The rapid advance of information technology over the past decade has led to breakthroughs in quality, continuity, and efficiency of e-health and e-psychology care (Drigas, Koukianakis, & Papagerasimou, 2011). One promising application is the use of natural language processing (NLP) and text mining techniques to identify clinical information contained in unstructured free text documents and to codify this information into structured data (Trusko et al., 2010). For instance, Pakhomov, Chacon, Wicklund, and Gundel (2011) extracted clear patterns of decline in grammatical complexity in language production affected by neurodegenerative disorders. Day, Christensen, Dalto, and Haug (2007) used an NLP system to classify trauma patients based on their clinical histories. Ando, Morita, and O’Connor (2007) identified cancer patients by using transcripts from structured interviews and employing a text-mining approach, finding considerable differences in how they look back at their life by age and gender.

Individuals’ speech and writing patterns can provide a window into their emotional and cognitive worlds (Tausczik & Pennebaker, 2010). Over the years, much evidence has suggested that the words and expressions in patients’ self-narratives are very informative for early detection of mental diseases (e.g., Franklin & Thompson, 2005; Gottschalk & Gleser, 1969; Rosenberg & Tucker, 1979; Smyth, 1998). In such cases, language becomes an important medium by which clinical psychologists attempt to understand the patients. With the increasing application of online measures, the automated identification of patients based on their self-narratives is, therefore, expected to be a promising step toward an effective e-health system, especially for the screening and diagnosis process (e.g., Alvarez-Conrad, Zoellner, & Foa, 2001; Coppersmith, Harman, & Dredze, 2014; Howes, Purver, & McCabe, 2014).

However, despite the great potential in mining information from self-narratives, it is not commonly used in clinical practice and has not been sufficiently described yet in the literature largely because of two sets of challenges. First, due to different backgrounds such as educational level, social status, living conditions, and so on, people often use various words to express the same concept. The openness and diversity of words may cause difficulties in mapping synonyms into a standardized reference terminology and
extracting robust information that represents an identical domain (Trusko et al., 2010). Second, unlike the numeric data collected from questionnaires, textual data are often unstructured, neither having a predefined data model nor fitting well into relational patterns. The result is irregularities and ambiguities that make direct analysis through traditional quantitative methods difficult.

The purpose of the present study was to develop an automated system that would screen patients for posttraumatic stress disorder (PTSD) through NLP and text-mining techniques. The process entailed asking trauma victims to write about their traumatic events and symptoms online rather than conducting face-to-face interviews with item-based questionnaires. Based on their textual input, the respondents could be classified into PTSD (i.e., high risk to develop PTSD) and non-PTSD (i.e., low risk to develop PTSD) groups. Those identified as PTSD at this initial stage were to be invited to a more extensive test for further and more precise diagnosis. Therefore, the textual screening procedure would be helpful if it could maximize the accuracy of finding potential PTSD patients or reliably exclude non-PTSD individuals from the follow-up tests. This approach promises to reduce the amount of time and expense associated with the diagnosis of PTSD and to help identify patients with PTSD at an earlier stage. Early identification of PTSD is critical to timely initiation of treatment.

Given previous efforts in development of a keyword-based textual assessment method (He & Veldkamp, 2012; He, Veldkamp, & de Vries, 2012), the present study sought to apply the text-mining techniques with higher order n-grams (i.e., keywords and expressions with multiple word components) in PTSD screening and evaluate their efficiency in conjunction with different text classification models. Two specific objectives were addressed here: first, to overview the procedure of automated textual assessment on patients’ self-narratives for PTSD screening and second, to compare the performances of different classification models in conjunction with n-gram representations in the screening process.

**Method**

**Participants**

A total of 300 trauma survivors were selected for the present study. Of the 300 participants, 150 were diagnosed as PTSD patients and the other half as non-PTSD. (We set the target line of data collection as $n=150$ per group.) All participants had been diagnosed as PTSD or non-PTSD by at least two clinical practitioners via structured interviews with standardized instruments including the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders–Fourth edition PTSD module (First, 1997) and the Clinician Administered PTSD Scale (Blake et al., 1995). When the two practitioners differed about a patient’s diagnosis, a third practitioner was involved. The diagnoses from the practitioners were used as true standards and compared with the results derived from computerized textual classifications in this study. The age of participants ranged from 19 to 63 years, with a mean of 30.06 years ($SD=11.3$). Of the 300 participants, 195 were female (65%), 63 were male (21%), and 42 did not report gender (14%). All participants reported having had at least 3 months’ experience using the Internet and did not encounter problems using the online survey system.

**Instrumentation**

The data were collected via an online survey that was embedded in an open forum dedicated for those seeking aid for mental health issues. All participants were asked to register before logging into the survey. The survey consisted of two parts: textual writing in response to an open-ended question and a demographic questionnaire. Two specific requirements were set for the self-narrative: (a) it must be written by the participant himself or herself rather than friends or family members and (b) it should describe the traumatic events and their impact on the participant’s daily life. The open-ended question used in the current study and an example of a response were presented as follows:

**Question:** What are the events that caused you most problems? What are their major impacts to your daily life? Would you please share your story?

**Answer:** I was only 15 when I was attacked by a group of men on the way home from school. They took turns screaming abuse at me and then they each raped me. Finally, they tried to stab me to death and would almost certainly have succeeded had the police not arrived on the scene. For months after this horrifying event, I was not myself. I was unable to keep the memories of the attack out of my mind. At night I would have terrible dreams of rape, and would wake up screaming. I had difficulty walking back from school because the route took me past the site of the attack, so I would have to go the long way home. I felt as though my emotions were numbed, and as though I had no real future. At home I was anxious, tense, and easily startled. I felt “dirty” and somehow shamed by the event, and I resolved not to tell close friends about the event, in case they too rejected me.

We consulted with experienced clinical practitioners on the textual data to ensure all three parts of the open-ended question were addressed in the self-narratives. The average
length (i.e., word counts) of the 300 self-narratives collected in the current study was 257 words ($SD=236$). In the PTSD corpus, the length of self-narratives ranged from 46 to 1,968 words, with a mean of 284 ($SD=291$). Comparatively, slightly shorter stories were generally found in the non-PTSD group, ranging from 51 to 973 words, with a mean of 229 ($SD=160$). The context of stressful events written by the 300 participants covered eight types: child abuse, sexual abuse, traffic accident, war, domestic violence, death of a loved one, robbery, and fire.

