To Scale or Not to Scale: Comparing Popular Sentiment Analysis Dictionaries on Educational Twitter Data

Conrad Borchers
University of Tübingen
conrad.borchers@student.uni-tuebingen.de

Joshua M. Rosenberg
University of Tennessee, Knoxville
jmrosenberg@utk.edu

Ben Gibbons
Emory University
ben.gibbons@emory.edu

Macy Alana Burchfield
University of Tennessee, Knoxville
mburchf3@vols.utk.edu

Christian Fischer
University of Tübingen
christian.fischer@uni-tuebingen.de

ABSTRACT

The extraction of sentiment from text requires many methodological decisions to make inferences about mood, opinion, and engagement in informal learning contexts. This study compares sentiment software (SentiStrength, LIWC, tidytext, VADER) on $N = 1{,}382{,}493$ tweets in the context of the Next Generation Science Standards reform ($N = 546{,}267$) and U.S. State Educational Twitter Hashtags ($N = 836{,}226$). Automated sentiment classifications were validated on $N = 300$ hand-coded tweets. Additionally, we developed a discrepancy measure to identify tweet features associated with scale inconsistency. Results indicated that binary sentiment classifications (positive/neutral vs. negative) were more accurate than trinary classifications (positive, neutral, negative). Combined tidytext dictionaries and VADER outperformed LIWC for negative sentiment, which was overall difficult to classify reliably while positive sentiment was classified with high accuracy across all four dictionaries. Thus, researchers are encouraged to (a) consider employing overall sentiment scales or positive/neutral to negative ratios based on binary classification to characterize their sample, (b) aggregate multiple dictionaries or use domain-specific sentiment dictionaries, and (c) be aware of the current limitations of detecting negativity through dictionary-based sentiment analysis in educational contexts.

Keywords

Social media data, sentiment analysis, online communities

1. INTRODUCTION

Sentiment analysis extracts positive and negative emotions from text. Its many applications include stock market prediction [22], marketing research [29], and, recently, investigating public sentiment on educational reforms on Twitter [32, 38]. Sentiment analysis typically requires numerous methodological decisions, such as deciding whether to use a dictionary-based or a supervised machine learning approach and determining how sentiment measures are suited to the investigation of a particular domain (e.g., VADER for social media data) [13, 30].

User-defined sentiment dictionaries (UDDs) rely on matches of word occurrences with a value in their dictionary, with little overlap often yielding less valid results. Whereas many sentiment measure validation studies investigate binary (i.e., positive and negative) sentiment classifications [1, 25, 28], less research has systematically compared trinary classifications (i.e., positive, neutral, negative) and sentiment scales. Furthermore, as sentiment classifiers do not generalize well across domains [2], sentiment validation studies are needed to inform educational researchers utilizing the increased availability of big data in education [7]. This study examines the performance of popular sentiment analysis methods in the context of a particular, social media-based data source: large education-related Twitter communities.

The motivation of this study is two-fold. First, sentiment measures can give insight into how and why teachers engage in online communities on Twitter, a potentially novel form of informal teacher learning [6, 33]. Second, public opinion and sentiment can be viewed as a proxy for successful reform implementation [4, 27]. Wang and Fikis applied SentiStrength on more than 660,000 tweets related to the Common Core State Standards, finding sentiment, including that expressed by teachers, to be largely negative [38]. In contrast, Rosenberg et al. found largely positive sentiment in 570,000 NGSS-related tweets through the same SentiStrength algorithm [32]. However, the validity of the utilized sentiment methods was not examined.

2. RESEARCH BACKGROUND

2.1 Sentiment Analysis Methods and Tools

Sentiment analysis is frequently carried out through *user-defined dictionaries* (UDDs) [9]. UDDs contain sets of labeled words that are rated on affect dimensions (e.g., valence, potency, activity) and matched to word occurrences in texts [23]. Researchers can either use pre-defined dictionaries or create their own dictionaries [9]. UDD methods

examine words individually, potentially neglecting figurative language and ambiguous phrases [36]. This study examines four popular examples of UDD software: (a) SentiStrength, (b) Linguistic Inquiry and Word Count (LIWC), (c) the R-package tidytext, and (d) the social-media attuned software VADER.

