Reagan et al. *EPJ Data Science* (2017) 6:28
DOI 10.1140/epjds/s13688-017-0121-9

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs

Andrew J Reagan[1*], Christopher M Danforth[1], Brian Tivnan[2], Jake Ryland Williams[3] and Peter Sheridan Dodds[1]

*Correspondence:
andy@andyreagan.com
[1]University of Vermont, 85 South Prospect St, Burlington, VT 05405, USA
Full list of author information is available at the end of the article

Abstract

The emergence and global adoption of social media has rendered possible the real-time estimation of population-scale sentiment, an extraordinary capacity which has profound implications for our understanding of human behavior. Given the growing assortment of sentiment-measuring instruments, it is imperative to understand which aspects of sentiment dictionaries contribute to both their classification accuracy and their ability to provide richer understanding of texts. Here, we perform detailed, quantitative tests and qualitative assessments of 6 dictionary-based methods applied to 4 different corpora, and briefly examine a further 20 methods. We show that while inappropriate for sentences, dictionary-based methods are generally robust in their classification accuracy for longer texts. Most importantly they can aid understanding of texts with reliable and meaningful word shift graphs if (1) the dictionary covers a sufficiently large portion of a given text's lexicon when weighted by word usage frequency; and (2) words are scored on a continuous scale.

Keywords: sentiment; sentiment analysis; sentiment dictionaries; language; natural language processing; data visualization; text visualization

1 Introduction

As we move further into what might be called the Sociotechnocene — with increasingly more interactions, decisions, and impact being made by globally distributed people and algorithms — the myriad human social dynamics that have shaped our history have become far more visible and measurable than ever before. Of the many ways we are now able to characterize social systems in microscopic detail, sentiment detection for populations at all scales has become a prominent research arena. Attempts to leverage online expression for sentiment mining include prediction of stock markets [1–4], assessing responses to advertising, real-time monitoring of global happiness [5], and measuring a health-related quality of life [6]. The diverse set of instruments produced by this work now provide indicators that help scientists understand collective behavior, inform public policy makers, and, in industry, gauge the sentiment of public response to marketing

campaigns. Given their widespread usage and potential to influence social systems, understanding how these instruments perform and how they compare with each other has become imperative. Benchmarking both their ability to provide insight into sentiment and their classification performance focuses future development and provides practical advice to non-experts in selecting a sentiment dictionary.

We identify sentiment detection methods as belonging to one of three categories, each carrying their own advantages and disadvantages:

1 Dictionary-based methods [5, 7–11],
2 Supervised learning methods [10], and
3 Unsupervised (or deep) learning methods [12].

Here, we focus on dictionary-based methods, which all center around the determination of a text T's average happiness (sometimes referred to as *valence*) with sentiment dictionary D through the equation:

$$h_{\mathrm{D}}^{T} = \frac{\sum_{w\in D} h_{\mathrm{D}}(w)\cdot f^{T}(w)}{\sum_{w\in D} f^{T}(w)} = \sum_{w\in D} h_{\mathrm{D}}(w)\cdot p^{T}(w), \tag{1}$$

where we denote each of the words in a given sentiment dictionary D as w, word sentiment scores as $h_{\mathrm{D}}(w)$, word frequency as $f^{T}(w)$, and normalized frequency of w in T as $p^{T}(w) = f^{T}(w)/\sum_{w\in D} f^{T}(w)$. In this way, we measure the happiness of a text in a manner analogous to taking the temperature of a room. While other simple sentiment metrics may be considered, we will see that analyzing individual word contributions is important and that this equation allows for a straightforward, meaningful interpretation.

Dictionary-based methods offer two distinct advantages which we find necessary: (1) they are in principle corpus agnostic (applicable to corpora without ground truth data available) and (2) in contrast to black box (highly non-linear) methods, they offer the ability to 'look under the hood' at words contributing to a particular score through *word shift graphs* (defined fully later; see also [13, 14]). Indeed, if we are concerned with understanding why a particular scoring method varies — e.g., our undertaking is scientific — then word shift graphs are essential tools. In the absence of word shift graphs, or similar devices, explanations of sentiment trends can and often will miss crucial information.

As all methods must, dictionary-based 'bag-of-words' approaches suffer from various drawbacks, and three are worth stating up front. First, they are only applicable to corpora of sufficient size, well beyond that of a single sentence [15] (widespread usage in this fashion does not suffice as a counterargument). We directly verify this assertion on individual tweets, finding that while some sentiment dictionaries perform admirably, the average (median) F1-score on the STS-Gold data set is 0.50 (0.54) from all datasets (Table S1). Others having shown similar results for dictionary methods with short text [15]. Second, state-of-the-art learning methods with a sufficiently large training set for a specific corpus will outperform dictionary-based methods on the same corpus [16]. However, in practice the domains and topics to which sentiment analysis are applied are highly varied, such that training to a high degree of specificity for a single corpus may not be practical and, from a scientific standpoint, will severely constrain attempts to detect and understand universal patterns. Third, words may be evaluated out of context or with the wrong sense. A simple example is the word 'miss' occurring frequently when evaluating articles in the Society

section of the New York Times. This kind of contextual error is something we can readily identify and correct for through word shift graphs, but could remain hidden to users of nonlinear learning methods.

We lay out our paper as follows. We list and describe the dictionary-based methods we consider in Section 2.1, and outline the corpora we use for tests in Section 2.2. We present our results in Section 3, comparing all methods in how they perform for specific analyses of the New York Times (NYT) (Section 3.1), movie reviews (Section 3.2), Google Books (Section 3.3), and Twitter (Section 3.4). In Section 3.5, we make some basic comparisons between dictionary-based methods and machine learning approaches. We provide concluding remarks in Section 4 and bolster our findings with figures, tables, and additional analysis in the Supplementary Material (supplied as Additional file 1).

2 Sentiment dictionaries, corpora, and word shift graphs

2.1 Sentiment dictionaries

The words 'sentiment dictionary,' 'lexicon,' and 'corpus' are often used interchangeably, and for clarity we define our usage as follows.

Sentiment Dictionary	Set of words (possibly including word stems) with ratings.
Corpus	Collection of texts which we seek to analyze.
Lexicon	The words contained within a corpus (often said to be 'tokenized').

We test the following six sentiment dictionaries in depth:

labMT	language assessment by Mechanical Turk [5].
ANEW	Affective Norms of English Words [7].
WK	Warriner and Kuperman rated words from SUBTLEX by Mechanical Turk [11].
MPQA	The Multi-Perspective Question Answering (MPQA) Subjectivity Dictionary [9].
LIWC{01,07,15}	Linguistic Inquiry and Word Count, three versions [8].
OL	Opinion Lexicon, developed by Bing Liu [10].

We also make note of 18 other sentiment dictionaries:

PANAS-X	The Positive and Negative Affect Schedule Expanded [17].
Pattern	A web mining module for the Python programming language, version 2.6 [18].
SentiWordNet	WordNet synsets each assigned three sentiment scores: positivity, negativity, and objectivity [19].
AFINN	Words manually rated –5 to 5 with impact scores by Finn Nielsen [20].
GI	General Inquirer: database of words and manually created semantic and cognitive categories, including positive and negative connotations [21].
WDAL	Whissel's Dictionary of Affective Language: words rated in terms of their Pleasantness, Activation, and Imagery (concreteness) [22].
EmoLex	NRC Word-Emotion Association Lexicon: emotions and sentiment evoked by common words and phrases using Mechanical Turk [23].
MaxDiff	NRC MaxDiff Twitter Sentiment Lexicon: crowdsourced real-valued scores using the MaxDiff method [24].

HashtagSent	NRC Hashtag Sentiment Lexicon: created from Tweets using Pairwise Mutual Information with sentiment hashtags as positive and negative labels (here we use only the unigrams) [25].
Sent140Lex	NRC Sentiment140 Lexicon: created from the 'sentiment140' corpus of Tweets, using Pairwise Mutual Information with emoticons as positive and negative labels (here we use only the unigrams) [26].
SOCAL	Manually constructed general-purpose sentiment dictionary [27].
SenticNet	Sentiment dataset labeled with semantics and 5 dimensions of emotions by Cambria *et al.*, version 3 [28].
Emoticons	Commonly used emoticons with their positive, negative, or neutral emotion [29].
SentiStrength	an API and Java program for general purpose sentiment detection (here we use only the sentiment dictionary) [30].
VADER	method developed specifically for Twitter and social media analysis [31].
Umigon	Manually built specifically to analyze Tweets from the sentiment140 corpus [32].
USent	set of emoticons and bad words that extend MPQA [33].
EmoSenticNet	extends SenticNet words with WNA labels [34].

All of these sentiment dictionaries were produced by academic groups, and with the exception of LIWC, they are provided free of charge. In Table 1, we supply the main aspects — such as word count, score type (continuum or binary), and license information — for the sentiment dictionaries listed above. In the GitHub repository associated with our paper, https://github.com/andyreagan/sentiment-analysis-comparison, we include all of the sentiment dictionaries except LIWC.

