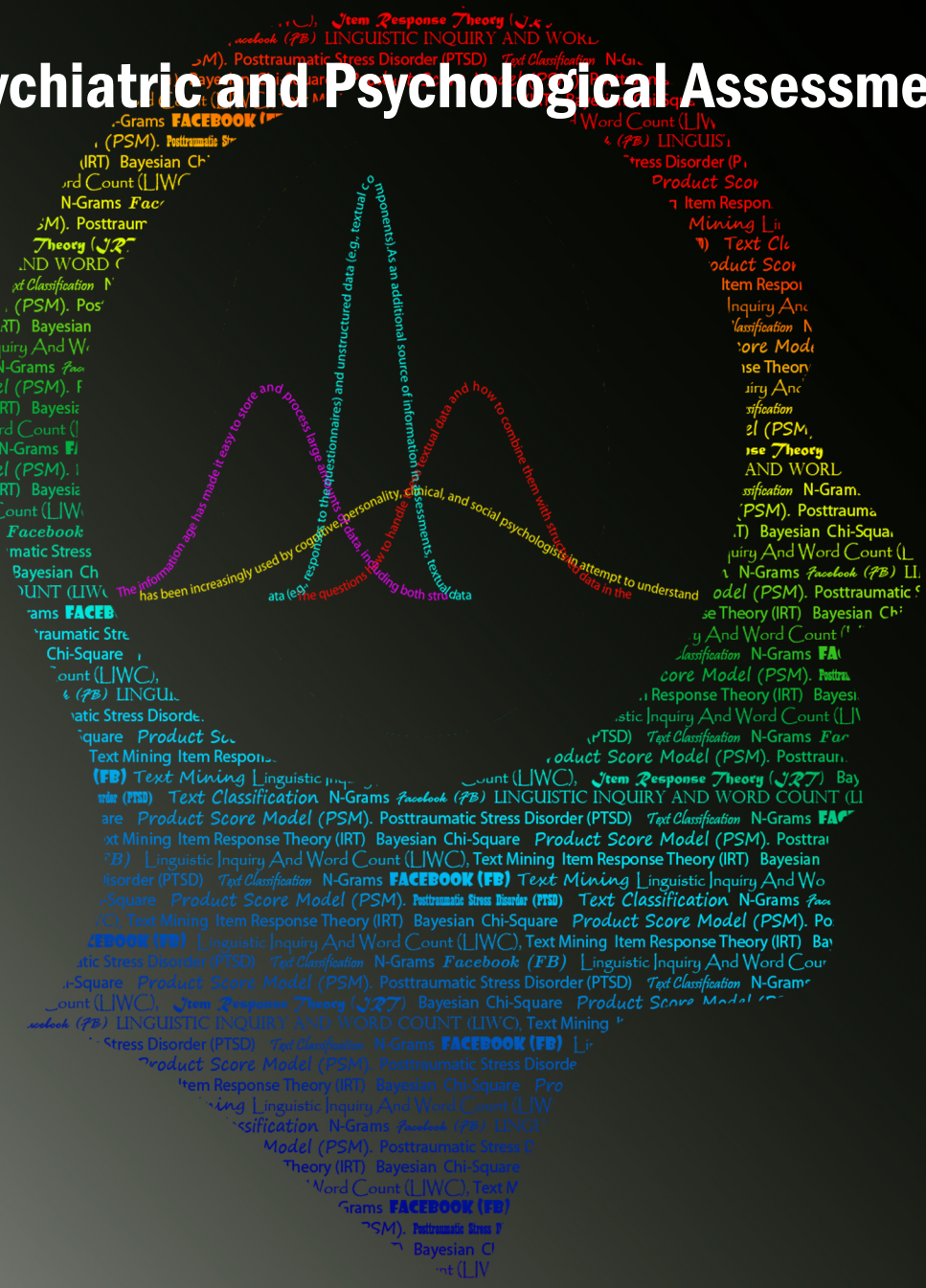
Text Mining and IRT for Psychiatric and Psychological Assessment



Qiwei He

TEXT MINING AND IRT FOR PSYCHIATRIC AND PSYCHOLOGICAL ASSESSMENT

Qiwei He

Graduation Committee

Chairman	Prof. Dr. K. I. van Oudenhoven-van der Zee
Promotor	Prof. Dr. C. A. W. Glas
	Prof. Dr. T. de Vries
Assistant promotor	Dr. Ir. B. P. Veldkamp
Members	Prof. Dr. K. Schreurs
	Prof. Dr. M. Ijzerman
	Prof. Dr. H. J. van den Herik
	Prof. Dr. R. Meijer
	Dr. E. Vermetten
	Prof. Dr. P. A. M. Vierhout

Qiwei He
Text Mining and IRT for Psychiatric and Psychological Assessment
Ph.D. thesis, University of Twente, Enschede, the Netherlands
ISBN: 978-90-365-0056-2
Doi: 10.3990/1.9789036500562
Printed by Ipskamp Drukkers, B. V., Enschede
Cover designed by Weihua Zhou

This research was partially supported by the Stichting Achmea Slachtofferhulp Samenleving, the Netherlands.

TEXT MINING AND IRT FOR PSYCHIATRIC AND PSYCHOLOGICAL ASSESSMENT

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday, October 3rd, 2013 at 14:45

by

Qiwei He
born on March 28th, 1979
in Beijing, China

This dissertation has been approved by the promotors:
Prof. Dr. C. A. W. Glas
Prof. Dr. T. de Vries
Dr. Ir. B. P. Veldkamp

“Language is a process of free creation; its laws and principles are fixed, but the manner in which the principles of generation are used is free and infinitely varied. Even the interpretation and use of words involves a process of free creation.”

— *Noam Chomsky*

Acknowledgements

This thesis is the result of almost five enjoyable years spent in the department of Research Methodology, Measurement, and Data Analysis (OMD) at the University of Twente. I am deeply appreciative of the many individuals who have supported my work and continually encouraged me through the writing of this dissertation. Without their time, attention, encouragement, thoughtful feedback, and patience, I would not have been able to see it through.

Above all, I would like to thank my promoters. Cees Glas, I have learned a great deal from his unique perspective on research, his sharp insight on almost any issue, and his personal integrity and expectations of excellence. I am amazed at his knowledge of mathematical statistics and enthusiasm of psychometrics. Theo de Vries, I am grateful for his inspirational and timely advice and constant encouragement over the last several years. He has been a great advisor for me outside of the academic world as well.

I am very fortunate to have had Bernard Veldkamp as my daily supervisor. He has always been ready with brilliant ideas, honest advice and encouraging words whenever I needed them. I enjoy working with him on the novel text mining techniques and all other areas. He has always been patient when explaining the nuances of English words in writing, and has shared with me many witty jokes, metaphors, and lessons on how to strive for work-life balance. I am looking forward to working with him more in the future.

I would also like to thank the other members of my graduation committee, Prof. dr. K. Schreurs, Prof. dr. M. Ijzerman, Prof. dr. H. J. van den Herik, Prof. dr. R. Meijer, Dr. E. Vermetten, and Prof. dr. P. A. M. Vierhout, for taking the time to read my thesis.

I am very grateful to Maaike van Groen, my first Dutch classmate and friend, and Hanneke Geerlings, my excellent roommate for the last year, for assisting me during the defense as paranymphs.

I am indebted to other (former) members in the OMD: Stephanie van den Berg for guiding me to use Winbugs and forcing me to keep practice of the Dutch language; Jean-Paul Fox for teaching me hierarchical IRT models and multilevel linear regressions; Hans Vos and Wim Tielen for giving me opportunities to assist teaching in the pre-master and master courses related to IRT and data analysis; Theo Eggen for teaching me a broad range of research methodologies in computerized adaptive testing

and inviting me to join all the activities in the Research Center for Examination and Certification (RCEC); Job van de Palen for advising me the research lines in the medical field. Also, I want to thank Rinke, Josine, Marianna, Muirne, Caroline and Connie, for their encouragement and accompaniment. Special thanks go to our secretaries, Birgit and Lorette for their supports and patience for all the daily issues.

I'd also like to thank other faculty members and colleagues outside of OMD. In particular, I want to thank Rieks op den Akker and Mariet Theune from the Department of Human Machine Interaction (HMI) for introducing me into the fascinating world of text mining and machine learning. My appreciation also goes to Saskia, Maarten, and all other colleagues from RCEC for their sharing and fun.

Life would not have been as colorful without the many good friends I met in the Netherlands. I would like to extend my thanks to Jie Zhao, Yali Zhang, Lixian Xu, Hongping Luo, Yang Cui, Haishan Cao, Meng Wang, Tingting Luan, Shi Pu, Lei Bai, Xiaofeng Sui, Yuying Gao, and many others. Special thanks go to Sinie Jissink, who is my sweet "grandma" in the Netherlands. She has always made me feel at home in a country outside of China. Thanks for teaching me to bake delicious Dutch cookies and warmly inviting my parents (in-law) for visits.

Last but not least, I want to thank my loving and caring family. Thanks to my father Peitong He and my mother Jiayun Deng for teaching me to be curious and sincere to life, and for always being there for me. Thanks to my in-laws, Zexing Zhou and Xiaoyan Hua, for giving me constant love and encouragement and for treating me as their own daughter. Words cannot express my appreciation and love for my son, Albert. Thanks for being healthy and eating well. To my husband, Weihua Zhou, thank you for designing the excellent book cover, always patiently standing by, supporting me in my academic pursuits and for the wonderful life that we share together.

Qiwei He
Enschede, the Netherlands
August 30th, 2013

Contents

Chapter 1

Introduction

The information age has made it easy to store and process large amounts of data, including both structured data (e.g., responses to the questionnaires) and unstructured data (e.g., natural language or prose). Structured data is defined as an organized database where specific information is stored based on a methodology of columns and rows. In contrast, unstructured data, mainly textual data, has no identifiable structure, neither having a predefined data model, nor fitting well into relational patterns. In addition to structured information in assessments, textual data has been increasingly used as a medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings (Tausczik & Pennebaker, 2010).

1.1 Psychiatric and Psychological Assessment

A psychiatric or psychological assessment is a process of gathering information about a person within a psychiatric or psychological service with the purpose of establishing a clinic diagnosis or making a cognitive evaluation (Meyer et al., 2001). The assessment can be supplemented by the use of ample data, including records from direct observations, responses from specific psychological tests, narratives, demographic information and so forth. In clinical settings, psychiatric assessment is typically a multidisciplinary process involving psychologists, occupational therapists, social workers, nurses and psychometricians.

The initiative of the current research project originates from the construction of a screening test for posttraumatic stress disorder (PTSD) by using the lexical features in patients' self-narratives. The objective of such a test is to detect PTSD in an early

phase for a timely treatment. The test can, for example, be administered either via Internet or via offices of general practitioners. Compared to itemized questionnaires, self-narratives provide patients with opportunities to express themselves freely, and they are easier to be interpreted by clinicians. That is, patients may describe the traumatic events and symptoms in their own styles without limitations set by the item options. From texts, clinicians may understand the contents straightforwardly without having to consult a psychometrician for interpretation of the scale parameters.

The context of the present study is related to two developments in research regarding the use of language in psychiatric and psychological assessment. First, language is the most common and reliable way for people to translate their internal thoughts and emotions into a form (Tausczik & Pennebaker, 2010). Numerous evidence has suggested that the words and expressions coming from patients' self-narratives are quite informative for early detection of epidemics and mental diseases (e.g., Gottschalk & Gleser, 1969; Rosenberg & Tucker, 1979; Smyth, 1998; Franklin & Thompson, 2005). With the increasing application of text-based measures, the automated identification of patients based on their self-narratives are, therefore, expected as a promising step toward an effective screening and diagnosis process. Secondly, but beyond the scope of the present study, writing of self-narratives has proved to be an effective treatment approach to relieve PTSD patients' stress (e.g., Batten, Follette, Hall, & Palm, 2002; Orsillo, Batten, Plumb, Luterek, & Roessner, 2004) and an impactful method to recall memory for psycholinguistic studies (e.g., Foa, Feske, Murdock, Kozak, & Mccarthy,1991; Brebion, Ohlsen, Pilowsky, & David, 2011). It is expected to be beneficiary also for the PTSD screening process.

In the next sections, the research questions and the research methodologies applied in this thesis will be discussed. The chapter will end with a structured overview of the remaining chapters.

1.2 Research Questions

Information technology demonstrated a number of important breakthroughs in handling unstructured textual data during the past decades. A promising technique for handling these data is text mining, which exploits information retrieval, information extraction, and corpus-based computational linguistics based on natural language processing (NLP). Although this technique has been generally used in information retrieval systems, for example in search engines like "Google" or in customer

investigations by companies like “Amazon”, its application is relatively new in psychiatric and psychological assessment.

Some challenges have to be confronted in this area. First, the size of the datasets used in psychiatric or psychological assessment is relatively small in comparison to the datasets employed by information retrieval systems. Therefore, the machine learning algorithms which perform well in a big information retrieval systems (like e.g., Google) might not yield a comparable performance. Secondly, the problems of identifying textual features that can robustly classify texts (e.g., patients’ self-narratives) into different categories (e.g., PTSD and NONPTSD) for the purpose of psychiatric or psychological assessments have been rarely addressed in the literature. Thirdly, itemized tests or questionnaires are predominantly used in psychiatric and psychological assessments. This leads to the question how the outcomes of text mining relate to the outcomes of these tests. Last but not least, both textual analysis and item-based measures have their own limitations. To combine these two methods in a systematic framework might be an approach to strengthen the benefits from both.

These considerations lead to the following two main research questions:

(1) How can we apply text mining to narratives collected in the framework of psychiatric and psychological assessment to make classification decisions; and

(2) How can we simultaneously model the outcome of text mining and the IRT-based outcome of responses to questionnaires to validate the text mining procedure and enhance the quality of the measurement and classification procedure.

The issue whether a text classification model can be developed to satisfy the specific requirements in psychiatric and psychological assessments is the first focus in this thesis. The performance of the new model is compared to existing models for text classification and compared to human raters (i.e., psychiatrists and psychologists). (Chapter 2, 3, 4, and 7 address these issues.)

Secondly, the question what textual features enable the application of text classification in a number of very different psychology or psychiatry related contexts, such as the context of reporting life stories of undergraduate students (Chapter 2), the context of PTSD detection (Chapter 3-4), and the context of assessment of self-monitoring skills via Facebook posts (Chapter 7), was studied as well.

Thirdly, the results from the text classification model could be validated using the outcomes of existing questionnaires. To follow this approach, the questionnaires are required to be validated themselves. This leads to a number of additional questions:

How can we evaluate the validity of these itemized instruments? How can we validate the responses to items, especially in a self-reporting (internet) environment? (Chapter 5 and 7 address these questions by means of item response modeling.)

Finally, the question how the textual assessment and item based measures can be combined in one systematic framework is explored. Whether the adding of textual assessment increases the accuracy of assessment is also examined in this thesis. (This issue is the focus of Chapter 6)

1.3 Research Methodologies

To answer the research questions, three types of research methodologies are used in this thesis: text mining techniques for textual analysis (Feldman & Sanger, 2007), item response theory (IRT; Rasch, 1960; Lord, 1980) for estimating latent abilities based on responses to items in structured tests, and a combination of text mining and item response modeling via a Bayesian approach.

1.3.1 Text Mining

Text mining is one of the new research areas originating from statistical NLP in Computer Science. It can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of pronounced patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections (Feldman & Sanger, 2007).

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 & Schütze, 1999). Feature extraction and machine learning are the two essential sections in text classification, playing influential roles in classification efficiency. During feature extraction, textual components are transformed into structured data and labeled with one or more classes. Based on these encoded data, the most discriminative lexical features are extracted by using computational statistic models, such as the chi-square selection algorithm (Oakes, Gaizauskas, Fowkes, Jonsson, & Beaulieu, 2001) and

likelihood ratio functions (Dunning, 1993). In machine learning, documents are allocated into the most likely classes by applying machine learning algorithms such as decision trees (DT), naïve Bayes (NB), support vector machines (SVM), and the *K* nearest neighbor model (KNN). Although many machine learning classifiers have been tested efficiently in text classification, alternative models are still being explored to further improve text classification performance and accelerate the speed of word processing (see more in Vapnik, 1998; Duda, Hart, & Stork, 2001; Kotsiantis, 2007).

Text classification is generally divided into two categories: supervised text classification where the actual status (label) of the training data has been identified ("supervised") and unsupervised text classification where the actual status of the data has not been identified, namely, any labels associated with objects are obtained solely from the data (e.g., clustering). This thesis mainly focuses on the supervised text classification approach. Chapter 2, 3, 4, 6 and 7 discuss the topic related to text mining and supervised text classification.

1.3.2 Item Response Theory

IRT has been generally applied in educational, psychological and psychiatric assessments for almost three decades. IRT models are based on the idea that psychological constructs are latent, that is, not directly observable, and that knowledge about these constructs can only be obtained through manifest responses of persons to a set of items (e.g., Embretson & Reise, 2000; Sijtsma & Molenaar, 2002). In contrast to the traditional sum score based classical test theory, the IRT models measure the latent trait at item level, which makes "tailored" testing possible in such a way that the individual's ability is assessed more flexibly by presenting a smaller selection of items tailored to his latent trait level.

Commonly applied unidimensional IRT models in psychiatric and psychological assessments are the Rasch model or one parameter logistic model (1PL; Rasch,1960) and the two parameter logistic model (2PL; Birnbaum, 1968). In the unidimensional 2PL model, that is, the probability of a score in category "yes" ($X_{ni}=1$) of item *i* is given by the item response function

$$P(X_{ni}=1\,|\,\theta_n)=\frac{\exp\left[\alpha_i(\theta_n-\beta_i)\right]}{1+\exp\left[\alpha_i(\theta_n-\beta_i)\right]}, \qquad (1.1)$$

where θ_n is the latent trait level of person n, β_i is an item location parameter representing the difficulty level of each item, and α_i is an item discrimination parameter indicating the extent to which the item response is related to the latent scale. A special version of the model is the Rasch model where the item discrimination parameter α_i is fixed as 1. Other popular models in psychological assessments are the polytomous IRT models (e.g., Ostini & Nering, 2005), but they are not the focus of this study.

Applications of IRT models are generally related to validation of measurement instruments, test equating, differential item functioning (DIF; Camilli & Shepard, 1994), data validation and computerized adaptive testing (CAT; Wainer et al., 1990; van der Linden & Glas, 2000). This thesis mainly focuses on two applications of IRT, (1) detection of DIF in Chapter 5 and (2) validation of the Internet data in Chapter 7. DIF examines the measurement equivalence across groups, that is, it explores the relationship between item response and a group variable defined by demographic attributes (e.g., gender, race), conditional on a measure of an underlying construct such as PTSD severity (Teresi & Fleishman, 2007). The data validation is to examine whether there is significant difference between the observations and the predictions estimated by the IRT model, namely, to evaluate to what extent an IRT model fits an instrument in a particular population (Glas & Dagohoy, 2007). The detailed explanation of IRT models and their applications can be found in the corresponding chapters (Chapter 5 and Chapter 7).

1.3.3 Combination of Text Mining and Item Response Modeling

Combining the text mining and item response modeling is a new approach to handle both structured and unstructured data in one systematic framework. See, for example, a schematic representation of the PTSD screening process in Figure 1.1.

As is shown in Figure 1.1, the PTSD screening process consists of two phases, text classification and IRT based analysis. The final output is a posterior estimate of the latent PTSD level of each individual. Bayesian methods are especially useful for the estimation of such a hierarchical structure. Following the arrows in the flowchart, the textual analysis and item response modeling are combined in a Bayesian framework, where the text score of each self-narrative obtained in the text classification is used as an informative prior and the IRT model is used to formulate

the likelihood of the responses. The posterior distribution of the latent PTSD level is proportional to the product of prior and the likelihood, that is

$$P(\theta \mid x, y) \propto p(x \mid \theta, \alpha, \beta) g(\theta \mid y), \tag{1.2}$$

where y is the text score for each individual, $g(\theta \mid y)$ is the population model given the covariate of textual assessments, α and β are the fixed item discrimination and difficulty parameters of the questionnaire, $p(x \mid \theta, \alpha, \beta)$ is the likelihood function.

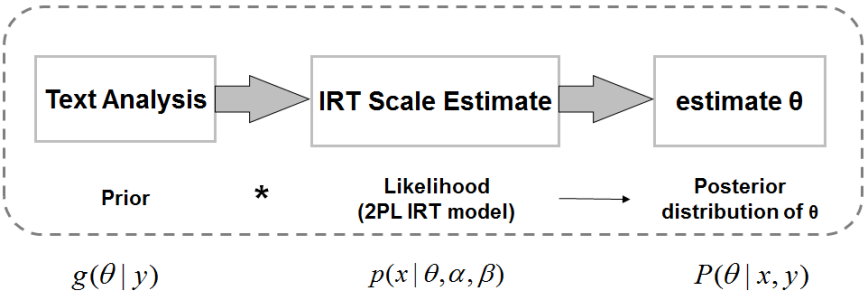


Fig. 1.1. A Bayesian framework that combines textual analysis and IRT scale estimates

1.4 Structure of the Thesis

In this introductory chapter, two main research questions have been formulated. Three research methodologies (text mining, IRT and the combination of text mining and IRT) to handle numerical and textual data have been discussed. The focus of the rest of this thesis will be on applications of these three research methodologies: text mining (Chapter 2–4), item response theory (Chapter 5), and combination of text mining and item response modeling (Chapter 6–7). The chapters follow a logical order, but have been written to be self contained. Hence, overlap could not be avoided.

Chapter 2 describes a general procedure for text classification using text mining techniques and presents an alternative machine learning algorithm for text classification, named the product score model (PSM). To evaluate the performance of PSM, a comparative study is conducted between PSM and two commonly used classification models, decision trees and naïve Bayes in conjunction with the bag-of-

words representation (unigrams). An application of these three models is illustrated for real textual data related to undergraduates' life stories.

A textual assessment method based on PSM is developed in Chapter 3 with the aim to screen the PTSD patients using lexical features in their self-narratives. Using 300 self-narratives collected online, the most discriminative keywords are extracted using the chi-square algorithm. In order to examine whether the performance of the textual screening method is comparable to that of the item-based screening instruments, the mean performance of thirteen commonly used screening instruments is used as the baseline.

An extension of the data representation model from unigrams to *n*-grams, where the occurrences of sets of *n* consecutive words are counted, is further discussed in Chapter 4. Based on the same sample used in the preceding chapter, the PSM together with decision trees and naïve Bayes are applied in conjunction with five representation models, unigrams, bigrams, trigrams, a combination of uni- and bigrams and a mixture of *n*-grams to identify the PTSD patients.

Chapter 5 explores the generalizability of DSM-IV diagnostic criteria for PTSD to various subpopulations by using IRT techniques. Besides identifying differential symptom functioning related to various background variables such as gender, marital status and educational level, this study also emphasizes the importance of evaluating the impact of DIF on population inferences as made in health surveys and clinical trials, and on the diagnosis of individual patients. The DIF items are detected by using an item oriented Lagrange multiplier (LM) item fit statistic (Glas, 1998, 1999). The impact of DIF is evaluated by introducing group-specific parameters and comparing the results via two approaches, one approach that measures the differences between group distributions and the other that measures the change in accuracy of diagnosis of individuals.

A combination of automated textual assessment of patients' self-narratives and structured interviews is proposed in Chapter 6. Text mining and item response modeling are used to analyze the patients' writings and responses to standardized questionnaires, respectively. The whole procedure is combined in a Bayesian framework where the textual assessment functions as an informative prior for the estimation of PTSD latent trait. The performance of combination model is compared with that obtained by text classification and IRT alone. The item parameters that has been calibrated earlier is fixed in this study, which makes the θ-estimates resulting from both of the two approaches, i.e., with and without textual priors, on a common scale and thus can be combined.

Chapter 7 extends the application of text classification from psychiatric assessments to psychological assessments based on the Internet data. Besides emphasizing the importance of validating data collected from the Internet, this chapter explores the relationship between self-monitoring skills and textual posts on the Facebook Wall and examines to what extent the lexical features of the posts can predict individual's self-monitoring skills. The textual analysis is conducted via both structured and unstructured approach. A commonly used psychology-related linguistic software package, called the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2001) is introduced in the structured textual analysis. To explore the links between the structured and unstructured textual analysis, the keywords extracted by the text mining techniques are also mapped onto the framework of LIWC.

This thesis concludes with answering to the research questions. A discussion is included on the significance of the results and some suggestions for future research are also given.

Chapter 2

Classifying Unstructured Textual Data Using the Product Score Model: An Alternative Text Mining Algorithm

Abstract

Unstructured textual data such as students' essays and life narratives can provide helpful information in educational and psychological measurement, but often contain irregularities and ambiguities, which creates difficulties in analysis. Text mining techniques that seek to extract useful information from textual data sources through identifying interesting patterns are promising. This chapter describes the general procedures of text classification using text mining and presents an alternative machine learning algorithm for text classification, named the product score model (PSM). Using the bag-of-words representation (single words), we conduct a comparative study between PSM and two commonly used classification models, decision tree and naïve Bayes. An application of these three models is illustrated for real textual data of students' life stories. The results show that the PSM performs the most efficiently and stably in classifying text. Implications of these results for the PSM are further discussed and recommendations about its use are given.

Keywords: text classification, text mining, product score model, unstructured data

This chapter has been published as: He, Q., & Veldkamp, B. P. (2012). Classifying unstructured textual data using the Product Score Model: An alternative text mining algorithm. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 47-62). Enschede, the Netherlands: RCEC.

2.1 Text Classification

Language is magic that diversifies our lives. The way individuals talk and write provides a window into their emotional and cognitive worlds. Yet despite the interesting attributes of textual data, analyzing them is not easy. One of the major reasons is that textual data are generally more diverse than numerical data and are often unstructured, neither having a predefined data model nor fitting well into relational patterns. The irregularities and ambiguities make it even harder to classify textual data compared with structured data stored in field form in databases. Thus, to address the challenge of exploiting textual information, new methods need to be developed.

The development of information technology demonstrated breakthroughs in handling unstructured textual data during the past decade. A promising technique is text mining, which exploits information retrieval, information extraction, and corpus-based computational linguistics. Analogous to data mining, text mining seeks to extract useful information from textual data sources by identifying interesting patterns. However, a preprocessing step is required to add transforming unstructured data stored in texts into a more explicitly structured intermediate format (Feldman & Sanger, 2007).

Text mining techniques are used, for example, for text classification, where textual objects from a universe are assigned to two or more classes (Manning & Schütze, 1999). Common applications in educational measurement classify students' essays into different grade levels with automated scoring algorithms, e.g., Project Essay Grade (PEG; Page, 2003) and automated scoring of open answer questions, e.g., E-raters (Burstein, 2003). Feature extraction and machine learning are the two essential sections in text classification, playing influential roles in classification efficiency. During feature extraction, textual components are transformed into structured data and labeled with one or more classes. Based on these encoded data, the most discriminative lexical features are extracted by using computational statistic models, such as the chi-square selection algorithm (Oakes et al., 2001) and likelihood ratio functions (Dunning, 1993). In the machine learning section, documents are allocated into the most likely classes by applying machine learning algorithms such as decision trees (DT), naïve Bayes (NB), support vector machines (SVM), and the *K* nearest neighbor model (KNN). Although many machine learning classifiers have been tested efficiently in text classification, new alternative models are still being explored to further improve text classification performance and accelerate the speed of word processing (see more in Vapnik, 1998; Duda et al., 2001; Kotsiantis, 2007).

This chapter briefly describes the general procedure for supervised text classification where the actual status (label) of the training data has been identified ("supervised"), introduces an effective and much used feature extraction model, i.e., the chi-square selection algorithm, and presents an alternative machine learning algorithm for text classification, named the product score model (PSM). To evaluate the PSM performance, a comparative study was conducted between PSM and two standard classification models, DT and NB, based on an example application for real textual data. The research questions focus on (a) whether the PSM performs more efficiently in classifying text compared to the standard models, and (b) whether the PSM maintains a stable and reliable agreement with the human raters' assessment.

2.2 Supervised Text Classification

Supervised text classification is a commonly used approach for textual categorization, which generally involves two phases, a training phase and a testing phase (Jurafsky & Martin, 2009; see Figure 2.1). During training, the most discriminative keywords for determining the class label are extracted. The input for the machine learning algorithm consists of a set of prespecified keywords that may potentially be present in a document and labels classifying each document. The objective of the training phase is to "learn" the relationship between the keywords and the class labels. The testing 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 that were not used during training. In the testing procedure, the keywords extracted from the training are scanned in each new input. Thus, the words that were systematically recognized are fed into the "trained" classifier model, which predicts the most likely label for each new self-narrative. To ensure proper generalization capabilities for the text classification models, a cross-validation procedure is generally applied.

To improve the efficiency of the training and testing procedure, a preprocessing routine is often implemented. This involves screening digital numbers, deducting noninformative "stop words" (e.g., "I", "to"), common punctuation marks (e.g., ".", ":"), and frequently used abbreviations (e.g., "isnt", "Im"), and "stemming" the rest of words, for instance, with the Porter algorithm (Porter, 1980) to remove common morphological endings. For example, the terms "nightmares," "nightmaring," and

"nightmared," though in variant lexical forms, are normalized in an identical stem "nightmar"[1] by removing the suffixes and linguistic rule-based indicators.

