Exploring Kolmogorov Arnold Networks for Interpretable Mental Health Detection and Classification from Social Media Text

Ajay Surya Jampana, Mohitha Velagapudi, Neethu Mohan, Sachin Kumar S

Amrita Vishwa Vidyapeetham

Abstract

Mental health analysis from social media text demands both high accuracy and interpretability for responsible healthcare applications. This paper explores Kolmogorov Arnold Networks (KANs) for mental health detection and classification, demonstrating their superior performance compared to Multi-Layer Perceptrons (MLPs) in accuracy while requiring fewer parameters. To further enhance interpretability, we leverage the Local Interpretable Model-Agnostic Explanations (LIME) method to identify key features, resulting in a simplified KAN model. This allows us to derive governing equations for each class, providing a deeper understanding of the relationships between texts and mental health conditions.

1 Introduction

Mental health illness, which gained significant importance in recent years is still one of the serious health concerns worldwide (Rehm and Shield, 2019).There are multiple mental illnesses from stress, depression, and anxiety to more severe ones like schizophrenia, Autism spectrum disorder (ASD) etc., affecting millions globally. Not only does it influence an individual's health, but also has a significant impact on society's health and well-being as a whole. In 2019, it was reported by the World Health Organization (WHO) that a staggering 1 out of 8 individuals were suffering from one or the other mental disorders , the most prominent ones being anxiety and depression.[1] Some studies have revealed that suicidal risk in an individual is strongly connected to his mental health condition (Windfuhr and Kapur, 2011). As per the reports published in [2] , a staggering 1.6 million individuals in England were waiting for mental health care services. Thus, comprehending various mental disorders necessitates developing a robust and accurate classification system for early detection. There are numerous ways in which individuals express their moods, feelings, and personal life experiences, with social media platforms being the most prominent one's (Zarrinkalam et al., 2020; Saha et al., 2022; Prieto et al., 2014).These writings can be used to make inferences about one's mental state leveraging NLP techniques (Chereddy et al., 2024) since around 80% of people tend to disclose their suicidal thoughts on these platforms (Golden et al., 2009) . In particular, for applications like healthcare, an accurate and transparent decision-making process is needed because incorrect predictions might lead to life-altering consequences.Despite the superior performance LLM's and deep-learning methods are still treated as black-boxes due to their lack of interpretability (Castelvecchi, 2016).

Our work embarked on exploring the power of Kolmogorov-Arnold Networks (KAN) to enhance classification accuracy over the traditional multilayer perceptron(MLP). Simultaneously, to interpret the outputs and provide transparency we employed LIME (Local Interpretable Model-agnostic Explanations) as an XAI approach.The novelty of our work lies in the application of KAN for multiclass classification as well as for binary classification in the realm of mental health followed by using XAI approach to gain insights into the explainability of our model. Our work stands out as the first to combine KAN with explainability techniques for mental health classification.

2 Related Works

Earlier studies focused on traditional machine learning methods typically following a pipeline approach. Among these methods, supervised learning is predominantly employed including SVM (Ziwei and Chua, 2019; Prakash et al., 2021; Coello et al., 2019), k-Nearest Neighbors (KNN) (Verma et al., 2021; Tlachac et al., 2019) Logistic Model Tree

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] https://is.gd/abioF_theguardian

(LMT) (Briand et al., 2018)Decision Tree (Marerngsit and Thammaboosadee, 2020; Fodeh et al., 2019), Logistic Regression (Németh et al., 2020; Benton et al., 2017)and several ensemble models (Chadha and Kaushik, 2021; Sekulic et al., 2018; Kumar et al., 2019). Reinforcement learning was also used in depression detection by giving attention to only useful information instead of noisy data (Gui et al., 2019). Deep learning techniques have started gaining more attention for detecting mental illness in recent years. (Murarka et al., 2021) used RoBERTa to classify 5 mental health conditions from Reddit posts (anxiety, bipolar disorder, ADHD, depression, and PTSD). Numerous studies have focused on the analysis of language used in social media for classifying emotions (Kumar et al., 2022) or aiming to identify depressed individuals. By focusing on linguistic choices made by individuals from controlled user groups and depressed user groups the researchers were able to classify the depressed individuals (Choudhury et al., 2013; De Choudhury et al., 2021; Coppersmith et al., 2014). Subsequently, (Coppersmith et al., 2015) examined the likelihood of generation of characters as a sequence by employing character-level language model. (Husseini Orabi et al., 2018) incorporated word embeddings with NN models like CNN and RNN for depression detection. In (Sekulic and Strube, 2019) experimented with Hierarchial Attention Networks (HAN) as part of a multi-class classification and designed a binary classifier for each disorder. Besides their remarkable performance using deep learning techniques, they are unable to provide transparency in predictions. Most of the research in mental health detection is majorly centered around improving accuracy over a model's ability to interpret as shown in (Su et al., 2020; Greco et al., 2023). To the extent of our understanding, only a handful of papers, as noted in (Song et al., 2018; Turcan et al., 2021; Sekulic and Strube, 2019) have focused on explainability.

