Four best practices for measuring news sentiment using ‘off-the-shelf’ dictionaries: a large-scale p-hacking experiment

Chung-hong Chan, Joseph Bajjalieh, Loretta Auvil, Hartmut Wessler, Scott Althaus, Kasper Welbers, Wouter van Atteveldt, Marc Jungblut[1]

CCR 3 (1): 1–27

DOI: 10.5117/CCR2021.1.001.CHAN

Abstract

We examined the validity of 37 sentiment scores based on dictionary-based methods using a large news corpus and demonstrated the risk of generating a spectrum of results with different levels of statistical significance by presenting an analysis of relationships between news sentiment and U.S. presidential approval. We summarize our findings into four best practices: 1) use a suitable sentiment dictionary; 2) do not assume that the validity and reliability of the dictionary is ‘built-in’; 3) check for the influence of content length and 4) do not use multiple dictionaries to test the same statistical hypothesis.

Keywords: sentiment analysis, p-hacking, news sentiment, agenda setting, text-as-data, validity

This paper uses a p-hacking experiment to demonstrate how different conclusions can be drawn using an array of 37 different dictionary-based sentiment scores from the same corpus. The two purposes of this paper are to 1) show the often overlooked validity problem of using these sentiment scores and; 2) suggest ways to mitigate the problem.

The main focus of this paper is dictionary-based sentiment analysis. It is a technique that uses a *dictionary* (list of words) to classify a piece of text by positive or negative sentiment.[2] The method was proposed as a solution in computer-assisted content analysis (Stone and Hunt 1963) and later adopted as a marketing tool by computer scientists. For example, one of the earliest papers in computer science literature on dictionary-based methods summarizes the polarity of user reviews of the products of an online shop (Hu and Liu 2004). Such applications were subsequently extrapolated for new analysis. Following previous studies (e.g. Ribeiro et al. 2016; Boukes et al. 2019), we call these applications *'off the shelf'* to mark the fact that researchers use dictionaries developed by other scholars without adjusting them for their own particular use.

Most of these dictionaries were not developed and validated for news texts, but researchers still use them in news analysis. This off-the-shelf dictionary-based sentiment analysis has been used quite heavily in political communication literature (e.g. Boukes et al. 2019; Young and Soroka 2012). New dictionaries such as Lexicoder (Young and Soroka 2012), VADER (Gilbert and Hutto 2014) and crowd-sourcing-based sentiment dictionaries (Haselmayer and Jenny 2016) were developed for application in communication science.

The advantages of these off-the-shelf methods are obvious: compared with traditional content analysis, these methods require no human input. In addition, the results are very easy to interpret. Moreover, in the primary studies dealing with dictionary development, some developers found very strong agreement between dictionary-based classification and human judgments in the contexts of their intended applications (e.g. Haselmayer and Jenny 2016; Gilbert and Hutto 2014; Young and Soroka 2012). Because of their apparent validity, many authors use these off-the-shelf sentiment dictionaries in their work with their own data, assuming that such an application should obtain similar levels of reliability and validity. However, scholars have criticized such use of off-the-shelf dictionary-based methods on two fronts: methodological and theoretical.

Methodologically, these sentiment analysis tools rely on two very simple assumptions: the bag-of-words assumption and the additivity assumption (Young and Soroka 2012). The bag-of-words assumption maintains that the order of the words in a text does not matter. Therefore, *'my cat is bad'* has the same sentiment level as its nonsensical rearrangements, such as *'bad my is cat'* and *'is my cat bad'.* Many, but not all, of these sentiment dictionaries do not consider the grammatical functions of words and even suggest converting all text to lowercase. One example is the inclusion of the word *trump* (as a verb as in the sentence '*machine learning methods trump dictionary-based methods*' or as a noun as in the sentence '*he plays the trump*') as a positive word in Bing Liu's dictionary (Hu and Liu 2004). When the grammatical functions of the word *trump* are ignored, as with the bag-of-words assumption, the sentence 'Trump is bad', wherein 'Trump' is a proper noun, is rated as neutral (the negativity of the word '*bad*' is canceled by '*trump*') while these same parameters situate the similarly constructed sentence 'Hillary is bad' as negative. Meanwhile, the additivity assumption maintains that text with a higher frequency of sentiment words has a higher level of actual sentiment. For example, '*my cat is bad and ugly*' is more negative than '*my cat is bad*'. This assumption usually ignores grammatical elements such as adverbs (e.g. '*my cat is very bad*' should be more negative than '*my cat is bad*', but most methods cannot handle the amplification effect of the adverb '*very*'). Most widely used dictionaries have acknowledged the weaknesses of these two assumptions. For example, Lexicoder (Young and Soroka 2012) provides a negated version of the dictionary (e.g. "not good") and an R preprocessing script to to remove special cases of language use (e.g. "good bye" should not be classified as positive). Many older ones, e.g. Bing Liu and LIWC (Tausczik and Pennebaker 2009), still rely on these two simple assumptions.

Moreover, off-the-shelf dictionary-based methods are sensitive to the features of source material, a limitation known as the domain-specificity problem. Previous benchmarks revealed that these methods demonstrated limited validity and reliability when applied to new datasets (González-Bailón and Paltoglou 2015; Ribeiro et al. 2016). This domain-specificity problem was addressed in the literature with technical solutions such as machine learning methods, which have been proposed (González-Bailón and Paltoglou 2015) and further developed (Rudkowsky et al. 2018). Other scholars suggest tuning dictionaries according to the source material (Diesner and Evans 2015; Grimmer and Stewart 2013) by, for example, adding domain-specific words to an existing dictionary and/or deleting words that have a different connotation in a new domain. In addition, Barberá et al. (2016) criticize these methods as 'independent of any actual human input on the document level'. It is possible to revalidate the performance of dictionary-based methods by human coding for every application (Grimmer and Stewart 2013; Ribeiro et al. 2016).

Beyond the methodological criticism, some scholars also question what dictionary-based methods actually measure in theoretical terms. For this, we need to go back to the fundamental question of '*what is sentiment*?'. According to the literature, 'sentiment' can mean different things (Puschmann and Powell 2018). For example, computer science literature defines 'sentiment' as the writer's 'appraisal or feelings towards an entity or an event' (Liu 2010; and a similar definition by Munezero et al. 2014) because the original intended use case of such tools was for product reviews with obvious targets (i.e. obvious entities or events). Other definitions include ‘affect expressed in a text’ and ‘the emotional state of a text’s author’ (Puschmann and Powell 2018, p.1). Puschmann and Powell (2018) argue that the ‘measurement of something called ’sentiment’ frequently fails to establish what sentiment might actually mean’. They base their criticism on the fact that researchers have used sentiment analysis to extract subjective emotional states from raw text using tools originally intended for uncovering the polarity of product reviews. The original developers of LIWC (Tausczik and Pennebaker 2009), for example, argue that language and behavior are linked and thus that their dictionary-based method can infer the emotional states of authors. However, some computer scientists reject such inference (Liu 2010; Pang and Lee 2008).