SentiStrength outputs two truncated five-point scales [37] which is different from many other UDD implementations. It offers feature selection options and measures sentiment weight, which is the intensity or strength of positivity or negativity in a text, as opposed to simply comparing the frequency of sentiments in a text [14].

LIWC is possibly the most popular text analysis software [17]. It uses a well-validated default dictionary [16] and contains around eighty subdictionaries of topics for which it outputs individual scales [26]. LIWC has been used to infer psychological processes and constructs (e.g., emotional expressions) from text [36].

Tidytext [34] does not provide its own default dictionary. At its core, it strives to pre-process input text which is then analyzed through any input dictionary [35]. Tidytext provides functions for converting text into a "one-token-per-document-per-row" format which may ease text analysis.

VADER (Valence Aware Dictionary for Sentiment Reasoning) features multiple subdictionaries and considers word order and degree modifiers (e.g., "very", "slightly", "somewhat") [5]. It performs well in sentiment analyses of social media content (including from Twitter) while remaining applicable to other contexts [5, 13]. That said, we found the R implementation of VADER to take around 80 times longer to compute compared to the other methods.

2.2 Research Questions

This study examined the validity of SentiStrength, LIWC, tidytext, and VADER in the context of educational Twitter data with the following research questions (RQs):

RQ1: How valid are the employed sentiment measures with respect to human coding of sentiment?

RQ2: How discrepant are sentiment scales and are correlations among scales consistent with these discrepancies?

RQ3: Which features of texts (i.e., the number of words, likes, retweets, and context) account for scale discrepancy?

3. METHOD

3.1 Sample

The study utilized tweets related to the Next Generation Science Standards reform and large educational state-wide hashtags (1,382,493 tweets, 156,446 users) posted between July 2008 and October 2020. Search terms included the #NGSSchat hashtag (N = 175,094 tweets, N = 67,060 of which being inside of designated chat-sessions), the terms "ngss" (without #NGSSchat, N = 312,167 tweets) or "next gen[eration] science standard[s]" (N = 59,006 tweets). In addition, we included tweets from 47 State Educational Twitter Hashtags (N = 836,226). Tweets not recognized as of English language by the Twitter API were omitted (5.0%).

3.2 Sentiment Measures

To investigate the validity of different sentiment measures, we used SentiStrength [37], LIWC [26], tidytext [34], and VADER [13] to obtain (a) binary and trinary classifications, (b) unidimensional (positive and negative) sentiment scales, and (c) a bidimensional sentiment scale rating for all tweets. While SentiStrength has binary and trinary classification methods, we subtracted negativity ratings from positivity ratings to obtain overall scores and defined a tweet as neutral if that overall rating was 0 (over 0 as positive, under 0 as negative) for LIWC and tidytext. For tidytext, we used the NRC [24], Loughran-McDonald [20], AFINN [10], and Bing [12] dictionaries, standardizing ratings by the number of words in each tweet and averaging across available ratings. The remaining non-matches were assigned a 0. For VADER, we used its internal compound score as overall scale and classified tweets as neutral if that score was between -0.05 and 0.05 (instead of 0) [13]. Binary classification combined positive and neutral tweets, such that neutral tweets were coded as positive, as done in previous validation studies [8] and since we observed that SentiStrength always classified tweets rated neutral in its trinary method positive in its binary method. Additionally, we defined ambiguity measures for all sentiment dictionaries as the sum of the absolute values of their positivity and negativity ratings.

3.3 Additional Variables

Continuous predictor variables included the number of likes, retweets, and words (excluding links and user mentions) of each tweet. To account for some features of the specific data sets we analyzed, we created a *categorical predictor variable* indicating whether a tweet was from the NGSS or SETHs data set (and, for the NGSS data set, whether the tweet was posted inside of #NGSSchat, designated chat-sessions of #NGSSchat, or included the term "ngss").

3.4 Data Analysis

3.4.1 Hand-coding and validation

To provide a validation set of tweets to investigate how UDDs compare to human-evaluated sentiment (RQ1), two raters hand-coded 300 randomly sampled tweets on two 1-5 scales for positivity and negativity, similar to SentiStrength. Our two raters reached a consensus of $\kappa = 0.728$ for positivity and $\kappa = 0.689$ for negativity after coding 70 tweets independently, fulfilling common thresholds for satisfactory agreement [21]. After discussing and resolving any disagreements, an additional 230 tweets were coded independently. The binary and trinary sentiment classifications of human coders were assigned analogously to how they were created for the other UDDs. We calculated accuracy, precision, recall, and F-score for each category in each classification method (binary and trinary).