The labMT, ANEW, and WK sentiment dictionaries have scores ranging on a continuum from 1 (low happiness) to 9 (high happiness) with 5 as neutral, whereas the others we test in detail have scores of ± 1, and either explicitly or implicitly 0 (neutral). We will refer to the latter sentiment dictionaries as being binary, even if neutral is included. Other non-binary ranges include a continuous scale from –1 to 1 (SentiWordNet), integers from –5 to 5 (AFINN), continuous from 1 to 3 (GI), and continuous from –5 to 5 (NRC). For coverage tests, we include all available words, to gain a full sense of the breadth of each sentiment dictionary. In scoring, we do not include neutral words from any sentiment dictionary.

We test the labMT, ANEW, and WK dictionaries for a range of stop words (starting with the removal of words scoring within $\Delta_h = 1$ of the neutral score of 5) [14]. The ability to remove stop words — a common practice for text pre-processing — is one advantage of dictionaries that have a range of scores, allowing us to tune the instrument for maximum performance, while retaining all of the benefits of a dictionary method. We will show that, in agreement with the original paper introducing labMT and looking at Twitter data, a $\Delta_h = 1$ is a pragmatic choice [14].

Since we do not apply a part of speech tagger, when using the MPQA dictionary we are obliged to exclude words with scores of both +1 and –1. The words and stems with both scores are: blood, boast* (we denote stems with an asterisk), conscience, deep, destiny, keen, large, and precious. We choose to match a text's words using the fixed word set from each sentiment dictionary before stems, hence words with overlapping matches (a fixed word that also matches a stem) are first matched by the fixed word.

Table 1 Summary of dictionary attributes used in sentiment measurement instruments. We provide all acronyms and abbreviations and further information regarding sentiment dictionaries in Section 2.1. We test the first 6 dictionaries extensively. The range indicates whether scores are continuous or binary (we use the term binary for sentiment dictionaries for which words are scored as ±1 and optionally 0).

Dictionary	# Entries	Range	Construction	License	Ref.
labMT	10,222	$1.3 \rightarrow 8.5$	Survey: MT, 50 ratings	CC	[5]
ANEW	1,034	$1.2 \rightarrow 8.8$	Survey: UF Intro Psych	Free for research	[7]
LIWC07	4,483	$[-1, 0, 1]$	Manual	Paid, commercial	[8]
MPQA	7,192	$[-1, 0, 1]$	Manual + ML	GNU GPL	[9]
OL	6,782	$[-1, 1]$	Dictionary propagation	Free	[10]
WK	13,915	$1.3 \rightarrow 8.5$	Survey: MT, 14-18 ratings	CC	[11]
LIWC01	2,322	$[-1, 0, 1]$	Manual	Paid, commercial	[8]
LIWC15	6,549	$[-1, 0, 1]$	Manual	Paid, commercial	[8]
PANAS-X	20	$[-1, 1]$	Manual	Copyrighted paper	[17]
Pattern	1,528	$-1.0 \rightarrow 1.0$	Unspecified	BSD	[18]
SentiWordNet	147,700	$-1.0 \rightarrow 1.0$	Synset synonyms	CC BY-SA 3.0	[19]
AFINN	2,477	$[-5, -4, \ldots, 4, 5]$	Manual	ODbL v1.0	[20]
GI	3,629	$[-1, 1]$	Harvard-IV-4	Unspecified	[21]
WDAL	8,743	$0.0 \rightarrow 3.0$	Survey: Columbia students	Unspecified	[22]
EmoLex	14,182	$[-1, 0, 1]$	Survey: MT	Free for research	[23]
MaxDiff	1,515	$-1.0 \rightarrow 1.0$	Survey: MT, MaxDiff	Free for research	[24]
HashtagSent	54,129	$-6.9 \rightarrow 7.5$	PMI with hashtags	Free for research	[25]
Sent140Lex	62,468	$-5.0 \rightarrow 5.0$	PMI with emoticons	Free for research	[26]
SOCAL	7,494	$-30.2 \rightarrow 30.7$	Manual	GNU GPL	[27]
SenticNet	30,000	$-1.0 \rightarrow 1.0$	Label propogation	Citation requested	[28]
Emoticons	132	$[-1, 0, 1]$	Manual	Open source code	[29]
SentiStrength	2,615	$[-5, -4, \ldots, 4, 5]$	LIWC + GI	Free for research	[30]
VADER	7,502	$-3.9 \rightarrow 3.4$	MT survey, 10 ratings	Freely available	[31]
Umigon	927	$[-1, 1]$	Manual	Public Domain	[32]
USent	592	$[-1, 1]$	Manual	CC	[33]
EmoSenticNet	13,188	$[-10, -2, -1, 0, 1, 10]$	Bootstrapped extension	Non-commercial	[34]

2.2 Corpora tested

For each sentiment dictionary, we test both the coverage and the ability to detect previously observed and/or known patterns within each of the following corpora, noting the pattern we hope to discern:

1. The New York Times (NYT) [35]: Goal of understanding differences between sections and ranking by sentiment (Section 3.1).
2. Movie reviews [36]: Goal of discerning how emotional language differs in positive and negative reviews and how these differences influence classification accuracy (Section 3.2).
3. Google Books [37]: Goal of understanding time series (Section 3.3).
4. Twitter: Goal of understanding time series (Section 3.4).

For the corpora other than the movie reviews and small numbers of tagged Tweets, there is no publicly available ground truth sentiment, so we instead make comparisons between methods and examine how words contribute to scores. We note that measuring how patterns of sentiment compare with societal measures of well being would also be possible [38]. We offer greater detail on corpus processing below, and we also provide the relevant scripts on GitHub at https://github.com/andyreagan/sentiment-analysis-comparison.

2.3 Word shift graphs

Sentiment analysis is often applied to classify text as positive or negative. Indeed if this were the only use case, the value added by sentiment analysis would be limited. We use

sentiment analysis as a lens that allows us to see how the emotive words in a text shape the overall content. This is accomplished by first analyzing each word to find its individual contribution to the difference in sentiment scores between two texts. The most important and final step is to examine the words themselves, ranked by their individual contribution. Of the four corpora that we analyze, three rely on this type of qualitative analysis: using the sentiment dictionary as a tool to better understand the sentiment of the corpora rather than as a binary classifier.

To make this possible, we must first find the contribution of each word individually. Starting with the ANEW sentiment dictionary and two texts which we label reference and comparison, we take the difference of their sentiment scores $h_{\text{ANEW}}^{(\text{comp})}$ and $h_{\text{ANEW}}^{(\text{ref})}$, rearrange terms, and arrive at

$$h_{\text{ANEW}}^{\text{comp}} - h_{\text{ANEW}}^{\text{ref}} = \sum_{w \in \text{ANEW}} \underbrace{\left[h_{\text{ANEW}}(w) - h_{\text{ANEW}}^{\text{ref}}\right]}_{+/-} \underbrace{\left[p^{\text{comp}}(w) - p^{\text{ref}}(w)\right]}_{\uparrow/\downarrow}.$$

Each word w in the summation contributes to the sentiment difference between the texts according to (1) its sentiment relative to the reference text (+/− = more/less positive), and (2) its change in frequency of usage (↑ / ↓= more/less frequent). As a first step, it is possible to visualize this sorted word list in a table, along with the basic indicators of how its contribution is constituted. We use word shift graphs to present this information in the most accessible manner to advanced users. For further detail, we refer the reader to our instructional post and video at http://www.uvm.edu/storylab/2014/10/06/hedonometer-2-0-measuring-happiness-and-using-word-shifts/.

3 Results

In Figure 1, we show a direct comparison between word scores for each pair of the 6 dictionaries tested. Overall, we find strong agreement between all dictionaries with the exceptions we note below. As a guide, we will provide more detail on the individual comparison between the labMT dictionary and the other five dictionaries by examining the words whose scores disagree across dictionaries shown in Figure 2. We refer the reader to the S2 Appendix for the remaining individual comparisons.

To start with, consider the comparison of the labMT and ANEW dictionaries on a word-for-word basis. Because these dictionaries share the same range of values, a scatterplot is the natural way to visualize the comparison. Across the top row of Figure 1, which compares labMT to the other 5 dictionaries, we see in Panel B for the labMT-ANEW comparison that the RMA best fit [40] is

$$h_{\text{labMT}}(w) = 0.92 * h_{\text{ANEW}}(w) + 0.40$$

for words w in both labMT and ANEW. The 10 words farthest from the line of best fit shown in Panel B of Figure 2 are (with labMT, ANEW scores in parenthesis): lust (4.64, 7.12), bees (5.60, 3.20), silly (5.30, 7.41), engaged (6.16, 8.00), book (7.24, 5.72), hospital (3.50, 5.04), evil (1.90, 3.23), gloom (3.56, 1.88), anxious (3.42, 4.81), and flower (7.88, 6.64). We observe that these words have high standard deviations in labMT. While the overall agreement is very good, we should expect some variation in the emotional associations of words, due to chance, time of survey, and demographic variability. Indeed, the Mechanical Turk users who scored the words for the labMT set in 2011 are evidently different from the University of Florida students who took the ANEW survey in 2000.

