Natural Language Processing Framework for Lowering the Incidence of Public Health Disorders

Manish Nandy[1], Ahilya Dubey[2]

[1]*Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India*
[2]*Research Scholar, Department of CS & IT, Kalinga University, Raipur, India*

KEYWORDS
Natural Language Processing, Public Health, Digital Health Interventions, Mental Health

ABSTRACT

Using Digital Technology (DT), which includes collecting, analyzing, and using data from different digital devices, can lower the number of diseases people get and improve their mental health. Digital Health Interventions (DHIs) can help with certain conditions quickly and successfully in a way that is both cost-effective and based on science. Natural Language Processing (NLP) gives ways to analyze writing, better understand interventions' effects, and make therapy decisions. This study aimed to develop a way to use technology to make it easier to automatically analyze both types of written data that are common in DHIs. This method creates textual traits and allows statistical models to predict goal factors like user involvement, condition change, and treatment outcomes. The study supports locating together outcome-optimizing teams that use data from various sources. The research uses complex data analysis and new methods to develop techniques and approaches that make prevention and therapy measures more widely available, accepted, used, and effective.

1. Introduction

Overview of Public Health and Natural Language Processing

Digital Technology (DT), which includes collecting, analyzing, and using data from different digital devices and sources, can lower the number of Mental Health (MH) problems and improve people's overall MH [1]. To do this, medical care and community service networks need to include preventive measures to lower the number of new cases and medical therapies to reduce the severity of current cases. Both of these are required to reduce the general incidence of MH. DT is helpful because it collects a lot of information that can be used to lead methods for prevention and management. The data needs to be reviewed and used to change and improve processes, even digital ones, to get the best results [2]. Improving results needs to be flexible, and DT needs to be aware of how quickly it is used, customer styles and wants, laws, and other factors that can change. It should allow quick changes to the text and delivery of the treatments. To increase the chances of future adoption and spread, working on measures with partners and users is essential.

Digital Health Interventions (DHI) are more scalable, meaning one doctor can care for many people [3]. As the use of DHIs grows, people need to be better at data analytics to judge the success of care and figure out what makes people participate and feel better. A lot of DHIs collect well-organized data that can be used to see how well people are following their treatment plan and see how their symptoms change over time. This involves indicators of symptoms, the number of finished conversations, and how often people can utilize the app [4]. A skilled practitioner (coach) is part of a DHI. The coach helps the customer understand the topic of the intervention, keeps track of their progress, and helps them get past barriers to change.

The next section gives a summary of text analysis methods used in DHIs. This study provides a structured way to use Natural Language Processing (NLP) in this field [5]. It shows how it can be used in a case study combining a DT for Eating Disorders (EDs) and a Healthcare Body Image (HBI) Project [7].

2. Methodology

NLP is a valuable set of tools for analyzing written data generated by DHIs and making forecasting models. NLP helps to understand how online treatments work and how to improve and tailor DHIs. This will lead to more automated technology-based initiatives and lower costs in the long run.

Feature Engineering

The feature engineering process is based on getting text specimens from the interventions or messages that clients and teachers send and receive [6]. To standardize the representation of intervention snippets and communications, which vary in length, the objective was to create a fixed-length vector that captures the essence of each text fragment. This involves changing all text fragments into numerical values or categorical categories.

Metadata

Metadata characteristics include non-contextual descriptive attributes of text snippets independent of meaning or interpretation [8]. Metadata refers to precise characteristics of text, including the quantity and size of phrases, sentences, and paragraphs, the incorporation of punctuation and specialized symbols, the proportion of capital letters, and the arrangement of the content.

Word Usage

It denotes the utilization of specific terminology [10]. Preprocessing encompasses many procedures, including tokenization (i.e., dividing the text into individual terms), stemming (i.e., translating associated terms to a shared base form), transforming terms to lowercase form, eliminating often appearing phrases (sometimes referred to as stop phrases), and substituting synonyms.

Word Embeddings

Word embeddings encode individual words using low-dimensional numerical matrices [12]. This graphical representation is derived by evaluating extensive text corpus and examining the co-occurrences of terms in texts. The underlying notion is that keywords that appear together in texts have shared attributes.