In this study, we used a simpler definition of sentiment as ‘emotions expressed in a text.’[3] In this understanding, sentiment is communicated through text, regardless of whether it reflects the actual subjective state of the text’s author. More specifically, we define news sentiment as ‘emotions expressed in a news article’. This definition does not include any target or inference, and is in line with the tradition in communication science of studying news tone, news negativity, news frames and ‘media affect’ (Young and Soroka 2012). We share the conviction of some computer scientists that it is very difficult to infer an author’s emotional state from a text and thus sentiment might reflect the subjective state of the text’s author. Authors can deliberately choose to express something that does not reflect their mood. Moreover, when we study journalistic text, it is difficult to attribute a piece of work to one author because a piece of news text can be an intellectual product of many people, such as reporters, journalistic assistants, copy-editors, fact checkers and editors. Here, it is helpful to note that we chose not to use the word ‘affect’ in our definition of news sentiment, as in previous papers (Puschmann and Powell 2018; Young and Soroka 2012), because affect is a non-conscious experience and thus is difficult to realize in language alone. Munezero et al. 2014 presents a useful discussion on the differences between affect, emotion, sentiment and opinion. In the rest of this paper, the word *sentiment* refers to the latent construct of ‘emotions expressed in a text’ that we measure by sentiment analysis.

Validation

Given the problem of domain-specificity, the validity of applying an off-the-shelf dictionary to one’s domain application could at best be face validity. Notably a recent delineation of validity (Van Atteveldt and Peng 2018, p.86–87) situates such claims of face validity as insufficient: 'The validity of a method or tool is dependent on the context in which it is used, so even if a researcher uses an existing off-the-shelf tool with published validity results it is vital to show how well it performs in a specific domain and on a specific task.' Failing to provide such revalidation can have dire consequences because systematic biases introduced by invalid measurements can spoil subsequent analyses.

The current study addresses the common problems that can stem from employing off-the-shelf dictionaries and demonstrates that unvalidated off-the-shelf applications of these methods are not robust enough to prevent dubious conclusions when applied to solve communication science problems. In doing so, we show that the validity of these methods for measuring news sentiment is not self-evident. We then demonstrate the seriousness of the problem by showing how different conclusions can be easily derived from such approaches.

In the first part of the study, we analyzed a set of dictionary-based sentiment scores as if they were a set of psychometric test items. Here, we reasoned that the psychometric properties of those tools could serve as measurements for the hidden construct of news sentiment. Based on classical test theory, a partial list of validity measures were studied, including i) convergent validity (are they correlated with each other?) and ii) structural validity (are they loaded into a unidimensional latent variable?). The second part of the paper puts those validity-challenged sentiment scores into action. In previous papers, sentiment scores extracted from news text are presented as time series (e.g. Haselmayer and Jenny 2016; Leetaru 2011; Young and Soroka 2012). In this part of the study, we demonstrate that time series analyses of news sentiment can yield misleading conclusions using a *p-hacking* approach; we based this work on an analysis done by Cohen (2004). Accordingly, we applied the same analysis to each of our 37 sentiment scores to test the same hypothesis and harvest those with a significant p-value.

The relationship between news sentiment and presidential approval

For the p-hacking experiment, our hypothesis was derived from Cohen (2004). He argued that *both* good and bad presidential news can impact the approval rating of US presidents; therefore, the direction of influence can sometimes be counterintuitive. One example mentioned by Cohen (2004) relates to the high popularity of Bill Clinton after his sex scandal. Building on Cohen's (2004) argument, in our own study the extremes in news sentiment (positive or negative) are assumed to be associated with *subsequent* extremes in presidential approval (but not the reverse direction of influence). Put it in the terminology of time series analysis, extremes in news sentiment are a *Granger-cause* (*G-cause*) of the extremes in presidential approval.

Although our hypothesis is derived from Cohen (2004), the hypothesis of the analysis in our p-hacking experiment is different. We would like to emphasize that the purpose of this study is not to replicate or extend Cohen's argument. Instead, we use our hypothesis as a case study to demonstrate the properties of sentiment scores based on off-the-shelf sentiment dictionaries and the risks of using them in domain applications without first establishing their validity for addressing the study's research questions (as proposed in Van Atteveldt and Peng 2018). Thus, we have no 'ground truth' and do not present a theoretical expectation on how the two variables (news sentiment and presidential approval) should behave; thus, we do not consider which p-value from our p-hacking experiment is 'wrong'. Instead, we aim to demonstrate that a large variety of conclusions can be derived using these dictionaries (which could be cherry picked) and the possible explanations behind this high variety of conclusions.

Methods

In the following two sections, we outline the operationalizations of presidential approval and news sentiment. Moreover, we also provide the validation procedures for the 37 sentiment scores.

Presidential approval rating time series

The presidential approval rating data were curated by the American Presidency Project (n.d.) hosted at the University of California, Santa Barbara. The presidential approval ratings from the Gallup Poll since 1943 were openly accessible online. The frequency of polling was irregular and ranged from every few weeks to every few days. In order to generate a regular time series, a daily time series of presidential approval ratings was created using spline interpolation between polls (as in Fu and Chan 2013).

News sentiment time series

The NYT data for this study was collected from ProQuest Historical Newspapers. We selected the NYT instead of another newspaper because it is an American 'newspaper of record'. We used the date of publication, content length (number of words) and sentiment scores extracted from the

NYT corpus. The articles represented the entire publication output of the NYT from June 1, 1980 to January 31, 2006. All articles were converted to lowercase and tokenized. The tokenized version of articles was used for extracting sentiment scores. In total, the sentiment scores of 2,246,177 articles were available.

The sentiment scores extracted were all based on widely-used off-the-shelf dictionaries.[4] Most of them have been used at least once in previous studies to quantify news sentiment,[5] although many of them are neither designed to measure news sentiment (e.g. measure moral foundations) nor measure sentiment in news text (e.g. measuring sentiment in product reviews). These dictionaries were General Inquirer (GI), Bing Liu (BL), Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW), Dictionary of Affect in Language (DAL), Moral Foundation Dictionary (MFD), NRC Word-Emotion Association Lexicon (NRC) and Lexicoder Sentiment Dictionary (LSD). An at-a-glance summary of these scores is available in Appendix A.

General Inquirer

General Inquirer (GI) is one of the oldest computer-assisted content analysis systems available (Stone and Hunt 1963). The system conducts content analyses on any kind of text and can use various dictionaries. Recent literature (e.g. Young and Soroka 2012) recognizes GI's capacity for sentiment analysis using a sentiment dictionary curated by a group of researchers from Harvard. The original system can output raw sentiment scores (raw frequency of matching words) and standardized scores (raw frequency divided by word count). In this study, the raw frequency was used. Two scores were calculated using this dictionary: GI + and GI -.

Linguistic Inquiry and Word Count

Linguistic Inquiry with Word Count (LIWC) is the most widely used off-the-shelf text analysis tool (Tausczik and Pennebaker 2009; Pennebaker et al. 2015). As mentioned previously, the authors of LIWC argue that the words a writer uses provide information on the writer's psychological state. Some researchers have adopted the tool as a measure of news sentiment (e.g. Ji et al. 2018; Walter 2019). For our purposes, it is important to note that LIWC is a proprietary software suite with several editions of the bundled dictionaries. We only had access to the 2007 edition of the dictionary, which has 64 categories of words. In this study, we selected 6 dimensions of LIWC related to news sentiment, namely, total affect, positive emotions, negative emotions, anxiety, anger and sadness. Thus, 6 scores were calculated using LIWC

(LIWC affect, LIWC +, LIWC -, LIWC anxiety, LIWC sadness). By default, the software gives standardized scores derived from raw frequency divided by word count.

Bing Liu

Bing Liu (BL) dictionary contains two lists of words with positive and negative sentiments (Hu and Liu 2004). The dictionary was proposed to quantify polarity of opinions from product reviews based on the frequency of matching words in a piece of text. In the original paper (Hu and Liu 2004), the 'orientation' of a text is quantified based on the difference between positive and negative word frequencies. This dictionary has been used to quantify news sentiment (e.g. Leetaru 2011; Walter 2019). One score was calculated using this dictionary: BL.