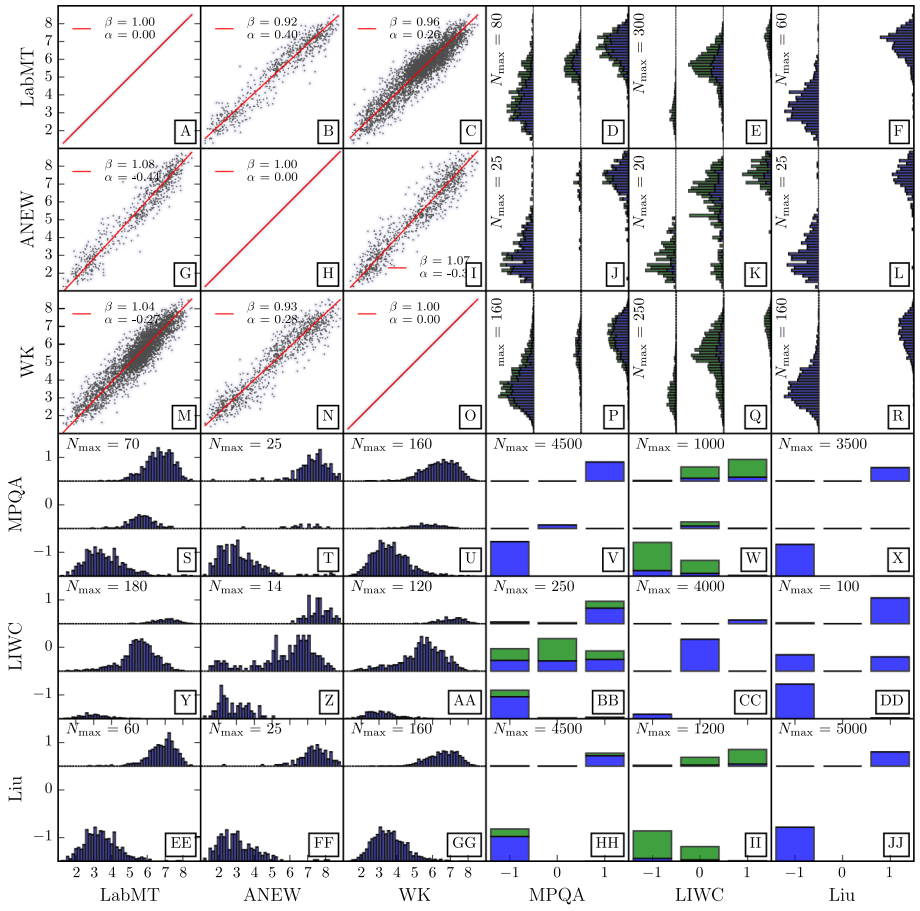

Figure 1 Direct comparison of the words in each of the dictionaries tested. For the comparison of two dictionaries, we plot words that are matched by the independent variable '*x*' in the dependent variable '*y*'. Because of this, and cross stem matching, the plots are not symmetric across the diagonal of the entire figure. Where the scores are continuous in both dictionaries, we compute the RMA linear fit. When a sentiment dictionary contains both fixed and stem words, we plot the matches by fixed words in blue and by stem words in green. The axes in the bar plots are not of the same height, due to large mismatches in the number of words in the dictionaries, and we note the maximum height of the bar in the upper left of such plots. Detailed analysis of Panel C can be found in [39]. We provide a table for each off-diagonal panel in the S2 Appendix with the words whose scores exhibit the greatest mismatch, and a subset of these tables in Figure 2.

Comparing labMT with WK in Panel C of Figure 1, we again find a fit with slope near 1, and with a smaller positive shift: $h_{\text{labMT}}(w) = 0.96 * h_{\text{WK}}(w) + 0.26$. The 10 words farthest from the best fit line, shown in Panel B of Figure 2, are (labMT, WK): sue (4.30, 2.18), boogie (5.86, 3.80), exclusive (6.48, 4.50), wake (4.72, 6.57), federal (4.94, 3.06), stroke (2.58, 4.19), gay (4.44, 6.11), patient (5.04, 6.71), user (5.48, 3.67), and blow (4.48, 6.10). Like labMT, the WK dictionary used a Mechanical Turk online survey to gather word ratings. We speculate that the variation may in part be due to differences in the number of scores required for each word in the surveys, with 14-18 in WK and 50 in labMT. For an in depth comparison of these sentiment dictionaries, see reference [39].

To compare the word scores in a binary sentiment dictionary (those with ± 1 or $\pm 1, 0$) to the word scores in a sentiment dictionary with a 1-9 range, we examine the distribution of the continuous scores for each binary score. Looking at the labMT-MPQA comparison in Panel D of Figure 1, we see that more of the matches are between words with-

A: LabMT comparison with ANEW

Word	h_{LabMT}	h_{ANEW}	h_{diff}
lust	4.64	7.12	1.72
bees	5.60	3.20	1.66
silly	5.30	7.41	1.43
engaged	6.16	8.00	1.20
book	7.24	5.72	1.15
hospital	3.50	5.04	1.15
evil	1.90	3.23	1.09
gloom	3.56	1.88	1.05
anxious	3.42	4.81	1.05
flower	7.88	6.64	1.00

B: LabMT comparison with WK

Word	h_{LabMT}	h_{WK}	h_{diff}
sue	4.30	2.18	1.39
boogie	5.86	3.80	1.39
exclusive	6.48	4.50	1.36
wake	4.72	6.57	1.35
federal	4.94	3.06	1.25
stroke	2.58	4.19	1.24
gay	4.44	6.11	1.23
patient	5.04	6.71	1.22
user	5.48	3.67	1.21
blow	4.48	6.10	1.20

C: LabMT comparison with MPQA's negative words

Word	h_{LabMT}	h_{MPQA}
fine	6.74	-1
game	6.92	-1
cartoon	7.20	-1
eternal	7.20	-1
moon	7.28	-1
fun	7.96	-1
comedy	7.98	-1
laugh	8.22	-1
laugh	8.22	-1
laughter	8.50	-1

D: LabMT comparison with MPQA's neutral words

Word	h_{LabMT}	h_{MPQA}
screaming	2.96	0
pressures	3.49	0
pressure	3.66	0
plead	3.67	0
mean	3.68	0
baby	7.28	0
precious	7.34	0
strength	7.40	0
surprise	7.42	0
surprise	7.42	0

E: LabMT comparison with MPQA's positive words

Word	h_{LabMT}	h_{MPQA}
vulnerable	3.34	+1
court	3.78	+1
conviction	4.10	+1
craving	4.46	+1
excuse	4.58	+1
bull	4.62	+1
striking	4.70	+1
offset	4.72	+1
admit	4.74	+1
repair	4.76	+1

F: LabMT comparison with LIWC's neutral words

Word	h_{LabMT}	h_{LIWC}
lack	3.16	0
couldn't	3.32	0
cannot	3.32	0
never	3.34	0
against	3.40	0
rest	7.18	0
greatest	7.26	0
couple	7.30	0
million	7.38	0
billion	7.56	0

Figure 2 The specific words from Panels G, M, S and Y of Figure 1 with the greatest mismatch. Only the center histogram from Panel Y of Figure 1 is included. We emphasize that the labMT dictionary scores generally agree well with the other dictionaries, and we are looking at the marginal words with the strongest disagreement. Within these words, we detect differences in the creation of these dictionaries that carry through to these edge cases. Panel A: The words with most different scores between the labMT and ANEW dictionaries are suggestive of the different meanings that such words entail for the different demographic surveyed to score the words. Panel B: Both dictionaries use surveys from the same demographic (Mechanical Turk), where the labMT dictionary required more individual ratings for each word (at least 50, compared to 14) and appears to have dampened the effect of multiple meaning words. Panels C-E: The words in labMT matched by MPQA with scores of –1, 0, and +1 in MPQA show that there are at least a few words with negative rating in MPQA that are not negative (including the happiest word in the labMT dictionary: 'laughter'), not all of the MPQA words with score 0 are neutral, and that MPQA's positive words are mostly positive according to the labMT score. Panel F: The function words in the expert-curated LIWC dictionary are not emotionally neutral.

out stems (blue) than those with stems (green), and that each score in –1, 0, +1 from MPQA corresponds to a wider range of scores in labMT. We examine the shared individual words from labMT with high sentiment scores and MPQA with score –1, both the happiest and the least happy in labMT with MPQA score 0, and the least happy when MPQA is 1 (Figure 2 Panels C-E). The 10 happiest words in labMT matched by MPQA words with score –1 are: moonlight (7.50), cutest (7.62), finest (7.66), funniest (7.76), comedy (7.98), laughs (8.18), laughing (8.20), laugh (8.22), laughed (8.26), laughter (8.50). This is an immediately troubling list of evidently positive words rated as –1 in MPQA. We observe the top 5 are matched by the stem 'laugh*' in MPQA. The least happy 5 words and happiest 5 words in labMT matched by words in MPQA with score 0 are: sorrows (2.69), screaming (2.96), couldn't (3.32), pressures (3.49), couldnt (3.58), and baby (7.28), precious (7.34), strength (7.40), surprise (7.42), and song (7.58). We see that these MPQA word scores are departures from the other dictionaries, warranting further concern. The least happy words in labMT with score +1 in MPQA that are matched by MPQA are: vulnerable (3.34), court (3.78), sanctions (3.86), defendant (3.90), conviction (4.10), backwards (4.22), courts (4.24), defendants (4.26), court's (4.44), and correction (4.44).

While it would be simple to adjust these ratings in the MPQA dictionary going forward, we are naturally led to be concerned about existing work using MPQA that does not ex-

amine words contributing to overall sentiment. We note again that the use of word shift graphs of some kind would have exposed these problematic scores immediately.