3 Materials and Methods

In this section, we provide a comprehensive overview of the materials and methodologies employed in our study. We detail the preparation of datasets used for mental health detection and classification, describe the feature extraction methods, and elaborate on how Kolmogorov Arnold Networks (KAN) were trained and compared with other existing models.

3.1 Dataset Summary

The Mental Health Corpus (MHC) [3]. and the Reddit Mental Health Dataset (RMH) (Low et al., 2020) were chosen due to their strong alignment with the task of mental health detection and classification. Both datasets consist of text data directly related to mental health conditions, making them highly relevant for the task at hand. The MHC dataset is suitable for binary classification tasks, focusing on distinguishing individuals with mental health concerns from others. On the other hand, the RMH dataset provides a diverse range of mental health conditions (ADHD, anxiety, depression, and suicidewatch), allowing for multi-class classification. Additionally, the real-world, unstructured text from social media platforms such as Reddit captures the natural language used by individuals to discuss their mental health, enhancing the model's real-world applicability. These datasets provide a balance of simplicity (binary classification) and complexity (multi-class classification), ensuring that the model can handle both general and specific mental health detection tasks.

3.2 Data Preprocessing

A standardized text preprocessing pipeline was applied uniformly across both datasets. To begin with, all text was converted to lowercase to ensure consistency and reduce redundancy. Subsequently, URLs were removed using a regular expression pattern to eliminate potential noise. Punctuation was then stripped to further clean the text, followed by elimination of stopwords. Finally, Porter Stemmer is used to stem words to their root form, thereby standardizing the text and reducing dimensionality. These preprocessing steps were crucial in preparing the raw text for effective feature extraction and model training.

3.3 Feature Extraction

After performing the necessary preprocessing, it is crucial to extract meaningful features for training our models. For this purpose, we employed DistilBERT (Jose and Harikumar, 2022), a lighter, faster and a distilled variant of BERT (Bidirectional Encoder Representations from Transformers) which runs 60% faster while preserving over 95 % of BERT's performance, excels at capturing contextualized word embeddings. This approach is

