Synergizing Digital Twin Technology for Advanced Depression Categorization in Social-Media through Data Mining Analysis

Sajawal Khan[1], Muhammad Dawood khan[2], Muhammad Awais Ashraf[1], Aamir Munir[1], Muhammad Zunair Zamir[1]

[1] School of Information Engineering, Chang'an University, Xi'an, Shaanxi Province, China
[2] Computer Science at Layyah Campus GCUF, Pakistan
* **Correspondence**: Sajawal khan, sk0013126@gmail.com

The progression from negative emotions to depression is a significant concern, marked by persistent sadness and an inability to cope with challenging circumstances. Regrettably, it can lead to the extreme step of suicide. According to the World Health Organization (WHO), 4.4% of the global population currently grapples with depression. Shockingly, 700,000 individuals worldwide took their own lives in 2023, and this tragic number continues to escalate. Our objective is to detect signs of depression in individuals through their social media posts, SMS, or comments. We collected nearly 10,000 pieces of information from Twitter comments, Facebook posts, and remarks. Employing data mining and machine learning algorithms has proven instrumental in swiftly discerning individuals' emotional states. To predict depression versus non-depression, we employed six classifiers, with support vector machines (SVMs) demonstrating the highest accuracy. A comparison between SVM and Naïve Bayes revealed that SVM yielded superior results in our study.

Keywords: Digital twin technology, Data Mining, Sentiment Analysis, Social Media, Depression, Categorization.

Introduction:

Social media's widespread use in recent years has changed the communication landscape by giving people a digital platform to communicate their ideas, feelings, and experiences. The abundance of data produced by these platforms presents a singular chance to investigate and comprehend many facets of human behavior, including mental health. The identification and classification of depression via data mining analysis of social media material is one important field of research [1]. The most familiar kind of depressive malady is termed Major Depressive Disorder (MDD), which conflicts diversely, plus the ability to function and examine hunger, slumber, and pleasure. It's the best way to diagnose the problems associated with people, their health, interests, likes, and dislikes from social media where everyone shares every movement of life happy, sad, etc. [2]. Social media is one of the best sources for real-time prediction [3]. Depression is classified as a feeling of jumble. It could be defined as emotions of grief, failure, or resentment that conflict with an individual's daily exercises. It is estimated by the Centers for Disease Control and Prevention (CDC) granted authorizations that 8.1% of American grown-ups who are in their twenties or above had depression during the provided two-week span from 2013 to 2016. Persons encounter distress in diverse fashions. It may impede regular tasks ending up in wasted time and lower potency. It may also affect relations. Situations that can become more chronic because of depression comprise. It is normal for someone to feel tired after hard work [4]. It is important to understand that it is essential for health. Everyone has to deal with different problems so, someone becomes sad and upset. The sign of depression occurs when someone constantly feels down or miserable [5].

Classification:

The order of distress is a general matter to hold infantry of psychiatrists debating lovingly for times. This document will concentrate specifically on the ideas of worrying depression and its connection to other classes [6]. we will not deal directly with the discrimination between stress and depression, which is included elsewhere in this assembly. Our area of study is anxiety in the classification of samples that are explained as a qualified model for dealing with depression [7].

Beyond the Blues:

In our lives, we go through such conditions when we feel tired. We assume that there is no way to move forward. If this sensation of melancholy, dullness, and vacuumed characterized by a scarcity of attention continues for longer than a couple of weeks, it indicates that a person is experiencing depression. It should perpetually be retained that depression has diverse modes and every class may manifest various kinds of marks. The best treatment for depression is to manage it by prescribed medications, counseling sittings, and speech therapy. It should always be remembered [8]. Major Depressive Disorder Significant depressive disorder also known as clinical depression, is the highly recognized depression type that strikes the populace in enormous numbers. It's accepted that over 16 million grown-ups worldwide have encountered in any event one scene of a significant burdensome issue at one point in their lives. Finding a significant burdensome problem typically includes a specialist searching for five essential manifestations that affect you're a person`s considerations activities and conduct [7]. Different side effects typically are related to issues are bitterness, absence of excitement in seeking after any sort of action, not having the option to rest, experiencing difficulty deciding, being unable to focus, being exorbitantly lethargic and encountering self-destructive contemplations or activities [9]. Public research on comorbidity shows that common sex contrasts first emerge about the age of 10 years and continue until midlife, after which they disappear. Consequently, during their childbearing years, women have the most significant chance of developing burdensome issues [2]. Symptoms of depression: The main indication is having a depressed condition of consciousness, for a substantial part of the time.

