Mental Health Risk Detection Using Linguistic Analysis with SAFM-Net

Ark Saraf
Dept. of Computing Technologies,
SRM Institute of Science and Technology,
Chennai, India.
as8227@srmist.edu.in

Kartik Gawade
Dept. of Computing Technologies,
SRM Institute of Science and Technology,
Chennai, India.
kg4409@srmist.edu.in

Dr. U.V Anbazhagu
Dept. of Computing Technologies,
SRM Institute of Science and Technology,
Chennai, India.
anbazhau@srmist.edu.in

Abstract*— Depression, and mental health issues in general, are a big problem for people everywhere, affecting how they feel, how well they work, and their lives overall. The way we usually diagnose these things is by a doctor or therapist evaluating the person, and this can be based on opinion, take a long time, and isn't available to many. Because of how quickly we all communicate digitally, the text we write has become a useful way to understand someone's mental state. In this research, we're presenting SAFM-Net (Sentiment-Augmented Fusion Model with Multi-Granular Attention) which is a new, in-depth learning system for figuring out how at risk someone is for mental health problems from their writing. It puts together the meaning of words in context (from a "transformer") with aspects of language that psychologists find important for understanding depression. This lets it identify both what the writing means, and the ways of thinking connected with depression. It uses something called 'cross-attention' to connect the language aspects to each individual bit of text, and then a 'gated fusion' to blend these details. On top of that, it's built with two parts working at once to both say if someone has depression (yes or no) and to put them into a risk category: low, medium, or high. To help the predictions be more consistent, a simple set of rules (a 'stability layer') is used to deal with the murkiness of the middle ranges. The system is trained with a combined method of working out errors, using focal loss, ordinal binary cross-entropy, and Brier score to be as accurate and well-tuned as possible. The experiments show it's 97.40% correct and has a Brier score of 0.04595, proving it works well, is something you can count on, and could be used to assess many people's mental health.

Keywords: Finding mental health issues, categorizing depression, transformer models, cross-attention fusion, language aspects, ranking, SAFM-Net, figuring out levels of risk

I. INTRODUCTION

Lots of people all over the world are struggling with their mental health, and depression is a main reason why people are disabled.[1] What's more, many with these problems don't get help when they need it. This is because of things like people's judgements of them, not understanding enough about mental health, and difficulties finding healthcare. Because of this difference between people who need help and people getting it, we really need ways to find out when someone is starting to have problems, and we need those ways to work for lots of people and be able to be used easily.

People are sharing what they think, how they feel and what's happening to them in writing, more and more as digital communication has zoomed ahead. This writing frequently has little hints in the way the language is used that show what's going on in someone's mind. Because of this, Natural Language Processing now looks like a really good way to understand mental health from text. The first studies in this area used older machine learning with features someone had to specifically design, things like simply listing words, how often a word appears versus how common it is in all documents (TF-IDF), and positive or negative values for words.[2] Although they gave us a first understanding of the topic, these methods couldn't really get at the meaning within the context of the writing or more complicated language structures.

Transformers, and especially models like BERT and DistilBERT,[3][4] have become much better at understanding how words relate to each other in a sentence. They do this with something called self-attention, which lets them see connections between words even if those words are far apart in a longer piece of writing; this leads to a more thorough grasp of language.[5] But, only using transformers doesn't usually pick up on the specific ways people use language in certain areas of life, and for looking at mental health, this is a problem. Research in psychology tells us that depressed people tend to say things in particular ways: they use "I", "me", and "my" more often, they're more likely to use words that are negative, or very all-or-nothing, and they negate things a lot. Standard transformers don't really 'notice' these kinds of language features.

People are now using both the detail of deep learning with a deliberate selection of which parts of language to pay attention to, because of the difficulties mentioned. However, a lot of systems currently just indicate if a person is at risk or if they aren't, and this doesn't communicate the degree of that risk. Mental health as we experience isn't a straightforward yes or no, it exists on a spectrum. Understanding someone's precise position on that spectrum leads to a considerably improved, more helpful assessment of how much danger they are in.

SAFM-Net is our new system, and it combines the way transformers grasp the meaning of words in their surrounding text with how the human brain handles language. They work together by focusing on each other. SAFM-Net improves on previous systems by ranking how severe a threat is, from lowest to highest. And to keep its predictions stable, reasonable and understandable, particularly in ambiguous situations, we've included a section based on a defined set of rules.

