Computerized Coding System for Life Narratives to Assess Students' Personality Adaption

Q. HE, B.P. VELDKAMP, G.J. WESTERHOF
University of Twente, the Netherlands

The present study is a trial in developing an automatic computerized coding framework with text mining techniques to identify the characteristics of redemption and contamination in life narratives written by undergraduate students. In the initial stage of text classification, the keyword-based classifier algorithm – product score model – performed more sensitive in finding the "change" elements in life stories comparing to the Decision Tree and Naïve Bayes models. The verbal features of life narratives were also discussed to enhance the classification accuracy further in assessing students' personality adaption.

Key Words and Phrases: Text mining, text classification, personality adaption, life narratives

1. INTRODUCTION

In the first decade of the 21st century, narrative approaches to personality have moved to the center of the discipline of personality psychology (McAdams, 2008). Besides the traditional qualitative methods and case studies, an increasing number of quantitative studies have been used in narrative research. For instance, McAdams and his colleagues developed objective coding systems for assessing narrative tone, themes of agency and communion in life-narrative accounts, redemption and contamination sequences, and goal articulation (see www.sesp.northwestern.edu/foley). However, such coding systems are specified for manual rating rather than computer-based assessment, which results in heavy workloads for the human raters.

The present study is to develop an automatic computerized coding framework corresponding to McAdams's manual coding system in identifying the characteristics of redemption and contamination based on a sample of life narratives written by the undergraduate students. Redemption and contamination are the two most important sequences to reveal the "change" tendency in people's emotional well-being through writing. 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. There are two specific objectives in the current research: (1) to develop a text mining model to accurately categorize the students' life narratives; and (2) to extract verbal features in life narratives to enhance the prediction further.

2. METHOD

The collected life stories were written by 271 undergraduate students in the United States, containing 656 life stories in total. The target classification was labeled into four categories: redemption (RED), contamination (CON), both of redemption and contamination (BOTH), neither redemption nor contamination (NEITHER). Three experienced human raters (kappa=0.67) labeled each story corresponding to the manual coding system. These results were set as the "standards" in the performance evaluation.

Based on a common feature – the "change" tendency in both redemption and contamination life stories, we constructed a hierarchical classification framework shown in the Figure 1. On the first stage, all the input were divided into two groups, "change" and "nochange", by identifying the presence and absence of "change" features in the stories. Based on these preliminary results, a more detailed classification was conducted

Authors' addresses: Q. He, Department of Methodology, Measurement and Data Analysis, University of Twente, the Netherlands. E-mail: q.he@gw.utwente.nl; B.P. Veldkamp, Department of Methodology, Measurement and Data Analysis, University of Twente, the Netherlands. E-mail: b.p.veldkamp@gw.utwente.nl; G.J. Westerhof, Department of Health and Technology, University of Twente, the Netherlands; E-mail: g.j.westerhof@utwente.nl

on the second stage to label each story based on the patterns extracted for redemption and contamination.

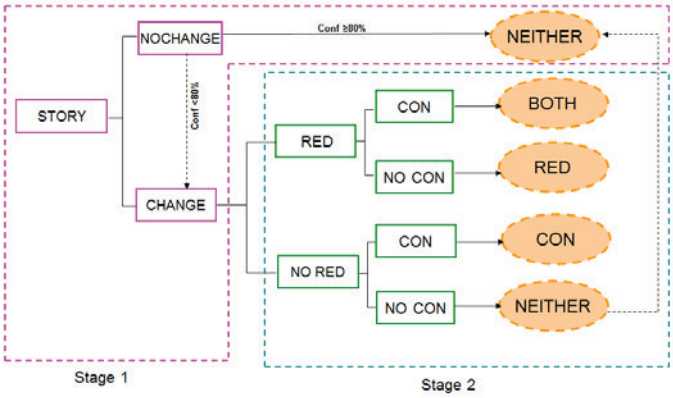

Fig 1: The hierarchical classification framework for life story computerized coding system

The dataset was split into two sets by a random selection of 70% as training set and 30% as test set. During training, the most discriminating keywords to determine the story with or without the tendency of "change" were extracted using the Chi-Square selection algorithm (Oakes et al., 2001). Afterwards, pairs of feature sets and labels were fed into a machine learning algorithm to "learn" their relationship and generated a classifier model. During prediction, the same feature extractor was used to convert new inputs to feature sets. These features were then fed into the "trained" classifier model, which predicted the most likely label for the new input. We utilized two standard machine learning algorithms, Decision Tree and Naïve Bayes in the textual classification process and also constructed an alternative model named product score, which is analogous to the log-likelihood ratio test.

3. RESULTS

The relationship between computerized coding method at various thresholds and the gold standard was assessed through 2-by-2 tables with four indicators, precision, recall, overall correct classification, and F1 score. In the initial stage of text classification, the product score model with the cutoff threshold at (-4) yielded the highest sensitivity in finding the "change" elements in life stories compared with the Decision Tree and Naïve Bayes models, with a precision rate around 85% but sacrificing the recall rate around 50%. Further analysis for the second stage is still in process.

4. CONCLUSIONS

The present study represents an initial development of text classifier model for computerized coding system in life narratives. This trial demonstrated that the text mining technique was very promising in assessing the people's personality adaption with verbal features. Some latent semantic analysis and sequence measurement model are also expected to add in the future research.

REFERENCES

McAdams. D. P. 2008. Personal narratives and the life story. In John, Robins, & Pervin (Eds.), *Handbook of Personality: Theory and research (3rd Ed.)* NY: Vilfort Press.

Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, A., Wan, V. & Beaulieu, M. 2001. A method based on Chi-Square test for document classification. In *Proceedings of the 24th Annual International ACM SIGIR Comference on Research and Development in Information Retrieval (SIGIR 21)*, 440-441.