Another indication is mass tumult or hindrance with emotional exhaustion & energy loss [10]. The next inclination of responsibility is futility with learning obsession, imagining, or deciding on a decision to transform into a difficult endeavor. The next indication is difficulty in having sleep or rest extravagantly and the last indication is the main manifestation that doesn't need to exist practically consistently, the indication is considering demise, self-destruction endeavor, or intending to end [11]. In this way, the anxieties increased by day-by-day life occasions that may build the odds of wretchedness. Five underneath manifestations happen consistently that include 1. Pessimistic Condition 2. Lack of curiosity 3. Lethal considerations 4. Thinking of inefficiency or depression   5. Deteriorated capacity to consider and analyze. Maybe different causes like qualities and clan ancestry may prompt depression [12]. The tendency of women to suffer is believed to be correlated with a few natural cycles, including hereditary vulnerability, hormonal vacillations identified with various sections of regenerative potential, and excessive affectability to certain hormonal shifts in frameworks of mind that intercede with burdensome states [8].

Objective:
The objective of this research is to develop a digital twin technology model for the categorization of depression through social media platforms considering age and marital status.

Novelty statement:
All other ways consider the text data which is saved into the database for application of the model to extract actionable information [9] [10]. However, this research considers real-time data to examine the degree of depression that may result in time-saving to rescue the lives of individuals [13].

Material and Methods:
In the first step, we collected data from Twitter, processed it, and classified it through classification algorithms, and finally analyzed data to examine the level of depression.
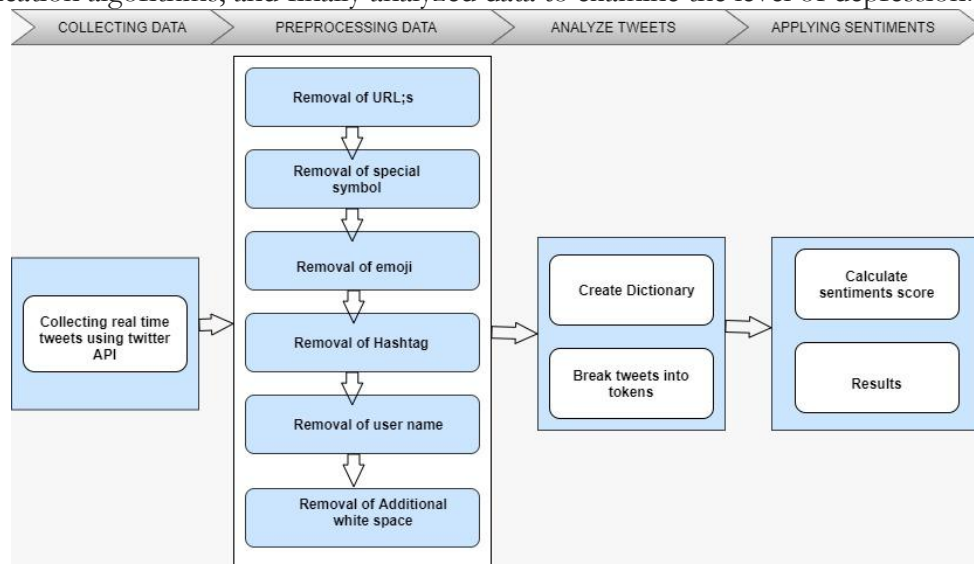
Figure 1: Methodology diagram of this project

Data Collection: We collected data from Twitter accounts of different people of different age groups along with their marital status. Many other attributes were also considered as input parameters including assets, profession, source of income, hobby, profession, and others. Twin technology is used (Get data from the twitter by developer account by twitter provide some security keys through this we can get data for research purpose)

