Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 1, February 2025, pp. 638~646
ISSN: 2302-9285, DOI: 10.11591/eei.v14i1.8380

A novel hybrid SMOTE oversampling approach for balancing class distribution on social media text

Nareshkumar Raveendhran, Nimala Krishnan
Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

Article Info

Article history:

Received Feb 23, 2024
Revised Oct 3, 2024
Accepted Oct 17, 2024

Keywords:

Bidirectional long short-term memory
Deep learning
Depression
Social online posts
Synthetic minority oversampling technique

ABSTRACT

Depression is a frequent and dangerous medical disorder that has an unhealthy effect on how a person feels, thinks, and acts. Depression is also quite prevalent. Early detection and treatment of depression may avoid painful and perhaps life-threatening symptoms. An imbalance in the data creates several challenges. Consequently, the majority learners will have biases against the class that constitutes the majority and, in extreme situations, may completely dismiss the class that constitutes the minority. For decades, class disparity research has employed traditional machine learning methods. In addressing the challenge of imbalanced data in depression detection, the study aims to balance class distribution using a hybrid approach bidirectional long short-term memory (BI-LSTM) along with synthetic minority over-sampling and Tomek links and synthetic minority over-sampling and edited nearest neighbors' techniques. This investigation presents a new approach that combines synthetic minority oversampling technique with the Kalman filter to provide an innovative extension. The Kalman-synthetic minority oversampling technique (KSMOTE) approach filters out noisy samples in the final dataset, which consists of both the original data and the artificially created samples by SMOTE. The result was greater accuracy with the BI-LSTM classification scheme compared to the other standard methods for finding depression in both unbalanced and balanced data.

Corresponding Author:

Nimala Krishnan
Department of Networking and Communications
School of Computing College of Engineering and Technology, SRM Institute of Science and Technology
Kattankulathur, Chennai, Tamil Nadu, India
Email: nimalak@srmist.edu.in

1. INTRODUCTION

Depression is one of the most severe mental illnesses since it often results in suicide; thus, it is crucial to identify and synthesize the body of research on depression symptom identification on social media utilizing the information supplied by users. Mental diseases are a global health concern that impacts a significant number of individuals and are responsible for a substantial number of deaths each year. According to a survey published in 2017 by the World Health Organisation (WHO), approximately 284 million people are affected by anxiety, while depression affects around 264 million individuals. Bipolar disorder impacts 46 million people, schizophrenia affects 20 million people, and eating disorders are prevalent among 16 million individuals [1]. Different authors have been examining the detection of unwanted noise and offensive language on the internet using typical machine learning (ML) algorithms. For instance, statistical topic modelling and feature engineering techniques can be utilised to recognise depression tweets. Similarly, train numerous classifiers to differentiate between general objectionable tweets and tweets containing hate speech. In recent years, deep artificial neural networks, also known as deep learning (DL), have been used for various text categorisation tasks. One of these tasks is the detection of offensive and hostile words. Make advantage of recurrent neural networks (RNN), for instance, to identify abusive language used in tweets. Transfer learning combined with convolutional neural networks (CNN) should be used to categorise objectionable tweets on Twitter (X) data [2].

This research project will provide a method for identifying and classifying writing that exhibits depressed symptoms as its primary contribution. In order to provide a more accurate representation of the data, researchers have used both under sampling and over sampling. After that, we trained and labeled the provided text with the help of an intelligence classifier [3]. Finding and synthesising up-to-date material on studies of depression symptom identification on social media is the goal of this literature review, which makes use of linguistic feature extraction, DL, computer tools, and statistical analysis approaches. There are other works that have been produced that deal with a topic that is comparable to the one that this work does. Examples of class imbalance that occur in the real world include the sending of spam emails, the diagnosis of illnesses, credit card fraud, cyberattacks, projections of manufacturing equipment, and the recovery of information. Class imbalance is a common issue. It is challenging to employ ML in the applications that are now being used in the real world due to the unbalanced datasets. This issue is receiving a growing amount of attention, and it has been recognized as one of the ten most difficult challenges in data mining. In order to find a solution to this issue, several different strategies have been suggested. One of the methods that may be used to fix an unbalanced dataset is called resampling, and it does so by altering the dataset itself. Resampling may be executed by numerous techniques, the most prevalent of which are undersampling and oversampling.

Scientists have investigated ML methods for identifying offensive language and unwanted noise on the internet. Additionally, they have employed DL techniques such as RNNs for tasks involving categorizing text. This project aims to create a strategy that can accurately identify and categorize written works that display symptoms of depression. This will be achieved using undersampling and oversampling techniques to improve data representation and train an intelligent classifier. Class imbalance, a prevalent problem in practical scenarios like disease diagnosis and fraud detection, continues to be complicated. Resampling techniques such as undersampling and oversampling have been suggested to address the bias towards the majority classes in imbalanced datasets.

The problem of imbalance in binary datasets was addressed by Tripathi *et al.* [4], who provided a unique method for solving the problem. The authors began by using synthetic minority oversampling technique (SMOTE), and then they partitioned the data that they had collected by using a Gaussian-Mixture method that was based on the clustering approach. At end, they used the cluster's given weight to choose false samples. In their tests, an support vector machine (SVM) was employed to categorize the data. In a recent paper, Elreedy *et al.* [5] research looks at how well the K-nearest neighbor (kNN) algorithm and the SMOTE work together to fix the problem of class inequality in user reviews of Tokopedia, Indonesia's biggest online market. Looking at 5,000 data points with an imbalance of 3,975 bad reviews and 1,025 positive reviews, the SMOTE-kNN method got a higher accuracy rate (90%) than kNN alone (82%), showing that it works better with unbalanced datasets.