For the labMT-LIWC comparison in Panel E of Figure 1 we examine the same matched word lists as before. The 10 happiest words in labMT matched by words in LIWC with score –1 are: trick (5.22), shakin (5.29), number (5.30), geek (5.34), tricks (5.38), defence (5.39), dwell (5.47), doubtless (5.92), numbers (6.04), shakespeare (6.88). From Panel F of Figure 2, the least happy 5 neutral words and happiest 5 neutral words in LIWC, matched in LabMT from LIWC words (i.e., using the word stems in LIWC to match across LabMT, directionality matters), are: negative (2.42), lack (3.16), couldn't (3.32), cannot (3.32), never (3.34), millions (7.26), couple (7.30), million (7.38), billion (7.56), millionaire (7.62). The least happy words in labMT with score +1 in LIWC that are matched by LIWC are: merrill (4.90), richardson (5.02), dynamite (5.04), careful (5.10), richard (5.26), silly (5.30), gloria (5.36), securities (5.38), boldface (5.40), treasury's (5.42). The +1 and –1 words in LIWC match some neutral words in labMT, which is not alarming. However, the problems with the 'neutral' words in the LIWC set are evident: these are not emotionally neutral words [39].

For the labMT-OL comparison in Panel E of Figure 1 we again examine the same matched word lists as before (except the neutral word list because OL has no explicit neutral words). The 10 happiest words in labMT matched by OL's negative list are: myth (5.90), puppet (5.90), skinny (5.92), jam (6.02), challenging (6.10), fiction (6.16), lemon (6.16), tenderness (7.06), joke (7.62), funny (7.92). The least happy words in labMT with score +1 in OL that are matched by OL are: defeated (2.74), defeat (3.20), envy (3.33), obsession (3.74), tough (3.96), dominated (4.04), unreal (4.57), striking (4.70), sharp (4.84), sensitive (4.86). Despite nearly twice as many negative words in OL as positive words (at odds with the frequency-dependent positivity bias of language [5]), after examining the words which are the most differently scored and seeing how quickly the labMT scores move into the neutral range, we can conclude that these dictionaries generally agree with the exception of only a few bad matches.

Our direct comparisons between the word scores in sentiment dictionaries, while perhaps tedious, have brought to light many problematic word scores. Our analysis also serves as a template for further comparisons of the words across new sentiment dictionaries. The six sentiment dictionaries under careful examination in the present study are further analyzed in the Supporting Information. Next, we examine how each sentiment dictionary can aid in understanding the sentiments contained in articles from the New York Times.

3.1 New York Times word shift analysis

The New York Times corpus [35] is split into 24 sections of the newspaper that are roughly contiguous throughout the data from 1987-2008. With each sentiment dictionary, we rate each section and then compute word shift graphs (described below) against the baseline, and produce a happiness ranked list of the sections.

To gain understanding of the sentiment expressed by any given text relative to another text, it is necessary to inspect the words which contribute most significantly by their emotional strength and the change in frequency of usage. We do this through the use of word shift graphs, which plot the percentage contribution of each word w from the sentiment dictionary (denoted $\delta h_{\text{ANEW}}(w)$) to the shift in average happiness between two texts, sorted by the absolute value of the contribution. We use word shift graphs to both

analyze a single text and to compare two texts, here focusing on comparing text within corpora. For a derivation of the algorithm used to make word shift graphs while separating the frequency and sentiment information, we refer the reader to Equations 2 and 3 in [14]. We consider both the sentiment difference and frequency difference components of $\delta h_{\text{ANEW}}(w)$ by writing each term of Eq. (1) as in [14]:

$$\delta h_{\text{ANEW}}(w) = 100 \frac{h_{\text{ANEW}}(w) - h^{\text{ref}}_{\text{ANEW}}}{h^{\text{comp}}_{\text{ANEW}} - h^{\text{ref}}_{\text{ANEW}}} \left[p(w)^{\text{comp}} - p(w)^{\text{ref}} \right]. \tag{2}$$

An in-depth explanation of how to interpret the word shift graph can also be found at http://hedonometer.org/instructions.html#wordshifts.

To both demonstrate the necessity of using word shift graphs in carrying out sentiment analysis, and to gain understanding about the ranking of New York Times sections by each sentiment dictionary, we look at word shift graphs for the 'Society' section of the newspaper from each sentiment dictionary in Figure 3, with the reference text being the whole of the New York Times. The 'Society' section happiness ranks 1, 1, 1, 18, 1, and 11 within the happiness of each of the 24 sections in the dictionaries labMT, ANEW, WK, MPQA, LIWC, and OL, respectively. These graphs show only the very top of the distributions which range in length from 1,030 (ANEW) to 13,915 words (WK).

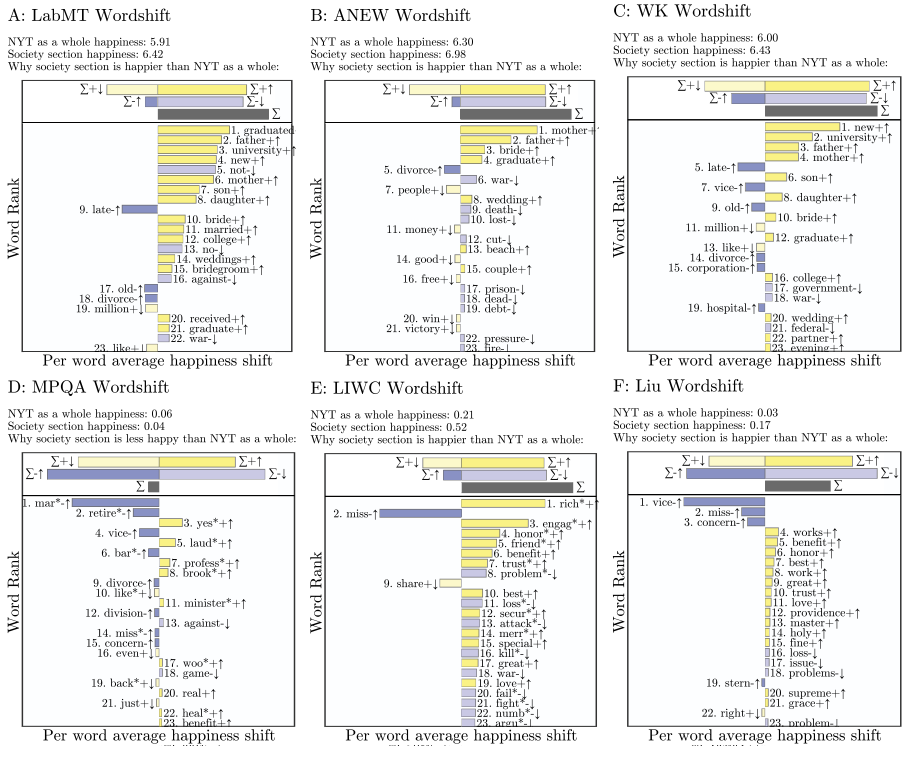

Figure 3 New York Times (NYT) 'Society' section shifted against the entire NYT corpus for each of the six dictionaries listed in tiles A-F. We provide a detailed analysis in Section 3.1. Generally, we are able to glean the greatest understanding of the sentiment texture associated with this NYT section using the labMT dictionary. We note that the labMT dictionary has the most coverage quantified by word match count (Figure in S3 Appendix), that we are able to identify and correct problematic words scores in the OL dictionary, and that we see that the MPQA dictionary disagrees entirely with the others because of an overly broad stem match.

First, using the labMT dictionary, we see that the words 'graduated', 'father', and 'university' top the list, which is dominated by positive words that occur more frequently (+ ↑). These more frequent positive words paint a clear picture of family life (relationships, weddings, and divorces), as well as university accomplishment (graduations and college). In general, we are able to observe with only these words that the 'Society' section is where we find the details of these events.

From the ANEW dictionary, we see that a few positive words have increased frequency, lead by 'mother', 'father', and 'bride'. Looking at this shift in isolation, we see only these words with three more ('graduate', 'wedding', and 'couple') that would lead us to suspect these topics are present in the 'Society' section.

The WK dictionary, with the most individual word scores of any sentiment dictionary tested, agrees with labMT and ANEW that the 'Society' section is the happiest section, with a somewhat similar set of words at the top: 'new', 'university', and 'father'. Low coverage of the New York Times corpus (see Figure S3) resulted in less specific words describing the 'Society' section, with more words that go down in frequency in the shift. With the words 'bride' and 'wedding' up, as well as 'university', 'graduate', and 'college', it is evident that the 'Society' section covers both graduations and weddings, in consensus with the other sentiment dictionaries.

The MPQA dictionary ranks the 'Society' section 18th of the 24 NYT sections, a departure from the other rankings, with the words 'mar*', 'retire*', and 'yes*' the top three contributing words (where '*' denotes a wildcard 'stem' match). Negative words increasing in frequency (– ↑) are the most common type near the top, and of these, the words with the biggest contributions are being scored incorrectly in this context (specifically words 'mar*', 'retire*', 'bar*', 'division', and 'miss*'). Looking more in depth at the problems created by the first out of context word match, we find 1,211 unique words match 'mar*'. The five most frequent, with counts in parenthesis, are married (36,750), marriage (5,977), marketing (5,382), mary (4,403), and mark (2,624). The score for these words in MPQA is –1, in stark contrast to the scores in other sentiment dictionaries (e.g., the labMT scores are 6.76, 6.7, 5.2, 5.88, and 5.48). These problems plague the MPQA dictionary for scoring the New York Times corpus, and without using word shift graphs would have gone completely unseen. In an attempt to fix contextual issues by fixing corpus-specific words, we remove 'mar*', 'retire*', 'vice', 'bar*', and 'miss*' and find that the MPQA dictionary ranks the Society section of the NYT at 15th of the 24 sections.