This paper's main achievements are these. We've created a new way of using 'cross-attention' to let language characteristics and the surrounding meaning of something work together. We also present a design with two parts: one to decide yes or no (binary classification), and another to measure the degree of risk. Plus, we've added a 'stability layer' to make predictions more similar to what you'd get in real life. And after lots of testing, we've shown our method is good at being accurate and giving likely probabilities.

II. RELATED WORKS

Lots of research has been going on to find mental health problems by looking at people's writing, and this is thanks to improvements in how computers process and learn from language. At first, people mainly used statistics and machine learning. They'd look at things in the text like how often words appear, whether the writing is positive or negative, and how many different kinds of words are used, and then use these to 'teach' classifiers - Support Vector Machines, Naïve Bayes, and

Logistic Regression are examples - to identify issues. These earlier ways of analysing text for mental health gave us a starting point, but they couldn't really grasp the full meaning of what was being said, or the complicated ways people use language.[6]

In deep learning, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks[7] were developed to handle text sequences by modelling their temporal relationships. However, the vanishing gradients in RNNs and inefficiency in handling longer input sequences became major drawbacks. Transformers are a family of architectures that address these issues, by enabling parallelization of computation and by employing self-attention to be able to capture global context information.[8]

Variant of BERT, RoBERTa and DistilBERT have recently reached new state-of-the-art results on a variety of natural language processing tasks including sentiment and text classification. These models first get deep language understanding by training them on a huge corpus of text, thereafter, fine-tuned on a specific task. For achieving high results on such tasks, they however lack a crucial component: domain-specific knowledge. There are interesting exceptions, like models for mental health, though.[9]

As deep learning models have reached high accuracy for mental health detection, a lot of research has been conducted on incorporating additional linguistic and psychological features into the input of neural networks. For the extraction of psychological features a variety of tools have been employed, with VADER, a rule-based analytical tool for sentiment,[10] a prevalent choice due to its interpretable results and widely researched functionality. Previous studies have also found a subset of linguistic features that correlate with states of depression. Here we expand on this subset, particularly focusing on function words and absolutist language.[11]

In the post-processing section another important area of research that was discussed at the workshop was ordinal classification. While many classification problems are treated as binary for simplicity, in many medical and psychological domains the outcome has an intrinsic order. So, even going beyond the simple binary classification to learning an ordered classification is very powerful. However, there is much work to be done in incorporating ordinal models into transformer-based architecture.[12]

The works presented in this book also address the problem of probability calibration in classification problems. Model accuracy is not enough, and the ability to evaluate the quality of the predicted probabilities is crucial. Probability calibration is especially important in critical applications, such as mental health diagnosis for intervention, where well-calibrated probabilities to the uncertainty of the model outputs are crucial for good decision making.[13]

The suggested SAFM-Net framework capitalizes on all these advances by integrating transformer contextual embedding, linguistic feature extraction, cross attention, ordinal classification, and probability calibration to design an end-to-end system.[14] The design takes care of deficiencies associated with the state-of-the-art models.

III. PROPOSED METHODOLOGY

SAFM-Net, the system we're suggesting, is built from a combination of approaches. It brings together understanding of what someone is saying in its context with the way language is known to relate to how our minds work, and this is to more reliably spot potential mental health issues. The entire system works in stages: first it gathers important information (feature extraction), then combines things with a focus on what's important (attention-based fusion), after that it learns how to represent information with a sort of control over what's included (gated representation learning).

A. System Overview

Our system does not require any structured input and simply accepts free form text, and for each input, it generates a probability score corresponding to the risk it represents and classifies the risk into Low, Medium or High. Our system has been implemented as a two-phase pipeline, where the first phase, the Training Pipeline, deals with data preprocessing, feature engineering and training of the deep learning model, and the second phase, the Inference Pipeline, is responsible for production ready deployment of the trained model. A high-level overview of the entire system is shown in the system diagram, and it is also deployed as a web interface using a flask server.

B. SAFM-Net Architecture

SAFM-Net is equipped with five components that work in a synergistic manner to generate predictions. The first two components, namely Transformer Encoder and Linguistic Feature Extractor, process input text and contextualize it and extract linguistic features corresponding to domain of risk assessment respectively. The Cross-Attention Fusion Module is then employed to fuse the contextualized representations of the two components by allowing them to attend to each other. The Gated Feature Integration module is utilized to regulate the degree of contribution of features from each component. The Dual Prediction Heads utilize the fused features to generate two types of outputs, namely continuous probability score and discrete tier classification.