Affective Norms for English Words

Affective Norms for English Words (ANEW) is a dictionary based on human evaluation of 1,030 English words (Bradley and Lang 1999). Each word contains a numerical ANEW rating from 1 to 9 to capture the absence or presence of valence (pleasant to unpleasant), arousal (calm to excited) and dominance (controlled to dominated). Subsequent studies adopted the dictionary as a sentiment evaluation tool by totaling (Naveed et al. 2011) or averaging (Dodds and Danforth 2009) the ANEW rating of matching words in a sentence. In this study, the averaging approach was used. This dictionary has been in previous studies to quantify news sentiment, e.g. Gonzalez-Bailon et al. (2014). Three scores were calculated using this dictionary: ANEW valence, ANEW arousal and ANEW dominance.

Dictionary of Affect in Language

Dictionary of Affect in Language (DAL, Whissell 1989) is a dictionary similar to ANEW, in which every word in the dictionary has a set of DAL scores ranging from 1 to 3 to capture the absence or presence of pleasantness, activation and imagery. The original developer applied the dictionary to different categories of text using the averaging approach (e.g. Whissell 2008). In this study, we also average raw scores. Three scores were calculated using this dictionary: DAL pleasantness, DAL activation and DAL imagery.

Moral Foundation Dictionary

The Moral Foundation Dictionary (MFD, Graham, Haidt, and Nosek 2009) is a dictionary based on the moral foundation theory proposed by the same group of authors (e.g. Haidt 2012). Under that theory, there are five fundamental moral values: care/harm, fairness/cheating, ingroup loyalty/betrayal, authority/subversion, and purity/degradation. Similarly, the MFD classified words into these five axes with positive (virtue) and negative (vice) categories. Therefore, 10 categories of words are available. The original development of the dictionary was based on an expert evaluation of the words (Graham, Haidt, and Nosek 2009). As a validation, Graham, Haidt, and Nosek (2009) demonstrated the difference in word usage in religious texts between liberals and conservatives. The dictionary was subsequently used to quantify the moral rhetoric of news text (e.g. Fulgoni et al. 2016). Some studies billed the moral rhetoric of text as *moral sentiment* (e.g. Dainas, Munot, and Tsutsui 2015). It is worth mentioning that the original developers adjusted the frequency of sentiment words by the total number of words in a piece of text (Graham, Haidt, and Nosek 2009), but this is not always practiced (e.g. Dainas, Munot, and Tsutsui 2015). In this study, we use the unadjusted version of the MFD score. In total, 10 scores were calculated using this dictionary: MF Harm+ (Care), MF Harm -, MF Fairness +, MF Fairness – (cheating), MF Ingroup + (loyalty), MF Ingroup – (betrayal), MF Authority +, MF Authority – (subversion), MF Purity +, and MF Purity – (degradation).

NRC Word-Emotion Association Lexicon

NRC Word-Emotion Association Lexicon (NRC) is a dictionary created by crowdsourcing the emotional meanings of words (Mohammad and Turney 2012). The dictionary has categories of words about joy, anticipation, trust, surprise, fear, anger, disgust and sadness. These categories can be combined into two general categories of positive and negative emotions. The original paper does not provide a way to quantify the sentiment strength of a piece of text based on the dictionary. Subsequent studies (e.g. Vosoughi, Roy, and Aral 2018) use a measure of length-adjusted frequency. In total, 10 scores were calculated: NRC Joy, NRC Anticipation, NRC Trust, NRC Surprise, NRC Fear, NRC Anger, NRC Disgust, NRC Sadness, NRC + and NRC -.

Lexicoder Sentiment Dictionary

Lexicoder Sentiment Dictionary (LSD) is a dictionary specifically developed for measuring news affect (Young and Soroka 2012). Among all of the sentiment dictionaries included in this study, the development of LSD is the most comprehensive because it has been validated against human-coded media content and can take care of negation automatically. The dictionary contains words in two broad categories: positive and negative. The negated version of words (e.g. *not good*) is also considered. In the original paper, the developers suggested two ways of quantifying tone: *net tone*, calculated as the difference between proportions of positive words and negative words in a piece of text and another measurement, which was not named in the original article, calculated akin to BL's absolute difference in positive and negative word frequencies. We name this latter measurement *LSD absolute*. Both scores have been validated by the original developers and have been used as a measurement of news sentiment in time series analyses (Young and Soroka 2012). In total, 2 scores were calculated: LSD nettone and LSD absolute.

Validity measurements

With 37 sentiment scores from our 2,246,177 articles (GI: 2, LIWC: 6, BL: 1, ANEW: 3, DAL: 3, MFD: 10, NRC: 10, LSD: 2), the following validity measurements were calculated: 1) convergent validity (the correlation matrix of 37 sentiment scores was created to evaluate how the scores correlate with each other) and 2) structural validity (singular value decomposition (SVD) was conducted to evaluate the latent structure).

Time series analysis

For each of the 37 sentiment scores, we aggregated the sentiment of all NYT news stories by day and generated a daily regular time series of news sentiment (let n_{ti} represent the number of news stories and their sentiment score S for a given day t_i, with the aggregated sentiment score $\bar{S}$ of day ti is calculated using Equation). All the time series of $\overline{S_{ti}}$ were mean-centered and made the absolute values of $\overline{S'_{ti}}$ (Equations 2 to 4).

The time series of presidential approval was similarly processed (mean-centered with absolute value as per Equations 2 to 4).

$$\bar{S}_{t_i} = \frac{\sum_{j=i}^{n_{t_i}} S_{t_{ij}}}{n_{t_i}} \qquad (1)$$

$$\bar{\bar{S}} = \frac{\sum_{k=i}^{t} \bar{S}_{t_k}}{t} \qquad (2)$$

$$\sigma_{\bar{S}} = \sqrt{\frac{\sum_{l=1}^{t} \bar{S}_{t_l} - \bar{\bar{S}}}{t-1}} \qquad (3)$$

$$\bar{S}'_{t_i} = \left| \frac{\bar{S}_{t_i} - \bar{\bar{S}}}{\sigma_{\bar{S}}} \right| \qquad (4)$$

Granger causality

A bivariate Granger causality test was performed for each of the 37 sentiment scores with presidential approval according to the Direct Granger Method suggested by Soroka (2002) for studying agenda setting.[6] The same statistical procedure was conventionally used in many previous studies to study agenda setting (e.g. Lee 2014). The maximum order was chosen at 30 days because previous time series studies identified that the agenda-setting power of traditional mass media can last for four weeks (Walgrave, Soroka, and Nuytemans 2007).

In the true spirit of p-hacking, we hacked p-values even further by repeating the Granger causality analysis with the subset of NYT stories with the names of US presidents as a proxy of presidential news (using the same selection method as in Eshbaugh-Soha 2010); this p-hacking-in-disguise aligns with Cohen's argument (2004). Additionally, we also changed the dependent variable from presidential approval to University of Michigan Consumer Sentiment Index and even some random noise. This part of the analysis is reported in Appendix C.

