



Machine learning to detect invalid text responses: Validation and comparison to existing detection methods

Ryan C. Yeung¹ · Myra A. Fernandes¹

Accepted: 17 January 2022 / Published online: 17 February 2022
© The Psychonomic Society, Inc. 2022

Abstract

A crucial step in analysing text data is the detection and removal of invalid texts (e.g., texts with meaningless or irrelevant content). To date, research topics that rely heavily on analysis of text data, such as autobiographical memory, have lacked methods of detecting invalid texts that are both effective and practical. Although researchers have suggested many data quality indicators that might identify invalid responses (e.g., response time, character/word count), few of these methods have been empirically validated with text responses. In the current study, we propose and implement a supervised machine learning approach that can mimic the accuracy of human coding, but without the need to hand-code entire text datasets. Our approach (a) trains, validates, and tests on a subset of texts manually labelled as valid or invalid, (b) calculates performance metrics to help select the best model, and (c) predicts whether unlabelled texts are valid or invalid based on the text alone. Model validation and evaluation using autobiographical memory texts indicated that machine learning accurately detected invalid texts with performance near human coding, significantly outperforming existing data quality indicators. Our openly available code and instructions enable new methods of improving data quality for researchers using text as data.

Keywords Machine learning · Text classification · Text as data · Careless responding · Autobiographical memory

Detecting invalid data is a vital part of conducting research. In particular, psychological research can be vulnerable to invalid data in that participants' responses may not always reflect "true" values (Curran, 2016). Common threats to validity include participants misunderstanding instructions, participants misrepresenting themselves, or participants responding carelessly or with insufficient effort (Curran, 2016; Groves et al., 2009; Johnson, 2005; Krosnick, 1999). In practice, many participants show evidence of careless/insufficient effort (C/IE) responding; typical estimates are around 10% of responders (Maniaci & Rogge, 2014; Meade & Craig, 2012), though some report rates much higher (46%; Brühlmann et al., 2020). This is worrying because even relatively small percentages of C/IE respondents (~5–10%) can cause measurable losses in data quality or even alter the

conclusions drawn from a study (Arias et al., 2020; Credé, 2010; Huang et al., 2015; Wood et al., 2017; Woods, 2006).

Text responses appear to be no exception. In a study analysing ~7000 text responses to a survey, 8% of these responses were manually labelled as irrelevant or miscellaneous text and removed from further analysis (Cunningham & Wells, 2017). In another sample of over 600,000 text responses, 28% were manually identified as either meaningless, or meaningful but not useful (Etz et al., 2018). As such, invalid data detection has been recognized as important when analysing open-ended or free text responses (Etz et al., 2018; Müller et al., 2014), especially for use in computational text analysis (Banks et al., 2018). To date, many methods for detecting invalid responses have been proposed (for reviews, see Curran, 2016; Meade & Craig, 2012). However, only some of these data quality indicators are applicable to *text* responses, including response speed or completion time (Curran, 2016; Huang et al., 2012; Leiner, 2019; Maniaci & Rogge, 2014), character or word count (Chen, 2011; Etz et al., 2018), and manual coding (Kennedy et al., 2020). In the present study, we propose an alternative method of detecting invalid text responses using supervised machine learning (ML). We first review past methods of

✉ Ryan C. Yeung
rcyeung@uwaterloo.ca

¹ Department of Psychology, University of Waterloo,
Psychology, Anthropology, and Sociology (PAS) Building,
200 University Avenue West, Waterloo, ON N2L 3G1,
Canada

detecting invalid responses and then provide an overview to our current approach.

Time-based data quality indicators

Past studies have considered response speed or completion time to be an indicator of response validity. The rationale is that participants who answer exceptionally quickly are unlikely to have read the instructions or given substantial thought to an item. Though extreme response speeds on either end of the spectrum (i.e., fast or slow) are potential signals of C/IE responding, only very fast respondents are typically excluded (Huang et al., 2012; Wood et al., 2017).

Some researchers have used static cutoffs for completion time, whereby invalid responses were defined as those submitted in under two seconds (Kaczmarek et al., 2017; Wise & Kong, 2005). Expanding upon this strategy, Huang et al. (2012) used cutoffs where invalid respondents were those who completed the study in less than 2 s *per item* (or less than 52 s per page), whereas Niessen et al. (2016) used 3 s per item. However, these static cutoffs have been largely designed with closed-ended responses (e.g., Likert-type scales) in mind, rather than open-ended or text responses. Relative cutoffs for completion time might be more applicable to text responses since relative cutoffs for valid/invalid classification are based on one's own data. For example, Maniaci and Rogge (2014) proposed labelling responses as invalid if they are completed in less than half the mean completion time, and these authors emphasized recalibrating this cutoff for one's own studies. A comparable approach was suggested by Leiner (2019), which involves computing speed factors (median completion time divided by the participant's completion time) for the total study and/or the pages containing the responses of interest. Similar to Maniaci and Rogge (2014), participants completing pages more than twice as quickly as the median (speed factor > 2) are flagged as suspicious using this cutoff (Leiner, 2019).

Text-based data quality indicators

Text response length has also been thought to signal invalid responding in that very short answers are unlikely to contain enough content to constitute a meaningful response to the item. Further, common types of invalid text responses are typically short (e.g., “don't know” responses; Behr et al., 2014; Holland & Christian, 2009; Kennedy et al., 2020; Scholz & Zuell, 2012). Despite general agreement that low character count can indicate poor data quality, static cutoffs for character count have differed considerably in the literature. In one case, responses that were 50 or fewer characters long (Kaczmarek et al., 2017) were treated as invalid,

whereas other researchers have implemented minimums of 200 (Banks et al., 2018) or 900 characters (Jones et al., 2021). As an alternative strategy, some authors have used multiple rounds of cutoffs to filter out text responses shorter than either six, 16, or 51 characters (Etz et al., 2018), which flagged the most suspicious cases first (fewer than six characters) and the less suspicious cases after (fewer than 16 or 51 characters). Finally, some relative cutoffs have been used, such that responses below the 50th percentile on character count have been filtered out before analysis (Viani et al., 2021).

Like character count, low word count has also been conceptualized as a signal of invalid responses to open text questions (Chen, 2011). Some minimum cutoffs for word count have included one (Kaczmarek et al., 2017) or three words (Gagnon et al., 2019), though no empirical support is presented for these cutoffs. Some more deliberate approaches have involved explicitly asking participants to write at least 50 words, and then later checking whether responses met that 50-word minimum (Banks et al., 2018; Brühlmann et al., 2020). However, other studies without an instructed minimum word count may find this cutoff inappropriate. Indeed, Banks et al. (2018) note that “a minimum number of words cannot be recommended because this issue is context dependent” (p. 452). Researchers appear to have agreed with this sentiment, given that cutoffs for character/word count have varied so widely in the literature. The wide range of static cutoffs proposed in past research might also reflect the diversity of tasks and items used across studies. As such, relative cutoffs might be preferable since they are dependent on researchers' own datasets, rather than static cutoffs determined by others' datasets.

The final text-based data quality indicator is content analysis of the text response, or manual coding (Brühlmann et al., 2020; Kennedy et al., 2020; Scholz & Zuell, 2012; Sischa et al., 2020; Smyth et al., 2009). Theoretically, coding is the most direct measure of whether a text response is valid or not – while response time and character/word count might *signal* an invalid response, the content of the text *dictates* the response's validity. Coding schemes also allow researchers to decide what constitutes a valid or invalid response in their own contexts. For instance, Brühlmann et al. (2020) manually labelled text responses for whether they were “thematically substantive” (p. 5), consisted of complete sentences, answered subquestions of the original question, and answered subquestions elaboratively. In terms of invalid responses, many researchers have arrived at similar codes, including refusals to respond, “don't know” or “no opinion” responses, irrelevant or non sequitur responses, and incomprehensible or nonsensical responses (Behr et al., 2014; Kennedy et al., 2020; Scholz & Zuell, 2012). Despite these advantages of using coding to detect invalid responses, it remains difficult to conduct compared to other

data quality indicators due to being more time- and labour-intensive. Coding typically requires a multitude of human hours for manual analysis of the content within texts. While this approach may be feasible for studies consisting of small sample sizes, it becomes untenable when texts are too long or too numerous.