1) Transformer Encoder

Given an input text sequence:

$$X = \{x_1, x_2, x_3, \ldots, x_T\}$$

the DistilBERT encoder produces contextual embedding:

$$H = \text{DistilBERT}(X), H \in \mathbb{R}^{T \times d}$$

where:

- T= sequence length
- $d = 768$(hidden size)

The pooled representation is obtained from the CLS token:

$$h_{cls} = H[0]$$

This vector encodes global semantic meaning of the input.

2) Linguistic Feature Extraction

In addition to deep embeddings, SAFM-Net extracts psychologically relevant linguistic features:

$$L = [f_0, f_1, \ldots, f_7] \in \mathbb{R}^8$$

The features are defined across 8 dimensions that can be considered psychologically motivated. The first three sentiments are captured by the overall VADER compound score for each message, as well as two sentiment-specific features (positive and negative). The final five features capture additional stylistic and cognitive characteristics, including the first-person pronoun ratio, absolutist word ratio, negation ratio, exclamation density, and average word length. These features capture strong psychological signals associated with depressive cognitive styles.

3) Cross-Attention Fusion Mechanism

Unlike simple concatenation, SAFM-Net uses a multi-head cross-attention mechanism to fuse linguistic features with contextual embeddings.
First, linguistic features are projected:

$$Q = W_L L$$

where $Q \in \mathbb{R}^d$
Attention is computed as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where:

- $K = H$
- $V = H$

This allows linguistic features to dynamically attend to relevant tokens.

4) Gated Fusion Mechanism
The model combines semantic and attention representations using a learned gate:

$$g = \sigma(W_g[h_{cls}; A])$$
$$h_{fused} = g \cdot h_{cls} + (1 - g) \cdot A$$

This adaptive fusion enables the model to weigh contextual and linguistic signals dynamically.

5) Dual-Head Prediction
Binary Classification Head:

$$y_{bin} = \text{softmax}(W_b h_{fused})$$

This predicts depression probability.

Ordinal Head:

$$y_{ord} = \sigma(W_o h_{fused})$$

Outputs:

- $P(Y \geq \text{Medium})$
- $P(Y \geq \text{High})$

C. Ordinal Risk Modeling
To derive class probabilities:

$$P_{high} = P(Y \geq High)$$
$$P_{med} = P(Y \geq Med) - P(Y \geq High)$$
$$P_{low} = 1 - P(Y \geq Med)$$

Expected risk score:

$$R = 0 \cdot P_{low} + 0.5 \cdot P_{med} + 1 \cdot P_{high}$$

D. Final Probability Estimation
The final probability is computed as:

$$P = 0.65 \cdot R + 0.35 \cdot P_{binary}$$

This ensures that the ordinal structure contributes stability to the prediction, while the binary head contributes confidence, resulting in a balanced and reliable final estimate.

E. Stability Layer (Hybrid Logic)
To further ensure real-world consistency, a rule-based stabilization layer is implemented on top of the learned model's outputs. The presence of a few crisis phrases ensures a minimum probability of P ≥ 0.80. The presence of distress indicators is balanced with functioning language, pinning the output in the medium risk band. Finally, any protective language detected in the input causes the model's probability to be reduced. This results in a hybrid learned and rule-based model that achieves good results in real-world scenarios.

F. Algorithmic Workflow
The complete training and inference procedures of SAFM-Net are formally described in Algorithm 1 and Algorithm 2 respectively.
Algorithm 1: Training Pipeline
Input: dataset.csv
Output: Trained SAFM-Net model and tokenizer

1. Load dataset from dataset.csv
2. Inject hard negative samples into the non-depression class
3. Apply data cleaning pipeline
4. Balance classes via under sampling
5. Split dataset into training and test subsets
6. For each text sample do
7. Extract eight-dimensional linguistic feature vector
8. Compute severity score and assign ordinal label
9. Tokenize text using DistilBERT tokenizer
10. End for
11. Initialize SAFM-Net model parameters
12. For each training epoch do
13. Perform forward pass through SAFM-Net
14. Compute composite loss: L = L_focal + L_ordinal_BCE + L_Brier
15. Perform backpropagation
16. Update model weights via optimizer step
17. End for
18. Evaluate model on held-out test set
19. Save trained model weights and tokenizer