Results

Validity measurements

Figure 1 shows the correlation matrix of the 37 sentiment scores. There are many abnormalities. When we group the sentiment scores by their polarity (Figure 1, bottom left and bottom right; as a histogram in Figure 2), not all sentiment scores with the same polarity have a correlation with each other. Some pairs, e.g. NRC + and ANEW Valence, have negative correlation. Only 40 pairs of positive sentiment scores (out of 91, 43.9%) and 51 pairs of negative sentiment scores (out of 105, 48.6%) have a positive correlation coefficient larger than 0.1. Median correlation coefficients for positive sentiment scores, negative sentiment scores, and all sentiment scores are 0.067 and 0.096 and 0.016 respectively.

Some pairs of positive and negative scores are strongly correlated (Figure 1, top). For example, the GI+ and GI- scores exhibit a positive correlation coefficient of 0.85. This correlation may indicate that: 1) positive and negative news sentiment occurs simultaneously or 2) both scores correlate with an unmeasured third variable.

Many of these abnormalities can be explained by the theory that both scores correlate with an unmeasured third variable. Firstly, whether or not a particular sentiment score adjusts for article length determines its

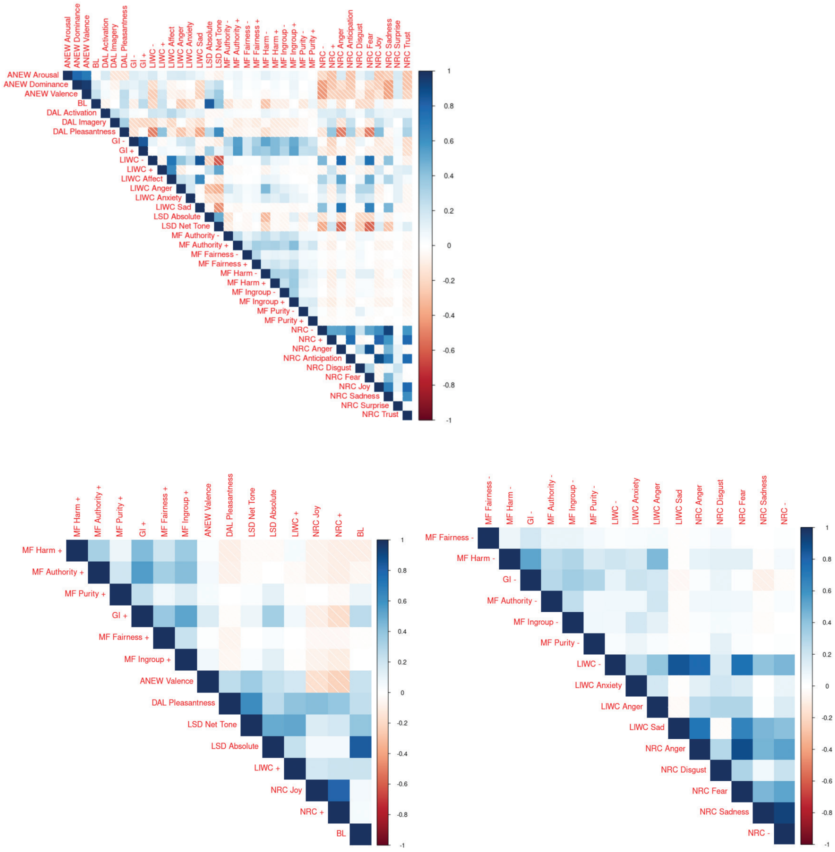

Figure 1: Correlation matrices of 37 sentiment scores (top), a subset of 14 positive sentiment scores (bottom left) and a subset of 15 negative sentiment scores (bottom right).

Notes: For the below two correlation matrices, the sentiment scores are ordered by a clustering algorithm based on their correlations with each other. Positive sentiment scores include all virtue scores of MFD, GI +, ANEW Valence, DAL Pleasantness, LSD Net Tone, LSD Absolute, LIWC +, NRC Joy, NRC + and BL. Negative sentiment scores include all vice scores of MFD, GI-, LIWC-, LIWC Anxiety, LIWC Sad, NRC Anger, NRC Disgust, NRC Fear, NRC Sadness, and NRC-. Some scores are not included in either positive and negative score matrices (e.g. NRC Anticipation, ANEW Dominance, ANEW Arousal, DAL Imagery, LIWC Affect) because their polarities are uncertain.

correlation with article length (Table 1). As indicated by a correlation coefficient larger than 0.1 between the sentiment score and article length (Table 1), 18 scores (including GI+ and GI-) have a positive correlation with article length.

Secondly, the exploratory factor analysis (Figure 3) also aligns with the theory that both scores correlate with an unmeasured third variable. In

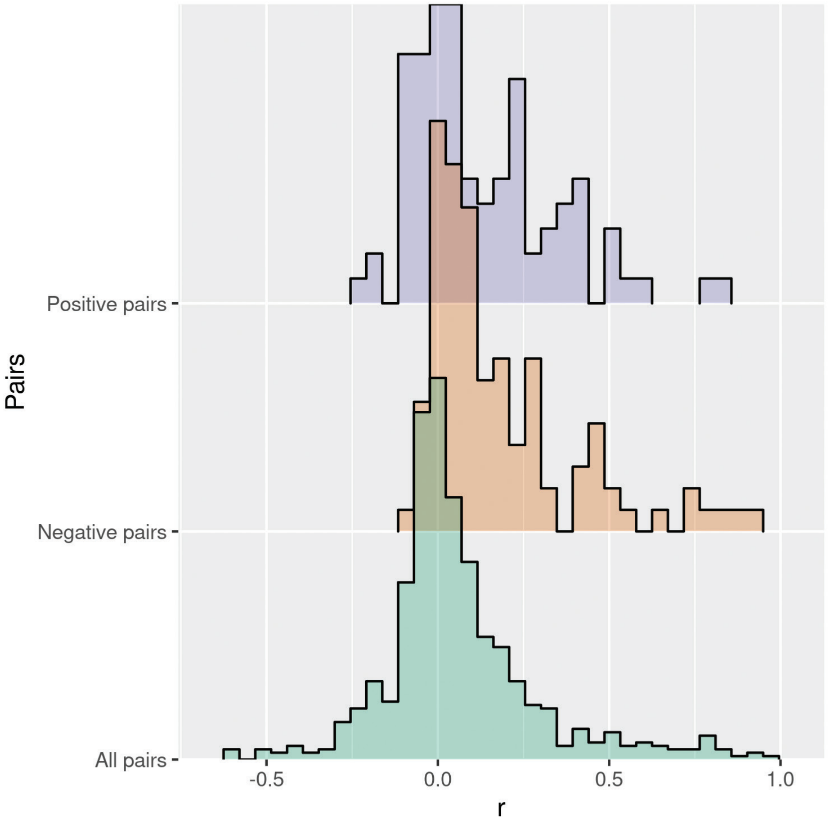

Figure 2: Histogram of correlation coefficients from positive pairs (top), negative pairs (middle), and all pairs (bottom).

this analysis, we extract the first component which explains most of the variance from these 37 sentiment scores. This component is helpful to test the structural validity, i.e. do these 37 sentiment scores collectively measure the latent construct of news sentiment? However, such a component score very strongly correlates with the article length (r=-0.933, Figure 3). Therefore, sentiment scores that do not adequately adjust for article length simply measure a 'latent construct' of unmeasured article length.