3.4.2 Scale consistency and discrepancy index

To quantify scale discrepancy for RQ2, we normalized the sentiment scales to $M = 0$ and $SD = 1$, accounting for SentiStrength's truncation of scales at |5| (contrasting LIWC, tidytext and VADER). As a discrepancy index, we calculated the absolute difference between normalized scales for positivity, negativity, and overall scales for all six pairs of sentiment measures. For each tweet and scale type, the total scale discrepancy was summed up and divided by the

number of comparisons. As a robustness check for our discrepancy measure, we calculated pairwise scale correlations between methods.

3.4.3 Predictive modeling of scale discrepancy

To examine RQ3, we conducted three ordinary least square linear regression models to predict discrepancy in the (a) positivity, (b) negativity, and (c) overall scales through various tweet properties. Model assumptions (normal distribution of residuals, homoscedasticity, linearity assumptions and leverage) were investigated through graphical model tests in R. Robust standard errors (HC3 estimator [19]) were used to address residual heteroscedasticity for discrepancy in the positivity scales. Independent variables included tweet context and the number of words, likes, and retweets of a tweet. We also included binary classifications (0: negative, 1: positive or neutral) to investigate whether scale discrepancies varied with sentiment polarity and ambiguity ratings to estimate whether tweets being high in positivity and negativity were less consistently rated than tweets with less emotional valence. All independent variables had a generalized variance inflation factor (GVIF) of less than 5 [3].

4. RESULTS

4.1 Validation of Sentiment Measures (RQ1)

4.1.1 Dictionary coverage

Coverage characterizes the fit of user-defined dictionaries with the data. Coverage is the relative frequency of texts that had a least one match inside a specific dictionary. We observed a coverage of 58.91% for SentiStrength, 55.7% for LIWC, and 67.7% for VADER. The combined tidytext dictionaries had a coverage of 84.9%. As subdictionaries, coverage was 70.5% for NRC, 62.0% for AFINN, 56.9% for Bing, and 34.9% for the Loughran-McDonald dictionary.

4.1.2 Hand-coded tweets and validation

Comparing human coders with SentiStrength's scale ratings (LIWC, tidytext and VADER do not output 1-5 scales; we describe these later in this section), we found a moderate two-way random effects ICC for absolute agreement [18] for both the positivity scale, $ICC2k = 0.690$ [0.57, 0.77] and the combined, overall scale, $ICC2k = 0.683$ [0.59, 0.75]. The negativity scale exhibited worse agreement, $ICC2k = 0.448$ [0.31, 0.56]. Notably, Cohen's kappa ratings were not satisfactory with $\kappa = 0.301$ for positivity, $\kappa = 0.270$ for the overall scale, and $\kappa = 0.183$ for negativity [21].

Tables 1 and 2 describe the validity of the binary and trinary classifications for SentiStrength, LIWC, tidytext, and VADER. We found trinary classifications to have higher accuracy scores than binary classification (ranging from 85.00% to 88.33% and 56.33% to 67.00%, respectively). Notably, we found classifications of negative tweets to be less accurate than for positive tweets, with F-scores of tidytext and VADER (0.45 and 0.44, respectively) being higher compared to SentiStrength and LIWC (0.38 and 0.29, respectively). To test whether these differences were significant or random, we ran permutation tests with 250,000 simulations [39]. Tidytext and VADER improved compared to LIWC, but not to SentiStrength, (albeit marginally) significantly ($p = .058$ and $p = .047$, respectively), although only 11.67% of tweets were rated as negative by human coders.

