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", "something") [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 (1,382,493 tweets, 156,446 users) or "next generation science standard[s]" (1,382,493 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 positive and neutral tweets, such that neutral tweets were 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 text 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, \( ICC_{2k} = 0.690 \) \([0.57, 0.77]\) and the combined, overall scale, \( ICC_{2k} = 0.683 \) \([0.59, 0.75]\). The negativity scale exhibited worse agreement, \( ICC_{2k} = 0.448 \) \([0.31, 0.56]\). 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.

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

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

<table>
<thead>
<tr>
<th>Table 1: Binary validation results</th>
<th>SentiStr.</th>
<th>LIWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>85.50</td>
<td>88.33</td>
</tr>
<tr>
<td>Pos/Neut</td>
<td>0.92</td>
<td>0.36</td>
</tr>
<tr>
<td>Neg</td>
<td>0.90</td>
<td>0.50</td>
</tr>
<tr>
<td>Recall</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>( F)-Score</td>
<td>0.91</td>
<td>0.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Trinary validation results</th>
<th>SentiStr.</th>
<th>LIWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>66.00</td>
<td>67.00</td>
</tr>
<tr>
<td>Pos</td>
<td>0.66</td>
<td>0.75</td>
</tr>
<tr>
<td>Neut</td>
<td>0.36</td>
<td>0.65</td>
</tr>
<tr>
<td>Neg</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Recall</td>
<td>0.77</td>
<td>0.63</td>
</tr>
<tr>
<td>( F)-Score</td>
<td>0.71</td>
<td>0.71</td>
</tr>
</tbody>
</table>

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).
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 across more educational contexts.

---

1 Code: https://github.com/jrosen48/comparing-sentiment
6. REFERENCES


