was good but way overpriced! The best thing about **the hotel** was **the location** - city centre, 2min from a metro stop.

## Named entities

- this NH hotel
  - NH Brussels
  - NH Zurich
- [http://en.wikipedia.org/wiki/NH\\_Hoteles](http://en.wikipedia.org/wiki/NH_Hoteles)  
[http://dbpedia.org/page/NH\\_Hoteles](http://dbpedia.org/page/NH_Hoteles)  
<http://en.wikipedia.org/wiki/Brussels>  
<http://en.wikipedia.org/wiki/Zurich>

# A hotel review

- Nothing special really. Comfortable and clean but very boring decor in comparison to **other NH hotels**. I stayed in **NH** in **Brussels** and **Zurich** and I really liked **them** because of **their modern and stylish** design and big rooms. **This one** was just like any other hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and **20 euros** for breakfast!!! It was good but way overpriced! The best thing about **the hotel** was **the location** - city centre, 2min from a metro stop.

Named entities:

- this NH hotel
- NH Brussels
- NH Zurich

Co-references:

- other
- them
- this one

Properties:

- décor
- design
- room
- clean (hygiene)

- service
- rate
- Internet
- breakfast

Sentiments:

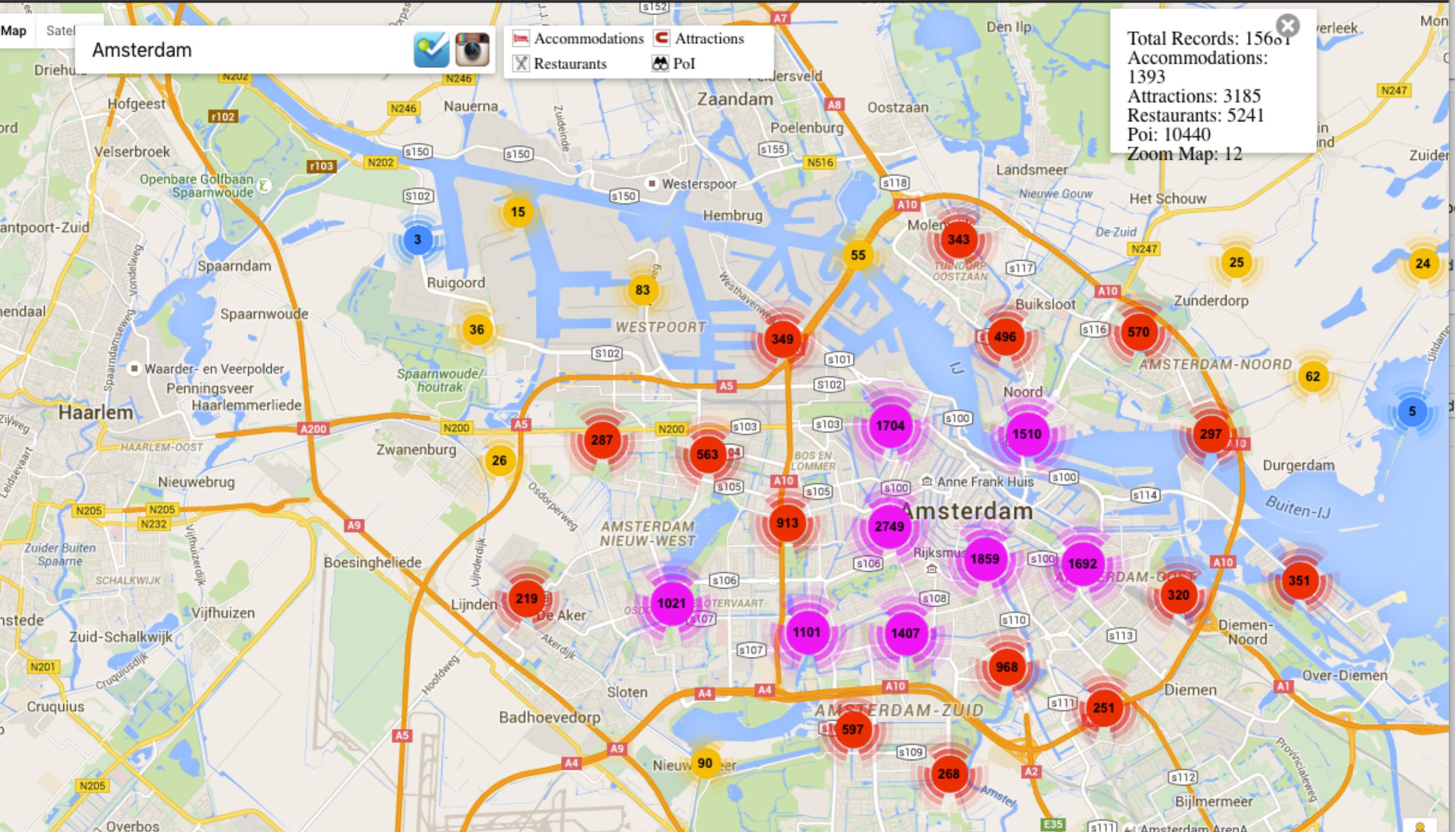
- nothing special really
- comfortable and clean
- boring
- really liked
- modern and stylish
- big
- just
- basic and dull