Data Preprocessing:

During preprocessing, the data was made noise-free using the code as under:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1429 entries, 0 to 1428
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Survey_id             1429 non-null   int64
 1   tweet                 1429 non-null   object
 2   Gender                1429 non-null   int64
 3   Age                   1429 non-null   int64
 4   Married               1429 non-null   int64
 5   Number_children       1429 non-null   int64
 6   education_level       1429 non-null   int64
 7   total_members         1429 non-null   int64
 8   gained_asset          1429 non-null   int64
 9   durable_asset         1429 non-null   int64
 10  save_asset            1429 non-null   int64
 11  living_expenses       1429 non-null   int64
 12  other_expenses        1429 non-null   int64
 13  incoming_salary       1429 non-null   int64
 14  incoming_own_farm     1429 non-null   int64
 15  incoming_business     1429 non-null   int64
 16  incoming_no_business  1429 non-null   int64
 17  incoming_agricultural 1429 non-null   int64
 18  farm_expenses         1429 non-null   int64
 19  labor_primary         1429 non-null   int64
 20  lasting_investment    1429 non-null   int64
 21  no_lasting_investmen  1409 non-null   float64
 22  depressed             1429 non-null   int64
 23  category              1429 non-null   object
dtypes: float64(1), int64(21), object(2)
memory usage: 268.1+ KB
```

Figure 2: Data preprocessing image

Data Classification:

SVM:

Each data item was plotted as a point in n-dimensional space in the SVM algorithms where the number of elements is denoted by 'n'. The value of each character is the value of a specific coordinate. Hyper-plane distinguishes the two classes competently, so, we find it to do the classification [14]. The coordinates of particular observances are Support Vectors. A frontier that best differentiates between the two classes including hyper-line/plane is called the SVM classifier.

Naïve Bayes:

The naive Bayes approach is recommended if we are playing with millions of catalogs having some characteristics. When this approach is utilized for textual data exploration, it generates incredible conclusions. It is meaningful to understand the Bayes Theorem for Naive Bayes classifier just like NLP. Conditional probability is the basis of this theorem [15]. The possibility of an incident employing its preliminary information can be evaluated by conditional probability. The formula for calculating the conditional probability is as under:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Naive Bayes is based on Bayes' theorem, which