

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

we want a **automatized method** that transfers psychological scales to documents and maintain validity and reliability

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

```
1 'Did Crooked Hillary help disgusting (check out sex tape and past) Alicia M become a U.S. citizen  
2 so she could use her in the debate?'  
3  
4 Positive sex, citizen  
5 Negative crooked, hillary, disgusting, out  
6 Sentiment Score (2+1) + (-2-1-3-1) = -4  
7 Sentiment Polarity Negative  
8 Overall Score Sum of all sentence scores
```

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

sentiment analysis a set of methods for extracting the affective (primarily) components from unstructured data

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

Originate in psychological and sociological scale studies

Three general approaches:

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

A **dictionary** is basically a set of words with ratings

Ratings can be **binary** (± 1 or 0/1) or based on **continuum** (1, 2 ... 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

across dictionaries we find inconsistencies, that is, words that seem incorrectly rated

NegativeMPQA : {*moonlight, cutest, finest, funniest, comedy, laugh**}

PositiveLIWC : {*dynamite, careful, richard*, silly, gloria, securities, boldface*}

origin of dictionary mismatches

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

word ratings for dictionaries are based on **more or less principled procedures**

- ▶ survey-based: random samples or crowd sourcing
- ▶ manual: expert or naive (~convenience)

rating issues

- ▶ space and time specificity (ANEW from 2000, NRC depends on Google Trans.)
- ▶ dependencies between raters
- ▶ the *Western, Educated, Industrialized, Rich, and Democratic* issue (LIWC is based on American undergraduates)
- ▶ uncorrected ratings

pros of dictionary-based sentiment analysis

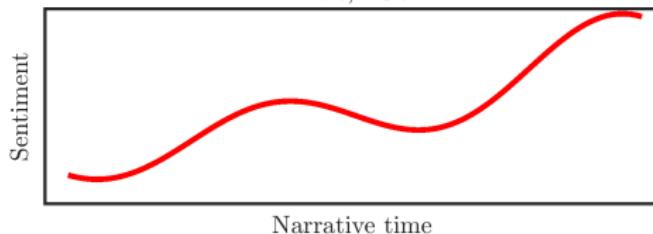
- ▶ computationally efficient way of accessing affective components
- ▶ corpus agnostic in comparison to ML (can be applied without training)
- ▶ transparent technique that **avoids black boxing** the solution

cons of dictionary-based sentiment analysis

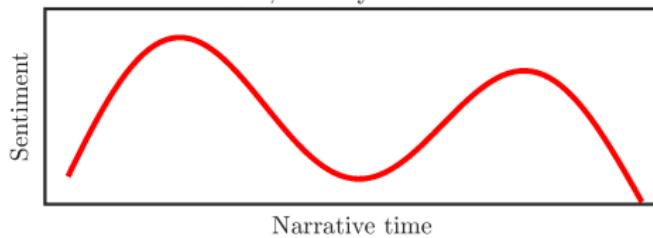
- ▶ bag-of-words assumption
- ▶ accuracy depends on **large data sets** (single sentence or paragraphs are useless)
- ▶ context sensitivity of word meaning ($miss_{\downarrow}$, $vice_{\downarrow}$) and negations
($\{not_{\downarrow} good_{\uparrow}\}_{neutral}$)
- ▶ lower accuracy than supervised learning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



Bible, KJV



Koran, Arberry Translation



language Assessment by Mechanical Turk (labMT)

```
1 import pandas as pd
2 url = 'http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/
3     journal.pone.0026752.s001'
4 labmt = pd.read_csv(url, skiprows=2, sep=' ', index_col=0)
5
6 dictionary = labmt.happiness_average.to_dict()
7 labmtbar = labmt.happiness_average.mean()
8 dictionary_0 = (labmt.happiness_average - labmtbar).to_dict()
9
10 def sent_scr(string):
11     tokens = string.split()
12     return sum([dictionary.get(token.lower(), 0.0) for token in tokens]) / len(tokens)
13
14 sent = 'Did Crooked Hillary help disgusting (check out sex tape and past) Alicia M become a U.S.
15 citizen so she could use her in the debate?'
16
17 print sent_scr(sent)
```

Danish dictionary from Finn Årup Nielsen

```
1 from afinn import Afinn
2 afinn = Afinn(language='da')
3 sent = 'Røvhul! Dumme svin! Du behandler mig som lort!
4     Nu kan du rende og hoppe jeg vil ikke tale med dig nogensinde igen'
5 print afinn.score(sent)
```

Matt Jockers' package for plot analysis

```
1 library(tm)
2 library(syuzhet)
3
4 text.v <- paste(scan(filepath, what = 'character', sep='\n', encoding = 'UTF-8'), collapse = " ")
5
6 text_sent <- get_sentences(text.v)
7
8 text_syuzhet <- get_sentiment(text_sent, method = 'syuzhet')
9
10 text_sent[which(text_syuzhet == max(text_syuzhet))]
11 text_sent[which(text_syuzhet == min(text_syuzhet))]
12 text_sent[which(text_syuzhet > (mean(text_syuzhet)+sd(text_syuzhet)*2))]
13 text_sent[which(text_syuzhet < (mean(text_syuzhet)-sd(text_syuzhet)*2))]
14
15 text_syuzhet_val <- get_percentage_values(text_syuzhet, bin = 100)
```

sentiment analysis ala Vermont (from labMTsimple/storyLab.py)

```
1 ignoreWords = ["nigga", "nigger", "niggaz", "niggas"];
2
3 for word in ignore:
4     ignoreWords.append(word)
5 newVec = copy.copy(tmpVec)
6 for i in range(len(score_list)):
7     if abs(score_list[i]-center) < stopVal:
8         newVec[i] = 0
9     if word_list[i] in ignoreWords:
10        newVec[i] = 0
11
12 return newVec
```