



Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach

Qiwei He^{*}, Bernard P. Veldkamp, Theo de Vries

Department of Research Methodology, Measurement and Data Analysis, Behavioral Sciences Faculty, University of Twente, P. O. Box 217, GW, 7500AE Enschede, The Netherlands

ARTICLE INFO

Article history:

Received 1 June 2011

Received in revised form 10 November 2011

Accepted 31 January 2012

Keywords:

Posttraumatic stress disorder

Text mining

Self narratives

Text classification

Screening

ABSTRACT

Much evidence has shown that people's physical and mental health can be predicted by the words they use. However, such verbal information is seldom used in the screening and diagnosis process probably because the procedure to handle these words is rather difficult with traditional quantitative methods. The first challenge would be to extract robust information from diversified expression patterns, the second to transform unstructured text into a structuralized dataset. The present study developed a new textual assessment method to screen the posttraumatic stress disorder (PTSD) patients using lexical features in the self narratives with text mining techniques. Using 300 self narratives collected online, we extracted highly discriminative keywords with the Chi-square algorithm and constructed a textual assessment model to classify individuals with the presence or absence of PTSD. This resulted in a high agreement between computer and psychiatrists' diagnoses for PTSD and revealed some expressive characteristics in the writings of PTSD patients. Although the results of text analysis are not completely analogous to the results of structured interviews in PTSD diagnosis, the application of text mining is a promising addition to assessing PTSD in clinical and research settings.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Accurately assessing and diagnosing Posttraumatic Stress Disorder (PTSD) are fundamental in improving treatment of the disorder and increasing our understanding of the construct of PTSD. But unfortunately, the approach to obtain precise results is made difficult by its complexity in relation to cost, logistics and training (Franklin and Thompson, 2005; Silove et al., 2007).

These substantial challenges imply that there is often an important role to be played by screening instruments that can be used to detect adverse psychological responses. Unlike diagnostic measures such as structured clinical interviews, screening instruments need not include items corresponding to specific diagnostic criteria but may be based on any measure (e.g., demographic, biological, or self-report items) that successfully predicts the criterion diagnosis (Brewin, 2005). Numerous screening instruments have been developed for PTSD during the past three decades, including the Hopkins Symptom Checklist (HSCL; Derogati et al., 1974), the Harvard Trauma Questionnaire (HTQ; Mollica et al., 1992), the Screen for Posttraumatic Stress Symptoms (SPTSS; Carlson, 2001), the Self-Rating Scale for Posttraumatic Stress Disorder (SRS-PTSD; Carlier et al., 1998), and etc. These instruments ideally contain the minimal number of items necessary for accurate case identification, have simple decision rules for determining who passes and fails the screening, and are applicable

to populations with varying prevalence of PTSD and experiencing different traumas (Brewin, 2005).

Despite significant advances in PTSD screening instruments, open questions regarding trauma history and physical symptoms are seldom utilized in the screening process. Although writing of self narratives has been commonly used as a treatment task to relieve the PTSD patients' stress (e.g., Batten et al., 2002; Orsillo et al., 2004) or as an approach to recalling memory for psycholinguistic studies (e.g., Foa et al., 1991; Brebion et al., 2011), little attention has been given to its benefits in the screening process. Over the past three decades, researchers have provided evidence to suggest that people's physical and mental health can be predicted by the words they use (Gottschalk and Gleser, 1969; Rosenberg and Tucker, 1979). For instance, Franklin and Thompson (2005) found that patients with PTSD may engage in various response styles when reporting symptoms in clinical interviews or on self-report measures of psychopathology. Smyth (1998) emphasized that verbal features in self narratives contained helpful information that would be important to analyze the PTSD patients' behavior. However, the procedure to handle such textual information is rather complicated and difficult (Smyth, 1998). The challenges mainly exist from two aspects: first, due to different background, e.g., educational level, social status, living conditions and etc., 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). Secondly, unlike the numeric data collected from questionnaires, textual data is often unstructured, neither having a pre-defined data model nor fitting well into relational

^{*} Corresponding author. Tel.: +31 53 4893650; fax: +31 53 4894239.

E-mail addresses: q.he@gw.utwente.nl (Q. He), b.p.veldkamp@gw.utwente.nl (B.P. Veldkamp), tdevries@wxs.nl (T. de Vries).

patterns. This results in irregularities and ambiguities that make it difficult to be directly analyzed using traditional quantitative methods. A process to transform the unstructured text into a structured dataset is necessarily added. Therefore, to address these challenges, new methods need to be developed.

The purpose of the present study is to develop a textual assessment method for PTSD screening using verbal features in self narratives with text mining techniques. In contrast to classical data analysis, text mining seeks to extract useful information from document collections through the identification and exploration of patterns among unstructured textual data (Feldman and Sanger, 2007). This technique has been widely applied in the field of natural language processing and information retrieval system, such as “Google” search engine and “Amazon” customer investigation, and recently, extended into educational and psychological studies. For example, Chung and Pennebaker (2008) identified dimensions of people thinking about themselves by developing an automated meaning extraction method for natural language. Gara et al. (2010) investigated the reasons for disagreement in schizophrenia diagnosis by mining the textual transcripts of structured interviews.

