Journal of Universal Computer Science, vol. 30, no. 13 (2024), 1872-1904
submitted: 30/1/2024, accepted: 25/6/2024, appeared: 28/12/2024 CC BY-ND 4.0

Detecting Suicidality from Reddit Posts Using a Hybrid CNN - LSTM Model

Seyedeh Aridis Ahadi
(Department of Information Technologies, School of Applied Sciences, Cyprus International University, Via Mersin 10, Nicosia
https://orcid.org/0009-0005-2108-4340, aridis.ahadi@gmail.com)

Kian Jazayeri
(Department of Management Information Systems, School of Applied Sciences, Cyprus International University, Via Mersin 10, Nicosia
https://orcid.org/0000-0003-2843-7354, kjazayeri@ciu.edu.tr)

Sahand Tebyani
(Department of Management Information Systems, School of Applied Sciences, Cyprus International University, Via Mersin 10, Nicosia
https://orcid.org/0009-0008-8183-5494, sahand.tebyani@yahoo.com)

Abstract: The identification of individuals who indicate suicidal behaviors on social media platforms has become more significant in recent years. The utilization of textual data may help in the development of systems aimed at predicting individuals' mental health. This article proposes an integrated framework for the identification of suicidal thoughts in social media through the implementation of a layered classifier model consisting of a convolutional neural network (CNN) and a long short-term memory (LSTM) model. Various combinations of embedding techniques, activation functions, and solver algorithms are applied to the network. The mixture of these techniques forms 82 distinct methodologies employed, followed by comparing the results obtained. A collection of approximately 60,0000 user posts from 2018 to 2020 was compiled from Reddit for the study. It has resulted in the combination of TF-IDF (word embedding), RReLU (activation function), and Adam (solver algorithm) reaching the highest overall performance. The model achieved impressive accuracy, F1 Score, and AUC of 86%, with precision and recall score of 91% and 82% respectively. It was fitted in just 8.69 seconds, demonstrating its time efficiency as well. This approach has great potential for creating a platform in real life to not only reduce the social impacts of suicidality and mental illness, but also increase social access to mental health resources for all individuals.

Keywords: Adam, Activation Function, Bagging, Classification, CNN, LSTM, Machine Learning, Mental Health Disorder, Neural Network, NLP, Reddit, RNN, RReLU, Social Media, Stacking, Suicidality, Suicide Detection, Word Embedding
Categories: I.7, J.4
DOI: 10.3897/jucs.119828

1 Introduction

The timely detection of suicidal thoughts in individuals experiencing depression facilitates the provision of critical medical intervention and support, ultimately leading to the preservation of lives. Numerous studies have been undertaken to clarify the

impact of social media on the manifestation of suicidal ideation [Haque, 21; Tadesse, 19]. A clear and direct relationship exists between using words and the meanings they express. Automating diagnosis with precision poses challenges due to the inherent difficulty in acquiring data in suitable formats, resulting in limited datasets [Walsh, 11]. An increasing number of academics are employing online forums for diagnostic purposes, specifically, neural network-based methodologies requiring large datasets for efficient training [Haque, 21]. Recently, Reddit has emerged as a notable platform for identifying mental health concerns. To safeguard privacy and ensure anonymity, Reddit allows users to establish secondary, disposable accounts [Shen, 17]. This phenomenon promotes sharing and facilitates individuals with limited social connections to access internet support [De Choudhury, 14]. Many researchers have applied machine learning (ML) to detect mental illness via social media. These research studies focus on constructing and evaluating ML models, generally using artificial neural networks (ANNs) to mimic the human brain's ability to find patterns in incoming data. The utilization of social media data analysis, when combined with artificial intelligence (AI) and ML methodologies, enables the prediction of an individual's inclination toward suicide. Incorporating deep learning, feature engineering, and sentiment analysis in social content identification is paramount [Ji, 20].

As more people use social media to discuss and seek mental health treatment, academics are using natural language processing (NLP) and ML to aid. NLP processes and interprets text data using ML and deep learning for language comprehension, production, and information retrieval [Ameer, 22]. The NLP system uses algorithms like word embedding to interpret text. Word embedding in NLP captures semantic links and contextual data by representing words as dense vectors in a continuous vector space. It helps NLP tasks by allowing machines to understand and use words in context and meaning [Pennington, 14].

CNNs are typically employed to evaluate image data but can also interpret text. Text CNNs use 1-dimensional convolutions to find regional patterns and attributes in text sequences. These new traits are used for named entity identification, sentiment analysis, and text categorization. Text CNNs are important for NLP because they automatically extract relevant information from sequential input, such as text [Kim, 20]. NLP, speech recognition, and time series forecasting can benefit from Long Short-Term Memory (LSTM) because it can store and retrieve information over long sequences and capture dependencies and patterns in time-series data [Friedman, 01]. Some studies used CNNs and LSTMs with word embeddings to determine suicidality [Tadesse, 19]. ML ensembles integrate models to improve prediction. Bagging and stacking ensembles receive the CNN-LSTM model to enhance the method.

In this article, a hybrid neural network was trained and evaluated using several word embeddings, activation functions, and optimization algorithm mixes, allowing different fitting approaches. The models were validated using the ROC curve (Receiver Operating Characteristic), iterations learning curve, and train size validation curve. Various time and performance measurements were used to compare and sort valid models.

The following computer specifications and software tools were employed to achieve the research purposes. "Python" coding language in "Visual Studio 2022" has been implemented on an "ASUS TUF gaming" laptop equipped with Windows 11 Pro 2022, 11th-gen Intel(R) Core (TM) i7 processor, and 16GB of Installed RAM.

1.1 Importance of Study

Many experts have studied suicide detection due to rising suicide rates [Venek, 17], including clinical approaches like patient-clinician contact and automated detection using user-generated information, mostly text. People increasingly express their emotions, grief, and suicidal ideas online. Internet platforms now monitor users for self-harm thoughts, and social media data mining can prevent suicide [Ji, 18]. AI and ML can predict a person's suicide risk based on social media data, letting us understand motivations and intervene early. Social content recognition relies on deep learning, feature engineering, and sentiment analysis. Heuristics are needed to choose features or develop ANN architectures for rich representation.

The motivation behind the work is:

1. The issue of suicide has garnered significant attention throughout society, and the number of people who commit suicide is increasing daily.
2. More people are driven to communicate their emotions on social media.
3. Using ML and NLP techniques makes it feasible to create a hybrid deep learning network inspired by studies on suicidality detection.
4. Detecting suicidality in the early stages from the expressions published on social media helps save lives and improves society's mental health.

The aim of this study is:

1. To apply effective embedding and a hybrid model for detecting suicidality.
2. To identify mental health status from social media posts and comments.
3. Comparison of multiple embedding, activation, and optimizing methods.
4. To introduce multiple accurate and efficient hybrid classifiers.

2 Literature Review

Globally, an estimated 450 million individuals are experiencing various mental health challenges, such as depression, schizophrenia, attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), and other related conditions [Marcus, 12, Ameer, 22]. According to Tadesse, the suicide rate is 10.5 per 100,000 individuals, making it the second leading cause of mortality among young individuals. It is believed that a multitude of intricate elements influence suicide [Tadesse, 19]. While it is true that individuals with depression have a higher risk of suicide, it is worth noting that even those without mental health issues can provide a broad overview of the proposed framework [Ji, 20; Vioules, 18]. Early mental health detection is essential for understanding and treating mental health difficulties [Hamilton, 67]. Clinical psychology research links unhappy people's language to their words [Pataky, 20; Li, 18].

In this era, people can easily share their daily acts, experiences, hopes, emotions, etc., online and generate a lot of data and news thanks to the internet's rapid growth.

Sharing text, audio, video, and photographs affects social media users' thoughts and actions, improving or worsening their mental health. How people feel is intimately tied to how they use and communicate words. Textual data can be utilized to predict mental health. People may also express their opinions on social media due to limited social connections. Many people share their views and seek support [Murarka, 21]. Numerous studies have shown that social media feeds can detect depression and other mental health concerns [Walsh, 11]. Early detection of suicide ideation in depressed patients and medical treatment may save their lives [Tadesse, 19].

2.1 Dataset

Social media users' thoughts and postings must be mined and analyzed using advanced data analytic tools and AI. Many researchers have used ML to diagnose mental illness via social media. AI and ML can be used to create and construct clever automatic early detection tools. These research studies mainly focus on AI and ML model construction and evaluation [De Ocampo, 18]. ML has grown fast in computer science and beyond in the past decade. Manually creating them is often appealing [Manyika, 11; Domingos, 12]. In ML, classification problems can be grouped as binary, multi-labeled, or hierarchical [Sokolova, 2006]. Several studies have examined AI and ML in straightforward, early experiments.

On the other hand, some complex advanced ones include feature selection, feature-generating algorithms, and ML or deep learning processing [Low, 20; Zhou, 20]. Deep neural networks have more than three layers, including input and output. Simple neural networks usually have three layers or less [Hastie, 09; Krogh, 08]. Labeled datasets, or supervised learning, can guide deep ML algorithms but are not required. This decreases human involvement and allows for more extensive data collection. Neural networks, a subfield of ML, include deep learning [Hastie, 9; Rumelhart, 86].

Recently, social media users have sought mental health help. Researchers have employed data and NLP/ML techniques to help needy individuals [Ameer, 22]. Language patterns in most languages must be identified using NLP methods such as word representations [Zhang, 22]. The field of NLP has a well-established background in acquiring continuous word representations [Rumelhart, 88]. Various techniques in NLP have been used to extract relevant elements from data obtained from social media platforms. Reddit data was used in multiple investigations by Pirina and Çöltekin. According to Pirina and Çöltekin, the Reddit source is the best setting for achieving high-quality results [Pirina, 18]. Shen & Rudzicz, who studied anxiety disorders, preferred Reddit for research on many themes. They found combinational models useful [Shen, 17].

2.2 Model

In ML, it is essential to construct feature representations, as variables that cannot discriminate might result in suboptimal and incorrect model performance. [Tay, 22; Mikolov, 13]. The unstructured data derived from social media platforms necessitates preprocessing before its utilization as features and attributes in classification models. To get information from the text, Symeonidis used methods like tokenization, stop word removal, negation word detection, elongation word correction, and part of speech (POS) lemmatization [Symeonidis, 18].

Data splitting is a common practice in data science and machine learning, wherein the provided dataset is divided into many subsets. This division enables a model's training, testing, and evaluation [Hastie, 09]. In NLP, a word embedding is a way to stand for a word. The embedding is used for text analysis. Word meanings are usually stored in the form of a real-valued vector. Words close to each other in the vector space are thought to have similar meanings [Jurafsky, 00].

In some data mining applications, deep feedforward neural networks outperform classical ML models [Amjad, 20; Amjad, 21]. Adding external knowledge bases and a suicide-related ontology to a CNN model's text representation enhanced performance [Gaur, 19]. Coppersmith used a self-attention mechanism, bidirectional LSTM sequence encoding, and deep learning word embedding to capture the most informative subsequence [Coppersmith, 18]. Ji recommended CNN and LSTM model aggregation to detect suicidal ideation in private chat rooms. However, decentralized training uses chat room coordinators to designate user posts for supervised training, which is only helpful in simple circumstances [Ji, 19].