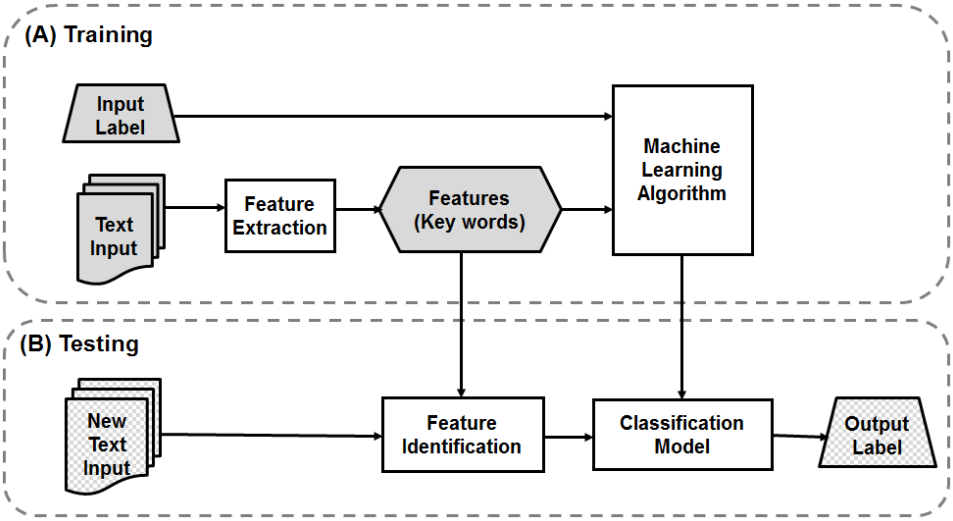


Fig. 2.1. The framework of supervised text classification

2.3 Chi-Square Feature Selection Algorithm

A classifier extraction can be designed to capture salient words or concepts from texts using a feature selection algorithm that compares the frequency of each word type in the text corpus[2] of interest to the frequency of that word type in the whole text corpora (Conway, 2010). Forman (2003) reviewed many feature selection methods for text classification, in which the chi-square selection algorithm (Oakes et al., 2001) was recommended for use due to its high effectiveness in finding robust keywords and testing for the similarity between different corpora. Thus, we briefly introduce this algorithm here and then apply it in the example data.

[1] The stemming algorithm is used to normalize lexical forms of words, which may generate stems without an authentic word meaning, such as "nightmar."

[2] A body of texts is usually called a text corpus. 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.

To apply the chi-square algorithm for feature selection, the K word types in the training set are compiled into an K-by-2 table, schematically shown in Table 2.1. The two columns correspond to the two corpora, C_1 and C_2. Each row corresponds to a particular word i. The number of word occurrences in C_1 and C_2 is indicated by n_i and m_i, respectively. The sum of the word occurrences in each corpus is defined as the corpus length,

$$len(C_1) = \sum_{i=1}^{k} n_i, \quad len(C_2) = \sum_{i=1}^{k} m_i \ . \tag{2.1}$$

Table 2.1. Structuralizing Textual Data in a Binary Classification

	C_1	C_2
Word 1	45	1
Word 2	23	0
⋮	⋮	⋮
Word i	n_i	m_i
⋮	⋮	⋮
Word k	n_k	m_k
Total	$len(C_1)$	$len(C_2)$

Table 2.2. Confusion Matrix for Word i in the 2-by-2 Chi-Square Score Calculation

	C_1	C_2
Word i	n_i	m_i
¬ Word i	$len(C_1) - n_i$	$len(C_2) - m_i$

Each word is then compiled into its own 2-by-2 contingency table as shown in Table 2.2. The values in each cell are called the observed frequencies (O_{ij}). Using the assumption of independence, the expected frequencies (E_{ij}) are computed from the marginal probabilities. 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, as the following formula:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \quad (2.2)$$

To ensure the reliability of the calculation, as Manning and Schütze (1999) suggested, in practice features or words that occur fewer than five times are usually eliminated. However, for a small sample, the number of word occurrences could be even lower, perhaps three times. Based on the chi-square scores, all words are ranked in a descending order, and those standing at the top are extracted as robust classifiers[3]. Further, if the ratio n_i / m_i is larger than the ratio $len(C_1) / len(C_2)$, the word is regarded as more typical of corpus C_1 (as a "positive indicator"); otherwise, it is more typical of corpus C_2 (as a "negative indicator") (Oakes et al., 2001).

2.4 Text Classification Models

Training text classifiers is the procedure where machines "learn" to automatically recognize complex patterns, to distinguish between exemplars based on their different patterns, and to make intelligent predictions on their class. Among various machine learning algorithms, decision trees (C4.5; Quinlan, 1993) and naïve Bayes are two of the most widely used text classification models (see more algorithms in Kotsiantis, 2007).

2.4.1 Decision Trees

A decision tree is a well-known machine learning approach to automatically induce classification trees based on training data sets. In the tree structures, 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. There are numerous methods for finding the feature that best divides the training data such as information gain (Hunt, Marin, & Stone, 1966) and the Gini index (Breiman, 1984).

[3] Since we are interested only in ranking the chi-square score for each word to find the optimal classifier, assessing the significance of the chi-square test is not important in this way.

The objects at each node are split into piles in a way that gives maximum information gain and stopped until they are categorized into a terminate class.

2.4.2 Naïve Bayes

Naïve Bayes is a probabilistic classifier applying Bayes's theorem with strong (naïve) independence assumptions (Lewis, 1998). It is simple but effective in practice (Hand & Yu, 2001). The basic idea is to estimate the conditional probability of the class C given the word vectors $\mathbf{w}$ with the assumption of word independence. Namely,

$$P(C \mid \mathbf{w}) = \frac{p(C)p(w_1 \mid C)p(w_2 \mid C)...p(w_k \mid C)}{p(w_1,...,w_k)} = \frac{p(C)\prod_{i=1}^{k} p(w_i \mid C)}{p(\mathbf{w})}, \quad (2.3)$$

where $p(C)$ is the prior probability of a certain class, and $p(w_i \mid C)$ is the conditional probability of a word occurs in a certain class, which is generally estimated with maximum likelihood. In the binary classification, the two probabilities from categories C_1 and C_2 could be simply compared in a ratio R. That is,

$$R = \frac{P(C_1 \mid \mathbf{w})}{P(C_2 \mid \mathbf{w})} = \frac{p(C_1)\prod_{i=1}^{k} p(w_i \mid C_1)}{p(C_2)\prod_{i=1}^{k} p(w_i \mid C_2)}. \quad (2.4)$$

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

2.4.3 Product Score Model

The product score model (He, Veldkamp, & de Vries, 2012) is an alternative machine learning algorithm, which features in assigning two weights for each keyword (in binary classification) – the probability of the word i occurs in the two separate corpora, U_i and V_i – to indicate to how much of a degree the word can represent the two classes. The weights are calculated by

$$\begin{cases} U_i = (n_i + a) / len(C_1) \\ V_i = (m_i + a) / len(C_2) \end{cases} . \tag{2.5}$$

Note that a smoothing constant a (we use $a = 0.5$ in this study) is added to the word occurrence in Formula (2.5) to account for words that do not occur in the training set, but might occur in new texts. (For more on smoothing rules, see Manning & Schütze, 1999; Jurafsky & Martin, 2009.)

The name product score comes from a product operation to compute scores for each class, i.e., S_1 and S_2, for each input text based on the term weights. That is,

$$\begin{cases} S_1 = P(C_1) \cdot \prod_{i=1}^{k} U_i = P(C_1) \cdot \prod_{i=1}^{k} \left[(n_i + a) / len(C_1) \right] \\ S_2 = P(C_2) \cdot \prod_{i=1}^{k} V_i = P(C_2) \cdot \prod_{i=1}^{k} \left[(m_i + a) / len(C_2) \right] \end{cases} , \tag{2.6}$$

where a is a constant, and $P(C)$ is the prior probability for each category given the total corpora. The classification rule is defined as:

$$\text{choose} \begin{cases} C = 1 & \text{if } \log(S_1 / S_2) > b \\ C = 2 & \text{else} \end{cases} , \tag{2.7}$$

where b is a constant[4].

To avoid mismatches caused by randomness, unclassification rules are also taken into account. As mentioned above, based on the chi-square selection algorithm, the keywords are labeled as two categories, positive indicator and negative indicator. Thus, we define a text as "unclassified" when either one of the following conditions is met: (a) no keywords are found in the text; (b) only one keyword is found in the text; (c) only two keywords are found in the text, and one is labeled as a positive indicator while the other as a negative indicator.

[4] In principle, the scope of threshold *b* could be set to be infinite. However, in practice, (–5,+5) is recommended as *a priori* for *b*.

2.5 Example Application

2.5.1 Data

As part of a larger study exploring the relationship between life narratives and students' personality adaption, 656 life stories were collected from 271 undergraduate students at Northwestern University in the United States. The classification target was to label the life stories into four categories: redemption (RED), contamination (CON), redemption and contamination (BOTH), and neither redemption nor contamination (NEITHER). In the narrative research in the discipline of personality psychology, redemption and contamination are the two most important sequences for revealing the "change" tendency in people's emotional well-being through writing (McAdams, 2008). In a redemption sequence, a demonstrably "bad" or emotionally negative event or circumstance leads to a happy outcome, whereas in a contamination scene, a good or positive event or state becomes bad or negative. Three experienced experts were invited to label each story based on McAdams's manual coding system (McAdams, 2008). The Kappa agreement among the three human raters was 0.67. The label for each story was defined as the decision made by at least two human raters, and was identified as the "standard" for the training process. According to the human raters' assessment, 231 stories were labeled "change" (i.e., redemption or contamination or both), and 425 stories were labeled "no change" (i.e., neither redemption nor contamination).

2.5.2 Method

Given concerns about the common feature – "the change" tendency – in the redemption and contamination sequences, a two-stage classification framework was constructed. On the first stage, all the input was divided into two groups, "change" and "no change." A further detailed classification was conducted at the second stage to categorize the preliminary results as redemption and contamination. To illustrate the application of the text classification models, we focused only on the first stage in the present study.

The dataset was randomly split into a training set and a testing set, 70% and 30%, respectively. The "stop word list" and the Porter algorithm were used in the preprocessing to deduct the noninformative words and normalize the words into their common lexical forms. The robust classifiers were extracted by using the chi-square

selection algorithm. Three machine learning models, DT, NB, and PSM, were applied for a comparative study.

Six performance metrics, accuracy, sensitivity (recall), specificity, positive predictive value (precision) (PPV), negative predict value (NPV), and F1 measure, were used to evaluate the efficiency of the three employed machine learning algorithms. A contingency table was used to perform calculations (see Table 2.3). All six indicators are defined in definitions (1) through (6), respectively. Accuracy, the main metric used in classification, is the percentage of correctly defined 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 (i.e., proportion of "change" and "no change" texts of the total) in the corpus, and hence are more indicative of real-world performance. The predictive values, PPV and NPV, are estimators of the confidence in predicting correct classification; that is, the higher predictive values, the more reliable the prediction would be. The F1 measure combines the precision and recall in one metric, which is often used in information retrieval to show classification efficiency. This measurement can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst value at 0. Further, to check the stability of the three classification models, we explored all the metrics with an increasing number of word classifiers by adding 10 keywords, five from positive classifiers (i.e., "change") and five from negative classifiers (i.e., "no change"), each time. The number of keywords included in the textual assessment ranged from 10 to 2,600.

2.5.3 Results

Among the top 20 robust positive classifiers (i.e., keywords representing a "change" tendency), the expressions with negative semantics, e.g., "death," "depress," "scare," "lost," "anger," "die," "stop," took a one-third proportion; whereas among the top 20 robust negative classifiers (i.e., keywords representing "no change" tendency), expressions with positive semantics, e.g., "peak," "dance," "high," "promo," "best," "excite," "senior," accounted for the most, around 35%. This result implies that people generally describe life in a happy way. The words with negative semantics would be informative for detecting the "change" tendency in the life stories.

Table 2.3. Contingency Table for Calculating Classification Metrics

	True Standard	
	C_1	C_2
Assigned C_1	*a*	*b*
Assigned C_2	*c*	*d*

Note. a is a true positive value (TP), b is a false positive value (FP), c is a false negative value (FN), and d is a true negative value (TN).

$$Accuracy = \frac{a+d}{a+b+c+d} \quad (1)$$

$$Sensitivity = \frac{a}{a+c} \quad (2)$$

$$Specificity = \frac{d}{b+d} \quad (3)$$

$$Positive\ Predictive\ Value\ (PPV) = \frac{a}{a+b} \quad (4)$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{d}{c+d} \quad (5)$$

$$F1\text{-}score = \frac{2 \times Sensitivity \times PPV}{Sensitivity + PPV} \quad (6)$$

The performances of three classification models are shown in Figure 2.2 with six metrics. Note that the three models resulted in a similar overall accuracy rate of around 70%, although the PSM was a bit superior to the other two, yet not robust. Further, the PSM ranked the highest in the F1 measure, which suggested that this model performed more efficiently than the DT and the NB in the text classification. In the sensitivity analysis, the NB yielded the highest specificity (more than 90%) but sacrificed too much in sensitivity (around 10%). The PSM performed worst on specificity (around 75%) but yielded the best result in sensitivity (around 60%). The PSM was more sensitive in detecting “change” life stories but a bit less capable of finding “no-change” stories than the other two models. However, among the three models, the PSM was the most balanced between sensitivity and specificity; that is, this model showed relatively satisfactory sensitivity without losing too much specificity. Another noticeable point was that the PSM showed the highest value in the NPV. This implies that we could

have the most reliable prediction to deduct “no-change” life stories from the further stage by using the PSM rather than the DT and the NB. In the PPV plot, the NB curve ranked highest but it waved substantially with the increasing number of keywords, whereas the DT and the PSM remained stable throughout the whole processing. The PSM and DT showed relatively low PPV values (around 60%), suggesting that the confidence for reliable prediction of “change” life stories was not that strong. However, since at this preliminary stage we targeted discarding the “no-change” life stories from further classification, PPV is less important than NPV in this sense.

2.6 Discussion and Conclusion

The example study demonstrated that the PSM is a promising machine learning algorithm for text (binary) classification. Although the three classification models showed a similar overall accuracy rate, the PSM performed the best in the F1 measure and remained stable as the number of keywords increased, implying better efficiency in text classification and more reliable agreement with the human raters’ assessment than the other two standard models. Similar results were found in a recent study by He et al. (2012), where the PSM was validated in text classification for posttraumatic stress disorder (PTSD) patients’ self-narratives regarding their stressful events and physical and mental symptoms. Analogous to the example application, the PSM successfully classified the self-narratives written by individuals with PTSD and non-PTSD in high agreement (82%) with the psychiatrists’ diagnoses and presented stable results as the number of keywords increased.

Further, to help practitioners select an optimal algorithm for their own problems, the following pros and cons of each model can be considered and compared. The DT model is one of the most comprehensive models for visually tracking the path in classification. It is easily understood why a decision tree classifies an instance as belonging to a specific class. However, this model may result in low accuracy, especially for a small sample dataset. The DT uses splits based on a single feature at each internal node. Thus, many features are necessary to extract from the training set. Another frequent problem that may occur in applying DT algorithms is the overfitting. The most straightforward way of using them is to preprune the tree by not allowing it to its full size (Kotsiantis, 2007) or establish a nontrivial termination criterion such as a threshold test for the feature quality metric (see more in Elomaa, 1999; Bruha, 2000).

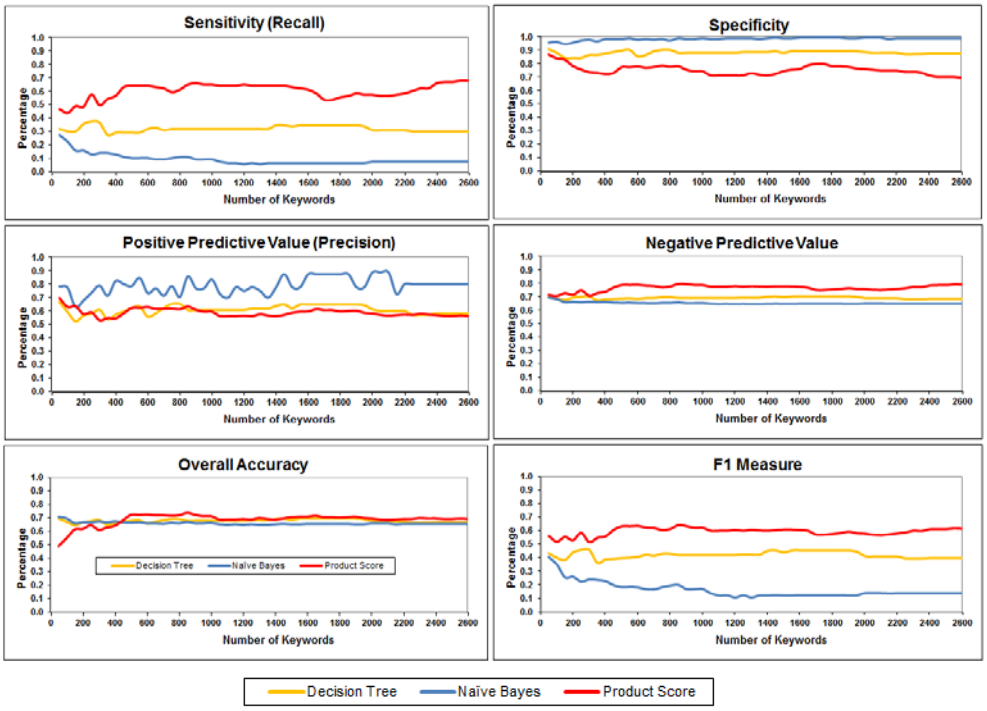


Fig. 2.2. Comparisons of text classification models, decision tree, naïve Bayes and product score model based on the example application.

The major advantages of NB are its short computational time for training and its simple form of a product with the assumption of independence among the features. However, 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, e.g., DT, instance-based learning, and rule induction, on standard benchmark datasets, and found it to be sometimes superior to the other learning schemes, even on datasets with substantial feature dependencies.

Despite adopting the same assumption of word independence in the NB, the PSM has more flexibility in the model decision threshold. As shown in Formula (2.7), the decision threshold b could be set as an unfixed constant in practice. For instance, in a clinical setting such as the PTSD screening process, on the one hand, psychiatrists may want to exclude people without PTSD from further tests, which needs a relatively

higher specificity value; on the other hand, when psychiatrists 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. With the example data in the current study, to yield satisfactory sensitivity in finding the "change" elements in life stories without sacrificing too much specificity, an optimal threshold of PSM log ratio score could be set at $b = -4$. However, since the PSM allocates a set of term weights for each key feature, more time and more storage space are expected in the training and validation process, which might reduce the PSM's effectiveness in a large sample.

In addition to the applications of text classification within the field of psychology and psychiatry, the PSM is also expected to extend its usage in educational measurement. For instance, this model might be used as an alternative approach to classify students' essays into different grade levels, to retrieve information about students' noncognitive skills by analyzing their writing components, e.g., diaries, posts, blogs, and short messages, and further to extract patterns among students' noncognitive skills and their academic grades.

In conclusion, the present study introduced the general procedure of text classification within the framework of text mining techniques and presented an alternative machine learning algorithm, the PSM, for text classification. In the comparative study with two standard models, DT and NB, the PSM was shown to be quite promising in text (binary) classification. It might be interesting to extend the PSM into a generalized multiple classification algorithm in future research, and to find out whether and how educational measurement could benefit from this new procedure.

Chapter 3

Screening for Posttraumatic Stress Disorder Using Verbal Features in Self-Narratives: A Text Mining Approach

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 transform unstructured text into a structured dataset, the second to extract robust information from diversified expression patterns. The present study develops 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 extract highly discriminative keywords with the chi-square algorithm and construct a textual assessment model to classify individuals with the presence or absence of PTSD. This results in a high agreement between computer and psychiatrists' diagnoses for PTSD and reveals 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.

Keywords: posttraumatic stress disorder; text mining; self-narratives; text classification; screening

This chapter has been published as: He, Q., Veldkamp, B. P., & de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self-narratives: A text mining approach. *Psychiatry Research, 198*(3), 441-447.

3.1 The Use of Self-Narratives

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 the approach to obtain precise results is made difficult by its complexity in relation to cost, logistics and training (Franklin & 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 and physical checks, screening instruments 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, such as the Hopkins Symptom Checklist (HSCL; Derogati.Lr, Lipman, Rickels, Uhlenhut.Eh, & Covi, 1974), the Harvard Trauma Questionnaire (HTQ; Mollica et al., 1992) , the Screen for Posttraumatic Stress Symptoms (SPTSS; Carlson, 2001), and the Self-Rating Scale for Posttraumatic Stress Disorder (SRS-PTSD; Carlier, Lamberts, van Uchelen, & Gersons, 1998). 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 & Gleser, 1969; Rosenberg & 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 backgrounds, e.g., educational level, social status and living conditions, 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 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 & 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 and his colleagues (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 classify the individuals accurately 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.

3.2 Method

3.2.1 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 (American Psychiatric Association [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.

3.2.2 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 & Schütze, 1999). The supervised text classification generally involves two phases, a training phase and a prediction phase (Jurafsky & Martin, 2009). During training, the most discriminative keywords to determine the presence or absence of PTSD are extracted. The input for the machine learning algorithm consists of (1) a set of prespecified features that may potentially be present in a document and (2) labels that classify each document (i.e.,

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.

3.2.3 Procedure

3.2.3.1 Preprocessing

To improve the efficiency of the training and testing 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”, “nightmaring” and “nightmared”, though in variant lexical forms, were normalized in an identical stem “nightmar” 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) by approximately 60% and 35%, respectively.

3.2.3.2 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 corpus 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 & Pedersen, 1997; Conway, Doan, Kawazoe, & Collier, 2009). In general, a chi-square test is used to assess whether paired observations on two variables are independent of each other (Jurafsky & 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 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 3.1) and calculated the χ^2 score for each word with the chi-square equation for a 2-by-2 table:

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}, \qquad (3.1)$$

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. Table 3.2 gives the top 20 keywords with the highest χ^2 scores used in this study.

Table 3.1. 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
¬ 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.

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 3.2). 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 the label.

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.

Table 3.2. 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, 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.

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.

3.2.3.3 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.

3.2.4 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.

The relationship between text screening method at various thresholds and the standard, 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 value (PPV), the prediction confidence that someone who has a positive test result is authentically diagnosed as PTSD, (d) negative predictive value (NPV), 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 & Zimmerman, 1993, for a complete discussion of calculating and interpreting diagnostic performance statistics.)

3.3 Results

3.3.1 Sample Characteristics

The descriptive statistics of the sample are shown in Table 3.3. Using a 95% confidence interval, the independent sample *t*-test showed that the PTSD patients used more words ($t = 2054$, $df = 232$, $p = 0.04$) and had a significantly larger vocabulary ($t = 3056$, $df = 254$, $p < 0.01$) 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.

Table 3.3. 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.

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.

3.3.2 Performance Metrics

A range of thresholds for the PTSD textual screening method, along with corresponding sensitivity, specificity, PPV, NPV, overall correct classification rates, and Kappa statistics are presented in Table 3.4. 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 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 3.4 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 82%.

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, such as the Impact of Event Scale (IES; Horowitz, Wilner, & Alvarez, 1979), the PTSD Checklist (PCL-C; Weathers, Huska, & Keane, 1991), the Davidson Trauma Scale (DTS; Davidson et al., 1997), were systematically reviewed on 22 PTSD studies with the sample size in a range from 65 to 422 (see the last row in

Table 3.4). 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 PPV (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 NPV (Mean = 0.85, S.D. = 0.07) were a bit lower but kept at a moderate level.

Table 3.4. Performance Metrics for Various Thresholds of the Textual Screening Method Compared with the Mean Performance of Screening Instruments

Threshold	Sensitivity	Specificity	PPV	NPV	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	

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 Figure 3.1. 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) 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.

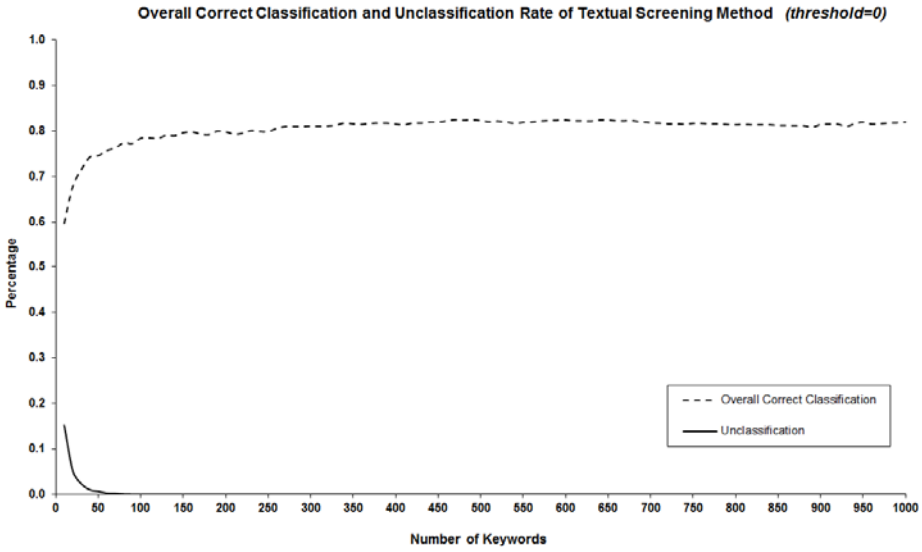


Fig. 3.1. The overall correct classification (accuracy) and unclassification rate of the textual screening method with an increasing number of keywords from 10 to 1,000 using a threshold of 0
Note. All the values are the averaged results from 15-fold cross validation.

3.4 Discussion and Conclusion

The results demonstrated that the textual screening method developed in the field of text mining was quite 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 value 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, Finkelstein, & Arana, 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 and associates (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., Smyth, 1998; Mendoza et al., 2011). Thus, in practice, a psychiatrist might be reluctant to ensure an accurate diagnosis when merely relying on the verbal features from self-narratives. Given this concern, Shrout and colleagues (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, Gudmundsdottir, Beck, Palyo, & Miller, 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 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, Zhao, Yang, & Lu, 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 & 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 & 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 two 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", 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 and his colleagues (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 quite 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 PPV 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. All in all, the textual screening method is helpful since it can make accurate prediction comparably to the human raters and significantly exclude people with low risks as PTSD from follow-up tests. 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. In addition, 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.

Chapter 4

Textual Assessment of Self-Narratives for Posttraumatic Stress Disorder Screening Using *N*-Gram Mining

Abstract

Patients' narratives of traumatic experiences and symptoms can support flexibility and depth of a psychiatric screening and diagnostic procedure. In recognition of the growing interests in analyzing such textual data, in the current study, we present an innovative computerized textual assessment system to screen patients for posttraumatic stress disorder (PTSD) via an *n*-gram text mining approach. Three machine learning algorithms, including decision tree (DT), naïve Bayes (NB), and a self-developed alternative – product score model (PSM), are used to extract the patterns between verbal features in self-narratives and psychiatric diagnoses. The performance of these three models is compared in conjunction with five representation models, unigrams, bigrams, trigrams, a combination of uni- and bigrams, and a mixture of *n*-grams. With the sample at hand, the PSM with unigrams attains the highest prediction accuracy compared to psychiatrists' diagnoses in structured interviews. The addition of *n*-grams contributes most to enhance the reliability of prediction and to balance the performance metrics, i.e., resulting in a fairly high sensitivity with the least sacrifice for specificity. This chapter further demonstrates that the computerized textual assessment system is a promising tool to analyze patients' expression behaviors, thus help psychiatrists identify the potential patients from an early stage.

Keywords: posttraumatic stress disorder; text mining; self-narratives; natural language processing; *n*-grams

This chapter has been submitted as: He, Q., Veldkamp, B. P., Glas, C. A. W. & de Vries, T. (2013). Automated assessment of patients' self-narratives for posttraumatic stress disorder (PTSD) screening using text mining.

4.1 An Automated Screening System for PTSD

The rapid advance of information technology over the last decade led to breakthroughs on quality, continuity, and efficiency of public health care (Drigas, Koukianakis, & Papagerasimou, 2011). One promising application is the use of natural language processing (NLP) and text mining techniques to identify the clinical information contained in unstructured free text documents and to codify this information into structuralized data (Trusko et al., 2010). For instance, Pakhomov and his colleagues (2011) extracted clear patterns of decline in grammatical complexity in language production affected by neurodegenerative disorders. Day and his group (2007) used NLP system to classify trauma patients based on their clinical histories. Ando and his colleagues (2007) identified cancer patients by using the transcripts through structured interviews with a text mining approach, and found considerable differences in the focus of life review by age and gender.

The individual's talk and write provide windows into their emotional and cognitive worlds (Tausczik & Pennebaker, 2010). Over the years, numerous evidence suggested that the words and expressions source from patients' self-narratives are quite informative for early detection of epidemics and mental diseases (e.g., Gottschalk & Gleser, 1969; Rosenberg & Tucker, 1979; Smyth, 1998; Franklin & Thompson, 2005). In such a case, language becomes an important medium by which clinical psychologists attempt to understand the patients. The automated identification of patients based on their self-narratives are, therefore, expected as a promising step toward the screening and diagnosis process. However, despite the great potential in mining information from self-narratives, they are not commonly used in clinical practice and have not been sufficiently described yet in the literature. There are two challenges that might limit their use. Firstly, since the self-narratives are not official records in the electronic health records (EHR), they are rarely made available outside the corporate settings that collected them. The formal studies on such texts are few and sparse. Secondly, the application of NLP technique meets more problems in patients' self-narratives than the standardized biomedical texts (i.e., academic texts in biomedical literature) and clinical texts (i.e., texts written by clinicians in the clinical settings, describing patients, pathologies and findings made during interviews and etc.). The main difficulties are their unstandardized formats, loose structures, and over-diversified words used by people with various backgrounds.