[3] https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus

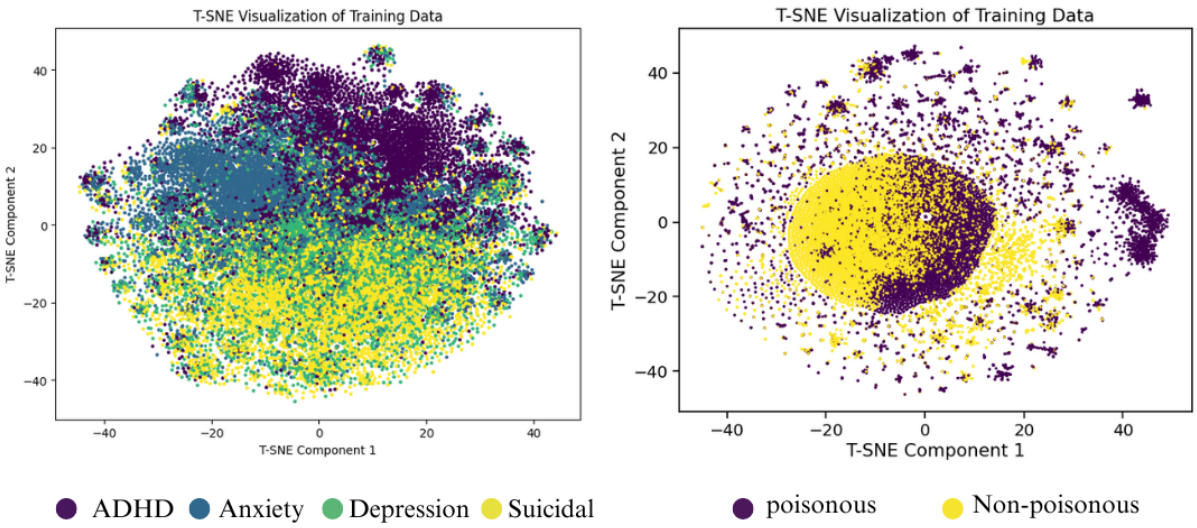

Figure 1: t-SNE Plot of DistilBERT Features for RMH (left) and MHC (right) Datasets

particularly advantageous for our datasets, which consist of diverse, context-rich social media texts. DistilBERT's ability to comprehend nuanced language helps in identifying mental health indicators such as depression, ADHD, anxiety, and suicidal thoughts, providing deep, contextual insights into the data. Furthermore, t-SNE (t-distributed Stochastic Neighbor Embedding) plots were generated for both datasets (Fig.1) to visualize document distributions in a lower-dimensional space (Nadella et al., 2023), demonstrating how DistilBERT embeddings effectively differentiate between various mental health conditions discussed in social media.

3.4 Komlogrov Arnold Networks

Komlogrov-Arnold Networks (KANs) are a type of neural network that uses the Komlogrov-Arnold representation (KAR) theorem (Akashi, 2001) instead of the universal approximation theorem found in neural networks. The activation functions in a neural network are static and computed at the nodes. In contrast, KANs have learnable activation functions on edges ("weights"). As a result, KANs have no linear weight matrices at all: instead, each weight parameter is replaced by a learn able 1D function parametrized as a spline. KANs' nodes simply sum incoming signals without applying any non-linearities.(Liu et al., 2024). KAN's are inspired from the Komlogrov-Arnold Representation theorem (KAR), which states that any multivariate function 'f' can be expressed as a finite composition of continuous functions of a single variable, combined with the binary operation of addition. Mathematically, the theorem can be articulated as follows:

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right) \quad (1)$$

where $f(x_1, \ldots, x_n)$ is a multivariate function, $\phi_{q,p}(x_p)$ are the univariate functions, and Φ_q takes the univariate functions and combines them. Eq. (1) implies that the learning of f(x) is complete if we can find appropriate univariate functions $\phi_{q,p}$ and Φ_q. Since all functions to be learned are univariate functions, each 1D function can be parameterized as a B-spline curve, with learnable coefficients of local B-spline basis functions. For comparison, a Multi-Layer Perceptron (MLP) can be written as an interleaving of affine transformations weights and non-linear activation functions, whereas, if Φ_l is the function matrix corresponding to the l-th Kolmogorov-Arnold Network (KAN) layer, a general KAN network is a composition of L layers. Given an input vector $x_0 \in R^{n_0}$, the output of KAN is:

$$\text{KAN}(x) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \cdots \circ \Phi_1 \circ \Phi_0)x. \quad (2)$$

KANs utilize residual activation functions by combining a basis function $b(x)$ and a spline function. The activation function $\phi(x)$ is defined as:

$$\phi(x) = w_b b(x) + w_s \,\text{spline}(x) \quad (3)$$

where w_b and w_s are learnable weights. The basis function $b(x)$ is set to the SiLU (Sigmoid Linear Unit) activation function:

$$b(x) = \text{silu}(x) = \frac{x}{1 + e^{-x}} \tag{4}$$

The spline function is parametrized as a linear combination of B-splines:

$$\text{spline}(x) = \sum_i c_i B_i(x) \tag{5}$$

where c_i are trainable coefficients and $B_i(x)$ are B-spline basis functions. Each activation function in KANs is initialized such that $w_s = 1$ and the spline function is initialized close to zero, ensuring a smooth start to learning. The weight w_b is initialized using the Xavier initialization method, which helps stabilize the gradients during the initial training phase.

4 Proposed Framework

The steps involved in the classification of mental health diseases for both datasets using the proposed method is presented as a block diagram in Fig.2.