Algorithm 2: Inference Pipeline
Input: Raw user-provided text
Output: Risk tier classification and associated probability score

1. Normalize input text
2. Tokenize normalized text using DistilBERT tokenizer
3. Extract eight-dimensional linguistic feature vector
4. Perform forward pass through SAFM-Net
5. Compute binary depression probability from classification head
6. Compute ordinal probabilities from ordinal prediction head
7. Convert ordinal probabilities to tier-level probabilities
8. Compute final probability estimate via weighted combination
9. Apply rule-based stability layer
10. Return risk tier and final probability score

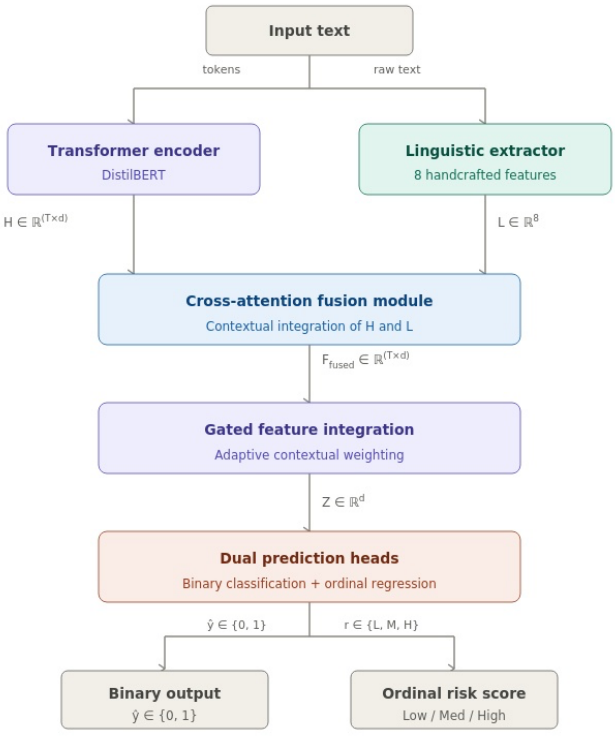

Fig 1: Architecture of SAFM – Net

IV. DATASET AND PREPROCESSING

Within this experiment, the dataset that is used consists of textual examples labeled by means of binary labels with respect to their depression statuses; these examples have been extracted from an open-source social media dataset.[15]

A. Dataset Statistics

The size of the initial dataset included 7731 examples, consisting of almost equal numbers of non-depression examples (3900 exam The samples went through pre-processing step which reduced the samples to 6795 samples (3851 non-depression samples and 2944 depression samples). Before training any model it is very important to keep the class balanced. We used under-sampling strategy which resulted in 5898 samples with 2944 samples of each class.[16]

B. Hard Negative Injection

To prevent false positives, we used curated "well-being" samples in the non-depression class. Hard negatives, i.e., well-being texts of higher quality, were injected into the model to improve its ability to generalize and to prevent false classifications of positive language as depressed in real world use cases.

C. Data Cleaning

A cleaning pipeline that is systematic was used to guarantee the quality of data. Blanks and blank text were eliminated, and filtering of token length was applied to maintain only those samples between the range of 3 to 200 words. Redundant records were then determined and removed in order to avoid leakage of data between training and evaluation splits.[17]

D. Feature Extraction

Any given sample was converted to two complementary representations. The former is a token embedding sequence generated by the DistilBERT encoder, which encodes rich contextual semantics. The second is eight-dimensional linguistic features encoding which encodes the psychologically motivated surface-level attributes as stated in Section II.

E. Ordinal Label Assignment.

An ordinal labelling scheme was implemented on each sample to allow the prediction of a risk that had been graduated. The number was then used to compute a severity score and the median of the resulting distribution was taken as the classification threshold. The samples were then coded using three ordinal items; Low, which was encoded as [0, 0]; Medium, which was encoded as [1, 0]; and High, which was encoded as [1, 1]. The overall proportion of the label distribution in the dataset was 2,944 Low-risk, 1,468 Medium-risk, and 1,476 High-risk, which showed a systematic and understandable stratification of the severity of depression.