The purpose of the present study is to develop an automated screening system for posttraumatic stress disorder (PTSD) on patients' self-narratives using text mining

techniques. The general idea was 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 procedure will be helpful if it can maximize the accuracy in finding the potential PTSD patients or significantly excluding the NONPTSD individuals from the follow up tests.

Given previous efforts in development of a keyword-based textual assessment method (He & Veldkamp, 2012; He et al., 2012), the present study sought to apply the text mining techniques on 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: (1) to provide an overview of the procedure of automated textual assessment on patients' self-narratives for PTSD screening; (2) to compare the performances of different classification models in conjunction with *n*-gram representations in the screening process.

4.2 Method

4.2.1 Participants

As part of a larger study developing an alternative intake procedure for PTSD, 308 self-narratives written by trauma survivors were collected via an online survey embedded in an open forum that is dedicated to patients with mental health issues. The requirements for the self-narratives 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) the self-narratives must include both of traumatic events and symptom description. The data with missing diagnoses were discarded, which resulted in a total of 300 narratives for the final set, 150 participants were diagnosed as PTSD patients and the other 150 participants as NONPTSD. (We stopped collecting data at n=150 per group.) We did

consult with experienced psychiatrists on the textual data to ensure that participants in this study were diagnosed correctly based on DSM-IV criteria (APA, 2000). Besides, all the participants reported that they had at least three-month experience using Internet and did not meet problems in using the online survey system. The context of stressful events covered the following eight types: child abuse, sexual abuse, traffic accident, war, domestic violence, death of a beloved person, robbery, and fire.

4.2.2 *N*-grams

The textual data are usually encoded via a data representation model, namely, 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 that considers the interaction effect of two, three or more consecutive words is proposed as a way to expand the standard unigram representation model (e.g., Bekkerman & Allan, 2003; Tan, Wang, & Lee, 2002). 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 as 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 "get depressed").

4.2.3 Procedure

The textual screening procedure generally consists of three phases – preparation, training and testing, as shown in Figure 4.1. During the preparation phase, the textual data are generally divided into a training and a test set and preprocessed according to several linguistic rules. The training and testing phases are the essential parts to classify the self-narratives into PTSD and NONPTSD groups, where the text mining techniques are mainly applied. During training, the most discriminative features (e.g., keywords or key vectors) for determining the class label are extracted. The input for

the machine learning algorithm consists of a set of prespecified features that may potentially be present in a document and labels that classify each document. The objective of the training phase is to "learn" the relationship between the features and the class labels. The testing 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 that were not used during training. In the testing procedure, the features extracted from the training are scanned in each new input. Thus, the features (e.g., keywords or key vectors) that were systematically recognized are fed into the "trained" classifier model to predict the most likely label for each new document.

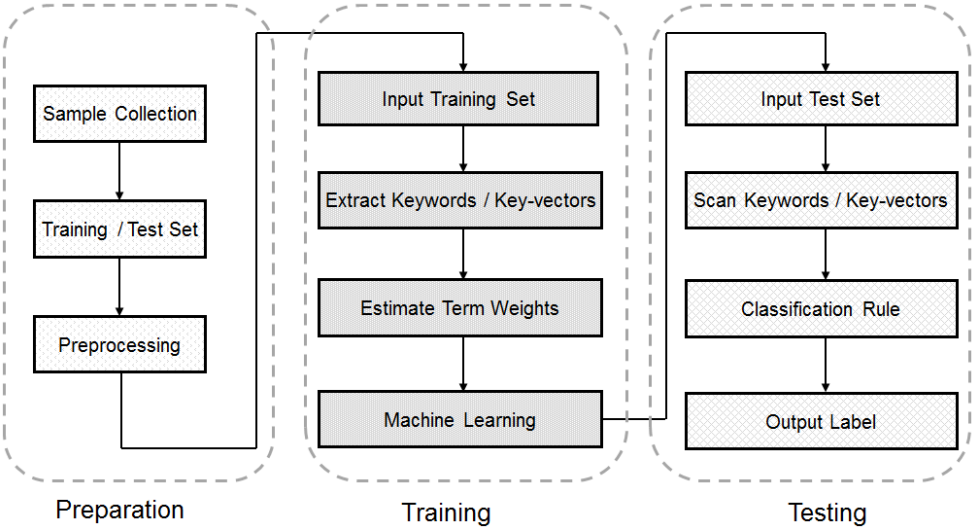


Fig. 4.1. The overview of text classification procedure for PTSD screening

4.2.3.1 Preparation

The dataset was split into a training set and a test set – the training set consisted of 200 self-narratives, 100 randomly selected from PTSD corpus and 100 from NONPTSD corpus, while the remaining 100 narratives, 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.

To improve the efficiency of training and testing procedures, a preprocessing routine was implemented. This involved screening digital numbers, deducting noninformative "stop words"[5] (e.g., "I", "to"), common punctuation marks (e.g., ".", ":"), and frequently used abbreviations (e.g., "isnt", "Im"), and "stemming" the rest of words, for instance, with the Porter algorithm (Porter, 1980) to remove common morphological endings. For example, the terms "nightmares", "nightmaring", and "nightmared", 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. Afterwards, each component of the *n*-gram was stemmed with the Porter algorithm, for example, "I worked" was standardized as "I work".

4.2.3.2 Training

Training text classifiers is the procedure where machines "learn" to automatically recognize complex patterns, to distinguish between exemplars based on their different patterns, and to make intelligent predictions on their class. The input texts were represented by five data representation models: unigrams, bigrams, trigrams, a combination of uni- and bigrams, and a mixture of uni-, bi-, and trigrams. The chi-square selection algorithm (Oakes et al., 2001) was used to perform feature extraction. This model is recommended for use due to its high effectiveness in finding robust keywords and testing for the similarity between different text corpora (Manning & Schütze, 1999; for more feature selection models refer to Forman, 2003). To apply the chi-square algorithm, each word is compiled into its own 2-by-2 contingency table as shown in Table 2.2 (Chapter 2). The number of word occurrences in two corpora C_1 (i.e., PTSD corpus) and C_2 (i.e., NONPTSD corpus) is indicated by n_i and m_i, 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 to determine how far C_1 departs from C_2. Under the null hypothesis, the two corpora are similar, so their distribution of words are 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 (O_{ij}). Under the

[5] The current study used the standard "English Stop Word List" (127 words) in Python NLTK (Natural Language Toolkit) to deduct the noninformative words.

null-hypothesis, that is, using the assumption of independence, the expected frequencies (E_{ij}) are computed from the marginal probabilities, i.e., from the totals of the columns and rows converted into proportions, using the formula $E_{ij} = (column_i\ total \times row_j\ total) / grandtotal$. Afterwards, 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, $X^2 = \sum_{i,j}[(O_{ij} - E_{ij})^2 / E_{ij}]$.

To ensure the reliability of the calculation, words that occur fewer than five times are usually eliminated (Manning & Schütze, 1999; Oakes et al., 2001). However, for a small sample, the number of word occurrences could be even lower, perhaps three times. Based on the chi-square scores, all word (vectors) are ranked in a descending order, and those standing on the top are extracted as robust classifiers. Further, if the ratio n_i / m_i is larger than the ratio $len(C_1) / len(C_2)$, the word is regarded as more typical of 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).

Three machine learning algorithms were used in the current study, including two commonly used models, decision trees (DT; Quinlan, 1993) and naïve Bayes (NB), and one alternative product score model (PSM; He & Veldkamp, 2012; He et al., 2012). 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 stopped until they are categorized into a terminate class.

The NB is a probabilistic classifier applying Bayes's theorem with strong (naïve) independence assumptions (Lewis, 1998). It is simple but effective in practice (Hand & Yu, 2001). The basic idea is to estimate the conditional probability of the class C given the word vectors $\mathbf{w}$ with the assumption of word independence. Namely,

$$P(C \mid \mathbf{w}) = \frac{p(C)\,p(w_1 \mid C)\,p(w_2 \mid C)...p(w_k \mid C)}{p(w_1,...,w_k)} = \frac{p(C)\prod_{i=1}^{k} p(w_i \mid C)}{p(\mathbf{w})}\ , \qquad (4.1)$$

where $p(C)$ is the prior probability of a certain class, and $p(w_i \mid C)$ is the conditional probability of a word occurs 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

$$R = \frac{P(C_1 \mid \mathbf{w})}{P(C_2 \mid \mathbf{w})} = \frac{p(C_1)\prod_{i=1}^{k} p(w_i \mid C_1)}{p(C_2)\prod_{i=1}^{k} p(w_i \mid C_2)} \quad . \tag{4.2}$$

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

The PSM is a self-developed alternative machine learning algorithm, derived from Naïve Bayes by using a smoothing constant for computing word probability. It features in assigning two weights for each keyword (in binary classification) – the probability of the word i occurs in the two separate corpora, U_i and V_i – to indicate to how much of a degree the word can represent the two classes, that is, $U_i = (n_i + a)/len(C_1)$ and $V_i = (m_i + a)/len(C_2)$. Note that a smoothing constant a (we use $a = 0.5$ in this study) is added to the word frequency to account for words that do not occur in the training set but might occur in new texts (for more smoothing rules refer to Manning & Schütze, 1999; Jurafsky & Martin, 2009). The name product score comes from a product operation to compute scores for each class, i.e., S_1 and S_2, for each input text based on the term weights. That is,

$$\begin{cases} S_1 = P(C_1) \cdot \prod_{i=1}^{k} U_i = P(C_1) \cdot \prod_{i=1}^{k} [(n_i + a)/len(C_1)] \\ S_2 = P(C_2) \cdot \prod_{i=1}^{k} V_i = P(C_2) \cdot \prod_{i=1}^{k} [(m_i + a)/len(C_2)] \end{cases} \quad . \tag{4.3}$$

The classification rule is defined as:

$$\text{choose} \begin{cases} C = 1 & \text{if } \log(S_1 / S_2) > b \\ C = 2 & \text{else} \end{cases} \quad , \tag{4.4}$$

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 that 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 or key vectors were used to accomplish the classification task, unclassification rules were needed to be considered. A text was defined as "unclassified" when either one of the following conditions 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, and one was labeled as a positive indicator (i.e., PTSD) while the other as a negative indicator (i.e., NONPTSD).

4.2.3.3 Testing

A testing 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 system, it first checked whether the extracted features (i.e., keywords or key vectors) existed and then 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.

4.2.4 Analytical Strategy

To ensure the proper generalization capabilities for the textual classifier model, a 15-fold cross validation procedure was applied. We arbitrarily divided the original 300 self-narratives into six subgroups, each subgroup consisting of 50 self-narratives, 25 from PTSD and 25 from NONPTSD. Each time, two-thirds of the subgroups were drawn as the training set, while the remaining one-third were used as the test set. We performed the multiple evaluations on the 15 folds and averaged the results.

Six performance metrics, accuracy, sensitivity (recall), specificity, positive predictive value (precision; PPV), negative predictive value (NPV), and F-score measurement, were used to evaluate the efficiency of the three employed machine learning algorithms. Accuracy, the main metric used in classification, is the percentage of correctly defined 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 (i.e., proportion of "PTSD" and "NONPTSD" texts of the total) in the corpus, and hence are more indicative of real-world performance. The predictive values, PPV and NPV, are estimators of the

confidence in predicting correct classification; that is, the higher predictive values, the more reliable the prediction would be. The F-score measurement combines the precision and recall in one metric, which is often used in information retrieval to show classification efficiency. This measurement can be interpreted as a weighted average of the precision and recall, where an F-score reaches its best value at 1 and worst value at 0. A contingency table was used to perform calculations of these six indicators (see Table 2.3 in Chapter 2). Statistical significance levels were reported based on paired *t*-tests using a 95% confidence interval.

Further, to check the stability of the three classification models, we explored all the metrics with an increasing number of keywords each time. That is, the analysis started with 10 keywords with the highest chi-square scores, i.e., 5 keywords labeled as PTSD classifiers and 5 keywords labeled as NONPTSD classifiers, and ended with 2600 keywords, i.e., 1300 keywords from either classifier label.

4.3 Results

Table 4.1 summarizes the averaged results and standard deviations from the 15-fold cross validation. The lowest and highest averaged values in each column are highlighted in shadow and bold forms, respectively. The DT performed the poorest among the three models, which yielded the lowest values in sensitivity (Mean=0.582, S.D.=0.071), NPV (Mean=0.572, S.D.=0.044) and F-score (Mean=0.576, S.D.=0.044) in conjunction with unigrams and the lowest accuracy rate (Mean=0.568, S.D.=0.033) with trigrams. The NB with trigrams produced the highest value in sensitivity (Mean=0.922, S.D.=0.040), but sacrificed dramatically in specificity (Mean=0.278, S.D.=0.078), and showed the lowest PPV (Mean=0.569, S.D.=0.027) among all. The PSM with unigrams attained the highest in accuracy rate (Mean=0.816, S.D.=0.053) and F-score (Mean=0.821, S.D.=0.053), implying the best agreement with psychiatrists' diagnoses using item-based questionnaires through traditional structured interviews.

The addition of *n*-grams made contributions to balancing performance metrics in the textual screening procedure. For instance, with the joint representation of unigrams and bigrams, the PSM moderately lowered the sensitivity value from 0.852 (unigram) to 0.814, but benefited in a significant raise of specificity from 0.780 (unigram) to 0.810. Further, compared with unigrams, the combination of *n*-grams helped enhance the reliability of prediction. The most reliable prediction of PTSD was suggested by the highest value of PPV (Mean=0.813, S.D.=0.039) produced by the PSM with a

mixture of uni-, bi-, and trigrams. The highest confidence in excluding NONPTSD from further assessment was indicated by the highest value of NPV (Mean=0.868, S.D.=0.058), which was shown in NB with a mixture of *n*-grams. However, the overall accuracy rate was not significantly improved by the introduction of *n*-grams and was marginally decreased when the bigrams and trigrams were used alone.

The stabilities of three models were presented by overall classification accuracy in Figure 4.2. 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 run on the top, closely following by the NB and substantially beyond the DT. Both of the PSM and NB started from a relatively low value when only a few keywords were used, but quickly rose up when more keywords were added. After the "elbow point" around 50 keywords, the PSM remained relatively stable, whereas the 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 NONPTSD 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 poor.

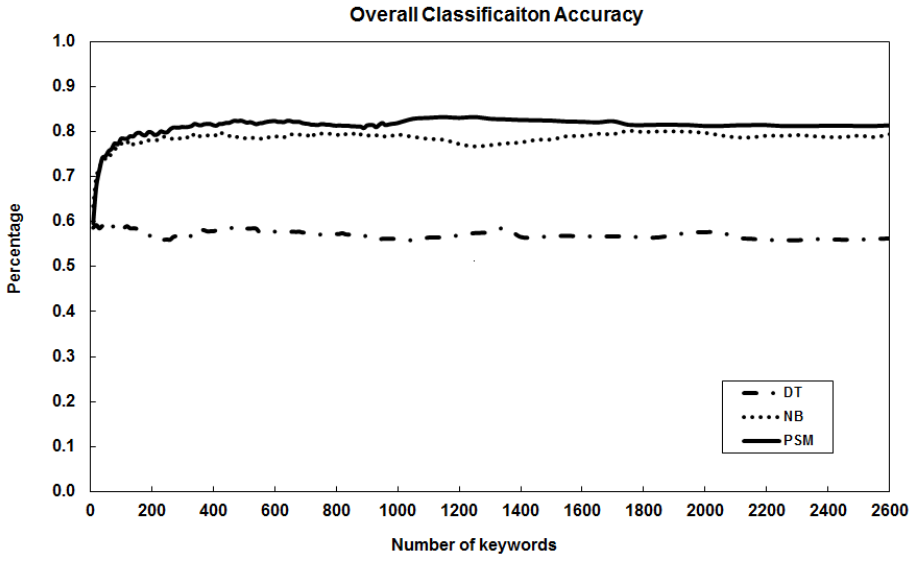


Fig. 4.2. Overall classification accuracy of three text classification models, decision trees, naïve Bayes and product score model in conjunction with unigrams

Table 4.1. Averaged Results from Three Classification Models: DT, NB and PSM Based on 15-fold Cross Validation

		Accuracy	Sensitivity	Specificity	PPV	NPV	F score
Decision Tree	Unigrams	0.570 (0.039)	0.582 (0.071)	0.558 (0.084)	0.571 (0.038)	0.572 (0.044)	0.576 (0.044)
	Bigrams	0.596 (0.045)	0.584 (0.077)	0.607 (0.080)	0.600 (0.046)	0.595 (0.049)	0.589 (0.052)
	Trigrams	0.568 (0.033)	0.621 (0.150)	0.512 (0.174)	0.573 (0.043)	0.582 (0.053)	0.583 (0.061)
	Uni + Bi	0.583 (0.041)	0.598 (0.065)	0.568 (0.085)	0.583 (0.042)	0.586 (0.046)	0.588 (0.041)
	Uni + Bi + Tri	0.580 (0.037)	0.595 (0.060)	0.564 (0.079)	0.580 (0.039)	0.583 (0.038)	0.585 (0.037)
Naïve Bayes	Unigrams	0.788 (0.032)	0.779 (0.063)	0.796 (0.071)	0.799 (0.050)	0.789 (0.046)	0.779 (0.029)
	Bigrams	0.680 (0.036)	0.888 (0.057)	0.472 (0.110)	0.637 (0.043)	0.832 (0.059)	0.736 (0.020)
	Trigrams	0.600 (0.030)	**0.922 (0.040)**	0.278 (0.078)	0.569 (0.027)	0.827 (0.075)	0.699 (0.018)
	Uni + Bi	0.782 (0.039)	0.865 (0.061)	0.699 (0.100)	0.752 (0.056)	0.851 (0.052)	0.799 (0.031)
	Uni + Bi + Tri	0.767 (0.032)	0.895 (0.057)	0.640 (0.073)	0.718 (0.036)	**0.868 (0.058)**	0.793 (0.027)
Product Score	Unigrams	**0.816 (0.053)**	0.852 (0.078)	0.780 (0.073)	0.799 (0.054)	0.848 (0.073)	**0.821 (0.053)**
	Bigrams	0.758 (0.043)	0.758 (0.087)	0.769 (0.051)	0.768 (0.039)	0.767 (0.060)	0.760 (0.053)
	Trigrams	0.670 (0.050)	0.635 (0.081)	0.781 (0.065)	0.747 (0.071)	0.683 (0.056)	0.684 (0.072)
	Uni + Bi	0.812 (0.046)	0.814 (0.097)	**0.810 (0.054)**	0.805 (0.042)	0.820 (0.075)	0.810 (0.055)
	Uni + Bi + Tri	0.802 (0.051)	0.801 (0.097)	0.804 (0.052)	**0.813 (0.039)**	0.810 (0.076)	0.800 (0.060)

Note. The value within the blankets presents the standard deviation in the 15-fold cross validation. PPV represents positive predictive value, NPV represents negative predictive value. The cells highlighted in shade and bold forms represent the lowest and highest value within each column, respectively.

4.4 Discussion and Conclusion

This chapter presented a computerized textual assessment system to screen for PTSD based on patients' self-narratives via an *n*-gram text mining approach. In the comparative study among three classification models, DT, NB and PSM, in conjunction with five data representations – unigrams, bigrams, trigrams, a combination of uni- and bigrams and a mixture of *n*-grams, with the sample at hand, the PSM with unigrams attained the highest prediction accuracy (81.6%) compared with psychiatrists' diagnoses in structured interviews. 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.

Bekkerman and Allan (2003) summarized that there exist two major approaches to incorporate *n*-grams into document representation. The first one excludes unigrams from the representation and bases the representation on *n*-grams ($n > 1$) only, while the second one applies *n*-grams together with unigrams. It turns out that the first approach leads in most cases to a decrease in the classification results in comparison to 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 datasets 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 baseline (i.e., unigram with PSM) has already reached a high agreement between computer and psychiatrists' diagnoses, implying that the unigrams were powerful enough to represent the relatively small and "simple" corpus. Therefore, the classification accuracy was not apparently enhanced with the addition of *n*-grams. It might be interesting to apply the *n*-gram text mining method on a larger and more complex dataset in a future study and include the textual structure features as well, such as grammatical properties and part-of-speech to supplement the frequency-based representation model.

To help practitioners select an optimal algorithm in their own studies, the following pros and cons of each model can be considered and compared. The DT model is one of the most comprehensive models for visually tracking the path in classification. It is easily understood why a decision tree classifies an instance as belonging to a specific class. However, this model may result in low accuracy, especially for a small sample dataset, and meet the problem of overfitting when the tree

grows too big based on the accidental properties of the training set. The most straightforward way of adjustment is to preprune the tree to reduce its size (Kotsiantis, 2007) or establish a nontrivial termination criterion such as a threshold test for the feature quality metric (for more on pruning trees refer to Elomaa, 1999; Bruha, 2000; Kirschstein, Liebscher, & Becker, 2013).

The major advantages of NB are its short computational time for training and its 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 other algorithms, e.g., DT, instance-based learning, and rule induction, on standard benchmark datasets, and found it to be sometimes superior to the other learning schemes, even on datasets with substantial feature dependencies.

Note that the results generated from PSM are very close to those from NB, the probable reasons might be their similar model structure and a common assumption of word independence. In the equation of NB (Formula 4.1), the product of conditional probability of a word i given a category j is taken over by $p(w_i \mid C_j)$, which is analogous to the definition of PSM (Formula 4.3) where these probabilities are replaced by the relative frequencies of words with a smoothing constant (i.e., term weight). However, the PSM has more flexibility in the model decision threshold than NB. As shown in the Formula 4.4, the decision threshold b could be set as an unfixed constant in practice. For example, in a clinical setting, on one hand, psychiatrists may want to exclude people without PTSD from further tests, which needs a relatively higher specificity value. On the other hand, when psychiatrists 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, since the PSM allocates a set of term weights for each keywords, more time and more storage space are demanded in the training and validation process. It might reduce the PSM's effectiveness when using a large sample or performing a multiple categorization.

Computer-delivered and Internet-based diagnoses 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; Proudfoot et al., 2004; Owen, Hanson, Preddy, & Bantum, 2011). The development of textual screening system for psychiatric patients was initiative in both fields of psychiatry and applied linguistics. The whole procedure could be easily embedded in

an Internet-based test as an additional module to online psychiatric diagnosis. With its help, people living in remote areas, people with restricted mobility, or people reluctant to seek face-to-face interviews could complete web-based tests in a private, flexible and relaxed way (Naglieri et al., 2004; Maheu, Pulier, McMenamin, & Posen, 2012). Besides the application in screening for PTSD presented here, the textual assessment method is expected to be useful in a broader domain with similar background and makeup, for instance, in screening for depression and multiple other mental diseases according to the DSM criteria (APA, 2000).

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 (1) the textual assessment on self-narratives achieved a high agreement with psychiatrists' diagnoses, and (2) the addition of higher order *n*-grams could help balance the classification metrics and enhance the reliability of classification prediction. This chapter further demonstrates that the computerized textual assessment system is a promising tool for e-health platforms to analyze patients' expression behaviors, thus help psychiatrists identify the potential patients from an early stage.

Chapter 5

Assessing Impact of Differential Symptom Functioning on Posttraumatic Stress Disorder Diagnosis Using Item Response Theory

Abstract

This chapter explores the generalizability of DSM-IV diagnostic criteria for posttraumatic stress disorder (PTSD) to various subpopulations. Besides identifying the differential symptom functioning (also referred to differential item functioning [DIF]) related to various background variables such as gender, marital status and educational level, this study emphasizes the importance of evaluating the impact of DIF on population inferences as made in health surveys and clinical trials, and on the diagnosis of individual patients. Using a sample from the National Comorbidity Study-Replication (NCS-R), four symptoms for gender, one symptom for marital status, and three symptoms for educational level are significantly flagged as DIF, but their impact on diagnosis is fairly small. From the results, we may conclude that the DSM-IV diagnostic criteria for PTSD do not produce substantially biased results in the investigated subpopulations, and there should be few reservations regarding their use. Further, although the impact of DIF is found to be quite small in the current study, we recommend that diagnosticians always perform a DIF analysis of various subpopulations using the methodology presented here to ensure the diagnostic criteria valid in their own studies.

Keywords: posttraumatic stress disorder; item response theory; differential item functioning; NCS-R

This chapter has been published as: He, Q., Glas, C. A. W., & Veldkamp, B. P. (2013). Assessing Impact of Differential Symptom Functioning on Posttraumatic Stress Disorder (PTSD) Diagnosis. *International Journal of Methods in Psychiatric Research.* In press.

5.1 The PTSD Diagnostic Criteria

The diagnostic criteria for posttraumatic stress disorder (PTSD) have been revised several times since the debut in the 1980s (Spitzer, First, & Wakefield, 2007). The current Diagnostic and Statistical Manual of Mental Disorders IV (DSM-IV; APA, 2000) summarizes the PTSD diagnostic criteria in six clusters: (A) exposure to a traumatic event (A1-A2); (B) reexperience such as flashbacks and nightmares (B1-B5); (C) avoidance of stimuli associated with the trauma and numbing (C1-C7); (D) increased arousal such as anger and hypervigilance (D1-D5); (E) duration of symptoms (more than one month); and (F) significant impairment in social life. A person is diagnosed as PTSD when he meets both conditions within Criterion A, has a specified number of symptoms from each of the three symptom clusters (for Criterion B at least one symptom, for Criterion C at least three symptoms, and for Criterion D at least two symptoms), and meets the conditions of Criteria E and F (for the complete list of PTSD criteria see the DSM-IV, APA, 2000).

Numerous studies have been conducted to validate the DSM-IV PTSD diagnostic criteria using either factor analytic methods (e.g., Andrews, Joseph, Shevlin, & Troop 2006; Asmundson et al., 2000; Blake et al., 1995) or item response theory (IRT; e.g., Betemps, Smith, Baker, & Rounds-Kugler, 2003; Conrad et al., 2004; Palm, Strong, & MacPherson, 2009). Evidence shows that demographic factors (e.g., gender, age, education) could influence the generalizability of the DSM-IV in such a way that patients with different backgrounds may have differential symptoms (Gavranidou & Rosner, 2003). For instance, Peters and his group (2006) found that men were less likely than women to have particular symptoms (e.g., easily startled) that are considered by society to be a sign of weakness. To ensure the accuracy of cross-group comparisons and the appropriateness of the PTSD diagnosis, differential symptom functioning needs to be identified and its potential influence needs to be estimated.

Examination of differential item functioning (DIF; Camilli & Shepard, 1994) based on IRT (Lord, 1980; Rasch, 1960) is a method to address these concerns. DIF examines the measurement equivalence across groups; that is, it explores the relationship between item response and a group variable defined by demographic attributes (e.g., gender, race), conditional on a measure of an underlying construct such as PTSD severity (Teresi & Fleishman, 2007). In the context of PTSD assessment, an item (i.e., a symptom) is identified to exhibit DIF if respondents from different subgroups with the same degree of PTSD do not have the same probability of endorsing each response category (e.g., yes/no) of that item. For example, Palm et al.

(2009) examined differential PTSD diagnostic symptoms between genders based on a national sample from the National Comorbidity Study-Replication (NCS-R; Kessler et al., 2004). Of the 17 symptoms comprising symptom clusters B through D, seven symptoms were identified as having significant gender bias (i.e., marked as DIF items). The results suggested that among PTSD sufferers, women are more likely than men to report feeling emotionally distant and easily startled, whereas men are more likely than women to report having a lack of a plan for the future, unwanted memories, unpleasant dreams, and a short temper.

However, statistical significance is not the only important aspect of a DIF analysis; the subsequent consequences in terms of diagnosis are also of interest. That is, in addition to identifying specific items exhibiting DIF, it is also important to assess whether DIF impacts the diagnosis of PTSD. Little investigation in this area has been done in earlier studies. Therefore, the objective of this study is to further explore the generalizability of DSM-IV PTSD diagnostic criteria in two ways: (a) to identify differential symptom functioning by the background variables of gender, marital status, and educational level; and (b) to introduce two methods for estimating the impact of DIF on PTSD diagnosis.

5.2 Method

5.2.1 Sample

The data used in the present study were from the PTSD module in the NCS-R . The NCS-R, conducted between February 2001 and April 2003 in the United States, is a nationally representative community household survey of the prevalence and correlates of mental disorders (Kessler et al., 2004). A total of 9282 interviews were completed in the main survey (Part I of the NCS-R). Questions regarding PTSD symptomology were included in Part II of the NCS-R interview, which assessed disorders that were either considered to be secondary importance or that were very time-consuming to assess (Kessler et al., 2004). Part II was administered to only 5692 of the total respondents, over-sampling those with clinically significant psychopathology (Kessler et al., 2004).