In sum, these sentiment scores might show convergent validity as indicated by the correlations among them. However, we have a very convincing alternative explanation for these correlations, namely, the influence of the unmeasured third variable of article length. The exploratory factor analysis indicates that these sentiment scores have low construct validity, that is,

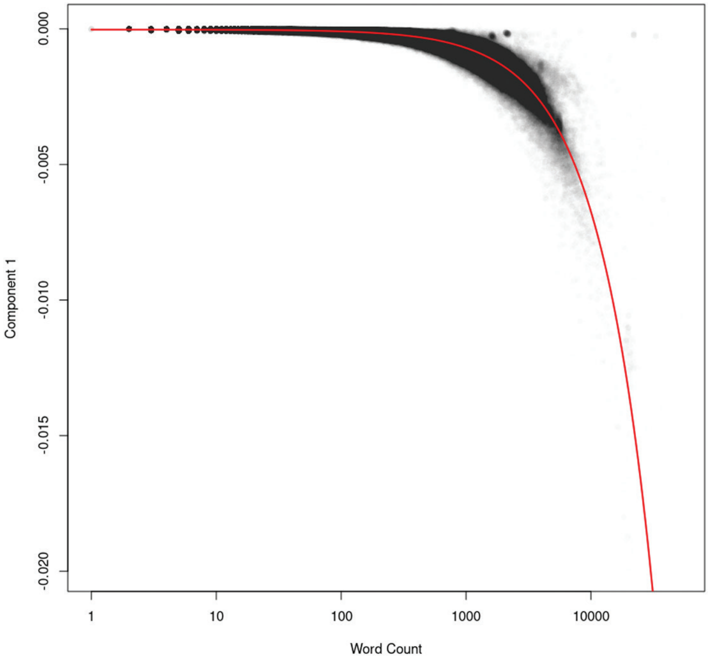

Figure 3: Scatterplot of the first component from the factor analysis and content length (*r* = *-0.933*).

the measurement has a poor ability to effectively measure what it purports to be measuring. Based on both analyses, we cannot reliably tell whether these sentiment scores are measuring sentiment, article length or a murky mixture of both. In other words, the validity of these sentiment scores as a measurement of sentiment is questionable.

Granger causality: p-hacking attempt

The results of the Granger causality test for predicting presidential approval are presented in Table 1. Using the conventional p<0.05 as the threshold of statistical significance, 9 scores (LIWC+, LIWC Anger, LIWC Anxiety, GI+, Bing Liu, MF Ingroup + MF Harm -, MF Authority – and LSD Absolute) emerge as statistically significant. As many sentiment scores tested here were not adjusted for article length, we performed an additional ad-hoc robustness analysis that takes into account the article length. Surprisingly, article length is a Granger cause of presidential approval (p<0.001). We attempted to adjust the four significant Granger causes found in the previous analysis by dividing the scores with the article length. We found that only

Table 1: ***Correlation of 37 sentiment scores and Granger causality tests for all sentiment scores***

Score	Correlation with content length - Pearson's r	Granger causality test: P (unadjusted)	Granger causality test: P (content- length adjusted)
NRC +	-0.228	0.999	
NRC Trust	-0.215	1.000	
DAL Imagery	-0.160	0.249	
NRC Sadness	-0.128	1.000	
NRC Anticipation	-0.128	1.000	
NRC Joy	-0.120	1.000	
NRC -	-0.117	1.000	
NRC Fear	-0.089	0.991	
LIWC Sad	-0.079	1.000	
NRC Anger	-0.068	0.993	
DAL Pleasantness	-0.066	0.302	
LIWC Affect	-0.043	0.322	
LIWC +	-0.033	0.001	
LIWC -	-0.011	0.984	
DAL Activation	-0.008	0.552	
NRC Disgust	-0.002	0.625	
LSD Net Tone	0.012	0.216	
NRC Surprise	0.021	0.760	
LIWC Anger	0.043	0.044	
ANEW Arousal	0.103	0.999	
LIWC Anxiety	0.121	0.000	
ANEW Valence	0.144	0.980	
ANEW Dominance	0.157	1.000	
MF Fairness -	0.163	0.857	
BL	0.178	0.000	0.686
MF Authority -	0.212	0.019	0.686
LSD Absolute	0.240	0.000	0.216
MF Purity -	0.245	0.875	
MF Purity +	0.267	0.993	
MF Ingroup -	0.294	0.002	0.068
MF Fairness +	0.315	0.646	
MF Harm +	0.374	0.863	
MF Harm -	0.384	0.002	0.003
MF Ingroup +	0.532	0.788	
MF Authority +	0.533	0.807	
GI -	0.868	0.053	
GI +	0.919	0.007	0.997
Article Length	1.000	0.000	

Note. The sentiment scores are sorted by their correlation with article length. The analysis from p-hacking should not be used to support or reject any substantive theory because it proceeds in an atheoretical manner.

one of these forcibly adjusted sentiment scores (MF Harm -) remained a significant Granger cause (p = 0.003).

Further p-hacking by using the subset of NYT content mentioning presidents' last names (Appedix C) also shows article length and MF harm vice as content-length adjusted significant Granger causes. In addition, using presidential news only, LSD Net Tone emerges as a new Granger cause for presidential approval. It is unclear whether this represents a genuine relationship or a fluke. In any case, the results concerning article length as an independent Granger cause for presidential approval disqualify all sentiment scores that do not adjust for article length. To be sure, even the remaining sentiment scores should not be picked based on the statistical significance we conducted in our p-hacking experiment.

Conclusion

Our analyses of 37 sentiment scores suggest that using off-the-shelf sentiment dictionaries can lead to unexpected validity problems. In this discussion, we organize our concerns about using off-the-shelf sentiment dictionaries by presenting four best practices for using off-the-shelf sentiment dictionaries for studying news sentiment. These four best practices are hardly original: most of them have been proposed in previous best practice articles (e.g. Grimmer and Stewart 2013; Van Atteveldt and Peng 2018; Barberá et al. 2020). With our empirical findings, this discussion illustrates the importance of these best practices.

Best practice #1: do not use dictionaries unsuitable for your task

A wrong choice of dictionary can lead to uninterpretable conclusions. Because this is a theoretical problem, we turn to it here first.

Some dictionaries, although used in previous studies as tools of sentiment analysis, were not created for sentiment analysis. For example, MFD was created to measure word choice in texts and determine the moral foundations dominant in different communities. Here, it is helpful to note that the variable being measured by MFD, as named by the original authors, is *moral foundation endorsement* (Graham, Haidt, and Nosek 2009). The inappropriateness of using MFD as a measurement of general news sentiment is best illustrated with the ways in which some findings from the p-hacking Granger analysis may be misinterpreted. For example, the MF harm – score emerged as a significant Granger cause of change in presidential approval. However, we have very strong reservations about interpreting this score as a measurement of news sentiment or news tone. A review of the lexicons that fall into the MF harm vice group reveals that nearly all of them are nouns and verbs about war and conflicts (e.g. *war, suffering, attack,* etc.). They are mostly not stylistic text features conveying emotions, such as adjectives (e.g. painful, sad, depressing, hopeless) and adverbs (e.g. painfully, sadly, depressingly, hopelessly). Instead, these words are the entities and events themselves. The MF harm vice score is very likely not a measurement of news sentiment, but rather of media salience of conflict events. Many previous studies have shown the relationship between conflict events and presidential approval, that is, the rally around the flag effect (Schubert, Stewart, and Curran 2002). Due to the construction of the dictionaries, many sentiment scores actually indicate topics and therefore may not be good indicators of 'emotions expressed in a text' when researchers want to study news texts covering different topics: news articles on some topics (e.g. conflict events) will then automatically have higher sentiment scores than other topics, purely due to the ways some dictionaries are constructed.