calculates the probability of a hypothesis based on the probabilities of the observed evidence. The "naive" part of Naive Bayes comes from the assumption that the features used to describe an observation are independent of each other, which may not always be true in real-world scenarios. Despite its simplicity and the independence assumption, Naive Bayes often performs well in various classification tasks, such as spam filtering and text categorization. It is particularly suitable for situations with a relatively small amount of data [16].

In Figure 3, attributes that were considered to conduct this research have been mentioned.

|	Survey_id	tweet	Gender	Age	Married	Number_children	education_level	total_members	gained_asset	durable_asset	...	incoming_own_farm	incoming_business	incoming_no_business	incoming_agricultural	farm_expenses	labor_primary	lasting_investment	no_lasting_investmen	depressed	category
0	926	damn taking this personality quiz and realizin...	1	28	1	4	10	5	28912201	22861940	...	0	0	0	30028818	31363432	0	28411718	28292707.0	0	major
1	747	gohan in general needs more love lol will neve...	1	23	1	3	8	5	28912201	22861940	...	0	0	0	30028818	31363432	0	28411718	28292707.0	1	major
2	1190	damn taking this personality quiz and realizin...	1	22	1	3	9	5	28912201	22861940	...	0	0	0	30028818	31363432	0	28411718	28292707.0	0	major
3	1065	my depression is really kicking my ass right n...	1	27	1	2	10	4	52667108	19698904	...	1	0	1	22288055	18751329	0	7781123	69219765.0	0	major
4	806	gohan in general needs more love lol will neve...	0	59	0	4	10	6	82606287	17352654	...	0	0	0	53384566	20731006	1	20100562	43419447.0	0	major

Figure 3: Data set diagram
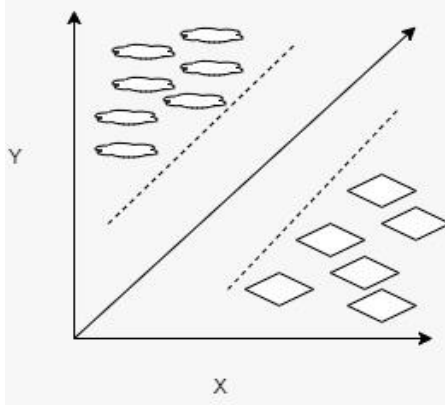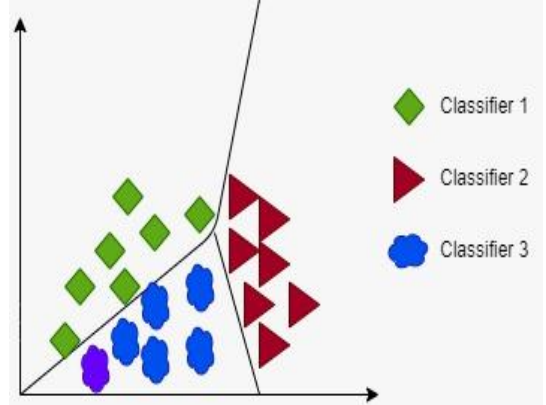
Figure 4: SVM results diagram

Figure 5: Naïve Bayes results diagram

Result and Discussion:

Figures 4 and 5 show data dispersion in form of a plot. Moreover, there are approximately 1200 total posts where 200 are depressed and others are not depressed. Here, 0 shows negative and 1 shows positive.
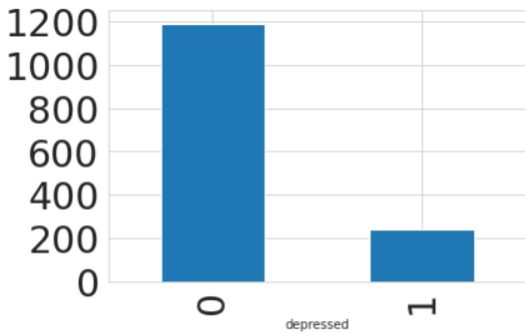

**depressed: [Zero: No depressed] or [One: depressed] 0 = negitive 1= positive

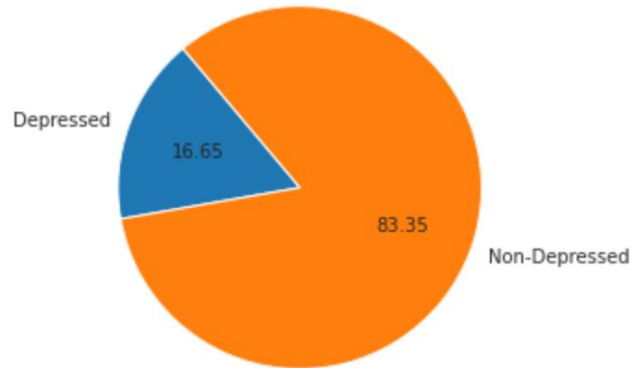**Figure 6:** Positive and Negative depression

Figure 7: Circle diagram how much depression or non-depress

As per the circular diagram, 83.35% were non-depress and 16.65% were depressed. In next various other attributes were also considered including Gender, Age, No of Children, educational level, category, and income of each participant.