V. TRAINING STRATEGY

A. Loss Function

The model is trained with a composite loss function which is a sum of three complementary objectives, formally defined as $L = \alpha L_{focal} + \beta L_{ordinal} + \gamma L_{brier}$, and, α, β and γ are weighting coefficients which are scalar values that regulate the contribution made by each of the loss components. This multi-objective formulation guarantees that the model is optimized at the same time in terms of classification accuracy, ordinal consistency and probabilistic calibration.

B. Focal Loss

The focal loss is defined as $L_{focal} = -(1 - p_t)^{\gamma} \log(p_t)$, p t represents the probability of the true class of the model estimated and gamma a focusing parameter. This loss is also used to down-weight easy-to-classify samples, and is especially useful in mental health classification problems where it is often clinically important to classify borderline cases.[18]

C. Ordinal Loss

The ordinal loss is defined as $L_{ordinal} = BCE(y_{ord}, y_{target})$, where binary cross-entropy is used in the ordinal label encoding. This formulation punishes the model against violations of the ordinal structure, and makes sure that the predicted risk changes between the Low, Medium, and High levels are monotonically consistent and explicable by clinicians.

D. Brier Score Loss

The Brier score loss is defined as $L_{brier} = (p - y)^2$ in which p denotes the forecasted probability and y is the actual binary label. This name punishes the squared distance between the estimated probabilities and the ground truth values to induce the model to make highly-calibrated estimates of the probabilities instead of just optimizing classification accuracy.

E. Optimization Strategy

The optimization methodology uses varying learning rates in encoder and prediction heads, which is a standard practice in applying a lower learning rate to the pretrained transformer layers to conserve learned representations but have the task-specific heads learn faster. In particular, the learning rate of 2×10^{-5} is fed to the DistilBERT encoder, and a larger learning rate of 5×10^{-4} is set on the prediction heads. The initial part of the training is a warmup that occupies 10% of total training steps, where the learning rate is decayed by a cosine schedule to allow a gradual convergence, and reduce the probability of oscillation in later training.[19]

F. Training Configuration

Training and evaluation were performed with a batch size of 8 and 16 respectively. The process of all experiments was carried out on a CPU machine, during which 2,064 training steps and a warmup time of 206 steps were completed. The reason behind this choice of configuration was to trade-off between computational feasibility and adequate exposure to the training data, which guarantees a stable and reproducible convergence of the model.

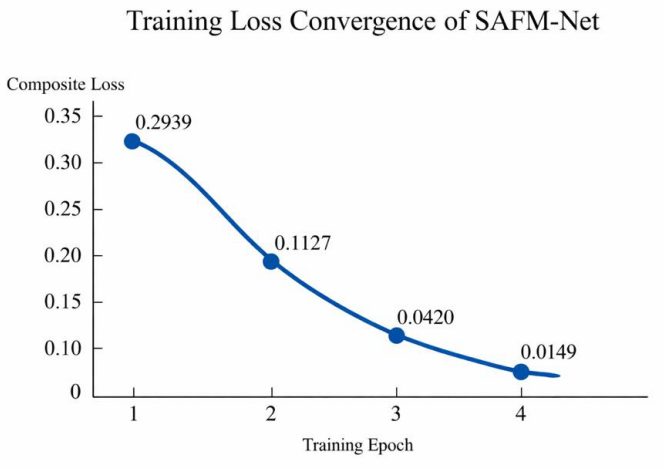

Fig 2: Training Loss Convergence of SAFM - Net

VI. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed SAFM-Net model was tested on a stratified test split of the data. The evaluation was based on accuracy of classification and probability calibration so that the model will give interpretable and reliable results.[20]

A. Training Dynamics and Convergence

The model was optimized with a total of four epochs and a composite loss of focal loss, ordinal binary cross-entropy, and

Brier score minimization. The convergence of the training process was stable, and the loss of the training process dropped significantly with the number of epochs. In particular, the loss on training decreased in the first epoch to 0.2939 and then to 0.1127 in the second and fourth epochs respectively, which indicates a consistent and speedy convergence of the optimization process.

Class	Precision	Recall	F1-score	Support
0 (Non-depression)	0.97	0.98	0.97	442
1 (Depression)	0.98	0.97	0.97	442

Table 1: Training Dynamics and Convergence

The performance in terms of validation increased gradually as the training loss decreased, and the best model checkpoint was determined by the minimum observed Brier score. The highest validation accuracy of 97.40, and the highest Brier score of 0.045955 showed that the model not only had high classification accuracy, but also generated calibrated probability estimates of all risk levels.