# A hotel review

- Nothing special really. Comfortable and clean but very boring decor in comparison to **other NH hotels**. I stayed in **NH** in **Brussels** and **Zurich** and I really liked them because of **their modern and stylish** design and big rooms. **This one** was just like any other hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and **20 euros** for breakfast!!! It was good but way overpriced! The best thing about **the hotel** was **the location** - city centre, 2min from a metro stop.

Scale 1 (negative) to 5 (positive)

|             | Overall | Design | Room | Service | Price | Location | Transport |
|-------------|---------|--------|------|---------|-------|----------|-----------|
| This NH     | 2       | 1,5    | 2    | 3       | 1     | 4        | 4         |
| NH Brussels | 4       | 4      | 4    |         |       |          |           |
| NH Zurich   | 4       | 4      | 4    |         |       |          |           |



<http://tour-pedia.org/about/index.html>



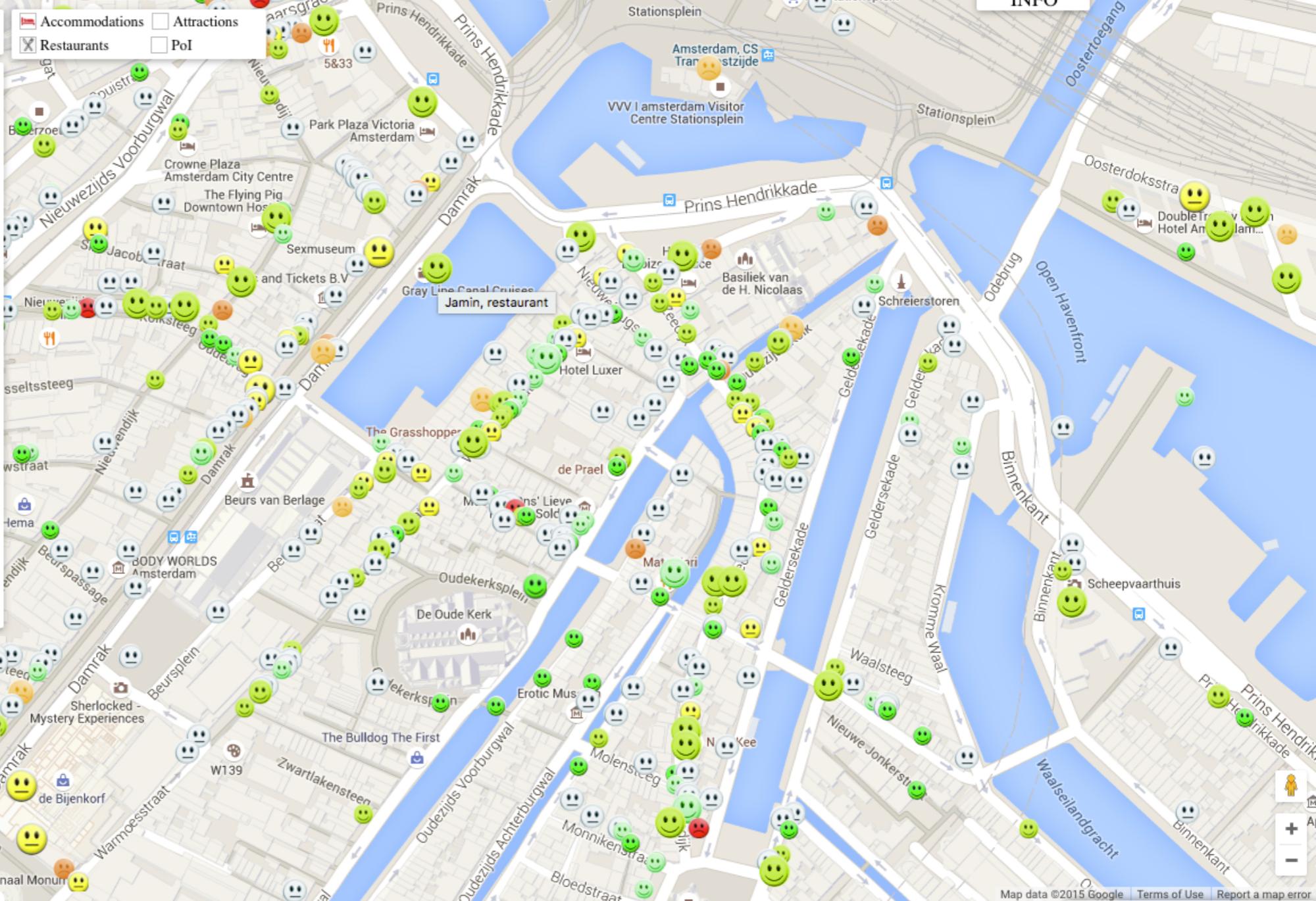
Map

Satel

## Amsterdam

Name:  
JaminAddress:  
Damrak 25, Amsterdam, NetherlandsOfficial website:  
<http://www.jamin.nl/>Links:  
[Foursquare Page](#)  
[GooglePlaces Page](#)

Category: restaurant

Reviews   
Num Reviews: 30  
Polarity: 7Awesome food, low price, definitely  
recommend!

Map

Satellite

## Amsterdam



X

Name:  
Allstars Steakhouse

Address:  
Damrak 32

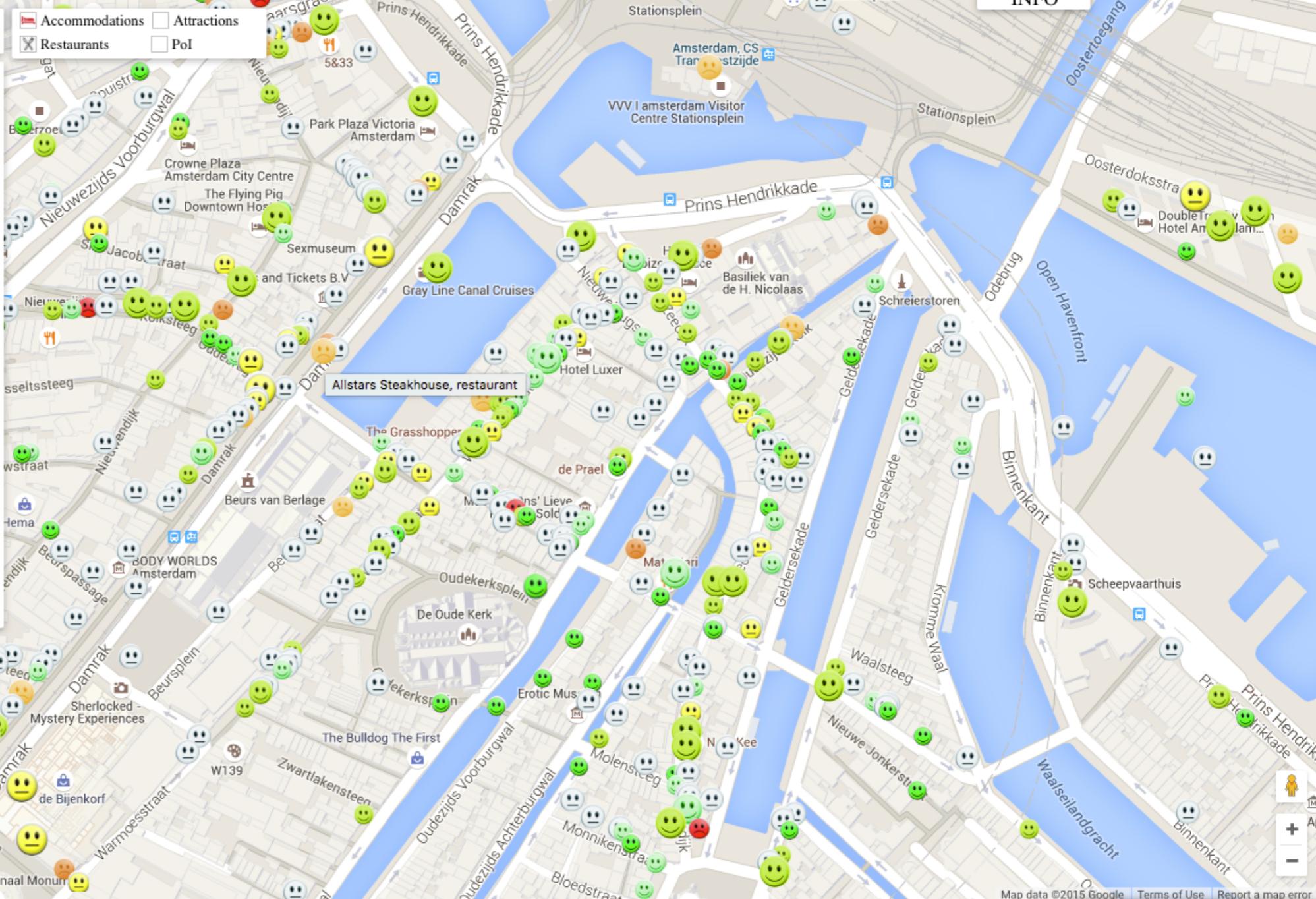
Official website:  
<http://www.allstars-restaurant.nl>