**N-Grams**

The textual data are usually encoded via a data representation model. More specifically, each document is generally represented as a vector of (possibly weighted) word counts (Manning & Schütze, 1999). The simplest and most commonly used data representation model is the “bag of words” (BOW), where each word in a document collection acts as a distinct feature. As an extension of BOW, $n$-gram—which considers the interaction effect of two, three, or more consecutive words—is proposed as a way to expand the standard unigram representation model (Jurafsky & Martin, 2009). For instance, in the sentence “I cry because I am frightened,” there are seven unigrams (e.g., “I,” “cry”), six bigrams (e.g., “I cry,” “cry because”), and five trigrams (e.g., “I cry because,” “cry because I”). $N$-grams are also used because they may help reduce the problems presented by polysemous words (e.g., “look like” vs. “like swimming”), identify concepts highly characteristic of a text domain (e.g., “World Health Organization” is more meaningful than its separated elements “world,” “health,” and “organization” in a context of public health), and interpret meaning of extracted features (e.g., the word “get” is hard to interpret by its single presence but makes more sense when it co-occurs with an adjective in a phrase such as “get depressed”).

**Procedures**

The textual screening procedure generally consists of three phases—preparation, training, and testing—as shown in Figure 1. During the preparation phase, the textual data are divided into a training set and a test set and preprocessed according to several linguistic rules. In this study, the training set consisted of 200 self-narratives—100 randomly selected from PTSD corpus and 100 from non-PTSD corpus—while the remaining 100 narratives, with 50 from either corpus, were used as the test set. The diagnoses made by practitioners for each participant were set as the “standard labels” (i.e., PTSD or non-PTSD) for the inputs. The training and testing phases were the essential parts to classify the self-narratives into PTSD and non-PTSD groups, where the text-mining techniques were primarily applied. During training, the most discriminative features (e.g., keywords or key vectors) for determining the class labels were extracted. The input for the machine-learning algorithm consisted of a set of prespecified features that may potentially be present in a document and labels that classify each document.

The goal of the training phase was to “learn” the relationship between the features and the class labels. The testing phase played an important role in checking how well the trained classifier model performed on a new data set. The

![Figure 1. The overview of text classification procedure for PTSD screening.](image)

*Note. PTSD = posttraumatic stress disorder.*
test set consisted of data that were not used during training. In the testing procedure, when a new set of inputs was presented to the system, the system first checked whether the extracted features existed and then followed the machine-learning algorithms to predict the label (i.e., PTSD or non-PTSD) for each text based on the “training” it had received.

**Preprocessing**

To improve the efficiency of training and testing procedures as well as to increase the ability to generalize to previously unseen data, a preprocessing routine was implemented. This involved screening digital numbers, deducting noninformative “stop-words” (e.g., “I,” “to”), common punctuation marks (e.g., “,” and “.”), and frequently used abbreviations (e.g., “isn’t,” “I’m”) and “stemming” the rest of the words, using the Porter (1980) algorithm, to remove common morphological endings. For example, the terms “nightmarest,” “nightmaring,” and “nightmare,” though in variant lexical forms, were normalized in an identical stem “nightmar” by removing the suffixes and linguistic rule-based indicators. An n-gram was deducted when all the components were included in the stop-word list. For instance, the bigram “I am” had to be removed because both “I” and “am” were on the stop-word list. Afterward, each component of the n-gram was stemmed with the Porter algorithm.

Discussions about the effect of stop-words such as pronouns (e.g., “I,” “he”) in textual analysis, especially related to psychological assessments, are far from conclusive. Campbell and Pennebaker (2003) argued that the flexibility of common words—particularly personal pronouns—when writing about traumatic memories was related to positive health outcomes. The exclusion of the “junk” words that people used in writing or speech might result in losing a tremendous amount of information about how people were thinking. On the contrary, Luther et al. (2011) applied an iterative term refinement strategy, which used a standard stop-word list followed by clinical review to eliminate non-clinical terms and those not related to PTSD. This approach yielded a slight improvement in screening accuracy compared with no stop-word removal. Similar results were found in the study of Torri et al. (2011). To get more accurate measurements in the current study, we conducted the textual analysis with both the inclusion and exclusion of stop-words. A slight increase (2.3%) in overall accuracy in binary classification of PTSD and non-PTSD was found when using the stop-words. Hence, we decided to continue the textual analysis with stop-words in the preprocessing. All the results derived from this point on are from the analysis with the stop-word list applied.

**Feature Selection**

In text categorization, feature selection is a strategy that aims at identifying the key features that contribute to accurate and efficient classification (Manning & Schütze, 1999). A number of feature selection methods have been widely used in NLP and text classification such as document frequency, information gain, mutual information, chi-square test, binormal separation, and weighted log likelihood ratio (see more in Forman, 2003; Li, Xia, Zong, & Huang, 2009; Nigam, McCallum, Thrun, & Mitchell, 2000; Vapnik, 1998; Yang & Pederson, 1997).

The motivation for choosing chi-square selection algorithm in the present study was to benefit from its effectiveness in finding robust keywords and testing for the similarity between different corpora (Manning & Schütze, 1999). In the current study, the input texts were represented by five data representation models: unigrams, bigrams, trigrams, a combination of unigrams and bigrams, and a mixture of unigrams, bigrams, and trigrams. To apply the chi-square algorithm, each word was compiled into its own 2-by-2 contingency table as shown in Table 1. The number of word occurrences in two corpora C1 (i.e., PTSD corpus) and C2 (i.e., non-PTSD corpus) is indicated by ni and m, respectively. The sum of the word occurrences in each corpus is defined as the corpus length, len(C). The idea of this method is to compare the two corpora and determine how far C1 departs from C2. Under the null hypothesis, the two corpora are similar, so their distribution of words is proportional to each other. A chi-square is computed to evaluate the departure from this null hypothesis. The table is defined as follows: The values in each cell are called the observed frequencies (Oij).

Under the null hypothesis, that is, using the assumption of independence, the expected frequencies (Eij) are computed from the marginal probabilities, that is, from the totals of the columns and rows converted into proportions, using formula $E_{ij} = \frac{(column_i \times row_j \times total)}{grandtotal}$. The chi-square statistic sums the differences between the observed and the expected values in all squares of the table, scaled by the magnitude of the expected values, $X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. To ensure the reliability of the calculation, words that occur fewer than five times are usually eliminated (Manning & Schütze, 1999; Oakes, Gaizauskas, Fowkes, Jonsson, & Beaulieu, 2001). Based on the chi-square scores, all word vectors are ranked in a descending order, and those located at the top are extracted as robust classifiers (named as keywords or key vectors). Furthermore, if the ratio ni / m is larger than the ratio len(C1) / len(C2), the word is regarded as more typical of

<table>
<thead>
<tr>
<th>Word i</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>¬ Word i</td>
<td>n</td>
<td>m</td>
</tr>
<tr>
<td>len(C1) - ni</td>
<td>len(C2) - m</td>
<td></td>
</tr>
</tbody>
</table>

Note. C represents the class label of text corpus, and n and m represent the number of occurrences of a word i in two corpora, respectively.