Machine learning to detect invalid responses

To overcome these limitations, we suggest an approach that leverages ML to keep the advantages of coding, but without the need to manually examine entire datasets (Nelson et al., 2018; Sebastiani, 2002). Using ML to detect C/IE responding has been a very recent development (Gogami et al., 2021; Schroeders et al., 2021) and these studies have not yet attempted to detect invalid text responses. Specifically, Schroeders et al. (2021) used ML to detect careless responders using a combination of existing data quality indicators (e.g., Mahalanobis distance, even-odd consistency). Though this approach appeared to work well with simulated data (balanced accuracy = .86–.95), performance dropped considerably with data from actual participants (balanced accuracy = .60–.66). A similar approach was taken by Gogami et al. (2021), who also used ML to detect careless responders, but they logged additional data (such as scrolling behaviour, deleting text, and reselecting response options) in concert with traditional data quality indicators (e.g., character count, response time). Implementing supervised ML with those data was reported to effectively detect careless responders to their survey, which included open text items (recall/sensitivity = .86; Gogami et al., 2021). However, these past studies attempted to detect invalid *respondents* (rather than invalid *responses*), and neither used the content of text responses as data. In the present study, we ask if ML can detect invalid text responses *based on the content of participants' text responses alone*. It is one thing to flag individuals who are carelessly responding to the survey overall; it is another to identify single text responses that are meaningless or irrelevant to the question being asked.

This distinction is relevant from both theoretical and practical standpoints. For one, it is theoretically important to consider that invalid text responses may not represent C/IE responding as it is typically conceptualized for other response types (e.g., Likert-type scales). It is certainly plausible that some participants will write meaningless text to finish the study faster. Importantly, it's equally plausible that some participants will write meaningful but not useful text (Etz et al., 2018; Kennedy et al., 2020). For example, participants may simply fail to understand the question or write that they are declining to respond (Groves et al., 2009; Scholz & Zuell, 2012). In such cases, participants may take their time writing thoughtful or lengthy responses, causing

them to pass basic time- or text-based data quality indicators – yet these texts would ultimately still be invalid and still need to be detected. Further, a practical consideration is that researchers may not have the ability to administer or analyse items beyond the text response itself. Researchers currently have scarce options if other safeguards were not included at the time of testing; text- or content-based measures are the only guarantee.

We propose that invalid text responses can be effectively detected using ML. Classifying texts into specific categories based on their content alone has been a long-established problem in ML literature (Aggarwal & Zhai, 2012; Yang, 1999), to which many effective solutions have been proposed. For example, researchers have used ML to successfully classify texts as sarcastic or nonsarcastic (Sarsam et al., 2020), abusive or nonabusive (Nobata et al., 2016), or expressing positive or negative sentiment (Liu & Zhang, 2012). Similar to these tasks, the current approach can be conceptualized as a supervised binary classification task (Chicco, 2017): it is *supervised* in that there are manually labelled values (valid or invalid) that we train, validate, and test the model upon; and it is *binary classification* in that there are exactly two possible target values (e.g., 0 = *valid*, 1 = *invalid*) for each text response. As such, we suggest a supervised ML approach that (a) trains, validates, and tests on a subset of texts manually labelled as valid or invalid, (b) calculates performance metrics to help select the best model, and (c) predicts whether unlabelled texts are valid or invalid based on the text alone. Finally, we compare the ML model's predictions to existing data quality indicators to assess its performance relative to traditional methods (e.g., time- or text-based cutoffs).

Here, we examined the effectiveness of our ML approach using text responses from an online survey, in which we asked participants about any recurrent involuntary autobiographical memories (IAMs) they may have experienced (Yeung & Fernandes, 2020). These are memories that spring to mind unintentionally and repetitively (Berntsen, 1996; Berntsen & Rubin, 2008), which occur commonly in daily life among large and diverse samples (Berntsen & Rubin, 2008; Brewin et al., 1996; Bywaters et al., 2004; Marks et al., 2018; Yeung & Fernandes, 2020; Yeung & Fernandes, 2021). During our online survey, if participants reported that they had experienced recurrent IAMs within the past year, we asked them to write a text description of their most frequently recurring one. Importantly, these text data seem prone to common concerns about validity. For instance, in order to provide a valid response, participants must have (a) read and understood the instructions, (b) accurately judged whether they had experienced this type of memory, (c) retrieved relevant information about a specific memory, and (d) decided to describe this memory in writing. Declining to perform any of these steps (or performing them carelessly)

would have likely led to invalid responses (Groves et al., 2009; Krosnick, 1999). Others have also demonstrated that an ML approach can successfully classify autobiographical memory texts into categories (i.e., specific vs. nonspecific memories; Takano et al., 2017). As such, we believe that our dataset, consisting of involuntary autobiographical memories, is a realistic example of psychological research involving text as data, from which invalid responses could be detected using ML.

Methods

Participants

Undergraduate students were recruited from the University of Waterloo, who participated in the current study in return for course credit. Data were collected in five waves between September 2018 and February 2020, with each wave occurring at the start of an academic term (i.e., Fall/September, Winter/January, Spring/May). In total, 6187 unique participants were recruited, and they produced 3624 text responses. Of these participants, 71% were women, 28% were men, and 1% were nonbinary, genderqueer, or gender nonconforming. Mean age was 19.9 ($SD = 3.3$, range = 16–49).

Materials

Recurrent Memory Scale The Recurrent Memory Scale (Yeung & Fernandes, 2020) assessed participants' recurrent IAMs. Participants indicated if they had experienced at least one recurrent IAM within the past year, not within the past year, or never (Berntsen & Rubin, 2008). If they had experienced at least one within the past year, they wrote a brief description¹ of their one most frequently recurring IAM and rated it on a series of five-point Likert scales (e.g., frequency of recurring, valence; see Yeung & Fernandes, 2020 for details). Only participants' text responses describing their memories were used in the current study; no ratings were analysed.

Procedure

Undergraduate students enrolled in at least one psychology course at the University of Waterloo self-registered for the 60-minute online study. After providing informed consent,

participants completed a battery of questionnaires in a randomized order, including the Recurrent Memory Scale. For the current study, we analysed only the text responses from this scale; all other measures were unrelated to the current study. All procedures were approved by the University of Waterloo's Office of Research Ethics (Protocol #40049).

Data preparation

Text data were preprocessed following recommended standards (Banks et al., 2018; Kobayashi et al., 2018), including stop word removal (removing highly frequent words) and lemmatization (reducing words to their roots). During preprocessing, 50 of the 3624 responses were removed because they no longer contained any text (e.g., the original text contained only nonalphabetic characters or stop words), resulting in 3574 processed text responses.

The processed texts from the labelled subset ($n = 940$) were then split into training, validation, and test sets. Specifically, 20% of these processed and labelled text responses were partitioned off as the test set ($n = 188$), to be held out for final evaluation of the model (i.e., never used during training or validation; Chicco, 2017). The remaining 80% of the processed and labelled text responses constituted the training ($n = 676$ – 677) and validation sets ($n = 75$ – 76), which were split ten times using stratified K -fold cross-validation ($K = 10$; Geisser, 1975; Marcot & Hanea, 2021).

