Optimization of Life Satisfaction Prediction Models Based on Text Data Augmentation

Authors: Chen Jiajing, Hu Dingding, Song Rui, Tan Shiqi, Li Yuqing, Zhang Shengnan, Zhu Tingshao, Zhao Nan, Zhu Tingshao, Zhao Nan

Date: 2024-02-29T13:31:45Z

Abstract

Objective: With the development of internet big data and machine learning methods, an increasing number of studies combine text analysis and machine learning to predict satisfaction. In research on building life satisfaction prediction models, to address the challenge of obtaining large amounts of effective labeled data, this study proposes text data augmentation to optimize life satisfaction prediction models. Methods: After adapting the Dalian University of Technology dictionary, using 357 life status descriptions as original text, with self-reported scores from the life satisfaction scale as labels, performing text data augmentation via EDA and back-translation, prediction models were built using traditional machine learning algorithms. Results: Results show that after adapting the Dalian University of Technology dictionary, the predictive capability of all models improved substantially; following data augmentation, enhancement effects of back-translation and EDA were only observed in linear regression models. The ridge regression model trained on original data achieved the highest Pearson correlation coefficient between predicted and actual values, reaching 0.4131. Conclusion: Improving feature extraction accuracy can optimize current life satisfaction prediction models, but for life satisfaction prediction models built with word frequency as features, text data augmentation based on back-translation and EDA may not be highly suitable.

Full Text

Optimization of a Life Satisfaction Prediction Model Based on Text Data Augmentation

Chen Jiajing[12], Hu Dingding[12], Song Rui[12], Tan Shiqi[12], Li Yuqing[12], Zhang Shengnan[12], Zhu Tingshao[1], *Zhao Nan[1]* [1] Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract

[**Objective**] With the development of network big data and machine learning methods, an increasing number of studies have begun combining text analysis with machine learning to predict individual satisfaction. In research focused on building life satisfaction prediction models, obtaining large amounts of valid labeled data presents a significant challenge. This study proposes using text data augmentation to optimize life satisfaction prediction models and address the difficulty of acquiring extensive annotated datasets.

[**Method**] After adapting the Dalian University of Technology (DLUT) emotion dictionary, we used 357 life status descriptions as original text data, with self-reported scores from the Life Satisfaction Scale as labels. We applied Easy Data Augmentation (EDA) and back-translation for text data augmentation, then established prediction models using traditional machine learning algorithms.

[**Results**] Results demonstrated that after adapting the DLUT dictionary, the predictive capability of all models improved substantially. Following data augmentation, enhancement effects from both back-translation and EDA were observed only in the linear regression model. The ridge regression model trained on original data achieved the highest Pearson correlation coefficient between predicted and actual values, reaching 0.4131.

[**Conclusion**] Improvements in feature extraction accuracy can optimize current life satisfaction prediction models. However, text data augmentation based on back-translation and EDA may not be particularly suitable for life satisfaction prediction models built using word frequency features.

Keywords: Life Satisfaction; DLUT-Emotionontology; Text Data Augmentation; Back-translation; EDA; Machine Learning

1. Introduction

Life satisfaction refers to an individual's subjective evaluation of their quality of life based on self-established criteria, representing a comprehensive judgment of one's own life (Papadopulos et al., 2007). The most common measurement approach for life satisfaction is questionnaire surveys, such as Diener's Satisfaction with Life Scale (SWLS). Although these scales demonstrate high reliability and validity, researchers have noted that questionnaire-based measurement of life satisfaction is influenced by contextual factors, memory effects, and participants' willingness to respond, resulting in insufficient ecological validity.

In recent years, with the rise of machine learning, researchers have proposed using text analysis and machine learning to establish satisfaction prediction models. The widely adopted machine learning approach is supervised learn-

ing, which involves annotating training data before applying machine learning models to predict outcome variables and achieve high prediction accuracy (Li et al., 2021; Peng et al., 2021). Building on this foundation, existing studies have employed dictionary-based segmentation and sentiment analysis to model and predict individual subjective well-being (Li et al., 2015; Wang et al., 2020), environmental satisfaction (Z. Wang et al., 2021), and electronic product satisfaction (Chatterjee et al., 2021), achieving Pearson correlation coefficients of 0.3–0.5. However, in current life satisfaction research, obtaining annotated data is extremely difficult, and smaller datasets may lead to model overfitting. Consequently, how to acquire large amounts of effective labeled data represents an urgent problem that needs to be solved when using machine learning to build life satisfaction models.

