Unveiling the Digital Mind: Leveraging Social Media Data for Detecting Depression and Anxiety

1st Dina Khalid
Systems and Biomedical Engineering
Cairo university
Cairo, Egypt
dina.salama00@eng-st.cu.edu.eg

2nd Alaa yasser
Systems and Biomedical Engineering
Cairo university
Cairo, Egypt
alaa.hameed01@eng-st.cu.edu.eg

3rd Omnia Sayed
Systems and Biomedical Engineering
Cairo university
Cairo, Egypt
omnia.hassaunien99@eng-st.cu.edu.eg

4th Mohammad Nasser
Systems and Biomedical Engineering
Cairo university
Cairo, Egypt
mohamed.mohamed0116@eng-st.cu.edu.eg

***Abstract*—This paper investigates the classification of mental health states, including normal, anxiety, and depression, by combining two distinct datasets: one focused on distinguishing between normal and depressed individuals, and the other on discriminating between anxiety and depressed individuals. Leveraging machine learning algorithms, we achieve high testing accuracies of approximately 95% on both datasets. By integrating the results from these individual classifications, we obtain a final classification model capable of accurately categorizing individuals into normal, anxiety, or depressed states. Further investigations can explore additional datasets and refine the classification model for enhanced accuracy and broader applicability.**

***Index Terms*—Machine learning, mental health, classification, social media, depression, anxiety, dataset integration, natural language processing, text analysis.**

I. Introduction

This paper examines the potential of utilizing social media data for detecting and addressing the prevalent issues of depression and anxiety. The exponential growth of user-generated content on social media platforms offers an unprecedented opportunity to leverage this data for proactive mental health support. Through the application of machine learning and natural language processing techniques, we seek to identify indicators of depression and anxiety within these vast amounts of user-generated content. The outcome of this research contributes to enhancing mental health interventions and fostering greater awareness of these widespread conditions in the digital era. The input for this study is social media data, specifically user-generated content from platforms such as Twitter, Facebook, or Reddit. The output is the identification of indicators or patterns that signify depression and anxiety within the analyzed social media content.

II. Related Work

This section presents a literature review on classifying mental health states using social media data, categorizing and analyzing relevant papers to highlight the state-of-the-art and compare them to our work.

A. Social Media Use and Depression and Anxiety Symptoms: A Cluster Analysis (Shensa et al., 2018) [1]

This study identified patterns of social media use (SMU) and examined their association with depression and anxiety symptoms. Using cluster analysis on a representative sample of US adults, five SMU patterns were identified. The "Wired" and "Connected" patterns were found to increase the odds of elevated depression and anxiety symptoms. The study's strengths include a large sample size and the identification of distinct SMU patterns. However, limitations include its cross-sectional design and reliance on self-reported measures. Compared to our work, which focuses on machine learning-based mental health classification using social media data, this study contributes insights into specific SMU patterns and their relationship with mental health outcomes.

B. Machine learning models to detect anxiety and depression through social media: A scoping review (Arfan Ahmed et al., 2022) [2]

The reviewed studies explored machine learning models to detect anxiety and depression through social media data, with a focus on language patterns and online activity. The majority of the studies were conducted during the peak of the COVID-19 pandemic. The strengths of these studies include the use of diverse social media platforms and the potential for early detection of symptoms. Similar to our work, they highlight the potential of machine learning in complementing traditional screening methods. The state-of-the-art involves leveraging social media data and AI technology to develop predictive models for mental health detection, particularly during challenging times such as the COVID-19 pandemic.

C. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study (Qijin Cheng et al., 2017) [3]

The study aims to assess suicide risk and emotional distress through computerized language analysis of Chinese social

media (Weibo). Participants' Weibo posts were analyzed using Simplified Chinese-Linguistic Inquiry and Word Count (SC-LIWC) categories. The associations between language features and suicide risk factors were examined, and a support vector machine (SVM) model was applied for classification. The strengths of this study include the use of Chinese social media data and the identification of unique language markers. However, the machine classifiers' performance needs further optimization. Compared to our work, this study focuses specifically on suicide risk and emotional distress assessment in Chinese social media. It highlights the potential of language analysis methods in identifying risk factors and offers new hypotheses for future research. The state-of-the-art involves leveraging computerized language analysis techniques to automatically assess suicide risk and emotional distress, aiming for early intervention and prevention in natural settings.

D. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts (Raymond Chiong et al.,2021) [4]

This study focuses on using machine learning to detect signs of depression in social media users by analyzing their posts, even when specific keywords are not used. They employ various text preprocessing and textual-based featuring methods along with machine learning classifiers. The study investigates the performance of the proposed approach on different social media platforms. Similar to our work, they leverage machine learning and text analysis for depression detection. The state-of-the-art involves using social media data and machine learning techniques to identify depression indicators, even without explicit keywords.

E. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media (Stevie Chancellor et al., 2019 [5]

This paper highlights potential risks and ethical considerations related to dehumanization in HCML. It emphasizes the importance of incorporating human insights and expertise into data-driven predictions while ensuring ethical practices. Compared to our work, which focuses on machine learning models for detecting mental health conditions using social media data, the examined paper provides a critical and qualitative perspective on the representation of human subjects in HCML research and addresses ethical considerations in the field.

III. Dataset and Features

A common practice is to use a test size of 20-30%, which means that 70-80% of the data is used for training the model. However, this is not a hard and fast rule and the test size may vary depending on the specific problem and dataset and in our problem we found that 20% test size is better as our data is not large.

Preprocessing:

The preprocessing is divided into some steps:

1. The first step is to remove punctuation from the input text

2. Next, the text is converted to lowercase, By converting all text to lowercase, we can ensure that words with the same meaning are treated as the same entity, regardless of their case. This can help to reduce the number of unique words in the dataset and make it easier to identify patterns and relationships between words.

3. The third step is to remove stop words from the text, by removing common words that are unlikely to be useful in predicting depression. These words are often referred to as "stop words" and include words such as "the", "a", "an", "and", "of", and "in". Removing stop words can help reduce the dimensionality of the data and improve the accuracy of the model.

4. The fourth step is to remove frequent words from the text, by removing frequently occurring words that may not be useful in distinguishing between depressed and non-depressed text. This function can be used to remove words that occur in a large proportion of the text, such as words like "I", "you", and "me".

5. The fifth step is to remove rare words from the text, by removing words that occur in only a few instances in the text and are not likely to be useful in predicting depression.

6. Finally, the text is stemmed. This is done by removing suffixes and prefixes from words, which can help reduce the dimensionality of the data and improve the accuracy of the model. For example, the stem of the words "running", "runner", and "runners" is "run".

We tried data augmentation but the result was accuracy decreased as the same text has different class in each dataset.

Example of Dataset

text	label
Gr gr dreaming of ex crush to be my game, God	0.0
wkwkwk what a joke	0.0
Leaves are also standby in front of the PC ... because the office is no longer on leave	0.0
Thank God even though it's just a ride through	0.0
wedding teaser concept using the song day6 - only, sounds good ga siiih	0.0

Fig. 1. Undepressed data set

text	label
I've shifted my focus to something else but I'm still worried	1.0
Have you ever felt nervous but didn't know why?	1.0
Sometimes it's your own thoughts that make you anxious and afraid to close your eyes until you don't sleep	1.0
Every time I wake up, I'm definitely nervous and excited, until when are you going to try™,	1.0

Fig. 2. Depressed data set

IV. Methods

The support vector classifier (SVC) model with a radial basis function (RBF) kernel was used to classify the text data into categories of depressed or not depressed. SVC finds the optimal hyperplane that maximizes the margin between the two categories. The RBF kernel maps the text data into a higher dimensional space, allowing the SVC to fit more complex shapes.

$$\phi(\mathbf{x}, \mathbf{c}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{c}||^2}{2\sigma^2}\right)$$

In this equation, $\phi(\mathbf{x}, \mathbf{c})$ represents the transformed feature vector of the input vector $\mathbf{x}$ using the RBF kernel, with $\mathbf{c}$ representing the center of the kernel and σ representing the width of the kernel. The RBF kernel is a popular choice for kernel methods in machine learning, as it can capture nonlinear relationships between the input features and can effectively handle high-dimensional data.

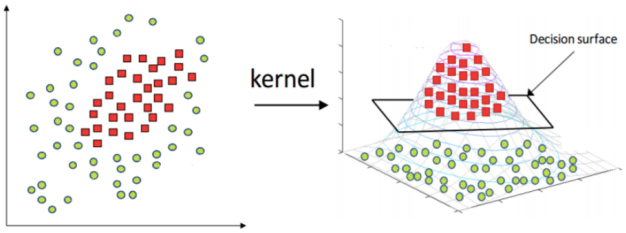

Fig. 3. Depressed data set

The SVC model was trained on TF-IDF features extracted from the text data. TF-IDF stands for term frequency-inverse document frequency, and it weights terms based on how frequently they appear in a document compared to the full corpus. This gives more weight to terms that are important for a specific document. The TF-IDF features were calculated as:

TF-IDF(t,d) = TF(t,d) * IDF(t)

Where TF(t,d) is the frequency of term t in document d, and IDF(t) is the inverse of the percentage of documents that contain term t. The TF-IDF vectors for each document were used as input to train the SVM.