According to Glorot, employing a nonlinear activation function is necessary to address nontrivial scenarios effectively [Glorot, 11]. The utilization of optimization algorithms has become imperative across several problem-solving domains. The parameters involved and the nature of the problem under consideration impact the choice of an appropriate optimization technique [Rumelhart, 86].

Combining multiple individual models, called base estimators, is called ensemble learning. This idea supports the "wisdom of crowds" theory that a larger group's judgment is frequently better than a single expert's [Zhou, 02]. Bagging, an ensemble learning method, randomly samples and replaces a training set's data, allowing numerous selections of the same data points [Ha, 05]. Stacking generalization, or piling in machine learning, uses all aggregated models' weights to generate a new model [Sridhar, 96; Wolpert, 92].

2.3 Evaluation

The fitting of an ML model indicates its ability to generalize effectively to data similar to the data it was trained on. The precision of the results is enhanced when utilizing a well-fitted model. According to LeCun, an underfitted model must adequately align with the available data. An overfitted model displays a high degree of resemblance to the data [LeCun, 98].

Calculating ML model run time improves efficiency and resource management, especially for large datasets and sophisticated models. Tang compared CNN model run time to different methods and examined how embedding strategies affect total run time [Tang, 15]. Time spent building word embeddings can affect NLP model training. More effective word embedding methods reduce training time, speeding model development [Getzen, 22]. Model training efficiency and iteration speed for best performance and deployment depend on neural network fitting time. Hessam Karimi evaluated neural network training time with multiple optimizers to find the optimal model setup [Karim, 18].

It is often advantageous to assess an algorithm's efficacy with numeric measures. This sometimes involves a comparative analysis to determine the most suitable algorithm for a specific application [Demšar, 06]. According to Sokolova and Lapalme, the commonly employed performance metrics for categorization problems include the

confusion matrix, accuracy, precision, recall, F1 score, and AUC (Area Under the Curve) [Sokolova, 09].

The ROC curve, a graphical tool, evaluates and visualizes a binary classification model like an ML algorithm or diagnostic test. It is notably helpful in imbalanced datasets when one class (typically the minority class) is much less frequent than the other. Learning curves are considered a valuable tool in ML and data analysis because they provide insights into a model's training process and identify potential issues such as overfitting, underfitting, and convergence challenges. A validation curve is a visual depiction illustrating the performance of a model over time or as the amount of training data increases for a given task or problem [Sokolova, 09].

2.4 Related Works

Society's suicide danger has increased dramatically in the past decade. Suicide is serious; hence, researchers have developed suicidality detection algorithms using social media user activity data to find latent suicide indicators early on [Lee, 22]. The prompt identification of suicide ideation in depressed patients allows for lifesaving treatment [Haque, 21]. Many recent studies have shown that social media affects suicide ideation [Tadesse, 19].

Reddit is very useful for mental health diagnosis [Shen, 17]. It allows users to create subreddits or groups for specific causes, discuss mental health issues, and give instances. Creating secondary throwaway accounts is another option to safeguard privacy and anonymity and help those without large social networks get online help [De Choudhury, 14]. Large-scale computer analysis of mental health issues is now possible due to the large user base, honesty of online environments, and regulated post-screening to ensure authenticity [Haque, 21]. The dataset contains user comments as text features and suicidality as labels.

Many studies have utilized machine learning to detect mental illness via social media. These researches focus on constructing and analyzing AI/ML models using Artificial Neural Networks (ANNs), which imitate the brain's ability to find patterns in incoming data. Researchers use NLP and machine learning to support mental health patients on social media [Ameer, 22; De Ocampo, 18]. AI and ML can predict suicide risk using social media data. This allows for early intervention and helps understand motives. Deep learning, feature engineering, and sentiment analysis are essential for social content recognition. Heuristics are used to choose features or design structures for ANNs to learn representations efficiently [Ji, 21].

3 Methodology

The proposed framework consists of four sequential processes, namely "Initiation," "Reddit Records", "Network Architecture", and "Outputs Analysis". The framework, as shown in Figure 1, is structured into distinct parts, each consisting of sub-steps that collectively contribute to the primary objective of identifying the presence of suicidal tendencies in user-posted texts on social media.

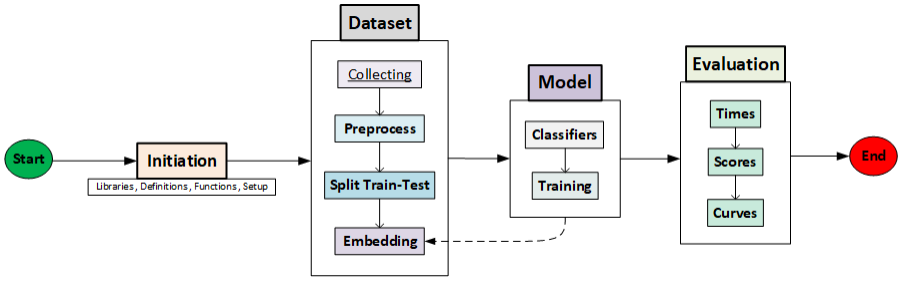


Figure 1: A broad overview of our suggested structure

3.1 Initiation

The initiation step determines model variable values (hyperparameters setup and input options) and defines libraries, paths, functions, methods, ensembles, punctuation, and other needed initial data. Selected methods are the combination of two embeddings (Count and TF-IDF), four activation methods (RReLU, Tanh, Mish, and ELU), and five solver algorithms. The technique combines these primary methods, resulting in 40 solutions (2*4*5=40). Bagging each combination adds 40 more models, totaling 80. The stacked classifier receives all estimators for each embedding; therefore, 82 combined approaches are employed.

Parameter	Regarding			Type	Value	Proof
gather	data	collecting	{data collection or data analysis}	boolean	True,False	*adquate to stage need*
extract			{number of records to be extracted from dataset}	integer	20000	*enough and not getting slow*
lim			{limit for density of mental classified texts (one user)}	integer	0	*bigger numbers cause lower labels ratio*
veclen	embedding	vector length	{length of embedding vector}	integer	50	*more length causes excess complexity*
alpha		learning rate	{learning rate of word vectorizer}	float	0.1	*default standard of library*
ngram		n words	{word range of vectorizers}	integer	1	*more range causes underfitting*
testsize	split	train-test	{train-test split ratio}	% percent	0.2	*between .3 and .2 is acceptable*
strtfy		stratify	{stratify or not}	boolean	True,False	*insurance of train-test label balances*
cnnlyrs	networks	convolutional	{architecture of convolution layers}	float list	[.5]	*more nodes or layers causes complexity*
hdnlyrs			{architecture of hidden fully connected layer}	float list	[]	*adding hidden layers causes complexity*
kernel			{kernel size of convolution}	integer	1	*more kernel size leads to high variance in the model*
drptrglz			{dropout layer ratio}	% percent	0.1	*less ratio causes overfitting , more leads to underfitting*
lstmlyrs		recurrent	{architecture of recurrent layers}	float list	[.5]	*more nodes or layers causes complexity*
nestbag	ensemble	bagging	{number of base estimators}	integer	5	*more estimators lead to overfitting*
smpl			{samples percentage drawn for each estimator}	% percent	0.2	*more samples cause overfitting*
ftrs			{features percentage drawn for each estimator}	% percent	0.8	*less features cause underfitting*
lr_adam	training	optimizers	{learning rate for adam}	float	0.01	*manual search for fit tuning*
lr_adamax			{learning rate for adamax}	float	0.01	*manual search for fit tuning*
lr_adadelta			{learning rate for adadelta}	float	2	*manual search for fit tuning*
lr_adagrad			{learning rate for adagrad}	float	0.01	*manual search for fit tuning*
lr_rmsprop			{learning rate for rmsprop}	float	0.001	*manual search for fit tuning*
l2wghtdcy			{weight regularization}	float	0.01	*manual search for fit tuning*
clipgrad			{weight clipping}	float	0.01	*insurance of preventing overfitting*
nptnc		early stopping	{max allowable number of early stopping counter}	integer	1	*more patience causes overfitting*
patience			{early stopping sensitivity}	float	0.0001	*the scale is synchronized to loss value*
tst_szs	validations	train size	{for plotting curve}	% percent list	[.99:.19]	*chosen set makes a smooth validation curve*

Table 1: Parameters' values

General definitions (output paths, applied methods, ensembles, punctuation, and numbers for preprocessing), variable values of the model (hyperparameters setup and input options), and functions (preprocesses such as lemmatizing, stemming, and stop-words; balancing, loss, CNN-LSTM network) will be explained along this part. Table 1 includes values of variables such as sizes, ratios, architecture, limits, and other hyperparameters. The variables regarding the architecture of the network are shown as lists. Each element of the lists is the proportion of the layer nodes' number to the embedding layers' length (first layer of the network).

The implemented libraries for executing different methods are listed in Table 2.

	Artificial Neural Network						**Classification Models**						**Word Embedding**		**Activation Function**				**Solver Algorithm**					**Model Validation**		
	Preprocess	Splitting	Embedding	Training	Times	Scores	CNN	LSTM	Bagging	Majority Voting	Stacking	Logistic Regression	Count	Tfidf	Elu	RRelu	Tanh	Mish	Adam	Adadelta	Adagrad	Adamax	RMSprop	ROC Curve	Learning Curve	Validation Curve
nltk	X																									
sklearn		X	X	X		X			X	X	X	X	X	X										X		
torch				X			X	X							X	X	X	X	X	X	X	X	X		X	X
time					X																					
plot																								X	X	X

Table 2: Implementation of libraries and their uses

We used the "Time" library to obtain duration measurement by calculating each process's beginning and ending. Measurements of total run time and model training speed can assist in uncovering more efficient strategies. The vocabulary fitting time can also be used to compare embedding vectorizers' speed. These numbers can be combined alongside other performance scores to evaluate models.

3.2 Reddit Records Preparation

This research describes NLP and text-to-classify methods to implement a deep learning classifier to improve language modeling and text classification for Reddit social media suicidal ideation detection. This technique uses a July 13, 2020, Reddit Mental Health Dataset[1] with postings from 28 subreddits (including 15 mental health support groups) during 2018-2020. In total, 59,996 user texts were composed for analysis.

To prepare the dataset, we first convert the suicidality grade of the raw data to binary labels of 0 or 1, considering any non-zero grade as label 1. Then, all comments of one person are grouped as corporate text samples, and if any of their comments have label 1, their merge comment is labeled 1. This stage must be executed just once per dataset. Then, 20,000 user texts were randomly selected from the dataset to study.

1 This dataset is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/ (Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (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.) DOI: 10.17605/OSF.IO/7PEYQ

Data preprocessing evaluates, filters, manipulates, and encodes data so a ML algorithm can understand and use the output; this stage consists of transforming the words to lowercase and original form (stemming and lemmatizing) and continued by eliminating numbers, punctuations, emojis, stop words, URLs, and mentions. Table 3 shows some random preprocessing examples from the dataset.