When datasets are small, data augmentation techniques can enable models to demonstrate better generalization capability and performance. Data augmentation refers to methods that increase data volume by making slight modifications to existing data to generate copies or creating new synthetic data from existing data (Li, Hou, & Che, 2021). Initially widely applied in computer vision (e.g., image flipping and rotation), these techniques were subsequently introduced to natural language processing (NLP) as text data augmentation. Currently, applications of data augmentation in NLP remain relatively limited. The main methods for text data augmentation include: (1) Lexical replacement, which replaces certain parts of the original text without changing the sentence's meaning (Ma & Langlang, 2020); (2) Easy Data Augmentation (EDA), which comprises four simple yet powerful operations: synonym replacement, random insertion, random swap, and random deletion—for a given training set, one of these four operations is randomly selected and applied (Wei & Zou, 2019); and (3) Back-translation, which uses machine translation to paraphrase and generate new text, typically by first translating into another language and then back to the original language, thereby expanding the dataset while preserving semantics (Lun, Zhu, Tang, & Yang, 2020; Ma & Langlang, 2020).

Text data augmentation has been applied in both traditional models and neural network models, demonstrating good effectiveness in improving predictive capability. In traditional models, Abdelrahman ElNaka (2021) and colleagues employed data augmentation methods including back-translation to expand datasets, resulting in significant performance improvements for random forest, support vector machine, and other models (ElNaka, Nael, Afifi, & Nada Sharaf, 2021). In neural network models, Jun Ma (2020) and colleagues also found that back-translation data augmentation could enhance the classification capability of deep learning models for Chinese text (Ma & Langlang, 2020). Additionally, Jiaqi Lun (2020) and colleagues demonstrated significant effects of text data augmentation in deep learning models for short answer scoring (Lun et al., 2020).

In summary, to address the problem of insufficient datasets in satisfaction machine learning modeling, this study employs back-translation and EDA for text

data augmentation to expand the dataset, aiming to establish a life satisfaction prediction model with high predictive capability.

2.1 Sample Sets

This study included four sample sets: the original sample set, the back-translation sample set, EDA sample set 1, and EDA sample set 2.

Original Sample Set: We randomly selected 392 graduate and doctoral students from the University of Chinese Academy of Sciences and asked participants to complete the Life Satisfaction Scale. We calculated their total scores and requested them to write a brief self-report in a txt document describing their evaluation of or thoughts about their current life situation, while annotating their life satisfaction scale scores and gender. Six screeners evaluated the 392 text documents based on criteria requiring self-reports to be at least 300 characters long, written in first-person narrative, not copied from others, and containing relevant descriptions of mood and life experiences. Additionally, the six screeners rated each text according to the five dimensions of the Life Satisfaction Scale (inter-rater reliability: Kendall's $W = 0.88$, $p < 0.001$). Texts were excluded if more than two raters' scores differed from the self-reported score by more than five points. After screening, 35 documents were eliminated, leaving 357 qualified texts (96 male participants, 261 female participants).

Back-translation Sample Set: We performed six rounds of back-translation on the training set, increasing the original text data sixfold and merging it with the original texts to obtain 2,499 samples, constituting the back-translation sample set.

EDA Sample Set 1: For each original text, we randomly performed synonym replacement, random insertion, random swap, or random deletion with a modification ratio (alpha) of 0.05 (Wei & Zou, 2019). To maintain the same size as the back-translation training set, we increased the existing text data sixfold through EDA rewriting, obtaining 2,499 samples that were merged with the original texts to form EDA sample set 1.

EDA Sample Set 2: Based on Wei & Zou (2019), the optimal augmentation multiplier for our original sample was 16. Therefore, using the same modification ratio (0.05) as EDA sample set 1, we increased the existing text data 16-fold through EDA and merged it with the original texts, obtaining 6,069 samples to form EDA sample set 2.

2.2 Tools

Satisfaction with Life Scale (SWLS) Chinese Version (Diener et al., 1985): This is a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), comprising five items. The sum of scores serves as the participant's total life satisfaction score. Reliability analysis showed a Cronbach's

coefficient of 0.78 and split-half reliability of 0.70, indicating good reliability for the Life Satisfaction Scale.