---

**Table 1. Confusion Matrix for Word i in the 2-by-2 Chi-Square Score Calculation.**

---

Downloaded from asmin.sagepub.com at Universiteit Twente on October 2, 2015
corpus $C_1$ (as a “positive indicator”); otherwise, it is more typical of corpus $C_2$ (as a “negative indicator”; for more details, refer to Oakes et al., 2001).

**Machine-Learning Text Classifiers**

Training text classifiers is the procedure where machines “learn” to automatically recognize complex patterns, distinguish between exemplars based on their different patterns, and make intelligent predictions on their class. Each narrative in the training corpus is converted into a feature vector (e.g., weighted frequencies of preselected words in the narrative), and then a machine-learning algorithm can be applied to build a classification model (Jurafsky & Martin, 2009; Manning & Schütze, 1999). Four machine-learning algorithms were deployed in the present study, including three widely used models—decision tree (DT; Quinlan, 1993), naive Bayes (NB), and support vector machine (SVM; Cortes & Vapnik, 1995)—as well as an alternative algorithm named product score model (PSM; He & Veldkamp, 2012).

The motivations for choosing these four algorithms varied according to their attributes. DTs are simple to understand and easy to interpret, which could facilitate practitioners visualizing the path of decision making through the textual analysis (Conway, Doan, Kawazoe, & Collier, 2009). NB classifiers that hold the assumption of independence among words are simple but effective in practice (e.g., Domingos & Pazzani, 1997; Hand & Yu, 2001). SVM is a powerful machine-learning paradigm that has been often reported to outperform other machine-learning classifiers (e.g., Conway et al., 2009; Jurafsky & Martin, 2009), making it worthwhile to include this state-of-the-art algorithm as a kind of baseline in the model comparison. PSM, as an attempt to address the smoothing issue in NB, showed promise in screening PTSD in a previous study (He et al., 2012), so we sought to better evaluate its performance with standard algorithms. In the following subsections, we review all of these four machine-learning algorithms and then compare their performances in classification of PTSD and non-PTSD.

**Decision Tree.** The DT is a well-known machine-learning approach to automatically induce classification trees based on training data sets. In a tree structure, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The feature that best divides the training data is the root node of the tree. The objects at each node are split into piles in a way that gives maximum information and stops when they are categorized into a terminal class. The rule that we used in this study was to split the objects at a node into two piles in the way that gave us maximum information gain. Information gain is as an information-theoretic measure defined as the difference of the entropy of the mother node and the weighted sum of the entropies of the child nodes, that is,

$$G(a, y) = H(t) - H(t | a) = H(t) - (p_H(t_1) + p_R(t_r)) \quad (1),$$

where $a$ is the attribute we split on, $y$ is the value of $a$, $t$ is the distribution of node that we split, $p_l$ and $p_r$ are the proportion of elements that are passed on to the left and right nodes, and $t_l$ and $t_r$ are the distributions of the left and right nodes.

**Naive Bayes.** The NB is a probabilistic classifier applying Bayes’s theorem with strong independence assumptions (Lewis, 1998). The basic idea is to estimate the conditional probability of the class $C$ given the word vectors $w$ with the assumption of word independence. Namely,

$$p(C | w) = \frac{\prod_{i=1}^{k} p(w_i | C)}{p(w)} \quad (2),$$

where $p(C)$ is the prior probability of a certain class in the whole corpus collection, and $p(w_i | C)$ is the conditional probability of a word occurring in a certain class, which is generally estimated with maximum likelihood. In binary classification, the two probabilities of categories $C_1$ and $C_2$ are compared in a ratio $R$ defined as

$$R = \frac{p(C_1 | w)}{p(C_2 | w)} = \frac{\prod_{i=1}^{k} p(w_i | C_1)}{\prod_{i=1}^{k} p(w_i | C_2)} \quad (3).$$

If $R > 1$, the object is classified in category $C_1$; otherwise, it is classified in category $C_2$.

**Support Vector Machine.** SVM classifiers can exploit a large number of features while avoiding overfitting to the training data (Jurafsky & Martin, 2009). Given two sets of instances belonging to two classes, SVM seeks a hyperplane that maximizes the margin between the two sets of instances. When instances are not linearly separable or a large margin is attainable by overlooking (misclassifying) some instances, the soft margin method can be used to allow misclassification at a defined cost for each misclassified instance. A nonlinear SVM classifier could be built, but past studies suggest that a linear SVM classifier is usually sufficient for text data (Yang & Liu, 1999). In the current study, we adopted the linear SVM model to select two hyperplanes in a way that they discriminately separate the data and then try to maximize their distance. Suppose the sample is $D = \{x', r'\}$, where $r' = +1$ if $x' \in C_1$ and $r' = -1$ if $x' \in C_2$. We would like to find $w$ and $w_0$ in the hyperplanes such that
\[ (w^T x' + w_0) \geq +1 \text{ for } r' = +1 \]
\[ (w^T x' + w_0) \leq -1 \text{ for } r' = -1 \]

which can be rewritten as \( r'(w^T x' + w_0) \geq 1 \). The distance from the hyperplane to the instances closest to it on either side is called the margin, which we would like to maximize for the best generalization (see more about SVM in Alpaydin, 2004).

**Product Score Model.** The PSM is an alternative machine-learning algorithm to address the smoothing issue of NB using a form of Laplace’s law (1995). This model was validated in a previous study (He et al., 2012). Holding the similar independence assumption as the NB model, the PSM features assigning two weights for each keyword (in binary classification), \( U_i \) and \( V_i \), to indicate how much of a degree the word can represent the two classes PTSD and non-PTSD, that is, \( U_i = (n_i + a) / \text{len}(C_i) \) and \( V_i = (m_i + a) / \text{len}(C_i) \). In this study, we used the smoothing constant \( a = 0.5 \), which was added to the word frequency to account for words that did not occur in the training set but might occur in new texts (for more smoothing rules, refer to Jurafsky and Martin, 2009; Manning and Schütze, 1999). An important additional step involved when a negative meaning was detected in a sentence, such as “I do not have flashbacks”; the term weights for PTSD and non-PTSD, that is, the values of \( U_i \) and \( V_i \) were then switched. The name product score comes from a product operation to compute scores for each class, \( S_1 \) and \( S_2 \), for each input text based on the term weights. That is,

\[
\begin{align*}
S_1 &= P(C_1) \prod_{i=1}^{k} U_i = P(C_1) \prod_{i=1}^{k} [(n_i + a) / \text{len}(C_i)] \\
S_2 &= P(C_2) \prod_{i=1}^{k} V_i = P(C_2) \prod_{i=1}^{k} [(m_i + a) / \text{len}(C_i)]
\end{align*}
\]

the classification rule is defined as:

\[
\text{choose } \begin{cases} 
C = 1 \text{ if } \log(S_1 / S_2) > b \\
C = 2 \text{ else }
\end{cases}
\]

where \( b \) is a constant. In this study, we set \( b = 0 \), because in the earlier study (He et al., 2012), it was found that during the PTSD textual screening procedure, the largest number of positive cases could be captured without unduly sacrificing specificity when the threshold was set at zero.