The objective of the SVC is to maximize the margin between the decision boundary and the instances from each class. It minimizes the cost function: Given the following hypothesis and cost function

$$h_\theta(x) = \begin{cases} 1 & \text{if } \theta^T x \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

$$J(\theta) = C \sum_{i=1}^{m} \left[y^{(i)} \cdot cost_1(\theta^T f^{(i)}) + (1 - y^{(i)}) \cdot cost_0(\theta^T f^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

where C is a hyperparameter that controls the trade-off between maximizing the margin and minimizing the classification error, $y^{(i)}$ is the true label of the i-th training example, $f^{(i)}$ is the feature vector of the i-th training example transformed by the RBF kernel, n is the number of features, and $cost_1$ and $cost_0$ are the cost functions for positive and negative examples, respectively. The second term is a regularization term that helps prevent overfitting.

The methodology utilizes the powerful machine learning technique of support vector machines to classify the complex data of natural language. The use of TF-IDF to engineer features from raw text data allows the model to achieve high accuracy and capture the semantic relationships in language.

The cost function of a naive Bayes classifier is the negative log-likelihood of the parameters given the data, which can be written as:

$$L(\theta) = -\sum_{i=1}^{n} \log P(y_i|x_i, \theta) - \sum_{j=1}^{k} \log P(\theta_j)$$

where n is the number of documents, k is the number of parameters, x_i is the feature vector of the i-th document, y_i is the class label of the i-th document (depressed or not), θ is the vector of parameters, and θ_j is the j-th parameter.

For a multinomial naive Bayes classifier, the parameters are the prior probabilities of each class ($P(y)$) and the conditional probabilities of each word given each class ($P(x_w|y)$), where w is an index of the vocabulary. The cost function can be simplified as:

$$L(\theta) = -\sum_{i=1}^{n} \log P(y_i) - \sum_{i=1}^{n} \sum_{w=1}^{V} x_{iw} \log P(x_w|y_i) - \sum_{j=1}^{k} \log P(\theta_j)$$

where V is the size of the vocabulary and x_{iw} is the count or frequency of word w in document i.

To minimize the cost function, one can use maximum likelihood estimation or maximum a posteriori estimation to find the optimal values of θ. For maximum likelihood estimation, there are closed-form solutions for θ, which are:

$$P(y) = \frac{N_y}{N}$$

where N_y is the number of documents with class label y and N is the total number of documents, and

$$P(x_w|y) = \frac{N_{wy} + 1}{N_y + V}$$

where N_{wy} is the total count of word w in documents with class label y. The term $+1$ in the numerator and $+V$ in the denominator are added for smoothing, to avoid zero probabilities.

In summary, we outline the methodology employed in our project to achieve the desired outcomes, for the main steps involved in our approach. Please refer to Figure 4 for the block diagram.

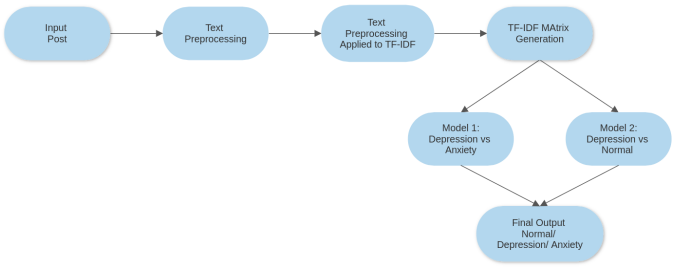

Fig. 4. Block Diagram

V. Experiments

A) Development of Sequential Models: In our study, we initially devised two separate models to address the task at hand. The first model aimed to classify whether a given text was indicative of a normal state or not, while the second model focused on distinguishing between anxiety and depression. To extract meaningful features from the text data, we employed two distinct approaches: word embedding utilizing GloVe embeddings and TF-IDF feature extraction. Our experimentation revealed that the TF-IDF approach yielded superior results compared to word embeddings.