Reddit User Text	Preprocessed Data
idk Is it possible to die if you take a bunch of xannies and mix it with alcohol?	idk possibl die take bunch xanni mix alcohol
will i die 160mg of valium @ 2:07pm 612.5 mg of promethazine 2:33pm 0.1 mg clonodine around same time	will die mg valium pm mg promethazin pm mg clonodin around time
i keep running suicide plans through my head constantly thinking and fantasizing of the perfect ways i'd wanna go.	keep run suicid plan head constantli think fantas perfect way id wanna go
It's my birthday It's my birthday and I want to kill myself. I've never had a good birthday in my life. I've been depressed and suicidial for the past 3 years. No one cares about and no one would care if I was dead.	birthday birthday want kill ive never good birthday life ive depress suicidi past year one care one care dead
they dont know i'm tired of pretending that everything is okay with me,i wake up everyday hating myself,i want to feel relief from this painful existence almost everyday before i sleep i think about how i'd kill myself,and the stress from it is maddening sorry for bad english.	dont know im tire pretend everyth okay mei wake everyday hat myselfi want feel relief pain exist almost everyday sleep think id kill myselfand stress madden sorri bad english

Table 3: Random preprocessing examples

The dataset is divided into training and testing subgroups. Although the label ratio is 53%, stratifying is employed at the splitting stage to ensure label balance. Statisticians use the term "stratify" to describe a data sampling or division process that keeps the same or nearly the exact proportions of features or categories as the original dataset. This study uses 80% of the data for training and 20% for test data.

Word embedding is a numerical vector representing a word within a space of reduced dimensions. N-grams are contiguous clusters of n text or speech components. The "learning rate" is a hyperparameter used in neural network-based NLP model training to determine how fast the model updates and moderates word vectors. The n-gram of word vectorizing in this study is set to 1. This text embedding uses a 0.1 learning rate and 50 vector length.

3.3 Network Architecture

Following tests with different mixes of CNN, RNN, and LSTM models, this study found that the CNN-LSTM combination worked best on the dataset regarding both time and accuracy. This combinational network's architectural characteristics and qualities are explained within the subsequent discourse.

The convolution layer is followed by activation, pooling, and normalization layers, forming a CNN of 25 nodes, or half of the initial 50 input features. Kernels are learnable or convolutional filters that extract features from one-dimensional sequences or time-series data. The kernel size for this CNN is set at 1. The network does not incorporate a fully linked hidden layer to mitigate excessive complexity that may result in overfitting. Subsequently, a dropout regularization technique is implemented on the output. Dropout regularization is a technique employed in neural networks to mitigate the occurrence of complex co-adaptations during the training process. Its primary objective is to address the issue of overfitting, a phenomenon wherein the model becomes too specialized for the training data.

The CNN output is subsequently transmitted to the LSTM network, along with its corresponding activation function. The LSTM network consists of 25 features, equivalent to the number of features in the CNN. Finally, a normalized, fully connected layer establishes a connection between the network and the output classification layer.

Adding activation functions to the network makes it possible to include nonlinearity, which makes simulating complex data relationships easier. Each neuron in a neural network receives a summation of its weighted inputs, and this activation function determines the neuron's output or level of activity. Activation functions play a crucial role in neural network topologies as they significantly influence the processing and transmission of information. RReLU, ELU, Tanh, and MISH are the activation functions selected for this study.

Optimization algorithms have become indispensable for a wide range of problem-solving scenarios. Each algorithm employs a distinct approach to identify the best solution for an optimization problem. The selection of the most suitable solver for a particular issue is contingent upon several factors, encompassing the parameters involved and the nature of the problem under consideration. The utilization of a hyperparameter known as the learning rate is prevalent in the training process of ML models, particularly in optimization strategies that rely on gradients. The purpose of this mechanism is to control the scale of adjustments made to the parameters of a model, including its weights and biases, throughout the training process. Adam, Adamax, Adadelta, Adagrad, and RMSprop are the solver algorithms that are applied in this paper.

Ensemble learning, a technique employed in the field of ML, involves combining multiple models, referred to as base estimators, to get predictions that are more accurate and dependable. The individual base learners are trained on distinct network models, subsets of the training data, or algorithms and are subsequently aggregated to generate a final prediction.

Bagging uses multiple copies of the same model trained on different portions of the training data to create a more accurate forecast. Each model considers different subsets of samples and attributes, making the ensemble more resilient to samples or features. Using "max samples" and "max features" adds unpredictability and variation

to ensemble models. For this investigation, the maximum sample value is set to 0.2, and the maximum feature value is 0.8. Barging classifiers use five base estimators.

Stacking ensembles are commonly utilized in ML to enhance a model's overall performance and accuracy by leveraging the benefits offered by many base models. This meta-model methodology aims to optimize accuracy and performance by aggregating predictions generated by the underlying base models. The base models and the meta-model are trained on numerous subsets of the original dataset to introduce diversity in the predictions.

To prevent overfitting, various regularization techniques are applied; these techniques include L2 weight decay, weight clipping, and early stopping in the optimization stage. Adding a dropout layer to the convolutional network helps improve the model's generalization. The regularization technique known as L2 weight decay involves the addition of a penalty word to the loss function. This penalty term aims to encourage the model to have less weight. Weight clipping is a technique used to restrict weights within a specific range to address the issue of excessive growth. The dropout layer is designed to randomly deactivate a subset of nodes, boosting independence among them and minimizing the risk of overfitting. Early stopping prevents the model from overfitting and determines when it can best generalize to new data. Neural network training stops after multiple retries if the loss function value doesn't decrease significantly (patience). In this study, the l2 weight decay value was equal to 0.01 for all optimizers. The weight values above one were clipped. The patience was 0.0001, and the number of retries was set to 1. Other configurations led the model to overfit or underfit (too early or too late stopping).

3.4 Analysis Approach

The results generated by the model are analyzed from different perspectives, including:

- Running Times (vectorizing times, fitting times, total run time)
- Metric Scores (accuracy, precision, recall, F1 score, and AUC)
- Receiver Operating Characteristic (ROC curve)
- Validation Curves (iterations learning curve, train size validation, lost curves)

The evaluation of an ensemble model necessitates training and assessment of several models, encompassing both individual base estimators and the resulting ensembles. Using advanced base estimators or a large-scale dataset may pose significant computational challenges. As a result, the ensemble only uses the fundamental estimators to validate its predictions. When computational resources are limited, evaluating base estimators can still provide valuable information and help design an effective ensemble model.

4 Findings and Discussion

Utilizing appropriate duration measures, performance scores, diverse validation approaches, and curves obtains insights into the models' speed and accuracy.

4.1 Total Run Time

Scanning and analyzing the dataset utilizing the combined methods mentioned earlier required a cumulative duration of 67 minutes and 54 seconds. This included 35 minutes

and 18 seconds dedicated to the model fitting, 30 minutes and 24 seconds allocated for validation, and the remaining 2 minutes and 12 seconds designated for various supporting processes (Table 4).

Total Run Time (min)	Fitting Time (min)	Validation Time (min)	Rest of the Process (min)
67.9	35.3	30.4	2.2

Table 4: Total run time

4.2 Vectorizing Time

The quantity and complexity of the text data and the preparation procedures used frequently affect the runtime of embedding methods. When selecting different methodologies for a specific text analysis assignment, it is essential to consider their applicability and the quality of embeddings they produce. Table 5 displays the duration required for vectorizing word embedding techniques when applied to a 20,000-user text entries dataset.

Embedding	Run Time (s)
TF-IDF	3.66
Count	2.88

Table 5: Embedding run time

4.3 Fitting Time

The fitting durations range from 0.33 to 63.92 seconds. The stacking model with count embeddings has the fastest fitting time. Forty hybrid models with several activation functions and solver techniques were utilized in this stacked estimator. The bagging model with count embedding, Mish activation function, and Adagrad solver had the most nondesirable fitting time.

Stacking models require little classifier-level training or learning; therefore, a pre-fitted stacking classifier can be changed quickly. A meta-classifier integrates the predictions of numerous base classifiers trained on the same dataset in a stacked ensemble. Base estimators are often already trained on the dataset, so fitting doesn't require extra training. In other words, the ensemble's base learners have been fitted, ending the time-consuming training procedure. This dramatically reduces the stacked classifier's fitting time. Training the meta classifier is crucial to fitting and is frequently faster and easier than training the basic classifiers.

In addition to stacked models, six models that fit in less than 10 seconds can be made by combining Adam solver techniques with different activation functions (Elu, RReLU, and Tanh) and both of the word embeddings (count and TF-IDF). Adam's optimization works quickly and well because it sets the learning rate automatically, lets you change the learning rate while fitting, uses little memory, converges quickly, handles sparse gradients well, and is widely accepted by deep learning (Table 6).

Model	Type	Embedding	Activation Function	Solver Algorithm	Base Model	Final Estimator	Fitting Time (s)
Stacking	Ensemble	Count	Multiple	Multiple	Hybrid (40)	Logistic Regression	0.33
Stacking	Ensemble	TF-IDF	Multiple	Multiple	Hybrid (40)	Logistic Regression	0.37
CNN+LSTM	NN	TF-IDF	Tanh	Adam	.....	.....	8.48
CNN+LSTM	NN	Count	Elu	Adam	.....	.....	8.54
CNN+LSTM	NN	Count	Tanh	Adam	.....	.....	8.54
CNN+LSTM	NN	TF-IDF	RRelu	Adam	.....	.....	8.69
CNN+LSTM	NN	TF-IDF	Elu	Adam	.....	.....	8.79
CNN+LSTM	NN	Count	RRelu	Adam	.....	.....	9.37
CNN+LSTM	NN	Count	Tanh	Adamax	.....	.....	9.97

Table 6: Fastest models' fitting times

4.4 Scores

The models' accuracy ranged from 74% to 86% on the test dataset. A total of thirteen models reached an accuracy of 86% on the test dataset. More than half of these models utilized TF-IDF for embedding, RReLU, and Tanh, which are used more than other activation functions integrated with Adam Solver. TF-IDF and Adam also do well on this measure score by combining the Elu activation function. Stacking models and neural networks employing Adam Solver, Tanh, and RReLU on both embeddings have a slightly low fitting time (less than 12 seconds).

The precision of the models ranged from 78% to 91%. Fourteen models scored a precision score of 91%. These models generally employed Adam and Adamax solvers with ELU activation functions and also included a hybrid stacking estimator with two bagging ensembles (Count embedding). Six of these models fit in under 10 seconds.

The models have a 70%–85% recall score. Six models reached the minimum recall of 84%. Most top bagging ensemble models employed RReLU activation functions with Adamax and RMSProp solvers, although their fitting times exceeded 30 seconds.

Model F1 scores on the test dataset varied from 74% to 86%. Stacking with TF-IDF embedding, several activation and optimization approaches, and a CNN-LSTM model with the RReLU activation function and Adam Solver algorithm have the most excellent F1 score (86%). They both fit in under 10 seconds.

The model AUC scores on the test dataset varied from 74% to 86%. Eighteen models had AUC scores over 86%. These models used Adam and Adamax optimizers on all embeddings and activation functions equally. There are also two stacked on the count and TF-IDF embedding with multiple activation functions and solver algorithms. These high-performance models fit in within 15 seconds.

We compare network approaches by calculating average accuracy and fitting time in different regions. Table 7 shows that Count and TF-IDF embeddings perform similarly for average accuracy and model fitting time. Mish takes longer to fit than other activation functions, but its precision is the same. Tanh, however, is the fastest model activation layer. Adam has the best average accuracy and fastest fitting time, whereas Adadelta has the lowest accuracy, and Adagrad is the slowest solver. Since it improves the base model, Stacking is the fastest and most accurate classification model.

On the other hand, the Bagging classifier fits slower and with worse accuracy on the dataset.