We propose the first best practice: when studying news sentiment, one should choose dictionaries intended for sentiment analysis of news content (e.g., Lexicoder). However, there is no 'one-size-fits-all' solution. It can be highlighted in the analysis using the University of Michigan Consumer Sentiment Indicator (Appendix C). To be sure, changing the dependent variable of the analysis from presidential approval to Consumer Sentiment Indicator can generate a different set of results (e.g. LSD-based scores are no longer significant). Instead of endorsing one sentiment dictionary, we recommend that researchers use theoretically informed dictionaries suitable for the task at hand.[7] Moreover, researchers should always check the lexicons in the dictionaries for topical words.

Best practice #2: do not assume that validity is a built-in feature of dictionaries; always revalidate

After choosing a suitable dictionary, one needs to test for validity and reliability of the dictionary. This suggestion is hardly new: previous studies have demonstrated how some sentiment scores lack criterion validity and have domain specificity problems. The current study identifies other undesirable psychometric properties to further demonstrate this point. The convergent validity (positive sentiment scores are positively correlated with other positive sentiment scores) and discriminant validity (positive sentiment scores are negatively correlated with negative positive sentiment scores) of these sentiment scores, as demonstrated in Figure 1, are also lacking. Negative sentiment scores and positive sentiment scores sometimes have a positive correlation. The structural validity for these sentiment scores is also difficult to interpret (Figure 3). Without closely scrutinizing the details, we may naïvely conclude that a hidden construct of news sentiment was actually measured by these sentiment scores; however, this naïve conclusion is unlikely to hold. For example, we show that the first component from the exploratory factor analysis is not a good measurement of the hidden construct of sentiment in text because it is actually tainted with the collective residual influence of article length (next paragraph). In sum, we cannot assume the validity of dictionaries are built-in. Not only these sentiment scores often lack criterion validity (whether or not they represent human understanding of sentiment, as reported in the previous validation studies), they also lack construct validity (whether or not they are measuring what they purport to be measuring). Therefore, we present a second best practice: one must always revalidate these dictionaries for the domain under study and publish the results of the revalidation with the subsequent analysis.

Best practice #3: check for the influence of article length on sentiment scores and outcomes

We found that many sentiment scores are mildly to strongly correlated with article length (Table 1). This residual influence is visualized in Figure 3, which shows that the first component—an indicator that can explain the variance of our 37 scores —has a strong correlation with article length. Such interpretation can also be used to interpret the positive correlation between positive and negative sentiment scores (Figure 1): both are strongly correlated with article length, which is only partially adjusted or even unadjusted.

As indicated by our p-hacking Granger analysis, many sentiment scores were found to be Granger causes of presidential approval (Table 1). Owing to the fact that many of the scores have not been completely adjusted for the effect of article length, we conducted an ad-hoc robustness test to take article length into account. As a result, many scores were no longer significant.

This influence of article length may not be a problem for content with less variability in length (e.g. tweets). However, in news analysis, this residual effect of article length is a problem: we found that article length is itself a Granger cause of presidential approval, which is of course a potentially meaningless artifact. This finding is surprising and, to our knowledge, has not yet been mentioned in the literature. We hypothesize that such a relationship can be explained by issue salience. Longer news articles, in general, may be indirect indicators of higher issue salience, although it is beyond the scope of this study to test this hypothesis. What is important to take away here for news analysis is that this problem of article length suggests that article length in itself may carry meaning.

Because these sentiment scores can be heavily correlated with article length and article length itself can potentially carry substantive meaning, we propose a third best practice: use the length-adjusted version of sentiment scores (e.g. LSD's Net Tone or averaged DAL scores), if available. However, it is important to note that even when using these length-adjusted sentiment scores, one still needs to check whether or not article length can still affect the results. This check involves two steps: 1) checking residual influence of content length; 2) checking if content length can affect the outcomes. We showed in this study that some length-adjusted sentiment scores can still have a residual influence from article length (e.g. NRC positive, LIWC Anxiety).

In addition, readers should be aware that these length-adjusted sentiment scores cannot be interpreted as a ratio scale. For example, a score of 0 does not indicate complete neutrality because length-adjusted sentiment scores are usually slightly biased towards either the positive or the negative due to the uneven baseline distribution of sentiment words in each category for a given dictionary. Therefore, the point of neutrality for these scores should always be calibrated before the scores are interpreted (Rauh 2018).

Best practice #4: do not use multiple dictionaries to test the same hypothesis

The wide availability of multiple off-the-shelf dictionaries can create a situation in which researchers can apply multiple dictionaries to the same piece of text. As in the current study, we used the same NYT text data to generate 37 different sentiment scores. Using the language of experimental design, one can generate multiple non-manipulated independent variables using essentially the same data. This freedom to increase non-manipulated independent variables has previously been criticized for incentivizing p-hacking (Simmons, Nelson, and Simonsohn 2011). Detection of p-hacking in literature is not trivial (Bishop and Thompson 2016) and therefore we do not—and will never—have any evidence to suggest that the availability of multiple off-the-shelf dictionaries leads researchers to p-hack. Thus, we are not accusing our fellow researchers for p-hacking. Instead, we address this problem as a hypothetical risk and focus on how to prevent such hypothetical risk from becoming a genuine risk to science.

From our p-hacking experiment, we found that using multiple dictionaries to test the same hypothesis can generate faulty—but significant—relationships. These off-the-shelf dictionaries are not resistant to domain-specific biases and to the influence of content length. But even without the aforementioned validity problems of these off-the-shelf dictionaries, one can expect to generate at least one statistically significant false positive result when one applies multiple dictionaries *en masse*. Hypothetically, it is entirely possible to use different off-the-shelf dictionaries to test the same statistical hypothesis until one obtains a statistically significant result. This is similar to the situation of 'physician shopping', where a patient visits multiple doctors to obtain medical opinions until he or she obtains an opinion that he or she wants to hear. Given the background of the ongoing replication crisis in science, this hypothetical 'dictionary shopping' could undermine the likelihood of valid conclusions and should thus be discouraged. One hedge against this 'dictionary shopping' risk in confirmatory studies is to enforce modern open science principles such as pre-registering research protocols. Studies that must use multiple dictionaries to test the same hypothesis should clearly document their usages and appropriately situate themselves as exploratory or hypothesis-generating studies.

Practically, one may not want to go 'dictionary shopping' but still apply multiple dictionaries to test the same hypothesis. For example, Walter (2019) first applied LIWC sentiment scores extracted from her news corpus to study the relationship between mentions of EU citizens and news sentiment in Brexit coverage. As a robustness check, she subsequently applied the BL sentiment score extracted from the same corpus and repeated the same analysis. Although this practice looks statistically sound, we discourage the comparison of one sentiment score with another as a robustness check because these sentiment scores are often measuring related but different concepts (see Appendix A, e.g. LIWC measures emotional states of the writer; BL extracts opinion from online reviews). The correlation between two sentiment scores can also be spurious, e.g. due to an unmeasured variable such as content length (Table 1). Thus, using two sentiment dictionaries to test the same hypothesis is not simply an alternative model specification as in a regular robustness test, but instead using two independent variables with different meanings.

We thus propose a fourth best practice: do not use multiple dictionaries to test the same statistical hypothesis. When possible, pre-register one's research protocol to resist the temptation of 'dictionary shopping'.