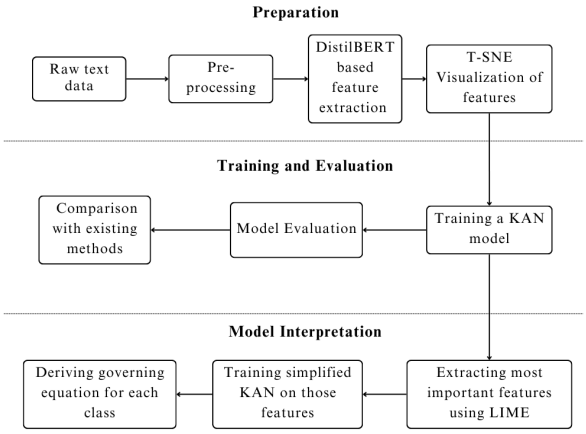

Figure 2: Model Flow chart

4.1 Hyperparameter Optimization

To determine the optimal set of hyperparameters for the KAN model, hyperparameter optimization was conducted using Bayesian Optimization. This method was selected for its efficiency in exploring the hyperparameter space while balancing exploration and exploitation. After extensive tuning, the optimal configuration was identified as a two-layer architecture with (128, 64) neurons for both the MHC and RMH datasets. A third-order basis spline function continued to be used for estimating the activations. The optimal learning rate, selected through Bayesian Optimization, was found to be 0.001, with a regularization parameter of $\lambda = 0.1$. The batch size was optimized to 32, balancing convergence speed and stability. For the MHC dataset, binary cross-entropy loss remained the most suitable for the binary classification task, while categorical cross-entropy was employed for the multi-class nature of the RMH dataset. The optimized model converged in 30 epochs for MHC and 50 epochs for RMH, outperforming the initial settings. The output layer sizes of 2 for MHC and 4 for RMH remained consistent with the number of classes in each dataset.

5 Results

In the proposed study, a set of well-established evaluation metrics for classification, including precision, recall, F1-score, and average accuracy, were used to assess the effectiveness of the proposed framework across multiple datasets. To benchmark the performance of Kolmogorov Arnold Networks (KAN), we compared it with the Support Vector Machine (SVM) utilizing the Neural Tangent Kernel (NTK) (Chen et al., 2022), as well as the Multilayer Perceptron (MLP). SVM with NTK was chosen due to its ability to perform well in small-sample scenarios and its theoretical connections to neural networks, which offer valuable insights when compared to KAN's neural structure. On the other hand, MLP is widely used in classification tasks due to its simplicity and capability to approximate any continuous function, making it a close structural competitor to KAN.

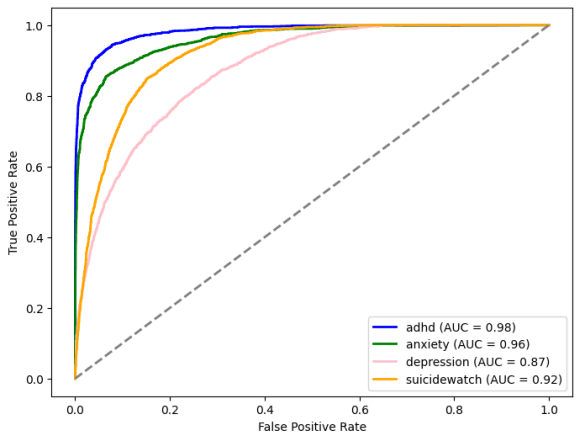

Figure 3: ROC curve for RMH Dataset using KAN

Fig. 3 presents the Receiver Operating Characteristic (ROC) curves for the KAN model, plotted for

Table 1: Performance of different models on the Reddit Mental Health dataset.

Model	Class	Precision	Recall	F1-score
SVM with Neural Tangent Kernel	ADHD	0.86	0.86	0.86
	Anxiety	0.82	0.78	0.80
	Depression	0.60	0.60	0.60
	Suicidewatch	0.68	0.71	0.70
Multi-Layer Perceptron (MLP)	ADHD	0.86	0.87	0.87
	Anxiety	0.84	0.75	0.79
	Depression	0.53	0.65	0.58
	Suicidewatch	0.68	0.60	0.64
Komlogrov Arnold Networks (KAN)	ADHD	0.91	0.86	0.88
	Anxiety	0.84	0.84	0.84
	Depression	0.64	0.60	0.62
	Suicidewatch	0.69	0.77	0.72

Table 2: Performance of different models on the Mental Health Corpus dataset.

