sentiment analysis

kristoffer l nielbo
kln@cas.au.dk
center for humanities computing

AARHUS UNIVERSITY

CENTER FOR HUMANITIES
COMPUTING AARHUS

JOHN DICKERSON: Did President Obama give you any advice that was helpful? That you think, wow, he really was–
DONALD TRUMP: – Well, he was very nice to me. But after that, we've had some difficulties. So it doesn't matter. You know, words are less important to me than deeds. And you– you saw what happened with surveillance. And everybody saw what happened with surveillance–
JOHN DICKERSON: Difficulties how?
PRESIDENT DONALD TRUMP: – and I thought that – well, you saw what happened with surveillance. And I think that was inappropriate, but that's the way–
JOHN DICKERSON: What does that mean, sir?
PRESIDENT DONALD TRUMP: You can figure that out yourself.
JOHN DICKERSON: Well, I– the reason I ask is you said he was– you called him "sick and bad".
PRESIDENT DONALD TRUMP: Look, you can figure it out yourself. He was very nice to me with words, but– and when I was with him – but after that, there has been no relationship.
JOHN DICKERSON: But you stand by that claim about him?
PRESIDENT DONALD TRUMP: I don't stand by anything. I just– you can take it the way you want. I think our side's been proven very strongly. And everybody's talking about it. And frankly it should be discussed. I think that is a very big surveillance of our citizens. I think it's a very big topic. And it's a topic that should be number one. And we should find out what the hell is going on.
JOHN DICKERSON: I just wanted to find out, though. You're– you're the president of the United States. You said he was "sick and bad" because he had tapped you– I'm just–
PRESIDENT DONALD TRUMP: You can take– any way. You can take it any way you want.
JOHN DICKERSON: But I'm asking you. Because you don't want it to be–
PRESIDENT DONALD TRUMP: You don't–
JOHN DICKERSON: –fake news. I want to hear it from–
PRESIDENT DONALD TRUMP: You don't have to–
JOHN DICKERSON: –President Trump.
PRESIDENT DONALD TRUMP: –ask me. You don't have to ask me.
JOHN DICKERSON: Why not?
PRESIDENT DONALD TRUMP: Because I have my own opinions. You can have your own opinions.
JOHN DICKERSON: But I want to know your opinions. You're the president of the United States.
PRESIDENT DONALD TRUMP: Okay, it's enough. Thank you. Thank you very much.

source: `http://www.cbsnews.com/news/president-trump-oval-office-interview-cbs-this-morning-full-transcript/`

- a document is an ordered set of words that (at least in part) expresses the cognitive and affective states of the author

- we want an automized method that transfers psychological scales to documents and maintain validity and reliablility

- preferably, the method should be scalable both in terms of quantity and context

- we can use a dictionary to extract cognitive and affective keywords from a collection of documents and apply a sentiment function

```
'Did Crooked Hillary help disgusting (check out sex tape and past) Alicia M become a U.S. citizen
so she could use her in the debate?'

Positive sex, citizen
Negative crooked, hillary, disgusting, out
Sentiment Score (2+1) + (-2-1-3-1) = -4
Sentiment Polarity Negative
Overall Score Sum of all sentence scores
```

- a sentiment vector is a vector of keyword frequencies weighted by sentiment scores

sentiment analysis

popular methods for rating the `affective content` of texts

used in business analytics and bio-NLP to predict market behavior, consumer preferences, happiness and quality of life

originate in psychometric and sociometric scale studies
three general approaches:

- `dictionary-based methods` (word counting)
- supervised learning (machine learning)
- unsupervised learning (machine learning)