Texts were then represented using a bag-of-words, unigram² approach (Grimmer & Stewart, 2013), which decomposes texts into lists of singular words without retaining word order. Specifically, processed text responses were transformed into document-term matrices (Grimmer & Stewart, 2013; Kobayashi et al., 2018; Welbers et al., 2017) and weighted³ using term frequency-inverse document frequency (TF-IDF; Salton & Buckley, 1988). In brief, TF-IDF accounts for how commonly a word is used across the entire dataset; words that are prevalent across many different texts are down-weighted since these common words are unlikely to distinguish between texts well (Salton & Buckley, 1988).

² We also implemented text representations using both unigrams (single tokens) and bigrams (pairs of tokens). MCCs obtained using this unigram & bigram model were lower than MCCs obtained using just unigrams. This is the case both for mean MCC across all models (.37 with unigrams & bigrams, compared to .40 with just unigrams) and MCC of the best performing model (.55 with unigrams & bigrams, compared to .60 with just unigrams). As such, we had no evidence to suggest that bigrams improved our models; all further analyses were conducted with just unigrams. For completeness, our code allows users to represent texts using either unigrams, or unigrams & bigrams. Further details are available on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/3HS56>).

³ Our code also allows users to opt out of TF-IDF weighting. By selecting this option, one can use term frequency (count) instead, which simply counts the instances of each word without accounting for the words' frequencies across the texts overall.

¹ Instructions for writing the text response were as follows: "Please think about the most frequent recollection that has recurrently popped into your mind by itself within the most recent year. The following questions will ask you about that memory. Please briefly describe your memory of the event in 3–5 sentences, without any identifying information" (Yeung & Fernandes, 2020).

Results

Manual coding of text response subset

As part of a prior study (Yeung & Fernandes, 2020), research assistants were trained to label a subset of the text responses ($n = 949$; 26% of responses collected) as either valid or invalid. For each text, two research assistants independently judged whether the text was valid (the participant's text described a memory) or invalid (the participant's text did not describe a memory). Specifically, they were instructed to judge whether the participant had chosen to describe a memory of a real, experienced event in their text response. One coder (C1) labelled all 949 texts, whereas the other three coders (C2, C3, C4) labelled approximately 316 (33%) texts each. All coders were naïve to the hypotheses of the study.

Interrater reliability between coders was calculated using intraclass correlation coefficients (ICCs) based on a mean-rating ($k = 2$), absolute agreement, one-way random effects model (Koo & Li, 2016). For the validity judgment, the average measures ICC between C1 and C2, C3, or C4 was .84 (95% CI [.82, .86]), indicating good reliability, $F(948, 949) = 6.2, p < .001$.

To compare interrater reliability across coder pairs, ICCs were also calculated between C1 and each of C2, C3, and C4 separately. Based on a mean-rating ($k = 2$), absolute agreement, two-way random effects model, the average measures ICC for C1–C2 was .78 (95% CI [.73, .81]); for C1–C3 was .88 (95% CI [.85, .90]); and for C1–C4 was .86 (95% CI [.83, .88]), indicating good reliability across coder pairs (all $ps < .001$).

Because confidence intervals did not overlap for C1–C2 compared to C1–C3 and C1–C4, we recalculated the ICC between C1 and C2, C3, or C4 (one-way random effects model) with C2's labels removed. The average measures ICC with C1–C2 excluded was .87 (95% CI [.85, .88]), $F(631, 632) = 7.6, p < .001$, which was not significantly different from the ICC with C1–C2 included (.84, 95% CI [.82, .86]). As such, we retained C2's labels in the following analyses. The final labelled subset consisted of 949 texts, 878 of which were labelled as valid by at least one coder (92.5%), and 71 of which were labelled as invalid by both coders⁴ (7.5%).

⁴ In prior work, we accepted a text as invalid if *at least one* of two coders judged the text as invalid (Yeung & Fernandes, 2020). Disagreements between coders on this judgment were relatively rare ($n = 47$). However, for the purposes of ML, we currently use a stricter threshold for manual labels since the quality of this labelled subset has measurable impacts on model performance (Song et al., 2020).

Invalid text detection

Supervised ML was implemented using Python (Harris et al., 2020; Lemaître et al., 2017; McKinney, 2010; Pedregosa et al., 2011; Seabold & Perktold, 2010; Waskom, 2021).

Model validation

Using stratified ten-fold cross-validation, we selected the best performing model based on the Matthews correlation coefficient⁵ (MCC; Chicco & Jurman, 2020; Luque et al., 2019), which indexes how well predictions are made on positive (invalid) as well as negative (valid) cases. While no single performance metric is exhaustive, the MCC offers many advantages in our context. First, it is unaffected by the high class imbalance (a 9:1 ratio of valid to invalid texts) present in our study (Chicco & Jurman, 2020). In contrast, other popular metrics such as accuracy (the proportion of correct classifications) are inadequate with imbalanced datasets since they are strongly impacted by the ratio of classes present in the data (Chawla et al., 2002; Chawla, 2009; He & Garcia, 2009; Menardi & Torelli, 2012). For example, trivially assigning the majority label (e.g., valid) to all cases can achieve impressive accuracy scores even though this strategy does not actually perform the task at hand (i.e., detecting invalid texts). Conversely, MCCs can only be high in a binary classification task if the model is correctly labelling both positive and negative cases (Chicco & Jurman, 2020). To supplement the information provided by MCCs, we also present macro F1 scores (the arithmetic mean over individual F1 scores; Opitz & Burst, 2019) in Table 1.

We defined the best model as the combination of classifier and resampler that resulted in the highest mean MCC across the ten folds (Lever et al., 2016), with preference given to simpler models (Chicco, 2017; Hand, 2006). Classifiers are algorithms that dictate the class to which each case is predicted to belong (Aggarwal & Zhai, 2012); resamplers are methods of temporarily modifying datasets with class imbalance (i.e., datasets where one class highly outnumbers another). Because our data were highly imbalanced, we used resampling methods to balance the classes' proportions during training (Estabrooks et al., 2004; He & Garcia, 2009). Classifiers tend to label minority cases poorly (in our case, invalid texts) unless this imbalance is accounted for, since

Footnote 4 (continued)

Hence, here we adopted a more stringent scheme of only accepting a text as invalid if *both* coders independently judged the text as invalid.

⁵ The MCC is a special case of the phi coefficient, which is in turn a special case of the Pearson product-moment correlation (see Footnote 6). Importantly, MCCs, phi coefficients, and Pearson product-moment correlations are mathematically equivalent in our case.

Table 1 Cross-validation performance of all classifiers and resamplers

Classifier	Resampler	Mean MCC (<i>SD</i>)	Mean Macro F1 Score (<i>SD</i>)
Naïve Bayes	SMOTE	.53 (.14)	.76 (.07)
	SVM-SMOTE	.60 (.29)	.80 (.14)
	ADASYN	.52 (.14)	.75 (.07)
Logistic Regression	SMOTE	.42 (.21)	.70 (.11)
	SVM-SMOTE	.42 (.25)	.68 (.12)
	ADASYN	.44 (.21)	.70 (.11)
Random Forest	SMOTE	.41 (.25)	.69 (.13)
	SVM-SMOTE	.43 (.27)	.69 (.13)
	ADASYN	.51 (.21)	.74 (.10)
Decision Tree	SMOTE	.30 (.19)	.64 (.09)
	SVM-SMOTE	.25 (.17)	.62 (.08)
	ADASYN	.20 (.12)	.59 (.06)
Gradient Boost	SMOTE	.28 (.22)	.63 (.11)
	SVM-SMOTE	.28 (.18)	.63 (.09)
	ADASYN	.25 (.22)	.62 (.11)
Linear SVM	SMOTE	.44 (.29)	.71 (.14)
	SVM-SMOTE	.44 (.29)	.71 (.14)
	ADASYN	.44 (.29)	.71 (.14)
Logistic Regression with Elastic Net Regularization ^a	SMOTE	.43 (.27)	.70 (.13)
	SVM-SMOTE	.41 (.32)	.69 (.15)
	ADASYN	.42 (.27)	.70 (.13)

Note. MCC = Matthews correlation coefficient, SD = standard deviation, SVM = support vector machine, SMOTE = synthetic minority oversampling technique, SVM-SMOTE = support vector machine SMOTE, ADASYN = adaptive synthetic sampling. Bolded mean MCCs and mean macro F1 scores denote the five highest values. ^a We used an elastic net mixing parameter (L1 ratio) of .15

classifiers can simply ignore the minority class and still achieve high performance metrics (Estabrooks et al., 2004).

