Depression score prediction model on Big Data

**Introduction:**

According to the World Health Organization released report in 2022 (World Mental Health Report: Transforming Mental health for All), around 970 million people around the world (~13% of the world's population) were living with mental disorders. This staggering statistic underscores the urgency of addressing mental health challenges on a global scale. The impact of mental health disorders extends beyond individual suffering; it also has serious economic consequences. It was reported that mental health disorders had an economic impact worth approximately over 800 billion USD in 2010, and this number is estimated to double by 2030, placing a significant strain on economies worldwide.

Depression stands out as one of the largest and most dangerous among all mental disorders. A staggering 29% of the global population is affected by this debilitating condition. What makes depression even more concerning is its ability to adversely affect chronic health conditions, such as cardiovascular disease, cancer, diabetes, and obesity. It’s complicated aetiology and chronic clinical features make it difficult for individuals to be consciously aware of their own depressive emotions, often leading to severe consequences, including suicide.

Looking ahead, it is projected that in the coming 20 years, depression will become the leading cause of disability in high-income nations. The World Health Organization (WHO) now ranks major depression as one of the most burdensome diseases in the world, further emphasizing the critical importance of addressing this issue. To mitigate the economic, physical health, and mental health losses associated with depression, it is imperative to develop innovative solutions for early detection and treatment.

Traditionally, depression diagnosis has relied on a psychiatrist's subjective analysis of questionnaires provided to patients. However, depression is characterized by a lack of clear boundaries or specific symptoms, making human intervention essential for accurate diagnosis. The alarming reality in low-and-medium income countries is that there is typically only one psychiatrist available to serve 200,000 or more people. This staggering ratio highlights the limited resources available to tackle mental health disorders in regions where they are most needed. Additionally, mental health services in these areas are often inaccessible or financially out of reach for many individuals.

In response to these challenges, this research project is poised to provide a groundbreaking solution for diagnosing depression without the need for human intervention. By harnessing the potential of big data analytics, we aim to create an automated depression detection system that can revolutionize mental health care worldwide. This innovative approach has the potential to improve early diagnosis, increase access to care, and ultimately alleviate the economic, physical, and mental health burdens associated with depression. In the pages that follow, we will delve deeper into the methodology and potential benefits of this pioneering research endeavour.

**Problem statement:**

Depression is a complex condition, and uncovering its underlying causes can be a challenging endeavour. This complexity arises from the interplay of numerous factors that influence its development. Broadly, these factors can be categorized into two major groups: (a) genomic factors and (b) environmental factors. Genomic factors are responsible for roughly one-third of the causes of depression, while environmental factors account for the remaining two-thirds. This implies that an individual's genetic makeup can lead to varied responses when exposed to situations that trigger depression, a phenomenon known as Gene × Environment interaction.

The World Health Organization (WHO) has identified a wide array of variables that can contribute to depression, including genetic predisposition, level of education, dietary habits, substance use (such as alcohol and drugs), vitamin D levels, sleep patterns, history of sexual abuse, experiences of bullying, belonging to ethnic minority groups, and a total of 28 such factors. Recognizing the multifaceted nature of depression and quantifying the extent to which each factor contributes to an individual's susceptibility to depression is essential. It holds the key to more accurate diagnosis, personalized interventions, and ultimately, improved mental health outcomes.

Notably, previous research in the field, particularly studies involving social media analytics and semantic analysis for depression prediction, have often focused on datasets with substantial volume but limited variety. In contrast, this research project takes a different approach by incorporating a diverse range of variables for analysis.

The core challenge that this research project seeks to address is as follows: to develop a methodology for quantifying a depression score, which serves as an indicator of an individual's susceptibility to depression due to various contributing factors. This approach aims to offer a comprehensive understanding of how these factors interact and impact an individual's mental well-being.

**Objectives:**

The aim of this research is to construct an automated depression score prediction model leveraging big data. It is important to note that our approach distinguishes itself from previous prediction models by the variety of variables we incorporate into our analysis.

To accomplish this goal, we have identified several specific sub-objectives that will guide our research efforts:

*Depression Score Calculation*: We aim to develop a model that can accurately calculate the depression score for an individual, expressed as a percentage. This score will serve as a quantifiable measure of an individual's susceptibility to depression based on a comprehensive assessment of various contributing factors.

*Identification of Significant Factors*: An essential aspect of our research is to pinpoint the factors that exert the most significant influence on an individual's likelihood of experiencing depression. By doing so, we intend to uncover the primary drivers behind this complex condition.

*Percentage-wise Contribution Analysis:* We will delve into the intricate interplay between different factors and their respective contributions to an individual's depression score. This analysis will provide a nuanced understanding of the relative importance of each factor in influencing mental well-being.

The ultimate outcome of this research will be the creation of a predictive model that can determine whether an individual is likely to be experiencing depression. This determination will be made by comparing the calculated depression score to a predefined threshold value. Importantly, the insights gained from our research have practical applications, particularly in regions with limited mental health resources, such as low and medium-income countries. In these settings, our model can play a vital role in aiding the diagnosis of depression, thereby facilitating timely intervention and support for individuals in need.