dictionary-based methods

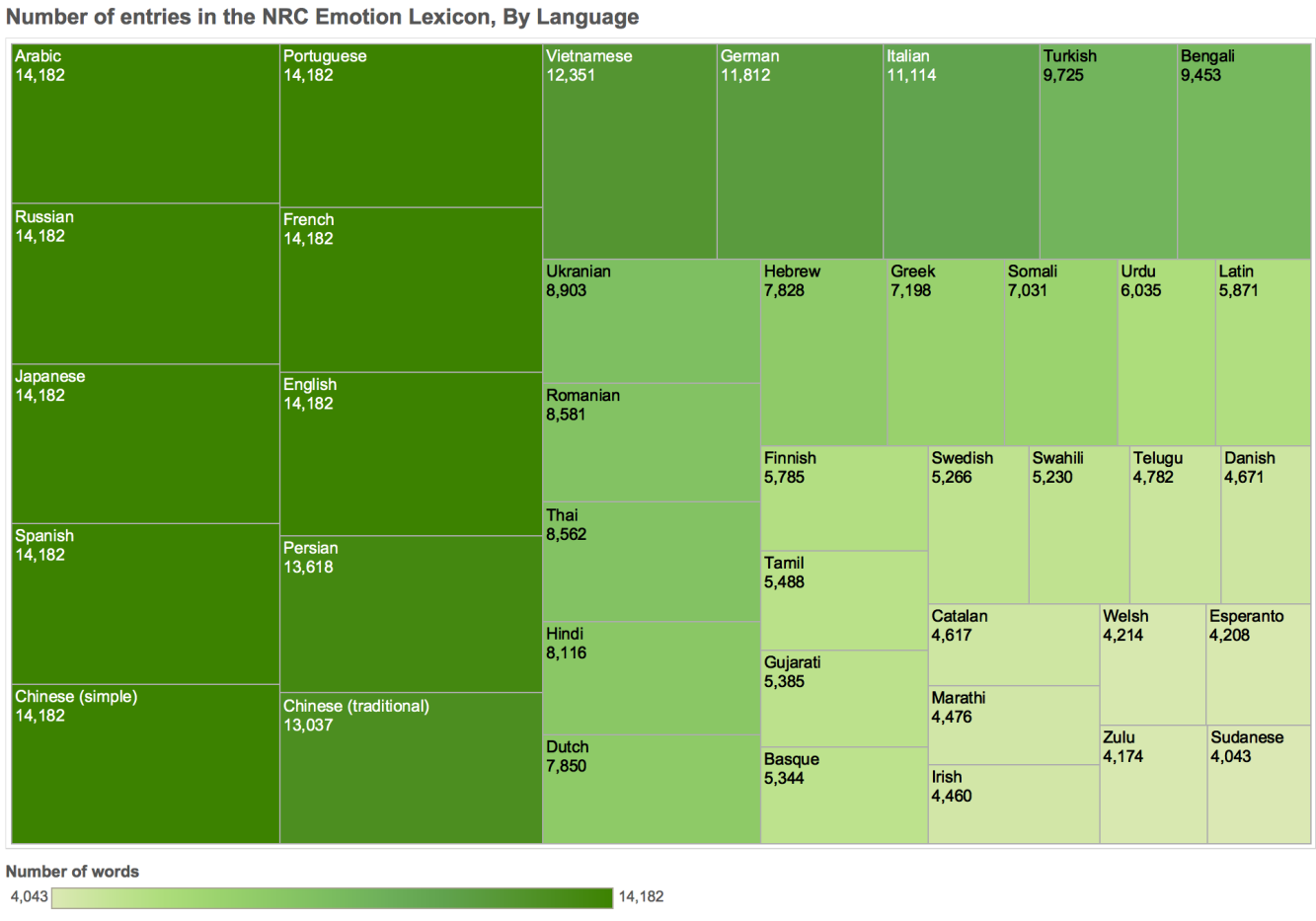
a dictionary is basically `a set of words with ratings`

ratings can be `binary` (± 1 or $0/1$) or based on `continuum` ($1, 2 \ldots m$ or $1 : m$)

compute corpus frequency for each dictionary word and multiply their sentiment rating (weight)

Dictionary	# Fixed	# Stems	Total	Range	# Pos	# Neg	Construction	License
LabMT	10222	0	10222	1.3→ 8.5	7152	2977	Survey: MT, 50 ratings	CC.
ANEW	1030	0	1030	1.25 → 8.82	580	449	Survey: FSU Psych 101	Free for research.
WK	13915	0	13915	1.26 → 8.53	7761	5945	Survey: MT, >14 ratings	CC.
MPQA	5587	1605	7192	-1,0,1	2393	4342	Manual + ML	GNU GPL.
LIWC	722	644	1366	-1,0,1	406	500	Manual	Paid, commercial.
Liu	6782	0	6782	-1,1	2003	4779	Dictionary propagation	Free.
PANAS-X	60	0	60	-1,1	10	10	Manual	Copyrighted paper
Pattern 2.6	1528	0	1528	-1,0,+1	528	620	Unspecified	BSD
SentiWordNet 2.6	147701	0	147701	-1 → 1	17677	20410	Synset synonyms	CC BY-SA 3.0
AFINN	2477	0	2477	-5, -4, . . . , 4, 5	878	1598	Manual	ODbL v1.0
General Inquirer	4205	0	4205	-1,+1	1915	2290	Harvard-IV-4	Unspecified
WDAL	8743	0	8743	1 → 3	6517	1778	Survey: Columbia students	Unspecified
NRC	1220176	0	1220176	-5 → 5	575967	644209	PMI with emoticons	Free for research

languages



AARHUS UNIVERSITY

CENTER FOR HUMANITIES COMPUTING AARHUS

pros and cons

advantages (in comparison to ML)

- corpus agnostic (can be applied without training)
- avoid *black boxing* the solution

assumptions and problems

- bag-of-words assumption
- large data: Accuracy depends on large data set (single sentence or paragraphs are useless)
- contextual errors: Context sensitivity of word meaning ($miss_{\downarrow}$, $vice_{\downarrow}$) and negations ($\{not_{\downarrow}\ good_{\uparrow}\}_{neutral}$)
- lower accuracy than supervised learning (but supervised learning needs class information and is corpus dependent)

word rating

words in dictionaries are rated according to more or less principled procedures:

- survey-based: Random samples or crowd sourcing (MTurk)
- manual: expert or naive (∼convenience)

Rating issues

- space and time specificity (e.g., ANEW is from 2000)
- dependencies between raters
- the `WIERD` problem (LIWC was based on American undergraduates)

mismatches

across dictionaries we find words that seem incorrectly rated

$Negative_{MPQA}$: {*moonlight*, *cutest*, *finest*, *funniest*, *comedy*, *laugh*∗}
$Positive_{LIWC}$: {*dynamite*, *careful*, *richard*∗, *silly*, *gloria*, *securities*, *boldface*}

- reliance on specific sample of raters
- 'dirty' ratings

political application