The general idea of this study is to ask trauma victims to write down their traumatic events and symptoms online, rather than conducting face-to-face interviews with item-based questionnaires. Based on their textual input, the respondents can be classified into PTSD (i.e., high risk to develop as PTSD) and NONPTSD (i.e., low risk to develop as PTSD) groups. Those who are identified as PTSD at this initial stage will be invited into a more extensive test for further precise diagnosis. Therefore, the textual screening method would be helpful if it could maximize the accuracy in finding the potential PTSD patients or significantly excluding the NONPTSD individuals from the follow up tests. As stated earlier, our present research has two specific objectives: (1) to develop a text mining model to accurately classify the individuals into PTSD and NONPTSD groups based on their self narratives; and (2) to extract verbal features in the PTSD patients' self narratives which would enhance the prediction further.

2. Method

2.1. Sample

The development of our text mining model was conducted on a set of self narratives written by trauma survivors that were collected via an online survey embedded in an open forum dedicating to the patients with mental health. The e-health survey for anxiety disorders was divided into two parts: two open questions regarding self narratives, including both traumatic events and symptom description, labeled Part 1, and the follow-up itemized questionnaires corresponding to DSM-IV criteria (APA, 2000), labeled Part 2. The requirements for the survey specified that: a) the participants must have experienced at least one traumatic event; b) the participants must have been diagnosed as PTSD or NONPTSD by at least two psychiatrists via the structured interviews with the standardized instruments, e.g., the Structured Clinical Interview for DSM-IV (SCID) PTSD module (First, 1997) and the Clinician Administered PTSD Scale (CAPS; Blake et al., 1995); c) it must be a first episode description; and d) both of the two open questions (i.e., traumatic events and symptom description) must be answered. In the present study, we used Part 1 data obtained from 308 participants. The data with missing diagnoses were discarded resulting in a total of 300 responses for the final set, among which 150 participants were confirmed as PTSD patients and the other 150 participants were NONPTSD. (We stopped collecting data at $n = 150$ per group.) We did consult with experienced psychiatrists on the textual data to ensure that the participants in the study were diagnosed correctly based on DSM-IV criteria. In addition, all the participants reported that they had at least three-month experience using internet-based services and did not meet problems in using the online survey system. The age of participants ranged from 18 to 75.

2.2. Text classification modeling

Text classification is a special approach in the field of text mining, aiming to assign textual objects from a universe to two or more classes (Manning and Schütze, 1999). The supervised text classification generally involves two phases, a training phase and a prediction phase (Jurafsky and Martin, 2009) (see Fig. 1). During training, the most discriminative keywords to determine the presence or absence of PTSD are extracted. The input for the machine learning algorithm is a set of pre-specified keywords

which may potentially be present in a self narrative. Further, the input consists of a label classifying each narrative (in the present case either “PTSD” or “NONPTSD”). The objective of the training phase is to “learn” the relationship between the keywords and class labels. The prediction phase plays an important role in checking how well the trained classifier model performs on a new dataset. The test set should consist of data which were not used during training. In the testing procedure, the keywords which were extracted from training are scanned in each new input. So the words which were systematically recognized are fed into the “trained” classifier model, which predicts the most likely label for each new self narrative.

2.3. Procedure

2.3.1. Preprocessing

To improve the efficiency of the training and prediction procedure, a preprocessing routine was implemented. This involved screening digital numbers, deducting non-informative “stop words” (e.g., “I”, “to”), common punctuations (e.g., “.”, “:”, “;”), and frequently used abbreviations (e.g., “isnt”, “Im”), and “stemming” the rest of words with the Porter Algorithm (Porter, 1980) to remove the common morphological endings. For example, the terms “nightmares”, “nightmarings” and “nightmared”, though in variant lexical forms, were normalized in an identical stem “nightmar”¹ by removing suffixes and linguistic rule-based indicators. As for the sample in hand, in comparison to the raw text, preprocessing efficiently reduced the number of word tokens (i.e., individual occurrences of words) and the number of word types (i.e., occurrences of different words)² by approximately 60% and 35%, respectively.

2.3.2. Training

The 300 self narratives were split into two sets. A training set consisting of 200 textual data was randomly extracted from the original data set, i.e., 100 from PTSD and 100 from NONPTSD text corpus³ respectively. The remaining 100 textual data, i.e., 50 from either corpus, were used as the test set. The diagnoses made by psychiatrists for each participant were set as the “standard labels” (i.e., PTSD or NONPTSD) for the inputs.

The input texts were represented by the so-called “bag-of-words” where each distinct word in the corpus acts as a feature. To simplify the computation procedure, each word in the “bag” was assumed independent. Feature extraction plays an important role in the training process. Forman (2003) reviewed abundant feature selection methods for text classification. The Chi-Square selection algorithm (Oakes, et al., 2001) is often applied because of the high effectiveness in finding robust keywords and testing the independency between corpora (Forman, 2003; Yang and Pedersen, 1997; Conway et al., 2009). In general, a Chi-Square test is used to assess whether paired observations on two variables are independent of each other (Jurafsky and Martin, 2009), for example, whether people from different regions differ in the frequency with which they report that they support a political candidate. Analogous to this example, suppose the paired observations are the occurrences of each word in the PTSD and NONPTSD corpora, the Chi-Square test can be used to assess whether individuals from different groups (PTSD or NONPTSD) differ in the occurrences of word (type) with which they report self narratives. If the ratios between each paired observations vary wildly, then the χ^2 score will be high, and we have evidence for a high degree of dissimilarity between the two corpora. The words with higher χ^2 scores make larger contributions in distinguishing one corpus from the other.