The second binary sentiment dictionary, LIWC, agrees well with the first three dictionaries and ranks the 'Society' section at the top with the words 'rich*', 'miss', and 'engage*' at the top of the list. We immediately notice that the word 'miss' is being used frequently in the 'Society' section in a different sense than was coded for in the LIWC dictionary: it is used in the corpus to mean 'the title prefixed to the name of an unmarried woman', but is scored as negative in LIWC (with the likely intended meaning 'to fail to reach an target or to acknowledge loss'). We would remove this word from LIWC for further analysis of this corpus (we would also remove the word 'trust' here). The words matched by 'miss*' aside, LIWC finds some positive words going up (+ ↑), with 'engage*' hinting at weddings. Without words that capture the specific behavior happening in the 'Society' section, we are unable to see anything about college, graduations, or marriages, and there is much less to be gained about the text from the words in LIWC than some of the other dictionaries we

have seen. Nevertheless, LIWC finds consensus with the 'Society' section ranked the top section, due in large part to a lack of negative words 'war' (rank 18) and 'fight*' (rank 22).

The OL sentiment dictionary departs from the consensus and ranks the 'Society' section at 11th out of the 24 sections. The top three words, 'vice', 'miss', and 'concern', contribute largely with respect to the rest of distribution, of which two are clearly being used in the wrong sense. For a more reasonable analysis we remove both 'vice' and 'miss' from the OL dictionary to score this text, and in doing so the happiness goes from 0.168 to 0.297, making the 'Society' section the second happiest of the 24 sections. Focusing on the words, we see that the OL dictionary finds many positive words increasing in frequency (+ ↑) that are mostly generic. In the word shift graph we do not find the wedding or university events as in sentiment dictionaries with more coverage, but rather a variety of positive language surrounding these events, for example, 'works' (4), 'benefit' (5), 'honor' (6), 'best' (7), 'great' (9), 'trust' (10), 'love' (11), etc. While this does not provide insight into the topics, the OL sentiment dictionary with fixes from the word shift graph analysis does provide details on the emotive words that make the 'Society' section one of the happiest sections.

In conclusion, we find that 4 of the 6 dictionaries score the 'Society' section at number 1, and in these cases we use the word shift graph to uncover the nuances of the language used. We find, unsurprisingly, that the most matches are found by the labMT dictionary, which is in part built from the NYT corpus (see S3 Appendix for coverage plots). Without as much corpus-specific coverage, we note that while the specifies of the text remain hidden, the LIWC and OL dictionaries still highlight the positive language in this section. Of the two that did not score the 'Society' section at the top, we are able to assess and repair the MPQA and the OL dictionaries by removing the words 'mar*', 'retire*', 'vice*', 'bar*', 'miss*' and 'vice', and 'miss', respectively. By identifying words used in the wrong sense/context using the word shift graph, we directly improve the sentiment score for the New York Times corpus from both MPQA and OL dictionaries closer to consensus. While the OL dictionary, with two corrections, agrees with the other dictionaries, the MPQA dictionary with five corrections still ranks the Society section of the NYT as the 15th happiest of the 24 sections.

In the first Figure in S4 Appendix we show scatterplots for each comparison, and compute the Reduced Major Axes (RMA) regression fit [40]. In the second Figure in S4 Appendix we show the sorted bar chart from each sentiment dictionary.

3.2 Movie reviews classification and word shift graph analysis

For the movie reviews corpus, we first provide insight into the language differences and secondly perform binary classification of positive and negative reviews. The entire dataset consists of 1,000 positive and 1,000 negative reviews, as rated with 4 or 5 stars and 1 or 2 stars, respectively. We show how well each sentiment dictionary covers the review database in Figure 4. The average review length is 650 words, and we plot the distribution of review lengths in S5 Appendix. We average the sentiment of words in each review individually, using each sentiment dictionary. We also combine random samples of N positive or N negative reviews for N varying from 2 to 900 on a logarithmic scale, without replacement, and rate the combined text. With an increase in the size of the text, we expect that the dictionaries will be better able to distinguish positive from negative. The simple statistic we use to describe this ability is the percentage of distributions that overlap the average.

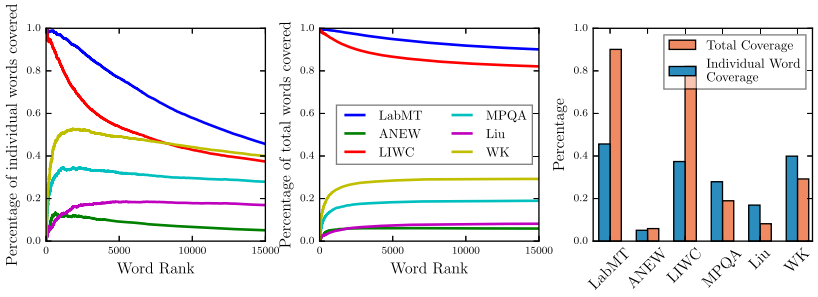

Figure 4 Coverage of the words in the movie reviews by each of the dictionaries. We observe that the labMT dictionary has the highest coverage of words in the movie reviews corpus both across word rank and cumulatively. The LIWC dictionary has initially high coverage since it contains some high-frequency function words, but quickly drops off across rank. The WK dictionary coverage increases across word rank and cumulatively, indicating that it contains a large number of less common words in the movie review corpus. The OL, ANEW, and MPQA have a cumulative coverage of less than 20% of the lexicon.

To analyze which words are being used by each sentiment dictionary, we compute word shift graphs of the entire positive corpus versus the entire negative corpus in Figure 5. Across the board, we see that a decrease in negative words is the most important word type for each sentiment dictionary, with the word 'bad' being the top word for every sentiment dictionary in which it is scored (ANEW does not have it). Other observations that we can make from the word shift graphs include a few words that are potentially being used out of context: 'movie', 'comedy', 'plot', 'horror', 'war', 'just'.

In the lower right panel of Figure 6, the percentage overlap of positive and negative review distributions presents us with a simple summary of sentiment dictionary performance on this tagged corpus. The ANEW dictionary stands out as being considerably less capable of distinguishing positive from negative. In order, we then see WK is slightly better overall, labMT and LIWC perform similarly better than WK overall, and then MPQA and OL are each a degree better again, across the review lengths (see below for hard numbers at 1 review length). Two Figures in the S5 Appendix show the distributions for 1 review and for 15 combined reviews.

Classifying single reviews as positive or negative, the F1-scores are: labMT 0.63, ANEW 0.36, LIWC 0.53, MPQA 0.66, OL 0.71, and WK 0.34 (see Table S4). We roughly confirm a rule-of-thumb that 10,000 words are enough to score with a sentiment dictionary confidently, with all dictionaries except MPQA and ANEW achieving 90% accuracy with this many words. We sample the number of reviews evenly in log space, generating sets of reviews with average word counts of 4,550, 6,500, 9,750, 16,250, and 26,000 words. Specifically, the number of reviews necessary to achieve 90% accuracy is 15 reviews (9,750 words) for labMT, 100 reviews (65,000 words) for ANEW, 10 reviews (6,500 words) for LIWC, 10 reviews (6,500 words) for MPQA, 7 reviews (4,550 words) for OL, and 25 reviews (16,250 words) for WK.

While we are analyzing the movie review classification, which has ground truth labels, we will take a moment to further support our claims for the inaccuracy of these methods at the sentence level. The OL dictionary, with the highest performance classifying individual movie reviews of the 6 dictionaries tested in detail, performs worse than guessing at classifying individual sentences in movie reviews. Specifically, 76.9/74.2% of sentences in the positive/negative reviews sets have words in the OL dictionary, and of these OL achieves