Section	Method	Average Accuracy	Average Fitting Time (s)
Classification Model	CNN+LSTM	0.85	14.78
Classification Model	Bagging	0.83	38.20
Classification Model	Stacking	0.86	0.35
Embedding	Count	0.84	26.06
Embedding	TF-IDF	0.84	25.64
Activation Function	RRelu	0.84	24.98
Activation Function	Tanh	0.84	23.93
Activation Function	Mish	0.84	31.91
Activation Function	Elu	0.84	25.14
Solver Algorithm	Adam	0.85	17.41
Solver Algorithm	Adamax	0.84	21.72
Solver Algorithm	Adagrad	0.84	34.08
Solver Algorithm	RMSprop	0.84	32.62
Solver Algorithm	Adadelta	0.83	26.61

Table 7: Accuracy and fitting time averages

Table 8 represents that how many of the models have reached the highest performance score. Their fitting time duration is also available in the same table.

Score	Accuracy	Precision	Recall	F1 Score	AUC
Number of Models with Maximum Score	13	14	6	2	18
Duration (Seconds)	<12	<10	<30	<10	<15

Table 8: Models' performance metric scores

Table 9 shows that which of the models has reached the highest performance in different metric scores simultaneously. For example, the combination of TF-IDF – RReLU – Adam hit the high score in Accuracy, Precision, F1 Score, and AUC; in this way, it is repeated four times as the highest score model (frequency column).

Ten of the best models were selected (Table 10) based on their frequency of appearance in top performance-ranked lists. Both stacked ensembles and combined models, as detailed in Table 10, are among these models.

Solution	Frequency	Accuracy	Precision	Recall	F1 Score	AUC
TF-IDF RReLU Adam	4	x	x		x	x
TF-IDF RReLU Adamax	3	x	x			x
TF-IDF Tanh Adam	3	x	x			x
TF-IDF Tanh Adamax	3	x	x			x
TF-IDF Elu Adam	3	x	x			x
TF-IDF Elu Adamax	3	x	x			x
Count Rrelu Adamax	3	x	x			x
Count Mish Adam	3	x	x			x
Stack TF-IDF	3	x			x	x
Stack Count	3	x	x			x

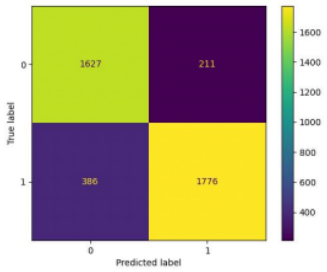
Table 9: Solution frequency in top models

Since all 82 models cannot be monitored for performance details, we continue analyzing the outputs using these ten combined method outputs. It is noticeable that stacked models implement a hybrid of 40 validated base models on each embedding, and to avoid unnecessary complexity, there is no separate validation plot for them. Although the fitting time reached a maximum of 64 seconds in the worst-case scenario, it was less than 12 seconds for these superior models.

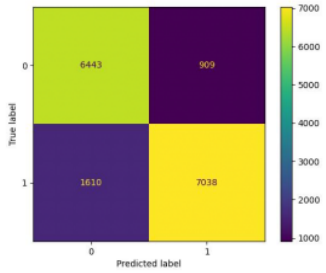
Model	Type	Embedding	Activation Function	Solver Algorithm	Base Model	Final Estimator	Accuracy	F1 Score	Precision	Recall	AUC	Fitting Time (s)
CNN+LSTM	Neural Network	TF-IDF	RRelu	Adam	.....	.....	[0.85, 0.86]	[0.86, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	8.69
CNN+LSTM	Neural Network	TF-IDF	Tanh	Adam	.....	.....	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	8.48
CNN+LSTM	Neural Network	TF-IDF	Elu	Adam	.....	.....	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	8.79
CNN+LSTM	Neural Network	TF-IDF	RRelu	Adamax	.....	.....	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	11.02
CNN+LSTM	Neural Network	TF-IDF	Tanh	Adamax	.....	.....	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.81]	[0.85, 0.86]	11.29
CNN+LSTM	Neural Network	TF-IDF	Elu	Adamax	.....	.....	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	11.15
CNN+LSTM	Neural Network	Count	Mish	Adam	.....	.....	[0.84, 0.86]	[0.85, 0.86]	[0.89, 0.91]	[0.81, 0.82]	[0.85, 0.86]	10.7
CNN+LSTM	Neural Network	Count	RRelu	Adamax	.....	.....	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	10.52
Stacking	Ensemble	TF-IDF	Multiple	Multiple	Hybrid (40)	Logistic Regression	[0.85, 0.86]	[0.86, 0.86]	[0.9, 0.9]	[0.82, 0.83]	[0.85, 0.86]	0.37
Stacking	Ensemble	Count	Multiple	Multiple	Hybrid (40)	Logistic Regression	[0.85, 0.86]	[0.85, 0.86]	[0.9, 0.91]	[0.81, 0.82]	[0.85, 0.86]	0.33

Table 10: Selected high-performance models

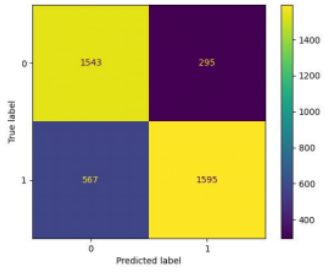
Confusion matrix is a vital ML approach for classifier evaluation that summarises classification model predictions compared to dataset labels. Although it can be adapted for multi-class situations, it is mainly used for binary classification. Various metric scores, such as Accuracy and F1 Score, can be calculated using the confusion matrix. The Confusion Matrixes of the top models are brought in Figure 2.



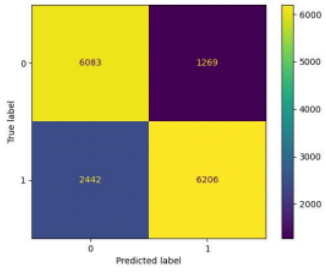
Count Mish Adam (test)



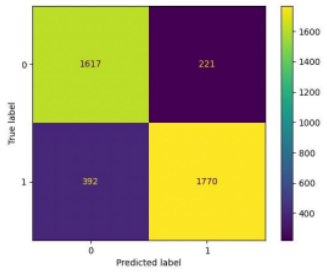
Count Mish Adam (train)



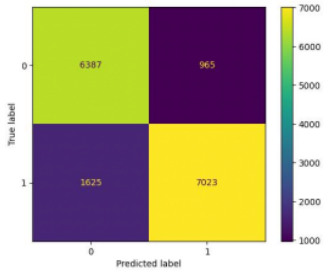
Count RReLUAdamax (test)



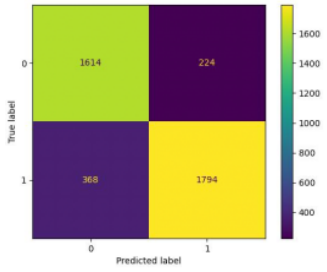
Count RReLUAdamax (train)



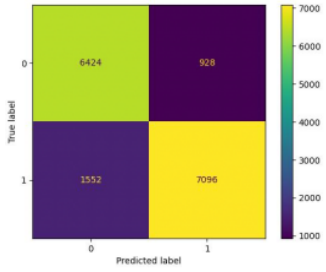
TF-IDF Elu Adam (test)



TF-IDF Elu Adam (train)



TF-IDF Elu Adamax (test)



TF-IDF Elu Adamax (train)

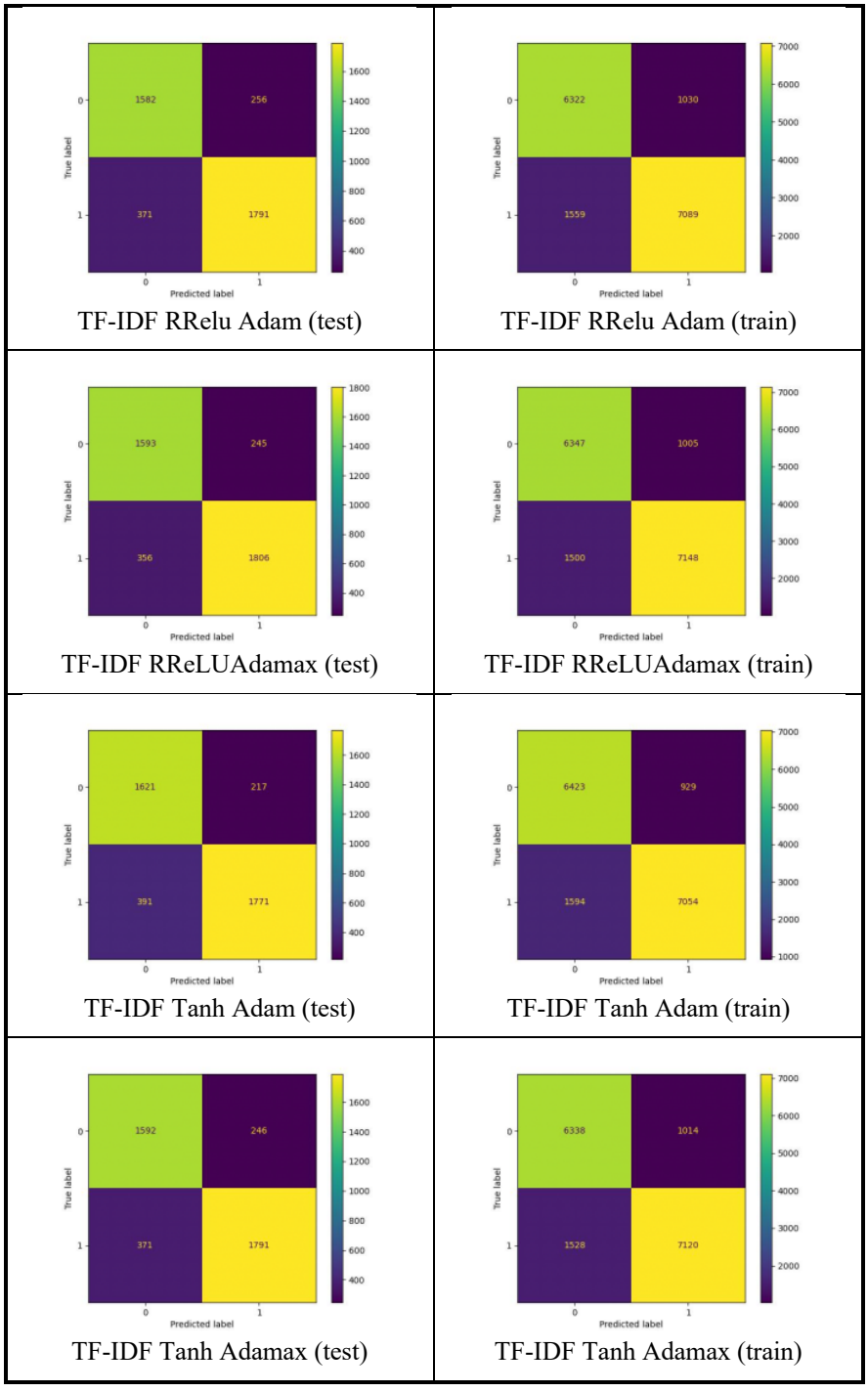


Figure 2: Top models' confusion matrix

4.5 ROC Curve

High-performance models on the test dataset had an average AUC of 86% among the top models (Figure 3).