Part-of-Speech Tagging

Part-of-speech (POS) tagging assigns a specific word class to every word in a given text sample, such as nouns, verbs, or adjectives [13]. This classification is based on the word's inherent characteristics and context. The current methodologies and software applications achieve POS categorization accuracies of over 95%.

Predictive and Inference Modeling

Supervised learning methods are used to (A) deduce the progression of symptom intensity over time, (B) forecast a therapeutic result, which might involve premature discontinuation, and (C) deduce the properties of the messages. The following models will be elucidated:

Model A involves deducing the intensity of symptoms as they progress over time. Model A attempts to create a correlation between the intensity of symptoms and text fragments that are contiguous in time. Given that the symptom measures and text fragments are arranged in a sequence, one possible method is dedicating the symptom measure based on the closest textual snippet in the period (before or after the textual snippet's creation). A different approach is to establish a predetermined time frame around a particular text excerpt and compute the mean value across symptom scales during this time frame.

Model B, designed to predict a therapy result, focuses on predicting one target variable for each user. It is desirable to determine at the midpoint of the intervention if a user is probable to continue making progress and what factors facilitate their continued improvement. Since these factors represent one result per user (i.e., symptom levels at the end of the interventions), the features produced by individual textual snippets must be combined by calculating the mean, variation, and linear or nonlinear developments throughout the intervention for each user. A trend measure, such as the mean emotion score per user over the duration, might indicate the treatment success or the progression of symptoms over time (model version A).

Model C involves inferring the qualities of a message based on textual fragments. User communication might be a question, a declaration, or a response to a prior inquiry from the coach. The research assigns a scale to each textual fragment that accurately represents the level of suicide risk for a user. Models of type C use the textual properties for every snippet to deduce the specific feature of interest. Since the textual snippets are associated with particular consumers, hierarchical modeling techniques are used for model categories A and C.

Prevalence Reduction

The prevalence of a condition in a community is determined by two main factors: the number of persons diagnosed with the illness and the number of people who acquire the disease during a specific timeframe. To decrease the occurrence of a disorder, two essential elements are required: proactive measures that lower the number of new cases and adequate medical treatments to ensure that individuals already affected by the disorder no longer connect the medical standards for the disorder.

The research acknowledges that treatments assessed under controlled conditions with specific, often easily accessible populations are not as practical or successful when implemented on a larger scale. To effectively address the prevalence of mental health issues in communities, interventions must be tailored and adjusted to meet the particular demands and interests of various specified groups in a flexible and organized manner [9].

Public Health Model

The approach intersects with various healthcare models, such as the highly influential Reach, Efficiency, Adoption, Deployment, and Management framework. Consistent with previous public health approaches, the research emphasized the significance of providing preventive and interventional strategies to decrease the prevalence [11]. This model distinguishes itself from other models by stressing explicitly the application of digital datasets and digital health treatments and innovations to reduce the prevalence of specific populations. The information and analytics must be tracked and understood by multiple systems codesign and transmission disciplines. This will allow for quick refinement and enhancement of the digital health remedy and its implementation. This strategy aligns with the suggestion that program assessment should be faster and more flexible, with iterative enhancements. The proposed methodology focuses on regularly monitoring and testing significant results to enhance these indices via continuing strategy refinement.

3. Results and discussion

Intervention

Student Bodies–Eating Diseases (SBED) was an online self-help course created explicitly for college-age female students with Eds [15]. Its purpose was to diminish ED pathophysiology and improve negative body image. The intervention consisted of 40 fundamental sessions completed at the user's speed and provided either an online platform or a dedicated application over eight months. The provided material consisted of directed self-help psychoeducational and cognitive behavioral treatment. The content was enhanced by the assistance of online mental health instructors who were either graduate pupils in clinical applications, postdoctoral colleagues, or study employees [14]. Licensed psychiatrists supervised these coaches. Trainers and their designated customers exchanged messages using text inquiries, which were sent over the SBED network.