Adapted DLUT-Emotionontology Dictionary: The existing 21-dimension DLUT emotion dictionary could not sufficiently cover emotional expression vocabulary. We first added a Weibo common emotion lexicon containing five dimensions to this dictionary. Further observation of segmentation results revealed that both the DLUT dictionary and the Weibo emotion lexicon lacked compound words in the form of "negation word + emotion word," yet such expressions appeared frequently. To improve feature extraction accuracy, we established a negation word library and identified all unique "negation word + emotion word" collocations from all texts. After discussion among a six-person team, we selected 1,125 unambiguous words from 1,496 compound emotion words to expand the emotion lexicon, adding three categories of compound emotion words: P (Positive), N (Negative), and Ne (Neutral). The final emotion lexicon contained 29 dimensions.

2.3.1 Data Augmentation

Back-translation: We called the Baidu Translation API in Python to perform six rounds of back-translation on the training set: Chinese-English-Chinese, Chinese-French-Chinese, Chinese-German-Chinese, Chinese-Russian-Chinese, Chinese-Korean-Chinese, and Chinese-Japanese-Chinese.

EDA: Adapted from Zhanlaoban's (2019) GitHub program in Python, we first used the jieba segmentation package to segment the original text, then performed synonym replacement, random insertion, random swap, or random deletion on the segmentation results of each text. Each EDA application used only one modification method. For synonym replacement, we called the Chinese synonyms package (synonyms) to find a series of synonyms for each of the n selected words (non-stop words) and randomly selected synonyms for replacement. Random insertion was completed by finding random synonyms for n words (non-stop words) in the sentence and inserting them at random positions. Random swap was completed by randomly selecting two words in the sentence and exchanging their positions, repeated n times. Random deletion was completed by deleting words from the text with probability p.

2.3.2 Feature Extraction

We used the jieba segmentation package and the adapted DLUT dictionary to segment the cleaned life satisfaction text data and removed words that provided relatively little information after segmentation. Subsequently, based on the adapted DLUT dictionary, we calculated the word frequency of 29 emotion word dimensions for each text, obtaining 29 features.

2.3.3 Model Building and Effectiveness Testing

To ensure consistent data partitioning, after each division of training set (80%) and test set (20%), we removed augmented texts from the test set so that only the training set text data was augmented, resulting in an augmented training set, an unaugmented training set, and a test set. We called Python's scikit-learn machine learning package to build six models: linear regression, ridge regression, random forest regression, decision tree, support vector regression, and Gaussian process regression. We trained each model using both the augmented training set and the unaugmented training set, obtaining six augmented models and six original models. We input the test set feature values into both augmented and original models to obtain test set predictions from 12 models. Pearson correlation analysis between these predictions and actual values yielded model prediction capability metrics. This entire process was repeated 100 times.

3.1 Impact of DLUT Dictionary Adaptation on Model Prediction Capability

We examined dictionary adaptation effects using the original sample set. After adding the Weibo lexicon, all models except ridge regression showed improved Pearson correlation coefficients between predicted and actual values. After screening all texts (including back-translated and rewritten texts) for unique "negation word + emotion word" collocations and expanding the dictionary, the Pearson correlation coefficients of all models increased by 0.09–0.13 (Table 1). The optimal prediction model was support vector regression, which achieved a maximum Pearson correlation coefficient of 0.5971 between predicted and actual values across 100 random training-test set partitions.

Table 1. Impact of Dictionary Adaptation on Model Prediction Capability (r) | Model | Original DLUT Dictionary | Weibo Lexicon + DLUT Dictionary | Adapted DLUT Dictionary | |——-|————————|————————|————————-| | Linear Regression | 0.23618 | 0.29207 | 0.34121 | | Ridge Regression | 0.30447 | 0.12607 | 0.42394 | | Random Forest | 0.29755 | 0.20791 | 0.41195 | | Decision Tree | - | - | 0.41216 |

Note: Values represent Pearson correlation coefficients (r) averaged across 100 random training-test set partitions.

3.2 Impact of Data Augmentation on Model Prediction Capability

Since feature extraction using the adapted DLUT dictionary yielded optimal model prediction capability, all subsequent feature extraction employed the adapted dictionary.

Pearson correlation analysis between model predictions and life satisfaction self-ratings revealed different effects of data augmentation across the six models:

linear regression, ridge regression, random forest regression, decision tree, support vector regression, and Gaussian process regression (Table 2).