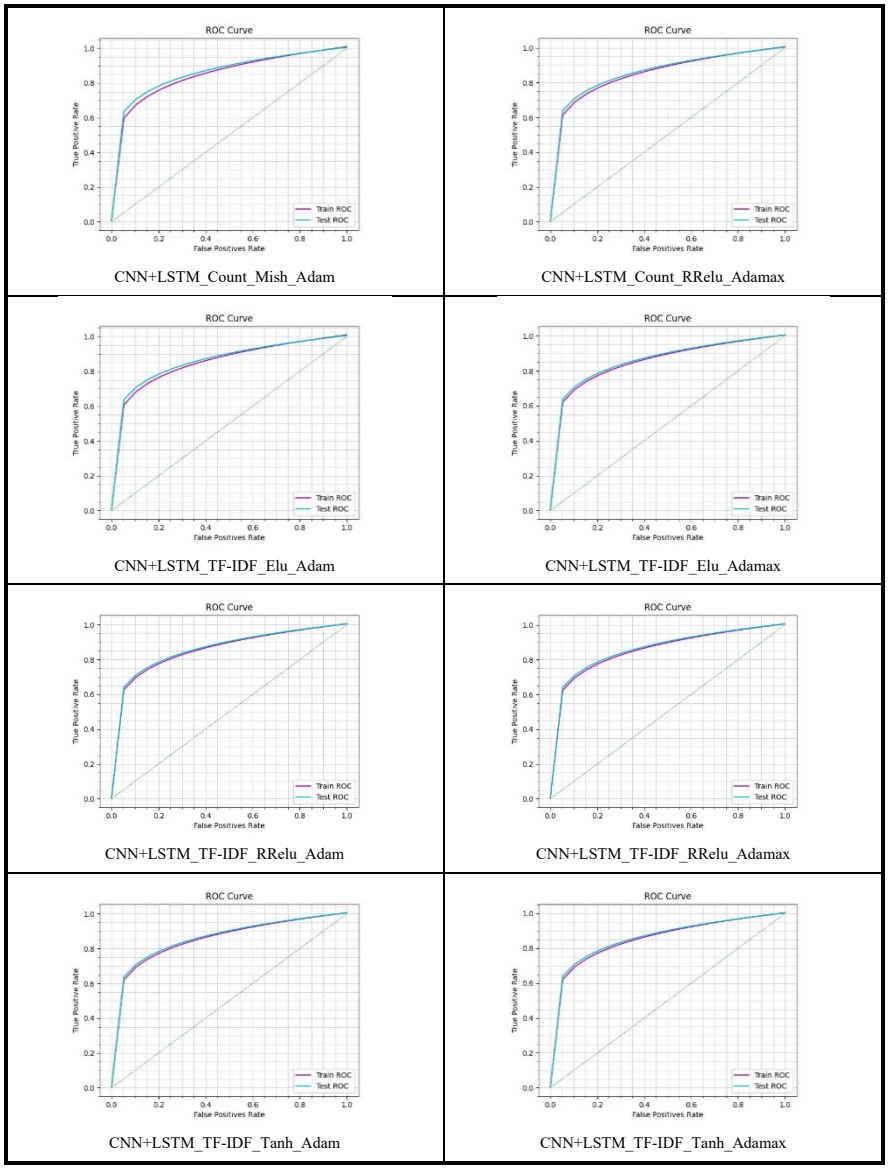


Figure 3: Top models' ROC curves

4.6 Validation Curves

Figures 4 to 10 illustrate the top performance models' iteration learning curves and train size validation curves. The shape and convergence of the curves indicate that the models are robust and generalized correctly. The average accuracy of the high-performance models on the test dataset was computed to be 86% (Figure 4).

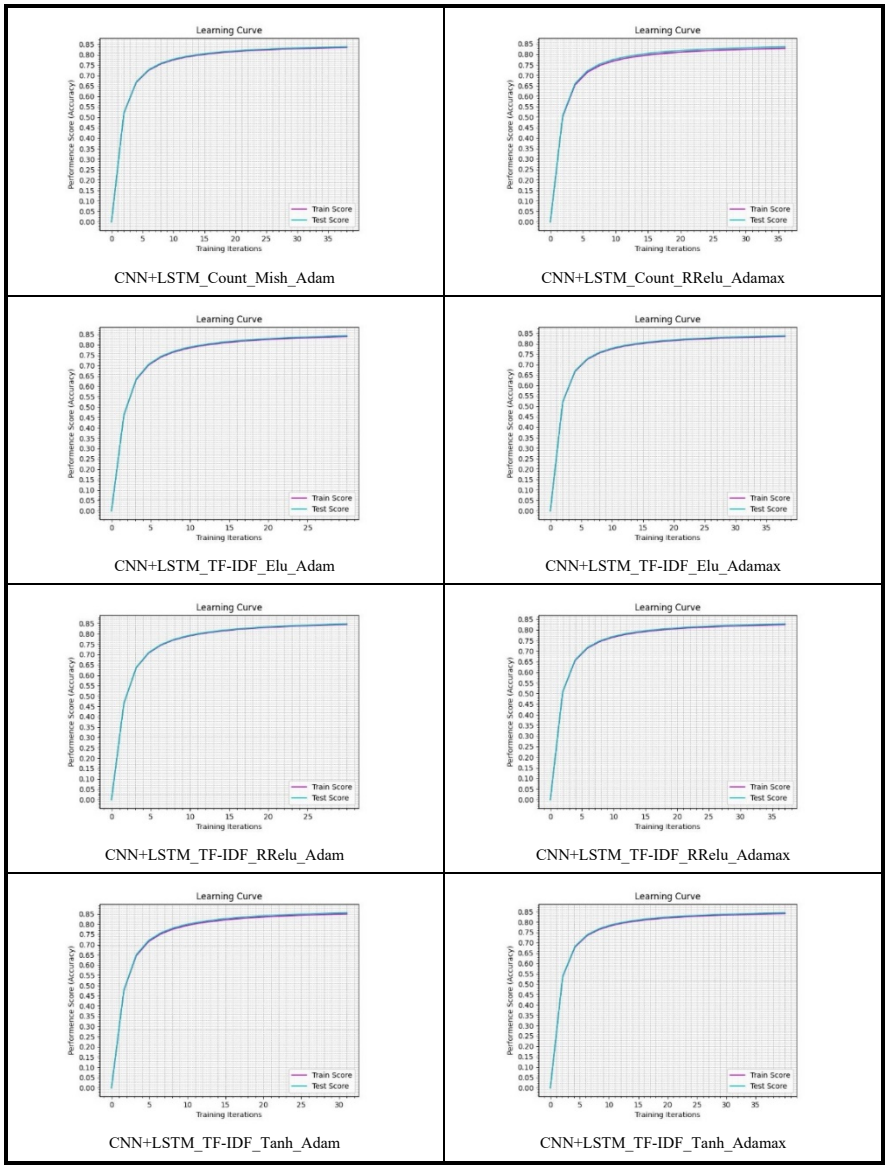


Figure 4: Accuracy learning curves

The precision of the high-performance models on the test dataset was determined to be 91% on average (Figure 5).

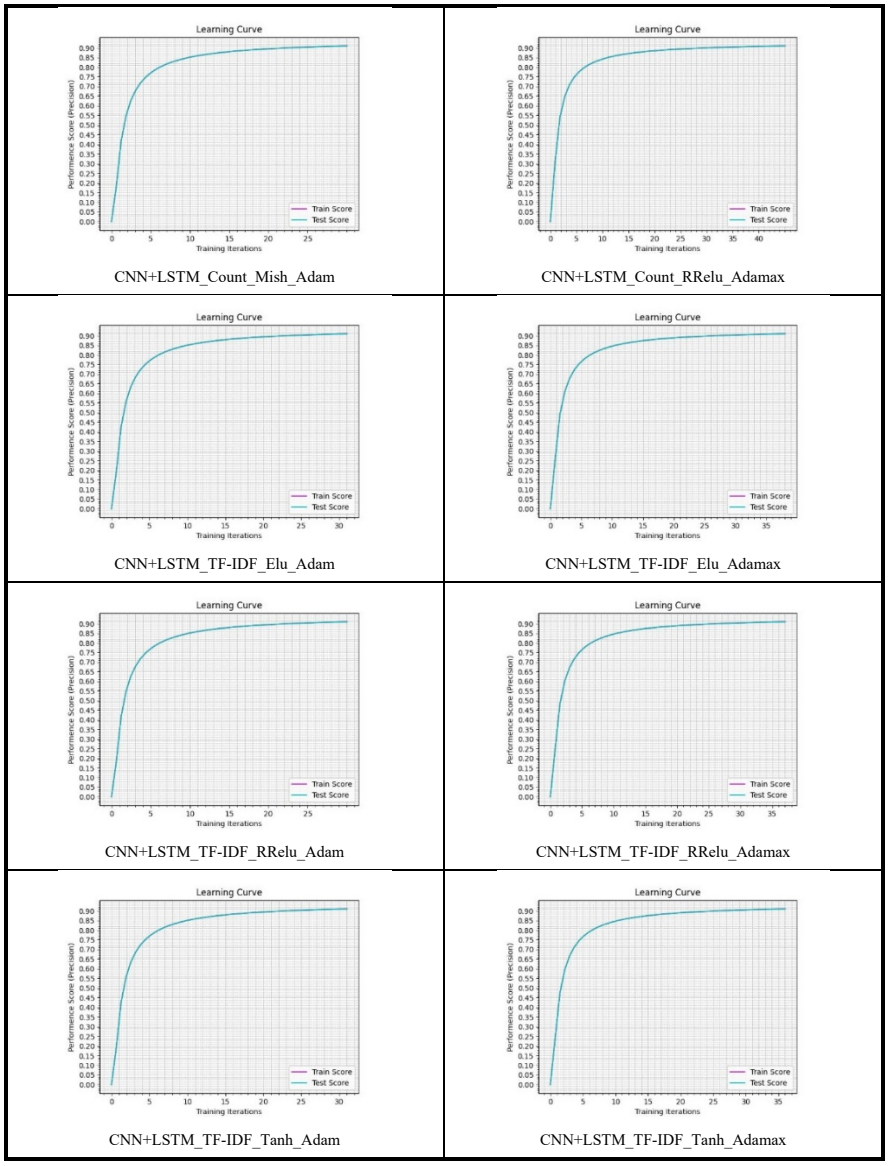


Figure 5: Precision learning curves

The recall metric for the high-performance models on the test dataset was computed to be 82% (see Figure 6).