To avoid mismatches caused by randomness, especially when a small number of keywords were used to accomplish the classification task, unclassification rules needed to be considered. We defined a text as “unclassified” when any condition among the following was met: (a) no keywords were found in the text; (b) only one keyword was found in the text; (c) only two keywords were found in the text, with one labeled as a positive indicator (i.e., PTSD) and the other as a negative indicator (i.e., non-PTSD).

**Analytical Strategy**

In evaluating machine-learning classifiers, cross-validation tests are commonly used in which a class-labeled data set is partitioned and classifiers are trained and evaluated on different partitions in a round-robin manner. To ensure the proper generalization capabilities for the text classification model, a 10-fold cross-validation procedure was applied in the current study. As defined by Jurafsky and Martin (2009), in this study, the 10-fold cross-validation was generally conducted as follows: The original sample was randomly partitioned into 10 subsamples of equal size. One subsample served as the validation data for testing the model each time through 10 iterations, with the remaining nine serving as training data. The 10 results from the folds were then averaged to produce a single estimation. The advantage of this rotation method over repeated random subsampling was that all observations were used for both training and validation, and each observation was used for validation exactly once.

To control the different percentage of PTSD in the data set, we set a range of prevalence as 5%, 15%, 25%, and 50%. The 10-fold cross-validation was applied to each study corresponding to the prevalence range. The purpose of this strategy was to present a fair comparison of the machine-learning algorithms with \( n \)-grams over a range that is comparable to the clinical practice. The range of prevalence of PTSD was comparable to the general base rates of PTSD recorded in the literature. For instance, in a review study made by Brewin (2005), the prevalence of PTSD ranged from 3% to 75% in 22 studies related to PTSD screening and diagnoses with the application of 13 different instruments. Of the 22 studies, the majority (nine studies) had the prevalence of PTSD between 45% and 55%, four studies with the prevalence of PTSD below 10%, seven studies with the prevalence of PTSD between 10% and 40%, and two studies with the prevalence of PTSD above 50%.

Five performance metrics—accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)—were used to evaluate the efficiency of the four machine-learning algorithms combined with \( n \)-gram models. The computation rules of these five indicators are presented in Table 2.
Accuracy, the main metric used in classification, is the percentage of correctly classified texts. Sensitivity and specificity measure the proportion of actual positives and actual negatives that are correctly identified, respectively. These two indicators do not depend on the prevalence in the corpus and hence are more indicative of real-world performance. The predictive values PPV and NPV address what the probability is that someone who has a positive test result will report a diagnosis of PTSD and that someone who has a negative test result will not receive a PTSD diagnosis. As Kessel and Zimmerman (1993) suggested, whereas sensitivity and specificity are independent of the prevalence of the disorder in the population, and thus can readily be compared across studies, PPV and NPV are sensitive to population prevalence. If there are very few cases to detect, the PPV of the test will suffer, whereas if the vast majority of the population is affected, its NPV will be correspondingly limited (Baldessarini, Finklestein, & Arana, 1983). In other words, at a low prevalence, a negative test result is more likely to be correct, whereas at a high prevalence, a positive result is more likely to be correct (Brewin, 2005). For a complete discussion of calculating and interpreting diagnostic performance statistics, refer to Kessel and Zimmerman (1993) and Baldessarini et al. (1983).

With the aim of examining whether the performance of the computerized textual screening system is comparable to that of the standard screening instruments used in daily practice, we used the results in Brewin’s (2005) review as a benchmark for PTSD screening in clinical settings. As mentioned earlier, in Brewin’s study, the performances of 13 commonly used screening instruments, such as the Impact of Event Scale (Horowitz, Wilner, & Alvarez, 1979), the PTSD Checklist (Weathers, Huska, & Keane, 1991), and the Davidson Trauma Scale (Davidson et al., 1997), were systematically reviewed in regard to 22 PTSD studies, with the sample sizes ranging from 65 to 422 participants. We compared the performance metrics of the four textual assessment models with the mean performance of screening instruments that were used in Brewin’s study.

Finally, to investigate whether the number of keywords used in the model influenced the performance of textual screening, we examined the changes of performance metrics for four assessment models with an increasing number of keywords (key vectors). Specifically, the experiment started with 10 keywords that had the highest chi-square scores, that is, five keywords labeled as PTSD classifiers and five keywords labeled as non-PTSD classifiers, and ended with 1,000 keywords—500 keywords from either classifier label.

**Results**

**Comparison Between Machine-Learning Text Classifiers With N-Grams**

The performances of four machine-learning text classifiers in conjunction with five textual representation models in the 10-fold cross-validation are summarized in Table 3. The value in each cell presents the averaged results and standard deviations from the cross-validation. The lowest and highest averaged values in each column are highlighted and are in bold forms, respectively. As shown in this table, the DT in conjunction with unigrams yielded the lowest values in sensitivity ($M = .58$, $SD = .07$) and NPV ($M = .57$, $SD = .04$). The NB with a mix of n-grams obtained the highest NPV ($M = .87$, $SD = .06$). The SVM with trigrams produced the highest value in sensitivity ($M = .95$, $SD = .04$) but sacrificed dramatically in specificity ($M = .10$, $SD = .08$) and showed the lowest PPV ($M = .52$, $SD = .03$) among all. The SVM also resulted in the lowest accuracy rate ($M = .53$, $SD = .02$) with the combined use of trigrams. The PSM with unigrams attained the highest in accuracy rate ($M = .82$, $SD = .05$), implying the best agreement with
practitioners’ diagnoses using item-based questionnaires through traditional structured interviews. The joint addition of $n$-grams (see the last two rows by each model in Table 3) showed a more balanced performance in text classification than with their single use. For instance, with the joint representation of unigrams and bigrams, the PSM moderately lowered the sensitivity value from .85 (unigrams) to .81 but benefited in a significant increase in specificity from .78 (unigrams) to .81. Furthermore, compared with unigrams, the combination of $n$-grams helped enhance the predictive power. The highest predictive power of PTSD was suggested by the highest value of PPV ($M = .81$, $SD = .04$) produced by the PSM with a mixture of unigrams, bigrams, and trigrams. The NB with a mixture of $n$-grams showed the best confidence in excluding non-PTSD from further assessments by the highest value of NPV ($M = .87$, $SD = .06$). However, the overall accuracy rate was not significantly improved by the addition of $n$-grams and was even decreased a bit when the bigrams and trigrams were singly used as key vectors in the text classification. We previously noted that the unigrams generally showed the best accuracy within the comparisons among $n$-grams within each machine-learning model except for the DT model. Consequently, we continued the comparison of performances among different machine-learning algorithms with a focus on the unigrams only in the following studies.