Links:  
[Foursquare Page](#)

Category: restaurant

Reviews   
Num Reviews: 16  
Polarity: 4

Worse restaurant ever !



# Cross-lingual/cultural sentiment

- 2,486 expressions in hotel reviews
- annotated in 6 languages
- 8 aspect groups (food, clean, price, behaviour, general evaluation, size, location, noise, size),

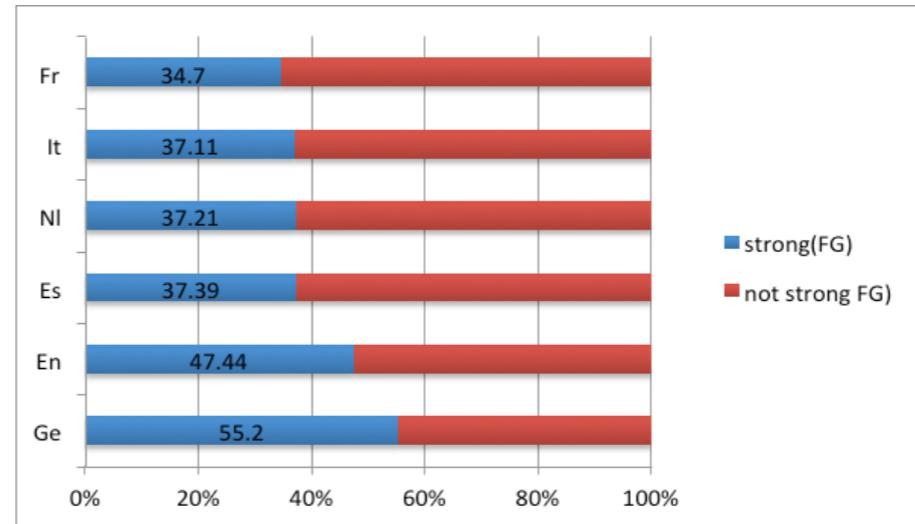
| GENERAL-EVALUATION 669 |                                 |                 |
|------------------------|---------------------------------|-----------------|
| en                     | definitely recommend;           | strong positive |
| es                     | recomendaríamos sin duda alguna | positive        |
| fr                     | recommander chaleureusement     | strong positive |
| it                     | consiglio vivamente             | positive        |

| NOISE 1117 |                        |                 |
|------------|------------------------|-----------------|
| en         | a lot of noise         | strong negative |
| fr         | énormement de bruit    | strong negative |
| it         | molto rumore           | negative        |
| nl         | veel lawaai            | negative        |
| nl         | ontzettend veel lawaai | strong negative |