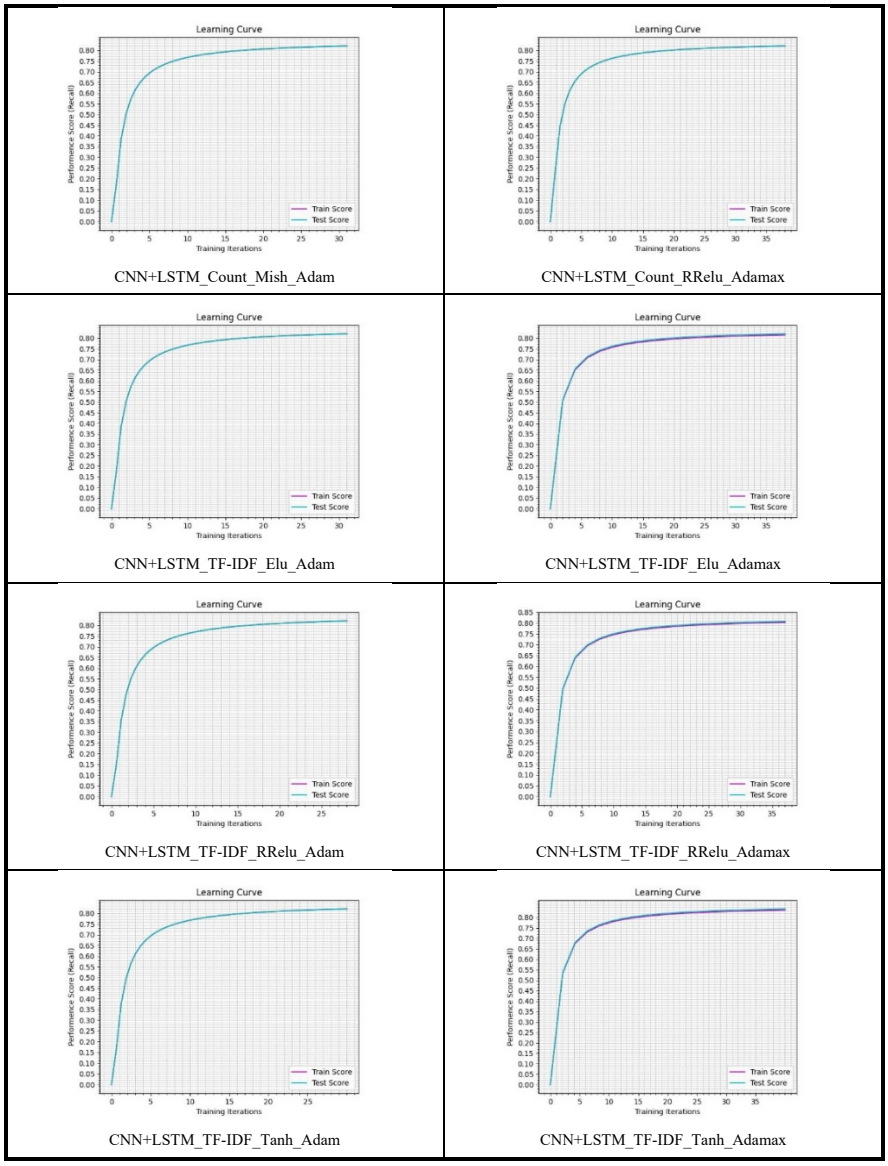


Figure 6: Recall learning curves

The high-performance models' F1 score on the test dataset was determined to be 86% on average, as shown in Figures 7.

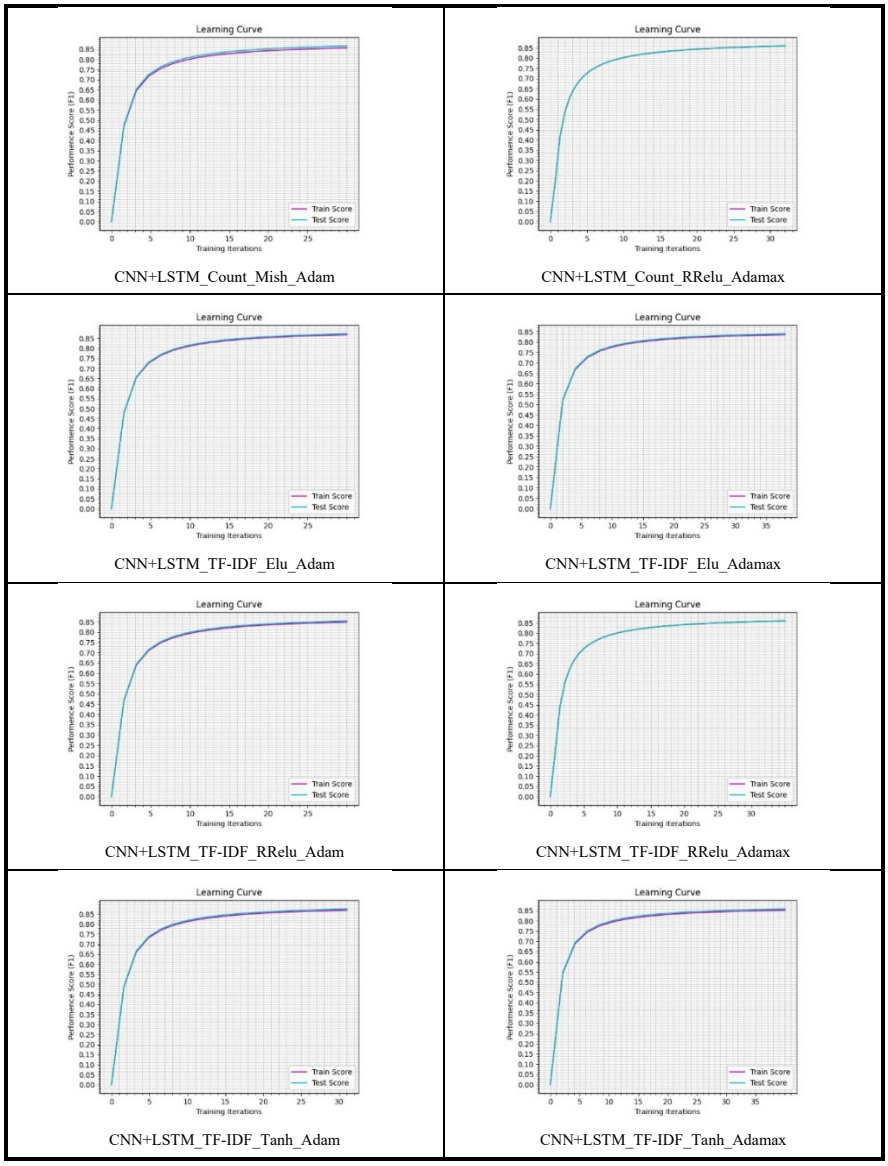


Figure 7: F1 Score learning curves

As indicated earlier, the mean area under the curve (AUC) of the high-performance model's on the test dataset was computed to be 86% (see Figures 8).

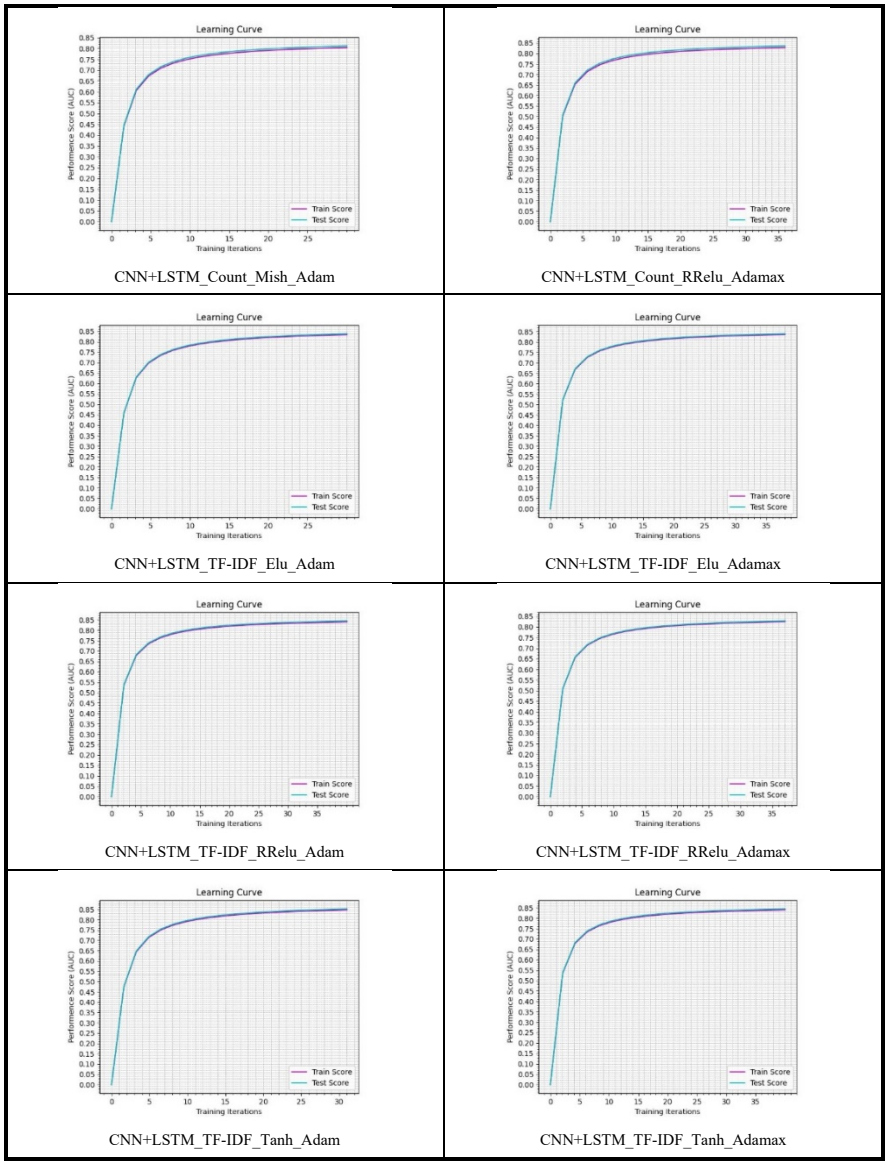


Figure 8: AUC score learning curves

Figure 9 shows the train size validation curves converging to the accuracy score.