Figure 2 presents the receiver operator characteristic curves for the four machine-learning algorithms in conjunction with the top 1,000 unigrams. A receiver operator characteristic graph is a technique for visualizing, organizing, and selecting classifiers based on their performance (Fawcett, 2006). The curve is created by plotting the true positive rate (also known as sensitivity) against the false positive rate (also known as 1–specificity) at various threshold settings (see Table 2, for the equations). The optimal thresholds, that is, those with perfect sensitivity and specificity, lie near the upper left corner. The area under the curve (AUC; Bradley, 1997; Hanley & McNeil, 1982) is a portion of the area of the unit square; its value is always between 0 and 1.0. However, because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of $.5$, no realistic classifier should have an AUC less than $.5$. In the current study, the AUCs of the four text classifiers—DT, NB, SVM, and PSM—were $.68$, $.89$, $.81$, and $.94$, respectively. These results implied that the PSM with unigrams was the most efficient approach to distinguish between PTSD and non-PTSD, while the DT was the least efficient classification approach. This result verified the findings presented in Table 3.

These are the unigram features that contributed significantly in the classification of PTSD and non-PTSD in the present study. In decreasing order of chi-square score, the top 10 keywords in the PTSD group were "emotion," Figure 2 presents the receiver operator characteristic curves for the four machine-learning algorithms in conjunction with the top 1,000 unigrams. A receiver operator characteristic graph is a technique for visualizing, organizing, and selecting classifiers based on their performance (Fawcett, 2006). The curve is created by plotting the true positive rate (also known as sensitivity) against the false positive rate (also known as 1–specificity) at various threshold settings (see Table 2, for the equations). The optimal thresholds, that is, those with perfect sensitivity and specificity, lie near the upper left corner. The area under the curve (AUC; Bradley, 1997; Hanley & McNeil, 1982) is a portion of the area of the unit square; its value is always between 0 and 1.0. However, because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of $.5$, no realistic classifier should have an AUC less than $.5$. In the current study, the AUCs of the four text classifiers—DT, NB, SVM, and PSM—were $.68$, $.89$, $.81$, and $.94$, respectively. These results implied that the PSM with unigrams was the most efficient approach to distinguish between PTSD and non-PTSD, while the DT was the least efficient classification approach. This result verified the findings presented in Table 3.

These are the unigram features that contributed significantly in the classification of PTSD and non-PTSD in the present study. In decreasing order of chi-square score, the top 10 keywords in the PTSD group were “emotion,” Figure 2 presents the receiver operator characteristic curves for the four machine-learning algorithms in conjunction with the top 1,000 unigrams. A receiver operator characteristic graph is a technique for visualizing, organizing, and selecting classifiers based on their performance (Fawcett, 2006). The curve is created by plotting the true positive rate (also known as sensitivity) against the false positive rate (also known as 1–specificity) at various threshold settings (see Table 2, for the equations). The optimal thresholds, that is, those with perfect sensitivity and specificity, lie near the upper left corner. The area under the curve (AUC; Bradley, 1997; Hanley & McNeil, 1982) is a portion of the area of the unit square; its value is always between 0 and 1.0. However, because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of $.5$, no realistic classifier should have an AUC less than $.5$. In the current study, the AUCs of the four text classifiers—DT, NB, SVM, and PSM—were $.68$, $.89$, $.81$, and $.94$, respectively. These results implied that the PSM with unigrams was the most efficient approach to distinguish between PTSD and non-PTSD, while the DT was the least efficient classification approach. This result verified the findings presented in Table 3.

These are the unigram features that contributed significantly in the classification of PTSD and non-PTSD in the present study. In decreasing order of chi-square score, the top 10 keywords in the PTSD group were “emotion,”

### Table 3. Averaged Results From Four Classification Models: DT, NB, SVM, and PSM Based on 10-fold Cross-Validation.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>Unigrams</td>
<td>.57 (0.04)</td>
<td>.58 (0.07)</td>
<td>.56 (0.08)</td>
<td>.57 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>.60 (0.04)</td>
<td>.58 (0.08)</td>
<td>.61 (0.08)</td>
<td>.60 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Trigrams</td>
<td>.57 (0.04)</td>
<td>.62 (0.15)</td>
<td>.51 (0.18)</td>
<td>.57 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams</td>
<td>.58 (0.04)</td>
<td>.60 (0.07)</td>
<td>.57 (0.09)</td>
<td>.58 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams + Trigrams</td>
<td>.58 (0.04)</td>
<td>.60 (0.06)</td>
<td>.56 (0.08)</td>
<td>.58 (0.04)</td>
</tr>
<tr>
<td>NB</td>
<td>Unigrams</td>
<td>.79 (0.03)</td>
<td>.78 (0.06)</td>
<td>.80 (0.07)</td>
<td>.80 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>.68 (0.04)</td>
<td>.89 (0.06)</td>
<td>.47 (0.11)</td>
<td>.64 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Trigrams</td>
<td>.60 (0.03)</td>
<td>.92 (0.04)</td>
<td>.28 (0.08)</td>
<td>.57 (0.03)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams</td>
<td>.78 (0.04)</td>
<td>.87 (0.06)</td>
<td>.70 (0.10)</td>
<td>.75 (0.06)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams + Trigrams</td>
<td>.76 (0.03)</td>
<td>.90 (0.06)</td>
<td>.64 (0.07)</td>
<td>.72 (0.04)</td>
</tr>
<tr>
<td>SVM</td>
<td>Unigrams</td>
<td>.80 (0.03)</td>
<td>.86 (0.05)</td>
<td>.74 (0.08)</td>
<td>.77 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>.57 (0.03)</td>
<td>.92 (0.06)</td>
<td>.23 (0.07)</td>
<td>.55 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Trigrams</td>
<td>.53 (0.02)</td>
<td>.95 (0.04)</td>
<td>.10 (0.08)</td>
<td>.52 (0.03)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams</td>
<td>.70 (0.06)</td>
<td>.84 (0.11)</td>
<td>.55 (0.07)</td>
<td>.66 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams + Trigrams</td>
<td>.69 (0.05)</td>
<td>.85 (0.10)</td>
<td>.53 (0.06)</td>
<td>.64 (0.04)</td>
</tr>
<tr>
<td>PSM</td>
<td>Unigrams</td>
<td>.82 (0.05)</td>
<td>.85 (0.08)</td>
<td>.78 (0.07)</td>
<td>.80 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>.76 (0.04)</td>
<td>.76 (0.09)</td>
<td>.77 (0.05)</td>
<td>.77 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Trigrams</td>
<td>.67 (0.05)</td>
<td>.64 (0.08)</td>
<td>.78 (0.07)</td>
<td>.75 (0.07)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams</td>
<td>.81 (0.05)</td>
<td>.81 (0.10)</td>
<td>.81 (0.05)</td>
<td>.81 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Unigrams + Bigrams + Trigrams</td>
<td>.80 (0.05)</td>
<td>.80 (0.10)</td>
<td>.80 (0.05)</td>
<td>.81 (0.04)</td>
</tr>
</tbody>
</table>