This was done by training the encoder and prediction heads with different learning rate schedules respectively, using both a warmup cosine learning rate scheduler and differentiated learning rates. The warmup period enabled the model to ramp up to its optimal learning rate and then the cosine decay schedule to gradually decrease the learning rate, avoiding sudden parameter changes during the initial training. To avoid distorting the pretrained representations of DistilBERT, the encoder learning rate was maintained at a relatively low level, whereas the learning rate of the fusion heads was increased to ensure quicker adaptation to the task-specific classification task. This asymmetric optimization approach was sufficient to reduce the risk of overfitting, and the results were constant improvement of the training loss reduction and validation performance improvement over the four epochs.

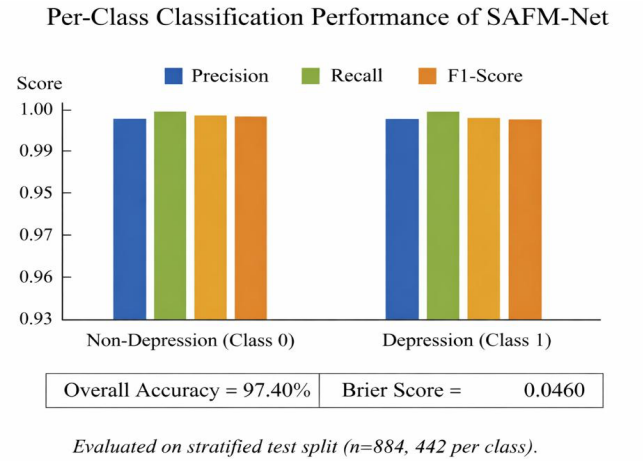

Fig 3: Per-Class Classification Performance of SAFM-Net

B. Final Model Performance

The final evaluation results are presented below:

Metric	Value
Accuracy	0.9740
Brier Score	0.04595
Precision (avg)	0.97
Recall (avg)	0.97
F1-score (avg)	0.97

Table 2: Final Model Performance

These findings suggest that the proposed model obtains high classification accuracy and probability predictions with good calibration.

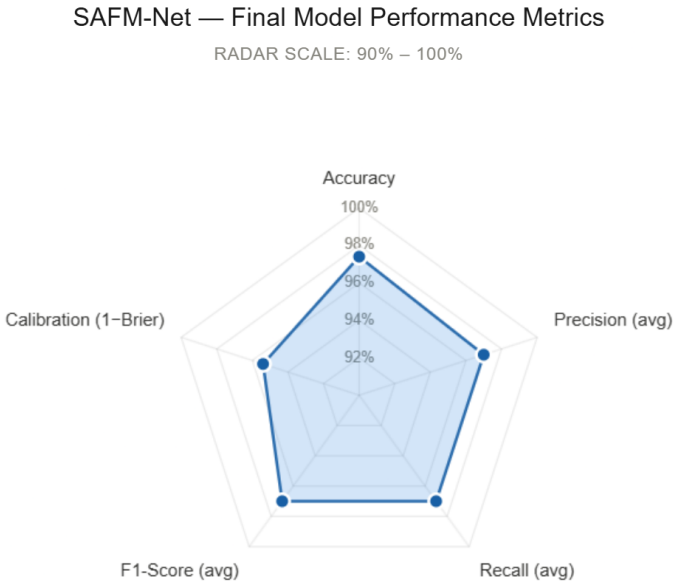

Fig 4: SAFM – Net – Final Model Performance Matrix

C. Classification Report

Equal performance in the two classes shows that the model is not biased towards the two classes, which is important in mental health use.

D. Confusion Matrix Analysis.

$$\begin{bmatrix} 433 & 9 \\ 14 & 428 \end{bmatrix}$$

As the confusion matrix shows, the model is very effective in reducing both forms of misclassification. False positives were also very low at 9 cases which means that the model does not tend to diagnose non-depressive cases as depressive hence lowering chances of over-diagnosis. On the same note, the false negative of 14 cases reaffirms the model rarely misses to identify the true cases of depression, a factor that curbs the chances of under-diagnosis. Combined, the results prove that this model represents an efficient and a clinically responsible balance between sensitivity and specificity that makes it a valuable instrument to be used in the real world to assess depression risk.