We took the benefits from the Chi-Square (χ^2) selection algorithm (Oakes et al., 2001) in the present study to extract the key word features. Using the 200 self-narratives in the training set, we compiled each word into a 2-by-2 confusion matrix (see Table 1) and calculated the χ^2 score for each word with the Chi-square equation for a 2-by-2 table:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}, \quad (1)$$

where N is the total number of word types in the training set, and O_{ij} represents the cell counts in the matrix. The counts O_{11} and O_{12} are the number of occurrences of a word in the PTSD and NONPTSD corpus, respectively, while O_{21} and O_{22} are the number of non-occurrences of this word in the two corpora. The words with higher χ^2 scores are more discriminative in classification, thus, we ranked the χ^2 score of each word in a descending order. The words ranked to the top were defined as the robust classifiers.⁴ Table 2 gives the top 20 keywords with the highest χ^2 scores used in this study.

¹ Stemming algorithm is used with the purpose to normalize lexical forms of words, which may generate stems without an authentic word meaning, such as “nightmar”.

² The frequency of words within the text corpus can be interpreted in two ways: word token and word type. Word token is defined as individual occurrence of words, i.e., the repetition of words is considered, whereas word type is defined as the occurrence of different words, i.e., excluding repetition of words.

³ A body of texts is usually called a text corpus.

⁴ Since we were only interested in ranking the Chi-Square scores for each word to find the robust classifiers, the assessment of significance of the Chi-Square tests was not important in this way.

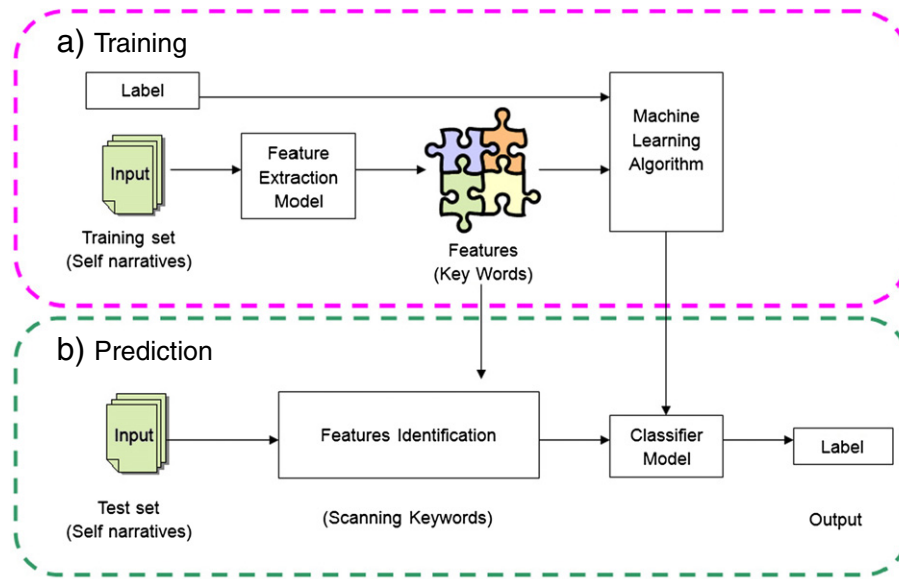


Fig. 1. The framework of supervised text classification. Note. The supervised text classification generally involves two phases, training and prediction. The objective of the training phase is to model (i.e., to learn) the relationship between the keywords and labels. The prediction phase is used to check how well the trained classifier model performs on a new dataset.

Given concerns on the imbalance of corpus size, we further assigned each word with two weights: PTSD weight and NONPTSD weight, that is, the number of occurrence of a word in either corpus against the total word tokens in the corresponding corpus. When the PTSD weight was larger than the NONPTSD weight, the word was labeled as a PTSD classifier. Conversely, when the PTSD weight was smaller than the NONPTSD weight, the word was labeled as a NONPTSD classifier (see the last column of Table 2). It was also important to implement a smoothing rule in the weight calculation – adding 0.5 to the number of word occurrence in both corpora – to avoid wrong estimation in zero probability that might happen when a word never occurs with a given label in the training set but possibly occurs in a new text. Without a smoothing rule, the new input will never be assigned this label, regardless of how well the other features fit it.

With the assumption of independency of words in the “bag”, we could obtain two final product scores for each self narrative by separately multiplying the PTSD weights and NONPTSD weights of all the keywords found in the text. We defined that when the logarithm of ratio between PTSD and NONPTSD product scores was larger than a certain threshold, the self narrative was determined as “PTSD”, otherwise, “NONPTSD” would be the correct label. In case of unequal group size in the training set, a correction factor – the proportion of group size against the total sample size – needs to be added to the calculation.

In addition, it was important to establish a set of unclassification rules to avoid mismatches caused by the randomness of word usage, especially when there were only a few keywords used in the text classification. We set three rules to identify the self narratives as “UNCLASSIFIED”, when either a) no keywords were found in the text, or b) only one keyword was found, or c) only two keywords were found but one labeled as a PTSD classifier while the other as a NONPTSD classifier.

2.3.3. Prediction

A prediction procedure is necessary to decide whether a classification model is accurately capturing the association between the words in self narratives and the labels assigned to them. As stated earlier, the test set followed the same format as the training set and was completely distinct from the training section, i.e., no self narratives in the training set was re-used in the prediction process. When a new set of inputs was

presented to the model, it started predicting the label (i.e., PTSD or NONPTSD) for each text, based on the “training” that it had received, as a human expert would do.