**Preliminary literature review:**

Depression represents a significant challenge within the realm of mental health, and researchers have previously employed Natural Language Processing (NLP) techniques to develop algorithms for detecting depression by analysing user-generated content, which includes opinions, emotions, and life events. In the field of mental health services, standardized tools are widely utilized for depression screening and assessment. These tools encompass various approaches, including questionnaires, psychological tests, clinical interviews, and even brain scanning. Each technique is briefly elucidated below:

Professional Interviews: This method is administered by trained mental health professionals who conduct structured interviews to assess a person's mental well-being.

Brain Scanning: Brain scanning techniques involve the examination of brain activity, neurons, and neurotransmitters to identify indicators of depression.

Previously, research incorporated sentiment analysis, utilizing vocabulary and predefined rules to ascertain tendencies indicative of depression. This model comprised ten selected features, categorized into three dimensions: micro-blog content, user interactions, and behavioral patterns. Validated through Bayesian analysis, decision trees, rules, ROC Area, and F-measure, this model demonstrated an 80% accuracy rate in detecting depression.

In the Machine Learning Algorithm module, several algorithms were applied. The three most accurate algorithms—Support Vector Machine (SVM), Random Forest, and Deep Learning—were chosen. Relevant features were identified, and a depression classifier model was constructed. During the testing phase, microblogs were processed without labels, and the extracted attributes were used as inputs for the depression classifier model. The resulting prediction label indicated the potential presence of depression in the user.

Following data pre-filtering, words were ranked based on their frequency and term frequency-inverse document frequency values. Several attributes, including emotional processes, temporal trends, and linguistic styles, were presented to characterize depressive behaviours among users.

For future research, the intention is to gather additional data to identify more relevant and valid features. The manual annotation of complex attributes through crowdsourcing and a deeper analysis of dimensions are also planned to enhance the depression detection algorithm. Additionally, alternative techniques for extracting paraphrases from various emotional features will be explored, along with the utilization of larger datasets to validate the efficiency and effectiveness of the techniques.

In the context of our Big Data project analysis, we are considering the implementation of Hadoop (MapReduce) to enhance our research endeavours. Hadoop leverages distributed storage and parallel processing to manage large-scale data and analytics tasks. It subdivides workloads into smaller units for concurrent execution, comprising the following key steps:

1. Splitting: Initial data splitting into manageable segments.

2. Mapping: Transformation of input datasets into intermediate datasets.

3. Sorting and Shuffling: Organization and grouping of intermediate data.

4. Reducing: Processing and reducing intermediate data into smaller units.

Furthermore, in various industries, particularly healthcare, Big Data is instrumental. It supports tasks such as diagnostics, treatment design, depression prediction, and the evaluation of human life quality. The voluminous and intricate data processed by healthcare organizations necessitates the use of Hadoop and MapReduce, enabling the seamless processing of terabytes of data.

**Methodology:**

From a broader perspective, Considering the depression data, we have taken different factors such as id, working status, marital status, divorce, number of children, age, highest year of school completed, sex, family income when 16 yrs, personal income, halting crime rate, happy, health, trust, confidence in education, unemployed in last 10 years, pregnant as the result of rape, divorce laws, pre-marital sex to interpret the depression score of an individual. To find and predict the desired outcome, the data is collected from various sources such as biological, Psychological, environmental, lifestyle, social, medical, and health factors, which are taken according to ethical guidance. As we are exploring the data, there are certain anomalies. To avoid that, we have methods such as encoding categorical variables into numerical variables and normalizing them to ensure that they are in the same scale for analysis as we have different types of attributes, and we find the correlation between variables. Considering the outcome is based on various factors in the dataset. We use ML techniques such as Random Forests, decision trees, and support vector machines to predict depression rates.

**References:**

[1] J. A. Naslund *et al.*, "Digital Innovations for Global Mental Health: Opportunities for Data Science, Task Sharing, and Early Intervention," (in eng), *Curr Treat Options Psychiatry,* vol. 6, no. 4, pp. 337-351, Dec 2019, doi: 10.1007/s40501-019-00186-8.

[2] T. C. Russ *et al.*, "How data science can advance mental health research," (in eng), *Nat Hum Behav,* vol. 3, no. 1, pp. 24-32, Jan 2019, doi: 10.1038/s41562-018-0470-9.

[3] E. Saravia, C. H. Chang, R. J. D. Lorenzo, and Y. S. Chen, "MIDAS: Mental illness detection and analysis via social media," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 18-21 Aug. 2016 2016, pp. 1418-1421, doi: 10.1109/ASONAM.2016.7752434.

[4] R. V. Saveanu and C. B. Nemeroff, "Etiology of depression: genetic and environmental factors," (in eng), *Psychiatr Clin North Am,* vol. 35, no. 1, pp. 51-71, Mar 2012, doi: 10.1016/j.psc.2011.12.001.

[5] C. Sudlow *et al.*, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," (in eng), *PLoS Med,* vol. 12, no. 3, p. e1001779, Mar 2015, doi: 10.1371/journal.pmed.1001779.

[6] S. Yang, P. Zhou, K. Duan, M. S. Hossain, and M. F. Alhamid, "emHealth: Towards Emotion Health Through Depression Prediction and Intelligent Health Recommender System," *Mobile Networks and Applications,* vol. 23, no. 2, pp. 216-226, 2018/04/01 2018, doi: 10.1007/s11036-017-0929-3.