# Cultural Normalization

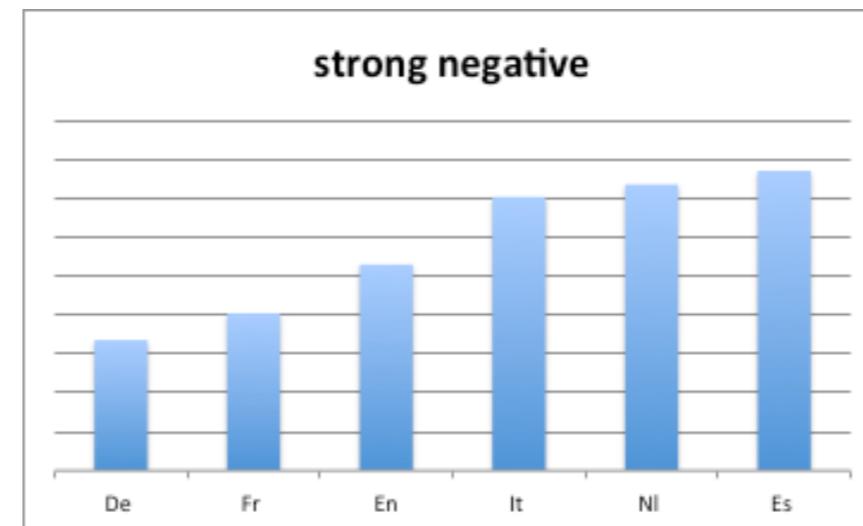
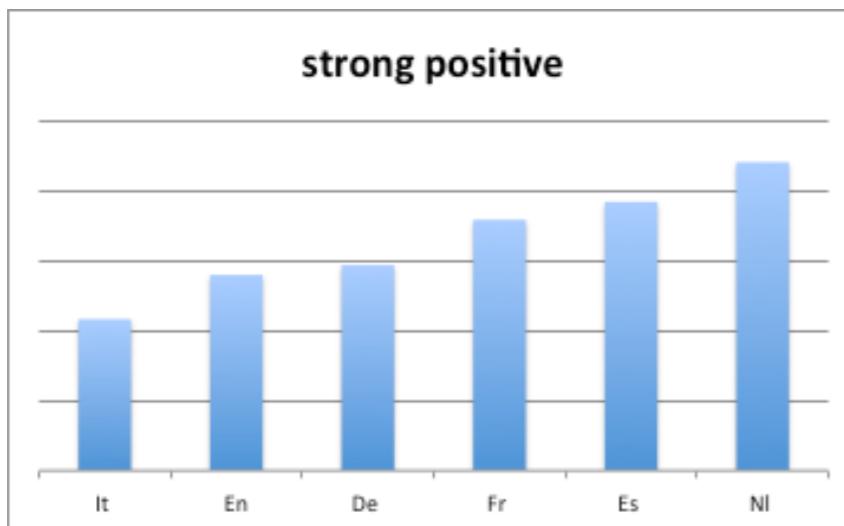
## Differences in ratings in hotel reviews:

- English and Germans give more extreme (strong positive and strong negative) ratings than the other cultures



## Differences in language use in hotel reviews:

- Dutch and Spanish are more expressive as they show a relative over-use of strong expressions
- German and English are less expressive as they show a relative under-use of strong expressions



=> People from German and English culture are quite strong when expressing their emotions in ratings, but they seem to downgrade their emotions when expressing them with language

# Reviewer & reader ratings

- The hotel seems rather outdated. The breakfast room is just not big enough to cope with the Sunday-morning crowds.



# Reviewer & reader ratings

- The hotel seems rather outdated. The breakfast room is just not big enough to cope with the Sunday-morning crowds.
- Maks and Vossen (RANLP-2013)

Review  
rating

7



Reader  
rating  
***negative***

Target= hotel, holiday

Target= aspects in the text

9% - 37% sentiment mismatch at document level

# Cross-lingual/cultural sentiment

- 2,486 expressions in hotel reviews
- annotated in 6 languages
- 8 aspect groups (food, clean, price, behaviour, general evaluation, size, location, noise, size),

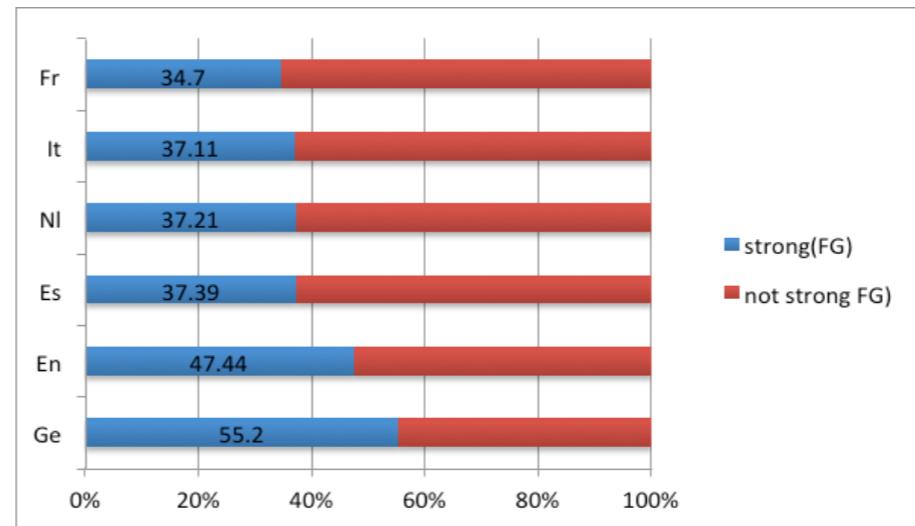
| GENERAL-EVALUATION 669 |                                 |                 |
|------------------------|---------------------------------|-----------------|
| en                     | definitely recommend;           | strong positive |
| es                     | recomendaríamos sin duda alguna | positive        |
| fr                     | recommander chaleureusement     | strong positive |
| it                     | consiglio vivamente             | positive        |