2.4. Analytic strategy

To ensure the proper generalization capabilities for the textual classifier model, a 15-fold cross validation procedure was applied. We subdivided the original corpus 15 times. Each time, two-thirds of the self narratives were randomly extracted as the training set, while the remaining one-third were used as the test set. We performed the multiple evaluations on the 15 different folds and averaged the results.

Table 2

The top 20 keywords with the highest Chi-Square scores used in the PTSD textual screening study.

Rank	Keywords	Chi-square score	Number of occurrences		Label
			PTSD	NONPTSD	
1	Wake	76.18	9	80	NONPTSD
2	Dream	56.26	34	105	NONPTSD
3	Feel	50.03	259	358	NONPTSD
4	Like	38.10	181	256	NONPTSD
5	Emotion	34.69	68	8	PTSD
6	Rape	33.87	43	0	PTSD
7	Anxiety	29.30	30	72	NONPTSD
8	Abuse	28.67	49	4	PTSD
9	Car	28.61	46	3	PTSD
10	Year	27.39	177	67	PTSD
11	Flashback	25.20	32	0	PTSD
12	Home	25.18	60	10	PTSD
13	Get	23.10	155	202	NONPTSD
14	Worry	23.09	18	49	NONPTSD
15	Head	20.73	22	52	NONPTSD
16	Breath	19.76	10	34	NONPTSD
17	Nightmare	19.57	37	4	PTSD
18	Fire	18.89	24	0	PTSD
19	Sometime	18.22	44	76	NONPTSD
20	Therapy	17.83	26	1	PTSD

Note. Each word is assigned two weights: PTSD weight and NONPTSD weight, i.e., the number of occurrence of a word in either corpus against the total word tokens in the corresponding corpus. When the PTSD weight is larger than the NONPTSD weight, the word is labeled as a PTSD classifier, vice versa.

Table 1

The confusion matrix for a word in the 2-by-2 Chi-Square score calculation.

	PTSD (C_1)	NONPTSD (C_2)
Word k	n_k	m_k
$\neg$ Word k	$len(C_1) - n_k$	$len(C_2) - m_k$

Note. C_1 = PTSD text corpus, C_2 = NONPTSD text corpus, n_k = the number of occurrence of word k in the PTSD corpus, m_k = the number of occurrence of word k in the NONPTSD corpus, $len(C_1)$ = the length of the PTSD corpus, i.e., the total number of word tokens in the PTSD corpus, $len(C_2)$ = the length of the NONPTSD corpus, i.e., the total number of word tokens in the NONPTSD corpus.

The relationship between text screening method at various thresholds and the standards, i.e., diagnoses by psychiatrists, was assessed through 2-by-2 tables. The performance metrics, including: a) sensitivity, that is, the probability that someone who has a PTSD diagnosis will have had a positive test result, b) specificity, the probability that someone who does not have a PTSD diagnosis will have had a negative test result, c) positive predictive power (PPP), the prediction confidence that someone who has a positive test result is authentically diagnosed as PTSD, d) negative predictive power (NPP), the prediction confidence that someone who has a negative test result will not receive a PTSD diagnosis, e) overall correct classification, the percentage of respondents correctly classified by the test as having or not having PTSD, and f) Kappa statistics, the level of agreement between the test in question and a standard beyond that accounted for chance alone, were calculated based on the averaged results from cross validation. (See Kessel and Zimmerman, 1993, for a complete discussion of calculating and interpreting diagnostic performance statistics.)

3. Results

3.1. Sample characteristics

The descriptive statistics of the sample are shown in Table 3. Using a 95% confidence interval, the independent sample *t*-test showed that the PTSD patients used more words ($t = 2054$, $d.f. = 232$, $p = 0.04$) and had a significantly larger vocabulary ($t = 3056$, $d.f. = 254$, $p < 0.01$) in description of traumatic events and physical symptoms than the individuals without PTSD. This was consistent with the results in the previous studies conducted by Gray and Lombardo (2001). The distribution of both word tokens and word types was spread out in the PTSD group which possessed the shortest (word token = 46) as well as the longest (word token = 1968) self narratives in the collected sample.

In the sample data, 1481 (38.37%) words were found only occurring in the PTSD corpus but never in the NONPTSD corpus; while 976 (25.28%) words were found only in the NONPTSD corpus but never in the other. Given concerns on the randomness of word selection in the utterance, we deducted the words with frequency less than 5 times, but still found 85 (2.20%) and 24 (0.62%) unique words remaining in the PTSD and NONPTSD corpus, respectively. In a decreasing order of word frequency, the ten unique words most used by the PTSD patients included “rape”, “flashback”, “fire”, “involve”, “avoid”, “incident”, “date”, “tower”, “men” and “fault”. The words “test”, “hardly”, “tumor”, “tight”, “excite”, “evil”, “pleasure”, “vision”, “frantic” and “funny” were found uniquely the top ten in the NONPTSD corpus. Analogous to the results obtained by Orsillo et al. (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 adjective, noun or verb.

3.2. Performance metrics

A range of thresholds for the PTSD textual screening method, along with corresponding sensitivity, specificity, positive predictive power, negative predictive power, overall correct classification rates, and Kappa statistics are presented in Table 4. A trade-off relationship was found between the sensitivity and specificity when the threshold of text classification model was justified from -5 to $+5$ as the actual

Table 3

The descriptive statistics of word tokens and word types in self narratives written by trauma victims with and without PTSD ($n = 300$).