Here, we compared the performance of common classifiers and resamplers that have been previously applied to text classification (Aggarwal & Zhai, 2012; Fernández et al., 2018). For classifiers, we compared multinomial naïve Bayes (McCallum & Nigam, 1998), logistic regression (Ng & Jordan, 2002), random forest (Breiman, 2001), decision tree (Safavian & Landgrebe, 1991), gradient boost (Friedman, 2001), linear support vector machine (SVM; Fan et al., 2008), and logistic regression with elastic net regularization (Zou & Hastie, 2005). For resamplers, we compared Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002), Support Vector Machine SMOTE (SVM-SMOTE; Nguyen et al., 2011), and adaptive synthetic sampling (ADASYN; He et al., 2008).

As shown in Table 1, mean MCCs were highest for naïve Bayes & SVM-SMOTE; naïve Bayes & SMOTE; and naïve Bayes & ADASYN. In our case, we selected naïve Bayes & SVM-SMOTE as the best model because of its highest mean MCC compared to the other models ($M_{\text{MCC}} = .60$). Performance of this cross-validation trained model was similar between test ($M_{\text{MCC}} = .62$, $SD = .04$) and validation (M_{MCC}

$= .60$, $SD = .29$), suggesting reasonable ability to generalize to unseen data.

Model evaluation

The best model was then trained on the full training and validation set (80% of the labelled data; $n = 752$) and evaluated on the test set (20% of the labelled data; $n = 188$; Sebastiani, 2002). The best model was never trained on any of the test data (Chicco, 2017). The correlation⁶ between the best model's predictions and human coding was positive, statistically significant, and large ($r = .57$, 95% CI [.46, 0.66], $t(186) = 9.35$, $p < .001$). To illustrate the model's performance, we present the confusion matrix (Table 2), which indicates how each text response was classified according to human coding versus the ML model. Two of the four cells reflect cases where human

⁶ Technically, this correlation is a phi coefficient since both variables are dichotomous. Phi coefficients are mathematically identical to Pearson product-moment correlations if one assigns distinct numerical values to each variable's dichotomous labels (Fleiss et al., 2003). As such, we report r instead of ϕ for interpretability's sake.

Table 2 Confusion matrix for best model's predictions on test data ($n = 188$)

		Model Predicted	
		Valid	Invalid
Human Labelled	Valid	169	7
	Invalid	4	8

Comparison to existing data quality indicators

As alternative methods of detecting invalid text responses, we implemented non-ML data quality indicators as recommended by existing literature (Curran, 2016; Etz et al., 2018; Leiner, 2019; Maniaci & Rogge, 2014; Meade & Craig, 2012). First, we detected univariate outliers for all indicator variables (total survey completion time, page completion

Table 3 Example text responses for best model's predictions on test data ($n = 188$)

		Model Predicted	
		Valid	Invalid
Human Labelled	Valid	"In general, situations where I have made a mistake or I have been uncomfortable pop into my head"	"memory was very accurate"
	Invalid	"It is about a dark future I imagine for some reason where my friend has passed away and what are the repercussions that follow."	"It is difficult to remember which events I remember when I am trying to remember them."

coding and the model agreed, which we interpret as the model making "correct" classifications. These include cases where both methods labelled a text response as invalid (true positives or hits; $n = 8$), and cases where both methods labelled a text response as valid (true negatives or correct rejections; $n = 169$). The other two cells reflect cases where human coding and the model disagreed, which we interpret as the model making "incorrect" classifications. These include cases where humans coded a text response as valid, but the model labelled it invalid (false positives or false alarms; $n = 7$), and cases where humans coded a text response as invalid, but the model labelled it valid (false negatives or misses; $n = 4$). Further, we also provide examples of correctly and incorrectly classified text responses from the test set (Table 3).

Finally, we extracted the words that were most predictive of each class according to the ML model (i.e., feature importance). For the invalid class, the ten most predictive words were "remember", "memory", "thing", "song", "listen", the negator "n't", "think", "happen", "sure", and "like". The ML model seems to have identified that invalid texts tended to contain terms related to the metacognitive aspects of remembering (e.g., being unsure if an event happened), rather than descriptions of the remembered event itself. In contrast, the most predictive words for the valid class included words like "ambulance", "awkward", "couple", "girlfriend", "lonely", "outburst", "shocked", "sister", "speak", and "visit". These words seem to reflect the ML model identifying that valid texts tended to contain specific details from remembered events (e.g., persons present, emotions felt).

time, character count, word count) in the labelled subset ($n = 949$) using the median absolute deviation (MAD) method (Leys et al., 2013; Leys et al., 2019). For each variable, values more than three MADs plus or minus the median were winsorized (i.e., replaced by the three MAD cutoff value). Problematic skew (1.1–30.2) and kurtosis (1.6–920.6) values were resolved following this process (skew = 0.8–0.9, kurtosis = -0.9–0.3).

Time-based measures Similar to Maniaci and Rogge (2014) and Leiner (2019), we calculated speed factors for both the total survey as well as the one page containing the text response of interest. Speed factors were computed by dividing the median completion time by the participant's completion time; a speed factor of 2 represents the participant completing the total survey or page twice as quickly compared to the median completion time. We used the relatively lenient cutoff of 2 suggested by Leiner (2019), such that instances of participants finishing either the total survey or the page more than twice as fast as the median time were labelled as invalid responses.

Text-based measures Following Etz et al. (2018), we used two cutoffs for character count: a minimum of 16 characters or 51 characters (inclusive). Both cutoffs have been described as effective methods of detecting meaningful but not useful data (Etz et al., 2018), where responses shorter than either 16 or 51 characters (exclusive) were labelled as invalid.

Table 4 Correlation matrix of data quality indicators

	1	2	3	4	5	6	7
1. Human coding	–						
2. Model predictions	.57***	–					
3. Character count (min 16)	.40***	.16*	–				
4. Character count (min 51)	.32***	.33***	.29***	–			
5. Word count (min 10)	.31***	.32***	.28***	.87***	–		
6. Total speed factor (max 2)	-.07	.01	-.03	.20**	.19**	–	
7. Page speed factor (max 2)	.02	-.08	.19*	.25***	.31***	.31***	–

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 5 Confusion matrices for existing data quality indicators on test data ($n = 188$)

		16-Character Cutoff		51-Character Cutoff		10-Word Cutoff	
		Valid	Invalid	Valid	Invalid	Valid	Invalid
Human Labelled	Valid	176	0	161	15	160	16
	Invalid	10	2	6	6	6	6

For word count, a cutoff of minimum 50 words has been suggested (Banks et al., 2018; Brühlmann et al., 2020). However, this cutoff was inappropriate for the current dataset, as our text responses were typically shorter than 50 words long (in the labelled subset, $M_{\text{words}} = 31.1$, $SD_{\text{words}} = 21.2$). Instead, we opted for a cutoff of minimum ten words long (inclusive), which was approximately 1 SD below the mean for word count⁷. Responses shorter than ten words (exclusive) were labelled as invalid based on this cutoff.