Model	Class	Precision	Recall	F1-score
SVM with Neural Tangent Kernel	No Mental Illness	0.86	0.91	0.88
	Mental Illness	0.90	0.85	0.87
Multi-Layer Perceptron (MLP)	No Mental Illness	0.88	0.88	0.88
	Mental Illness	0.87	0.87	0.87
Komlogrov Arnold Networks (KAN)	No Mental Illness	0.89	0.90	0.90
	Mental Illness	0.90	0.89	0.89

each class in the RMH dataset. These scores are indicative of the KAN model's strong classification performance, with ADHD achieving the highest AUC of 0.98, followed by anxiety (AUC = 0.96), suicidewatch (AUC = 0.92), and depression (AUC = 0.87). The higher AUC values signify that the model is able to maintain a high true positive rate with relatively fewer false positives, which is crucial in sensitive applications such as mental health classification, where misclassification can have serious consequences.

5.1 Reddit Mental Health (RMH) Dataset

Based on the training parameters discussed in the previous section, we trained the KAN and achieved an average test accuracy of 77%. For the MLP, the average test accuracy was 72%, and for the SVM with NTK, it was 74%. Regarding training time, MLP performed better, taking just 43 seconds, whereas KAN took 395 seconds. However, it was evident that KAN performed well on other evaluation measurements such as precision, recall, and F1-score, as seen from Table.1, when compared to MLP and SVM with NTK.

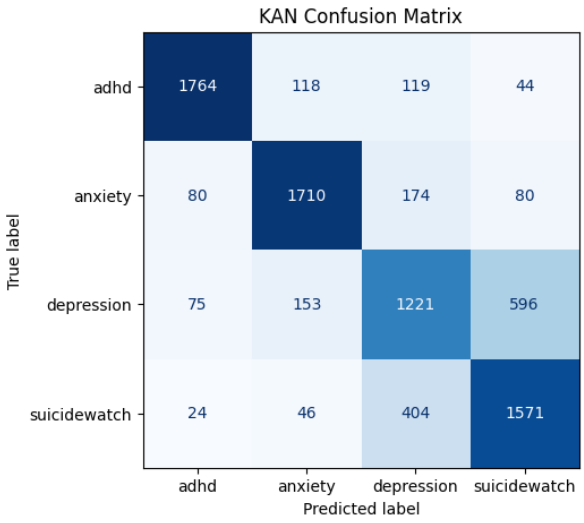

Figure 4: KAN Confusion matrix for RMH Dataset

Additionally, the confusion matrix for KAN showed significantly fewer false negatives as seen