	Number of document	Word token				Word type			
		Mean	S.D.	Max	Min	Mean	S.D.	Max	Min
PTSD	150	284	291	1968	46	149	86	507	37
NONPTSD	150	229	160	973	51	124	55	288	41
TOTAL	300	257	236	1968	46	136	73	507	37

Note. All the values were derived after preprocessing. The word token indicates the word count with repetition, while the word type indicates the word count without repetition.

requirements. A test for PTSD can be made highly sensitive by setting a very low threshold, i.e., setting the logarithm of ratio score between PTSD and NONPTSD at -5 or even lower, with the result that almost everyone who has a PTSD diagnosis will exceed this threshold. However, many people will exceed this threshold even though they do not have the disorder, with the result that the specificity of the test will be correspondingly low. Analogously, a test for PTSD can be made highly specific by setting a very high threshold, i.e., enhancing the logarithm of ratio score to $+5$ or even higher, with the result that almost nobody who does not have a PTSD diagnosis will exceed this threshold. However, many people will fall short of this threshold even though they do have the disorder, with the result that the sensitivity of the test will be correspondingly low.

To capture the greatest number of PTSD-positive cases without unduly sacrificing specificity, a target of approximately 0.85 sensitivity was selected a priori. Table 4 shows that when using a threshold of 0 as the cutoff of the logarithm ratio score between PTSD and NONPTSD corpora, the text classification model was able to classify 43 of the 50 PTSD-positive patients and 39 of the 50 PTSD-negative patients correctly for a sensitivity rate of 85%, a specificity of 78%, and an overall correct classification rate of 80%.

In order to examine whether the performance of textual screening method is comparable to that of the item-based screening instruments, we used the results in Brewin's (2005) study as a baseline, where the mean performance of 13 commonly used screening instruments, including the Impact of Event Scale (IES; Horowitz et al., 1979), the PTSD Checklist (PCL-C; Weathers et al., 1991), the Davidson Trauma Scale (DTS; Davidson et al., 1997), and etc., were systematically reviewed on 22 PTSD studies with the sample size in a range from 65 to 422 (see the last row in Table 4). Using a 95% confidence interval, the one sample *t*-test revealed that the sensitivity ($mean = 0.85$, $S.D. = 0.05$, $t = 3.46$, $p < 0.01$) and PPP ($mean = 0.80$, $S.D. = 0.05$, $t = 37.85$, $p < 0.01$) of the text classification model, with the logarithm ratio threshold at 0, were significantly higher than the mean performance of the reviewed instruments, while the specificity ($mean = 0.78$, $S.D. = 0.07$) and NPP ($mean = 0.85$, $S.D. = 0.07$) were a bit lower but kept at a moderate level.

To investigate whether the number of keywords used in the model influenced the performance of screening, we repeated these analyses with an increasing number of keywords from 10 to 1000 using a threshold score of 0. The averaged results of overall correct classification and unclassification rate from cross validation are plotted in Fig. 2. The horizontal axis indicates the number of keywords attached to the model and the vertical axis indicates the percentage of accuracy and unclassification. The accuracy curve starts at 60% when only 10 keywords (i.e., the top 10 words with the highest Chi-square scores)

Table 4

Performance metrics for various thresholds of the textual screening method compared with the mean performance of screening instruments reviewed by Brewin (2005).

Threshold	Sensitivity	Specificity	PPP	NPP	Overall correct classification	Kappa
-5	0.91	0.65	0.73	0.88	0.78	0.56
-4	0.90	0.68	0.74	0.88	0.79	0.58
-3	0.89	0.71	0.76	0.87	0.80	0.60
-2	0.88	0.73	0.77	0.86	0.81	0.61
-1	0.86	0.76	0.79	0.86	0.81	0.62
0	0.85	0.78	0.80	0.85	0.82	0.63
$+1$	0.84	0.80	0.81	0.84	0.82	0.64
$+2$	0.82	0.82	0.82	0.83	0.82	0.64
$+3$	0.80	0.84	0.83	0.81	0.82	0.63
$+4$	0.77	0.85	0.84	0.80	0.81	0.63
$+5$	0.75	0.87	0.85	0.79	0.81	0.62
Mean performance in Brewin's review	0.83	0.85	0.70	0.90	0.86	

Note. PPP and NPP indicate the positive predictive power and negative predictive power, respectively.

are used in the analysis, but quickly rises to 80% when the number of keywords is increased to 50. Afterwards, the curve turns in a flat gradient section and keeps stable, even though more keywords are used. Likewise, the unclassification curve also has a salient “elbow point” when the number of keywords is around 30 to 50. It falls down fast to 0 from 15% when only 10 keywords are used.

4. Discussion

The results demonstrated that the textual screening method developed in the field of text mining was very promising for making accurate and reliable screening for PTSD. With a text mining tool, psychiatrists can easily draw inferences from a large number of words used by the patients and describe the relative importance or contribution of each word by term weight. Although the verbal features are not formal criteria for DSM-IV diagnosis (APA, 2000), they may provide additional information in the assessment process as indirect predictors.

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 power are sensitive to population prevalence (Brewin, 2005). 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 (Baldessarini et al., 1983). Since the present study used a balanced sample size in the PTSD and NONPTSD group, all the results were reported at the prevalence of 50%. For those studies with prevalence lower or higher than 50%, the conclusions should be drawn carefully. Moreover, Baldessarini et al. (1983) commented that 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 shown in our study, when the threshold was set as 0 or lower than 0, the text classification model was highly sensitive and had a relatively high negative predictive power, which suggested that this model could perform well in excluding the individuals identified as NONPTSD from the follow-on tests.