Correlations to human coding To compare all data quality indicators' abilities to estimate human coding, we conducted correlations between human coding, model predictions, character count cutoffs, word count cutoff, total speed factor cutoff, and page speed factor cutoff (see Table 4). These analyses were confined to the test set ($n = 188$) to ensure that the ML model was never trained or validated on these data. To account for multiple comparisons, Holm corrections were applied.

The correlations illustrate that the ML model's predictions successfully estimated human coding. Of the non-ML data quality indicators, only the text-based measures

(character count cutoffs, word count cutoff) were significantly correlated with human coding ($r_s = .31$ – $.40$, $p_s < .001$). Time-based measures were not significantly correlated with human coding ($r_s = -.07$ – $.02$, $p_s = .4$ – $.8$).

We then tested whether the human–model correlation was significantly different from the correlations between human coding and existing text-based data quality indicators (Graham & Baldwin, 2014). To do so, we conducted Williams' tests to compare correlation coefficients derived from a single sample (Dunn & Clark, 1971; Neill & Dunn, 1975). These Williams' tests indicated that the human–model correlation was significantly stronger than the 16-character count cutoff ($t(186) = 2.19$, $p = .03$, Cohen's $q = .22$), the 51-character count cutoff ($t(186) = 3.45$, $p < .001$, Cohen's $q = .31$), and the ten-word count cutoff ($t(186) = 3.56$, $p < .001$, Cohen's $q = .32$).

Confusion matrices also illustrate the shortcomings of the existing text-based data quality indicators (Table 5). For instance, the 16-character count cutoff only labelled two responses out of 188 as invalid. Both were successful detections (true positives or hits), since these two responses were indeed labelled by humans as invalid – however, the 16-character count cutoff incorrectly labelled the vast majority of human-labelled invalid texts as valid (83%; false negatives or misses). In comparison, the 51-character count cutoff and the ten-word count cutoff labelled more responses as invalid (21 and 22, respectively). Unfortunately, most of these “invalid” responses were human-labelled as valid (71–73%; false positives or false alarms).

⁷ Note that using this same method for constructing a relative cutoff (mean minus 1 standard deviation) can be applied to character count as well as word count. For character count, this relative cutoff would be a 55-character cutoff (in the labelled subset, $M_{\text{characters}} = 165.4$, $SD_{\text{characters}} = 110.5$), which is very similar to the static 51-character cutoff already implemented. We have opted for the 51-character cutoff here since it has already been used and recommended by past researchers (Etz et al., 2018).

Table 6 Example text responses from the unlabelled subset ($n = 2634$) that were predicted as valid or invalid by the best model

Model Predicted Valid	Model Predicted Invalid
“felt very scared and small - felt Powerless - I usually start crying”	“Mostly a traumatizing event but I don't like talking about it.”
“My parents divorce”	“I prefer not to say”
“When I was younger, I climbed up a mountain via a hiking trail by myself during a vacation. I took a photo at a reasonable height up and then was lead back down after running into a couple. My parents were quite worried afterwards, as I did not tell them where I went.”	“I don't remember exactly what the event was but I remember something like this happening.”
“I was at a party, everybody was super drunk. it was dark and there were bright lights. I was dancing with this girl and she smelled great. and I remember us talking.”	“Can't remember off the top of my head as there has been countless of them.”
“Thoughts of someone who used to be very close to me would flash in my mind when certain things or object reminded me of them. This time, an image of said person turning around and smiling at me would appear. The environment/whereabouts of the person was unknown, but mostly it was the act of the person smiling.”	“I don't remember :(”
“overtime that I walk through any hallway, I will feels like the old dog is chasing me. as it happened before, I don't afraid of the dog or the hall way, but the details of that hallway will be much in detail which I can remember.”	“Being anxious, and having to move back from Montreal. The feeling of panic attacks.”

Predicting validity of unseen text responses

To apply the ML model to out-of-sample, unlabelled data, we trained the ML model on the full labelled dataset (including train, validation, and test sets) to increase training data (and theoretically, final model performance; Sebastiani, 2002).

Out of the 2634 unlabelled text responses, the final model predicted 170 as invalid (6.5%) and 2464 as valid (93.5%), which is similar to base rates produced by human coding (7.5% invalid and 92.5% valid out of 949 labelled text responses). The final model detected invalid texts containing a wide range of content, including text responses about declining to respond, text responses about misunderstanding instructions, and text responses that were irrelevant to the question (e.g., describing dreams when the question asked about memories). For a random sample of unlabelled texts predicted as valid or invalid, see Table 6.

Discussion

Best practices dictate that researchers using text as data should detect and remove text responses that are invalid (e.g., meaningless or irrelevant text; Banks et al., 2018; Etz et al., 2018; Müller et al., 2014), since even small proportions of invalid responses can measurably distort results (Arias et al., 2020; Credé, 2010; Huang et al., 2015; Wood et al., 2017; Woods, 2006). However, current literature offers little guidance as to how to detect these invalid text responses effectively or efficiently, and even less evidence as to which methods work well with text data. Here, we implemented

and assessed a supervised ML-based method of detecting invalid text responses and compared it to existing detection methods. Results indicated that our ML approach accurately classified valid and invalid autobiographical memory texts with performance near human coding, significantly outperforming previous non-ML approaches.

First, we found that only *some* of the data quality indicators recommended in the current literature could reliably discriminate valid from invalid responses in our text data. Specifically, only the minimum cutoffs for character count and word count were significantly correlated with human coding. Because labelling very short text responses as invalid approximated human coding to some extent in our dataset, it might be reasonable to flag extremely suspicious cases using character or word count (Etz et al., 2018). At the very least, using a 16-character count cutoff (Etz et al., 2018) did no harm: no valid responses (as labelled by humans) were incorrectly labelled as invalid based on this criterion in our test set. However, one should not expect this 16-character count cutoff to never result in false alarms, since even extremely short text responses (e.g., a few words) can be valid (see Table 6). Confusion matrices also demonstrate the limited effectiveness of using these text-based approaches to discriminate valid from invalid responses. Character and word count cutoffs suffered from very high miss rates (labelling responses as valid when humans labelled them as invalid), very high false alarm rates (labelling responses as invalid when humans labelled them as valid), or both (see Table 5).

Time-based cutoffs also offered little value for detecting invalid text responses in our current study. The maximum cutoffs for speed factor (for either the total survey or single page) were not significantly correlated with human coding.

This is interesting considering that time-based data quality indicators have effectively detected C/IE responding in other studies, leading many researchers to recommend their use (Curran, 2016; Huang et al., 2012; Jones et al., 2021; Maniaci & Rogge, 2014; Niessen et al., 2016). Importantly, some caveats to using response time have been reported: Meade and Craig (2012) noted that while time-based measures were better than nothing, using response time alone was ineffective at detecting invalid responses in their studies. Further, Leiner (2019) suggested that response time may fail to detect invalid responses when questions are fact- or knowledge-based – a potentially relevant distinction to the extent that text responses may rely more heavily on retrieving knowledge from memory compared to closed-ended responses (Holland & Christian, 2009; Kaczmirek et al., 2017; Krosnick, 1999; Scholz & Zuell, 2012). In our case, retrieval was necessary because the question asked participants to recall an autobiographical memory – given that individual differences in autobiographical memory function are well documented (Palombo et al., 2018; Rubin, 2021), responding “too quickly” may represent natural variance in individuals’ access to autobiographical memories more so than C/IE responding.