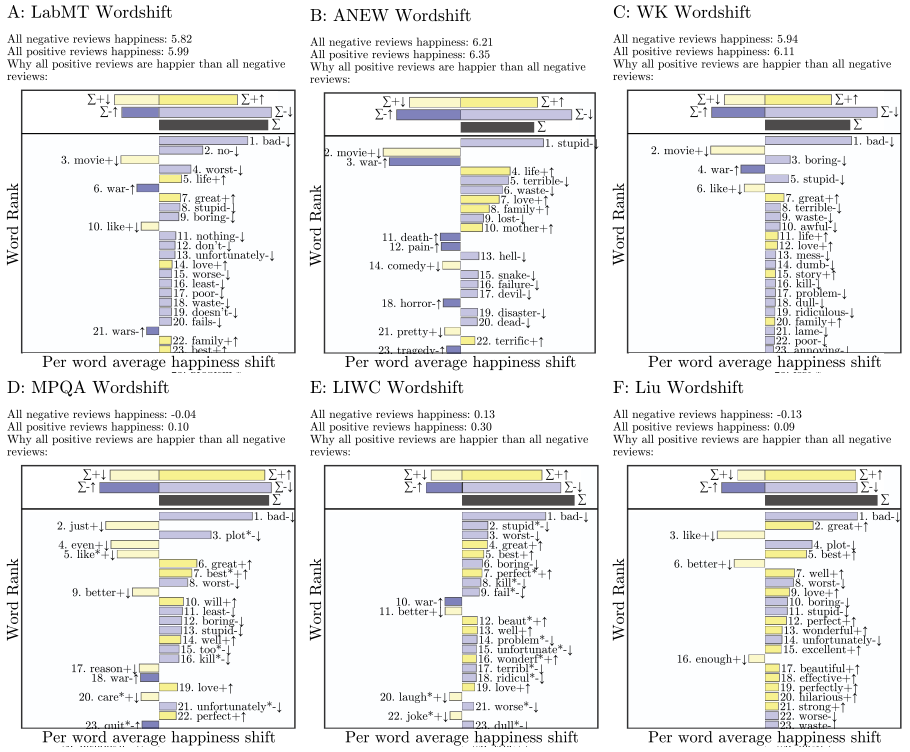

Figure 5 Word shift graphs for the movie review corpus. By analyzing the words that contribute most significantly to the sentiment score produced by each sentiment dictionary we are able to identify words that are problematic for each sentiment dictionary at the word-level, and generate an understanding of the emotional texture of the movie review corpus. Again we find that coverage of the lexicon is essential to produce meaningful word shift graphs, with the labMT dictionary providing the most coverage of this corpus and producing the most detailed word shift graphs. All words on the left hand side of these word shift graphs are words that individually made the positive reviews score more negatively than the negative reviews, and the removal of these words would improve the accuracy of the ratings given by each sentiment dictionary. In particular, across each sentiment dictionary the word shift graphs show that domain-specific words such as 'war' and 'movie' are used more frequently in the positive reviews and are not useful in determining the polarity of a single review.

an F1-score of 0.44. The results for each sentiment dictionary are included in Table S5, with an average (median) F1 score of 0.42 (0.45) across all dictionaries. While these results do cast doubt on the ability to classify positive and negative reviews from single sentences using dictionary based methods, we note that we need not expect the sentiment of individual sentences to be strongly correlated with the overall review polarity.

3.3 Google books time series and word shift analysis

We use the Google books 2012 dataset with all English books [37], from which we remove part of speech tagging and split into years. From this, we make time series by year, and word shift graphs of decades versus the baseline. In addition, to assess the similarity of each time series, we produce correlations between each of the time series.

Despite claims from research based on the Google Books corpus [41], we keep in mind that there are several deep problems with this beguiling data set [42]. Leaving aside these issues, the Google Books corpus nevertheless provides a substantive test of our six dictionaries.

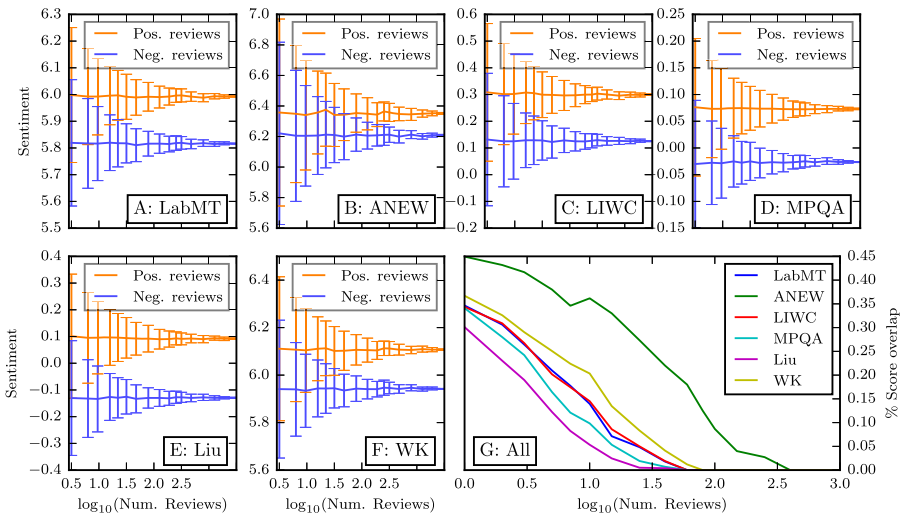

Figure 6 The score assigned to increasing numbers of reviews drawn from the tagged positive and negative sets. For each sentiment dictionary we show mean sentiment and 1 standard deviation over 100 samples for each distribution of reviews in Panels A-F. For comparison we compute the fraction of the distributions that overlap in Panel G. At the single review level for each sentiment dictionary this simple performance statistic (fraction of distribution overlap) ranks the OL dictionary in first place, the MPQA, LIWC, and labMT dictionaries in a second place tie, WK in fifth, and ANEW far behind. All dictionaries require on the order of 1,000 words to achieve 95% classification accuracy.

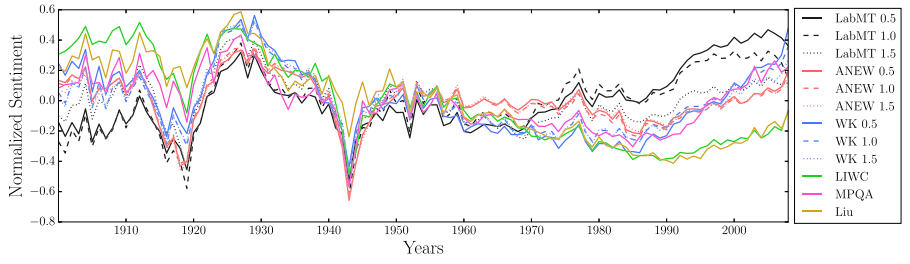

Figure 7 Google Books sentiment time series from each sentiment dictionary, with stop values of 0.5, 1.0, and 1.5 from the dictionaries with word scores in the 1-9 range. To normalize the sentiment score, we subtract the mean and divide by the absolute range. We observe that each time series has increased variance, with a few pronounced negative time periods, and trending positive towards the end of the corpus. The score of labMT varies substantially with different stop values, although remaining highly correlated, and finds absolute lows near the World Wars. The LIWC and OL dictionaries trend down towards 1990, dipping as low as the war periods.

In Figure 7, we plot the sentiment time series for Google Books. Three immediate trends stand out: a dip near the Great Depression, a dip near World War II, and a general upswing in the 1990's and 2000's. From these general trends, a few dictionaries waver: OL does not dip as much for WW2, OL and LIWC stay lower in the 1990's and 2000's, and labMT with $\Delta_h = 0.5, 1.0$ go downward near the end of the 2000's. We take a closer look into the 1940's to see what each sentiment dictionary is picking up in Google Books around World War 2 in Figure in S6 Appendix.

In each panel of the word shift Figure in S6 Appendix, we see that the top word making the 1940's less positive than the rest of Google Books is 'war', which is the top contributor for every sentiment dictionary except OL. Rounding out the top three contributing words

are 'no' and 'great', and we infer that the word 'great' is being seen from mention of 'The Great Depression' or 'The Great War'. All dictionaries but ANEW have 'great' in the top 3 words, and each sentiment dictionary could be made more accurate if we remove this word.

In Panel A of the 1940's word shift Figure in S6 Appendix, beyond the top words, increasing words are mostly negative and war-related: 'against', 'enemy', 'operation', which we could expect from this time period.

In Panel B, the ANEW dictionary scores the 1940's of Google Books lower than the baseline as well, finding 'war', 'cancer', and 'cell' to be the most important three words. With only 1,030 words, there is not enough coverage to see anything beyond the top word 'war,' and the shift is dominated by words that go down in frequency.

In Panel C, the WK dictionary finds the 1940's to be slightly less happy than the baseline, with the top three words being 'war', 'great', and 'old'. We see many of the same war-related words as in labMT, as well as some positive words like 'good' and 'be' are up in frequency. The word 'first' could be an artifact of first aid, a claim that could be substantiated with further analysis of the Google Books corpus at the 2-gram level but beyond the scope of this manuscript.

In Panel D, the MPQA dictionary rates the 1940's slightly less happy than the baseline, with the top three words being 'war', 'great', and 'differ*'. Beyond the top word 'war', the score is dominated by words decreasing in frequency, with only a few words up in frequency. Without specific words increasing in frequency as contextual guides, it is difficult to obtain a good glance at the nature of the text. Once again, having a higher coverage of the words in the corpus enables understanding.

In Panel E, the LIWC dictionary rates the 1940's nearly the same as the baseline, with the top three words being 'war', 'great', and 'argu*'. When the scores are nearly the same, although the 1940's are slightly higher in happiness here, the word shift is a view into how the words of the reference and comparison text vary. In addition to a few war related words being up and bringing the score down ('fight', 'enemy', 'attack'), we see some positive words also being up that could also be war related: 'certain', 'interest', and 'definite'. Although LIWC does not manage to find World War II as a low point of the 20th century, the words that contribute to LIWCs score for the 1940's compared to all years are useful in understanding the corpus.

In Panel F, the OL dictionary rates the 1940's as happier than the baseline, with the top three words being 'great', 'support', and 'like'. With 7 positive words up, and 1 negative word up, we see how the OL dictionary misses the war without the word 'war' itself and with only 'enemy' contributing from the words surrounding the conflict. The nature of the positive words that are up is unclear, and could justify a more detailed analysis of why the OL dictionary fails here.