Table 2. Impact of Data Augmentation on Model Prediction Capability in Test Sets (r) | Model | Back-translation Sample Set | EDA Sample Set 1 | EDA Sample Set 2 | |——-|————————-|————————|————————| | | Unaugmented | Augmented | Unaugmented | Augmented | Unaugmented | Augmented | | Linear Regression | - | - | - | - | - | - | | Ridge Regression | - | - | - | - | - | - | | Random Forest | - | - | - | - | - | - | | Decision Tree | - | - | - | - | - | - | | SVR | - | - | - | - | - | - | | GPR | - | - | - | - | - | - |

Note: Values represent Pearson correlation coefficients (r) averaged across 100 random training-test set partitions.

As shown in Table 1 and Figure 1, when using back-translation for data augmentation, only the linear regression model's Pearson correlation coefficient increased by 0.04, while all other models' coefficients decreased (by 0.002–0.029). When using 6x EDA augmentation, all models' Pearson correlation coefficients decreased (by 0.017–0.049). When using 16x EDA augmentation, only the linear regression model's Pearson correlation coefficient increased slightly by 0.010, while all other models' coefficients decreased (by 0.006–0.038).

Across all 100 training iterations, the ridge regression model achieved the highest Pearson correlation coefficient (r = 0.41647) when trained on the unaugmented sample set.

Figure 1. Changes in Model Prediction Capability in Test Sets Before and After Data Augmentation
(Note: Linear=Linear Regression, Ridge=Ridge Regression, RFR=Random Forest Regression, DT=Decision Tree, SVR=Support Vector Regression, GPR=Gaussian Process Regression)

Discussion

With continuous advancements in information technology, people now share their thoughts across various platforms, all presented in textual form. Analyzing these texts holds promise for predicting individuals' life satisfaction. To address the problem of insufficient datasets in satisfaction machine learning modeling, this study employed back-translation and EDA for text data augmentation to expand the dataset, aiming to establish a life satisfaction prediction model with high predictive capability. Results demonstrated that the adapted DLUT dictionary enhanced the predictive capability of traditional machine learning models. Regarding data augmentation, back-translation and EDA methods only improved the predictive capability of the linear regression model.

After adding the Weibo lexicon and compound emotion words to the DLUT dictionary, the optimal correlation coefficient between support vector regression predictions of individual life satisfaction and individuals' self-reported life satisfaction scores reached 0.5 without data augmentation. Previous research

indicates that correlation coefficients between different measurement tools in social and personality psychology range from 0.39 to 0.68 (Li et al., 2015). Additionally, studies using machine learning algorithms to predict individual subjective well-being have reported optimal results between 0.27 and 0.60 (Li et al., 2015). In this study, the correlation coefficient for model-predicted life satisfaction reached approximately 0.5, indicating good model performance. This result suggests that the approach combining text analysis and machine learning algorithms to predict individual life satisfaction is relatively reliable.

This study used back-translation and EDA methods to augment text data, finding that after augmentation, model performance varied, with only the linear regression model showing improved predictive capability. These findings differ somewhat from previous literature. Some researchers have used data augmentation methods including back-translation, EDA, and pretrained semantic models to expand datasets for prediction with various machine learning models, finding that linear regression and support vector machine models improved after data augmentation, while random forest models deteriorated (Ansari, Garg, & Saxena, 2021). However, other studies extracting word vectors as features and using support vector machines, random forests, and neural networks on unaugmented and back-translation-augmented text datasets found that all three machine learning algorithms performed better with back-translation-augmented data compared to unaugmented data (ElNaka et al., 2021). Regarding this discrepancy, some researchers suggest that data augmentation effects vary depending on the augmentation method—weak augmentation often improves prediction accuracy, while strong augmentation may reduce it (Min et al., 2021). Additionally, Raghunathan et al. (2020) found that training models with augmented data produces smaller robust errors but may generate larger standard errors. Furthermore, in natural language processing research, most studies using data augmentation for data preprocessing are based on deep learning models for prediction, possibly because deep learning models heavily rely on large data volumes to avoid overfitting (Wen et al., 2020; Shorten, Khoshgoftaar, & Furht, 2021). These findings suggest that data augmentation effects may depend on augmentation methods, feature extraction approaches, and model characteristics. Therefore, future research could explore feature extraction methods such as sentiment analysis and word vector computation, or employ deep learning models for training and prediction.

After adapting the DLUT dictionary, all models showed substantial improvements in predictive capability, demonstrating that enhanced feature extraction accuracy can improve current life satisfaction prediction models. However, after text data augmentation, data augmentation only improved model predictive capability in the linear regression model, suggesting that text data augmentation based on back-translation and EDA may not be particularly suitable for life satisfaction prediction models built using word frequency features.