Table 1: Binary validation results

SentiStr.			**LIWC**	
Accuracy	85.50		88.33	
	Pos/Neut	**Neg**	**Pos/Neut**	**Neg**
Precision	0.92	0.36	0.90	0.50
Recall	0.91	0.40	0.97	0.20
***F*-Score**	0.91	0.38	0.94	0.29
tidytext			**VADER**	
Accuracy	87.00		88.33	
	Pos/Neut	**Neg**	**Pos/Neut**	**Neg**
Precision	0.93	0.44	0.93	0.50
Recall	0.92	0.46	0.94	0.40
***F*-Score**	0.93	0.45	0.93	0.44

Note: Positive Tweets are either positive *or* neutral in binary classification. Support: 265 Pos/Neut, 35 Neg

Table 2: Trinary validation results

SentiStr.				**LIWC**		
Accuracy	66.00			67.00		
	Pos	**Neut**	**Neg**	**Pos**	**Neut**	**Neg**
Precision	0.66	0.75	0.36	0.65	0.71	0.50
Recall	0.77	0.63	0.40	0.78	0.69	0.20
***F*-Score**	0.71	0.69	0.38	0.71	0.70	0.29
tidytext				**VADER**		
Accuracy	56.33			65.33		
	Pos	**Neut**	**Neg**	**Pos**	**Neut**	**Neg**
Precision	0.50	0.88	0.44	0.59	0.79	0.50
Recall	0.90	0.33	0.46	0.84	0.57	0.40
***F*-Score**	0.64	0.48	0.45	0.69	0.66	0.44

Note: Support: 115 Pos, 150 Neut, 35 Neg

4.2 Consistency of Sentiment (RQ2)

4.2.1 Positivity scale

For positivity scales, LIWC and VADER were the most consistent with each other based on scale correlation ($r = .83$) and mean discrepancy (0.41 *SDs*) followed by tidytext and VADER ($r = .71$, 0.53 *SDs*) and LIWC and tidytext ($r = .63$, 0.61 *SDs*). On average, positivity scales yielded pairwise correlations of $r = .63$ and scale discrepancies of 0.60 *SDs* (Table 3).

4.2.2 Negativity scale

For negativity scales, LIWC and VADER were the most consistent with each other based on scale correlation ($r = .68$) and mean discrepancy (0.33 *SDs*) followed by SentiStrength and LIWC if based on scale correlation ($r = .61$, 1.14 *SDs*) and LIWC and tidytext if based on scale discrepancy ($r = .09$, 0.59 *SDs*). On average, negativity scales yielded pairwise correlations of $r = .35$ and scale discrepancies of 0.83 *SDs* (Table 3).

4.2.3 Overall scale

For overall scales, LIWC and VADER appeared to be closest based on scale correlation ($r = .69$) and mean discrepancy (0.54 *SDs*) followed by LIWC and tidytext ($r = .65$, 0.59 *SDs*), SentiStrength and VADER ($r = .64$, 0.65 *SDs*), and SentiStrength and LIWC ($r = .56$, 0.64 *SDs*), respectively. On average, overall scales yielded pair-wise correlations of $r = .61$ and scale discrepancies of 0.64 *SDs* (Table 3).

Table 3: Pairwise scale correlations (Corr) and discrepancy (Disc) for positivity, negativity, and overall scales of SentiStrength (SS), LIWC (LI), tidytext (TT), and VADER (VA)

	Pos		Neg		Scale	
	Corr	Disc	Corr	Disc	Corr	Disc
SS, LI	0.54	0.64	0.61	1.14	0.56	0.64
LI, TT	0.63	0.61	0.09	0.59	0.65	0.59
SS, TT	0.43	0.78	0.13	1.04	0.52	0.74
SS, VA	0.59	0.66	0.58	1.19	0.64	0.65
LI, VA	0.83	0.41	0.68	0.33	0.69	0.54
TT, VA	0.71	0.53	-.01	0.69	0.60	0.65
∅	0.63	0.60	0.35	0.83	0.61	0.64

Table 4: Linear models predicting aggregated scale discrepancy measures; N = 1,382,493