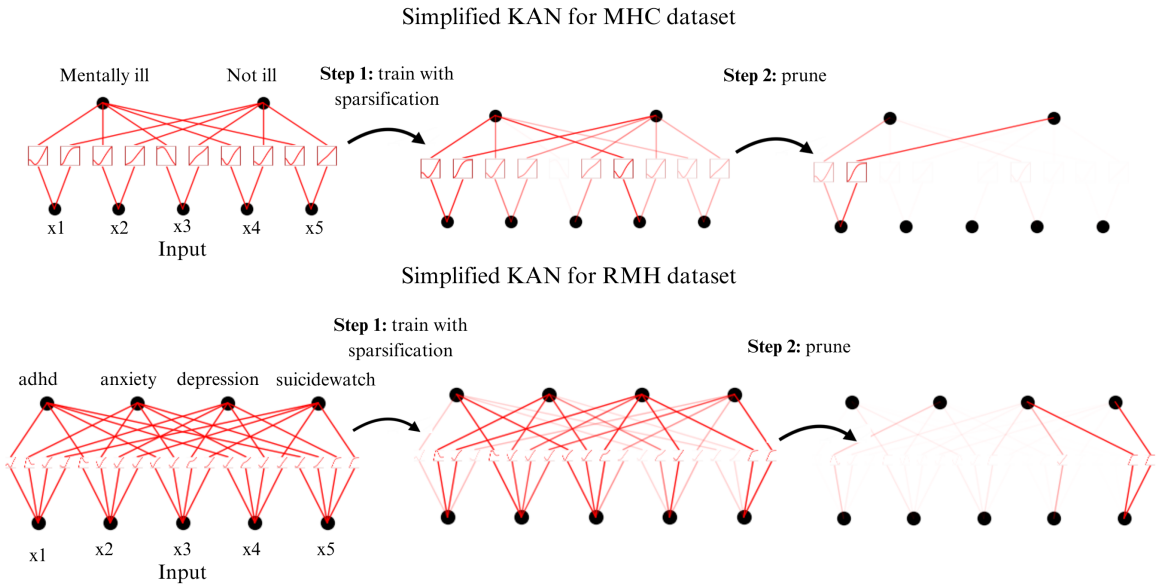

Figure 5: Simplified KAN's for both datasets

in Fig.4, which is crucial as it is important not to misdiagnose a person with a disease.

5.2 Mental Health Corpus (MHC) Dataset

For the MHC dataset, the average test accuracy for KAN was 90%, while for MLP and SVM with NTK, it was 88%. The training time for MLP was again better, taking 32 seconds, whereas KAN took 267 seconds.

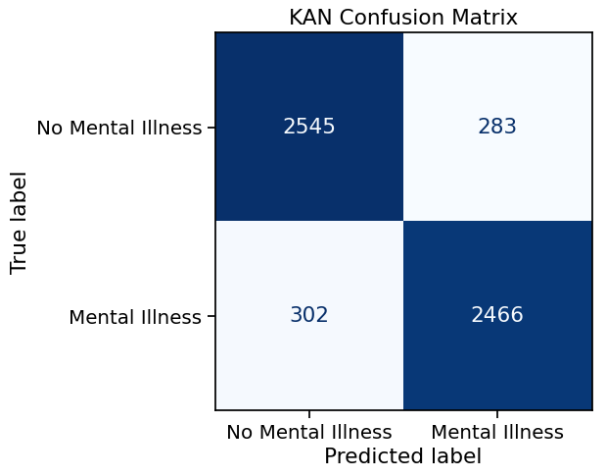

Figure 6: KAN confusion for MHC Dataset

Similarly, for this dataset, KAN performed well across all evaluation metrics compared to other methods as seen from Table.2. The confusion matrix (Fig.6) for KAN indicated that it effectively reduced false negatives, ensuring that individuals without mental illness were not misclassified as having a mental illness, which is critical.

6 Model Interpretation

After training the KAN model, we employed the LIME (Local Interpretable Model-agnostic Explanations) explainability technique to identify the most important features that contributed the most to computing the output. LIME generates a local surrogate model, typically a linear model, that approximates the predictions of the complex model around the instance of interest (Ribeiro et al., 2016). This method is particularly suitable for our problem because it allows us to interpret the model's behavior for specific predictions, making it easier to explain individual decisions, which is crucial for applications in mental health. The features that cause significant changes in the predictions are assigned higher importance scores. We fixed a threshold for the feature importance scores that selected the 5 most contributing features above this threshold as seen from Table. 3 & 4.

Table 3: Top contributing features in MHC dataset

Word	feel	life	depress	want	kill
Impact score	2.7	2.2	1.8	1.7	1.5

Table 4: Top Contributing Features in RMH Dataset

Word	adderal	suicide	life	scare	care
Impact score	0.7	0.5	0.4	0.3	0.3

Reddit Mental Health Dataset	
ADHD	$\sin(x_1-6)+\sin(7.2x_3+3.1)+\sin(4.5x_4-2.4)-\sin(5.8x_5+0.9)+\tanh(9.9x_2-0.6)-0.2$
Anxiety	$\sin(7.5x_1+2.6)+\sin(6.9x_2-3.3)+\sin(x_3-8.8)+\tanh(1.7x_4)+\tanh(3x_5-1.3)-0.53$
Depression	$\sin(5.2x_1+9.9)+\sin(5.7x_2-3)+\tanh(6.4x_3-0.5)+\tanh(7.4x_5-1)+\lvert 8.6x_4-1.8\rvert+0.16$
Suicidewatch	$(0.2-x_1)^2+\sin(5.2x_2+9.8)-\sin(4x_4+7.2)+\tanh(3.4x_3-1.8)+\tanh(7.4x_5-1)+0.45$
Mental Health Corpus Dataset	
No mental illness	$\sin(4.1x_3+7.3)+\lvert 9.9x_1-2.2\rvert+\lvert 4.5x_2-0.8\rvert+\lvert 9.1x_4-1.8\rvert+\lvert 6x_5-1.3\rvert-0.29$
Mental illness	$\sqrt{0.7x_5+1}+\sin(4x_4-8)+\tanh(6x_1-1)+\tanh(3x_2-1)+\tanh(1.3x_3+0.3)-0.56$