References

Li, A., Hao, B., Bai, S., & Zhu, T. (2015). Psychological computing based on network data analysis: Focusing on mental health status and subjective well-being. *Chinese Science Bulletin*, 60(11), 994–1001.

Li, J., Liu, D., Wan, C., Liu, X., Qiu, X., Bao, L., & Zhu, T. (2021). A survey on automatic mental health assessment for social network users. *Journal of Chinese Information Processing*, 35(2), 19–32.

Peng, J., Zhao, Y., & Wang, L. (2021). Research on video anomaly detection based on deep learning. *Laser & Optoelectronics Progress*, 58(6), 51–61.

Ansari, G., Garg, M., & Saxena, C. (2021). Data Augmentation for Mental Health Classification on Social Media. *arXiv preprint arXiv:2112.10064*.

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2021). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research*, 131(January 2020), 815–825. https://doi.org/10.1016/j.jbusres.2020.10.043

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75.

ElNaka, A., Nael, O., Afifi, H., & Sharaf, N. (2021). AraScore: Investigating Response-Based Arabic Short Answer Scoring. *Procedia Computer Science*, 189, 282–291.

Li, B., Hou, Y., & Che, W. (2021). Data Augmentation Approaches in Natural Language Processing: A Survey. *Journal of LATEX Templates*.

Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020). Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring. *The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence*, 13446–13453.

Ma, J., & Langlang. (2020). Data Augmentation For Chinese Text Classification Using Back-Translation. *Journal of Physics: Conference Series*, 1651(2020) 012039.

Min, Y., Chen, L., & Karbasi, A. (2021, December). The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence* (pp. 129–139).

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., & Liang, P. (2020). Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data*, 8(1), 1–34.

Wang, Y., Wu, P., Liu, X., Li, S., Zhu, T., & Zhao, N. (2020). Subjective well-being of Chinese Sina Weibo users in residential lockdown during the COVID-19 pandemic: Machine learning analysis. *Journal of Medical Internet Research*, 22(12), 1–13. https://doi.org/10.2196/24775

Wang, Z., Zhu, Z., Xu, M., & Qureshi, S. (2021). Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning. *Science of the Total Environment*, 776, 145908.

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv*.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2020). Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*.

ZaneMuir. (2017). *DLUT-Emotionontology*. https://github.com/ZaneMuir/DLUT-Emotionontology

Zhanlaoban. (2019). *EDA_NLP_for_Chinese*. https://github.com/zhanlaoban/EDA_NLP_for_Chinese/co

Author Contributions

Chen Jiajing: Research design, data cleaning, code implementation, data analysis, writing of methods, results, and conclusion sections, overall paper integration
Hu Dingding: Research design, data cleaning, code implementation, data analysis, discussion section writing
Song Rui: Research design, data cleaning, code implementation, data analysis, introduction section writing
Tan Shiqi: Research protocol discussion, data cleaning, introduction section writing
Li Yuqing: Research protocol discussion, data cleaning, data analysis, methods section writing, formatting revision
Zhang Shengnan: Research protocol discussion, data cleaning, introduction section writing
Zhu Tingshao, Zhao Nan: Original data collection, technical analysis support, paper guidance

Acknowledgments

We sincerely thank Lu Shengyou from the National Supercomputer Center in Guangzhou, School of Computer Science and Engineering, Sun Yat-sen University for technical support.

Figures

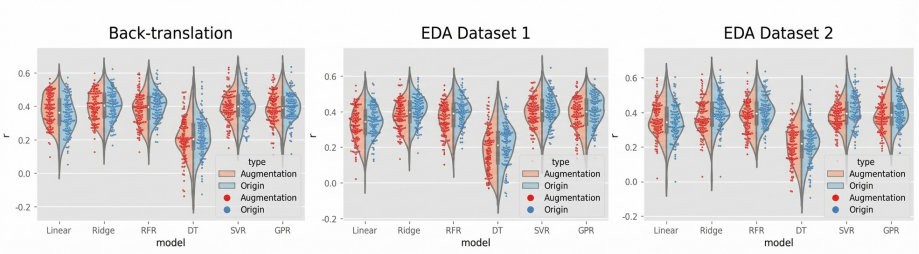

Figure 1: Figure 1

Source: ChinaXiv — Machine translation. Verify with original.