Gender plays a significant role, as studies have indicated differing prevalence rates between men and women, with societal expectations and biological factors influencing susceptibility. Age is another crucial factor, as the manifestation and severity of depression can vary across different life stages. The number of children in a participant's life can also impact their mental well-being, with parenting responsibilities and family dynamics influencing stress levels. Educational level is linked to mental health, with higher education often associated with better-coping mechanisms. Additionally, socioeconomic factors such as category and income contribute to depression, as individuals facing economic challenges may experience heightened stressors. Therefore, a holistic understanding of depression necessitates an exploration of these multifaceted parameters to develop targeted interventions and support systems for individuals experiencing depressive symptoms.

The choice between Naive Bayes and Support Vector Machines (SVM) for estimating depression based on comments from social media platforms depends on several factors, including the characteristics of your data, the size of your dataset, and the specific requirements of your task. Both algorithms have their strengths and weaknesses.

Naive Bayes is a simple and computationally efficient algorithm, making it well-suited for text classification tasks. It assumes independence between features, which may not hold true in the case of natural language, but it often performs surprisingly well in practice. Naive

Bayes is particularly useful when dealing with large datasets and can handle high-dimensional feature spaces efficiently.

On the other hand, Support Vector Machines are powerful classifiers that excel in finding complex patterns and nonlinear relationships in data. SVM can capture intricate relationships between words and expressions in text, making it suitable for tasks where feature interactions are crucial. However, SVMs can be computationally intensive, especially with large datasets.

For estimating depression from social media comments, where the language used can be complex and nuanced, and the dataset might be large, SVM might be a good choice due to its ability to capture intricate patterns. However, it's recommended to experiment with both algorithms on your specific dataset and evaluate their performance using metrics like accuracy, precision, recall, and F1-score to determine which one works better for your particular application. Additionally, other advanced techniques like deep learning models, such as recurrent neural networks (RNNs) or transformers, could also be explored for their effectiveness in capturing context and nuances in textual data.