'Revalidate, revalidate, revalidate'

In the early days of computational research, researchers were overwhelmed by the contradiction between the increasing volume of text data on the one hand and the fact that traditional methods, such as quantitative content analysis, do not scale up very well on the other. In that era, the scalability of a method might have *trumped* concerns with validity, and this might be why methods with limited validity were (and still are) popular. However, the field of computational research is maturing to a point where validity is equally, if not more, important than scalability.

Our findings support the observation that off-the-shelf dictionary-based methods come with significant pitfalls (Ribeiro et al. 2016). These methods might have been validated in the initial development. However, all such methods must be revalidated again by humans before applying them to new research questions and/or new text material, as indicated by the catchy motto '*validate, validate, validate*' (Grimmer and Stewart 2013). His point has been rightly recited in subsequent best practice papers for communication researchers, such as those by Boumans and Trilling (2015) and Van Atteveldt and Peng (2018). The details about how to validate these methods are available in Song et al. (2020). In Appendix D, we demonstrate how to implement best practice #2 and #3.

Song et al. (2020) based on their simulation study suggest that one should hand annotate at least 1% of the source material in a validation study. When the sample size of articles is not overwhelming, revalidation is a reasonable path to take. For example, the aforementioned study by Walter (2019) is a reasonable case for taking this revalidation path. Hand annotating 1% of articles in her study (n=19,367) amounts to only 194 articles.

As pointed out by Barberá et al. (2016), the revalidation of off-the-shelf dictionaries can be labour-intensive and can quickly outweigh the advantage of using those dictionaries. The revalidation path of off-the-shelf tools is no longer reasonable when the sample size is large. Using this study as an example and applying Song et al. (2020)'s suggestion, 22,461 articles would need to be hand annotated and that would cost a handsome amount of money.

If researchers had the resources to do so, then they may alternatively consider putting their energy towards creating new validated and customized sentiment assessment tools for their own research purposes, even though such tools may only be for one-time use (e.g. Fu and Chan 2013). We may thus approach such tools as we do syringes: it is safer to manufacture and use single-use, 'throw-away' syringes than reuse them. Crucially, using a ‘throw-away’ sentiment tool can also eliminate the risk of ‘dictionary shopping’ and guarantees the use of a validated sentiment tool. With human validation, new, more nuanced applications of dictionary-based sentiment tools have emerged. For example, Fogel-Dror et al. (2018) utilize off-the-shelf LSD in an analysis of sentiment against news entities using a validated, rule-based approach. If one has to hand annotate 1% of the material and that amounts to a few thousand articles, a new study shows that there is more than enough data to train and validate an accurate supervised machine learning model of news sentiment (Barberá et al. 2020). Regardless, all these new applications require heavy human validation.

Additionally, we encourage authors to replicate previous studies that make use of unvalidated off the shelf sentiment analyses. Using a validated sentiment analysis in the replication of these previous studies can certainly improve the strength of evidence supporting these previous findings.

Limitations

The current study has two important limitations.

We did not use length-adjusted versions of some scores, such as GI and MFD; instead, we used the unadjusted versions because they were used by previous studies. We replicated the exploratory factor analysis again with the length-adjusted version of GI and MFD scores and, as expected, the resultant first component exhibited a much weaker correlation with content length. This highlights the third best practice we present above. In our p-hacking attempt, using both the length-adjusted and unadjusted version would have only increased the false discovery rate of significant relationships.

Similarly, preprocessing is consequential to generated sentiment scores. Similar to another benchmark study using LSD (González-Bailón and Paltoglou 2015), this study has not studied the effect of preprocessing and for some dictionaries, e.g. LSD, we have not used the script provided by Young and Soroka (2012) which has been shown to improve dictionaries’ performance. We anticipate using that script would improve the performance of LSD but using that would also introduce an additional layer of heterogeneity in methodology. Also, we do not believe that would change our conclusion, particularly for those non-LSD sentiment scores. Although that preprocessing script is not used in this study, we still recommend users of LSD to use that preprocessing script in practical applications.

In sum, this study found some undesirable psychometric properties in 37 off-the-shelf sentiment scores extracted from a large corpus of NYT articles. Using these sentiment scores to study the relationship between news sentiment and presidential approval in a p-hacking manner, we demonstrated that it is possible to use multiple sentiment scores to test the same statistical hypothesis to generate statistically significant causal results due to the residual influence of the confounding content length. Even after we forcibly adjusted for the effect of content length, the conclusions remained very difficult to interpret due to the ambiguity of topic and style words in these off-the-shelf sentiment dictionaries. The current study shows the adverse outcomes of applying these sentiment scores without proper revalidation. We also propose four best practices and suggest alternatives to using off-the-shelf sentiment dictionaries.

Notes

[1] We would like to thank Dr. Stefanie Walter (Zentrum für Medien-, Kommunikations- und Informationsforschung, Universität Bremen) and Professor Sandra González-Bailón (Annenberg School for Communication, University of Pennsylvania) for sharing details about their studies. This project was funded by a research grant from the 1) German Research Foundation (Deutsche Forschungsgemeinschaft), 2) The Netherlands Organisation for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) and 3) the National Endowment for the Humanities, through the Trans-Atlantic Platform's Digging into Data Challenge funding program. This paper has been provisionally accepted by *Computational Communication Research*. Source code and data are available at https://github.com/chainsawriot/ots/ Online appendix is available at https://doi.org/10.17605/OSF.IO/UTXS5

Correspondence concerning this article should be addressed to Chung-hong Chan, A5, 6 (section A), 68159 Mannheim, Germany. E-mail: chung-hong.chan@mzes.uni-mannheim.de

[2] This paper deals with dictionary-based sentiment analysis only. Indeed, there are other applications of dictionary-based methods in the realm of communication studies, e.g. measurement of populism (Rooduijn and Pauwels 2011). Although these applications are not studied in this paper, in principle the findings from this study still apply.

[3] Emotions are defined here as 'preconscious social expressions of feelings and affect influenced by culture' (Munezero et al. 2014, 4).

[4] In this paper, a sentiment dictionary is simply a word list. A sentiment score is a score calculated based on a sentiment dictionary. A sentiment dictionary can have multiple categories of words. For instance, General Inquirer has positive and negative categories. Therefore, one can calculate 2 sentiment scores based on General Inquirer: "General Inquirer Positive" and "General Inquirer Negative" scores. Some dictionaries, e.g. Bing Liu, require one to use multiple categories of words to calculate one score.

[5] This paper only focuses on news articles. Therefore, dictionaries for short texts, e.g. VADER (Gilbert and Hutto 2014), were not considered.

[6] Please refer to Appendix B for the description of the statistical test.

[7] It is possible that a dictionary gives accurate results for a different task than it was developed for, especially if the tasks are conceptually similar. This can be confirmed through (re)validation, as discussed in the second best practice. However, we recommend caution in exploring which existing dictionaries can be reused for a different task. In particular, one should not simply validate many existing dictionaries to see which performs best on a given gold standard, due to concerns of overfitting and multiple comparisons.

References

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2020). "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis*, June, 1–24. https://doi.org/10.1017/pan.2020.8

Barberá, PBarberá, P., Boydstun, A., Linn, S., McMahon, R., & Nagler, J. (2016). "Methodological Challenges in Estimating Tone: Application to News Coverage of the Us Economy." In Meeting of the Midwest Political Science Association, Chicago, Il.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774

Bishop, D. V., & Thompson, P. A. (2016). "Problems in Usingp-Curve Analysis and Text-Mining to Detect Rate of P-Hacking and Evidential Value." *PeerJ* 4 (February): e1715. https://doi.org/10.7717/peerj.1715

Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2019). "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools." *Communication Methods and Measures* 14 (2): 83–104. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J. W., & Trilling, D. (2015). "Taking Stock of the Toolkit." *Digital Journalism* 4 (1): 8–23. https://doi.org/10.1080/21670811.2015.1096598

Bradley, M. M., & Lang, P. J. (1999). "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings." Technical report C-1, the center for research in psychophysiology.