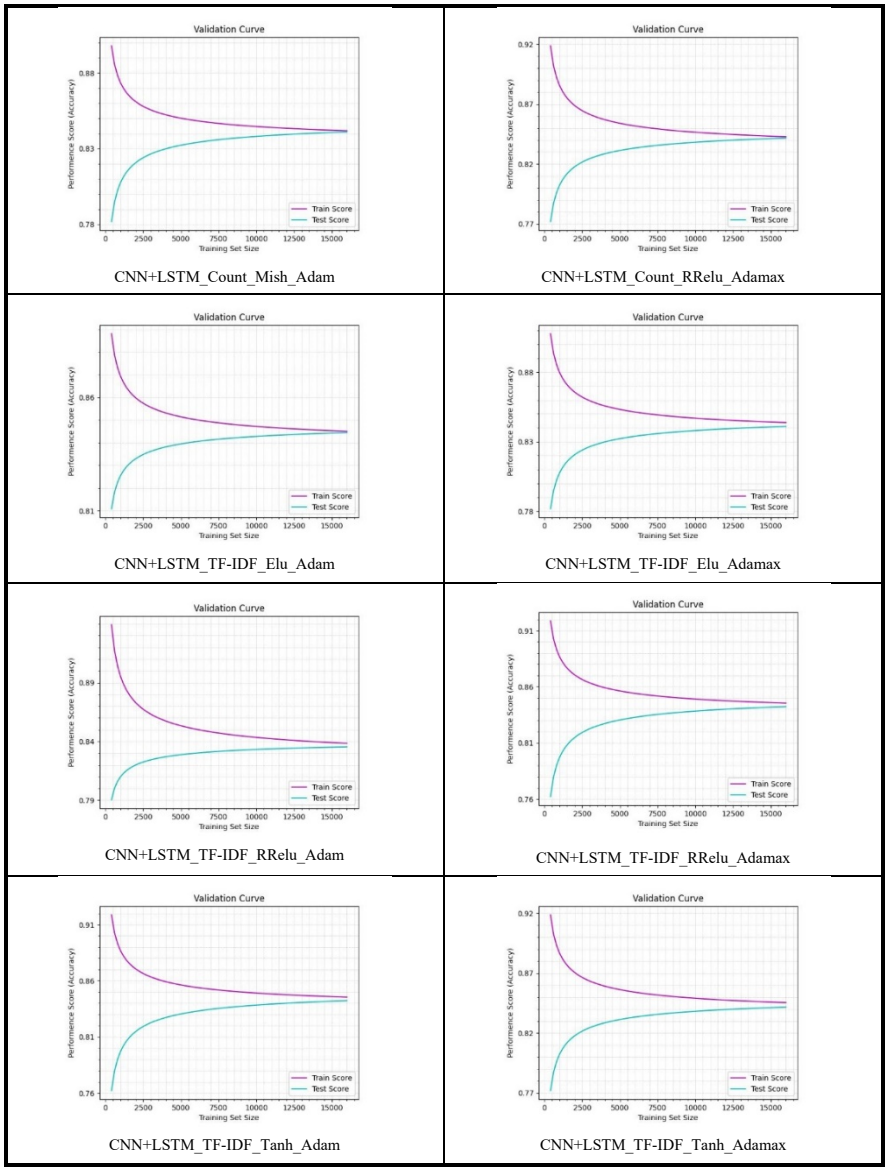
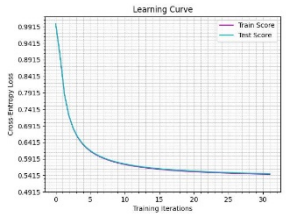
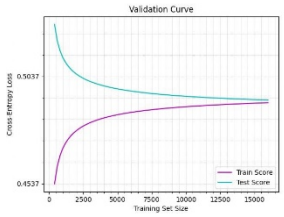


Figure 9: Train size validation curves

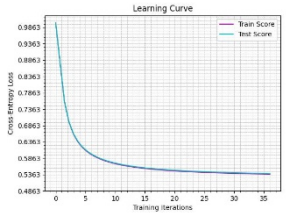
The learning curves and train size validation curves for loss values are shown in Figure 10. These cross-entropy training plots assess the model's performance at various training phases with varying quantities of training data, revealing how effectively the model generalizes to new data. If the model consistently performs well across different train sizes, it indicates that it does not overfit the training data and can make accurate predictions about new information.



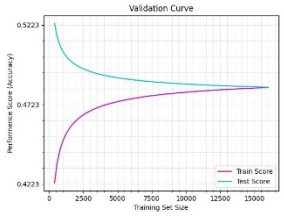
Count Mish Adam (learning curve)



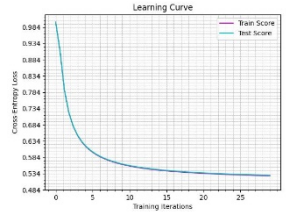
Count Mish Adam (train size curve)



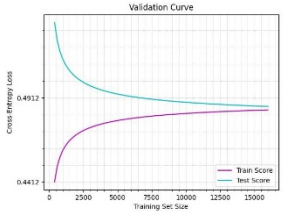
Count RreLU Adamax (learning curve)



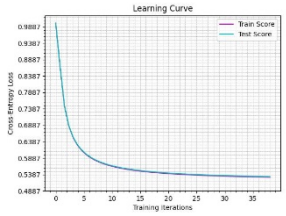
Count RreLU Adamax (train size curve)



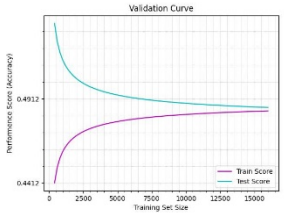
TF-IDF Elu Adam (learning curve)



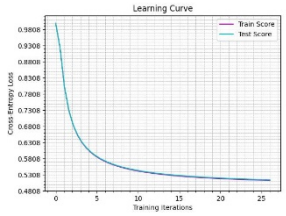
TF-IDF Elu Adam (train size curve)



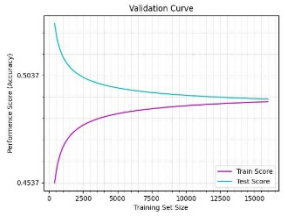
TF-IDF Elu Adamax (learning curve)



TF-IDF Elu Adamax (train size curve)



TF-IDF RRelu Adam (learning curve)



TF-IDF RRelu Adam (train size curve)

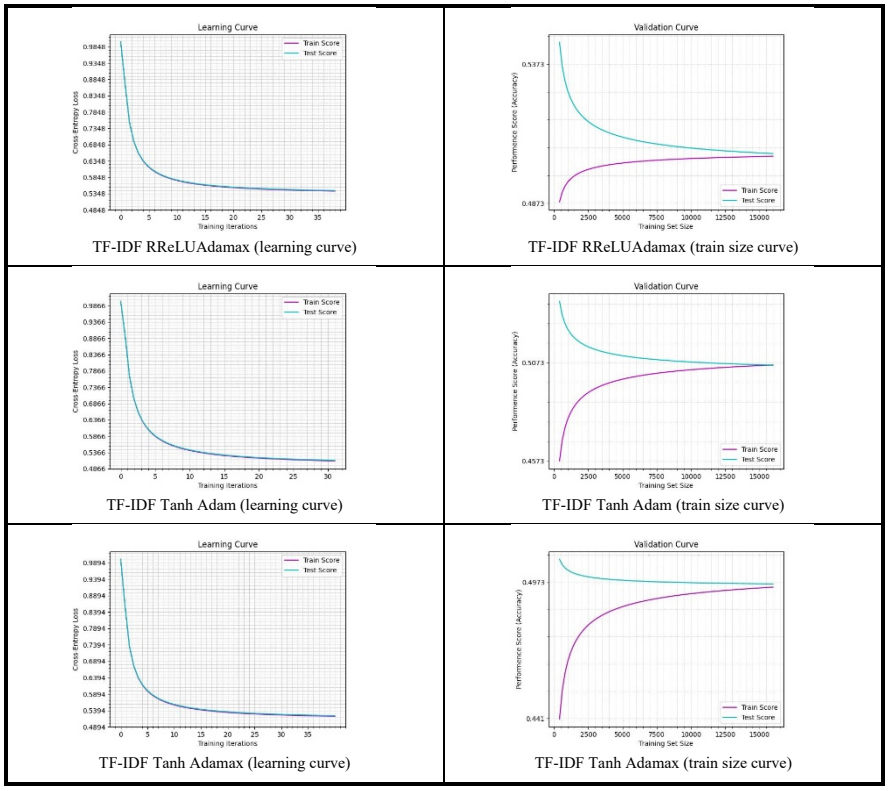


Figure 10: Training and validation loss curves

4.7 Discussion

It was shown in this study that Suicidal behavior or thoughts in online content can be identified in a social media-dominated environment. Social media's ability to identify those in need can significantly improve suicide prevention and mental health care. Early suicidality identification using NLP and ML improves society's mental health and may save lives. This article presented and assessed 82 hybrid CNN-LSTM models to detect text-based suicidality on a Reddit dataset.

4.8 Comparison of Results

Using various word embeddings, activation functions, and solver algorithms impacted accuracy by 12%, precision by 13%, recall by 15%, F1 score by 12%, and AUC by 12% on the test dataset (Table 11).

Score	Accuracy	Precision	Recall	F1 Score	AUC
min	0.74	0.78	0.7	0.74	0.74
max	0.86	0.91	0.85	0.86	0.86
range	0.12	0.13	0.15	0.12	0.12

Table 11: Models' performance metric scores

The top performance models included eight combination models and two stacking ensembles. These models are the most frequently topped in the metric scores' rankings, and each one fits in under 10 seconds. The most common methods among high-performance estimators are the TF-IDF word embedding, RReLU activation function, and Adam solver. Using Tanh instead of RReLU as the activation function results in the fastest-fitting model (Table 10).

The top model lists of all scores generally consist of the same approaches except for recall. Bagging ensembles only appear at the top of recall metric scores. This ensemble lowers false negatives and improves recollection. To increase positive example detection, this estimator learns on various data subsets. Data attributes and variances are captured through a majority vote or averaging. The ensemble's collective decision-making process boosts recall by considering multiple models' perspectives. Remember that a bagged recall score may be good but not precision, accuracy, or F1 score. Increasing memory by bagging reduces false negatives but may raise false positives. Thus, a bagged ensemble's needs and aims must be considered.

4.9 Supporting Studies

Hybrid and ensemble methods in machine learning have garnered significant interest from the scientific community in recent years. Research has demonstrated that using multiple ensemble learning models can yield significantly improved performance compared to using a single weak learner, particularly when dealing with difficult regression and classification tasks that involve high dimensions [Kazienko, 13].

There are many studies on using machine learning to identify mental illness via social media. Many of these researches focus on AI/ML model construction and evaluation. ANNs, which mimic the human brain, may find patterns and relationships in input data [De Ocampo, 18]. Recent years have seen people seeking mental health help on social media. This has led researchers to leverage data, NLP and ML techniques to help others [Ameer, 22]. Studies suggest depression detection systems using deep learning algorithms [Squarcina, 21]. NLP methods like word representations are needed to identify linguistic patterns in the majority of languages [Zhang, 22]. There is a huge need for better embedding approaches and learning algorithms that are able to identify mental illness from the literature [Tejaswini, 24].

Gaur et al. (2018) improved a CNN model by adding external knowledge sources and a suicide ontology to a textual representation [Gaur, 19]. Coppersmith et al. used a self-attention mechanism, a bidirectional LSTM for sequence encoding, and a deep learning model for word embedding to identify the most informative subsequence [Coppersmith, 18]. Sawhney et al. (2018) identified suicidal ideation using LSTM, CNN, and RNN [Sawhney, 18]. In 2019, Ji et al. created an attentive relation network using LSTM and topic modeling to encode text and danger indicators [Ji, 19]. In 2022,

they proposed model aggregation strategies to update CNNs and LSTMs to detect suicidal ideation in private chat rooms. Coordinators in chat rooms annotate user posts for supervised learning in decentralized training [Ji, 22]. Ji et al. (2019) suggested updating CNNs and LSTMs with model aggregation to detect suicidal ideation in private chat rooms [Ji, 19].

In "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech" and "Introduction to Text Classification" by Suryaningrum (2023) and Wendland et al. (2021), the TF-IDF technique yields better results than the Count vectorizer [Suryaningrum, 23; Wendland, 21]. In "Deep Sparse Rectifier Neural Networks" (2011), Bengio et al. state that rectified linear units (ReLU) outperform sigmoid and hyperbolic tangent activation functions in deep neural networks, according to this study. On several benchmark datasets, ReLU activations accelerate training-phase convergence and improve generalization [Bengio, 03; Glorot, 11]. Kingma and BA (2014) offer "Adam: A Method for Stochastic Optimisation" in this work that combines the benefits of the RMSProp and AdaGrad algorithms. Adam beats other optimization methods in convergence and generalization across various deep learning architectures and datasets [Kingma, 14]. Generally, the Adam optimizer exhibits superior performance in terms of both time and accuracy [Maurya, 22]. Kandel suggests that the Adamax optimizer demonstrated the fastest convergence time and the highest stability among all optimizers, with minimal deviation observed across varied learning rates [Kandel, 20].