Table 5: Governing Equations obtained from simplified KAN

6.1 Simplifying KAN's

To simplify the Kolmogorov Arnold Networks (KANs), we took as input the most important features determined by LIME in the previous step and employed multiple techniques aimed at improving the model's interpretability (see Fig. 5):

1. **Step 1: Sparsification:** We applied entropy regularization to the activation functions to promote sparsity in the model. This method encourages the network to favor simpler and more compact representations of the data.

2. **Step 2: Pruning:** After training the model with a sparsification penalty, we further pruned the network by removing unimportant neurons. Unlike traditional edge-level pruning, we implemented *node-level pruning*. For each neuron (say the i-th neuron in the l-th layer), we computed the incoming ($I_{l,i}$) and outgoing ($O_{l,i}$) scores and retained only those neurons whose scores exceeded a threshold hyperparameter ($\theta = 10^{-2}$). This technique reduces the network to a smaller subnetwork, retaining only the most important neurons.

Finally we were able to derive governing equations for each class as activation functions learned by KAN are symbolic and they capture the behavior of text data from each class. These equations as seen from Table. 5 can be incredibly useful, as they enable predictions without retraining the entire network. This capability saves a considerable amount of time and computational resources, making the model more efficient for future predictions.

7 Conclusion

Our study explored Kolmogorov Arnold Networks (KAN) as an interpretable alternative to traditional neural networks for mental health detection and classification from social media text. KAN outperformed Multilayer Perceptrons (MLP) and SVM with Neural Tangent Kernel (NTK) in terms of accuracy and other evaluation metrics across the RMH and MH Corpus datasets. KAN's unique architecture allows for model simplification and visualization through governing equations, enhancing interpretability—an essential feature for mental health prediction tasks. By utilizing the LIME explainability technique, we identified key features influencing the model's output, enabling the creation of a streamlined KAN model without sacrificing performance.

Limitations

Despite the high accuracy and interpretability offered by Kolmogorov Arnold Networks (KAN), several limitations must be acknowledged. KANs have higher computational complexity and longer training times compared to traditional models, which can make them less suitable for resource-constrained environments and limit their scalability to larger datasets. Additionally, while KANs provide clear interpretability benefits, the process of pruning nodes and simplifying the model may result in the loss of nuanced information critical for specific predictions.

References

Shigeo Akashi. 2001. Application of -entropy theory to kolmogorov—arnold representation theorem. *Reports*

on Mathematical Physics, 48(1):19–26. Proceedings of the XXXII SYMPOSIUM ON MATHEMATICAL PHYSICS.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Antoine Briand, Hayda Almeida, and Marie-Jean Meurs. 2018. Analysis of social media posts for early detection of mental health conditions. In *Canadian Conference on AI*.

Davide Castelvecchi. 2016. Can we open the black box of ai? *Nature*, 538:20–23.

Akshma Chadha and Baij Kaushik. 2021. Machine learning based dataset for finding suicidal ideation on twitter. pages 823–828.

Yilan Chen, Wei Huang, Lam M. Nguyen, and Tsui-Wei Weng. 2022. On the equivalence between neural network and support vector machine.

Spandana Chereddy, K Geetha, and A G Sreedevi. 2024. Tweeting the blues: Leveraging nlp and classification models for depression detection. In *2024 2nd International Conference on Advancement in Computation Computer Technologies (InCACCT)*, pages 875–880.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Web Science Conference*.

Laritza Coello, Rosa Ortega-Mendoza, Luis Villaseñor-Pineda, and Manuel Montes. 2019. *Crosslingual Depression Detection in Twitter Using Bilingual Word Alignments*, pages 49–61.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Samah Fodeh, Taihua Li, Kevin Menczynski, Tedd Burgette, Andrew Harris, Georgeta Ilita, Satyan Rao, Jonathan Gemmell, and Daniela Raicu. 2019. Using machine learning algorithms to detect suicide risk factors on twitter. pages 941–948.

