Popup Lab: Data Preparation

Frontiers in Digital Scholarship

DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

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:

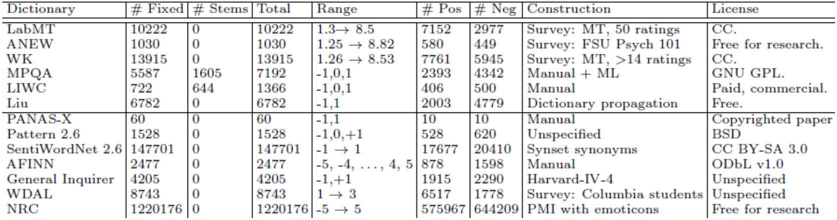
- Dictionary-based methods (word counting)
- Supervised learning (machine learning)
- Unsupervised learning (machine learning)

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 $1 : m$)

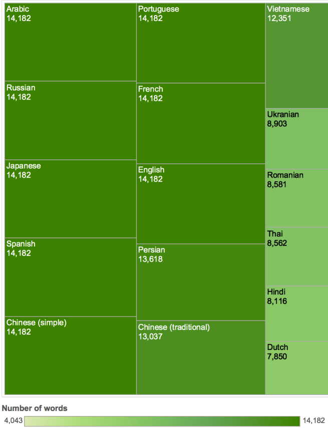
Compute corpus frequency for each dictionary word and multiply their sentiment rating (weight)



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

Number of entries in the NRC Emotion Lexicon, By Language



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

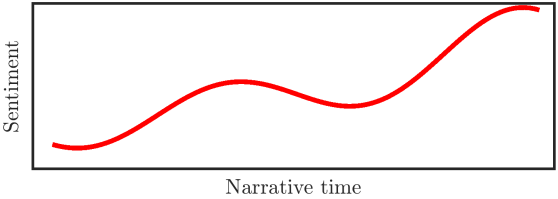
Across dictionaries we find words that seem incorrectly rated

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

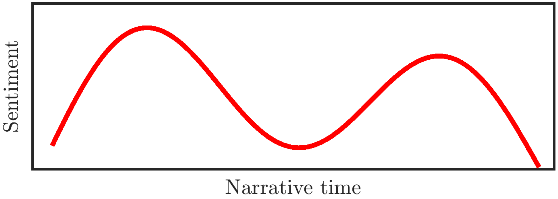
- Reliance on specific sample of raters
- 'Dirty' ratings

Literary applications

Bible, KJV



Koran, Arberry Translation


Sentiment
Narrative time