Critically, our analyses indicated that our ML approach was a significantly better proxy for human coding compared to existing data quality indicators. Though character and word count cutoffs were acceptable to some degree ($r_s = .31-.40$, $p_s < .001$), the ML model performed significantly better in all regards ($p_s < .03$). A closer inspection of the best non-ML data quality indicator in our study (16-character count cutoff) further illustrates its inadequacy for invalid text detection – only two out of a possible 188 texts (1%) were flagged this way. Our own data suggest that human coders detected far more invalid responses (7%) when examining the same set of texts. As such, this character cutoff is likely insufficient to address the problem of detecting invalid text responses, as only the most extreme cases are ever identified as suspicious. In comparison to existing methods, our ML approach was more accurate at both flagging text responses that humans labelled as invalid and retaining text responses that humans labelled as valid. Given its success here, we have made our code openly available (<https://github.com/ryancyeung/invalid-text-detect>) in the hopes of facilitating research involving text data. Our code is fully open source, takes less than one minute to run from start to finish on a modern computer, and can be implemented with any text data (see Appendix for further details).

Though the ML model significantly outperformed existing methods here, many considerations should be taken into account when deciding which method best fits with one’s own data and goals. As with any method of excluding data, researchers should weigh the potential benefits of removing invalid data against the potential costs of accidentally

discarding valid data. In some contexts, it might be critical that meaningless or irrelevant text be removed, lest they alter a study’s conclusions – if this is the case, researchers might find it acceptable if some valid data is misclassified and ultimately excluded prior to analysis. In other contexts, researchers might prefer retaining as much valid data as possible, even if it means that some invalid data is included in analyses. Recent literature suggests that the former case is more realistic, since many studies have shown that even small amounts of invalid responses can cause noticeable harm to data quality (Arias et al., 2020; Credé, 2010; Huang et al., 2015; Wood et al., 2017; Woods, 2006). Our ML approach might be particularly attractive in these situations, since the model shows a balanced ability to both retain valid data and detect invalid data. However, some may find the possibility of removing valid data (false positives/false alarms) too costly. In these circumstances, using a lenient non-ML method might be more appealing (e.g., a 16-character count cutoff). Alternatively, our ML approach can be supplemented with manual adjustments to the unlabelled texts predicted as invalid. If it is paramount to preserve as much data as possible (e.g., with small sample sizes), one should consider manually inspecting the cases predicted as invalid⁸ and adjusting as necessary. Any manual adjustments to the predicted cases should be conducted rigorously (e.g., with multiple coders) and reported transparently.

We also want to acknowledge that our current ML approach requires some up-front commitment of resources, since a labelled or manually coded subset is necessary. Recommendations to date for the size of this labelled subset have ranged between 100, 200, and 500 labelled texts (Figuerola et al., 2012; Grimmer & Stewart, 2013; Hopkins & King, 2010). When choosing how many texts to label, a factor to consider is how many invalid texts are to be expected – if humans label very few texts as invalid in the labelled subset, ML may have difficulty training well on these rare cases, especially after recommended splits into training, validation, and test sets (Chicco, 2017). For example, if an ML model is trained on very few invalid cases, it might fail to learn from enough cases to generalize well to new, unseen text responses. We mitigate the impact of class imbalance (e.g., few invalid texts compared to valid texts) using resampling techniques (Chawla et al., 2002; Estabrooks et al., 2004; He et al., 2008; He & Garcia, 2009; Nguyen et al., 2011), but researchers should still take caution to ensure that instances of the minority class (in our case, invalid texts) exist in the training, validation, and

⁸ We suggest inspecting the predicted invalid cases (rather than the predicted valid cases) because they make up the smaller proportion of cases and are therefore faster to work through. Notably, this can only reduce false positives (false alarms) – one would need to inspect the predicted valid cases to reduce false negatives (misses).

test sets. Devoting resources to labelling a subset of text responses may appear onerous to some; however, we note that researchers have already adopted coding as a method of detecting invalid text responses (Brühlmann et al., 2020; Etz et al., 2018; Kennedy et al., 2020; Scholz & Zuell, 2012; Sischa et al., 2020), which necessitates coding the *entire* dataset manually. Instead, our current ML approach only requires coding a *subset* of the text responses. Our tools here can easily cut the workload into a small fraction of what is typically undertaken.

Future directions

Other researchers have demonstrated that unsupervised ML can be used to detect irrelevant text responses to short-answer questions in an educational context (Gagnon et al., 2019). Specifically, it was found that irrelevant text responses clustered together, since the words commonly used in irrelevant answers tended to be dissimilar to the words commonly used in relevant answers. Future work could investigate unsupervised ML further, as this could eliminate the need to manually label a subset to train, validate, and test upon. However, the assumption that valid texts use similar words might not hold in a context where responses are more varied (e.g., autobiographical memories) compared to educational short-answer questions.

Alternatively, performance and/or selection of the best ML model might be improved by adding steps such as hyperparameter tuning (Chicco, 2017) or feature selection (Aggarwal & Zhai, 2012), which have generally improved text classification in other contexts (Desmet & Hoste, 2018; Takano et al., 2017). Text responses could also be represented in different ways to retain information about word order (e.g., trigram/skip-gram models). For instance, word embeddings can be used to represent words or phrases as vectors (Mikolov et al., 2013), which is thought to capture syntactic and semantic components of language better than the bag-of-words representations used in the current study. However, these more complex representations do not always result in better performance (Grimmer & Stewart, 2013). In fact, the simpler bag-of-words representations we implemented here (count, TF-IDF) have been shown to be well suited for relatively short texts (Joti et al., 2020; Padurariu & Breaban, 2019; Wang et al., 2017), such as those in our dataset ($M_{\text{words}} = 30.9$, $M_{\text{characters}} = 164.1$).

Estimates of ML model performance may also be made more reliable by using nested cross-validation instead of the flat cross-validation used in our present study (Varma & Simon, 2006). That is, instead of conducting cross-validation in a single round, it is possible to cross-validate in multiple, recursive rounds (i.e., one cross-validation nested within another cross-validation), which has the potential to reduce error in the performance estimates used for model selection.

Importantly, it remains worth noting that in practical settings such as our own, nested cross-validation often results in similar quality model selection to that of flat cross-validation (Wainer & Cawley, 2021). Future extensions of this work could explore whether nested cross-validation might impact the model selection process.

Conclusions

Researchers have recommended that those working with text as data ought to identify invalid text responses prior to analysis, in order to ensure that data are meaningful and analytically useful. Despite the importance of detecting invalid data, options in past literature are of limited applicability to text responses. While some options are relatively fast to compute (e.g., response time, character/word count), their effectiveness as detection methods has not been well-demonstrated. Conversely, more direct measures of text validity (e.g., coding) are not always practical due to the time and labour required. In the current work, we demonstrated that a supervised ML approach can approximate the accuracy of human coding, but without the need to manually label full datasets. Our ML approach also significantly outperformed existing data quality indicators when classifying text responses as valid or invalid. In the hopes of facilitating research involving text as data, we present our openly available code as a new option for detecting invalid responses that is both effective and practical.

Appendix

Notes on Workflow and Implementation

Folder structure

After downloading our Python code, users can put the code file into any folder (e.g., “invalid_text_detect_project” in Appendix Fig. 1). Two subfolders should then be created: (1) “input” and (2) “output”. We recommend forking and cloning the project’s GitHub repository (<https://github.com/ryanceyung/invalid-text-detect>), which will store a correctly formatted, up-to-date copy of the project on your machine locally.