Note. DT = decision tree; NB = naive Bayes; SVM = support vector machine; PSM = product score model. The value within the parentheses presents the standard deviation in the 10-fold cross-validation. PPV and NPV represent positive predictive value and negative predictive value, respectively. The cells highlighted in shade and bold forms represent the lowest and highest value within each column, respectively.
“rape,” “abuse,” “car,” “year,” “flashback,” “home,” “nightmare,” “fire,” and “therapy.” The top 10 keywords in the non-PTSD group were “wake,” “dream,” “feel,” “like,” “anxiety,” “get,” “worry,” “head,” “breath,” and “love.” Analogous to the results obtained by Orsillo, Batten, Plumb, Luterek, and Roessner (2004) in the research regarding emotion expressions of PTSD patients, the words favored by the PTSD patients had relatively stronger negative semantic tendency no matter the lexical forms as adjectives, nouns, or verbs.

The performance of computerized textual classification appeared to largely depend on the extracted keywords in the current analysis. It seemed that the textual screening approach did a good but not perfect job in the majority of cases. Thus, it would be appropriate to conduct further analysis on when and why this approach performed as it did. We illustrated this concern with the following two examples.

Example 1: I am 24 years old and was involved in a house fire two years ago. Since then I have split with my long-term boyfriend and not been able to form any other committed relationship. I have been suffering from insomnia regularly, which is impacting on my work situation. I do not think I have flashbacks. I am always aware of where I am but certain smells and sounds make me unable to think about anything else for days at a time and causes me to become really emotional and unable to focus on anything. I have been feeling really disconnected from my life for the last two years and I have finally come to the realization that I need to get help. I am just not sure where to go, or if this is something that will go away on its own.

Example 2: I was the victim of Domestic Violence, and as result of no help from anyone, I have so much anger, that I have become the Batter now. I have never been so ashamed of myself for my actions to someone that I love. He is an alcoholic who is sobering now for over 100 days, after a DUI accident. We were separated, he was arrested and asked me to come back to work on us. Help him through his ordeal. Prove my "LOVE" to him, though this time, and I went back, I knew within a few days it was all bogus, and I told him, I gave you my word that I would stand by you. You obviously are just using me while you don’t have a license. We argued a couple times, during this 3 month time, once I lost my self and hit him. Yes, I did hit him I struck him and slapped him I was furious; he was telling me again that he used me for his benefit. This man used to hit me

Figure 2. Receiver operator characteristic curves showing classification performance of the four machine-learning algorithms in conjunction with the top 1,000 unigrams.

Note. The diagonal line represents the strategy of randomly guessing a class.
has cause me to have surgery, YET, I feel like a horrible person because I am the one that’s homeless again from situation, and he lives on like nothing. Hopefully someday I will be able to do the same.

These two examples were both written by PTSD patients. The top 20 keywords—10 for the PTSD and 10 for the non-PTSD group—were used in the analysis. In the first example, eight keywords (bold font) were identified: three PTSD indicators and five non-PTSD indicators, while in the second example, five keywords (bold font) were identified, which were all non-PTSD indicators. Based on the calculation, it would not be difficult to label the first example as a PTSD self-narrative by classification algorithms because the PTSD indicators identified in this example had stronger weights than the non-PTSD indicators. However, in the second example, because only the non-PTSD indicators were found, a high probability of wrong classification into the non-PTSD group would be expected. For instance, the word “LOVE” seemed to have an opposite meaning.

Comparison Across a Range of Prevalence of PTSD

The averaged results of four machine-learning algorithms with unigrams that were derived from cross-validation with different prevalence of PTSD are presented in Table 4. Among the five performance metrics, it was found that the overall accuracy and specificity of each text classifier showed a slight change across different prevalence settings. A moderate decrease of around 10% was found in sensitivity and NPV of each model when the prevalence of PTSD increased. A substantial increase in PPV was noticed with the increase of prevalence of PTSD. It was interesting to find that when using the unigrams alone, the performances of NB, SVM, and PSM were quite comparable in the classification of self-narratives for PTSD and non-PTSD. Strictly speaking, the PSM was marginally better in overall accuracy than the other two. Consistent with the results shown in Table 3, the DT had the poorest performance across all the prevalence levels among the four machine-learning algorithms.

The prevalence is an important indicator when reporting the performance metrics of a screening method. Whereas sensitivity and specificity are independent of the prevalence of the disorder in the population, positive and negative predictive powers are sensitive to population prevalence (Brewin, 2005). Baldessarini et al. (1983) commented that in general, at low prevalence, a negative test result is more likely to be correct, whereas at high prevalence, a positive result is more likely to be correct. Moreover, highly sensitive tests (those having a low false negative rate), even with moderate specificity, are particularly useful when test results are negative and when the prevalence of the condition is low, that is, they should be helpful in excluding individuals from further assessment. As the results showed in our study, when the prevalence of PTSD was low, the SVM and PSM were very sensitive and had a high negative predictive power. These results suggested that these two models could perform well in excluding the individuals identified as non-PTSD from the follow-up tests.

<table>
<thead>
<tr>
<th>Prevalence of PTSD</th>
<th>Classifier</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>DT</td>
<td>.57</td>
<td>.65</td>
<td>.57</td>
<td>.13</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>.78</td>
<td>.72</td>
<td>.80</td>
<td>.28</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>.76</td>
<td>.93</td>
<td>.74</td>
<td>.29</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>PSM</td>
<td>.80</td>
<td>.94</td>
<td>.78</td>
<td>.32</td>
<td>.99</td>
</tr>
<tr>
<td>15%</td>
<td>DT</td>
<td>.56</td>
<td>.61</td>
<td>.56</td>
<td>.22</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>.78</td>
<td>.71</td>
<td>.80</td>
<td>.44</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>.76</td>
<td>.87</td>
<td>.74</td>
<td>.42</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td>PSM</td>
<td>.79</td>
<td>.85</td>
<td>.78</td>
<td>.46</td>
<td>.96</td>
</tr>
<tr>
<td>25%</td>
<td>DT</td>
<td>.57</td>
<td>.58</td>
<td>.56</td>
<td>.40</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>.77</td>
<td>.74</td>
<td>.78</td>
<td>.64</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>.78</td>
<td>.86</td>
<td>.74</td>
<td>.64</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td>PSM</td>
<td>.80</td>
<td>.84</td>
<td>.78</td>
<td>.67</td>
<td>.91</td>
</tr>
<tr>
<td>50%</td>
<td>DT</td>
<td>.57</td>
<td>.58</td>
<td>.56</td>
<td>.57</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>.79</td>
<td>.78</td>
<td>.80</td>
<td>.80</td>
<td>.79</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>.80</td>
<td>.86</td>
<td>.74</td>
<td>.78</td>
<td>.84</td>
</tr>
<tr>
<td></td>
<td>PSM</td>
<td>.82</td>
<td>.85</td>
<td>.78</td>
<td>.80</td>
<td>.85</td>
</tr>
<tr>
<td>Mean performance in Brewin’s review</td>
<td></td>
<td>.86</td>
<td>.83</td>
<td>.85</td>
<td>.70</td>
<td>.90</td>
</tr>
</tbody>
</table>

