Proceedings of the 2022 Pre-ICIS SIGDSA Symposium

Special Interest Group on Decision Support and Analytics (SIGDSA)

12-12-2022

Quantification of Gender-related Stereotypes in Psychotherapy Sessions

Sagar Mahesh Badve

Satvik Chekuri

Wenqi Shen

G. Alan Wang

Quantification of Gender-related Stereotypes in Psychotherapy Sessions

Research-in-Progress (Extended Abstract)

Sagar Mahesh Badve
Virginia Tech
sagarbadve@vt.edu

Wenqi Shen
Virginia Tech
shenw@vt.edu

Satvik Chekuri
Virginia Tech
satvikchekuri@vt.edu

Alan Wang
Virginia Tech
alanwang@vt.edu

Abstract

Gender-related stereotypes and biases can have severe consequences in the medical domain, especially in mental health therapy. In this study, we analyzed 91 psychotherapy transcripts from the Alexander Street database to investigate whether gender-related stereotypes differ in the treatment of patients by male versus female therapists using natural language processing and statistical analyses. We built a lexicon of ten high-level categories that capture sentence-level attributes and represent gender-related stereotypes. Our results suggest significant statistical differences in categories such as *active, negatives, positives*, etc., during the treatment of female patients by male therapists as compared to female therapists. We built logistic regression models using the ten high-level lexical categories to predict the gender of the therapist. We also provide recommendations on how our analytical methods can be used, along with other advanced deep-learning methods, to detect and reduce gender-related stereotypes in psychotherapy sessions.

Keywords

Gender-related stereotypes, psychotherapy sessions, natural language processing, and statistical analyses.

Introduction

Gender-related stereotyping and bias have severe consequences in healthcare. For example, female patients' treatment could involve less use of aggressive analgesics than males (Chen et al. 2008). Gender bias can cause significant issues in patient diagnosis, treatment (Eichler et al. 1992), and satisfaction. Gender-related stereotypes and gender blindness are characterized as the leading causes of gender bias (Hamberg 2008). Hence, identifying gender-related stereotypes is essential as they are considered the main indicators of gender bias. Gender-related stereotyping is also one of the underlying factors in people facing depression (Nugent and Jones 2005). In mental healthcare, identifying gender-related stereotypes is vital to safeguard the conversations between patients and psychotherapists.

In this research-in-progress paper, we investigate potential indicators of gender bias and attempt to quantify the gender-related stereotypes in psychotherapy sessions to gain more clarity on the gender-related stereotypes using natural language processing and text analytics approach. We investigate:

Research Question 1 (RQ1): Do gender-related stereotypes differ in the therapy of patients by therapists?

> Specifically, RQ1a. Do gender-related stereotypes differ in the therapy of male patients by male therapists as opposed to female therapists?

> RQ1b. Do gender-related stereotypes differ in the therapy of female patients by male therapists as opposed to female therapists?

RQ1c. Do gender-related stereotypes differ by male therapists as opposed to female therapists in psychotherapy sessions involving a therapist and a patient of the same gender?

Research Question 2: If gender stereotypes exist among therapists, can we predict a therapist's gender using the high-level lexical categories based on the adjectives and verbs used in the therapy sessions?

Our research methodology provides a good starting point to investigate gender-related stereotypes in psychotherapy sessions. We built ten high-level lexical categories for gender-related stereotypes in the psychotherapy domain. Additionally, we found statistically significant differences in the investigated gender-related stereotypes towards patients by male and female therapists. The following sections of the paper are organized with the literature review on gender stereotypes, research methodology, results, conclusion, limitation, and future work.

Literature Review

Prior studies show the existence of gender-related stereotypes in medical practice (Andersson et al. 2013; Broverman et al. 1970). This suggests a significant need to examine gender stereotypes in psychotherapy sessions where the therapists interact more with the patients than in other medical streams. Our proposed linguistic method is developed based on the syntactic categories in the linguistic theory (White 2005), such as nouns, pronouns, verbs, and adjectives, and their relationships with each other, which allow us to form grammatically correct sentences and also interpret their meaning correctly. The lexical categories developed in other domains may not be directly applied to healthcare, especially psychotherapy sessions, because domain-specific terms exist (Abrahams et al. 2015). We aim to identify gender-related stereotyping lexical categories for psychotherapy sessions that may be applied to another healthcare context by investigating works in various streams related to gender stereotyping.

Research Methodology

This section discusses the data, the building of the lexical categories to capture gender-related stereotypes, and the statistical methods used to answer our research questions.