3.4 Twitter time series analysis

For Twitter data, we use the Gardenhose feed, a random 10% of the full stream. We store data on the Vermont Advanced Computing Core (VACC), and process the text first into hash tables (with approximately 8 million unique English words each day) and then into word vectors for each 15 minutes, for each sentiment dictionary tested. From this, we build sentiment time series for time resolutions of 15 minutes, 1 hour, 3 hours, 12 hours, and 1 day. Along with the raw time series, we compute correlations between each time series to assess the similarity of the ratings between dictionaries.

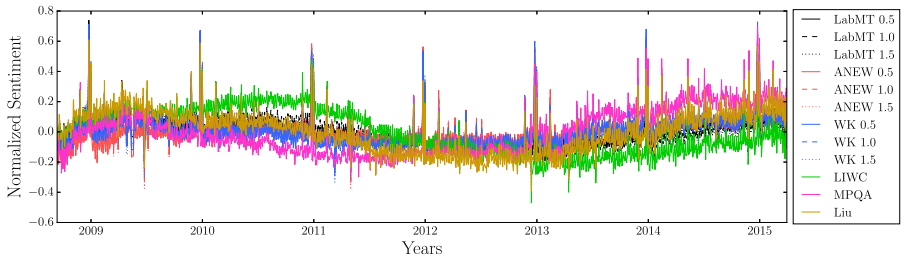

Figure 8 Normalized sentiment time series on Twitter using Δ_h of 1.0 for all dictionaries. To normalize the sentiment score, we subtract the mean and divide by the absolute range. The resolution is 1 day, and draws on the 10% gardenhose sample of public Tweets provided by Twitter. All of the dictionaries exhibit wide variation for very early Tweets, and from 2012 onward generally track together strongly with the exception of MPQA and LIWC. The LIWC and MPQA dictionaries show opposite trends: a rise until 2012 with a decline after 2012 from LIWC, and a decline before 2012 with a rise afterwards from MPQA. To analyze the trends we look at the words driving the movement across years using word shift Figures in S7 Appendix. An interactive version of this Figure using the labMT dictionary be found at http://hedonometer.org.

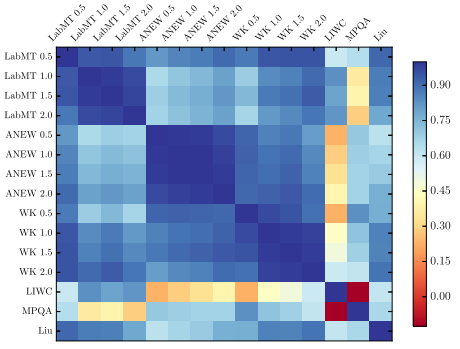

Figure 9 Pearson's *r* correlation between daily resolution Twitter sentiment time series for each sentiment dictionary. We see that there is strong agreement within dictionaries, with the biggest differences coming from the stop value of $\Delta h = 0.5$ for labMT and WK. The labMT and OL dictionaries do not strongly disagree with any of the others, while LIWC is the least correlated overall with other dictionaries. labMT and OL correlate strongly with each other, and disagree most with the ANEW, LIWC, and MPQA dictionaries. The two least correlated dictionaries are the LIWC and MPQA dictionaries. Again, since there is no publicly accessible ground truth for Twitter sentiment, we compare dictionaries against the others, and look at the words. With these criteria, we find the labMT and OL dictionaries to be the most robust with Tweets.

In Figure 8, we present a daily sentiment time series of Twitter processed using each of the dictionaries being tested. With the exception of LIWC and MPQA we observe that the dictionaries generally track well together across the entire range. A strong weekly cycle is present in all, although muted for ANEW. An interactive version of the plot in Figure 8 can be found at http://hedonometer.org.

We plot the Pearson's correlation between all time series in Figure 9, and confirm some of the general observations that we can make from the time series. Namely, the LIWC and MPQA time series disagree the most from the others, and even more so with each other. Generally, we see strong agreement within dictionaries with varying stop values Δh.

The time series from each sentiment dictionary exhibits increased variance at the start of the time frame, when Twitter volume was much lower in 2008 and into 2009. As more people join Twitter and the Tweet volume increases through 2010, we see that LIWC rates the text as happier, while the rest start a slow decline in rating that is led by MPQA in the negative direction. In 2010, the LIWC dictionary is more positive than the rest with words like 'haha', 'lol' and 'hey' being used more frequently and swearing being less frequent than all years of Twitter put together. The other dictionaries with more coverage find a decrease in positive words to balance this increase, with the exception of MPQA which finds many

negative words going up in frequency (see 2010 word shift Figure in Appendix S7). All of the dictionaries agree most strongly in 2012, all finding a lot of negative language and swearing that brings scores down (see 2012 word shift Figure in Appendix S7). From the bottom at 2012, LIWC continues to go downward while the others trend back up. The signal from MPQA jumps to the most positive, and LIWC does start trending back up eventually. We analyze the words in 2014 with a word shift against all 7 years of Tweets for each sentiment dictionary in each panel in the 2014 word shift Figure in Appendix S7: A. labMT scores 2014 as less happy with more negative language. B. ANEW finds it happier with a few positive words up. C. WK finds it happier with more negative words (like labMT). D. MPQA finds it more positive with less negative words. E. LIWC finds it less positive with more negative and less positive words. F. OL finds it to be of the same sentiment as the background with a balance in positive and negative word usage. From these word shift graphs, we can analyze which words cause MPQA and LIWC to disagree with the other dictionaries: the disagreement of MPQA is again marred by broad stem matches, and the disagreement of LIWC is due to a lack of coverage.

3.5 Brief comparison to machine learning methods

We implement a Naive Bayes (NB) classifier (sometimes harshly called idiot Bayes [43]) on the tagged movie review dataset to examine how individual words contribute in machine learning classification. While more advanced methods have better classification accuracy, we focus on the simplest example to illustrate how analysis at the individual word level aids in understanding sentiment analysis scores. We use a 70/30 split of the data into training and out-of-sample testing sets, and examine the model performance on 100 random permutations of this split. Again following standard best-practice, we remove the top 30 ranked words ('stop words') from the 5,000 most frequent words, and use the remaining 4,970 words in our classifier for maximum performance (we observe a 0.5% improvement). Our implementation is analogous to those found in common Python natural language processing packages (see 'NLTK' or 'TextBlob' in [44]).

As we should expect, at the level of single review, NB outperforms the dictionary-based methods with a classification accuracy of 72.4-76.1% averaged over 100 trials. As the number of reviews is increased, the overlap from NB decreases, and using our simple 'fraction overlapping' metric, the error drops to 0 with more than 200 reviews. Overall, with Naive Bayes we are able to classify a higher percentage of individual reviews correctly, but with more variance.

In the two Tables in S8 Appendix we compute the words which the NB classifier uses to classify all of the positive reviews as positive, and all of the negative reviews as positive. The Natural Language Toolkit (NLTK [44]) implements a method to obtain the 'most informative' words, by taking the ratio of the likelihood of words between all available classes, and looking for the largest ratio:

$$\max_{\text{all words } w} \frac{P(w|c_i)}{P(w|c_j)} \tag{3}$$

for all combinations of classes c_i, c_j. This is possible because of the 'naive' assumption that feature (word) likelihoods are independent, resulting in a classification metric that is linear for each feature. In S8 Appendix, we provide the derivation of this linearity structure.

We find that the trained NB classifier relies heavily on words that are very specific to the training set including the names of actors of the movies themselves, making them useful as classifiers but not in understanding the nature of the text. We report the top 10 words for both positive and negative classes using both the ratio and difference methods in Table in S8 Appendix. To classify a document using NB, we use the frequency of each word in the document in conjunction with the probability that that word occurred in each labeled class c_i. While steps can be taken to avoid this type of over-fitting, it is an ever-present danger that remains hidden without word shift graphs or similar.

We next take the movie-review-trained NB classifier and use it to classify the New York Times sections, both by ranking them and by looking at the words (the above ratio and difference weighted by the occurrence of the words). We ranked the Sections 5 different times, and among those find the 'Television' section both by far the happiest, and by far the least happy in independent tests. We show these rankings and report the top 10 words used to score the 'Society' section in Table S3.

We thus see that the NB classifier, a linear learning method, may perform poorly when assessing sentiment outside of the corpus on which it is trained. In general, performance will vary depending on the statistical dissimilarity of the training and novel corpora. Added to this is the inscrutability of black box methods: while susceptible to the aforementioned difficulty, nonlinear learning methods (unlike NB) also render detailed examination of how individual words contribute to a text's score more difficult.

4 Conclusion

We have shown that measuring sentiment in various corpora presents unique challenges, and that sentiment dictionary performance is situation dependent. Across the board, the ANEW dictionary performs poorly, and the continued use of this sentiment dictionary with clearly better alternatives is a questionable choice. We have seen that the MPQA dictionary does not agree with the other five dictionaries on the NYT corpus and Twitter corpus due to a variety of context, word sense, phrase, and stem matching issues, and we would not recommend using this sentiment dictionary. While the OL achieves the highest binary classification accuracy, in comparison to labMT, the WK, LIWC, and OL dictionaries fail to provide much detail in corpora where their coverage is lower, including all four corpora tested, the main goal of our analysis. Sufficient coverage is essential for producing meaningful word shift graphs and thereby enabling deeper understanding.