Predictor	Pos	Neg	Scale
(Intercept)	-0.56***	1.52***	-0.23***
Number of Words	-0.00***	0.01***	0.01***
Number of Likes	0.00	0.00	0.00
Number of Retweets	0.00***	0.00***	0.00***
Context [#NGSSchat]	0.02***	-0.01***	0.03***
Context [SETHs]	-0.01***	0.05***	-0.02***
Context [Chat Hour]	0.00	-0.03***	0.00
Ambiguity [SentiStr.]	0.07***	0.21***	0.07***
Ambiguity [LIWC]	0.11***	0.14***	0.11***
Ambiguity [tidytext]	0.08***	0.19***	0.10***
Ambiguity [VADER]	0.04***	0.06***	0.04***
SentiStr. Binary [1]	0.16***	-0.64***	0.00
LIWC Binary [1]	0.29***	-0.28***	0.30***
tidytext Binary [1]	0.30***	-0.47***	0.11***
VADER Binary [1]	0.14***	-0.48***	-0.03***
R^2	0.22	0.75	0.24

Note: ***p<0.001 **p<0.01 *p<0.05.

4.3 Understanding Scale Discrepancies (RQ3)

Linear models for evaluating scale discrepancies included four notable associations between tweet properties and scale discrepancies (Table 4). First, all four ambiguity measures were positively associated with scale discrepancy measures across all three models, most notably SentiStrength's ambiguity measure with negativity scale discrepancy, $\beta = 0.21$, $t(1382478) = 229.30$, $p < .001$. Second, for binary classifications (i.e., positive/neutral vs. negative), negative tweets tended to have higher negativity discrepancy and vice versa. For example, discrepancy in negativity scales was negatively associated with tweets classified as positive/neutral by SentiStrength, β = -0.64, $t(1382478)$ = -281.13, $p < .001$. Meanwhile, tidytext classifying tweets as positive was associated with increased positivity discrepancy, $\beta = 0.30$, $t(1382478) = 129.71$, $p < .001$. Third, text- and tweet-specific variables (e.g., number of words, likes, and retweets) did not seem to be associated with scale discrepancy, while tweet context had a small effect size. For instance, tweets from State Educational Twitter Hashtags were positively associated with negativity scale discrepancy, $\beta = 0.05$, $t(1382478) = 52.86$, $p < .001$. Fourth, the explained variance in scale discrepancy was highest for negativity scales at 75.3%, followed by overall scales (23.5%) and positivity scales (22.4%).

5. DISCUSSION

5.1 Key Findings

This study evaluates sentiment analysis methods on educational Twitter data. Our three main findings are as follows:

First, negative sentiment is difficult to reliably detect with dictionary approaches. This could be due to nuanced linguistic markers (e.g., sarcasm) that require advanced algorithms to be detected [31]. While this finding aligns with previous work [30], it contrasts initial validations of SentiStrength [37] on a set of around 1,000 MySpace comments [37]. Nonetheless, this highlights the importance of validating commonly used sentiment analysis tools across multiple contexts. Thus, we encourage researchers to carefully examine how negativity may be expressed in their study context.

Second, in the context of educational Twitter data, binary sentiment classifications that combine positive and neutral sentiment are substantially more robust than trinary classifications. Thus, researchers may consider computing the ratio of negative to positive/neutral tweets, similar to a recent Twitter study on the Common Core State Standards [38]. For a continuous variable, our findings suggest using an overall scale, as discrepancy in negativity was substantially associated with ambiguity and binary classifications.

Third, in the context of educational Twitter data, tidytext and VADER produce more accurate classifications of negative sentiment than LIWC. Notably, tidytext also has the highest dictionary coverage. Thus, educational researchers are encouraged to aggregate multiple dictionaries or to create domain-specific sentiment dictionaries for more reliable measures of negative sentiment, for instance, similar to a previous study investigating political expression [11].

5.2 Limitations

This study has two notable limitations. First, the sample size of the training data is relatively small ($N = 300$). That said, it is comparable to sample sizes of previous sentiment measure validation studies [28]. Similarly, the lack of negative tweets in our training data (11.67%) may limit inferences about that particular type of sentiment. The number of negative tweets is considerably smaller compared to previous validation studies utilizing text data from sources such as MySpace, Twitter, BBC forums, and YouTube that include up to 86.84% negative sentiment [8]. Therefore, future validation studies should deliberately sample more negative tweets [13] or sample from contexts with higher expected negativity, such as Common Core State Standards hashtags [38]. Second, this study focuses on dictionary-based sentiment analysis, while future studies might also consider feature extraction and word co-occurrence methods [15].