In succeeding parts of Figure 8, light colors show the maximum level and dark colors show the minimum level.
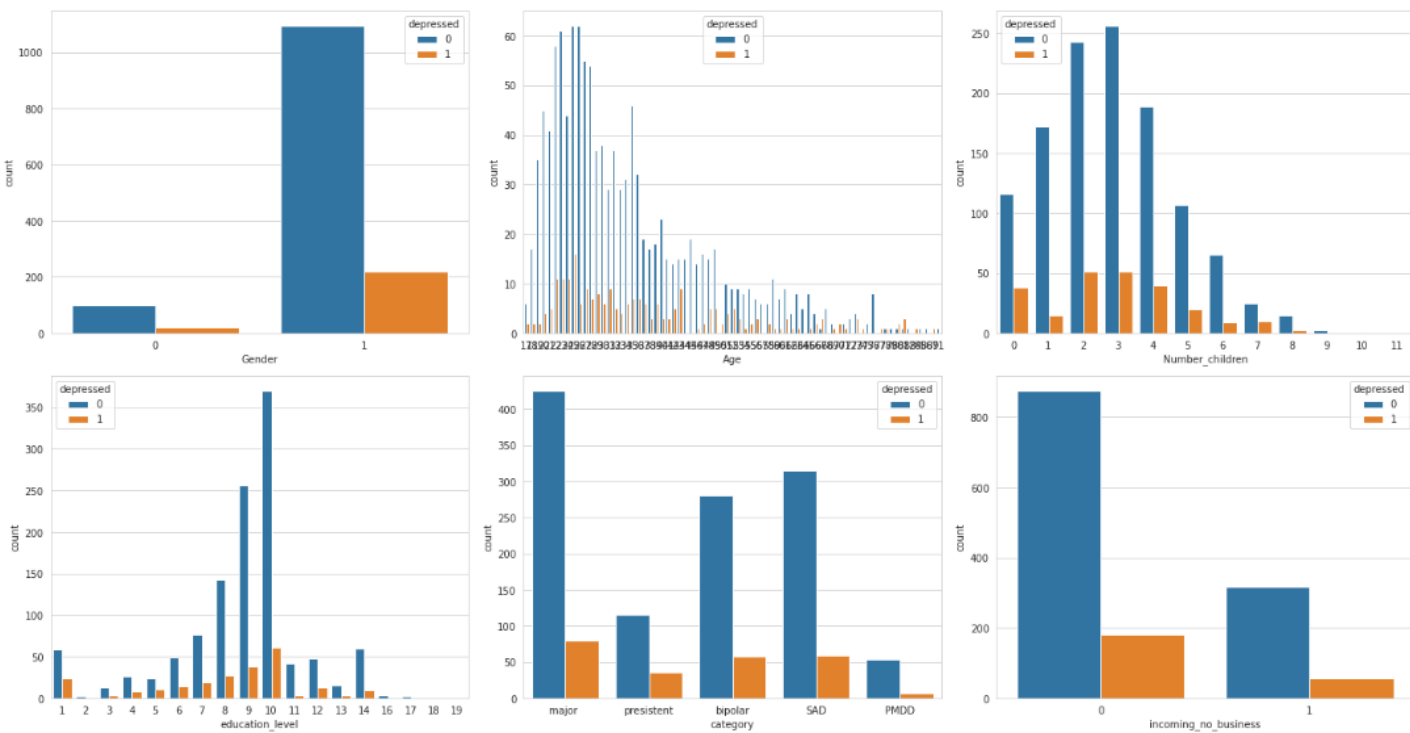
Figure 8: Graphical results of depression on the base of gender, age and different personals attributes

Discussion:

The investigation of depression by social media data mining is an example of how technology, mental health, and ethical issues are dynamically intersecting. The process of gathering data and applying machine learning models has revealed new information and shown the possibility of classifying and comprehending depression in the digital sphere.

Interpretation of Results:

The sentiment and keyword analyzed to train the machine learning models provide promising results in differentiating between social media posts that are symptomatic of depression and those that are not. The precision of the categorization process and dependability highlight data mining potential as a useful tool for mental health research. But

it's important to recognize the inherent difficulties in interpreting findings when dealing with subjective, complex diseases like depression.

Ethical Considerations and User Privacy:

When analyzing mental health discussions on social media, ethical considerations remain paramount. Striking the correct balance between acquiring valuable insights and safeguarding user privacy demands meticulous attention. Future studies should explore more reliable anonymization techniques, and ethical standards for responsible data utilization in digitally conducted mental health research need to be considered.

Potential for Intervention and Support:

While this research has primarily focused on classification and understanding, it also aims to offer responsible interventions. Identifying individuals at risk of depression provides an opportunity for timely and targeted treatments, ranging from the allocation of mental health resources to the establishment of virtual support groups. However, the ethical implications of these interventions must be carefully considered, underscoring the importance of user consent and respect for individual preferences.
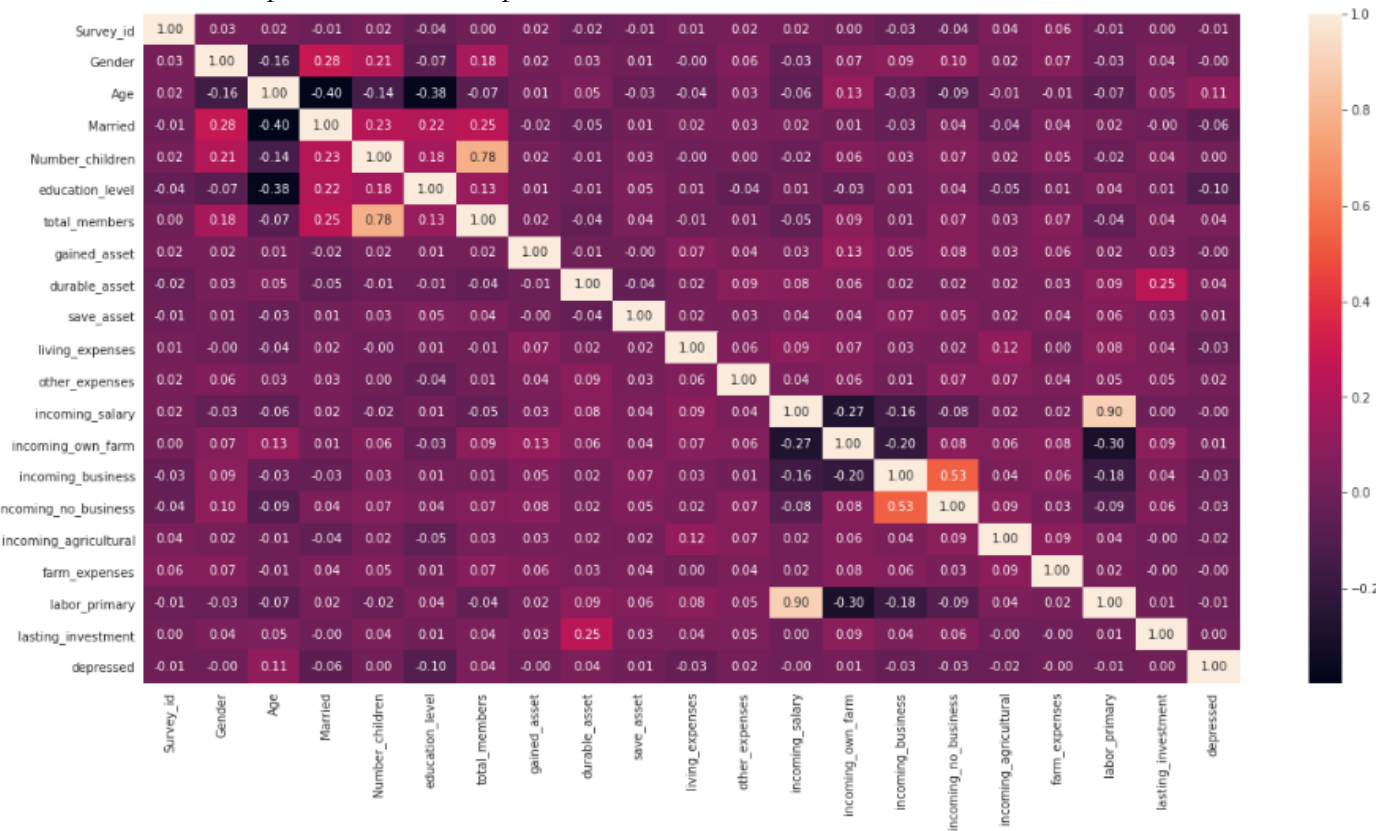

Figure 9: Results in the form of a Heat Map diagram

Conclusion:

In conclusion, the synergistic integration of Digital Twin technology, alongside sophisticated data mining and analysis techniques, holds immense promise for advancing the categorization of depression based on social media interactions. By creating a virtual representation of individuals and their online behaviors, Digital Twin technology offers a dynamic and real-time framework for capturing the evolving nature of mental health expressions. Coupled with powerful data mining algorithms, such as Support Vector Machines and Naive Bayes, this approach enables a nuanced understanding of complex linguistic patterns and contextual nuances within social media content. The fusion of these technologies not only facilitates more accurate and timely depression categorization but also opens avenues

for personalized interventions and targeted mental health support. As we continue to navigate the evolving landscape of digital interactions, the synergistic application of Digital Twin technology and data mining analysis emerges as a transformative strategy with the potential to enhance our comprehension and response to mental health challenges in the digital era.

Limitations and Future Directions:

The models generated may not entirely capture the spectrum of depressive experiences, given their reliance on readily available public data. Future research should explore methods to integrate diverse data sources and consider the nuanced differences in language and expression across various cultures. Additionally, due to the dynamic nature of social media platforms, regular updates to the models are essential to ensure their continued effectiveness in identifying emerging language patterns.

Acknowledgment:

References:

[1] M.-Y. Fan et al., "Acupoints compatibility rules of acupuncture for depression disease based on data mining technology," Zhongguo Zhen Jiu, vol. 43, no. 3, pp. 269–76, 2023, doi: 10.13703/J.0255-2930.20221103-K0001.

[2] D. Xue, Y. Zhang, Z. Song, X. Jie, R. Jia, and A. Zhu, "Integrated meta-analysis, data mining, and animal experiments to investigate the efficacy and potential pharmacological mechanism of a TCM tonic prescription, Jianpi Tongmai formula, in depression," Phytomedicine, vol. 105, p. 154344, Oct. 2022, doi: 10.1016/J.PHYMED.2022.154344.

[3] "Meme Detection of Journalists from Social Media by Using Data Mining Techniques | International Journal of Innovations in Science & Technology." Accessed: Nov. 26, 2023. [Online]. Available: https://journal.50sea.com/index.php/IJIST/article/view/404

[4] F. Alhussain, A. Bin Onayq, D. Ismail, M. Alduayj, T. Alawbathani, and M. Aljaffer, "Adjustment disorder among first year medical students at King Saud University, Riyadh, Saudi Arabia, in 2020," J. Fam. Community Med., vol. 30, no. 1, p. 59, 2023, doi: 10.4103/JFCM.JFCM_227_22.

[5] Ö. Baltacı, "The Predictive Relationships between the Social Media Addiction and Social Anxiety, Loneliness, and Happiness," Int. J. Progress. Educ., vol. 15, no. 4, pp. 73–82, Aug. 2019, doi: 10.29329/IJPE.2019.203.6.