Data and Feature Extraction

We collected 91 psychotherapy session transcripts between patients and therapists from the Alexander Street online repository (https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series), in which the therapist's and patients' genders were identifiable. Out of 91 psychotherapy session transcripts, 46 are for male patients, and 45 are for female patients to quantify the gender-related stereotyping by therapists, which are indicators of gender bias. The sentence-level analysis allows us to use the verbs and adjectives associated with pronouns related to the gender of the entity involved (in our case, the patients) as indicators in the sentences, as discussed in prior work on gender stereotypes (Fast et al. 2016b). Since behavior is related to characterizing stereotypes, we extracted the subject-verb relationships where the subject is one of the pronouns related to the patient's gender and lemmatized the verbs to remove redundancies and duplicates. We also extracted and lemmatized the adjective descriptors for the patient's gender because stereotypes can also manifest in how people are described (Fast et al. 2016b). The adjectives were extracted by targeting those related to words associated with male or female patients.

Building Lexical Categories to capture the Gender-related Stereotypes

Several prior studies have identified high-level lexical categories or studied various features related to gender-related stereotypes in domains like television, social media, and natural language (Fast et al. 2016b; Lauzen et al. 2008; Tang 2019; Towbin et al. 2004). Moreover, there are standard tools like the *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker et al. 2015) and Empath (Fast et al. 2016a) that are used widely to classify attributes into higher-level categories. Using all the above resources, two taggers from among the authors independently worked on identifying ten high-level categories to capture the gender-related stereotypes. We labeled the 630 most frequently occurring unique verbs and adjectives into those ten lexical categories, leaving us with a total of 11848 verbs and adjectives for male patients and a total of 22634 verbs and adjectives for female patients. The inter-rater reliability score of the two taggers for the labeling task was 80%, which is a good measure of consistency (Landis and Koch 1977). A third tagger also

assisted throughout the labeling process for further verification. The ten high-level categories, along with examples, are as follows: *communicative* (respond, talk, etc.), *insightful* (conscious, gauge, etc.), *dominant* (assertive, disruptive, etc.), *active* (persist, perform, etc.), *emotional* (love, care, etc.), *negatives* (jealous, insecure, etc.), *positives* (pleasant, supportive, etc.), *strong* (strong, energetic, etc.), *weak* (struggle, sensitive, etc.), and *wishful* (wish, hope, need, etc.).

Research Methods

To answer RQ1 (1a, 1b, and 1c), we tested for a difference in mean frequency counts, taken over the number of sessions, between the two groups for our ten stereotypical categories. We also used the two-sided Welch's t-test and applied a Bonferroni correction of alpha = 0.05/10 = 0.005. The statistically significant results (alpha of 0.005) are presented as odds ratios between means. The odds ratio indicates how likely an attribute (a verb or an adjective) is used to characterize males rather than females (Tang 2019).

To investigate RQ2, we built logistic regression models that used frequency counts from the ten high-level lexical categories, formed on the adjectives and verbs used in the therapy session, to predict the gender of the therapist during the session. These models would provide insights into which of the ten lexical categories are significant in identifying the therapist's gender. Previous studies have developed such models to gain further insights into linguistic signals (Fast et al. 2016b; Gilbert 2012).

Results

Research question 1a: The results for RQ1a suggest no significant difference (p-value > 0.005) in gender-related stereotyping of male patients by male therapists compared to gender-related stereotyping of male patients by female therapists.

Research question 1b: The results for RQ1b indicate that seven lexical categories – 'insightful', 'active', 'negatives', 'positives', 'strong', 'weak', and 'wishful' – are statistically significant for a p-value < 0.005. Prior work indicates that the categories like 'active' and 'strong' are more frequently associated with men than women (Emons et al. 2010; Lauzen et al. 2008). But in our analysis based on the odds, the usage of fewer words to describe female patients as 'active' or 'strong' is seen only in the case of male therapists as compared to female therapists.

Research question 1c: The results for RQ1c indicate that nine lexical categories – 'insightful', 'dominant', 'active', 'emotional', 'negatives', 'positives', 'strong', 'weak', and 'wishful' – are statistically significant for a p-value < 0.005. We find that the usage of 'dominant' words is significantly higher by female therapists for female patients (odds of 1.52) compared to the use of 'dominant' words by male therapists for male patients. These high odds are in contrast to prior research in which men are more frequently associated with 'dominant' words than women by people of both genders (Fast et al. 2016b).

Research question 2: The results indicate that three of the ten categories - 'active', 'negatives', and 'positives' - are statistically significant predictors of a therapist's gender (p-value < 0.05), with the odds for a female therapist of 1.037, 0.943, and 1.112 respectively for each of the three categories. The male therapists were identified with an F-measure of 0.75, and the female therapists were identified with an F-measure of 0.67.