Note. DT = decision tree; NB = naive Bayes; SVM = support vector machine; PSM = product score model; PTSD = posttraumatic stress disorder. PPV and NPV represent positive predictive value and negative predictive value, respectively.
Comparison With a Screening Benchmark in Clinical Practice

An additional comparison was made between the performance of four machine-learning algorithms and the mean performance of traditional screening measures that were presented in Brewin (2005). The last row in Table 4 exhibits the mean performance of 22 studies related to PTSD screening in Brewin’s review. When the prevalence of PTSD was around 50%, the performance of textual classifiers, especially SVM and PSM, was closer to that of the traditional screening measures. It appeared that the sensitivity and PPV of SVM and PSM remained higher than that of the traditional measures across the whole range of prevalence. However, their overall accuracy and specificity were a bit lower than the benchmark. The SVM and PSM resulted in a higher NPV compared with the traditional measures when the prevalence was below 50%, but the opposite results were found when the prevalence increased above 50%. The performance of NB was generally lower than the benchmark, except for the PPV, when the prevalence was 50%. As we expected, the performance of DT was too far from the benchmark, particularly in NPV, when the prevalence was lower.

Effect of Number of Keywords on Model Performance

We further examined the performance stability of the four text classifiers in conjunction with n-grams by increasing the number of keywords along the text classification procedure. Figure 3 presents an example of the overall classification accuracy predicted by the four models with unigrams. The horizontal axis indicates the number of keywords (i.e., unigrams) attached to the models and the vertical axis indicates the percentage of accuracy in classification. The PSM curve was at the top, closely followed by the SVM and NB and substantially beyond the DT. The PSM, SVM, and NB started from a relatively low value when only a few keywords were used but quickly rose when more keywords were added. After the “elbow point” around 50 keywords, the PSM remained relatively stable, whereas the SVM and NB exhibited obvious waves. The “elbow point” implied that the top 50 keywords with the highest chi-square scores played essential roles in classification, which explained the most variance between the PTSD and non-PTSD corpora. The DT curve kept flat but located on a fairly low level throughout the whole range, suggesting that this model could make a stable classification, but its classification performance was rather low.

Discussion

This article presented a new automated assessment system to screen for PTSD based on patients’ self-narratives using NLP and text-mining techniques. A comparative study was conducted among four classification models—DT, NB, SVM, and PSM—in conjunction with five data representations—unigrams, bigrams, trigrams, a combination of unigrams and...
bigrams, and a mixture of \(n\)-grams. With the sample at hand, it was found that the narrative classification accuracy (82%) was maximized with the PSM in conjunction with unigrams. Although the addition of \(n\)-grams has not significantly enhanced the overall classification accuracy, it did help balance the performance metrics of text classification and improve the reliability of prediction.

Slight prevalence effects were found in the overall accuracy of all the machine-learning algorithms; however, a substantial increase of PPV was noticed with the increase of prevalence of PTSD. When the prevalence of PTSD was low, the SVM and PSM had good sensitivity and high negative predictive power. This suggested that these two models could perform well in excluding the individuals identified as non-PTSD from the follow-up tests. Furthermore, the SVM and PSM seemed to be more sensitive in detecting PTSD than the traditional screening measures, but their ability at detecting non-PTSD was a bit lower than the benchmark in clinical practice.

Computer-delivered diagnosis and treatment interventions for those with middle to moderate mental health needs have been shown to have promise as an adjunct to more traditional forms (Graham, Franses, Kenwright, & Marks, 2000; Owen, Hanson, Preddy, & Bantum, 2011; Proudfoot et al., 2004). The development of textual screening system for psychiatric patients was initiated in both the fields of psychiatry and applied linguistics. The whole procedure could be easily administered in offices of clinicians or embedded in an Internet-based test as an additional module to online psychiatric diagnosis. With its help, people living in remote areas, those with restricted mobility, or those reluctant to engage in face-to-face interviews could complete web-based tests in a private, flexible, and relaxed setting (Naglieri et al., 2004). Furthermore, compared with itemized questionnaires, self-narratives provide patients with opportunities to express themselves freely, and they are easier to interpret by clinicians. That is, patients may describe the traumatic events and symptoms in their own style without limitations set by the item options. From texts, clinicians may understand the content straightforwardly without having to consult a psychometrician for interpretation of the scale parameters.

The textual assessment system developed in this study can be applicable to research methods with similar background and makeup, for instance, in screening for multiple mental diseases such as depression according to the \textit{Diagnostic and Statistical Manual of Mental Disorders} criteria (American Psychiatric Association, 2000). Depression has been forecast as the second-leading cause of disability by 2020 (World Health Organization, 2001) and is expected to be the largest contributor to disease burden by 2030 (World Health Organization, 2008). Moreover, early detection, either by a general practitioner or via an online screening test, could result in more effective and shorter treatment, potentially reducing cost. The textual screening method presented here would be an ideal approach to improve the cost effectiveness of a diagnostic procedure and reduce both patients’ burden and clinicians’ workload. In addition, new applications of text-mining techniques, for instance, speech recognition where patients’ spoken words can be automatically transferred into written forms, would bring extra benefit to both practitioners and patients. This application may especially help patients who are not able to express their feelings in writing for screening and diagnosis.

From the technical aspect, the effect of \(n\)-grams in text classification also merits discussion. Bekkerman and Allan (2003) summarized that there exist two major approaches to incorporating \(n\)-grams into document representation. The first excludes unigrams from the representation and bases the representation on \(n\)-grams \((n > 1)\) only, while the second one applies \(n\)-grams and unigrams together.

It turns out that the first approach, in most cases, leads to a deterioration in classification accuracy in comparison with the BOW due to the high dimensionality, low frequency, and high degree of synonymy. The second approach might improve the results in some cases, but statistical significance was usually shown on very specific data sets, where the baseline classification results were low or in domains with severely limited lexicons and high chances of constructing stable phrases (Lewis, 1992). In the current application, the unigrams with SVM as well as PSM have already reached a high agreement between computer and practitioners’ diagnoses, implying that the unigrams were powerful enough to represent the relatively small and “simple” corpus like the collection of patients’ self-narratives. Therefore, the classification accuracy apparently was not enhanced with the addition of \(n\)-grams. It might be interesting to apply the \(n\)-gram text-mining method to a larger and more complex data set in a future study and include the textual structure features as well, such as grammatical properties and parts of speech, to supplement the frequency-based representation model.