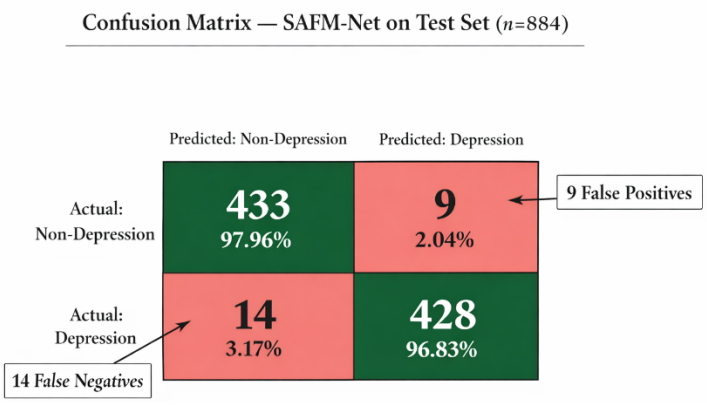

Fig 5: Confusion Matrix – SAFM-Net on Test Set

E. Calibration of probabilities (Brier Score Analysis)

The Brier value of 0.04595 is a good probability estimation. In contrast to the conventional measures of accuracy, the Brier score assesses the similarity between the predicted and the real probabilities. The lower the score, the better the calibration.[21]

The training goal with the Brier loss added was much more effective in enhancing the predictive reliability of the model and therefore suited risk sensitive applications.

VII. SYSTEM COMPONENT ANALYSIS

By reviewing each component of the structure in question, it will be possible to have an even clearer idea of how well this design works.

A. Effect of Language Features

The addition of special language characteristics to a model allows the model to significantly improve the recognition of meaningful patterns within text, associated with the thought processes and emotional states of individuals. The addition of language characteristics allows the model to recognize signals such as self-reference (i.e., reference to oneself), negation (e.g., "no") and absolutes (e.g., "never"), all of which may indicate the presence of negative thinking, based upon past research concerning language and psychology. While the primary function of the transformer encoder is to assess the overall context of the text including the relationships among individual words, the combined use of both types of knowledge provide the model with a comprehensive, accurate representation of the content expressed in the text, thereby providing greater clarity into assessing the underlying meaning of the text. In essence, this represents a connection between knowledge learned through data and knowledge acquired from psychological theories regarding the relationship between language and cognition.[22]

B. The language and token features are significantly linked through the cross-attention mechanism which, in fact, they work together and change with each other. This very fact enables the model to have the orientation to those areas of the text that are essential, according to the knowledge it has on human cognition, and that way, on the one hand, it becomes clearer to us why the model is doing something, and on the other hand, its efficiency is improved a lot.

C. Ordinal Modelling Benefits

Admitting that a thing is either yes or no is extremely simple. This tactic truncates the complicated mental health risk concept to just two parts and, does not indicate how much risk there is. However, the correction to this lie in the ordinal prediction. An ordinal model can put forward the estimate of the probability of the different levels of the problem- like low, medium, high, for instance. By getting a risk assessment with three levels, rather than just two, you paint a much more accurate picture of the case than a standard 'yes or no' model.[23]

D. Stability Layer Contribution

The system's reliability is significantly improved by the rule-based stabilization layer. The layer, by adding domain-informed constraints to the outputs of the model learned, ensures that the predictions stay the same for linguistically similar inputs, makes medium-class classifications more stable by decreasing the inevitable lower confidence of the model, and makes the model's output more similar to the actual clinical expectations. This method of combining learned probabilistic inference with deterministic rule enforcement leads to a more reliable and ready-to-deploy risk assessment system.[24]

The evidence from the data makes it clear that SAFM-Net has a successful method for acquiring linguistic and contextual features for mental state detection. Thanks to the high accuracy and well probability calibration, which are absolutely important in the real world, the model stays unique in its kind.

The negative samples are the main source of the model's stability, which they do by deleting the false high-risk predictions from normal text. Furthermore, the system is also ordinal modeled, which makes it possible to show risk levels that are more appropriate than the binary ones.

Nonetheless, it should be highlighted that the model is not designed to substitute clinical diagnosis in any form. It, on the contrary, serves as the main instrument for assisting with the early detection of and screening at-risk individuals.