Studies

This study incorporates information from two trials that examined the effectiveness of the SBED intervention. The HBI Program research is a comprehensive, multicenter, Randomized Controlled Trial (RCT) that aims to evaluate the effectiveness of SBED for undergraduate women with EDs. Participants from 31 higher education institutions in the United States who tested positive for an ED and obtained a medical recommendation were randomly assigned, at the school stage, either to get the intervention or a recommendation for standard treatment at their health facility. Three hundred

seventy-two college students participated in SBED via these activities and were paired with a coach for communication. Consumers in the combined database of both projects produced 38k intervention paragraphs and delivered 4.3k comments to their trainers.

Results

Feature engineering was independently applied to two distinct forms of text data, namely intervening snippets and user communications, because of their substantial variations in content and average duration. Varying hyperparameter selections, such as adjusting the frequency limits for the fraction of word use in all snippets to be covered, affect the amount of characteristics generated, such as the expressive dimension of the word insertions. When selecting hyperparameters for modeling A and C, it is advisable to have a more significant number of textual snippets than attributes. The choices made in this research led to the identification of 250 characteristics on the text fragment level for communications and 280 attributes for intervention messages.

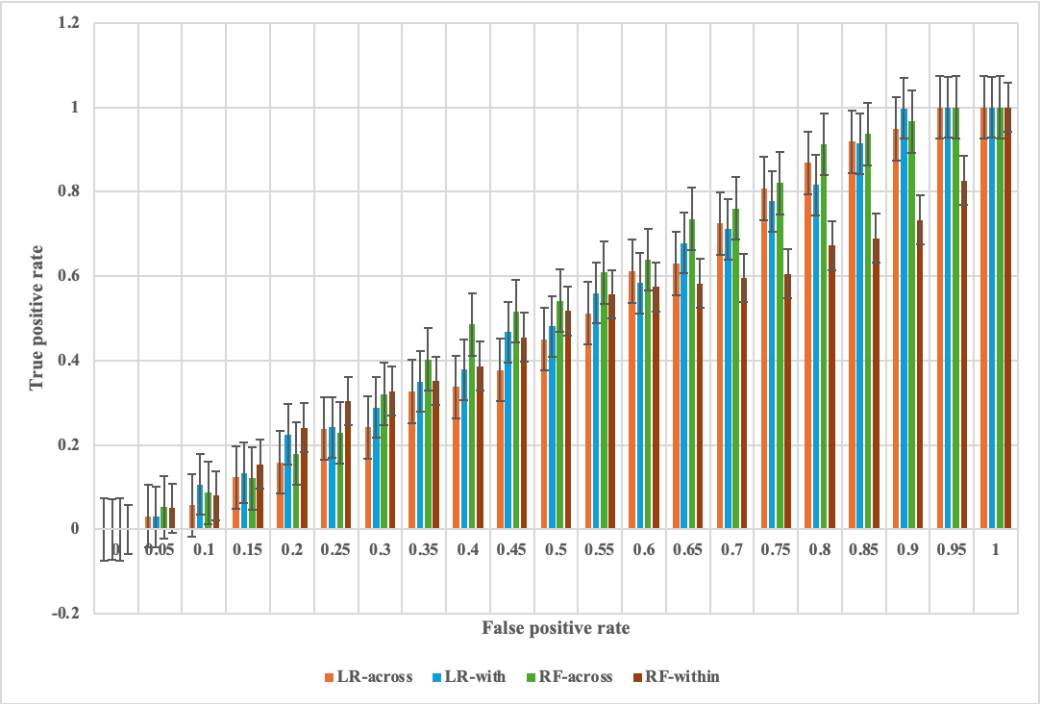

Figure 1. ROC analysis

The Region of Convergence (ROC) charts are calculated using the test information. Area Under the Curve (AUC) = 0.81 for within-user learning using the Random Forest (RF) method indicates that the intervention snippets accurately predict binge eating events over time. During the process of learning across several users, the RF model exhibited overfitting. The Logistic Regression (LR) model produced superior outcomes with an AUC value of 0.63. The ROC analysis findings can guide personalized microinterventions at each participant's level. By analyzing consumers' writing styles, the research recognizes those who are more likely to binge eat during the treatment. These individuals can then be provided with more tailored feedback, such as a brief online chat with a trainer or an additional level of treatment.