Input data

Input data should be formatted into a single comma-separated values (CSV) file and placed into the input folder. Three columns are necessary in this file: (1) unique identifiers for each text response, (2) the original text responses, and (3) manually coded labels for the labelled subset (0 =

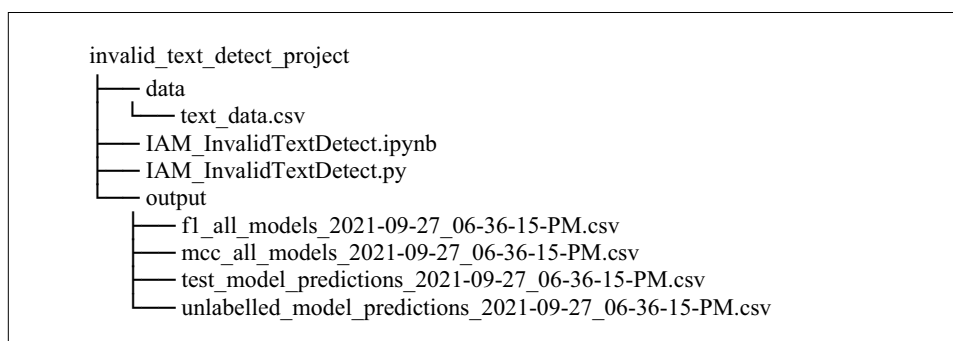


Fig. 1 Example folder structure

Table 7 Example structure of the input data file

doc_id	text	human_labelled
1	The quick brown fox jumps over the lazy dog.	0
2	Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.	1
3	Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.	
4	The five boxing wizards jump quickly.	

valid, 1 = *invalid*, blank = unlabelled). Headings for these columns (i.e., the first row) should be (1) “doc_id”, (2) “text”, and (3) “human_labelled”, respectively. As long as the three required columns are present (unique identifier, original text, manually coded labels), any number of additional columns can be included; extra columns will have no impact on running the code. As an example, see Appendix Table 7. In this mock dataset, text responses 1 and 2 were manually labelled as valid or invalid, forming the labelled subset. Text responses 3 and 4 were not manually coded, forming the unlabelled subset.

Running the Python code

Our Python code is accessible as either a Python script (.py) or a Jupyter notebook file (.ipynb), both of which can be run locally or using services like Google Colab (with Python v3+). Based on one’s results from the model validation section of the code, the user should edit the model evaluation section appropriately. In other words, users should inspect the performance metrics from the model validation section (e.g., MCC, macro F1 score, precision, recall, accuracy) to decide which model performed best (e.g., highest mean MCC; Chicco & Jurman, 2020; Luque et al., 2019), though “best” always depends on the user’s specific context. Once this best model is selected, users can edit the code to supply the name of the best model as an argument for the model evaluation and final predictions sections.

Output data

Output data are created in the output folder as the Python code runs. No files are necessary in the output folder to run the code; only the input data file in the input folder (in Appendix Fig. 1, “text_data.csv”) is needed to run the code. Our Python code is written to include the date and time that each output file was created in the output file names. In addition to helping with organization and version control, this helps prevent any accidental overwriting of output files⁹.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-022-01801-y>.

Code availability The code written for and/or used during the current study is available on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/3HS56>) and GitHub (<https://github.com/ryancyeung/invalid-text-detect>).

Authors' contributions Ryan Yeung: Conceptualization, data curation, formal analysis, funding acquisition, methodology, software, validation, writing – original draft, writing – review & editing. Myra Fernandes: Funding acquisition, methodology, resources, supervision, writing – review & editing.

⁹ An exception to this safeguard is if a user runs a code chunk that writes an output file multiple times in one second.

Funding This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through an Alexander Graham Bell Canada Graduate Scholarship (CGSD3-535024-2019) awarded to author Ryan C. Yeung, and an NSERC Discovery Grant (2020-03917) awarded to author Myra A. Fernandes.

Data availability The datasets generated and/or analysed during the current study are available on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/3H556>) and GitHub (<https://github.com/ryancyeung/invalid-text-detect>). This study was not preregistered.

Declarations

Conflicts of interest The authors have no potential conflicts of interest or competing interests to disclose.

Ethics approval All procedures were approved by the Office of Research Ethics at the University of Waterloo (Protocol #40049).

Consent to participate Informed consent to participate was obtained from all participants included in the study.

Consent for publication Informed consent for publication was obtained from all participants included in the study.

References

- Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163–222). Springer. https://doi.org/10.1007/978-1-4614-3223-4_6
- Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, 52(6), 2489–2505. <https://doi.org/10.3758/s13428-020-01401-8>
- Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, 33(4), 445–459. <https://doi.org/10.1007/s10869-017-9528-3>
- Behr, D., Braun, M., Kaczmarek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity*, 48(1), 127–148. <https://doi.org/10.1007/s11135-012-9754-8>
- Berntsen, D. (1996). Involuntary autobiographical memories. *Applied Cognitive Psychology*, 10(5), 435–454.
- Berntsen, D., & Rubin, D. C. (2008). The reappearance hypothesis revisited: Recurrent involuntary memories after traumatic events and in everyday life. *Memory & Cognition*, 36(2), 449–460. <https://doi.org/10.3758/mc.36.2.449>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brewin, C. R., Christodoulides, J., & Hutchinson, G. (1996). Intrusive thoughts and intrusive memories in a nonclinical sample. *Cognition & Emotion*, 10(1), 107–112. <https://doi.org/10.1080/026993996380411>
- Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2, 1–13. <https://doi.org/10.1016/j.metip.2020.100022>
- Bywaters, M., Andrade, J., & Turpin, G. (2004). Intrusive and non-intrusive memories in a non-clinical sample: The effects of mood and affect on imagery vividness. *Memory*, 12(4), 467–478. <https://doi.org/10.1080/09658210444000089>
- Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In O. Maimon, L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed., pp. 875–886). Springer. https://doi.org/10.1007/978-0-387-09823-4_45
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, P.-S. D. (2011). Finding quality responses: The problem of low-quality survey responses and its impact on accountability measures. *Research in Higher Education*, 52(7), 659–674. <https://doi.org/10.1007/s11162-011-9217-4>
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1), 1–17. <https://doi.org/10.1186/s13040-017-0155-3>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 1–13. <https://doi.org/10.1186/s12864-019-6413-7>
- Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70(4), 596–612. <https://doi.org/10.1177/0013164410366686>
- Cunningham, M., & Wells, M. (2017). Qualitative analysis of 6961 free-text comments from the first National Cancer Patient Experience Survey in Scotland. *BMJ Open*, 7(6), 1–16. <https://doi.org/10.1136/bmjopen-2016-015726>
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- Desmet, B., & Hoste, V. (2018). Online suicide prevention through optimised text classification. *Information Sciences*, 439–440, 61–78. <https://doi.org/10.1016/j.ins.2018.02.014>
- Dunn, O. J., & Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association*, 66(336), 904–908. <https://doi.org/10.1080/01621459.1971.10482369>
- Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1), 18–36. <https://doi.org/10.1111/j.0824-7935.2004.101-1-00228.x>
- Etz, R. S., Gonzalez, M. M., Eden, A. R., & Winship, J. (2018). Rapid sense making: A feasible, efficient approach for analyzing large data sets of open-ended comments. *International Journal of Qualitative Methods*, 17(1), 1–10. <https://doi.org/10.1177/1609406918765509>
- Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, 1871–1874.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer. <https://doi.org/10.1007/978-3-319-98074-4>
- Figuerola, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1), 1–10. <https://doi.org/10.1186/1472-6947-12-8>
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons. <https://doi.org/10.1002/0471445428>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