4.10 Restrictions and Limitations

Mental illness detection using ML in licensed psychologist-classified datasets is limited. Labeled datasets may be biased and inconsistent because specialists weigh symptoms and behaviors differently. Another challenge is the need for more variation in these databases. Specialist-labeled datasets may not account for demographic and cultural differences in mental illness presentation. Using broad and representative datasets, various views in labeling, and regular updates and checks for correctness and dependability can help researchers and developers overcome these restrictions.

This research suggests future text classification research to solve the urgent problems relevant individuals and communities face. By laying out a roadmap for future work, we hope to inspire more research, creativity, and progress in the field, tackling new issues and improving existing solutions.

5 Conclusions

CNN-LSTM models help identify suicidality in social media messages in the early stages. Hence, a hybrid classifier was designed, trained, and tested on the dataset. Stacked and bagging ensemble approaches were used to improve network accuracy and robustness. Ensembles generally reduce overfitting and handle complicated or high-dimensional data better. The generated network was tested on various word embeddings, activation functions, and solver methods. Combining different methods affected accuracy by 12%, precision by 12%, recall by 15%, F1 score by 6%, and AUC by 11%. The outcomes were assessed and compared. Thus, these high-performance approaches are listed:

- Classifier Models: CNN-LSTM and stacked ensembles
- Word Embeddings: TF-IDF
- Activation Functions: RReLU
- Solver Algorithms: Adam and Adamax

A CNN-LSTM model with TF-IDF embedding, RReLU activation function, and Adam optimizer topped all metric scores in our high-performance lists. Stacked ensembles on the Count and TF-IDF word embedding are the fastest models to fit. Except for recall, the top rankings for all other scores are mostly the same combined models. Bagging ensembles obtained a desirable output only on recall performance measures.

In conclusion, identifying mental illness is a challenging and complex task that requires careful attention. Based on the literature, text classification using neural network ML techniques to diagnose mental illness has shown promising results. Technological advances and enhanced collaboration between researchers and mental health providers have expanded the potential for neural network ML classifiers to improve early mental disease detection and intervention. This study examined numerous aspects and implications of neural network ML models, underlining their need for further analysis. More research is needed to address data imbalance, model interpretability, and ethical issues. Governments, researchers, and individuals must recognize the urgent need for ML in mental and psychological realms like suicide detection. This effort may lead to a future without mental illness and suicide.

6 Future Work

It is imperative to analyze the impact of neural networks on society's health and mental care. Machine learning algorithms for real-life suicidality detection are invaluable. These technologies allow social service providers, technological platforms, and mental health professionals to identify individuals who are at risk before they seek treatment. ML algorithms can provide valuable insights and caution by analyzing massive datasets like social media, text messages, and behavioral patterns. This allows timely interventions and may prevent suicides. This technology enhances traditional mental health evaluation, tracks large groups, and reduces the devastating effects of suicide on individuals, families, and communities.

Further study on text classification is necessary to address relevant individuals' and organizations' urgent concerns effectively. The strategic framework promotes further examination, innovation, and progress in the field, addressing emerging challenges and enhancing our comprehension.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Asst. Prof. Dr. Kian Jazayeri, for his unwavering guidance, assistance, and invaluable mentorship during this research endeavor. His knowledge and critique have been influential in shaping the direction of this thesis. I genuinely appreciate his dedication and commitment to fostering my academic growth.

References

[Ameer, 22] Ameer, I., Arif, M., Sidorov, G., Gòmez-Adorno, H., & Gelbukh, A. (2022). Mental illness classification on social media texts using deep learning and transfer learning. arXiv preprint arXiv:2207.01012.

[Amjad, 21] Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A., & Gelbukh, A. (2021). Threatening language detection and target identification in Urdu tweets. IEEE Access, 9, 128302-128313.

[Amjad, 20] Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I., & Gelbukh, A. (2020). "Bend the truth": Benchmark dataset for fake news detection in Urdu language and its evaluation. Journal of Intelligent & Fuzzy Systems, 39(2), 2457-2469.

[Bengio, 03] Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. Advances in neural information processing systems, 13.

[Coppersmith, 18] Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomedical informatics insights, 10, 1178222618792860.

[De Choudhury, 14] De Choudhury, M., & De, S. (2014, May). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 71-80).

[De Ocampo, 18] De Ocampo, A. L. P., & Dadios, E. P. (2018, November). Mobile platform implementation of lightweight neural network model for plant disease detection and recognition. In 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM) (pp. 1-4). IEEE.

[Demšar, 06] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research, 7, 1-30.

[Domingos, 12] Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

[Friedman, 01] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[Gaur, 19] Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., ... & Pathak, J. (2019, May). Knowledge-aware assessment of severity of suicide risk for early intervention. In The world wide web conference (pp. 514-525).

[Getzen, 22] Getzen, E., Ruan, Y., Ungar, L., & Long, Q. (2022). Mining for health: A comparison of word embedding methods for analysis of ehrs data. medRxiv, 2022-03.

[Glorot, 11] Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 315-323). JMLR Workshop and Conference Proceedings.

[Ha, 05] Ha, K., Cho, S., & MacLachlan, D. (2005). Response models based on bagging neural networks. Journal of Interactive Marketing, 19(1), 17-30.

[Hamilton, 67] Hamilton, M. A. X. (1967). Development of a rating scale for primary depressive illness. British journal of social and clinical psychology, 6(4), 278-296.

[Haque, 21] Haque, A., Reddi, V., & Giallanza, T. (2021). Deep learning for suicide and depression identification with unsupervised label correction. In Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30 (pp. 436-447). Springer International Publishing.

[Hastie, 09] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: Springer.

[Ji, 18] Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. Complexity, 2018.

[Ji, 19] Ji, S., Long, G., Pan, S., Zhu, T., Jiang, J., Wang, S., & Li, X. (2019). Knowledge transferring via model aggregation for online social care. arXiv preprint arXiv:1905.07665.

[Ji, 20] Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems, 8(1), 214-226.

[Ji, 22] Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications, 34(13), 10309-10319.

[Jurafsky, 00] Jurafsky, D. (2000). Speech & language processing. Pearson Education India.

[Kandel, 20] Kandel, I., Castelli, M., & Popovič, A. (2020). Comparative study of first order optimizers for image classification using convolutional neural networks on histopathology images. Journal of imaging, 6(9), 92.

[Karim, 18] Karim, H., Niakan, S. R., & Safdari, R. (2018). Comparison of neural network training algorithms for classification of heart diseases. IAES International Journal of Artificial Intelligence, 7(4), 185.

[Kazienko, 13] Kazienko, P., Lughofer, E., & Trawiński, B. (2013). Hybrid and ensemble methods in machine learning J. UCS special issue. J Univers Comput Sci, 19(4), 457-461.

[Kim, 20] Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. Scientific reports, 10(1), 1-6.

[Kingma, 14] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[Krogh, 08] Krogh, A. (2008). What are artificial neural networks?. Nature biotechnology, 26(2), 195-197.

[LeCun, 98] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[Lee, 22] Lee, D., Kang, M., Kim, M., & Han, J. (2022, July). Detecting suicidality with a contextual graph neural network. In Proceedings of the eighth workshop on computational linguistics and clinical psychology (pp. 116-125).

[Li, 18] Li, A., Jiao, D., & Zhu, T. (2018). Detecting depression stigma on social media: A linguistic analysis. Journal of affective disorders, 232, 358-362.

[Low, 20] Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (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.

[Manyika, 11] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity.

[Marcus, 12] Marcus, M., Yasamy, M. T., van Ommeren, M. V., Chisholm, D., & Saxena, S. (2012). Depression: A global public health concern.

[Maurya, 22] Maurya, M., & Yadav, N. (2022, May). A comparative analysis of gradient-based optimization methods for machine learning problems. In International Conference on Data Analytics and Computing (pp. 85-102). Singapore: Springer Nature Singapore.

[Mikolov, 13] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[Murarka, 21] Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021, April). Classification of mental illnesses on social media using RoBERTa. In Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (pp. 59-68).

[Pataky, 20] Pataky, E. A., & Ehlert, U. (2020). Longitudinal assessment of symptoms of postpartum mood disorder in women with and without a history of depression. Archives of Women's Mental Health, 23(3), 391-399.

[Pennington, 14] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[Pirina, 18] Pirina, I., & Çöltekin, Ç. (2018, October). Identifying depression on reddit: The effect of training data. In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task (pp. 9-12).

[Rumelhart, 86] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

[Rumelhart, 88] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Neurocomputing: Foundations of research.

[Sawhney, 18] Sawhney, R., Manchanda, P., Mathur, P., Shah, R., & Singh, R. (2018, October). Exploring and learning suicidal ideation connotations on social media with deep learning. In Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 167-175).

[Shen, 17] Shen, J. H., & Rudzicz, F. (2017, August). Detecting anxiety through reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality (pp. 58-65).

[Sokolova, 06] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Berlin, Heidelberg: Springer Berlin Heidelberg.

[Sokolova, 09] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

[Squarcina, 21] Squarcina, L., Villa, F. M., Nobile, M., Grisan, E., & Brambilla, P. (2021). Deep learning for the prediction of treatment response in depression. Journal of affective disorders, 281, 618-622.

[Sridhar, 96] Sridhar, D. V., Seagrave, R. C., & Bartlett, E. B. (1996). Process modeling using stacked neural networks. AIChE Journal, 42(9), 2529-2539.

[Suryaningrum, 23] Suryaningrum, K. M. (2023). Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech. Engineering, MAthematics and Computer Science (EMACS) Journal, 5(2), 79-83.

[Symeonidis, 18] Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110, 298-310.

[Tadesse, 19] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. Algorithms, 13(1), 7.

[Tang, 15] Tang, J., Qu, M., & Mei, Q. (2015, August). Pte: Predictive text embedding through large-scale heterogeneous text networks. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1165-1174).

[Tay, 22] Tay, H. E., Lim, M. K., & Chong, C. Y. (2022). SERCNN: Stacked Embedding Recurrent Convolutional Neural Network in Detecting Depression on Twitter. arXiv preprint arXiv:2207.14535.

[Tejaswini, 24] Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-20.

[Venek, 17] Venek, V., Scherer, S., Morency, L. P., & Pestian, J. (2017). Adolescent suicidal risk assessment in clinician-patient interaction. IEEE Transactions on Affective Computing, 8(2), 204-215.

[Vioules, 18] Vioules, M. J., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. IBM Journal of Research and Development, 62(1), 7-1.

[Walsh, 11] Walsh, R. (2011). Lifestyle and mental health. American Psychologist, 66(7), 579.

[Wendland, 21] Wendland, A., Zenere, M., & Niemann, J. (2021). Introduction to text classification: impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique. In Systems, Software and Services Process Improvement: 28th European Conference, EuroSPI 2021, Krems, Austria, September 1–3, 2021, Proceedings 28 (pp. 289-300). Springer International Publishing.

[Wolpert, 92] Wolpert, D. H. (1992). Stacked generalization. Neural networks, 5(2), 241-259.

[Zhang, 22] Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. NPJ digital medicine, 5(1), 1-13.

[Zhou, 02] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. Artificial intelligence, 137(1-2), 239-263.