[6] W. Lu et al., "Differences in cognitive functions of atypical and non-atypical depression based on propensity score matching," J. Affect. Disord., vol. 325, pp. 732–738, Mar. 2023, doi: 10.1016/J.JAD.2023.01.071.

[7] M. P. Valerio, J. Lomastro, A. Igoa, and D. J. Martino, "Clinical Characteristics of Melancholic and Nonmelancholic Depressions," J. Nerv. Ment. Dis., vol. 211, no. 3, pp. 248–252, Mar. 2023, doi: 10.1097/NMD.0000000000001616.

[8] F. Tasnim, S. U. Habiba, N. Nafisa, and A. Ahmed, "Depressive Bangla Text Detection from Social Media Post Using Different Data Mining Techniques," Lect. Notes Electr. Eng., vol. 834, pp. 237–247, 2022, doi: 10.1007/978-981-16-8484-5_21/COVER.

[9] M. Maes and A. F. Almulla, "Research and Diagnostic Algorithmic Rules (RADAR) and RADAR Plots for the First Episode of Major Depressive Disorder: Effects of Childhood and Recent Adverse Experiences on Suicidal Behaviors, Neurocognition and Phenome Features," Brain Sci. 2023, Vol. 13, Page 714, vol. 13, no. 5, p. 714, Apr.

2023, doi: 10.3390/BRAINSCI13050714.

[10] C. S. Wu, C. J. Kuo, C. H. Su, S. H. Wang, and H. J. Dai, "Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records," J. Affect. Disord., vol. 260, pp. 617–623, Jan. 2020, doi: 10.1016/J.JAD.2019.09.044.

[11] R. Vanlalawmpuia and M. Lalhmingliana, "Prediction of Depression in Social Network Sites Using Data Mining," Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2020, pp. 489–495, May 2020, doi: 10.1109/ICICCS48265.2020.9120899.

[12] C. K. Ettman, S. M. Abdalla, G. H. Cohen, L. Sampson, P. M. Vivier, and S. Galea, "Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic," JAMA Netw. Open, vol. 3, no. 9, pp. e2019686–e2019686, Sep. 2020, doi: 10.1001/JAMANETWORKOPEN.2020.19686.

[13] L. Smith et al., "Association between depression and subjective cognitive complaints in 47 low- and middle-income countries," J. Psychiatr. Res., vol. 154, pp. 28–34, Oct. 2022, doi: 10.1016/J.JPSYCHIRES.2022.07.021.

[14] S. L. Dubovsky, B. M. Ghosh, J. C. Serotte, and V. Cranwell, "Psychotic Depression: Diagnosis, Differential Diagnosis, and Treatment," Psychother. Psychosom., vol. 90, no. 3, pp. 160–177, Apr. 2021, doi: 10.1159/000511348.

[15] Ö. DEMİRCİ and E. INAN, "Postpartum Paternal Depression: Its Impact on Family and Child Development," Curr. Approaches Psychiatry, vol. 15, no. 3, pp. 498–507, Sep. 2023, doi: 10.18863/PGY.1153712.

[16] L. Sforzini et al., "A Delphi-method-based consensus guideline for definition of treatment-resistant depression for clinical trials," Mol. Psychiatry 2021 273, vol. 27, no. 3, pp. 1286–1299, Dec. 2021, doi: 10.1038/s41380-021-01381-x.