VIII. CONCLUSION

This document dives into a recently created hybrid neural network SAFM-Net, which is designed explicitly for the task of mental health risk identification through the use of textual data. The mechanism is presented as being based on the transformer-based contextual embeddings, combined with additional psychological and linguistic constraints through a mechanism of attention-based fusion. By making use of semantic comprehension from text and the inclusion of indicator words about language, the model is advanced to the level of capturing both the deep contextual meaning and the faint behavioral patterns that the individual presents, which are possible signs of psychological problems. Moreover, the dual-head architecture enables the implementation of the model in both the binary classification task and the ordinal risk modeling task, while a stability layer confirms the attainment of the more consistent and uniform predictions on different samples.

SAFM-Net has been proven to be excellent, attaining a high success rate of 97.40%, as well as having strong probability calibration, as shown by a Brier score of 0.04595. These outcomes indicate that the model suggested is both dependable and very efficacious in dealing with the constraints that are commonplace to many traditional machine learning and deep learning systems. The model's ability to contribute both semantic and psychological representations in a unified framework sets it apart from others in terms of performance and robustness.

The method is a reliable instrument for evaluating psychological risks in the patients in real-life settings. Besides, it may be useful to the healthcare people in the timely and accurate decision-making and the early screening and treatment of patients. An extension of veracity on larger and more mixed groups of subjects can make the SAFM-Net develop into a prominent player in the upcoming AI-mental health care systems and digital well-being platforms.[25]

IX. FUTURE WORK

The potential for further investigations on various themes is limitless. First of all we could test the method in multilingual situations since this is one of the most promising and important approaches to the topic. The situational language of mental health talks which obviously is so distinct and varied between the speakers of different languages and their cultures and The languages used by existing transformer architectures also based on mainly the English language and these might not apply to other languages is the main reason why this is a good option. The other one is the variety of data, it is expected to be the combination of textual inputs with prosodic information, speaking speed, and vocal emotions which are extracted from the spoken language.

Moreover, thirdly, the validation from the clinically annotated data set is one of the factors that should be considered as the datasets that are created by the community, e.g., through social media, do not always reflect the language and demographics of the patient populations. Additionally, the validation through clinical settings would imply that the current model could be compared with the existing screening mechanisms and the

regulatory factors that are involved in such decisions could also be discussed. The fourth point is that the present rule-based stability layer can be changed {{replaced}} through the end-to-end learning method which would be carried out through the use of differentiable constraints or additional loss functions that would encode the criteria of stability during the training stage, rather than during the testing stage. Finally, an app interface should be designed for mobile use to which access to mental health screening can be increased among the patients in the underserved areas only if precautionary measures are taken ahead of time.

REFERENCES

[1] A. Calvo, D. Milne, M. Hussain, and C. Nugent, "Natural Language Processing for Mental Health Interventions: A Systematic Review," *Computer Methods and Programs in Biomedicine*, vol. 156, pp. 45–52, 2018.

[2] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying Mental Health Signals in Twitter," in *Proc. ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," in *Proc. NeurIPS Workshop on Energy Efficient Deep Learning*, Vancouver, Canada, 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[6] M. Gkotsis et al., "Characterisation of Mental Health Conditions in Social Media Using Informed Deep Learning," *Scientific Reports*, vol. 7, no. 1, p. 45141, 2017.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Z. Ji, Z. Lee, F. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[9] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[10] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, Ann Arbor, MI, USA, 2014, pp. 216–225.

[11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *Proc. International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[12] P. Liang et al., "Holistic Evaluation of Language Models," *Transactions on Machine Learning Research*, 2022.

[13] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in *Proc. International Conference on Machine Learning (ICML)*, Bonn, Germany, 2005, pp. 625–632.

[14] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[15] K. Losada and F. Crestani, "A Test Collection for Research on Depression and Language Use," in *Proc. CLEF*, 2016.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, TX, USA, 2016, pp. 2383–2392.

[17] S. Miao, C. Liang, and K. Su, "A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020, pp. 975–984.

[18] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.

[19] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training Verifiers to Solve Math Word Problems," *arXiv preprint arXiv:2110.14168*, 2021.

[20] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[21] S. Kadavath et al., "Language Models (Mostly) Know What They Know," *arXiv preprint arXiv:2207.05221*, 2022.

[22] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 111–116.

[23] C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions," in *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, MN, USA, 2019, pp. 2924–2936.

[24] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 3214–3252.

[25] Y. Bang et al., "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity," *arXiv preprint arXiv:2302.04023*, 2023.