In the PTSD module of the interview, participants were asked about exposure to 27 different traumatic events (e.g., domestic violence, rape, life-threatening automobile accident). A total of 4984 endorsed at least one traumatic event and 1197 met criterion A (i.e., had experienced, witnessed or been confronted with a traumatic event and involved intense fear, helplessness, and /or horror) regarding the worst event. Of the

1197 participants, 317 were excluded because they endorsed neither of two filter items: “did you have any problems after the event like upsetting memories or dreams, feeling emotionally distant or depressed, trouble sleeping or concentrating, or feeling jumpy or easily startled?” and “did any of these reactions ever last for 30 days or longer?” PTSD symptoms were assessed by dichotomous items (i.e., “yes”=1, “no”=0). Each item corresponded to one symptom criterion listed in the DSM-IV. The present study focused on the 17 items regarding PTSD symptoms corresponding to Criteria B through D. PTSD criteria requiring the highest thresholds were assessed first in the NCS-R, that is, symptoms were assessed in the following order: Criterion C, B and D. The NCS-R implemented a “skip-out” principle: If the number of experienced symptoms for any criterion domain was below the threshold, e.g., the number of experienced symptoms for Criterion C was smaller than three, the administration of items stopped. The respondents who were skipped out of the survey were directly diagnosed as non-PTSD. A total of 880 respondents received all items in Criterion C regarding the worst event. 41 respondents did not continue in the assessment because they did not meet the threshold in Criterion C, and 65 respondents did not receive any items regarding symptoms in Criterion D because they did not meet threshold of Criterion B. 14 respondents stopped the assessment after giving all the negative responses to items in Criterion D, and 88 respondents did not go further because they reported that none of these reactions in Criterion C, B and D lasted over one month. Consequently, a total of 672 participants who completed the entire set of PTSD diagnostic items in the NCS-R were included in the present study. Of the 672 participants, 472 (70.2%) were diagnosed as PTSD and 200 (29.8%) diagnosed as non-PTSD. Table 5.1 gives a number of characteristics of the sample. The current sample had a mean age of 42.2 years (S.D.=13.97), and the majority of respondents was female (500, 74.4%). 156 (23.2%) participants had higher education (i.e., over 16-year education). 120 (18.2%) participants reported that they had never been married.

5.2.2 Dimensionality and Model Fit

The DIF analysis consisted of two parts: identifying DIF items and assessing the impact of DIF on PTSD diagnosis. Three background variables – gender, marital status, and educational level – were used for the analysis. All the results were obtained using the software program MIRT (Glas, 2010).

Table 5.1. Sample Characteristics (n=672)

Characteristics	Total	PTSD	NONPTSD
N	672	472	200
Gender			
Female	500	365	135
Male	172	107	65
Age (years)			
Mean	42.2	41.55	43.69
(S.D.)	(13.97)	(13.62)	(14.71)
Marriage status			
Married	351	243	108
Separated/widowed	201	142	59
Never married	120	87	33
Education level			
0-11 years	122	98	24
12 years	192	124	68
13-15 years	202	141	61
Over 16 years	156	109	47

Responses to the 17 items of the NCS-R were used as input for the statistical analysis. The item parameters in an IRT model were estimated by marginal maximum likelihood (MML; Bock & Aitkin, 1981). Given that the three symptom domains defined by the DSM-IV are used to index a general level of PTSD severity, we first considered a unidimensional two-parameter logistic model (2PL) underlying responses to the 17 symptoms (i.e., all 17 items on a single dimension). Next, given that the 17 symptoms are placed *a priori* into three separate criterion domains, we also considered a three-dimensional IRT model where each domain was associated with a separate dimension.

In the unidimensional 2PL model, that is, the probability of a score in category "yes" ($X_{ni}=1$) of item i is given by the item response function

$$P(X_{ni}=1|\theta_n)=\frac{\exp[\alpha_i(\theta_n-\beta_i)]}{1+\exp[\alpha_i(\theta_n-\beta_i)]}, \tag{5.1}$$

where θ_n is the latent PTSD level of person n, β_i is an item location parameter representing the severity level of each diagnostic symptom, and α_i is an item

discrimination parameter indicating the extent to which the item response is related to the latent scale. A special version of the model is the Rasch model or one-parameter logistic model (1PL; Rasch, 1960) where the item discrimination parameter α_i is fixed as 1. The Rasch model is often used in clinical applications as well (e.g., Elhai et al., 2011; Wong, Ungvari, Leung, & Tang, 2007).

In the multidimensional version of the 2PL model, the probability of a positive response depends on M latent variables, say $\theta_{n1},...,\theta_{nm},...,\theta_{nM}$. In the multidimensional case, in Formula (5.1), the product $\alpha_i\theta_n$ is replaced by $\sum_m \alpha_{im}\theta_{nm}$. Note that the discrimination parameters $\alpha_{i1},...,\alpha_{im},...,\alpha_{iM}$ can be viewed as factor loadings. In the present study, we used a between-items multidimensional model, where each item loaded on one of the three dimensions only.

Two approaches were used to investigate which model fitted the data better: a likelihood ratio-statistic and an item-oriented Lagrange multiplier (LM) test. First, the likelihood-ratio test of the 2PL model against the Rasch model yielded a value of the test statistic of $\chi^2 = 78.53, df = 16, p < 0.001$, while the multidimensional model against the unidimensional 2PL model yielded a value of $\chi^2 = 37.41, df = 3, p < 0.001$. It was concluded that the multidimensional model fit the data best and the 2PL fit the data significantly better than the Rasch model. However, though using a more complex model will generally result in better model fit, using a more parsimonious model might still lead to an adequate data description. To investigate this, a second approach was used. Under each model, item fit was evaluated using an LM item fit statistic (Glas, 1998, 1999). These statistics can be used to evaluate the fit of the expected item response function given by Formula (5.1) to the observed item responses.

Item fit was tested with a significance level of 0.01. For the Rasch model, the test was significant for six items, while no tests were significant for either the 2PL model or the multidimensional model. Further, the LM test statistic is accompanied by an effect size which measures the difference in observed and expected average item responses. For the 2PL model and the multidimensional model, these differences had the same magnitude. Hence, although a multidimensional IRT model fit the data better than 2PL in terms of the likelihood ratio test, it was not clearly superior in item fit. Therefore, the simpler unidimensional 2PL model was preferred over the more complicated multidimensional one.

5.2.3 Searching for DIF

Also DIF can be evaluated using an item-oriented LM item fit statistic (Glas, 1998, 1999). To compute the LM statistic, the sample of respondents is divided into subgroups labeled $g = 1, 2, ..., G$. For instance, to examine DIF items related to gender, we divided the sample into two groups (i.e., male and female), and $G = 2$. The statistic is based on the difference between average observed scores on every item *i* in the subgroups, namely, $S_{ig} = \frac{1}{N_g} \sum_{n|g}^{N_g} X_{ni}$ (where the summation is over the N_g respondents in subgroup g), and their expectations $E(S_{ig})$. The differences are squared and divided by their covariance matrix (for more details refer to Glas, 1998, 1999; Glas & Falcon, 2003). The LM statistic has an asymptotic chi-square distribution with $G-1$ degree of freedom. The statistics are accompanied by effect size $d_{ig} = \max_g \left| S_{ig} - E(S_{ig}) \right|$ showing the degree of model violation. Since the effect size d_{ig} is on a scale ranging from 0 to the maximum score m_i, effect size $d_{ig} < 0.10$, although somewhat arbitrary, is commonly suggested as an indicator of minor and acceptable model violation because of its effectiveness to detect a sufficient number of DIF items (van Groen, ten Klooster, Taal, van de Laar, & Glas, 2010). In the present study, an item was identified as DIF when the LM statistic was significant ($p < 0.05$) or the effect size was above the cutoff point ($d_{ig} = 0.10$).

When items are flagged as DIF, one way to ensure comparability across subpopulations is to remove these items. However, removing items may decrease the reliability and potentially threaten the validity of the test. An alternative is to assign group-specific item parameters to the DIF items. In this approach, it is assumed that the same construct is measured in all subgroups, but for some subgroups the item locations on the latent scale or the discrimination parameters vary. This procedure was, for instance, successfully applied by Weisscher, Glas, Vermeulen, & de Haan (2010), van Groen et al. (2010), and de Jong, Steenkamp, & Veldkamp (2009) to establish measurement equivalence in such areas as quality of life, physical disabilities, and marketing scales. An iterative procedure was conducted in this study, where the item with the largest significant LM statistic was given group-specific item parameters in each step. The iteration steps were repeated until no items were left with significant LM statistics (i.e., $p < 0.05$) or when the effect size was smaller than 0.10 (for more on the iteration procedure, refer to Glas & Verhelst, 1995).

5.2.4 Assessing the Impact of DIF

The impact of DIF was evaluated by introducing group-specific parameters and comparing the results via two approaches, one approach that measures the differences between group distributions and the other that measures the change in accuracy of diagnosis of individuals.

5.2.4.1 Measurement on Group Distributions

In MML estimation, all subgroups (e.g., genders) have their own normal distribution of latent trait values. In the gender-related case, the latent scale is identified by setting the mean and variance of the latent distribution of female respondents equal to 0 and 1, respectively. A common method is evaluating whether the means of the distributions of male and female respondents are different. This boils down to testing the hypothesis whether the mean of the males differs from zero. The impact of DIF can be evaluated by estimating the parameters of the population distributions, with and without group-specific item parameters.

5.2.4.2 Measurement on Diagnosis Accuracy on Individuals

The second method focuses on examining whether the introduction of group-specific item parameters can significantly enhance the accuracy of the diagnosis of individuals. Based on the diagnostic results from the NCS-R survey (i.e., diagnoses on the presence or absence of PTSD concerning DSM-IV criteria), the distribution of latent scores of PTSD and non-PTSD groups were obtained. A cutoff point on the latent scale was estimated to distinguish PTSD from non-PTSD.

To take the gender-related case as an example, we used three approaches to set the standard (i.e., obtain a cutoff point on the latent scale). The first approach entailed finding the midpoint between the medians of the two distributions (Cizek & Bunch, 2007). The medians of the distribution of PTSD and non-PTSD groups were 0.191 and −0.874, respectively, and the midpoint was −0.341. The second approach was the contrasting-groups method (Brandon, 2002), which uses logistic regression to determine the latent score point at which the probability of category membership is 50%. Setting the respondent status as a dichotomous variable coded 0 = non-PTSD and 1 = PTSD, we entered the latent scores of all the respondents into a general logistic regression equation; that is, $y^* = a + bx$, where y^* is the predicted value of the outcome

variable (respondent status) for a respondent and x is the respondent's observed score. The intercept and slope parameters were estimated as $a = 2.113$ and $b = 3.200$, respectively. Given $y^* = 0.5$, the classification cutoff point for PTSD and non-PTSD groups was obtained at $x = -0.504$. The third approach used the Bayesian discrimination function, which minimizes expected risk. Using the zero-one loss function, the decision boundary becomes

$$g_i(x) = P(C_i|x) = \frac{P(C_i)p(x|C_i)}{p(x)}, \tag{5.2}$$

where $P(C_i)$ is the prior probability (i.e., the prevalence of PTSD or non-PTSD in the total sample); $p(x \mid C_i)$ represents the class likelihood (we assumed the latent trait scores have a normal distribution); and $p(x)$ indicates the marginal probability of observation x. Given the assumption of normal distribution in both PTSD and non-PTSD groups, denoted as $PTSD \sim N(0.276, 0.731)$ and $NONPTSD \sim N(-0.874, 0.521)$ respectively, using Formula (5.2), we derived the cutoff point at -0.577. Finally, we calculated the average of these three cutoff points, $\left[(-0.341) + (-0.504) + (-0.577)\right]/3 = -0.474$ as the cutoff point on the latent scale.

The cutoff diagnostic classification results on the new latent scale are analogous to the original inferences. The impact of DIF can be evaluated by estimating the diagnosis agreement, e.g., Cohen's Kappa (Cohen, 1960), between classification with and without group-specific item parameters. Kappa represents the level of agreement between two tests (raters) beyond that accounted for by chance alone. The Kappa coefficient is defined as

$$k = \frac{\Pr(o) - \Pr(e)}{1 - \Pr(e)}, \tag{5.3}$$

where $\Pr(o)$ is the relative observed agreement among tests, and $\Pr(e)$ is the hypothetical probability of chance agreement. The Kappa coefficient is always less than or equal to 1 (1 = perfect agreement). The universal guideline for using Kappa is that 0–0.20 is slight, 0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is substantial, and 0.81–1 is almost perfect agreement (Landis & Koch, 1977).

Table 5.2. DIF Analysis in Gender Using 2PL Model

Item	Question in NCS-R	Item parameters		LM	Prob	Group Dif	Male		Female	
		α (SE)	β (SE)				Obs	Exp	Obs	Exp
B1	Did you ever have repeated unwanted memories of the event, that is, you kept remembering it even when you didn't want to?	1.12 (0.21)	-2.32 (0.19)	5.94	0.01	0.04	0.82	0.80	0.70	0.76
B2	Did you ever have repeated unpleasant dreams about the event?	0.88 (0.15)	-0.91 (0.11)	2.84	0.09	0.03	0.69	0.64	0.67	0.68
B3	Did you have flashbacks, that is, suddenly act or feel as if the event were happening over again?	1.18 (0.16)	-0.74 (0.12)	0.78	0.38	0.02	0.56	0.58	0.65	0.64
B4	Did you get very upset when you were reminded of the event?	1.13 (0.20)	-1.69 (0.16)	3.19	0.07	0.02	0.88	0.85	0.86	0.87
B5	When you were reminded of the event, did you ever have physical reactions like sweating, your heart racing, or feeling shaky?	1.30 (0.18)	-0.89 (0.13)	1.01	0.31	0.02	0.57	0.60	0.67	0.66
C1	After the event, did you try not to think about it?	0.67 (0.17)	-1.59 (0.13)	0.01	0.91	0.00	0.79	0.79	0.81	0.81
C2	After the event, did you purposely stay away from places, people or activities that reminded you of it?	1.17 (0.16)	-0.88 (0.13)	0.41	0.52	0.01	0.63	0.61	0.66	0.66
C3	After the event, were you ever unable to remember some important parts of what happened?	0.70 (0.11)	0.30 (0.10)	7.65	0.01	0.05	0.31	0.39	0.46	0.43
C4	After the event, did you lose interest in doing things you used to enjoy?	1.41 (0.19)	-0.87 (0.14)	0.18	0.68	0.01	0.60	0.59	0.64	0.65

C5	After the event, did you feel emotionally distant or cut-off from other people?	1.57 (0.22)	-1.45 (0.17)	4.90	0.03	0.03	0.63	0.68	0.75	0.73
C6	After the event, did you have trouble feeling normal feelings like love, happiness, or warmth toward other people?	1.68 (0.22)	-1.16 (0.16)	0.16	0.69	0.01	0.63	0.62	0.68	0.68
C7	After the event, did you feel you had no reason to plan for the future because you thought it would be cut short?	1.31 (0.16)	0.79 (0.12)	16.63	0.00	0.07	0.39	0.29	0.32	0.35
D1	During the time this event affected you most, did you have trouble falling or staying asleep?	0.79 (0.19)	-1.86 (0.15)	1.58	0.21	0.02	0.85	0.84	0.79	0.82
D2	During the time this event affected you most, were you more irritable or short-tempered than you usually are?	0.91 (0.13)	-0.49 (0.11)	4.03	0.05	0.04	0.58	0.60	0.61	0.55
D3	During the time this event affected you most, did you have more trouble concentrating or keeping your mind on what you were doing?	1.13 (0.19)	-1.53 (0.15)	0.02	0.90	0.00	0.78	0.77	0.73	0.74
D4	During the time this event affected you most, were you much more alert or watchful, even when there was no real need to be?	0.90 (0.16)	-1.18 (0.13)	1.43	0.23	0.02	0.72	0.73	0.73	0.69
D5	During the time this event affected you most, were you more jumpy or easily startled by ordinary noises?	1.22 (0.19)	-0.99 (0.14)	2.31	0.13	0.03	0.70	0.68	0.59	0.63

Note. The item parameters were estimated from 2PL model without taking DIF into consideration. The columns LM and Prob give the value of the LM statistics and the significance probabilities of the LM test, respectively. Group Dif gives the absolute value of effect size averaged across the two gender groups. Obs and Exp are the observed and expected scores by the model, respectively.

5.3 Results

5.3.1 DIF Related to Gender

The DIF detection results for gender are shown in Table 5.2. The third and fourth columns give the item discrimination (α) and difficulty (β) parameters and the standard errors in estimation. The columns labeled LM and Prob give the value of the LM statistics and the significance probabilities of the LM test, respectively. The last four columns give the observed and expected average item scores for male and female respondents, respectively. Finally, the column labeled Group Dif gives the effect size averaged across the two gender groups. Four items (symptom criteria), corresponding to Criteria B1, C3, C5, and C7, were identified as the DIF items (highlighted in gray) because of significant LM statistics ($p < 0.05$). No items had an effect size larger than 0.10. The maximum effect size (0.07) was found in Criterion C7 (i.e., sense of foreshortened future). It was noted that the majority of differential symptom functioning by gender occurred within symptom cluster C; that is, the greatest difference between the male and female PTSD respondents was in the area of avoidance and numbing.

Table 5.3 presents the estimates of gender-specific item parameters for DIF. The differences between group parameters suggested that female respondents were more likely to feel emotionally distant (C5) whereas male respondents were more likely to have a sense of a foreshortened future (C7) as well as repeated unwanted memories (B1). These findings were consistent with those of Palm et al. (2009). DIF was also found in item C3 ($\beta_{male} = 0.562, \beta_{female} = 0.177, DIF = 0.385, p < 0.05$); that is, female respondents were more likely to be unable to remember important aspects of the stressful event. It was also noticeable that the male group had a much higher discriminate parameter in C5 than the female group ($\alpha_{male} = 3.271, \alpha_{female} = 1.304$, $DIF = 1.967, p < 0.05$) which suggests that the symptom of feeling emotionally distant is more discriminative to distinguish PTSD and non-PTSD for males than females.

The results showing the impact of DIF regarding the population distributions are presented in Table 5.4. This table gives the estimates of mean and variance of the distribution of male respondents, without group-specific item parameters (row P0) and with group-specific item parameters (row P1) for the DIF items. Note that in row P0, the two-sided hypothesis that the estimated mean of male respondents equals zero has

to be rejected at 5% (the critical values are $\pm(0.107\times1.96)=\pm0.210$). In row P1, the hypothesis also must be rejected; that is, the estimated mean of male respondents is not equal to the mean of female respondents (at zero). So the introduction of gender-specific item parameters demonstrates that DIF did not have an impact in this sample.

Table 5.3. Estimation of Item Parameters for DIF Items in Gender

Item	Male				Female			
	α	$SE(\alpha)$	β	$SE(\beta)$	α	$SE(\alpha)$	β	$SE(\beta)$
B1	0.938	0.357	-2.337	0.260	1.248	0.255	-1.914	0.208
C3	1.301	0.369	0.562	0.226	0.550	0.121	0.177	0.102
C5	3.271	1.177	-0.998	0.736	1.304	0.214	-1.444	0.167
C7	1.612	0.398	0.102	0.232	1.474	0.197	1.050	0.149

Note. α indicates the item discrimination parameter, $SE(\alpha)$ is the standard error of α estimation. β indicates the item difficulty parameter, $SE(\beta)$ is the standard error of β estimation.

Table 5.4. Estimation of Gender (Male) Population Parameters

Pattern	$\hat{\mu}$	$SE(\hat{\mu})$	σ	$SE(\sigma)$	Specified DIF Items
P0	-0.261	0.107	0.891	0.094	
P1	-0.280	0.100	0.763	0.096	B1, C3, C5, C7

Note. The table shows the population parameters of male group. The female group is set as the reference group with mean and variance equal to 0 and 1, respectively. P0 is the baseline before the introduction of gender specific item parameters. P1 is the pattern with gender specific parameters for the DIF items defined in this study.

The implications of gender-related DIF on diagnostic results for individuals are presented in Table 5.5. Analogous to the previous analysis, we labeled the pattern without and with group-specific item parameters as P0 and P1, respectively. The

Kappa coefficient of agreement was nearly perfect (0.942) between P0 and P1, suggesting that the introduction of group-specific item parameters had little impact on the diagnostic results.

Table 5.5. Diagnostic Confusion Matrix Between P0 and P1

		P0		
		PTSD	NONPTSD	Total
P1	PTSD	422	8	430
	NONPTSD	10	232	242
Total		432	240	672

Kappa=0.942

Note. P0 is the pattern before the introduction of gender specific item parameters. P1 is the pattern with gender specific parameters for the DIF items defined in this study.

5.3.2 DIF Related to Marital Status

Only one item, B1 (i.e., repeated unwanted memories), was marked DIF in the analysis of marital status on the basis of the significant results of the LM test ($p < 0.05$). The item characteristic curves (ICC) of item B1 with and without group-specific parameters are exhibited in Figure 5.1. The horizontal axis indicates the latent score of individuals' PTSD level, while the vertical axis indicates the probability of individuals giving a positive response to the item. The solid line indicates the ICC without using the group-specific parameters, whereas the dashed line (with marriage experience) and dotted line (without marriage experience) indicate the ICCs with the group-specific parameters. Note that the item difficulty parameter differed moderately between those with and without marriage experience ($\beta_{mar} = -2.609, \beta_{nevmar} = -1.546, DIF = 1.063, p < 0.05$) when using the group-specific item parameters. It suggested that those who have been married (line veers to left of solid line) were more likely to have unwanted memories than those who have never been married (line veers to right of solid line).

Analogous to the process of estimating the impact of gender-related DIF, we identified the latent scale by setting the mean and variance of the latent distributions of

the never-married group equal to 0 and 1, respectively. The two-sided hypothesis that the estimated means of the married group ($\hat{\mu}_{Mar} = -0.156, SE(\hat{\mu}_{Mar}) = 0.133$) and separated/widowed group ($\hat{\mu}_{Sep} = -0.044, SE(\hat{\mu}_{Sep}) = 0.134$) are equal to zero was not rejected at 5%, regardless of whether or not the group-specific item parameters were used. Further, the Kappa coefficient for diagnosis agreement between tests with and without group-specific item parameters was 0.941, a nearly perfect agreement. Therefore, although we did detect the DIF item B1 by evaluating for marital status, the impact of DIF for this item on PTSD diagnosis was extremely small.

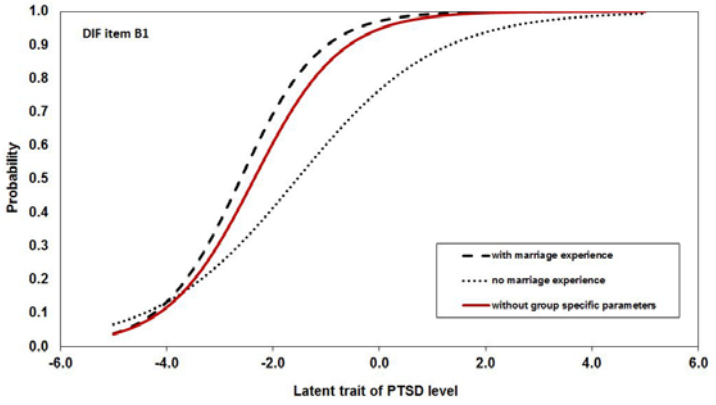


Fig. 5.1. Item characteristic curves of item B1 with and without marriage-specific item parameters
Note. The horizontal axis indicates the latent score of individuals' PTSD level while the vertical axis indicates the probability of individuals giving a positive response to item B1.

5.3.3 DIF Related to Educational Level

Three items, C5 (i.e., feeling of emotionally distant), D2 (i.e., irritability) and D5 (i.e., exaggerated startle response) were detected as DIF related to educational background on the basis of the significant results of the LM tests ($p < 0.05$). The ICCs of these three items with and without group-specific parameters are plotted in Figure 5.2. The results suggested that the lower-educated respondents identified as PTSD were more easily startled (D5) than the higher-educated respondents, whereas the higher-educated respondents were more likely to feel emotionally distant from others (C5) than the lower-educated patients. The respondents with 13–15 years of education were more likely to be irritable (D2) than the other groups.

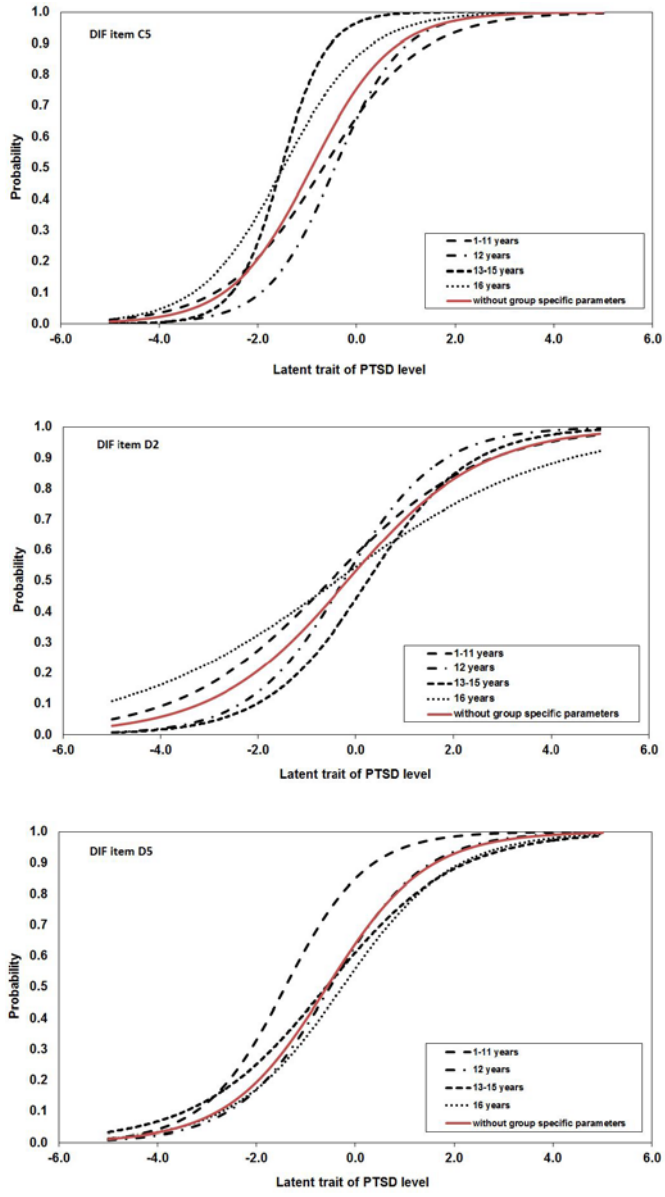


Fig. 5.2. Item characteristic curves of item C5, D2, and D5 with and without education-specific item parameters
Note. The horizontal axis indicates the latent score of individuals' PTSD level while the vertical axis indicates the probability of individuals giving a positive response to the item.

The impact of DIF related to educational level was also assessed via two methods. With the implication regarding population distributions, we identified the latent scale by setting the mean and variance of the latent distributions of the group with over 16 year of education equal to 0 and 1, respectively. The two-sided tests of the hypothesis that the means of the other groups were equal to zero were not rejected at 5%, regardless of whether or not the group-specific item parameters were used. Further, with respect to the implications for individuals, the Kappa coefficient of diagnosis agreement was more than 0.90 between tests with and without group-specific item parameters. Thus, we concluded that the impact of DIF related to educational background was negligible.

5.4 Discussion and Conclusion

This chapter explored DIF analyses for the background variables of gender, marital status, and educational level with respect to differential symptom functioning in persons with PTSD as determined by DSM-IV criteria. Besides detecting DIF items, this article emphasized the importance of evaluating the impact of DIF, and introduced two methods for estimating the impact of DIF on inferences regarding population distributions and diagnosis of individuals. The results showed that four symptoms for gender (B1, C3, C5, and C7), one symptom for marital status (B1), and three symptoms for educational level (C5, D2, and D5) were significantly identified as DIF. But the impact of DIF was marginal and had a negligible impact on diagnosis.

Current findings partly reconfirm the conclusions from the previous studies by Palm et al. (2009). In their study, only the comparison across gender was taken into account and 7 symptoms including B1, B2, B3, C5, C7, D2, and D5 were detected as DIF. Three symptoms (B1, C5 and C7) were repetitively identified as DIF in our study. To take a further look at the Table 5.2, it was noticeable that although the item B2 and D2 were not significantly labeled as DIF in the present study, they were pretty close to the significance level 0.05. Consistent with the findings of Martin et al. (2007), the symptom B1 was detected bias between the patients with and without marriage experience, that is, patients with marriage experience (married or divorced) were more likely to repeat the unwanted memories than those who never got married. The probable reason was interpreted as patients who have ever married might have more chance of exposure to traumatic events, such as marital rape and domestic violence (Martin, Taft, & Resick 2007). Dekel and Solomon (2006) further found that the PTSD

patients with marriage experience exhibited decreased marital satisfaction, increased verbal aggression and heightened sexual dissatisfaction than those in single status. Educational background is also an important factor to be included in the study of PTSD. For instance, Shih and his group (2010) examined prevalence rates of PTSD in a sample of 677 individuals experiencing different types of trauma and found that people who reported having less than a high school education were more likely to develop PTSD than their more educated counterparts. However, the differential endorsement of PTSD symptoms across education groups was seldom investigated in the existing literatures. Additionally, we noted that symptom B1 was marked DIF for both gender and marital status and that symptom C5 was marked DIF for both gender and educational level. Further studies could be conducted to determine whether there are interaction effects for symptoms B1 and C5 in these corresponding demographic variables.