Robert N Golden, Carla Weiland, and Fred Peterson. 2009. *The truth about illness and disease*. Infobase Publishing.

Candida M Greco, Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. 2023. Transformer-based language models for mental health issues: a survey. *Pattern Recognition Letters*, 167:204–211.

Tao Gui, Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang. 2019. Depression detection on social media with reinforcement learning. In *China National Conference on Chinese Computational Linguistics*.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Anagha Jose and Sandhya Harikumar. 2022. Covid-19 semantic search engine using sentence-transformer models. In *Computational Intelligence, Cyber Security and Computational Models. Recent Trends in Computational Models, Intelligent and Secure Systems*, pages 189–200, Cham. Springer International Publishing.

Akshi Kumar, Aditi Sharma, and Anshika Arora. 2019. Anxious depression prediction in real-time social data. *MatSciRN: Other Biomaterials (Topic)*.

Sruthi S. Kumar, S. Sachin Kumar, and K. P. Soman. 2022. Deep learning-based emotion classification of hindi text from social media. In *Advanced Machine Intelligence and Signal Processing*, pages 535–543, Singapore. Springer Nature Singapore.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. 2024. Kan: Kolmogorov-arnold networks.

Daniel M Low, Laurie Rumker, John Torous, Guillermo Cecchi, Satrajit S Ghosh, and Tanya Talkar. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Supawit Marerngsit and Sotarat Thammaboosadee. 2020. A two-stage text-to-emotion depressive disorder screening assistance based on contents from online community. *2020 8th International Electrical Engineering Congress (iEECON)*, pages 1–4.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and*

Information Analysis, pages 59–68, online. Association for Computational Linguistics.

Manish Rama Gopal Nadella, Venkata Krishna Rayalu Garapati, Eswar Sudhan S.k., Gouthami Jangala, Soman K.p., and Sachin Kumar. 2023. Enhancing Telugu news understanding: Comparative study of ML algorithms for category prediction. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 108–115, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Renáta Németh, Domonkos Sik, and Fanni Máté. 2020. Machine learning of concepts hard even for humans: The case of online depression forums. *International Journal of Qualitative Methods*, 19:160940692094933.

Ananya Prakash, Kanika Agarwal, Shashank Shekhar, Tarun Mutreja, and Partha Sarathi Chakraborty. 2021. An ensemble learning approach for the detection of depression and mental illness over twitter data.

Víctor Prieto, Sérgio Matos, Manuel Alvarez, Fidel Cacheda, and José Oliveira. 2014. Twitter: A good place to detect health conditions. *PloS one*, 9:e86191.

Jürgen Rehm and Kevin D. Shield. 2019. Global burden of disease and the impact of mental and addictive disorders. *Current Psychiatry Reports*, 21.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Koustuv Saha, Asra Yousuf, Ryan L. Boyd, James W. Pennebaker, and Munmun De Choudhury. 2022. Social media discussions predict mental health consultations on college campuses. *Scientific reports*, 12:123.

Ivan Sekulic, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on Reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Ivan Sekulic and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327, Hong Kong, China. Association for Computational Linguistics.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10.

M. L. Tlachac, Ermal Toto, and Elke A. Rundensteiner. 2019. You're making me depressed: Leveraging texts from contact subsets to predict depression. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4.

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909, Online. Association for Computational Linguistics.

Prashant Verma, Kapil Sharma, and Gurjit Walia. 2021. *Depression Detection Among Social Media Users Using Machine Learning*, pages 865–874.

Kirsten Windfuhr and Navneet Kapur. 2011. Suicide and mental illness: A clinical review of 15 years findings from the uk national confidential inquiry into suicide. *British medical bulletin*, 100:101–21.

Fattane Zarrinkalam, Guangyuan Piao, Stefano Faralli, and Ebrahim Bagheri. 2020. Mining user interests from social media. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3519–3520, New York, NY, USA. Association for Computing Machinery.

Bernice Yeow Ziwei and Hui Na Chua. 2019. An application for classifying depression in tweets. In *Proceedings of the 2nd International Conference on Computing and Big Data*, ICCBD 2019, page 37–41, New York, NY, USA. Association for Computing Machinery.