B) Dataset Merging Experiment: To explore the potential benefits of data augmentation and the feasibility of employing a single unified model, we conducted an experiment involving the merging of the two datasets. By combining the datasets, we sought to create a comprehensive dataset encompassing three classes: normal, anxiety, and depression. However, upon training the merged dataset on a single model, we observed a notable decline in classification accuracy by approximately 20%.

This decline in accuracy can be attributed to certain instances where the classification labels differed between the two datasets. Instances that were classified as normal in one dataset were classified as anxiety or depression in the other dataset. As a result, merging the datasets introduced inconsistencies that impeded the model's ability to accurately classify instances, leading to a decrease in overall performance.

In light of these findings, it is evident that careful consideration must be given to the quality and consistency of the datasets when merging multiple sources of data. It is crucial to ensure the harmonization of classification criteria and undertake meticulous data preprocessing to maintain the reliability and effectiveness of the classification model.

VI. Results

We experimented with various machine learning models, including Naive Bayes, Random Forest, Gradient Boosting, Decision Tree, Bagging Classifier, and Support Vector Classifier (SVC), to determine the best approach for our task of detecting depression and anxiety in social media data. After evaluating their performance, we found that the SVC model achieved the highest accuracy among them. Therefore, we selected SVC as our model of choice for detecting depression and anxiety in social media data.

The accuracy results for the different models and embedding methods are presented in Table I for embedding and Table II for TFIDF.

TABLE I
Accuracy Results for Embedding

Model	Accuracy
RandomForestClassifier	0.867
GradientBoostingClassifier	0.872
DecisionTreeClassifier	0.764
BaggingClassifier	0.836
SVC	0.882

TABLE II
Accuracy Results for TFIDF

Model	Accuracy
MultinomialNB	0.811
RandomForestClassifier	0.926
GradientBoostingClassifier	0.924
BaggingClassifier	0.882
SVC	0.960

Based on these results, we determined that the SVC model with TFIDF achieved the highest accuracy and, therefore, selected it as the best model for our task.

To further assess the performance of the SVC model, we generated confusion matrices for each model with the final preprocessing. and evaluated their results. The confusion matrix for the SVC model is presented see Figure 5.

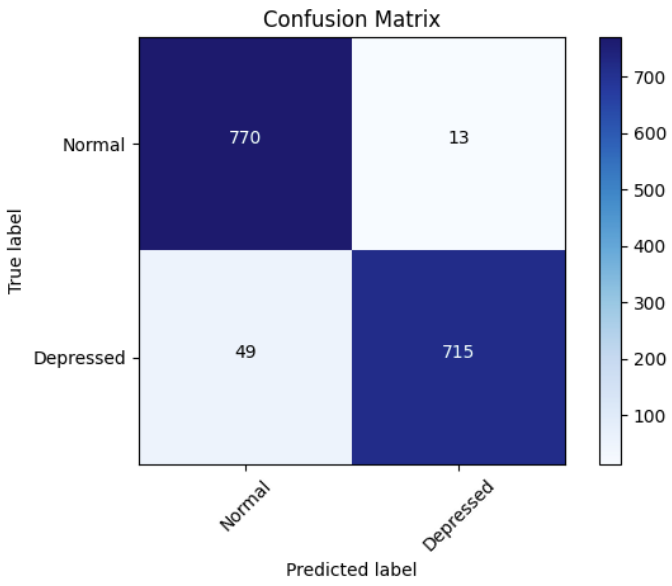

Fig. 5. Confussion matrix for SVC

The confusion matrix reveals that the SVC model achieved 770 true positive predictions, 715 true negative predictions, 13 false positive predictions, and 49 false negative predictions. These results demonstrate the high performance of the SVC model in accurately classifying instances of depression and anxiety in social media data.

In addition to achieving the highest accuracy, the SVC model's superior performance in the confusion matrix further confirms its effectiveness in detecting depression and anxiety. The model demonstrates a high ability to correctly identify positive cases (true positives) and negative cases (true negatives), while minimizing false positive and false negative predictions.

Utilizing both accuracy results and the confusion matrix, we can confidently conclude that the SVC model is the optimal choice for detecting depression and anxiety in social media data, providing accurate and reliable predictions.