Whether the DIF makes a difference to diagnosis is fairly important to clinical practice. Teresi and Fleishman (2007) emphasized that besides identifying specific items exhibiting DIF, additional research was also needed to examine the impact of DIF in health-related measures. The methods for detecting DIF and assessing DIF impact presented in this chapter can be generally applied to a broader scope in psychiatric and clinical fields for checking the validity of diagnostic criteria and instruments. For instance, the methods could be used in the functional disability measurement, where differences of symptoms across demographic groups (e.g., age, gender) could be identified and the impact of DIF could be estimated. The results from the DIF analysis can play an important role in estimating the anticipating demands for services (Fleishman, Spector, & Altman, 2002).

The major difference between the two methods for assessing DIF impact is the measurement level, one focuses on a group level and the other on an individual level. The first approach to assessing impact is in terms of effect size. That is, an analysis is conducted to examine how much mean group differences in total score distributions change with and without using specific group item parameters. This method is simple and straightforward to see whether the group means have significant changes in different patterns, but rather difficult to describe the DIF impact on individuals. The second approach focuses on an individual level, that is, to compare the accuracy rate with and without using group specific item parameters. It is a bit more complicated than the first one, since the optimal cut off points need to be estimated on a latent scale. In the daily practices, we recommend to use both of these two methods to make a thorough investigation.

Another issue that worth concerns is the prerequisite of using the methods. The LM test statistics are closely related to the analysis of cross tabulations (Glas & Falcon, 2003). For instance, the LM statistic targeted at DIF splits up the sample of respondents into G groups, and the count in each cell, that is, for each group, should be at least 5. However, such low frequencies do not result in an acceptable power to detect item misfit. Glas and Falcon (2003) suggested that a sample size of 500 respondents was required to obtain the acceptable power.

In summary, on the one hand we conclude that the DSM-IV PTSD diagnostic criteria do not produce substantially biased results in the three investigated subpopulations, and therefore there should be few reservations regarding their use. On the other hand, we recommend that diagnosticians always perform a DIF analysis of various subpopulations using the methodology presented here to ensure that the diagnostic criteria is valid to use for the target population.

Chapter 6

Combining Textual Assessments of Self-Narratives and Item-Based Measures to Screen for Posttraumatic Stress Disorder

Abstract

In this chapter, a new intake procedure for PTSD screening is developed, which combines an automated textual assessment of patients' self-narratives and measures on itemized instruments. Text mining and item response modeling are used to analyze the patients' writings and responses to standardized questionnaires, respectively. The whole procedure is combined in a Bayesian framework where the textual assessment functions as an informative prior for the estimation of PTSD latent trait. A sample of 99 trauma survivors is used to examine the performance of this combination. Results show that adding textual assessments to the item-based measures increase the screening accuracy, decrease the standard error of estimation, and can shorten the length of follow-up test. These findings highlight the benefits of combining textual assessment and item-based measures in a psychiatric screening process. From the results, we may conclude that this combination approach is promising to improve the cost-effectiveness of a screening procedure and to reduce both patients' burden and clinicians' workload.

Keywords: posttraumatic stress disorder; text mining; item response theory; Bayesian framework; self-narratives

This chapter has been submitted as: He, Q., Veldkamp, B. P., Glas, C. A. W, & vern der Berg, S. M. (2013). Combining textual assessments of self-narratives and item-based measures to screen for posttraumatic stress disorder (PTSD).

6.1 The Combination of Self-Narratives and Itemized Instruments

Epidemiological research on mental illnesses such as posttraumatic stress disorder (PTSD) requires efficient methods to identify cases in large population-based samples (Shrout & Yager, 1989). If the diagnosis of the disorder is difficult or expensive, a two-phase design may be employed, where in the first phase a screening measure is employed, which is then followed by a more detailed diagnostic procedure for a selected sub-sample (Diamond & Lilienfeld, 1962; Shrout, Skodol, & Dohrenwend, 1986).

Item-based self-report instruments are often considered efficient for PTSD screening, as they usually require short administration time and do not require the presence of a clinician (Wohlfarth, van den Brink, Winkel, & ter Smitten, 2003). In the last decades, a large number of self-report PTSD instruments have been developed, such as the Impact of Events Scale (IES; Horowitz et al., 1979), the PTSD Symptom Scale Self-Report version (PSS-SR; Foa, Riggs, Dance, & Rothbaum, 1993), the PTSD Checklist – Civilian version (PCL-C; Weathers et al., 1991), the Davidson Trauma Scale (DTS; Davidson et al., 1997), the Harvard Trauma Questionnaire (HTQ; Mollica et al., 1992), the Revised Civilian Mississippi Scale (Norris & Perilla, 1996), and many others (for a review, see Norris & Riad,1997; Brewin, 2005). 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).

As an alternative to such questionnaire-based screening, He et al. (2012) recently developed a computerized textual assessment system for PTSD screening using text mining techniques. These techniques can be used for open-ended writings regarding trauma history and physical symptoms. Based on their textual input, that is, based on the self-narratives describing the traumatic experiences and their impacts on personal life, participants were predicted to be either diagnosed with PTSD or not. The textual screening procedure resulted in a good agreement (82%) compared with a clinical structured interview in identifying the presence and absence of PTSD, and yielded a higher sensitivity and positive prediction power than an itemized screening instrument.

The focus of this chapter is to assess to what extent text mining techniques can be applied in the screening phase, and to establish the extent to which it results in better estimates and better prediction of true diagnosis, compared to the use of a questionnaire alone. We propose to combine text mining data and questionnaire data in

a Bayesian framework, where a score based on text mining serves as input for a prior distribution of a latent trait associated with PTSD that is measured by a number of questionnaire items using an Item Response Theory (IRT) model. Because of its flexibility, the IRT framework (Lord, 1980; Rasch, 1960) is increasingly applied in psychiatric assessments currently (e.g., He, Glas, & Veldkamp, 2013; van Groen et al., 2010; Weisscher et al., 2010). In contrast to the classical sum score methods, IRT provides improvement in scaling items and people. The IRT models generally scale the difficulty of items and the latent trait level of people on the same metric. Namely, the severity of prescribed symptoms and the latent degree of individuals' mental illness are set on a common scale, and thus can be meaningfully compared.

A combination of IRT and text mining is an interesting approach to increase measurement precision in the screening phase. Better precision in the screening phase will lead to better prediction, in turn leading to a smaller subsample that needs to be clinically assessed in the diagnostic phase. In this study, we propose to use Bayesian methods which are especially useful for the estimation of a hierarchical structure. The results derived from the textual assessments can function as an informative prior for estimation of a latent trait in an IRT model (refer to Mislevy, 1986; Zwinderman, 1991). Several studies in educational assessment use background variables, such as scores obtained by the examinees from other tests or testlets, socio-economic, and demographic variables, as informative priors to improve the accuracy of ability estimates (e.g., Matteucci & Veldkamp, 2012). Also in clinical assessment, van den Berg et al. (2013) combined self-report and clinical interview data on schizotypal symptoms in order to increase measurement precision. However, the inclusion of textual assessments as prior information has been rarely described in the literature.

To examine the performance of our proposed method, we conducted a study to compare the estimates for a latent trait associated with PTSD with and without the use of a text mining score. Two specific objectives were investigated (1) whether the addition of textual assessment could enhance the efficiency of PTSD detection, and (2) whether the length of a follow-up item-based test could be shortened as compensation from a text prior.

6.2 Method

6.2.1 Sample

Data were collected from 105 trauma survivors via an online survey embedded in an open forum that is dedicated to people with mental health issues. Before administering items from the survey, all the participants were asked to report whether they had been diagnosed as PTSD or NONPTSD by psychiatrists via structured interviews with standardized instruments. Those with missing diagnoses were discarded in the present study. Participants were informed that the objective of the research was to develop a more friendly and flexible intake procedure for PTSD diagnosis and were requested to give responses to all the questions following the instructions. The survey consisted of two parts, self-narrative writing and administration of dichotomous questions regarding PTSD symptoms. In the writing section, respondents were asked to write down their traumatic events and briefly describe the symptoms related to these experiences. A text length was suggested to be over 150 words. In the item section, the questionnaire that was employed for PTSD screening in the National Comorbidity Study-Replication (NCS-R; Kessler et al., 2004) was used. The NCS-R, conducted between February 2001 and April 2003 in the United States, is a nationally representative community household survey of the prevalence and correlates of mental disorders (see details of NCS-R refer to Kessler et al., 2004). In this questionnaire, the PTSD symptoms are assessed by dichotomous items (i.e., “yes”=1, “no”=0). Each item corresponds to one symptom criterion listed in the DSM-IV. A total of 21 items corresponding to Criteria A through F were included in the current study (see more PTSD diagnostic criteria refer to DSM-IV; APA, 2000).

Six of the 105 participants were excluded because two of them reported that they had never experienced traumatic events that were listed in the NCS-R, and four gave responses only to the item section but missed the writing section. This resulted in a total of 99 participants for the final set, among whom 34 were diagnosed as PTSD and 65 as NONPTSD. The sample had an age range between 19 and 63 years, with a mean of 30.06 (S.D.=11.30). The majority of participants were female (78.4%). Over 90% participants had a higher educational background (i.e., college / university or above). As for the marital status, 52.6% participants were single, 40.2% were married, and 6.2% were divorced.

6.2.2 Procedure

We estimated the individual's PTSD latent trait via three approaches: (1) using the IRT modeling on the 21-item NCS-R questionnaire, (2) using the text classification of self-narratives, and (3) combining textual analysis and IRT in a Bayesian framework. All analyses were made using the software WinBUGS 1.4.3 (Lunn, Thomas, Best, & Spiegelhalter, 2000).

6.2.2.1 Approach 1: IRT Modeling on the Itemized Questionnaire

Responses to the 21 PTSD related items in the NCS-R were used as input for the statistical analysis. A unidimensional 2PL model was used to estimate the individual's latent trait level θ_n (Lord, 1980). In this model, the probability of a score in category "yes" ($X_{ni} = 1$) of item i is given by the item response function

$$P(X_{ni} = 1 \mid \theta_n) = \frac{\exp[\alpha_i(\theta_n - \beta_i)]}{1 + \exp[\alpha_i(\theta_n - \beta_i)]}, \tag{6.1}$$

where θ_n is the latent PTSD level of person n, β_i is an item difficulty parameter representing the severity level of each diagnostic symptom, and α_i is an item discrimination parameter indicating the extent to which the item response is related to the latent θ-scale. The item parameters were fixed in the current study. They were calibrated by marginal maximum likelihood (MML; Bock & Aitkin, 1981) on a sample of 880 respondents in an earlier study (He et al., 2013). As shown in Table 6.1, the discrimination parameters varied in the interval [0.775, 1.864], with a mean value around 1.3. The difficulty parameters were included in the range [−4.447, 1.215], with a mean of −1.0. The respondents' latent traits were estimated by expected a posteriori (EAP) assuming a normal distribution.

6.2.2.2 Approach 2: Text Classification of Self-Narratives

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 & Schütze, 1999). Supervised text classification generally involves two phases, a training phase and a testing phase. During the training phase, the most discriminative keywords to determine the presence or absence of PTSD are extracted and the relationship between

the keywords and class labels is learned. The testing phase involves checking how well the trained classification model performs on a new dataset. In the testing procedure, each new input is scanned for the keywords which were extracted from training, and the most likely label for each new self-narrative is predicted.

Table 6.1. Item Parameters of 21 Questions Related to PTSD in NCS-R (Calibrated with n=880)

Item	Question in NCS-R	α	SE(α)	β	SE (β)
A2	Did you feel terrified or very frightened, helpless, shocked or horrified, numb at the time?	1.19	0.41	-4.45	0.48
B1	Did you ever have repeated unwanted memories of the event, that is, you kept remembering it even when you didn't want to?	1.82	0.20	-1.74	0.15
B2	Did you ever have repeated unpleasant dreams about the event?	1.24	0.14	-0.49	0.10
B3	Did you have flashbacks, that is, suddenly act or feel as if the event were happening over again?	1.41	0.15	-0.22	0.10
B4	Did you get very upset when you were reminded of the event?	1.64	0.18	-1.18	0.12
B5	When you were reminded of the event, did you ever have physical reactions like sweating, your heart racing, or feeling shaky?	1.68	0.17	-0.34	0.11
C1	After the event, did you try not to think about it?	0.95	0.12	-1.31	0.11
C2	After the event, did you purposely stay away from places, people or activities that reminded you of it?	1.34	0.14	-0.45	0.10
C3	After the event, were you ever unable to remember some important parts of what happened?	0.83	0.10	0.58	0.08
C4	After the event, did you lose interest in doing things you used to enjoy?	1.53	0.15	-0.39	0.10
C5	After the event, did you feel emotionally distant or cut-off from other people?	1.55	0.16	-0.88	0.11
C6	After the event, did you have trouble feeling normal feelings like love, happiness, or warmth toward other people?	1.86	0.18	-0.55	0.12
C7	After the event, did you feel you had no reason to plan for the future because you thought it would be cut short?	1.45	0.15	1.22	0.12

D1	During the time this event affected you most, did you have trouble falling or staying asleep?	1.14	0.18	-1.53	0.12
D2	During the time this event affected you most, were you more irritable or short-tempered than you usually are?	1.11	0.14	-0.16	0.09
D3	During the time this event affected you most, did you have more trouble concentrating or keeping your mind on what you were doing?	1.47	0.19	-1.10	0.11
D4	During the time this event affected you most, were you much more alert or watchful, even when there was no real need to be?	0.96	0.16	-0.85	0.10
D5	During the time this event affected you most, were you more jumpy or easily startled by ordinary noises?	1.28	0.17	-0.55	0.10
E1	Was any of these reactions continue to have at least one month?	0.78	0.30	-3.30	0.21
F1	Did these reactions cause distress to you?	1.55	0.26	-2.15	0.17
F2	Did these reactions disrupt or interfere with your normal, daily life?	1.02	0.16	-0.88	0.11

Note. The item parameters were estimated from unidimensional 2PL model on a sample of 880 respondents in the NCS-R. SE indicates the standard error of item parameter estimation.

He et al. (2012) developed a supervised text classification model for PTSD screening. In this previous study, 300 self-narratives, consisting of 150 written by the PTSD people and 150 written by the NONPTSD people, were used to develop a screening system. Three machine learning algorithms, including Decision Tree (DT), Naïve Bayes (NB) and a self-developed alternative, the product score model (PSM), were employed in conjunction with an *n*-gram representative model. The textual data are usually encoded via a data representation model. Unigram is the simplest and most commonly used data representation model where each word in a document collection acts as a distinct feature. *N*-gram considers the interaction effect of two, three or more consecutive words (Manning & Schütze, 1999). The PSM with unigrams resulted in the highest agreement with the psychiatrists' diagnoses in clinical practice. Therefore, we applied this model in the present study, that is, the self-narratives were classified using the PSM with the 1000 most predictive unigrams that were extracted in the study of He et al. (2012).

PSM is an alternative machine learning algorithm, derived from Naïve Bayes by using a smoothing constant for computing word probability. It features in assigning

two weights for each keyword (in binary classification) to indicate how popular the keywords are in the corpora of self-narratives written by either PTSD patients (corpus C_1) or self-narratives written by NONPTSD patients (corpus C_2). The name product score comes from a product operation to compute scores for each class, i.e., S_1 and S_2, for each input text based on the term weights. That is,

$$\begin{cases} S_1 = P(C_1) \cdot \prod_{w=1}^{k} [(u_w + a) / len(C_1)] \\ S_2 = P(C_2) \cdot \prod_{w=1}^{k} [(v_w + a) / len(C_2)] \end{cases}, \tag{6.2}$$

where u_w and v_w are the number of occurrences of keyword w in both corpora C_1 (i.e., PTSD corpus) and C_2 (i.e., NONPTSD corpus), respectively. $len(C)$ is the corpus length, namely, the sum of the word occurrences in each corpus. The classification rule is defined as:

$$\text{choose} \begin{cases} C = 1 & \text{if } \log(S_1 / S_2) > b \\ C = 2 & \text{else} \end{cases}, \tag{6.3}$$

where b is a constant. In this study we set $b = 0$, because in the previous study (He et al., 2012) it was found that during the PTSD textual screening procedure that the largest number of positive cases could be captured without unduly sacrificing specificity when the threshold was set at zero. The value of $\log(S_1 / S_2)$ was defined as the text score for each self-narrative (see also He & Veldkamp, 2012; He et al., 2012). For an easy comparison with the IRT scales, we standardized the text scores as $Z \sim N(0,1)$.

6.2.2.3 Approach 3: Combining Textual Analysis and IRT in a Bayesian Framework

Textual analysis and item response modeling were combined in a Bayesian framework using the Markov Chain Monte Carlo (MCMC) algorithm, where the text score of each self-narrative obtained in approach 2 was used as an informative prior. The posterior distribution of the latent PTSD level is proportional to the product of the prior and the likelihood, that is,

$$P(\theta \mid x, y) \propto p(x \mid \theta, \alpha, \beta) g(\theta \mid y), \tag{6.4}$$

where x is the vector of responses to the questionnaire, y is the text score for each individual, $g(\theta \mid y)$ is the prior given the covariate of textual assessments, α and β are the fixed discrimination and difficulty parameters of items, $p(x \mid \theta, \alpha, \beta)$ is the likelihood function of the IRT model. The relation between the PTSD latent trait θ of individual n and the text score y_n is given by the linear regression

$$\theta_n = b_0 + b_1 y_n + \varepsilon_n, \tag{6.5}$$

where b_0 and b_1 are the regression coefficients. The error terms are assumed to be independent and normally distributed as $\varepsilon_n \sim N(0, \sigma^2)$ with $n = 1, ..., N$ individuals. The assumption of a linear regression model is translated into a normal conditional distribution of θ_n given the text covariate as

$$\theta_n \mid y_n \sim N(b_0 + b_1 y_n, \sigma^2) \ . \tag{6.6}$$

Formula (6.6) represents an informative prior distribution of the PTSD latent trait. For each individual, the estimation of latent trait was performed by using 5000 MCMC iterations with the burn-in of length of 1000.

To determine whether the introduction of the prior distribution was effective, we compared the posterior distribution of θ_n in this approach with the estimation from approach 1. Because the item parameters in the IRT model were fixed, the θ-estimates resulting from both of the two approaches were on a common scale and thus could be compared.

6.2.3 Analytic Strategy

Two investigations were conducted to analyze the efficiency of the combination model. The first was to combine the textual assessments with the full range of 21 items of the NCS-R questionnaire. The main purpose was to explore whether adding the text prior would make significant impact on the accuracy of PTSD detection. The second

investigation pursued the question whether adding textual assessments to the questionnaire could result in a reduction of the number of items administered without sacrificing precision of the θ-estimates. Precision of the θ-estimates is inversely proportional to the amount of information provided by the items. Those items that provide peak information around the cutoff threshold are ideal for a shorter version of a mastery test (Thomas, 2011). Since the target of screening is to make classification decisions, a natural choice would be to maximize information at the chosen diagnostic cutoff (for more about item information refer to Lord, 1980). We estimated a cutoff point to distinguish the PTSD and NONPTSD group on the IRT scale at $\theta = -0.153$ using the standard setting methods presented in Chapter 5. Consequently, in the present study, we calculated the item information for all the 21 items at this cutoff point and ranked the items in a descending order, that is, starting from the item with the highest information until the least information (see Figure 6.1). The items were ordered as following: C6, B5, C4, B3, C5, C2, D5, B2, B4, D3, D2, F2, C7, D4, D1, C1, B1, C3, F1, E1, A2. We started to examine the performance of a combination of the text prior and the most informative item, that is, text prior with item C6 (i.e., "Did you have trouble feeling normal feelings like love, happiness, or warmth toward other people?") versus using item C6 alone. The second informative item (B5) was then added in for the comparison of the next pattern. The procedure continued until all the 21 items were included. Both test information and standard error of θ-estimates were calculated for each pattern (i.e., with and without text prior) with an increasing number of informative items. Since textual assessment was suggested as a sort of complementary information to predict people's physical and mental health (e.g., Gottschalk & Gleser, 1969; Rosenberg & Tucker, 1979; Franklin & Thompson, 2005; Smyth, 1998), the test information was expected to increase and the standard errors were expected to decrease when text priors were added.

Further, the performance of the three approaches was compared on five metrics; accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). The diagnoses made in the structured interviews by psychiatrists were used as the true standard in the comparison. Accuracy, the main metric used in classification, is the percentage of correctly defined individuals. Sensitivity and specificity are the proportion of actual positives and actual negatives that are correctly identified, respectively. These two indicators do not depend on the prevalence in the sample (i.e., proportion of "PTSD" and "NONPTSD" of the total), and hence are indicative of real-world performance. The predictive values, PPV and

NPV, are estimators of the confidence in predicting correct classification; that is, the higher predictive values are, the more reliable the prediction is.

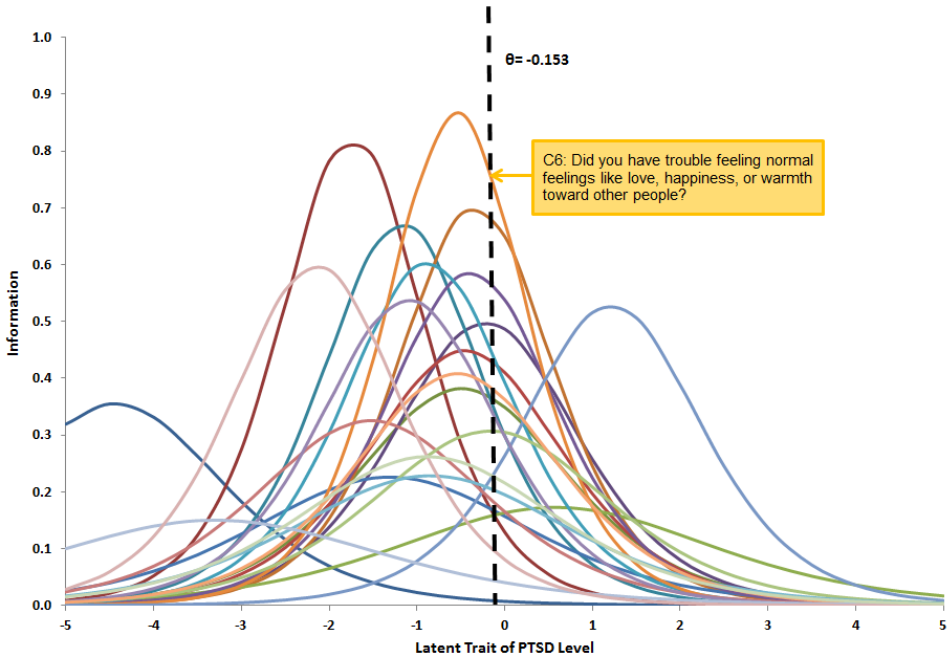


Fig. 6.1. Item information for 21 items in NCS-R questionnaire corresponding to DSM-IV PTSD diagnosis criteria
Note. The cutoff point was estimated at –0.153 on latent scale to distinguish PTSD and NONPTSD. Item C6 is the most informative item which has the highest intersection value with the cutoff line.

6.3 Results

For the sample of 99 participants, the latent trait estimation via approach 1 resulted in a normal distribution of latent trait levels θ_n, with a mean value of −0.39 and standard deviation of 2.31. The standardized text scores obtained from approach 2 resulted in a range [−2.923, 4.218]. In approach 3, the latent linear regression model given by Formula (6.4) and (6.5) was estimated using the item responses and the textual

covariates. The intercept and slope coefficients were obtained as −0.41 and 1.44, respectively. The error term had a normal distribution with a mean value of zero and variance as 3.57. Hence, the informative prior distribution of the PTSD latent trait was defined as $\theta_n \mid y_n \sim N(-0.41+1.44y_n, 3.57)$.

The correlations among the estimations from the three approaches are presented in Table 6.2. It was noted that the correlation between the EAP of θ-estimates via approach 1 and the text scores estimated via approach 2 was 0.558, suggesting that there was a positive and moderate relation between the self-narrative writing and the responses to the itemized questionnaire. This result reiterated the findings in the earlier studies that the words and expressions were capable to predict one's mental health status.

Table 6.3 shows the performance metrics of the three approaches. As our expectation, the diagnostic accuracy rate was fairly high as 0.94 when using 21-item questionnaire by the IRT alone, and was improved to 0.97 with an addition of textual assessment. Given concerns on only using the keywords as predictors to make the classification, the accuracy rate (0.84) produced by the textual assessment was satisfactorily high, although it was a bit lower than the other two approaches. The sensitivity and NPV were perfect for all the three approaches, implying that both the IRT and the textual assessments were sensitive for identifying PTSD patients. With the introduction of textual assessment, the specificity and PPV rose to 0.95 and 0.92, respectively. It suggested that the textual assessment played an effective role in detecting the NONPTSD and strengthened the power in identifying the PTSD from the population.

Figure 6.2 shows the relationship between the standard error of the estimate of θ and the number of items with the presence or absence of text prior. The horizontal axis indicates the number of items in the IRT model and the vertical axis indicates the average standard error of the latent trait estimation. The curve of standard error without using the text prior (i.e., the dash line), that is, using the IRT model alone via approach 1, starts around 1.6 and drops gradually to 0.68 when all the 21 items are included. The curve of standard error using a text prior (i.e., the solid line) follows the similar pattern but keeps on a lower level than the dash curve. It starts approximately to 1.4 and ends around 0.65. Using a 95% confidence interval, the paired sample *t*-test showed that the standard error of estimation with text prior was significantly lower than that without text prior ($t = 3.856,\ df = 98,\ p < 0.01$) when including the whole range of 21 items. With the increasing number of items, the differences between these two curves decreased from 0.2 to 0.03. It suggested that the textual assessment made an impact on

the latent trait estimation, and the effect was more apparently when using fewer items. The dotted line highlights the standard error when using 21 items without the text prior. It crosses the solid curve at 17 items, implying that with the introduction of the text prior, 17 items would be good enough to make the estimation as precise as using the whole range of 21 items. That is, by using the text priors, the questionnaire length can be shortened by 4 items.

Table 6.2. Correlations Among Estimates from Three Approaches: IRT, TX, IRT&TX

	IRT	TX	IRT & TX(21-item)
IRT	1.000		
TX	0.558	1.000	
TX&IRT(21-item)	0.994	0.623	1.000

Note. TX indicates the textual assessments. Correlation is significant at the 0.01 level (2-tailed).

Table 6.3. Performance Metrics Compared Among IRT, TX and a combination of IRT&TX

	Accuracy	Sensitivity	Specificity	PPV	NPV
IRT	0.94	1.00	0.92	0.87	1.00
TX	0.84	1.00	0.77	0.69	1.00
TX&IRT (21-item)	0.97	1.00	0.95	0.92	1.00

Note. TX indicates textual assessment. PPV and NPV represent the positive predictive value and negative predictive value, respectively.

The effect of the text prior on θ-estimation were further illustrated in Figure 6.3, where both of the test information and standard error of estimation with (i.e., red dots) and without (i.e., blue dots) using the text priors were presented with different test lengths, 5, 10 and 21 items included in the questionnaire. The standard error is inversely related to the test information. As shown in the test information plots, the curve shaped by the red dots was apparently higher than the one by the blue dots in the first panel when only 5 items were used. More overlap was found in the second panel

when 10 items were used. Finally, the curves of the blue and red dots were hardly distinguished in the third panel with the inclusion of the whole range of 21 items. It implied that the text prior played an informative role in the test. Its impact was shown more obviously using fewer items, which strengthened the findings from Figure 6.2. In case of an increasing number of items, the effects of text priors were gradually overwhelmed by the information provided by the new items.

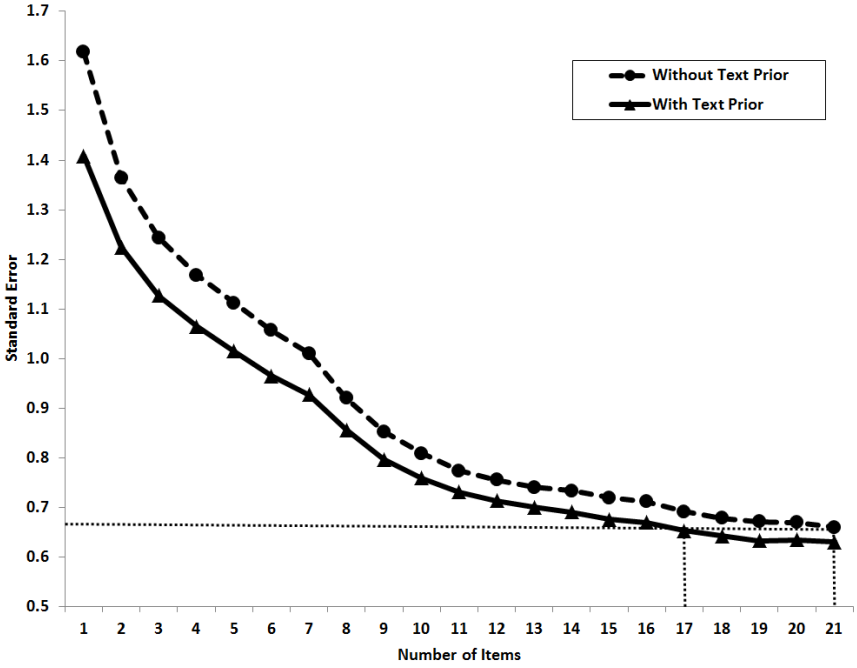


Fig. 6.2. The relationship between standard error of the estimate of θ and the increasing number of items with or without using text priors

6.4 Discussion and Conclusion

In this study, a new intake procedure for PTSD screening was developed that combined an automated textual assessment of patients' self-narratives and IRT based measures of responses to structured questionnaire. To determine whether the introduction of text information is effective, we identified PTSD cases via three approaches: (1) estimated PTSD latent trait by using IRT on a standardized

questionnaire, (2) classified patients' self-narratives into PTSD and NONPTSD groups by using a text mining technique, and (3) estimated the posterior distribution of PTSD latent trait by combining textual assessments and IRT in a Bayesian framework. With the sample at hand, the results showed that the combination model enhanced the accuracy of PTSD detection from 0.94 to 0.97, reduced the standard error of latent trait estimation and could shorten the questionnaire length by four items without sacrificing the accuracy rate.