With evidences from the current study, the number of keywords attached to the classification model could make impacts on the performance of screening. People might wonder “how many keywords are good enough 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 keep relatively stable after this point, which suggests that the efficiency of textual classification would not be enhanced much even though more keywords are added. This phenomenon can be explained by the Zipf's Law (1949) that reflects the fact that there is a small vocabulary which accounts for a large part of the tokens in the text. Therefore, generally speaking, in the 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 inclusion of the keywords.

Another issue that worth concerns is that the results from textual screening method are not completely analogous to those derived from the standard structured interviews. Some evidence has shown that the unstructured verbal data can provide very helpful information for an individual's health status, but not as total (e.g., Mendoza et al., 2011; Smyth, 1998). Thus, in practice, a psychiatrist might be reluctant to ensure an accurate diagnosis when merely relying on the verbal features from self narratives. To solve this problem, Shrout et al. (1986) recommended a two-step approach to identifying PTSD. In the first step, individuals are administered screening measures related to a particular disorder. If a predetermined cutoff score is exceeded, a more extensive and time-consuming diagnostic evaluation can be conducted in the second step. By administering a screening measure first to identify cases that are most likely to require additional assessment and possible treatments, clinicians can efficiently allocate clinical services where they are potentially most needed or refer cases to other health care providers for appropriate clinical services (Coffey et al., 2006). By following this two-step approach, the textual screening method would be set as the first step, which plays an important role in screening for PTSD and providing the prior information for the further tests. The development of a complete computerized intake procedure, namely the combination of text classification and optimal follow-up tests will be a topic for the further research.

In addition to the text classification algorithm, so-called “product score model”, developed in the present study, there are a number of

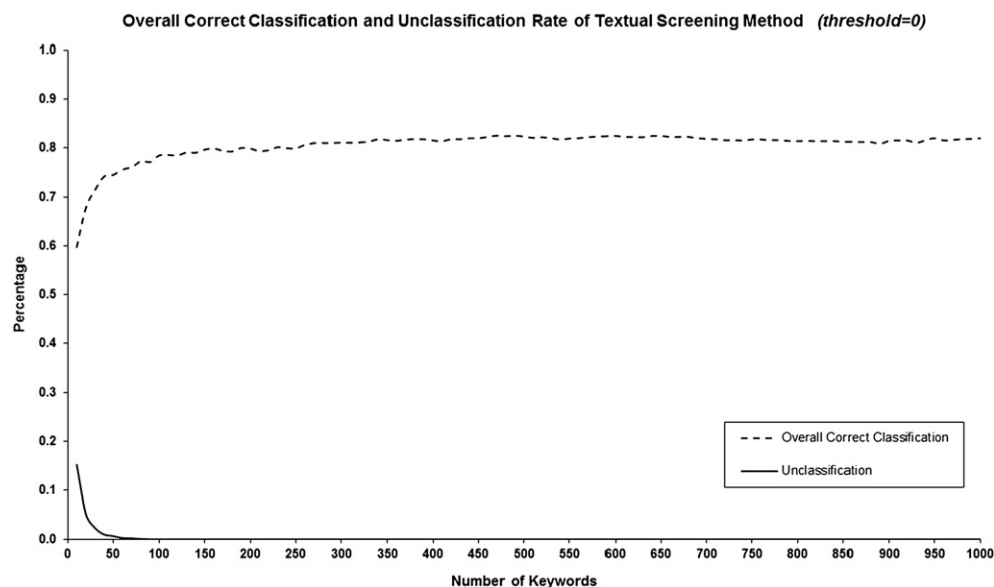


Fig. 2. The overall correct classification (accuracy) and unclassification rate of the textual screening method with an increasing number of keywords from 10 to 1000 using a threshold of 0. Note. All the values are the averaged results from 15-fold cross validation.

alternative machine learning classifiers commonly used in the text mining, such as Decision Tree, Naïve Classifier, Neural Networks and Support Vector Machines (SVM). The Decision Tree (also named as Classification Tree) is one of the most comprehensive models to visually track the path in classification, but may meet the problem of overfitting when the tree grows too big based on the accidental properties of the training set (Kotsiantis, 2007). Thus, some advanced decision tree models (e.g., algorithm C4.5; Quinlan, 1993) are usually built by first growing a large tree and then pruning it back to a reasonable size. The main advantage of Naïve Bayes is the efficient computation due to the simplicity of assumption on word independency (Huang et al., 2008). However, this assumption is not always true in practice, which might lead to a biased classification result. The Neural Networks is an attempt to come up with an alternative framework for the conceptualization and acquisition of hierarchical structure in language (Manning and Schütze, 1999). The SVM is usually used in binary classification that seeks a hyperplane in the feature space that maximizes the margin between the two sets of instances (Cortes and Vapnik, 1995). Generally, the SVM and Neural Networks tend to perform better when dealing with multidimensions and continuous features (Kotsiantis, 2007). For these two models, a large sample size is required in order to achieve its maximum prediction accuracy whereas Naïve Bayes may need a relatively small dataset (Kotsiantis, 2007). To further evaluate the performance of our model, a comparative study between the product score model and alternative classifiers would be a next step in our research.

Besides the positive results, there are some limitations that also merit discussion. First, the individuals collected in the present sample were survivors from mixed trauma, including child abuse, sexual abuse, traffic accident, war, domestic violence, death of beloved, robbery, and fire. Given concerns on the limited sample size of each trauma category, we developed a common text classification model for the population with mixed trauma, instead of generating separate models for each trauma type.