To further optimize the performance of the SVC model, we utilized grid search to determine the optimal hyperparameters. Grid search involved systematically evaluating various combinations of hyperparameters, such as kernel type, regularization parameter (C), and gamma value. The following parameter grid was used:

```
param_grid = {
'C': [0.1, 1, 10],'gamma': [0.1, 1, 10],
'kernel': ['linear', 'rbf']}
```

We compared the results obtained using the chosen hyperparameters and the default hyperparameters. However, the default hyperparameters provided the best performance, so we used them for our SVC model. The default hyperparameters were as follows:

```
{'C': 1.0, 'gamma':'scale', 'kernel':'rbf'}
```

VII. Code Notebook

The code implementation for this research project can be accessed at the following URL: https://colab.research.google.com/drive/1y9zgmw8p4c6RlSGlnn4epnd2QTN-aSQE?usp=sharing The notebook includes the code for data preprocessing, feature extraction, model training, and evaluation. It also provides instructions for replicating the experiments and reproducing the results. Feel free to explore the code and adapt it for your own research purposes.

Please note that the notebook is hosted on an external platform and may require specific dependencies or software versions to run successfully.

VIII. Conclusion

This study explored the utilization of social media data to detect indicators of depression and anxiety using machine learning and natural language processing techniques. Among the models tested, the Support Vector Classifier (SVC) performed the best, achieving an accuracy of approximately 0.96. The SVC model effectively classified social media content as indicative or non-indicative of depression and anxiety. The confusion matrix demonstrated the model's ability to accurately distinguish between positive and negative cases. These findings highlight the potential of using social media data and machine learning for proactive mental health support and increasing awareness of mental health conditions in the digital era. Further research is needed to optimize the model and evaluate its performance on larger datasets. Overall, leveraging machine learning and natural language processing can contribute to improving mental well-being in the digital age.

IX. Future Work

In terms of future work, if provided with additional time, team members, and computational resources, we aim to gather our own dataset specifically in Arabic. This would involve creating an Arabic platform where users can freely share their feelings and emotions, allowing us to collect data that is culturally and linguistically relevant to Arabic-speaking users.

To enrich the platform's functionality, we plan to incorporate various multimedia elements, including photos, audio recordings, and videos. These additional mediums of expression would empower users to communicate their emotions more effectively, leading to more accurate assessments and personalized feedback.

In addition to providing feedback, our future vision includes offering personalized tips and guidance based on users' expressed emotions. These recommendations could encompass coping mechanisms, self-care strategies, or exercises aimed at enhancing mental well-being. Furthermore, for users requiring professional assistance, we envision establishing connections with mental health practitioners or experts, enabling access to appropriate resources and support.

In this future iteration, each user post would serve as a therapeutic session, generating valuable feedback and insights. By engaging in this iterative process, users would foster self-awareness, make informed decisions about their mental well-being, and embark on a journey of personal growth and self-improvement.

By conducting our own data collection in Arabic and building the platform around it, we aim to create a valuable tool that empowers individuals in the Arabic-speaking community to actively engage in their mental well-being. This future work holds the potential to cultivate a supportive community and contribute to personal development on a broader scale.

References

[1] A. Shensa, J. E. Sidani, M. A. Dew, C. G. Escobar-Viera, and B. A. Primack, "Social Media Use and Depression and Anxiety Symptoms: A Cluster Analysis," *Am. J. Health Behav.*, vol. 42, no. 2, pp. 116-128, Mar. 2018.

[2] A. Ahmed, S. Aziz, C. T. Toro, M. Alzubaidi, S. Irshaidat, H. Abu Serhan, A. A. Abd-alrazaq, and M. Househ, "Machine learning models to detect anxiety and depression through social media: A scoping review," *Comput. Methods Programs Biomed.*, vol. 2, p. 100066, 2022.

[3] Q. Cheng, T. Li, C. Kwok, T. Zhu, and P. Yip, "Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study," *J. Med. Internet Res.*, vol. 19, no. 7, p. e243, Jul. 2017.

[4] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput. Biol. Med.*, vol. 135, p. 104499, Aug. 2021.

[5] S. Chancellor, E. P. S. Baumer, and M. De Choudhury, "Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, issue CSCW, article no. 147, pp. 1-32, Nov. 2019.

[6] Scikit-learn."sklearn.naive_bayes.GaussianNB", June. 2023.

[7] BaggingClassifier: Leo Breiman. Bagging predictors. Machine Learning, 24(2):123-140, 1996.

[8] RandomForestClassifier: Leo Breiman. Random forests. Machine Learning, 45(1):5-32, 2001.

[9] GradientBoostingClassifier: Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5):1189-1232, 2001.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, J.-M. Thirion, O. Grisel, et al. "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12.1 (2011): 2825-2830.