In the current study the diagnostic accuracy was already high (0.94) when using the itemized questionnaire alone (approach 1). However, a structured interview that generally employs questionnaires is time consuming in daily practice. The computerized textual assessment proposed in this study is relatively easy to conduct online. However, the detection accuracy is comparatively low (0.84), but well acceptable as a screening tool. Given concerns of the cost-effectiveness of the diagnosis, it would be interesting to combine these two approaches in a two-phase framework to reduce the clinical expense and improve the accuracy rate.

Further, according to the results in the previous study of He et al. (2012), the NPV of the textual assessments was satisfactorily high as 0.85 when the text classification algorithm PSM was applied in conjunction with unigrams. It meant that the textual screening tool was helpful to exclude the NONPTSD respondents from the follow up tests. For the 99 sample in the present study, taking the 85% confidence interval, 53 respondents could be excluded from the further tests actually.

In addition, some limitations in the present study merit discussion. First, the sample size was rather small of only 99 participants. Secondly, the sample had an unusually high level of education. This was mainly caused by the Internet data collection approach. People with a higher educational background are possibly easier accessed via a web-based test (Naglieri, 2004).

In conclusion, the current study presented an innovative trial to combine textual assessment of patients' self-narratives and measures on itemized instruments. Its aim was to reduce the respondents' burden and clinicians' workload. Adding textual prior information, the detection accuracy was increased and the test length could be shortened. The results demonstrated that the combination of a textual assessment and an IRT-based questionnaire is a promising approach to increase the cost-effectiveness in the PTSD screening and is expected to be applicable in a broader scope of both (online) screening and psychiatric diagnosis in the future.

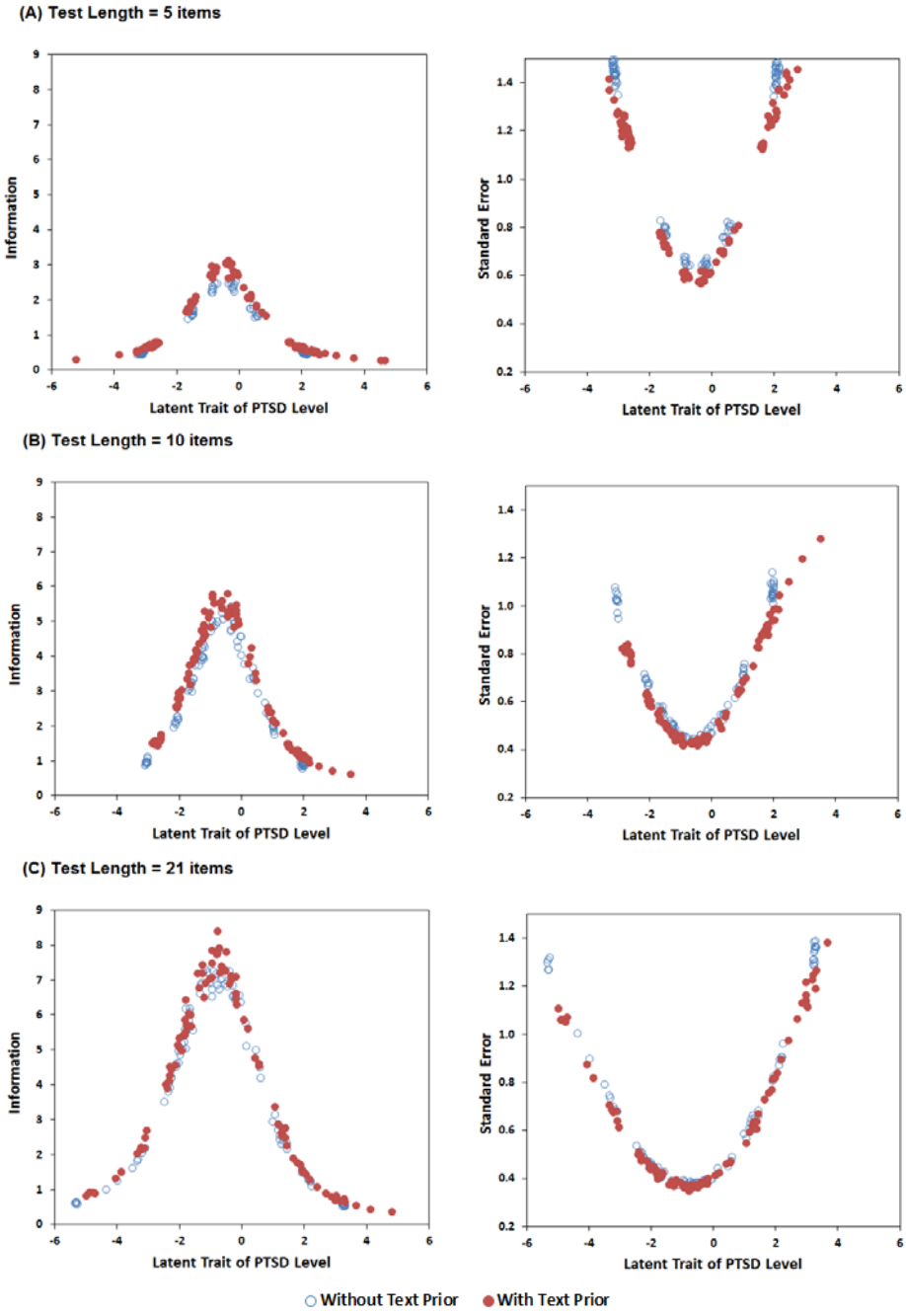


Fig. 6.3. Test information and standard error functions (test length=5, 10, 21 items)

Chapter 7

Predicting Self-Monitoring Skills Using Textual Posts on Facebook

Abstract

The popularity of the social networking site Facebook (FB) has grown unprecedented during the past five years. The research question investigated is whether posts on FB would also be applicable for the prediction of psychological features of the users. We present a model to evaluate the relationship between the posts and self-monitoring skills. The aim of this study is twofold: first, to evaluate the quality of responses to the Snyder's Self-Monitoring (SM) Questionnaire (1974) collected via the Internet; and secondly, to explore the textual features of the posts in different SM-level groups. The prediction of posts result in an approximate 60% accuracy compared with the classification made by Snyder's SM scale. The variable "family" is found the most significant predictor in structured textual analysis via Linguistic Inquiry and Word Count (LIWC). The emoticons and Internet slangs are extracted as the most robust classifiers in the unstructured textual analysis. From the results, we may conclude that the textual posts on the FB Wall can partially predict the users' SM skills. Besides, we recommend that the researchers always check the validity of Internet data using the methodology presented here to ensure the data is valid before being processed.

Keywords: Facebook; self-monitoring; text mining; item response theory; data validation; LIWC

This chapter has been submitted as: He, Q., Glas, C. A. W, & Veldkamp, B. P. (2013). Predicting self-monitoring skills using textual posts on Facebook.

7.1 An Online "Self" on Facebook

As the second most visited website on the Internet (Alexa Internet Inc., 2011), Facebook (FB) attracts a global audience of over 606 million people for a daily use (Gonzalez, 2011). When joining the FB community, the platform requires users to compose an online "self" and allows them to share their emotions and problems by posts on the Wall, which are viewed by the users' self-selected and mediated audience. This composition is inherently an act of self-presentation (Hall & Pennington, 2013), which is "the goal directed activity of controlling information of self in order to influence the impressions formed by audiences" (Schlenker, 2004).

Among all the multimedia formats of the posts, textual input is predominantly used for updating users' status on the FB. The "status updates" are short user-generated public messages that generally contain information about what the FB user is doing or thinking at that point of time, i.e., "what's on your mind?" (Ryan & Xenos, 2011). Such language is regarded as the most common and reliable way for people to translate their internal thoughts and emotions into a form that others (i.e., online audiences) can understand (Tausczik & Pennebaker, 2010). Similar as in a face-to-face environment, the self-monitoring (SM) skill that is applied in various self-presentation tactics plays an important role in controlling oneself to the online social appropriateness and reflecting individual's expression behavior (Hall & Pennington, 2013).

Past studies noticed that there was significant correlation between SM skills and the linguistic cues on FB (e.g., Pennebaker & King, 1999; Mairesse, Walker, Mehl, & Moore, 2007; Markovikj, Gievska, Kosinski, & Stillwell, 2013). Hall and Pennington (2013) found that the high self-monitors would be more extravert and used cues on FB that might promote extraversion on FB (i.e., to receive more "likes" by their FB friends). For instance, the high self-monitors were more likely to use profile pictures at a younger age and use shorthand abbreviations, such as OMG (oh my god) and BTW (by the way). In comparison, the low self-monitors were more likely to be honest on FB. They promoted a conscientious self to their FB public. For example, a higher proportion of family talks was found in this group in their FB status updates. Rosenberg and Egbert (2011) also suggested that the low self-monitors were less sensitive to social cues than the high self-monitors, and therefore were less skilled at assessing appropriate behaviors and self-presentation in various situations. Unlike high self-monitors who regulate their own words and behaviors perceived favorably by others, low self-monitors often chose actions and words in accordance with their dispositions in a social network.

The aim of this study is twofold. The first is to evaluate the quality of responses to the Snyder's Self-Monitoring (SM) Questionnaire collected via the Internet. This will be done using an item response theory (IRT) model. Besides giving an indication of the scalability and reliability of the responses, the model also provides estimates of the personal level of SM skills of the sampled FB users. Secondly, the textual features of the posts for different SM-level groups will be extracted using structured and unstructured textual analysis using a concurrent model of the measured SM skills and posts on the FB Wall.

7.1.1 Self-Monitoring

The SM construct was introduced by Snyder (1974) as a trait that describes and explains individual differences in the self-control of expressive behavior for the sake of the demands and norms of an audience or context (von Davier & Rost, 1996). There are striking and important individual differences in the extent to which individuals can and do monitor their self-presentation, expressive behavior, and nonverbal affective display (Snyder, 1974). The SM-scale developed by Snyder (1974) was specifically designed to discriminate individual differences in concern for social appropriateness, sensitivity to the expression and self-presentation of others in social situations as cues to social appropriateness of self-expression. This instrument covers 25 self-report items like "I find it hard to imitate the behavior of other people" that are usually analyzed in a quantitative fashion, i.e., by summing the item responses after coding all items in the same direction. A median-split (sum score = 12) is generally applied to these sum scores in order to differentiate between two groups of people, the high self-monitors and the low self-monitors (von Davier & Rost, 1996).

The high self-monitors are characterized as persons who behave strategically to obtain desired outcomes by regulating public presentations. That is, when persons are made certain of their emotional reactions, they look to the behavior of others for cues to define their emotional states and model the emotional expressive behavior of others in the same situation who appear to be behaving appropriately (Schachter & Singer, 1962). For instance, such a person would be more likely to laugh at a comedy when watching it with amused peers than when watching it alone (Fuglestad & Snyder, 2009). In comparison, the low self-monitors present themselves in ways that reflect their authentic attitudes, values and beliefs. They express it as they feel it rather than monitored, controlled, and molded to fit the situation (Snyder, 1974).

7.1.2 Online Assessment and Internet Data

The past decade has witnessed a rapid expansion of the Internet. The Internet has not only significantly changed the way people conduct business, communicate, and live, but also influenced the practice of psychology as it related to testing and assessment (Naglieri et al., 2004). The Internet brings benefits of speed, costs, convenience and flexibility to the online assessment but introduce new problems such as testing security and data validity as well. A major limitation of previous researches regarding FB is that the data primarily assessed through self-report and few attempts have been made to evaluate the criterion validity of these measures. In other words, few studies has been conducted yet to evaluate whether self-reported data on FB are, in fact, related to actual data collected offline. A recent study was conducted by Junco (2013) regarding the criterion validity of measures of FB frequency by comparing self-reported time spent on the site and number of logins against actual usage as measured by computer monitoring software. Although there was a strong positive correlation between self-reported and actual time spent on FB in that study, a significant discrepancy was also shown between the two: Students spent an average of 26 minutes per day on FB , significantly lower than the average of 145 minutes per day obtained through self-report. Researches in other areas of human behavior have also shown that self-report measures through Internet tests could raise the risks of inaccuracy when compared to actual behaviors. For example, online self-reported measures of physical activity underestimated health risk biomarkers by as much as 50% when compared to accelerometer measurements (Celis-Morales et al., 2012), and self-reported TV watching time was underestimated by an average of 4.3 hours per week when compared to data from a TV monitor (Otten, Littenberg, & Harvey-Berino, 2010).

Although some misrepresentation and outright deception unavoidably occurs on mediated platforms like FB, on the whole, users present themselves online in a manner that approximates their offline self (Gosling, Gaddis, & Vazire, 2007; Toma, Hancock, & Ellison, 2008). However, as there is growing interest in researching on the Internet data and the psychosocial effect of FB use, it is important to come up with measurement methods that are both accurate and useful. The requirement of validity is of utmost importance for such tests on the Internet (Buchanan, 2002). How to validate the Internet data would be an interesting and essential topic to be addressed before stepping into data processing. In the present study, a validation method based on IRT (Lord, 1980; Rasch, 1960) was introduced to examine the authenticity of the Internet data.

7.1.3 Structured Textual Analysis – LIWC

Computer-based textual analysis is generally divided into two categories: structured and unstructured one. Structured textual analysis usually involves tight structures from existing software, such as Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2001). LIWC is a textual analysis software program that looks for words and counts words in psychology-relevant categories across multiple text files, for instance, essays, emails, blogs, novels and so on. It has two central features – the processing component and the dictionaries. During the processing, the program goes through each file word by word. Each word in a given text file is compared with the dictionary file. A dictionary refers to the collection of words that define a particular category such as "family", "positive emotion" and "work". There are 80 word categories (i.e., variables) in LIWC. These variables are divided into five dimensions: (a) Linguistic Process, in which variables as word count, word count per sentence, first person pronouns, verbs are output; (b) Psychological Process, in which variables related to positive emotions, negative emotions, family words, anxiety words, and etc. are calculated; (c) Personal Concerns, in which variables related to hobbies, work, life and etc. are output, (d) Spoken Categories, which focus on elements used in spoken language, and (e) Punctuations (Tausczik & Pennebaker, 2010). Each word or word stem defines one or more word categories or subdictionaries. For example, the word "cried" is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. Hence, if it is found in the target text, each of these five subdictionary scale scores will be incremented.

7.1.4 Unstructured Textual Analysis – Text Mining

Unstructured textual analysis focuses on extraction of patterns from loose structures. The development of information technology demonstrated breakthroughs in handling unstructured textual data during the past decade. A promising technique is text mining, which exploits information retrieval, information extraction, and corpus-based computational linguistics. This technique also plays a fundamental role in extracting correlation patterns between personality and variety of user's data captured from multiple sources (Markovikj et al., 2013).

Supervised text classification is a commonly used approach for textual categorization with text mining techniques. It generally involves two phases, a training phase and a prediction phase (Jurafsky & Martin, 2009). During training, the most

discriminative keywords for determining the class label are extracted. The input for the machine learning algorithm consists of a set of pre-specified keywords that may potentially be present in a document and labels classifying each document. The objective of the training phase is to "learn" the relationship between the keywords and the 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 that were not used during training. In the testing procedure, the keywords extracted from the training are scanned in each new input. Thus, the words that were systematically recognized are fed into the "trained" classifier model, which predicts the most likely label for each new self-narrative. To ensure proper generalization capabilities for the text classification models, a cross-validation procedure is generally applied.

7.1.5 The Present Study

The objective of the present study is to investigate whether the textual features of posts on FB Wall can predict the users' SM skills. After a validity checking on the Internet data, both structured textual analysis by LIWC and unstructured textual analysis by text mining techniques were conducted in this study. The predictions from textual analyses were further compared with the binary classification i.e., high or low self-monitors defined by the SM-scale (Snyder, 1974). As stated earlier, the purpose of this study is twofold: first, to evaluate the quality of responses to the Snyder's SM Questionnaire (1974) collected via the Internet; and secondly, to explore the textual features of the posts in different SM-level groups.

7.2 Method

7.2.1 Dataset

A sample of 39218 instances from Facebook (activity and demographic data) with approximately 1.8 million status updates used in the present study was provided by the MyPersonality project (http://mypersonality.org/wiki; Celli, Pianesi, Stillwell, & Kosinski, 2013). All the instances participated the Snyder SM test on FB at least once within the time period from January, 2009 to April, 2011. The respondents were highly motivated to answer honestly and carefully, as the only gratification they received for their participation was the feedback on their results (Kosinski, Stillwell, & Graepel,

2013). Among the respondents, 37360 people took the test once, and 1858 people took the test at least twice. (We took the most recent responses into analysis for the duplicate cases.) Applying the median-split rule (threshold = 12) on the SM scale, 54.8% of the total are low self-monitors, while 45.2% are high self-monitors. The SM score followed a normal distribution with mean equaling to 11.97 and standard deviation equaling to 4.13.

Of the 39218 instances, 2972 respondents had at least one textual post in status updates during the collection time period, that is, they gave both textual expressions and responses to the SM-scale. To simplify the study, we only focused on English-speaking people, which resulted in 2655 participants. Further, to concentrate on the investigation of predictability of posts on FB, we followed the approach of Argamon, Dhawle, Koppel, & Pennebaker (2005) by only including respondents with extremely high or low scores on SM-scale and excluding the middle scorers. The extreme groups were defined as SM scores were above the 75% (SM>15) or below the 25% (SM<9) (Snyder, 1974). The typical group of extreme SM skills might be professional stage actors as high self-monitors and psychiatric patients as low self-monitors, respectively. This approach was testified in the study of Mairesse et al. (2007) that a 2-3% increase in overall accuracy scores yielded compared to datasets that included the middle scorers. The two groups with extreme SM scores, consisting of 1128 respondents, were finally used in the current study for textual analysis. The sum of their posts on FB Wall was approximately 140 thousand. Of the 1128 respondents, 552 (48.9%) were extremely low self-monitors, while 576 (51.1%) were extremely high self-monitors. The majority of respondents were female (55.7%). The age of respondents ranged from 18 to 60 years, with a mean of 25.7 and standard deviation as 9.1.

7.2.2 Validation of the Internet Data

The purpose of data validation is to determine that the data are valid, sensible, reasonable and secure before they are processed. In the present study, the data of the SM-scale collected from FB were validated by using an IRT model. In psychological and educational measurement, instruments are developed that are used in a population of persons and item fit is used to evaluate to what extent an IRT model fits an instrument in a particular population (Glas & Dagohoy, 2007). Analyses were carried out using the public domain software MIRT (Glas, 2010). The SM-scale (Snyder, 1974) was developed based on a population of Stanford University undergraduates (n = 533) and it is necessary to use item fit to evaluate whether the IRT model fits the SM-scale

for the FB users. If the model holds fit, it implies that the Internet data are as valid as the original data that were used for the instrument development; otherwise, the Internet data can be determined as invalid.

Responses of the whole sample 39218 FB users to the 25 items of the SM-scale were used as input for the statistical analysis. A unidimensional two parameter logistic (2PL) model was used to estimate the individual's latent trait of SM. In this model, that is, the probability of a score in category "yes" ($X_{ni}=1$) of item i is given by the item response function

$$P(X_{ni}=1\,|\,\theta_n)=\frac{\exp[\alpha_i(\theta_n-\beta_i)]}{1+\exp[\alpha_i(\theta_n-\beta_i)]}, \tag{7.1}$$

where θ_n is the latent SM level of person n, β_i is an item location parameter representing the difficulty level of each SM item, and α_i is an item discrimination parameter indicating the extent to which the item response is related to the latent scale. The item parameters in the IRT model were estimated by marginal maximum likelihood (MML; Bock & Aitkin, 1981).

We investigated item fit using Lagrange Multiplier (LM; Glas, 1999) tests. Given the size of the data set, the focus will not be on the significance probabilities of the test, but on the observed and expected response frequencies and the effect sizes on which the test is based. To compute the LM statistic, the sample of respondents is divided into subgroups labeled $g=1,2,...,G$. We defined the subgroups as three total-score level groups (i.e., Level 1: total scores 0-9, Level 2: total scores 10-14, and Level 3: total scores 15-25) which were formed in such a way that the numbers of respondents in each group were approximately the same. The statistic is based on the difference between average observed scores on every item i in the subgroups, namely, $S_{ig}=\frac{1}{N_g}\sum_{n|g}^{N_g}X_{ni}$ (where the summation is over the N_g respondents in subgroup g), and their posterior expectations $E(S_{ig})$. The differences are squared and divided by their covariance matrix (for more details refer to Glas, 1998, 1999; Glas & Falcon, 2003). The LM statistic has an asymptotic chi-square distribution with $G-1$ degree of freedom. The statistics are accompanied by effect size $d_{ig}=\max_g\left|S_{ig}-E(S_{ig})\right|$ that show the degree of model violation. Since the effect size d_{ig} is on a scale ranging from

0 to the maximum score m_i, effect size $d_{ig} < 0.10$, although somewhat arbitrary, is commonly suggested as an indicator of minor and acceptable model violation (He et al., 2013; van Groen et al., 2010). In the current study, an item was identified as misfit when the effect size was above the cutoff point $d_{ig} = 0.10$.

Besides the item fit analysis, person fit is also necessary to take into consideration when handling the Internet data, because specific persons may still produce patterns that are highly unlikely given the model, although the IRT model may generally fit the data. For instance, some persons may give random responses because they are unserious to take the test. Using person fit statistics, the fit of a score pattern can be determined under the null-hypothesis that the IRT model holds. To test the person fit, we used the LM test for the constancy of θ over response patterns for the 2PL model introduced by Glas & Dagohoy (2007). The LM person fit test is based on a split two subtests: say the first part of the test (1-11 item) and the second part (12-25 item) of the test. In addition, to show that the quality of the Internet data is appropriate, we also compared the distribution of SM scores of the Stanford undergraduates sample and the FB sample by applying non-parametric Wilcoxon Rank Sum Test (also called Mann-Whitney Test).

7.2.3 Textual Analysis

The textual posts on users' FB Wall were analyzed via both structured and unstructured approaches by using LIWC and text mining techniques, respectively. In the present study, we focused on the two extreme SM groups, the low self-monitors (LSM, SM < 9) and the high self-monitors (HSM, SM > 15). A sample of 1128 respondents with approximately 140 thousand posts were included.

7.2.3.1 Structured Textual Analysis Using LIWC

All the 80 variables in LIWC, including 26 variables in the dimension of Linguistic Process, 32 variables in the dimension of Psychological Process, 7 variables in the dimension of Personal Concerns, 3 variables in the dimension of Spoken Categories and 12 variables in the dimension of Punctuations, were input for the structured textual analysis. Two approaches, logistic regression and classification trees were used to classify the individuals into two categories – HSM and LSM – based on the features of their textual input.

In the logistic regression, the dependent variable was defined in a binary category, i.e., 0 (i.e., LSM) and 1 (i.e., HSM). We input all the 80 variables as predictors. The logistic regression model is defined as

$$\text{logit}(p) = \ln\left(\frac{p}{1\text{-}p}\right) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = b_0 + b_1 x \ , \qquad (7.2)$$

where $p = \pi(x)$ is the probability that the dependent variable equals 1, b_0 and b_1 are regression coefficients.

In the approach of classification trees, we used the classification and regression tree (CRT) growing method with maximum tree depth equaling to 100. CRT splits the data into segments that are as homogeneous as possible with respect to the dependent variable. A terminal node in which all cases have the same value for the dependent variable is a homogeneous and "pure" node (Kotsiantis, 2007). The minimum number of cases in each parent node and child node were set as 100 and 50, respectively. Pruning tree was used to avoid over-fitting (for more on pruning trees refer to Elomaa, 1999; Bruha, 2000). The minimum change in improvement of each depth was set at 0.0001. To ensure the proper generalization capabilities for the classification tree model, a 10-fold cross validation procedure was applied.

7.2.3.2 Unstructured Textual Analysis Using Text Mining Techniques

In the unstructured textual analysis, a preprocessing was conducted first to ensure the textual data following a standardized format. Unlike the normal preprocessing in text classification, we included the stop words (e.g., "I", "is", "the", and etc.) in the present study, because some literature mentioned that the inclusion of stop words could increase the classification accuracy in textual analysis of online blogs (e.g., Iacoblli, Gill, Nowson, & Oberlander, 2011). Further on, all the words were stemmed using Porter's stemming algorithm (Porter, 1980). We noticed that the Internet language was more casual, thus resulted in more spelling mistakes than the normal writings. For instance, the "wrong" spelling "soooooo big" was often used in FB posts to emphasize the degree of bigness. In the current study, we handled these typical Internet words by two steps: first, transforming them into the original status (e.g., "sooooo" was transformed into "so") and secondly stemming them by the Porter stemming algorithm. This approach avoided the mighty confusion in keywords extraction. For example,

"soooo" (i.e.,"so" with four o's) and "sooooooo" (i.e., "so" with seven o's) would be extracted as a unique stem "so" instead of two different ones. However, the writer's latent intention to emphasize the degree of bigness was lost in preprocessing.

We deployed a supervised text classification in the present study, that is, to divide the textual analysis into two phases, training and testing. 70% of the dataset were randomly selected into training data, while the remaining 30% of the dataset were used to test the trained model. During training, the most discriminative keywords to determine the SM-level were extracted by using chi-square feature selection model (Oakes et al., 2001). A recently developed machine learning algorithm, product score model (PSM; He et al., 2012) was employed in conjunction with three representative models – unigrams, bigrams, and a combination of uni- and bigrams – to learn the patterns between the extracted keywords and the labels.

The PSM is an alternative machine learning algorithm, which features in assigning two weights for each keyword (in binary classification) – the probability of the word w occurs in the two separate corpora, U_w and V_w – to indicate the degree that a word represents the two classes. The weights are calculated by

$$\begin{cases} U_w = (n_w + a) / len(C_1) \\ V_w = (m_w + a) / len(C_2) \end{cases}, \tag{7.3}$$

where n_w and m_w are the word occurrences in HSM Corpus (C_1) and LSM Corpus (C_2), respectively. Note that a smoothing constant a (we use $a = 0.5$ in this study) is added to the word occurrence in Formula (7.3) to account for words that do not occur in the training set, but might occur in new texts. (For more on smoothing rules, see Manning & Schütze, 1999; Jurafsky & Martin, 2009.) The name product score comes from a product operation to compute scores for each class, i.e., S_1 and S_2, for each input text based on the term weights. That is,

$$\begin{cases} S_1 = P(C_1) \cdot \prod_{w=1}^{k} U_w = P(C_1) \cdot \prod_{w=1}^{k} \left[(n_w + a) / len(C_1) \right] \\ S_2 = P(C_2) \cdot \prod_{w=1}^{k} V_w = P(C_2) \cdot \prod_{w=1}^{k} \left[(m_w + a) / len(C_2) \right] \end{cases}, \tag{7.4}$$

where $P(C)$ is the prior probability for each category given the total corpora. The classification rule is defined as:

$$\text{choose}\begin{cases} C=1 & \text{if } \log(S_1/S_2) > b \\ C=2 & \text{else} \end{cases}, \tag{7.5}$$

where b is a constant and was defined equal to 0 in the current study (for more on PSM, see He et al., 2012; He & Veldkamp, 2012).

To avoid mismatches caused by randomness, unclassification rules are also taken into account. Based on the chi-square selection algorithm, the keywords are labeled as positive indicators or negative indicators. We define a text as "unclassified" when either one of the following conditions is met: (a) no keywords are found in the text; (b) only one keyword is found in the text; (c) only two keywords are found in the text, and one is labeled as a positive indicator while the other as a negative indicator.

To generalize the results from unstructured textual analysis, a 5-fold cross validation was also applied.

7.2.4 Analytic Strategy

In the present study, we defined the label made by the SM-scale as "standard" and the label predicted via textual analysis as "test", respectively. The performances of the structured and unstructured textual analysis were compared on five metrics: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Accuracy, the main metric used in classification, is the percentage of correctly defined individuals. Sensitivity and specificity are the proportion of actual positives (HSM) and actual negatives (LSM) that are correctly identified, respectively. The predictive values, PPV and NPV, are estimators of the confidence in predicting correct classification; i.e., the higher predictive values are, the more reliable the prediction is.

7.3 Results

7.3.1 Validation of the Internet Data

Table 7.1 shows the item parameters that were calibrated by the MML on a sample of 39218 instances from FB. As shown in Table 7.1, the discrimination parameters varied in the interval [0.256, 1.543], with a mean value around 0.69 (S.D. = 0.37). The difficulty parameters were included in the range [−0.808, 1.243], with a mean of 0.09 (S.D. = 0.64).