Secondly, the current text classification model was developed based on unigrams – the units of analysis are individual words – but the interactions between words were not taken into consideration at this moment. This leads a vague interpretation in some extracted keywords. For example, the word “get” was extracted as a robust classifier to distinguish the NONPTSD from the PTSD patients. This word was generally used within a combination in daily expressions, such as “get happy”, “get up”, and “get depressed”, etc., but seldom used alone. We could infer that the word “get” more or less indicated a changing tendency of the writer's emotion, but such explanations need more substantiation.

Although the performance metrics demonstrated that the text classifier model was valid and reliable, they are still expected to improve if the multigrams – interactions between word sequences – are added to the text analysis. Conway et al. (2009) proved the effectiveness of a mixture of n -grams (i.e., a combination of unigrams and multigrams) used in the feature selection process in a text classification study regarding disease out-break reports. In the comparison of feature representations in conjunction with Naïve Bayes classification algorithm, they found that the classification sensitivity was 13% higher when using the mixture-based word classifiers (89%) than relying on a single type of unigrams (76%). Of the 100 most discriminative features, 50% were unigrams, 37% were bigrams, 8% were trigrams and 5% were semantic tags. Therefore, we would expect that the text classifier model might be beneficiary from the adding of multigrams in the future studies.

In conclusion, the present study concerns a new development of text classification model for PTSD screening based on verbal features from self narratives. This study demonstrated that the text mining technique was very promising in screening and diagnostic process for mental diseases. The main finding is that there was a good agreement between the textual screening procedure and clinical interview

in identifying the presence and absence of PTSD. The textual screening method yielded a higher sensitivity and positive prediction power than the itemized screening instruments in general. The results also suggested that the textual screening method performed well in excluding the NONPTSD individuals from further tests, but the structured interviews or other intensive tests might be needed in investigations during the second phase to make definitive diagnoses on the residual subgroup. In addition, with the help of text mining techniques, the verbal features in self narratives written by trauma survivors were also extracted for further understanding of the construct of PTSD. The textual assessment method developed in this study will be applicable to researches with similar background and makeup. Similar as the application in PTSD screening, the text classification model can be utilized in a broader scope of qualitative studies within the field of education, psychology, sociology in both research settings and practical use. Besides, new applications of text mining techniques, for instance, speech recognition where patients' addressing can be automatically transferred into written forms, would bring extra benefits for both patients and psychiatrists. It might be interesting to make intensive exploration on this new technique, challenge other researchers to improve our method, and to integrate it in an operational intake procedure in the further research.

Acknowledgement

The textual assessment for PTSD diagnosis research project was supported and sponsored by Stichting Achmea Slachtofferhulp Samenleving, the Netherlands.

Reference

- American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, 4th ed. The American Psychiatric Association, Washington, DC.
- Baldessarini, R.J., Finkelstein, S., Arana, G.W., 1983. The predictive power of diagnostic-tests and the effect of prevalence of illness. *Archives of General Psychiatry* 40 (5), 569–573.
- Batten, S.V., Follette, V.M., Hall, M.L.R., Palm, K.M., 2002. Physical and psychological effects of written disclosure among sexual abuse survivors. *Behavior Therapy* 33 (1), 107–122.
- Blake, D.D., Weathers, F.W., Nagy, L.M., Kaloupek, D.G., Gusman, F.D., Charney, D.S., Keane, T.M., 1995. The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress* 8 (1), 75–90.
- Brebion, G., Ohlsen, R.I., Pilowsky, L.S., David, A.S., 2011. Serial and semantic encoding of lists of words in schizophrenia patients with visual hallucinations. *Psychiatry Research* 186 (1), 5–10.
- Brewin, C.R., 2005. Systematic review of screening instruments for adults at risk of PTSD. *Journal of Traumatic Stress* 18 (1), 53–62.
- Carlier, I.V.E., Lamberts, R.D., Van Uchelen, A.J., Gersons, B.P.R., 1998. Clinical utility of a brief diagnostic test for posttraumatic stress disorder. *Psychosomatic Medicine* 60 (1), 42–47.
- Carlson, E.B., 2001. Psychometric study of a brief screen for PTSD: assessing the impact of multiple traumatic events. *Assessment* 8 (4), 431–441.
- Chung, C.K., Pennebaker, J.W., 2008. Revealing dimensions of thinking in open-ended self-descriptions: an automated meaning extraction method for natural language. *Journal of Research in Personality* 42 (1), 96–132.
- Coffey, S.F., Gudmundsdottir, B., Beck, J.G., Palyo, S.A., Miller, L., 2006. Screening for PTSD in motor vehicle accident survivors using the PSS-SR and IES. *Journal of Traumatic Stress* 19 (1), 119–128.
- Conway, M., Doan, S., Kawazoe, A., Collier, N., 2009. Classifying disease outbreak reports using n -grams and semantic features. *International Journal of Medical Informatics* 78 (12), 47–58.
- Cortes, C., Vapnik, V., 1995. Support vector networks. *Machine Learning* 20, 273–297.
- Davidson, J.R.T., Book, S.W., Colket, J.T., Tupler, L.A., Roth, S., David, D., Hertzberg, M., Mellman, T., Beckham, J.C., Smith, R.D., Davison, R.M., Katz, R., Feldman, M.E., 1997. Assessment of a new self-rating scale for posttraumatic stress disorder. *Psychological Medicine* 27 (1), 153–160.
- Derogati, L., Lipman, R.S., Rickels, K., Uhlenhuth, E.H., Covi, L., 1974. Hopkins symptom checklist (Hscl) – self-report symptom inventory. *Behavioral Science* 19 (1), 1–15.
- Feldman, R., Sanger, J., 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. The Cambridge University Press, Cambridge.
- First, M.B., 1997. *User's Guide for the Structured Clinical Interview for DSM-IV Axis II Personality Disorders: SCID-II*. The American Psychiatric Press, Washington, DC.
- Foa, E.B., Feske, U., Murdock, T.B., Kozak, M.J., Mccarthy, P.R., 1991. Processing of threat-related information in rape victims. *Journal of Abnormal Psychology* 100 (2), 156–162.
- Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* 3, 1289–1305.