Conclusion, Limitation, and Future Work

We found crucial and statistically significant results where gender stereotypes differed in the therapy of male and female patients by therapists. The analyses we conducted and the models like ours can be used to detect gender-related stereotyping, along with models based on neural networks and transformers that can automatically correct the words related to gender stereotypes with more gender-neutral words (Pryzant et al. 2020). But even if the sentence-level analysis has its advantages, it does not capture the contextual information in the dialogues. However, our detailed work presents a good starting point to analyze gender-related stereotypes in psychotherapy sessions and avoid unintended consequences for the patients seeking psychotherapy and the providers of psychotherapy. We will continue our research to gain further contextual insights from the therapy transcripts and crowdsource the annotation task from platforms like Amazon Mechanical Turk to capture the perceptions of a larger group of people.

References

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., and Jiao, J. 2015. "An Integrated Text Analytic Framework for Product Defect Discovery," *Production and Operations Management* (24:6), pp. 975-990.

Andersson, J., Salander, P., and Hamberg, K. 2013. "Using Patients' Narratives to Reveal Gender Stereotypes among Medical Students," *Academic Medicine* (88:7), pp. 1015-1021.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P. S., and Vogel, S. R. 1970. "Sex-Role Stereotypes and Clinical Judgments of Mental Health," *Journal of Consulting and Clinical Psychology* (34:1), pp. 1-7.

Chen, E. H., Shofer, F. S., Dean, A. J., Hollander, J. E., Baxt, W. G., Robey, J. L., Sease, K. L., and Mills, A. M. 2008. "Gender Disparity in Analgesic Treatment of Emergency Department Patients with Acute Abdominal Pain," *Academic Emergency Medicine* (15:5), pp. 414-418.

Eichler, M., Reisman, A. R., and Borins, E. M. 1992. "Gender Bias in Medical Research," *Women and Therapy* (12:4), pp. 61-70.

Emons, P., Wester, F., and Scheepers, P. 2010. "'He Works Outside the Home; She Drinks Coffee and Does the Dishes' Gender Roles in Fiction Programs on Dutch Television," *Journal of Broadcasting and Electronic Media* (54:1), pp. 40-53.

Fast, E., Chen, B., and Bernstein, M. S. 2016a. "Empath: Understanding Topic Signals in Large-Scale Text," in: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. pp. 4647-4657.

Fast, E., Vachovsky, T., and Bernstein, M. S. 2016b. "Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community," in: *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*. pp. 112-120.

Gilbert, E. 2012. "Phrases That Signal Workplace Hierarchy," in: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, pp. 1037-1046.

Hamberg, K. 2008. "Gender Bias in Medicine," *Womens Health (Lond)* (4:3), pp. 237-243.

Landis, J. R., and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," *Biometrics* (33), pp. 159-174.

Lauzen, M. M., Dozier, D. M., and Horan, N. 2008. "Constructing Gender Stereotypes through Social Roles in Prime-Time Television," *Journal of Broadcasting and Electronic Media* (52:2), pp. 200-214.

Nicolai, J., and Demmel, R. 2007. "The Impact of Gender Stereotypes on the Evaluation of General Practitioners' Communication Skills: An Experimental Study Using Transcripts of Physician-Patient Encounters," *Patient Education and Counseling* (69:1-3), pp. 200-205.

Nugent, F. A., and Jones, K. D. 2005. "Counseling in a Pluralistic World," in *Introduction to the Profession of Counseling*. Upper Saddle River, NJ: Pearson Education, Inc., pp. 323-364.

Pennebaker, J., Booth, R., Boyd, R., and Francis, M. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates.

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., and Yang, D. 2020. "Automatically Neutralizing Subjective Bias in Text," in: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. pp. 480-489.

Seabold, S., and Perktold, J. 2010. "Statsmodels: Econometric and Statistical Modeling with Python," in: *Proceedings of the 9th Python in Science Conference*. pp. 92-96.

Tang, S. 2019. "Understanding Unconscious Bias by Large-Scale Data Analysis." ProQuest Dissertations and Theses: UC Santa Barbara.

Towbin, M. A., Haddock, S. A., Zimmerman, T. S., Lund, L. K., and Tanner, L. R. 2004. "Images of Gender, Race, Age, and Sexual Orientation in Disney Feature-Length Animated Films," *Journal of Feminist Family Therapy* (15:4), pp. 19-44.

White, L. 2005. "Syntactic Categories in Second Language Acquisition," in *Handbook of Categorization in Cognitive Science,* H. Cohen, & Lefebvre, C. (ed.). Elsevier Science Ltd, pp. 515-533.

Zygmond, M. J., and Denton, W. 1988. "Gender Bias in Marital Therapy: A Multidimensional Scaling Analysis," *American Journal of Family Therapy* (16:3), pp. 262-272.