A further exploration was done on the interaction of PTSD-type events and robust textual classifiers. There were eight types of traumatic events experienced by the participants in our sample. Based on the 300 samples in hand, the sample size by each event type (an average of 37.5 stories) would be too small to generalize an accurate conclusion regarding the association between the word features and traumatic event type. But we found we could perform an analysis regarding individuals experiencing multiple traumatic events because there were 86 (28.7%) participants, consisting of 77 PTSD and 9 non-PTSD respondents, who reported they had at least two different traumatic events in their life (for instance, a childhood assault and a rape as an adult). Earlier studies showed evidence that those with multiple traumas were more likely to develop PTSD than those with a single trauma. Having two or more traumatic events involving an
assault during childhood was associated with a nearly fivefold greater risk that a traumatic event in adulthood would lead to PTSD (Breslau, Chilcoat, Kessler, & Davis, 1999) than someone who had not experienced trauma. Therefore, in the current study, we compared the textual features in self-narratives written by individuals with multiple trauma events with those with single trauma events.

It was found that in the PTSD group, the words related to event types (e.g., “rape” and “fire”) and timing (e.g., “year”) were more likely used by participants with multiple traumas, while the words more associated with PTSD symptoms (e.g., “flashback” and “nightmare”) were more likely used by those with single trauma events. Although in this study it would be challenging to thoroughly discuss which keywords are discriminating or robust by event type, the topic is worth future investigation based on a larger sample.

With evidence from the current study, the number of keywords attached to the classification model could make an impact on screening performance. People might wonder how many keywords are sufficient in the textual screening. We would recommend the “elbow point” as an ideal answer. The principal reason is that the values of performance metrics vary substantially before the “elbow point” but are relatively stable after it, which suggests that the efficiency of textual classification would not be much enhanced even with the addition of more keywords. This phenomenon can be explained by Zipf’s (1949) Law, which states there is a small vocabulary that accounts for a large part of the tokens in the text. Therefore, generally speaking, in daily practice, it is not necessary to include a whole set of keywords in the text classification. The “elbow point,” more or less, suggests an optimal number of inclusions of the keywords.

To help practitioners select an optimal algorithm in their own studies, the following pros and cons of each model should be weighed. The DT model is one of the most comprehensive models for visually tracking the path in classification. It is easy to understand how a DT classifies an instance as belonging to a specific class. However, this model may result in low accuracy, especially for a small data set, and encounters the problem of overfitting when the tree grows too large, based on the accidental properties of the training set.

The major advantages of NB are its short computational time for training and simple form of a product with the assumption of independence among the features. Unfortunately, the assumption of independence among words is not always correct, and thus, the NB is usually less accurate than other more sophisticated learning algorithms. However, the NB is still a very effective model in classification. Domingos and Pazzani (1997) performed a large-scale comparison of the NB with state-of-the-art algorithms—including DT, instance-based learning, and rule induction—on standard benchmark data sets and found it to be superior sometimes to the other learning schemes, even on data sets with substantial feature dependencies.

The SVM is a powerful machine-learning algorithm and has been standardized in many coding programs, for instance, Python and Program R. However, the linear SVM may result in a poor classification accuracy caused by the low-frequency features, such as trigrams, in the current study. A better solution might be to use a nonlinear SVM to get a better model fit.

The PSM has more flexibility in the model decision threshold than NB. As shown in Formula 6, the decision threshold $b$ could be set as an unfixed constant in practice. For example, in a clinical setting, practitioners may want to exclude people without PTSD from further tests, which needs a relatively higher specificity value. On the other hand, when practitioners focus on treatment for patients with PTSD, a more sensitive result from the text analysis is probably required to detect potential patients as precisely as possible. In addition, because the PSM allocates a set of term weights for each keyword or key vector, more time and storage space are needed in the training and validation process. It might reduce the PSM’s effectiveness when using a large sample or performing a multiple categorization.

While we found promising results in the current study, some limitations also merit discussion. First, because the participants in this study were all individuals seeking care for mental health concerns and who, by the nature of the form of data collection, were computer-literate, a potential source of selection bias might exist. The sample was likely not representative of the general population of those with PTSD, favoring individuals with access to a computer and a certain level of writing ability. To address this potential issue, new applications of text-mining techniques—for instance, speech recognition where patients’ words can be automatically transferred into written forms—would bring extra benefits for both patients and practitioners. With the use of speech recognition techniques, the potential PTSD patients would not have to write down their stories but could speak them out. Meanwhile, the data related to gestures, emotions, and languages could also be recorded for a comprehensive diagnosis.

Second, the primary language of participants was required to be English, but it was challenging to control for the language background of immigrants for this U.S.-based study. It is conceivable that nonnative speakers might have a hard time with vocabulary, which would negatively affect classifiers. There would also be cultural differences to consider in how people from different backgrounds describe traumatic events (e.g., Chen, 2005; Nussbaum, & Freund, 2009; Smyth, Hockemeyer, & Tulloch, 2008). It would be interesting for the future studies to investigate the language and cultural impacts on PTSD patients’ self-narratives.

In conclusion, the present study concerns the development of an n-gram-based computerized textual assessment.
system to screen for PTSD based on patients’ self-narratives. The results showed that the textual assessment on self-narratives achieved a high agreement with practitioners’ diagnoses, and the addition of higher order n-grams could help balance the classification metrics and enhance the reliability of classification prediction. This article further demonstrates that the automated textual assessment system is a promising tool for analyzing patients’ self-expression behaviors, thus helping practitioners identify potential patients at an early stage.

Acknowledgments
The authors would like to thank Larry Hanover for his help in reviewing this article.

Authors’ Note
Qiwei He was at University of Twente when she performed most of the work for this article.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The project of textual assessment for PTSD was supported by Stichting Achmea Slachtofferhulp Samenleving, the Netherlands.

Notes
1. The original data set included 308 participants. Eight participants withdrew from the study in the middle, because they had been diagnosed by one practitioner only. We discarded their cases from the current study. The result was a total of 300 participants for the final set.
2. A body of texts is usually called a text corpus.
3. Punctuation is often regarded as one gram when disassembling a sentence into n-grams.
4. The current study used the standard “English Stop-Word List” (127 words) in Python Natural Language Toolkit to deduct the noninformative words.
5. The stemming algorithm is used to normalize lexical forms of words, which may generate stems without an authentic word meaning such as “nightmar.”
6. Since we are interested only in ranking the chi-square score for each word to find the robust classifier, assessing the significance of the chi-square test is not important here.
7. In principle, the scope of threshold b could be set to be infinite. However, in practice, it is often recommended to be set within a range from −5 to +5.

References