The validity of Internet data was investigated between the observations and expectations predicted by the IRT 2PL model using the LM statistics. The observed total score is the sum score of the responses on all items except the item targets. Table 7.2 reports the outcomes of analysis of model fit. The columns Obs and Exp give the observed and expected scores under the model, respectively. The last column (Dif) gives the effect sizes d_{ig}. Note that the highest effect size was 0.05, which is well below the criterion of 0.10. Further, in the person fit analysis, the detection of inconstancy of θ-estimates identified 365 individuals. Thus, the detection rate was 0.009, which is far below the significance probability of 0.05. In addition, the reliability of the SM score predictions of FB users was 0.732, which is well acceptable. Finally, in the comparison of distribution between the two samples of Stanford undergraduates and the FB users, the Wilcoxon Rank Sum Test resulted in a *p*-value of 0.25. Therefore, the null hypothesis that there is no systematic difference between the two independent populations was not rejected. As shown in Figure 7.1, the SM scores of the FB users (solid line) follow a normal distribution, with a mean of 11.97 and standard deviation of 4.13. The SM scores of the Stanford undergraduates (dot line) also follow a normal distribution, though a bit condensed than the FB curve, with a mean of 12.41 and standard deviation of 3.48.

The overall conclusion is that the IRT model fitted the Internet data very well, and the hypothesis that a latent scale pertained to the FB users was not rejected. Thus, the Internet data used in this study was valid enough to be further processed.

7.3.2 Textual Analysis

In the structured textual analysis approach, the logistic regression using LIWC showed a model fit in the Hosmer and Lemeshow test of goodness-of-fit ($\chi^2 = 4.507, df = 8, p = 0.809$). This test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. Table 7.3 lists the logistic regression coefficients of the top twenty LIWC predictors based on the significance of the score statistic. It was found that only the first three predictors were significant in parameter estimation ($p < 0.05$), including two variables, i.e., “family” ($b = -0.496$) and “discrepancy” ($b = -0.311$), in the dimension of Psychological Process and one variable, i.e., “leisure” ($b = 0.198$), in the dimension of Personal Concerns.

Table 7.1. Item Parameters of 25-item Self-Monitoring Scale (Snyder, 1974)

Item	Question in NCS-R	Item parameters	
		α (SE)	β (SE)
1	I find it hard to imitate the behavior of other people. (F)	1.159 (0.020)	-0.532 (0.014)
2	My behavior is usually an expression of my true inner feelings, attitudes, and beliefs. (F)	0.271 (0.016)	1.238 (0.013)
3	At parties and social gatherings, I do not attempt to do or say things that others will like. (F)	0.651 (0.012)	-0.542 (0.012)
4	I can only argue for ideas which I already believe. (F)	0.544 (0.014)	0.352 (0.011)
5	I can make impromptu speeches even on topics about which I have almost no information. (T)	0.959 (0.017)	0.298 (0.013)
6	I guess I put on a show to impress or entertain people. (T)	1.137 (0.020)	0.587 (0.014)
7	When I am uncertain how to act in a social situation, I look to the behavior of others for cues. (T)	0.256 (0.014)	-0.578 (0.011)
8	I would probably make a good actor. (T)	1.543 (0.024)	-0.166 (0.015)
9	I rarely seek the advice of my friends to choose movies, books, or music. (F)	0.649 (0.015)	1.243 (0.013)
10	I sometimes appear to others to be experiencing deeper emotions than I actually am. (T)	0.257 (0.013)	0.288 (0.011)
11	I laugh more when I watch a comedy with others than when alone. (T)	0.300 (0.013)	0.151 (0.011)
12	In groups of people, I am rarely the center of attention. (F)	1.060 (0.019)	0.365 (0.013)
13	In different situations and with different people, I often act like very different persons. (T)	0.476 (0.015)	-0.034 (0.012)
14	I am not particularly good at making other people like me. (F)	0.812 (0.017)	-0.753 (0.013)
15	Even if I am not enjoying myself, I often pretend to be having a good time. (T)	0.318 (0.013)	-0.034 (0.011)
16	I'm not always the person I appear to be. (T)	0.405 (0.015)	-0.728 (0.012)
17	I would not change my opinions (or the way I do things) in order to please someone else or win their favor. (F)	0.384 (0.015)	1.057 (0.013)
18	I have considered being an entertainer. (T)	1.254 (0.020)	0.251 (0.014)
19	In order to get along and be liked, I tend to be what people expect me to be rather than anything else. (T)	0.343 (0.016)	1.164 (0.013)
20	I have never been good at games like charades or improvisational acting. (F)	1.203 (0.020)	-0.571 (0.014)
21	I have trouble changing my behavior to suit different people and different situations. (F)	0.908 (0.018)	-0.808 (0.013)
22	At a party, I let others keep the jokes and stories going. (F)	0.730 (0.016)	0.678 (0.013)

23	I feel a bit awkward in company and do not show up quite as well as I should. (F)	0.631 (0.015)	-0.050 (0.012)
24	I can look anyone in the eye and tell a lie with a straight face (if for a right end). (T)	0.663 (0.015)	-0.398 (0.012)
25	I may deceive people by being friendly when I really dislike them. (T)	0.440 (0.014)	-0.189 (0.011)

Note. The item parameters were estimated from 2PL model. α indicates the item discrimination parameter, β indicates the item difficulty parameter.

Table 7.2. Model Fit in Score Level Groups for Facebook Users (n=39218)

Item	Level 1		Level 2		Level 3		Dif.
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	
1	0.40	0.40	0.62	0.62	0.81	0.80	0.00
2	0.15	0.19	0.22	0.22	0.30	0.27	0.02
3	0.48	0.49	0.64	0.63	0.76	0.75	0.01
4	0.31	0.31	0.41	0.41	0.53	0.53	0.00
5	0.28	0.26	0.43	0.43	0.60	0.62	0.01
6	0.17	0.19	0.35	0.37	0.62	0.58	0.03
7	0.55	0.59	0.64	0.64	0.73	0.69	0.02
8	0.31	0.28	0.54	0.54	0.75	0.77	0.02
9	0.21	0.15	0.25	0.23	0.26	0.34	0.05
10	0.38	0.37	0.43	0.43	0.48	0.48	0.00
11	0.37	0.40	0.47	0.46	0.55	0.53	0.02
12	0.27	0.24	0.42	0.42	0.58	0.62	0.02
13	0.36	0.41	0.50	0.51	0.67	0.61	0.04
14	0.53	0.51	0.67	0.68	0.79	0.80	0.01
15	0.40	0.44	0.51	0.51	0.62	0.58	0.03
16	0.55	0.59	0.68	0.68	0.79	0.75	0.03
17	0.17	0.20	0.25	0.26	0.36	0.33	0.02
18	0.27	0.24	0.45	0.45	0.63	0.67	0.02
19	0.14	0.19	0.24	0.24	0.34	0.30	0.03
20	0.44	0.41	0.62	0.63	0.78	0.81	0.03
21	0.48	0.51	0.70	0.69	0.84	0.82	0.02
22	0.26	0.22	0.34	0.34	0.45	0.49	0.03
23	0.43	0.38	0.51	0.51	0.59	0.64	0.03
24	0.45	0.46	0.59	0.60	0.74	0.72	0.01
25	0.41	0.45	0.54	0.55	0.69	0.64	0.03

Note. The columns labeled Obs and Exp give the observed and expected scores under the model, respectively. The observed total score is the sum score of the responses on all items. Dif gives the absolute value of effect size averaged across the three score levels. Level 1: total scores 0-9, Level 2: total scores 10-14, Level 3: total scores 15-25. Degree of freedom equals to 2.

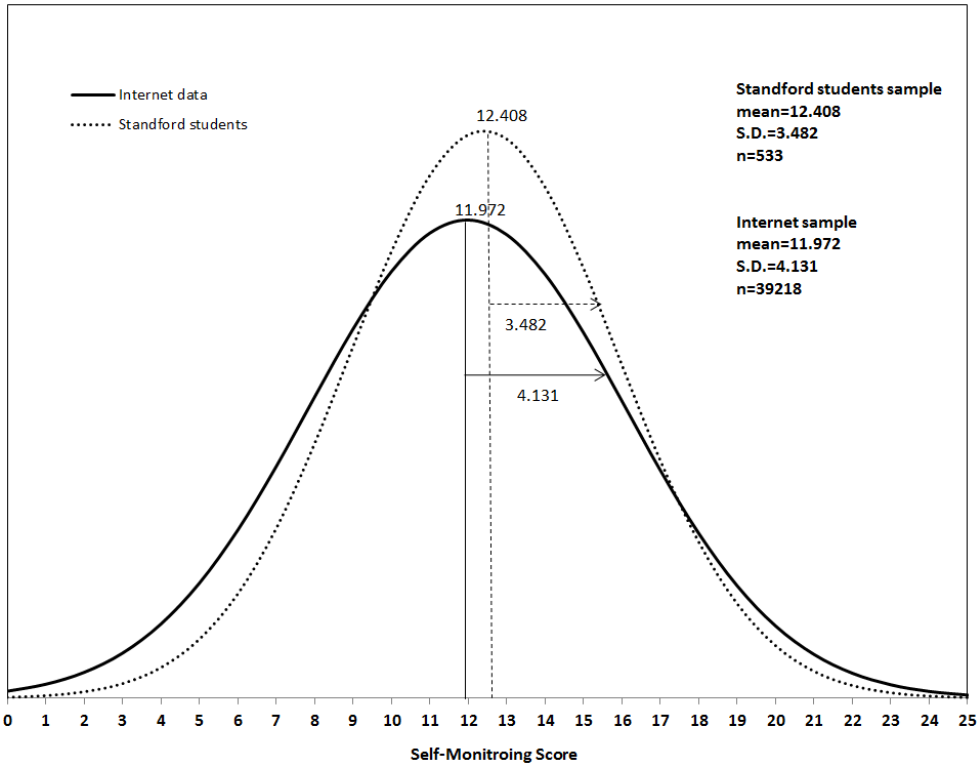


Fig. 7.1. SM score distribution of two samples: Stanford undergraduates (n=533) and FB users on the Internet (n=39218).

Figure 7.2 presents the classification tree model based on the 80 LIWC variables. The tree resulted in four depths, nine nodes and five terminal nodes. The four robust classifiers were extracted in a decreasing order: “anger”, “family”, “preps” and “word per sentence”. Note that the improvement in each depth was very marginal and the highest improvement was produced by “anger” as 0.012, implying that the predictors were not very powerful to make the decision. This might also be the reason of the shortness of the tree. Based on a 10-fold cross validation, the risk of misclassification was estimated as 0.38 with standard error of 0.014.

Further, we also investigated the correlations between each variable in the LIWC and the SM scores. Using a 95% confidence interval, thirteen variables were found significantly correlated with the SM scores, including eight variables had positive correlations and five variables had negative correlations, though the values were not

high (see Table 7.4). It was interesting to find that the words related to assent had the highest positive correlation with the SM scores, which implied that the higher SM skill a person has, the more often he/she may use assent words like “ok”, “yes”, “agree” in the posts. We also noticed that the lower SM skill a person had, the more likely he/she might use words related to the third episode and family terms (e.g., home, sister, brother) to update the status on FB. This result kept consistent with the findings in the study of Hall and Pennington (2013).

Table 7.3. Logistic Regression Coefficients of LIWC Predictors (Top 20 predictors based on the significance of the score statistic)

Rank	Predictor	b_1	p-value	95% Confidence Interval	
				low	high
1	family	-0.496	0.031	-1.105	-0.150
2	leisure	0.198	0.046	0.020	0.460
3	discrepancy	-0.311	0.049	-0.719	-0.003
4	quantifiers	0.244	0.063	-0.009	0.565
5	adverb	0.161	0.089	-0.032	0.397
6	humans	-0.299	0.094	-0.770	0.010
7	sad	-0.428	0.112	-1.033	0.190
8	verb	-0.080	0.117	-0.216	0.037
9	space	0.312	0.129	-0.103	0.868
10	exclamation mark*	0.128	0.138	-0.039	0.350
11	auxiliary verbs	0.154	0.161	-0.097	0.398
12	function words	-0.135	0.172	-0.390	0.074
13	religion	-0.156	0.172	-0.450	0.056
14	period*	0.112	0.189	-0.067	0.327
15	achievement	0.162	0.191	-0.154	0.435
16	question mark*	0.111	0.196	-0.071	0.332
17	parenthesis*	0.232	0.201	-0.173	0.673
18	quote mark*	0.111	0.206	-0.067	0.336
19	all punctuations*	-0.102	0.221	-0.314	0.078
20	assent	0.164	0.227	-0.102	0.486

Note. b_1 indicates the logistic regression coefficient for each predictor. * indicates predictors in the dimension of Punctuations in LIWC.

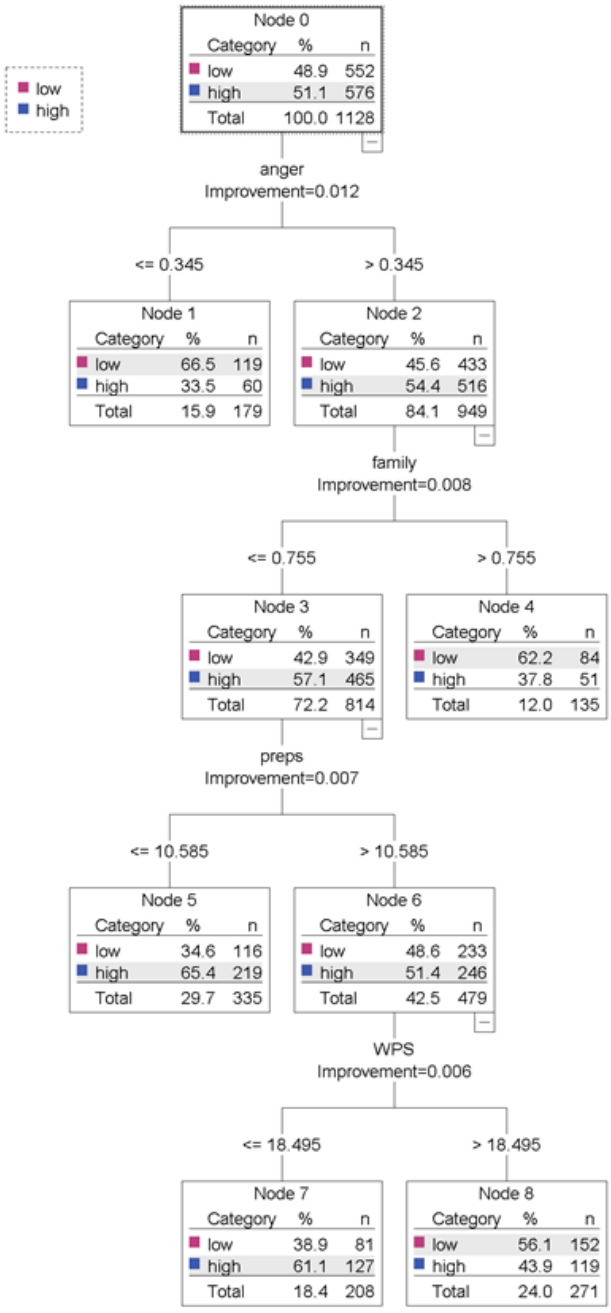


Fig. 7.2. Classification tree model on the 80 LIWC variables

Table 7.4. Correlation Between LIWC Predictors and SM Scores (significance $p<0.05$)

Predictor	Correlation with SM scores
assent	0.085
question mark*	0.065
exclusive	0.064
adverb	0.051
feel	0.046
cause	0.044
body	0.044
family	-0.079
religion	-0.068
they (3rd pers plural)	-0.060
inhibition	-0.051
dash mark*	-0.041

Note. * indicates predictors in the dimension of Punctuations in LIWC.

In the unstructured textual analysis, 1000 keywords, consisting of 500 keywords for HSM and 500 keywords for LSM were extracted as robust classifiers and used for text classification with text mining techniques. Table 7.5 presents the top 20 keywords (10 for HSM and 10 for LSM) extracted from the FB posts. The fourth column shows the chi-square score for each keyword. The last two columns give the number of occurrences of each keyword in the LSM and HSM corpora, respectively. It was noticed that among the top twenty keywords, eight were emoticons and four were Internet slangs. The robustness of emoticons and Internet slangs in prediction of SM skills aroused our special interests. An emoticon is a communicative pictorial representation of a facial expression to send the feelings of the user, for instance, ":)" indicates a happy face and ":(" indicates a sad face. The Internet slangs are expressions that coined and popularized by the Internet users to save time on keystrokes, for instance, "wow" is generally used to express astonishment or admiration, and "ugg" often indicates ugly. As shown in Table 7.5, the "happy faces",

e.g., ":)", "=)", ";)", and "^_^", were found the most significant classifiers for low self-monitors whereas the "sad or puzzling face", e.g., ":(" and "++", were the robust classifiers in the high self-monitors. In addition, we also found that the Internet slangs (e.g., "wow", "ugg", "lol", "omg") were used more often by the group of high self-monitors. This was similar as the findings in the study of Hall and Pennington (2013) where the high self-monitors were reported more likely to use the shorthand abbreviation.

Table 7.6 exhibits the performance metrics compared between structured and unstructured textual analysis. The logistic regression using variables from LIWC yielded the highest accuracy (0.629) among all the models. It also resulted in the highest sensitivity, specificity and PPV. In general, the structured textual analysis approach performed better than the unstructured one. However, giving concerns on classifying FB posts solely based on the keywords, the over 50% accuracy rate is acceptable. The PSM in conjunction with the unigrams performed the best in the unstructured textual analysis. Although the PSM with a combination of unigrams and bigrams resulted in the highest NPV (0.678) among all, it was compensated by the lowest PPV (0.328). The bigram was not as powerful as it was shown in the study of Iacobelli and his group (2011) where the bigrams were found as the robust classifiers to represent the bloggers' personality. The probable reason might be that the blogs and FB posts are substantially different in textual format, though both of them are collected online. For instance, the blogs are generally long texts and describe a relatively complete story while the posts are comparatively short and express thinking in mind in a relatively casual way.

7.4 Discussion and Conclusion

The present study evaluated the quality of responses to the Snyder's SM Questionnaire (1974) collected via the Internet, and explored the textual features of the posts in different SM-level groups and extracted patterns between FB users' SM skills and their posts on the FB Wall. By using both structured and unstructured textual analysis, we demonstrated that the textual posts on the FB Wall could partially predict the users' SM skills. The variable of "family" was found the most significant predictor in LIWC in both of the logistic regression and classification tree approach. The accuracy of classification of the SM groups was above 60% when using LIWC and above 50% with the text mining algorithm PSM in conjunction with unigrams. The emoticons and

Internet slangs were extracted as the most robust classifiers in the unstructured textual analysis.

Table 7.5. Top 20 Keywords Extracted by Chi-square Feature Selection Model in Unstructured Textual Analysis

Label	Rank	Keyword	Chi-square score	Number of occurrences	
				LSM	HSM
LSM	1	=)	93.60	102	6
	2	god	83.28	137	27
	3	grace	76.04	88	7
	4	:-)	74.94	134	30
	5	;)	66.48	79	7
	6	work	65.74	293	142
	7	“--	63.47	69	4
	8	bless	53.88	212	96
	9	(^-^)	49.96	46	0
	10	repost	48.67	172	73
HSM	1	“:	232.27	13	291
	2	wow	123.97	158	460
	3	!!!	88.48	1783	2583
	4	:(	86.76	41	193
	5	wit	79.73	50	202
	6	ugg	63.44	1	72
	7	fuck	55.86	426	727
	8	++	52.40	1	60
	9	lol	47.70	40	140
	10	omg	46.88	160	332

Note. LSM and HSM indicate low self-monitors and high self-monitors, respectively.

Table 7.6. Performance Metrics Compared Between Structured and Unstructured Textual Analysis

	Method	Accuracy	Sensitivity	Specificity	PPV	NPV
LIWC						
	Logistic regression	0.629	0.629	0.628	0.663	0.592
	Classification Tree	0.621	0.637	0.607	0.601	0.643
PSM						
	Unigrams	0.537	0.541	0.542	0.642	0.437
	Bigrams	0.521	0.601	0.487	0.558	0.530
	Uni+Bigrams	0.499	0.510	0.496	0.328	0.678

Note. The categories determined by the SM-scale are used as true standard in the classification. The structured and unstructured textual analysis was conducted by using the software LIWC and machine learning algorithm PSM, respectively. PPV and NPV represent the positive predictive value and negative predictive value, respectively.

It was found that the text classification of FB posts via an unstructured approach did not perform as well as the past researches that focused on story-based documents. He et al. (2012) used the PSM to analyze the patients' self-narratives to detect the posttraumatic stress disorder (PTSD) patients. The accuracy of text classification in that study was fairly high as 0.82. In another study of He & Veldkamp (2012), the PSM was applied to analyze undergraduates' life stories in order to understand their personality adaption. The computerized text classification resulted in over 70% accuracy compared with the human-raters' results. The reasons of a relatively low classification accuracy in the current study might be addressed from three aspects. First, the contents of posts on FB Wall were much more diversified than the story-based documents. The FB users can express anything in their mind on the Wall whereas the respondents in the PTSD psychiatric screening are asked to focus on description of the traumatic events and related symptoms. Secondly, the Internet textual posts may have more loose linguistic structures. For instance, on the FB, it is more likely to see a sentence without a subjective or use consecutive punctuations to express the emotions, e.g., "?!!!!!". Thirdly, the posts on FB are generally written in a more casual way, e.g., using slangs, shorthand abbreviations and emoticons and might have more spelling mistakes and coined words, e.g., "soooooo big". These wording variations bring new challenges in the unstructured textual analysis.

To link the extracted keywords with the psychology-oriented predictors in the LIWC, we mapped the 1000 keywords into the 80 categories in LIWC. The top five LIWC predictors that were most frequently matched by the keywords were “affect” (affective processes), “posemo” (positive emotion), “relativ” (relativity), “social” (social process) and “cogmech” (cognitive process). All these five predictors located in the second dimension of LIWC, i.e., Psychological Processes, which implying that the keywords extracted from the text mining approach were mainly the words with psychological attributes. Moreover, we also noted that the keywords extracted from the group of HSM appeared more often in the predictors “negemo” (negative emotion) and “percep” (perceptual processes). The group of LSM had obviously more keywords mapped in the predictors of “funct” (function words), “verb”, “present” (present tense), “home” and “relig” (religion). These findings demonstrated that the predictors in LIWC and keywords extracted from the text mining could be mutual supplements. To obtain the benefits from both methods, it might be interesting to put the 80 LIWC variables and the extracted keywords altogether in a pool of predictors for an entire pattern exploration in the future study. Further, as mentioned above, the emoticons and Internet slangs were extracted as the robust classifiers to distinguish the low and high SM groups. However, in the current version of LIWC, it is hard to map them into corresponding categories. With the increasing research interests in textual posts on social communication networks, like FB and Twitter, we would recommend to extend the dictionaries in the LIWC to a larger scope which could include the attributes of special Internet-related languages, such as emoticons and Internet slangs.

In conclusion, the current study demonstrated that a relation between the textual posts on FB Wall and SM skills exist. Moreover, the textual posts can partially predict the users’ SM skills. Both, the LIWC and text mining techniques, were promising in handling the Internet textual posts. This study also introduced a method to investigate the validity of Internet data. We recommend that the researchers always perform a data validation study using the methodology presented here to ensure the data is valid before being processed.

Chapter 8

Epilogue

Text mining is one of the new and exciting research areas in statistical natural language processing (NLP). Although this technique has been generally used in information retrieval (IR) systems, such as “Google” and “Amazon”, its application is rather new in the area of psychiatric and psychological assessments. Textual data collected in psychiatric assessments may have several differences from the data used in the IR systems. For instance, patients’ self-narratives are often less structured than, for example, the biomedical texts (i.e., academic texts in biomedical literature) and clinical texts (i.e., texts written by clinicians in the clinical settings, describing patients, pathologies and findings made during interviews and etc.). The corpus size resulting from self-narratives is usually much smaller than the ones used for IR such as Brown News Corpus that possesses over 1 million words. These features challenged existing methods in text mining, and some alternatives had to be sought for.

8.1 Answers to Research Questions

The present thesis answers two main research questions that are proposed in Chapter 1. That is, (1) how can we apply text mining techniques in psychiatric and psychological assessment to make classification decisions, and (2) how can we validate the text mining procedure and combine it with itemized measures in the framework of psychiatric and psychological assessment.

Chapter 2 introduces the chi-square features selection model and presents an alternative machine learning algorithm for the binary text classification, named the

product score model (PSM). The PSM features in assigning two weights for each keyword to indicate to how much of a degree the word can represent the two classes and possessing more flexibility in the model decision threshold. In this thesis, the PSM is proposed for a binary categorization. A generalized model to multiple categories is expected in future research.

Text mining and the PSM algorithm are applied in Chapter 3 to develop a textual screening tool for patients with posttraumatic stress disorder (PTSD) based on lexical features (bag-of-words) in their self-narratives. The text mining approach results 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 concluded as a promising addition to assessing PTSD in clinical and research settings.

In Chapter 4, the data representation model is extended from unigrams, where a document is represented as a vector of single word counts (Manning & Schütze, 1999), to *n*-grams, where the document is represented by counts of groups of *n* consecutive words (e.g., Bekkerman & Allan, 2003; Tan et al., 2002). Using the same sample as in Chapter 3, the PSM together with two commonly used machine learning algorithms, decision trees and naïve Bayes are applied in conjunction with five representation models, unigrams, bigrams, trigrams, a combination of uni- and bigrams and a mixture of *n*-grams to identify the PTSD patients. Although the PSM with unigrams attains the highest prediction accuracy compared with psychiatrists' diagnoses in structured interviews, it is interesting to find that the addition of *n*-grams contributed most to enhance the reliability of prediction and balance the performance metrics. This chapter further demonstrates that the computerized textual assessment system is a promising tool to analyze patients' expression behaviors, thus helps psychiatrists identify the potential patients from an early stage.

Despite the unique attributes of open questions, itemized instruments are predominantly used in the psychiatric and psychological assessments. In contrast to the sum score based classical test theory, item response theory (IRT) measures a latent trait at the item level. To detect differential item functioning (DIF) among various subpopulations is one of the important applications of IRT. In Chapter 5, the generalizability of DSM-IV diagnostic criteria for PTSD is explored for different groups related to background variables such as gender, marital status and educational level. This chapter also emphasizes the importance of evaluating the impact of DIF on population inferences and introduces two approaches to estimate the DIF impact. It is

concluded that the DSM-IV diagnostic criteria for PTSD does not produce substantially biased results in the investigated subpopulations, and there should be few reservations regarding their use.

Given concerns of the positive effects in either text mining or IRT as discussed in Chapter 2 through 5, a combination of these two methods is proposed in Chapter 6 to further strengthen the benefits from both sides. Chapter 6 combines text mining techniques and an IRT model in a Bayesian framework, where the textual assessment functions as an informative prior for the estimation of the PTSD latent trait. Note that the data used in this chapter are different from those used in the previous chapters. The item parameters that are calibrated in Chapter 5 are fixed in this study, which make the two scales with and without textual priors comparable. Results show that adding textual assessment to the itemized measures significantly increase the diagnosis accuracy, decrease the standard error of estimation, and can be used to shorten the length of a follow-up test.

Chapter 7 extends the model application from psychiatric datasets to an Internet dataset, which consists of both textual posts and responses to the scales on Facebook (FB). This chapter emphasizes the importance of validating data collected from the Internet and explores the relationship between self-monitoring skills and textual posts on the FB Wall. Textual analysis is conducted via both structured and unstructured approaches. To link the results from these two approaches, the keywords extracted by the text mining techniques are mapped onto the framework of Linguistic Inquiry and Word Count (LIWC), a commonly used psychology-related linguistic software package. The variable of “family” is found the most significant predictor in LIWC. The emoticons and Internet slangs are extracted as the most robust classifiers in the unstructured textual analysis. The conclusion is drawn that textual posts on the FB Wall can partially predict the users’ self-monitoring skills. The accuracy rate is expected to enhance if variables from LIWC and keywords extracted from text mining can be combined to use in the future studies.

With respect to both research questions, three general conclusions can be drawn from this thesis. (1) Text mining techniques are promising in analyzing unstructured data in psychiatric and psychological assessments. The use of an alternative algorithm adapted for the specific requirements in psychiatric and psychological assessments, like for instance the PSM presented in this thesis, can improve the accuracy of classification decisions. (2) The results derived from text mining procedures can be validated by itemized measures using item response theory modeling. It is important to check the validity of the itemized measures as well before using them. (3) The

outcomes of text mining and itemized measures can be simultaneously modeled via a Bayesian approach presented in this thesis to enhance the quality of the measurement.

8.2 Further Application and Future Recommendation

The final remarks of this epilogue concern the further application of text mining techniques in psychiatric and psychological assessments and some recommendations for future research. The textual assessment method developed in this thesis will be applicable to researches with similar background and makeup. For instance, the text classification model developed for PTSD screening can be also utilized in an initial detection for individuals' with potential risk in depression. Depression has been reported as the second cause of world disability by 2020 (World Health Organization [WHO], 2001) and is expected to be the largest contributor to disease burden by 2030 (WHO, 2008). Moreover, an early detection, either by a general practitioner or by an online screening test, will result in a more effective and shorter treatment, compared to a later one. This leads to a reduction in the cost of treatment. 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' addressing can be automatically transferred into written forms, would bring extra benefits for both psychiatrists and patients. This application may especially facilitate the patients who are not able to express their feelings by writing in participation of screening and diagnosis.