In each case, to analyze the output of the dictionary method, we rely on the use of word shift graphs. With this tool, we can produce a finer grained analysis of the lexical content, and we can also detect words that are used out of context and can mask them directly. It should be clear that using any of the dictionary-based sentiment detecting methods without looking at how individual words contribute is indefensible, and analyses that do not use word shift graphs or similar tools cannot be trusted. The poor word shift performance of binary dictionaries in particular gravely limits their ability to reveal underlying stories.

In sum, we believe that dictionary-based methods will continue to play a powerful role — they are fast and well suited for web-scale data sets — and that the best instruments will be based on dictionaries with excellent coverage and continuum scores. To this end, we urge that all dictionaries should be regularly updated to capture changing lexicons, word usage, and demographics. Looking further ahead, a move from scoring words to scoring both phrases and words with senses should realize considerable improvement for many

languages of interest. With phrase dictionaries, the resulting phrase shift graphs will allow for a more nuanced and detailed analysis of a corpus's sentiment score [6], ultimately affording clearer stories for sentiment dynamics.

Additional material

Additional file 1: Supplementary Material. (pdf)

Funding
Not applicable.

Abbreviations
NYT, New York Times; MT, Mechanical Turk; ML, Machine Learning; BSD, Berkeley Software Distribution; CC, Creative Commons; GNU, GNU's Not Unix; GNU GPL, GNU General Public License; RMA, Reduced Major Axes; VACC, Vermont Advanced Computing Core; NB, Naive Bayes; NLTK, Natural Language Toolkit; labMT, language assessment by Mechanical Turk [5]; ANEW, Affective Norms of English Words [7]; WK, Warriner and Kuperman rated words from SUBTLEX by Mechanical Turk [11]; MPQA, The Multi-Perspective Question Answering (MPQA) Subjectivity Dictionary [9]; LIWC{01,07,15}, Linguistic Inquiry and Word Count [8]; OL, Opinion Lexicon, developed by Bing Liu [10]; PANAS-X, The Positive and Negative Affect Schedule Expanded [17]; AFINN, Words manually rated –5 to 5 with impact scores by Finn Nielsen [20]; GI, General Inquirer [21]; WDAL, Whissel's Dictionary of Affective Language [22]; EmoLex, NRC Word-Emotion Association Lexicon [23]; MaxDiff, NRC MaxDiff Twitter Sentiment Lexicon[24]; HashtagSent, NRC Hashtag Sentiment Lexicon [25]; Sent140Lex, NRC Sentiment140 Lexicon [26]; SOCAL, Semantic Orientation CALculator [27]; VADER, Valence Aware Dictionary for sEntiment Reasoning [31]; USent, set of emoticons and inappropriate words that extend MPQA [33].

Availability of data and materials
In the following GitHub repository associated with our paper, we include all of the sentiment dictionary data (except LIWC). In addition, we also provide the scripts to reproduce our analysis. The repository is publicly available on GitHub at https://github.com/andyreagan/sentiment-analysis-comparison.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AJR, CMD, BT, JRW, and PSD designed research. AJR performed research. JRW contributed analytic tools. AJR, CMD, BT, and PSD analyzed data. AJR, CMD, and PSD wrote the paper. All authors read and approved the final manuscript.

Author details
[1]University of Vermont, 85 South Prospect St, Burlington, VT 05405, USA. [2]The MITRE Corporation, 7525 Colshire Drive, McLean, VA 22102, USA. [3]Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA.

Publisher's Note

Received: 24 January 2017 Accepted: 14 September 2017 Published online: 30 October 2017

References

1. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1-8
2. Si J, Mukherjee A, Liu B, Li Q, Li H, Deng X (2013) Exploiting topic based Twitter sentiment for stock prediction. In: ACL, vol 2, pp 24-29
3. Chung S, Liu S (2011) Predicting stock market fluctuations from Twitter. Berkeley, California
4. Ruiz EJ, Hristidis V, Castillo C, Gionis A, Jaimes A (2012) Correlating financial time series with micro-blogging activity. In: Proceedings of the fifth ACM international conference on web search and data mining. ACM, New York, pp 513-522
5. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, Mitchell L, Harris KD Kloumann IM, Bagrow JP, Megerdoomian K, McMahon MT, Tivnan BF, Danforth CM (2015) Human language reveals a universal positivity bias. Proc Natl Acad Sci USA 112(8):2389-2394
6. Alajajian SE, Williams JR, Reagan AJ, Alajajian SC, Frank MR, Mitchell L, Lahne J, Danforth CM, Dodds PS (2017) The Lexicocalorimeter: gauging public health through caloric input and output on social media. PLoS ONE 12(2):e0168893. doi:10.1371/journal.pone.0168893
7. Bradley MM Lang PJ (1999) Affective norms for english words (ANEW): stimuli, instruction manual and affective ratings. Technical report c-1, University of Florida, Gainesville, FL
8. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: LIWC 2001. Erlbaum, Mahway
9. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of human language technologies conference/conference on empirical methods in natural language processing (HLT/EMNLP 2005)
10. Liu B (2010) Sentiment analysis and subjectivity. In: Handbook of natural language processing 2nd edn. pp 627-666

11. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav Res Methods 45(4):1191-1207. doi:10.3758/s13428-012-0314-x
12. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp 1631-1642. Citeseer
13. Dodds PS, Danforth CM (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. J Happiness Stud 11(4):441-456. doi:10.1007/s10902-009-9150-9
14. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE 6(12):26752. doi:10.1371/journal.pone.0026752
15. Ribeiro FN, Araújo M, Gonçalves P, Gonçalves MA, Benevenuto F (2016) SentiBench — a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci 5(1):23. doi:10.1140/epjds/s13688-016-0085-1
16. Liu B (2012) Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. Morgan & Claypool, San Rafael, pp 1-167
17. Watson D, Clark LA (1999) The PANAS-X: manual for the positive and negative affect schedule-expanded form. PhD thesis, University of Iowa
18. De Smedt T, Daelemans W (2012) Pattern for Python. J Mach Learn Res 13(1):2063-2067
19. Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC'10, pp 2200-2204
20. Nielsen F (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Rowe M, Stankovic M, Dadzie A-S, Hardey M (eds) Proceedings of the ESWC2011 workshop on 'making sense of microposts': big things come in small packages CEUR workshop proceedings, vol 718, pp 93-98
21. Stone PJ, Dunphy DC, Smith MS (1966) The general inquirer: a computer approach to content analysis. MIT Press, Cambridge
22. Whissell C, Fournier M, Pelland R, Weir D, Makarec K (1986) A dictionary of affect in language: iv. Reliability, validity, and applications. Percept Mot Skills 62(3):875-888
23. Mohammad SM Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436-465
24. Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723-762
25. Zhu X, Kiritchenko S, Mohammad SM (2014) NRC-Canada-2014: recent improvements in the sentiment analysis of tweets. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 443-447. Citeseer
26. Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the seventh international workshop on semantic evaluation exercises (SemEval-2013), Atlanta, Georgia, USA
27. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267-307
28. Cambria E, Olsher D, Rajagopal D (2014) Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence. AAAI Press, Menlo Park, pp 1515-1521
29. Gonçalves P, Araújo M, Benevenuto F, Cha M (2013) Comparing and combining sentiment analysis methods. In: Proceedings of the first ACM conference on online social networks. ACM, New York, pp 27-38.
30. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61(12):2544-2558
31. Hutto CJ, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media. AAAI Press, Menlo Park
32. Levallois C (2013) Umigon: sentiment analysis for tweets based on terms lists and heuristics. In: Second joint conference on lexical and computational semantics (*SEM), vol 2, pp 414-417
33. Pappas N, Katsimpras G, Stamatatos E (2013) Distinguishing the popularity between topics: a system for up-to-date opinion retrieval and mining in the web. In: International conference on intelligent text processing and computational linguistics. Springer, Berlin, pp 197-209
34. Poria S, Gelbukh A, Hussain A, Howard N, Das D, Bandyopadhyay S (2013) Enhanced senticnet with affective labels for concept-based opinion mining. IEEE Intell Syst 28(2):31-38
35. Sandhaus E (2008) The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia
36. Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL
37. Lin Y, Michel J-B, Aiden EL, Orwant J, Brockman W, Petrov S (2012) Syntactic annotations for the Google books ngram corpus. In: Proceedings of the ACL 2012 system demonstrations, pp 169-174. Association for Computational Linguistics
38. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM (2013) The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. PLoS ONE 8(5):64417. doi:10.1371/journal.pone.0064417
39. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, Mitchell L, Harris KD, Kloumann IM, Bagrow JP, Megerdoomian K, McMahon MT, Tivnan BF, Danforth CM (2015) Reply to Garcia et al.: common mistakes in measuring frequency-dependent word characteristics. Proc Natl Acad Sci USA 112(23):2984-2985. http://www.pnas.org/content/112/23/E2984.full.pdf. doi:10.1073/pnas.1505647112
40. Rayner JMV (1985) Linear relations in biomechanics: the statistics of scaling functions. J Zool 206:415-439
41. Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL (2011) Quantitative analysis of culture using millions of digitized books. Science 331(6014):176-182
42. Pechenick EA, Danforth CM, Dodds PS (2015) Characterizing the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. arXiv preprint arXiv:1501.00960
43. Hand DJ, Yu K (2001) Idiot's Bayes — not so stupid after all? Int Stat Rev 69(3):385-398
44. Bird S (2006) Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on interactive presentation sessions, pp 69-72. Association for Computational Linguistics