| NOISE 1117 |                        |                 |
|------------|------------------------|-----------------|
| en         | a lot of noise         | strong negative |
| fr         | énormement de bruit    | strong negative |
| it         | molto rumore           | negative        |
| nl         | veel lawaai            | negative        |
| nl         | ontzettend veel lawaai | strong negative |

# Cultural Normalization

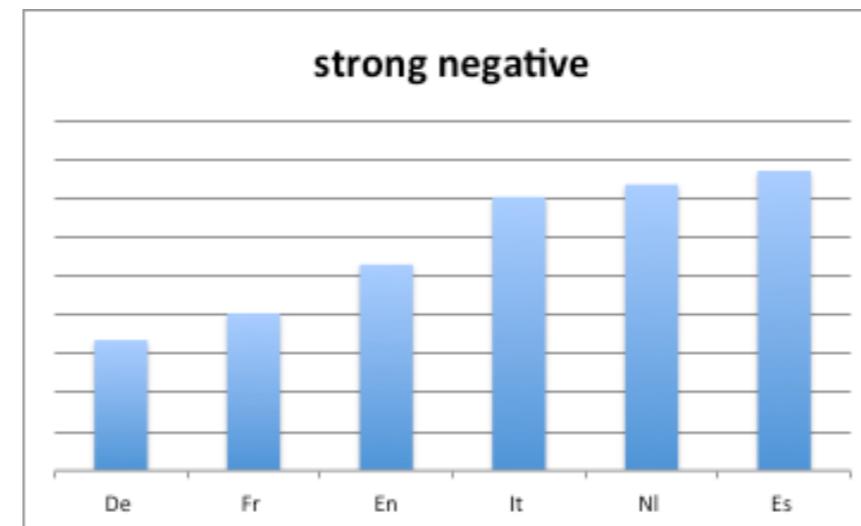
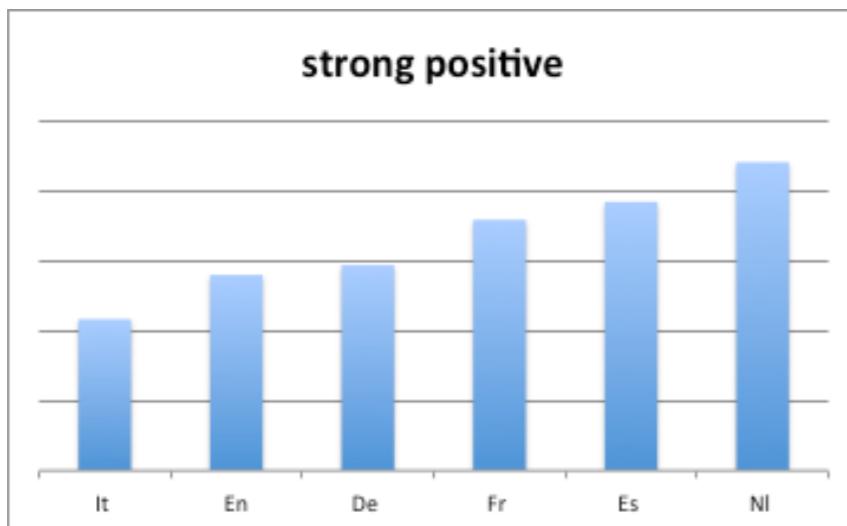
## Differences in ratings in hotel reviews:

- English and Germans give more extreme (strong positive and strong negative) ratings than the other cultures



## Differences in language use in hotel reviews:

- Dutch and Spanish are more expressive as they show a relative over-use of strong expressions
- German and English are less expressive as they show a relative under-use of strong expressions



=> People from German and English culture are quite strong when expressing their emotions in ratings, but they seem to downgrade their emotions when expressing them with language