- Franklin, C.L., Thompson, K.E., 2005. Response style and posttraumatic stress disorder (PTSD): a review. *Journal of Trauma & Dissociation* 6 (3), 105–123.
- Gara, M.A., Vega, W.A., Lesser, I., Escamilla, M., Lawson, W.B., Wilson, D.R., Fleck, D.E., Strakowski, S.M., 2010. The role of complex emotions in inconsistent diagnoses of schizophrenia. *The Journal of Nervous and Mental Disease* 198 (9), 609–613.
- Gottschalk, L.A., Gleser, G.C., 1969. *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. The University of California Press, Berkeley.
- Gray, M.J., Lombardo, T.W., 2001. Complexity of trauma narratives as an index of fragmented memory in PTSD: a critical analysis. *Applied Cognitive Psychology* 15 (7), 171–186.
- Horowitz, M., Wilner, N., Alvarez, W., 1979. Impact of event scale — measure of subjective stress. *Psychosomatic Medicine* 41 (3), 209–218.
- Huang, W., Zhao, Y., Yang, S., Lu, Y., 2008. Analysis of the user behavior and opinion classification based on the BBS. *Applied Mathematics and Computation* 205, 668–676.
- Jurafsky, D., Martin, J.H., 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. The Pearson Prentice Hall, Upper Saddle River.
- Kessel, J.B., Zimmerman, M., 1993. Reporting errors in studies of the diagnostic performance of self-administered questionnaires: extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment* 5 (4), 395–399.
- Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. *Informatica* 31, 249–268.
- Manning, C.D., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge.
- Mendoza, R., Cabral-Calderin, Y., Dominguez, M., Garcia, A., Borrego, M., Caballero, A., Guerra, S., Reyes, M.M., 2011. Impairment of emotional expression recognition in schizophrenia: a Cuban familial association study. *Psychiatry Research* 185 (1–2), 44–48.
- Mollica, R.F., Caspiyavin, Y., Bollini, P., Truong, T., Tor, S., Lavelle, J., 1992. The Harvard Trauma Questionnaire — validating a cross-cultural instrument for measuring torture, trauma, and posttraumatic-stress-disorder in Indo-Chinese refugees. *The Journal of Nervous and Mental Disease* 180 (2), 111–116.
- Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, A.W.V., Beaulieu, M., 2001. A method based on chi-square test for document classification. Paper Presented at the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 21). Morristown, New Jersey.
- Orsillo, S.M., Batten, S.V., Plumb, J.C., Luterek, J.A., Roessner, B.M., 2004. An experimental study of emotional responding in women with posttraumatic stress disorder related to interpersonal. *Journal of Traumatic Stress* 17 (3), 241–248.
- Porter, M.F., 1980. An algorithm for suffix stripping. *Program-Automated Library and Information Systems* 14 (3), 130–137.
- Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco.
- Rosenberg, S.D., Tucker, G.J., 1979. Verbal-behavior and schizophrenia — semantic dimension. *Archives of General Psychiatry* 36 (12), 1331–1337.
- Shrout, P.E., Skodol, A.E., Dohrenwend, B.P., 1986. A two-stage approach for case identification and diagnosis, first stage instruments. In: Barrett, J.E., Rose, R.M. (Eds.), *Mental Disorders in Community: Progress and Challenge*. The Guilford Press, New York, pp. 286–303.
- Silove, D., Manicavasagar, V., Mollica, R., Thai, M., Khiek, D., Lavelle, J., Tor, S., 2007. Screening for depression and PTSD in a Cambodian population unaffected by war — comparing the Hopkins Symptom Checklist and Harvard Trauma Questionnaire with the structured clinical interview. *The Journal of Nervous and Mental Disease* 195 (2), 152–157.
- Smyth, J.M., 1998. Written emotional expression: effect sizes, outcome types and moderating variables. *Journal of Consulting and Clinical Psychology* 66 (1), 174–184.
- Trusko, B., Rosenbloom, S.T., Montella, D., Jackson, J.C., Fitzhenry, F., Brown, S.H., Elkin, P.L., Fielstein, E., Kotter, K., Tuttle, M., Iannelli, R.J., Speroff, T., 2010. Are posttraumatic stress disorder mental health terms found in SNOMED-CT medical terminology? *Journal of Traumatic Stress* 23 (6), 794–801.
- Weathers, F.W., Huska, J.A., Keane, T.M., 1991. *The PTSD Checklist — Civilian Version (PCL-C)*. The National Center for PTSD, Boston Veterans Affairs Medical Center, Boston.
- Yang, Y., Pedersen, J., 1997. A comparative study on feature selection in text categorization. Paper Presented at the 14th International Conference on Machine Learning (ICML97). Nashville, Tennessee.
- Zipf, G.K., 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. The Addison-Wesley Press, Cambridge.