- Gagnon, V., Labrie, A., Bhatnagar, S., & Desmarais, M. C. (2019). Filtering non-relevant short answers in peer learning applications. *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 556–559).
- Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350), 320–328.
- Gogami, M., Matsuda, Y., Arakawa, Y., & Yasumoto, K. (2021). Detection of careless responses in online surveys using answering behavior on smartphone. *IEEE Access*, 9, 53205–53218. <https://doi.org/10.1109/access.2021.3069049>
- Graham, Y., & Baldwin, T. (2014). Testing for significance of increased correlation with human judgment. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 172–176). <https://doi.org/10.3115/v1/d14-1020>
- 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), 267–297. <https://doi.org/10.1093/pan/mps028>
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley-Blackwell.
- Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1–14. <https://doi.org/10.1214/088342306000000060>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiene, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/tkde.2008.239>
- He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 1322–1328). <https://doi.org/10.1109/ijcnn.2008.4633969>
- Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review*, 27(2), 196–212. <https://doi.org/10.1177/0894439308327481>
- Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1), 229–247. <https://doi.org/10.1111/j.1540-5907.2009.00428.x>
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828–845. <https://doi.org/10.1037/a0038510>
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, 39(1), 103–129. <https://doi.org/10.1016/j.jrp.2004.09.009>
- Jones, A., Caes, L., Rugg, T., Noel, M., Bateman, S., & Jordan, A. (2021). Challenging issues of integrity and identity of participants in non-synchronous online qualitative methods. *Methods in Psychology*, 5, 1–5. <https://doi.org/10.1016/j.metip.2021.100072>
- Joti, D., Young, J. C., & Rusli, A. (2020). A comparison of supervised text classification and resampling techniques for user feedback in Bahasa Indonesia. *Proceedings of the 2020 Fifth International Conference on Informatics and Computing (ICIC)* (pp. 1–6). IEEE. <https://doi.org/10.1109/icic50835.2020.9288588>
- Kaczmarek, L., Meitinger, K., & Behr, D. (2017). *Higher data quality in web probing with EvalAnswer: A tool for identifying and reducing nonresponse in open-ended questions*. *GESIS – Leibniz Institute for the Social Sciences*. Retrieved September 4, 2021, from <https://web.archive.org/web/20210904214259/https://www.ssoar.info/ssoar/handle/document/51100>
- Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2020). *Assessing the risks to online polls from bogus respondents*. Pew Research Center. Retrieved August 19, 2021, from <https://web.archive.org/web/20210819101207/https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polls-from-bogus-respondents/>
- Kobayashi, V. B., Mol, S. T., Berkens, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods*, 21(3), 733–765. <https://doi.org/10.1177/1094428117722619>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50(1), 537–567. <https://doi.org/10.1146/annurev.psych.50.1.537>
- Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, 13(3), 229–248. <https://doi.org/10.18148/srm/2019.v13i3.7403>
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(1), 559–563.
- Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. *Nature Methods*, 13(9), 703–704. <https://doi.org/10.1038/nmeth.3968>
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
- Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, 32(1), 1–10. <https://doi.org/10.5334/irsp.289>
- Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 415–463). Springer. https://doi.org/10.1007/978-1-4614-3223-4_13
- Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216–231. <https://doi.org/10.1016/j.patcog.2019.02.023>
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83. <https://doi.org/10.1016/j.jrp.2013.09.008>
- Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, 36(3), 2009–2031. <https://doi.org/10.1007/s00180-020-00999-9>
- Marks, E. H., Franklin, A. R., & Zoellner, L. A. (2018). Can't get it out of my mind: A systematic review of predictors of intrusive memories of distressing events. *Psychological Bulletin*, 144(6), 584–640. <https://doi.org/10.1037/bul0000132>

- McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In: *AAAI-98 Workshop on Learning for Text Categorization* (Vol. 752, No. 1, pp. 41–48).
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. <https://doi.org/10.1037/a0028085>
- Menardi, G., & Torelli, N. (2012). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1), 92–122. <https://doi.org/10.1007/s10618-012-0295-5>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of Advances in Neural Information Processing Systems* (Vol. 26, pp. 3111–3119).
- Müller, H., Sedley, A., & Ferrall-Nunge, E. (2014). Survey research in HCI. In J. Olson, W. Kellogg (Eds.), *Ways of knowing in HCI* (pp. 229–266). Springer. https://doi.org/10.1007/978-1-4939-0378-8_10
- Neill, J. J., & Dunn, O. J. (1975). Equality of dependent correlation coefficients. *Biometrics*, 31(2), 531–543. <https://doi.org/10.2307/2529435>
- Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2018). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1), 202–237. <https://doi.org/10.1177/0049124118769114>
- Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In: *Advances in Neural Information Processing Systems* (pp. 841–848).
- Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1), 4–21. <https://doi.org/10.1504/ijkesdp.2011.039875>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63, 1–11. <https://doi.org/10.1016/j.jrp.2016.04.010>
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. <https://doi.org/10.1145/2872427.2883062>
- Opitz, J., & Burst, S. (2019). *Macro F1 and macro F1*. arXiv. <https://arxiv.org/abs/1911.03347v3>
- Padurariu, C., & Breaban, M. E. (2019). Dealing with data imbalance in text classification. *Procedia Computer Science*, 159, 736–745. <https://doi.org/10.1016/j.procs.2019.09.229>
- Palombo, D. J., Sheldon, S., & Levine, B. (2018). Individual differences in autobiographical memory. *Trends in Cognitive Sciences*, 22(7), 583–597. <https://doi.org/10.1016/j.tics.2018.04.007>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rubin, D. C. (2021). Properties of autobiographical memories are reliable and stable individual differences. *Cognition*, 210, 1–13. <https://doi.org/10.1016/j.cognition.2021.104583>
- Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660–674.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. *International Journal of Market Research*, 62(5), 578–598. <https://doi.org/10.1177/1470785320921779>
- Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Social Science Research*, 41(6), 1415–1428. <https://doi.org/10.1016/j.ssresearch.2012.07.006>
- Schroeders, U., Schmidt, C., & Gnambs, T. (2021). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 1–28. <https://doi.org/10.1177/00131644211004708>
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference* (Vol. 57, p. 61).
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47. <https://doi.org/10.1145/505282.505283>
- Sischka, P. E., Décieux, J. P., Mergener, A., Neufang, K. M., & Schmidt, A. F. (2020). The impact of forced answering and reactance on answering behavior in online surveys. *Social Science Computer Review*, 1–21. <https://doi.org/10.1177/0894439320907067>
- Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, 73(2), 325–337. <https://doi.org/10.1093/poq/nfp029>
- Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., 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), 550–572. <https://doi.org/10.1080/10584609.2020.1723752>
- Takano, K., Ueno, M., Moriya, J., Mori, M., Nishiguchi, Y., & Raes, F. (2017). Unraveling the linguistic nature of specific autobiographical memories using a computerized classification algorithm. *Behavior Research Methods*, 49(3), 835–852. <https://doi.org/10.3758/s13428-016-0753-x>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1). <https://doi.org/10.1186/1471-2105-7-91>
- Viani, N., Botelle, R., Kerwin, J., Yin, L., Patel, R., Stewart, R., & Velupillai, S. (2021). A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Scientific Reports*, 11(1), 1–12. <https://doi.org/10.1038/s41598-020-80457-0>
- Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182, 115222. <https://doi.org/10.1016/j.eswa.2021.115222>
- Wang, Y., Zhou, Z., Jin, S., Liu, D., & Lu, M. (2017). Comparisons and selections of features and classifiers for short text classification. In: *IOP Conference Series: Materials Science and Engineering* (Vol. 261, No. 1, p. 012018). IOP Publishing. <https://doi.org/10.1088/1757-899X/261/1/012018>
- Waskom, M. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
- Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, 11(4), 245–265. <https://doi.org/10.1080/19312458.2017.1387238>

- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2
- Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, 8(4), 454–464. <https://doi.org/10.1177/1948550617703168>
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186–191. <https://doi.org/10.1007/s10862-005-9004-7>
- Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1, 69–90. <https://doi.org/10.1023/A:1009982220290>
- Yeung, R. C., & Fernandes, M. A. (2020). Recurrent involuntary autobiographical memories: Characteristics and links to mental health status. *Memory*, 28(6), 753–765. <https://doi.org/10.1080/09658211.2020.177731>
- Yeung, R. C., & Fernandes, M. A. (2021). Recurrent involuntary memories are modulated by age and linked to mental health. *Psychology and Aging*, 36(7), 883–890. <https://doi.org/10.1037/pag0000630>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

Open Practices Statement All data and materials for this study are available on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/3HS56>) and GitHub (<https://github.com/ryancyeung/invalid-text-detect>). This study was not preregistered.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.