Text mining together with IRT is expected to be a promising tool in psychological and psychiatric assessments in the future decades. The involvement of text mining provides a new perspective to handle structured and unstructured data in a common framework. With the unprecedented popularity of social communication networks like Facebook, numerous interesting research ideas on the subject of psychology, such as personality, self-presentation and online honesty are springing up. The techniques applied and discussed in this thesis might play an important role in exploring these exciting fields.

References

Alexa Internet Inc. (2011). Alexa top 500 global sites. (http://www.alexa.com/topsites) accessed February 3rd, 2011.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV (4th ed.)*. Washington, DC: American Psychiatric Association.

Ando, M., Morita, T., & O'Connor, S. J. (2007). Primary concerns of advanced cancer patients identified through the structured life review process: A qualitative study using a text mining technique. *Palliative and Support Care*, *5*(3), 265-271.

Andrews, L., Joseph, S., Shevlin, M., & Troop, N. (2006). Confirmatory factor analysis of posttraumatic stress symptoms in emergency personnel: An examination of seven alternative models. *Personality and Individual Differences*, *41*(2), 213-224.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical predictors of personality type. In: *Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America*, St. Louis, Missouri, USA.

Asmundson, G. J. G., Frombach, I., McQuaid, J., Pedrelli, P., Lenox, R., & Stein, M. B. (2000). Dimensionality of posttraumatic stress symptoms: A confirmatory factor analysis of DSM-IV symptom clusters and other symptom models. *Behaviour Research and Therapy*, *38*(2), 203-214.

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.

Bekkerman, R., & Allan, J. (2003). *Using bigrams in text categorization*. Center for Intelligent Information Retrieval, University of Massachusetts, Technical Report IR-408. (http://www.cs.umass.edu/~ronb?papers/bigrams.pdf) accessed December 27th, 2003.

Betemps, E. J., Smith, R. M., Baker, D. G., & Rounds-Kugler, B. A. (2003). Measurement precision of the clinician administered PTSD scale (CAPS): A Rasch model analysis. *Journal of Applied Measurement*, *4*(1), 59-69.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison–Wesley.

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.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika*, *46*(4), 443-459.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development*, *35*(3), 167-181.

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.

Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Brewin, C. R. (2005). Systematic review of screening instruments for adults at risk of PTSD. *Journal of Traumatic Stress*, *18*(1), 53-62.

Bruha, I. (2000). From machine learning to knowledge discovery: Survey of preprocessing and postprocessing. *Intelligent Data Analysis*, *4*, 363-374.

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice*, *33*(2), 148-154.

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Mahwah, NJ: Erlbaum.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

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.

Celis-Morales, C. A., Perez-Bravo, F., Ibañez, L., Salas, C., Bailey, M. E. S., & Gill, J. M. R. (2012). Objective vs. self-reported physical activity and sedentary time: Effects of measurement method on relationships with risk biomarkers. *PLoS ONE, 7*(5), e36345. http://dx.doi.org/10.1371/journal.pone.0036345.

Celli, F., Pianesi, F., Stillwell, D. S., & Kosinski, M. (2013). Workshop on Computational Personality Recognition. In: *Proceedings of the 7th international AAAI conference on weblogs and social media*, Boston, MA, USA.

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.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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.

Cohen, J. (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, *20*, 37-46.

Conrad, K. J., Wright, B. D., McKnight, P., McFall, M., Fontana, A., & Rosenheck, R. (2004). Comparing traditional and Rasch analyses of the Mississippi PTSD scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement*, *5*(1), 15-30.

Conway, M. (2010). Mining a corpus of biographical texts using keywords. *Literary and Linguistic Computing*, *25*(1), 23-35.

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.

Day, S., Christensen, L. M., Dalto, J., & Haug, P. (2007). Identification of trauma patients at a level 1 trauma center utilizing natural language processing. *Journal of Trauma Nursing*, *14*(2), 79-83.

de Jong, M. G., Steenkamp, J. B. E. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science*, *28*, 674-689.

Dekel, R., & Solomon, Z. (2006). Marital relations among former prisoners of war: Contribution of posttraumatic stress disorder, aggression, and sexual satisfaction. *Journal of Family Psychology*, *20*(4), 709-712.

Derogati.Lr, Lipman, R. S., Rickels, K., Uhlenhut.Eh, & Covi, L. (1974). Hopkins Symptom Checklist (Hscl) - self-report symptom inventory. *Behavioral Science*, *19*(1), 1-15.

Diamond, E. L., & Lilienfeld, A. M. (1962). Effects of errors in classification and diagnosis in various types of epidemiological studies. *American Journal of Public Health*, *52*, 1137-1144.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2-3), 103-130.

Drigas, A., Koukianakis, L., & Papagerasimou, Y. (2011). Towards an ICT-based psychology: E-psychology. *Computers in Human Behavior*, *27*, 1416-1423.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York, NY: Wiley.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61-74.

Elhai, J. D., de Francisco Carvalho, L., Miguel, F. K., Palmieri, P. A., Primi, R., & Frueh, B. C. (2011). Testing whether posttraumatic stress disorder and major depressive disorder are similar or unique constructs. *Journal of Anxiety Disorders*, *25*, 404-410.

Elomaa, T. (1999). The biases of decision tree pruning strategies. *Advances in Intelligent Data Analysis Proceedings*, *1642*, 63-74.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, England: Cambridge University Press.

First, M. B. (1997). *User's guide for the structured clinical interview for DSM-IV Axis II personality disorders: SCID-II*. Washington, DC: American Psychiatric Press.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences, 57B*(5), 275-284.

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.

Foa, E. B., Riggs, D. S., Dance, C. V., & Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Traumatic Stress, 6*(4), 459-473.

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 and Dissociation, 6*(3), 105-123.

Fuglestad, P. T., & Snyder, M. (2009). Self-monitoring. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of Individual Differences* (pp. 574-591). New York, NY: Guilford.

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. *Journal of Nervous and Mental Disease, 198*(9), 609-613.

Gavranidou, M., & Rosner, R. (2003). The weaker sex? Gender and post-traumatic stress disorder. *Depression and Anxiety, 17*(3), 130-139.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*(3), 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273-294.

Glas, C. A. W. (2010). MIRT: Multidimesional Item Response Theory (computer software) (http://www.utwente.nl/gw/omd/afdeling/Glas).

Glas, C. A. W., & Dagohoy, A. V. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*(2), 159-180.

Glas, C. A. W., & Falcon, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87-106.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications* (pp. 69-96). New York, NY: Springer.

Gonzalez, N. (2011). Facebook marketing statistics, demographics, reports and news. (http://www.checkfacebook.com) accessed February 3rd, 2011.

Gosling, S., Gaddis, S., & Vazire, S. (2007). Personality impressions based on Facebook profiles. In B. Berendt (Eds.), *Authorial analysis*. Symposium conducted at the international conference on weblogs and social media, Boulder, CO, USA.

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley, CA: University of California Press.

Graham, C., Franses, A., Kenwright, M., & Marks, I. (2000). Psychotherapy by computer: A postal survey of responders to a teletext article. *The Psychiatrist*, *24*(9), 331-332.

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.

Hall, J. A., & Pennington, N. (2013). Self-monitoring, honesty, and cue use on Facebook: The relationship with user extraversion and conscientiousness. *Computers in Human Behavior*, *29*, 1556-1564.

Hand, D. J., & Yu, K. M. (2001). Idiot's Bayes - Not so stupid after all? *International Statistical Review*, *69*(3), 385-398.

He, Q., Glas., C. A. W., & Veldkamp., B. P. (2013). Assessing Impact of Differential Symptom Functioning on Posttraumatic Stress Disorder (PTSD) Diagnosis. *International Journal of Methods in Psychiatric Research*. In press.

He, Q., & Veldkamp, B. P. (2012). Classifying unstructured textual data using the Product Score Model: An alternative text mining algorithm. In T. J. H. M. Eggen & B. P.

Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 47-62). Enschede, Netherlands: RCEC.

He, Q., Veldkamp, B. P., & de Vries T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*, *198*(3), 441-447.

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.

Hunt, E. B., Marin, J., & Stone, P. J. (1966). *Experiments in induction*. New York, NY: Academic Press.

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In S. D'Mello, A. Graesser, B. Schuller, & J-C. Martin (Eds.), *Proceedings of affective computing and intelligent interaction (ACII 2011, Part II)* (pp. 568-577), Lecture Notes in Computer Science, 6975. New York, NY: Springer.

Junco, R. (2013). Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior*, *29*, 626-631.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Pearson Prentice Hall.

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.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B. E., Walters, E. E., Zaslavsky, A., & Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R) design and field procedures. *International Journal of Methods in Psychiatric Research*, *13*(2), 69-92.

Kirschstein, T., Liebscher, S., & Becker, C. (2013). Robust estimation of location and scatter by pruning the minimum spanning tree. *Journal of Multivariate Analysis*, *120*, 173-184.

Kosinski, M., Stillwell, D. J., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 110*(15), 5802-5805.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31,* 249-268.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lewis, D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In N. Belkin, P. Ingwersen, & A. M. Pejtersen (Eds.), *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 37-50). New York, NY: ACM.

Lewis, D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nedellec & C. Rouveirol (Eds.), *Machine learning: ECML-98, Proceedings from the 10th European conference on machine learning* (pp. 4-15). New York, NY: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, CA: Erlbaum.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing, 10,* 325-337.

Maheu, M. M., Pulier, M. L., McMenamin, J. P., & Posen, L. (2012). Future of telepsychology, telehealth, and various technologies in psychological research and practice. *Professional Psychology: Research and Practice, 43*(6), 613-621.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research, 30*(1), 457-500.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. S. (2013). Mining Facebook data for predictive personality modeling. In: *Proceedings of the 7th international AAAI conference on weblogs and social media (ICWSM 2013),* Boston, MA, USA.

Martin, E. K., Taft, C. T., & Resick, P. A. (2007). A review of marital rape. *Agression and Violent Behavior, 12*(3), 329-347.

Matteucci, M., & Veldkamp, B. P. (2013). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods and Application*, *22*(2), 243-267.

McAdams, D. P. (2008). Personal narratives and the life story. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 242-264). New York, NY: Guilford.

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.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128-165.

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. *Journal of Nervous and Mental Disease*, *180*(2), 111-116.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist*, *59*(3), 150-162.

Norris, F. H., & Perilla, J. L. (1996). The revised Civilian Mississippi Scale for PTSD: reliability, validity, and cross-language stability. *Journal of Traumatic Stress*, *9*(2), 285-298.

Norris, F. A., & Riad, J. K. (1997). Standardized self-report measures of civilian trauma and posttraumatic stress disorder. In J. P. Wilson & T.M. Keane (Eds.), *Assessing psychological trauma and PTSD* (pp. 7-42). New York, NY: Guilford.

Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, W. A. V., & Beaulieu, M. (2001). A method based on chi-square test for document classification. In: *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 440-441). New York, NY: ACM.

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.

Ostini, R., & Nering, M. L. (2005). *Polytomous Item Response Theory Models*. Quantitative Applications in the Social Sciences 144. SAGE.

Otten, J. J., Littenberg, B., & Harvey-Berino, J. R. (2010). Relationship between self report and an objective measure of television-viewing time in adults. *Obesity*, *18*(6), 1273-1275.

Owen, J. E., Hanson, E. R., Preddy, D. A., & Bantum, E. O. (2011). Linguistically-tailored video feedback increases total and positive emotional expression in a structured writing task. *Computers in Human Behavior*, *27*(2), 874-882.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Erlbaum.

Pakhomov, S., Chacon, D., Wicklund, M., & Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disesase: A case study of Iris Murdoch's writing. *Behavior Research Methods*, *43*(1), 136-144.

Palm, K. M., Strong, D. R., & MacPherson, L. (2009). Evaluating symptom expression as a function of a posttraumatic stress disorder severity. *Journal of Anxiety Disorders*, *23*(1), 27-37.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count.* Mahwah, NJ: Erlbaum.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*, 1296-1312.

Peters, L., Issakidis, C., Slade, T., & Andrews, G. (2006). Gender differences in the prevalence of DSM-IV and ICD-10 PTSD. *Psychological Medicine*, *36*(1), 81-89.

Porter, M. F. (1980). An Algorithm for suffix stripping. *Program-Automated Library and Information Systems*, *14*(3), 130-137.

Proudfoot, J., Ryden, C., Everitt, B., Shapiro, D. A., Goldberg, D., Mann, A., Tylee, A., Marks, I., & Gray, J. A. (2004). Clinical efficacy of computerized cognitive-behavioral therapy for anxiety and depression in primary care: Randomized controlled trial. *The British Journal of Psychiatry*, *185*(1), 46-54.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127-137.

Rosenberg, J., & Egbert, N. (2011). Online impression management: Personality traits and concerns for secondary goals as predictors of self-presentation tactics on Facebook. *Journal of Computer-Mediated Communication*, *17*, 1-18.

Rosenberg, S. D., & Tucker, G. J. (1979). Verbal-behavior and schizophrenia - semantic dimension. *Archives of General Psychiatry*, *36*(12), 1331-1337.

Ryan, T., & Xenos, S. (2011). Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, *27*, 1658-1664.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinant of emotional state. *Psychological Review*, *69*, 379-399.

Schlenker, B. (2004). Self-presentation. In M. R. Levy & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 492-518). New York, NY: Guilford.

Shih, R. A., Schell, T. L., Hambarsoomian, K., Belzberg, H., & Marshall, G. N. (2010). Prevalence of posttraumatic stress disorder and major depression after trauma center hospitalization. *Journal of Trauma: Injury, Infection and Critical Care*, *69*(6), 1560-1566.

Shrout, P. E., Skodol, A. E., & Dohrenwend, B. P. (1986). A two-stage approach for case identification and diagnosis, first stage instruments. In J. E. Barrett & R. M. Rose (Eds.), *Mental disorders in community: Progress and challenge* (pp. 286-303). New York, NY: Guilford.

Shrout, P. E., & Yager T. J. (1989). Reliability and validity of screening scales: Effects of reducing scale length. *Journal of Clinical Epidemiology*, *42*(1), 69-78.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

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. *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.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, *30*(4), 526-537.

Spitzer, R. L., First, M. B., & Wakefield, J. C. (2007). Saving PTSD from itself in DSM-V. *Journal of Anxiety Disorders*, *21*(2), 233-241.

Tan, C. M., Wang, Y. F., Lee, C. D. (2002). The use of bigrams to enhance text categorization. *Information Processing & Management*, *38*(4), 529-546.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psycology*, *29*(1), 24-54.

Teresi, J., & Fleishman, J. (2007). Differential item functioning and health assessment. *Quality of Life Research*, *16*(1), 33-42.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*, 291-307.

Toma, C., Hancock, J., & Ellison, N. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, *34*, 1023-1036.

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.

van den Berg, S. M., Paap, M. C. S., Derks, E. M., & GROUP (2013). Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. *Psychiatry Research*, *206*(1), 75-80.

van der Linden, W. J., & Glas, C. A. W. (2000). Computerized adaptive testing: Theory and practice. Boston, MA: Kluwer Academic Publishers.

van Groen, M. M., ten Klooster, P. M., Taal, E., van de Laar, M. A. F. J., & Glas, C. A. W. (2010). Application of the health assessment questionnaire disability index to various rheumatic diseases. *Quality of Life Research*, *19*(9), 1255-1263.

Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY: Wiley.

von Davier, M. & Rost, J. (1996). Self Monitoring - A class variable? In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 296-305). Münster, Germany: Waxmann.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weathers, F. W., Huska, J. A., & Keane, T. M. (1991). *The PTSD checklist - cilivilian version (PCL-C).* Boston, MA: The National Center for PTSD, Boston Veterans Affiaris Medical Center.

Weisscher, N., Glas, C. A. W., Vermeulen, M., & de Haan, R. J. (2010). The use of an item response theory-based disability item bank across diseases: Accounting for differential item functioning. *Journal of Clinical Epidemiology*, *63*(5), 543-549.

Wohlfarth, T., van den Brink, W., Winkel, F. W., & ter Smitten, M. (2003). Screening for posttraumatic stress disorder: An evaluation of two self-report scales among crime victims. *Psychological Assessment, 15*(1), 101-109.

Wong, E., Ungvari, G. S., Leung, S. K., & Tang, W. K. (2007). Rating catatonia in patients with chronic schizophrenia: Rasch analysis of the Bush-Francis Catatonia Rating Scale. *International Journal of Methods in Psychiatric Research*, *16*(3), 161-170.

World Health Organization (2001). *The World health report 2001 – Mental health, new understanding, new hope.* Geneva, Switzerland: WHO.

World Health Organization (2008). *WHO Global Burden of Disease*. (http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf) accessed February 5th, 2012.

Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. In: *Proceedings of the 14th international conference on machine learning (ICML 97)*, Nashville, Tennessee, USA.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589-600.

Summary

The information age has made it easy to store and process large amounts of data, including both structured data (e.g., responses to questionnaires) and unstructured data (e.g., natural language or prose). As an additional source of information in assessments, textual data has been increasingly used by cognitive, personality, clinical, and social psychologists in attempt to understand human beings. The questions how to handle these textual data and how to combine them with structured data in psychiatric and psychological assessments are the major themes in this thesis.

The thesis starts with a brief introduction to the initiatives and rationales of developing new methods in handling textual data in psychiatric and psychological assessments in Chapter 1. Besides overviewing the structure of the whole thesis, this chapter proposes two main research questions: (1) how can we apply text mining to narratives collected in the framework of psychiatric and psychological assessment to make classification decisions; and (2) how can we simultaneously model the outcome of text mining and the IRT-based outcomes of responses to questionnaires to validate the text mining procedure and enhance the quality of the measurement and classification procedure. The three main research methodologies applied in the study, that is, text mining for handling unstructured data, item response theory (IRT) for handling structured data, and the combination of these two methods using a Bayesian framework are discussed as well.

Chapter 2 introduces the chi-square feature selection model and presents an alternative machine learning algorithm for binary text classification, named the product score model (PSM). The PSM features in assigning two weights for each keyword (in a binary classification) to indicate the degree that a word represents the two classes. For both classes, a product score can be calculated by multiplying the weights of keywords that occur in a text. Finally, the PSM assigns each text to the class that obtains the highest product score.

In Chapter 3 through Chapter 6, a new intake procedure is developed for detection of posttraumatic stress disorder (PTSD), which combines the utilization of advanced text mining techniques and item response modeling in one framework. The research mainly consists of three parts: (a) computerized text classification on patients' self-narratives to screen for PTSD (Chapter 3 & 4); (b) exploring the generalizability of DSM-IV diagnostic criteria for PTSD using item response modeling (Chapter 5); and (c) combining textual assessment on patients' self-narratives and structured interviews in the PTSD identification process (Chapter 6).

Using 300 self-narratives collected online, a textual assessment method based on the PSM is developed in Chapter 3 to distinguish people with high or low risks to develop as PTSD. The text mining approach results in a high agreement (82%) with the psychiatrists' diagnoses and reveals 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, it can be concluded that the application of text mining is a promising addition to assessing PTSD in clinical and research settings.

An extension of the data representation model from unigrams (i.e., single words) to *n*-grams, where the occurrences of sets of n consecutive words are counted, is further discussed in Chapter 4. Based on the same sample used in the preceding chapter, the PSM, decision trees and naïve Bayes are applied in conjunction with five representation models, unigrams, bigrams, trigrams, a combination of uni- and bigrams and a mixture of *n*-grams, to identify the PTSD patients. Although the PSM with unigrams attains the highest prediction accuracy compared with psychiatrists' diagnoses in structured interviews, it is interesting to find that the addition of *n*-grams contributed most to enhance the reliability of prediction and balance the performance metrics, i.e., resulting in a fairly high sensitivity with the least sacrifice for specificity.

Chapter 5 explores the generalizability of DSM-IV diagnostic criteria for PTSD to various subpopulations by using IRT techniques. Besides identifying differential symptom functioning related to various background variables such as gender, marital status and educational level, this study also emphasizes the importance of evaluating the impact of DIF on population inferences as made in health surveys and clinical trials, and on the diagnosis of individual patients. It is concluded that the DSM-IV diagnostic criteria for PTSD does not produce substantially biased results in the investigated subpopulations, and there should be few reservations regarding their use.

Considering the positive effects in either text mining or IRT as discussed earlier, a combination of these two methods is proposed in Chapter 6 to further strengthen the

benefits from both sides. Text mining and item response modeling are used to analyze patients' writings and responses to standardized questionnaires, respectively. The whole procedure is combined in a Bayesian framework where the textual assessment functions as an informative prior for the estimation of the PTSD latent trait. Results show that adding textual prior information, the detection accuracy is increased and the test length can be shortened.

Chapter 7 extends the application of the model from psychiatric datasets to an Internet dataset, which consisted of both textual posts and responses to the scales related to self-monitoring skills on Facebook (FB). This chapter emphasizes the importance of validating data collected from the Internet and explores the relationship between self-monitoring skills and textual posts on the FB Wall. Textual analysis is conducted via both structured and unstructured approaches. To link the results from these two approaches, the keywords extracted by the text mining techniques are mapped onto the framework of Linguistic Inquiry and Word Count (LIWC), a commonly used psychology-related linguistic software package. The variable of "family" was found the most significant predictor in LIWC. Emoticons and Internet slangs were extracted as the most robust classifiers in the unstructured textual analysis. The conclusion was drawn that textual posts on the FB Wall could partially predict the users' self-monitoring skills. The accuracy rate is expected to enhance if variables from LIWC and keywords extracted from text mining could be used in combination in future studies.

Finally, the thesis concludes the work presented in this thesis in the epilogue. The involvement of text mining provides a new perspective to handle structured and unstructured data in a common framework. Text mining together with IRT is expected to be a promising tool in psychological and psychiatric assessments in the future.

Samenvatting

Het informatietijdperk biedt mogelijkheden om grote hoeveelheden gegevens over personen, zowel gestructureerde data (bijvoorbeeld antwoorden op een vragenlijst) als ongestructureerde data (zoals tekstfragmenten), op te slaan en te verwerken. Deze tekstfragmenten worden door psychologen steeds vaker gebruikt als een extra bron van informatie in hun streven om mensen beter te begrijpen. De vraag hoe deze tekstfragmenten geanalyseerd moeten worden en hoe de analyse van teksten gecombineerd kan worden met de analyse van gestructureerde data voor het meten van psychologische en psychiatrische constructen, staat centraal in dit proefschrift.

Het proefschrift begint in hoofdstuk 1 met een korte introductie in de achtergrond van het ontwikkelen van nieuwe methoden voor het gebruik van tekst data in psychiatrische en psychologische toetsing. De twee belangrijkste onderzoeksvragen worden geïntroduceerd: (1) Hoe kan tekst mining toegepast worden om teksten te classificeren die verzameld zijn in de context van psychiatrische en psychologische toetsing? (2) Hoe kunnen de uitkomsten van tekst mining en van met IRT geanalyseerde vragenlijsten gecombineerd worden om de tekst mining procedure te valideren en de kwaliteit van de meetprocedure te verhogen? In dit hoofdstuk wordt verder een overzicht gegeven van het hele proefschrift en worden de drie belangrijkste methodes beschreven die zijn gebruikt: tekst mining voor het analyseren van ongestructureerde data, item respons theorie (IRT) voor het analyseren van gestructureerde data en het combineren van beide methodes binnen een Bayesiaans framework.

In hoofdstuk 2 wordt een model geïntroduceerd om kenmerken, dat wil zeggen woorden of groepen woorden, uit een tekst te selecteren voor tekst mining op basis van een Chi-kwadraat statistiek. Daarnaast wordt een alternatief tekst mining model gepresenteerd, het Product Score Model (PSM), dat ontwikkeld is om teksten dichotoom te classificeren, bijvoorbeeld ‘wel’ of ‘geen’ stoornis. Binnen het PSM

worden voor elk van de categorieën gewichten toegekend aan een kenmerk. Vervolgens wordt een score berekend door de gewichten van de kenmerken die in de tekst voorkomen met elkaar te vermenigvuldigen. Het PSM wijst een tekst toe aan de categorie met de hoogste score.

In hoofdstuk 3 tot hoofdstuk 6 wordt beschreven hoe in dit promotieonderzoek een nieuwe intake procedure ontwikkeld is voor het detecteren van Post Traumatische Stressstoornis (PTSS). De procedure combineert het gebruik van tekst mining technieken met IRT. Het onderzoek heeft plaatsgevonden in drie fasen. Allereerst is gewerkt aan het geautomatiseerd herkennen van PTSS patiënten door het analyseren van hun blog teksten (hoofdstuk 3 en 4). Vervolgens is verkend in hoeverre IRT gebruikt kan worden om data te analyseren die verkregen zijn met vragenlijsten die gebaseerd zijn op de DSM-IV criteria voor PTSS (hoofdstuk 5). Tenslotte wordt in hoofdstuk 6 beschreven hoe tekst mining en IRT gecombineerd kunnen worden in één procedure.

Uitgangspunt, in hoofdstuk 3, waren 300 verhalen van patiënten waarin ze beschreven hoe het met ze ging en wat ze meegemaakt hadden. Het PSM werd ontwikkeld om onderscheid te maken tussen mensen die een hoog of een laag risico lopen op het ontwikkelen van PTSS. De overeenkomst tussen de classificatie van het model en de diagnose van een psychiater was hoog (82%). Daarnaast gaf het PSM informatie over welke woorden indicatief waren voor PTSS. Alhoewel de classificatie niet helemaal overeenkwam met de diagnose, kan geconcludeerd worden dat tekst mining een veelbelovende toevoeging is aan de bestaande methodes voor PTSS herkenning op klinisch- en onderzoeksgebied.

In hoofdstukken 2 en 3 wordt een tekst geanalyseerd door te tellen hoe vaak losse woorden (unigrams) in de tekst voorkomen. In hoofdstuk 4 wordt de tekst representatie uitgebreid tot *n*-grams – groepen van n opeenvolgende woorden. Dezelfde verhalen als in het voorgaande hoofdstuk worden geanalyseerd met het PSM, beslisbomen en naieve Bayes classificatie, waarbij de teksten worden gerepresenteerd door unigrams, bigrams, trigrams, een combinatie van unigrams en bigrams, of een combinatie van unigrams, bigrams en trigrams. Alhoewel het PSM toegepast op unigrams het risico op PTSS het meest accuraat voorspelde, was het interessant om te zien dat het gebruik van *n*-grams zorgde voor een hogere betrouwbaarheid.

De vraag in hoeverre de DSM-IV criteria voor PTSS gegeneraliseerd kunnen worden naar subpopulaties is onderzocht met IRT (zie hoofdstuk 5). Naast het identificeren van criteria die differential symptom functioning (DSF) laten zien voor subpopulaties gebaseerd op geslacht, burgerlijke staat, of hoogst genoten opleiding,

benadrukte dit onderzoek ook het belang van het bestuderen van DSF bij grootschalige gezondheidsenquetes, bij klinische trials en bij het stellen van individuele diagnoses. In dit onderzoek werd geen systematische bias voor subpopulaties gevonden en de criteria kunnen zonder al te veel reserveringen bij deze subpopulaties worden toegepast.

Gegeven de meerwaarde van zowel tekst mining als IRT, is gewerkt aan het combineren van beide methodes (zie hoofdstuk 6). Een combinatie van methodes is gebruikt om zowel verhalen van patiënten als uitkomsten van vragenlijsten gezamenlijk te analyseren binnen een Bayesiaans kader. De uitkomsten van tekst mining zijn daarbij gebruikt om een empirische prior op te stellen voor de latente PTSS trek. Het toevoegen van tekst mining resulteerde in een hogere accuraatheid, verkleinde de standaard meetfout en kon gebruikt worden om de lengte van de vragenlijst te verkorten.

In hoofdstuk 7 is het gebruik van het PSM uitgebreid van datasets uit de psychiatrie naar een internet dataset, die bestond uit posts op Facebook en vragenlijsten die op facebook waren ingevuld. In dit hoofdstuk wordt het belang van het valideren van data die afkomstig zijn van het internet benadrukt en wordt de relatie tussen Self-Monitoring vaardigheden en posts op Facebook onderzocht. De tekst analyse werd zowel uitgevoerd met het software pakket Linguistic Inquiry and Word Cound (LIWC, een binnen de psychologie veelgebruikt linguistisch software pakket), als met het PSM. Om beide methodes te vergelijken zijn de gevonden PSM kenmerken gelinkt aan het LIWC framework. Bij LIWC was de belangrijkste voorspeller van Self-Monitoring skills de categorie ‘familie’. Bij PSM waren de sterkste voorspellers het gebruik van emoticons en voor social media typerende afkortingen. Er werd aangetoond dat Self-Monitoring skills in redelijke mate voorspeld kunnen worden op basis van iemands posts op Facebook. De accuraatheid van de voorspelling zou verhoogd kunnen worden door de sterkste voorspellers van beide methodes te combineren.

Tenslotte wordt in de epiloog een terugblik gegeven op het onderzoek. De combinatie van tekst mining en IRT biedt nieuwe perspectieven voor het gecombineerd gebruik van gestructureerde en ongestructureerde data. Het is een veelbelovende methode voor psychologische en psychiatrische assessment voor de nabije toekomst.