4. Conclusion and future scope

Text data enhances and broadens the understanding of the people who provide and use psychiatric therapies delivered digitally. The work presented in this study demonstrates innovation in several aspects. The research is a technological framework that integrates text data into analyzing and predicting significant outcomes in DHIs. The study provides a novel approach that uses word embeddings to analyze the results of interventions, which, to the understanding, has yet to be done.

The research enhances the existing paradigm by utilizing a case study that includes data from a substantial RCT and many text excerpts. Using this information, the research successfully showed that the textual characteristics accurately forecasted symptom changes over time using NLP.

While the study provided in this publication is in its early stages, The research urges other teams to evaluate the potential usefulness of the structure in making treatment decisions. Providing easily accessible, adaptable, cost-effective, and evidence-based DHIs that consider individual consumer preferences, features, and history will contribute to worldwide mental health care initiatives and alleviate the impact of mental illnesses.

Reference

[1] Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). It is understanding the role of digital technologies in education: A review. Sustainable operations and computers, 3, 275-285.

[2] Senbekov, M., Saliev, T., Bukeyeva, Z., Almabayeva, A., Zhanaliyeva, M., Aitenova, N., ... & Fakhradiyev, I. (2020). The recent progress and applications of digital technologies in healthcare: a review. International journal of telemedicine and applications, 2020(1), 8830200.

[3] Erku, D., Khatri, R., Endalamaw, A., Wolka, E., Nigatu, F., Zewdie, A., & Assefa, Y. (2023). Digital health interventions to improve access to and quality of primary health care services: a scoping review. International Journal of Environmental Research and Public Health, 20(19), 6854.

[4] Bobir, A.O., Askariy, M., Otabek, Y.Y., Nodir, R.K., Rakhima, A., Zukhra, Z.Y., Sherzod, A.A. (2024). Utilizing Deep Learning and the Internet of Things to Monitor the Health of Aquatic Ecosystems to Conserve Biodiversity. Natural and Engineering Sciences, 9(1), 72-83.

[5] Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172.

[6] Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for deep learning English language. IEEE Access, 8, 46335-46345.

[7] S. Neelima, Manoj Govindaraj, Dr.K. Subramani, Ahmed ALkhayyat, & Dr. Chippy Mohan. (2024). Factors Influencing Data Utilization and Performance of Health Management Information Systems: A Case Study. Indian Journal of Information Sources and Services, 14(2), 146–152. https://doi.org/10.51983/ijiss-2024.14.2.21

[8] Waterworth, D., Sethuvenkatraman, S., & Sheng, Q. Z. (2021). Advancing smart building readiness: automated metadata extraction using neural language processing methods. Advances in Applied Energy, 3, 100041.

[9] Alamer, L., Alqahtani, I. M., & Shadadi, E. (2023). Intelligent Health Risk and Disease Prediction Using Optimized Naive Bayes Classifier. Journal of Internet Services and Information Security, 13(1), 01-10.

[10] Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.

[11] Malathi, K., Shruthi, S.N., Madhumitha, N., Sreelakshmi, S., Sathya, U., & Sangeetha, P.M. (2024). Medical Data Integration and Interoperability through Remote Monitoring of Healthcare Devices. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(2), 60-72. https://doi.org/10.58346/JOWUA.2024.I2.005

[12] Chiu, B., & Baker, S. (2020). Word embeddings for biomedical natural language processing: A survey. Language and Linguistics Compass, 14(12), e12402.

[13] Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. Journal of Big Data, 9(1), 10.

[14] Stamenkovic, Saša, Stevovic, S., & Stamatovic, Milan. (2019). Res Capacity Increase in EU and Wind Project Sustainability with Case Study on Serbia and Montenegro Market. Archives for Technical Sciences, 1(20), 1–11.

[15] Kalindjian, N., Hirot, F., Stona, A. C., Huas, C., & Godart, N. (2021). Early detection of eating disorders: a scoping review. Eating and Weight Disorders-Studies on Anorexia, Bulimia, and Obesity, 1-48.