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. The Journal of Politics, 75 (3), 659–671. https://doi.org/10.1017/s0022381613000492

Cohen, J. E. (2004). If the news is so bad, why are presidential polls so high? Presidents, the news media, and the mass public in an era of new media. Presidential Studies Quarterly, 34 (3), 493–515. https://doi.org/10.1111/j.1741-5705.2004.00209.x

Dainas, A., Munot, V., & Tsutsui, S. (2015). The moral foundations in new york times. https://pdfs.semanticscholar.org/0c08/ab050e941e57de95433722895d8c1abd9064.pdf.

Diesner, J., & Evans, C. S. (2015). Little bad concerns: Using sentiment analysis to assess structural balance in communication networks. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 – ASONAM '15. https://doi.org/10.1145/2808797.2809403

Dodds, P. S., & Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness Studies, 11 (4), 441–456. https://doi.org/10.1007/s10902-009-9150-9

Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. Political Communication, 27 (2), 121–140. https://doi.org/10.1080/10584600903502623

Fogel-Dror, Y., Shenhav, S. R., Sheafer, T., & Van Atteveldt, W. (2018). Role-based association of verbs, actions, and sentiments with entities in political discourse. Communication Methods and Measures, 13 (2), 69–82. https://doi.org/10.1080/19312458.2018.1536973

Fu, K.-w., & Chan, C.-h. (2013). Analyzing online sentiment to predict telephone poll results. Cyberpsychology, Behavior, and Social Networking, 16 (9), 702–707. https://doi.org/10.1089/cyber.2012.0375

Gilbert, C., & Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment Analysis of Social Media Text. In Eighth International Conference on Weblogs and Social Media (icwsm-14)., 81:82. http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf.

Gonzalez-Bailon, S., De Francisci Morales, G., Mendoza, M., Khan, N., & Castillo, C. (2014). Cable news coverage and online news stories: A large-scale comparison of media bias. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2389525

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online Communication. *The ANNALS of the American Academy of Political and Social Science* 659 (1): 95–107. https://doi.org/10.1177/0002716215569192

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology* 96 (5): 1029–46. https://doi.org/10.1037/a0015141

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Vintage.

Haselmayer, M., & Jenny, M. (2016). Sentiment Analysis of Political Communication: Combining a Dictionary Approach with Crowdcoding. *Quality & Quantity* 51 (6): 2623–46. https://doi.org/10.1007/s11135-016-0412-4

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '04. https://doi.org/10.1145/1014052.1014073

Ji, Q., Raney, A. A., Janicke-Bowles, S. H., Dale, K. R., Oliver, M. B., Reed, A., . . . Raney, A.A. 2018. Spreading the Good News: Analyzing Socially Shared Inspirational News Content. *Journalism & Mass Communication Quarterly* 96 (3): 872–93. https://doi.org/10.1177/1077699018813096

Lee, H. S. (2014). Analyzing the Multidirectional Relationships Between the President, News Media, and the Public: Who Affects Whom? *Political Communication* 31 (2): 259–81. https://doi.org/10.1080/10584609.2013.815295

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, August. https://doi.org/10.5210/fm.v16i9.3663

Liu, B. 2010. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing* 2 (2010): 627–66.

Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a word-emotion association Lexicon. *Computational Intelligence* 29 (3): 436–65. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing* 5 (2): 101–11. https://doi.org/10.1109/taffc.2014.2317187

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad News Travel Fast. Proceedings of the 3rd International Web Science Conference on – WebSci '11. https://doi.org/10.1145/2527031.2527052

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval* 2 (1–2): 1–135. https://doi.org/10.1561/1500000011

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of Liwc2015

Puschmann, C., & Powell, A. (2018). Turning Words into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media. *Social Media + Society* 4 (3): 205630511879772. https://doi.org/10.1177/2056305118797724

Rauh, C. (2018). Validating a Sentiment Dictionary for German Political Language—a Workbench Note. *Journal of Information Technology & Politics* 15 (4): 319–43. https://doi.org/10.1080/19331681.2018.1485608

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. 2016. SentiBench – a Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods. *EPJ Data Science* 5 (1). https://doi.org/10.1140/epjds/s13688-016-0085-1

Rooduijn, M., & Pauwels, T. (2011). Measuring Populism: Comparing Two Methods of Content Analysis. *West European Politics* 34 (6): 1272–83. https://doi.org/10.1080/01402382.2011.616665

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More Than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures* 12 (2-3): 140–57. https://doi.org/10.1080/19312458.2018.1455817

Schubert, J. N., Stewart, P. A., & Curran, M. A. (2002). A Defining Presidential Moment: 9/11 and the Rally Effect. *Political Psychology* 23 (3): 559–83. https://doi.org/10.1111/0162-895x.00298

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., . . .Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication* 37 (4): 550–72. https://doi.org/10.1080/10584609.2020.1723752

Soroka, S. N. (2002). Agenda-Setting Dynamics in Canada. UBC press.

Stone, P. J., & Hunt, E. B. (1963). A Computer Approach to Content Analysis.” Proceedings of the May 21-23, 1963, Spring Joint Computer Conference on – AFIPS ’63 (Spring). https://doi.org/10.1145/1461551.1461583

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29 (1): 24–54. https://doi.org/10.1177/0261927x09351676

The American Presidency Project. (n.d.) Presidential Job Approval. https://www.presidency.ucsb.edu/statistics/data/presidential-job-approval

Van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures* 12 (2-3): 81–92. https://doi.org/10.1080/19312458.2018.1458084

Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. *Science* 359 (6380): 1146–51. https://doi.org/10.1126/science.aap9559

Walgrave, S., Soroka, S., & Nuytemans, M. (2007). The Mass Media’s Political Agenda-Setting Power. *Comparative Political Studies* 41 (6): 814–36. https://doi.org/10.1177/0010414006299098

Walter, S. (2019). Better Off Without You? How the British Media Portrayed Eu Citizens in Brexit News. *The International Journal of Press/Politics* 24 (2): 210–32. https://doi.org/10.1177/1940161218821509

Whissell, C. (2008). Emotional Fluctuations in Bob Dylan’s Lyrics Measured by the Dictionary of Affect Accompany Events and Phases in His Life. *Psychological Reports* 102 (2): 469–83. https://doi.org/10.2466/pr0.102.2.469-483

Whissell, C. M. (1989). The Dictionary of Affect in Language. *The Measurement of Emotions*, 113–31. https://doi.org/10.1016/b978-0-12-558704-4.50011-6

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication* 29 (2): 205–31. https://doi.org/10.1080/10584609.2012.671234

About the Authors

Chung-hong Chan and Hartmut Wessler, Mannheimer Zentrum für Europäische Sozialforschung, Universität Mannheim, Germany

Joseph Bajjalieh, Loretta Auvil and Scott Althaus, Cline Center for Advanced Social Research, University of Illinois Urbana-Champaign, USA

Kasper Welbers and Wouter van Atteveldt, Department of Communication Science, Vrije Universiteit Amsterdam, Netherlands

Marc Jungblut, Department of Media and Communication, Ludwig Maximilian University of Munich, Germany