5.3 Implications

This study highlights the importance of coverage, validity, and scale discrepancy in sentiment analysis, specifically for negative sentiment. For educational Twitter data, this study recommends using binary classifications or overall scales, preferably derived from tidytext or VADER, and encourages replication studies[1] across more educational contexts.

[1]Code: https://github.com/jrosen48/comparing-sentiment

6. REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38, 2011.

[2] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, volume 1, pages 2–1. Citeseer, 2005.

[3] T. A. Craney and J. G. Surles. Model-dependent variance inflation factor cutoff values. *Quality Engineering*, 14(3):391–403, 2002.

[4] A. Edgerton. Learning from standards deviations: Three dimensions for building education policies that last. *American Educational Research Journal*, 57(4):1525–1566, 2020.

[5] S. Elbagir and J. Yang. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 122, page 16, 2019.

[6] C. Fischer, B. Fishman, and S. Y. Schoenebeck. New contexts for professional learning: Analyzing high school science teachers' engagement on Twitter. *AERA Open*, 5(4), 2019.

[7] C. Fischer, Z. Pardos, R. Baker, J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1):130–160, 2020.

[8] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38, 2013.

[9] J. Grimmer and B. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

[10] L. Hansen, A. Arvidsson, F. Nielsen, E. Colleoni, and M. Etter. Good friends, bad news-affect and virality in twitter. In *Future information technology*, pages 34–43. Springer, 2011.

[11] M. Haselmayer and M. Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & quantity*, 51(6):2623–2646, 2017.

[12] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[13] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[14] M. Ibrahim. Extracting weight in Twitter SentiStrength dataset to determine sentiment polarity. *Journal of Information Systems Research and Innovation*, 10(3):245–265, 2016.

[15] R. Iliev, M. Dehghani, and E. Sagi. Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, 7(2):265–290, 2015.

[16] J. Kahn, R. Tobin, A. Massey, and J. Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, pages 263–286, 2007.

[17] M. Kern, G. Park, J. Eichstaedt, A. Schwartz, M. Sap, L. Smith, and L. Ungar. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4):507, 2016.

[18] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[19] J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[20] T. Loughran and B. McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[21] M. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[22] A. Mittal and A. Goel. Stock prediction using Twitter sentiment analysis. *Standford University, CS229*, 15, 2012.

[23] S. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*, pages 201–237. Woodhead Publishing, 2016.

[24] S. Mohammad and P. Turney. NRC emotion lexicon. Technical report, National Research Council, Canada, 2013.

[25] I. Mozetič, L. Torgo, V. Cerqueira, and J. Smailović. How to evaluate sentiment classifiers for Twitter time-ordered data? *PloS one*, 13(3):e0194317, 2018.

[26] J. Pennebaker, M. Francis, and R. Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[27] M. Polikoff, T. Hardaway, J. Marsh, and D. Plank. Who is opposed to Common Core and why? *Educational Researcher*, 45(4):263–266, 2016.

[28] R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.

[29] M. Rambocas, J. Gama, et al. Marketing research: The role of sentiment analysis. Technical report, Universidade do Porto, Faculdade de Economia do Porto, 2013.

[30] F. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[31] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714, 2013.

[32] J. Rosenberg, C. Borchers, E. Dyer, D. Anderson, and C. Fischer. Advancing new methods for understanding public sentiment about educational reforms: The case

of Twitter and the Next Generation Science Standards. *OSF Preprints*, 2020.

[33] J. Rosenberg, J. Reid, E. Dyer, M. Koehler, C. Fischer, and T. McKenna. Idle chatter or compelling conversation? the potential of the social media-based #ngsschat network for supporting science education reform efforts. *Journal of Research in Science Teaching*, 57(9):1322–1355, 2019.

[34] J. Silge and D. Robinson. tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, 1(3), 2016.

[35] J. Silge and D. Robinson. *Text mining with R: A tidy approach*. O'Reilly Media, Inc., 2017.

[36] Y. Tausczik and J. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[37] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

[38] Y. Wang and D. Fikis. Common Core State Standards on Twitter: Public sentiment and opinion leaders. *Educational Policy*, 33(4):650–683, 2019.

[39] P. Yeh. More accurate tests for the statistical significance of result differences. *arXiv preprint cs/0008005*, 2000.