Jiang *et al.* [6] an innovative methods to the process of oversampling, which focuses on the contribution degree of the classification. The authors completed the maths to work out what percentage of samples were positive, what percentage were in the minority group, and what percentage were in clusters formed using the k-means technique. SMOTE generates a certain amount of synthetic samples for each available sample, and this number is based on the degree to which safe neighbourhoods contribute to the classification. A resampling strategy was introduced by Pereira *et al.* [7] in order to enhance the overall performance of binary classification problems. As a solution to the problem of inconsistent data, the authors suggested using two methods of oversampling and two methods of undersampling. The findings of the examination indicate a substantial rise in levels of performance. Despite the fact that very little information is now available in this sector, Nareshkumar and Nimala [8] have the potential to be successful.

In Table 1, we have provided terms relevant to the field and depression and a few of the most recent applications based on our findings. The methods addressing imbalance in binary datasets face several significant limitations. Firstly, complexity and computational intensity are major concerns, as approaches like those proposed involve intricate, multi-step processes that demand substantial computational resources and expertise. Secondly, there is a dependency on clustering accuracy, mainly where the success of the Gaussian-Mixture clustering approach is crucial yet challenging to achieve. Thirdly, parameter tuning is critical, which requires precise adjustments that may need to generalize better to different datasets. Additionally, there is potential overfitting, especially with complex resampling strategies. Generalizability issues also arise, as some methods may not be readily applicable across various datasets and might necessitate significant modifications. Lastly, computational and implementation challenges are prevalent, particularly with methods involving genetic algorithms and complex mathematical procedures, making them difficult to implement and resource-intensive.

Table 1. Keyword and linked field of the assessment

Field	Keywords	Related term
Health	Anxiety	Disorder
Online media	Stress	Online users
Online interactions	Depression	Online web
		Platforms
		Sentiment analysis
		Facebook Reddit Instagram Weibo

In addition to introducing a bidirectional long short-term memory (BI-LSTM) method for identifying depression comments in unbalanced datasets, this research gives the following benefits:

– This investigation introduces a unique approach to classifying depression-related comments, BI-LSTM, a novel method in the field. Bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF) are employed as feature extraction method employing unbalanced and imbalanced datasets, respectively, for assessment. Synthetic minority over-sampling and Tomek links (SMOTE Tomek) and synthetic minority over-sampling and edited nearest neighbours' techniques (SMOTE+ENN) are both used in order to accomplish the goal of producing a fair distribution of data across the datasets.
– The suggested model also incorporates The Kalman-SMOTE (KSMOTE) method, which eliminates noisy samples from the final dataset, including both the original data and the artificially generated samples via SMOTE. The system was evaluated both with and without SMOTE Tomek and SMOTE ENN and achieved superior results compared to earlier models in the task of depression categorization.

The document is structured as described below and section 2 discusses the content of our contribution. In the next section 3 discuss about metric setup and the model's results. The work will be finished with section 4, where we will reiterate the importance of our research and its potential impact on mental health.

2. A NOVEL HYBRID ALGORITHM GROUPING SMOTE TOMEK AND ENN BASED ON BI-LSTM

Our research employed data mining method to examine data and make prediction about whether it exhibited characteristics of depression or not. This section describes the methodology used to identify posts or comments on social media that indicate signs of depression. This research introduces an innovative method for categorizing tweets linked to depression by utilizing a BI-LSTM network and complex algorithms for extracting features and resampling data. The study opened by collecting and preparing a benchmark dataset. It then used BoW and TF-IDF to extract features. The investigation applied SMOTE-Tomek, SMOTE-ENN, and a novel Kalman-SMOTE (KSMOTE) technique to address the class imbalance. The BI-LSTM model underwent training using the revised datasets, and its performance was assessed, revealing superior outcomes in depression identification compared to other baseline models. This demonstrates the model's efficacy in healthcare applications.

First, the raw data was preprocessed, which involved cleaning and transforming the data to prepare it for analysis. This was followed by data extraction, where relevant features such as mood, behavior, and sentiment were extracted from the preprocessed data. The text data was then processed using text analysis techniques including tokenization, stemming, and lemmatization so that it could be further analysed. To address imbalanced data, the SMOTE Tomek and SMOTE ENN method was used to oversample the minority class (depressive posts or comments) and balance the dataset. TF-IDF was then used to weight the importance of each term in the text data and reduce the impact of common words that may not be useful in identifying depressive posts or comments. Finally, various classification models such as LSTM networks were trained and evaluated on the preprocessed, TF-IDF weighted [9], and balanced data to predict whether a post or comment exhibited signs of depression or not.

2.1. Implement process

To implement this work, we followed several steps outlined in Figure 1. Here Bench mark depression dataset consisting of 10,282 data points, each representing either depression or non-depression, which were collected from online text sources such as comments and posts. Specifically, focused on collecting short one-line texts. Research used a straightforward data pre-processing pipeline that involved several steps. Firstly, converted all text to lowercase letters to ensure consistency in the dataset. Next, filtered out URLs, usernames, punctuation marks, irrelevant characters, and emojis from the text. Finally, split the pre-processed text into individual word-level tokens. The method of "stemming," which is a kind of natural language processing, is used to condense words down to their essential. The primary objective of stemming is to normalise words in such a way that distinct forms of a word are analysed and retrieved as if they were variants of the same word. In order to determine a word's fundamental structure, stemming algorithms strip away frequent prefixes and suffixes before analysing the resulting word. For instance, when stemming the phrases "jumping," "jumps," and "jumped," a stemming algorithm may convert all of those terms to the root word "jump."

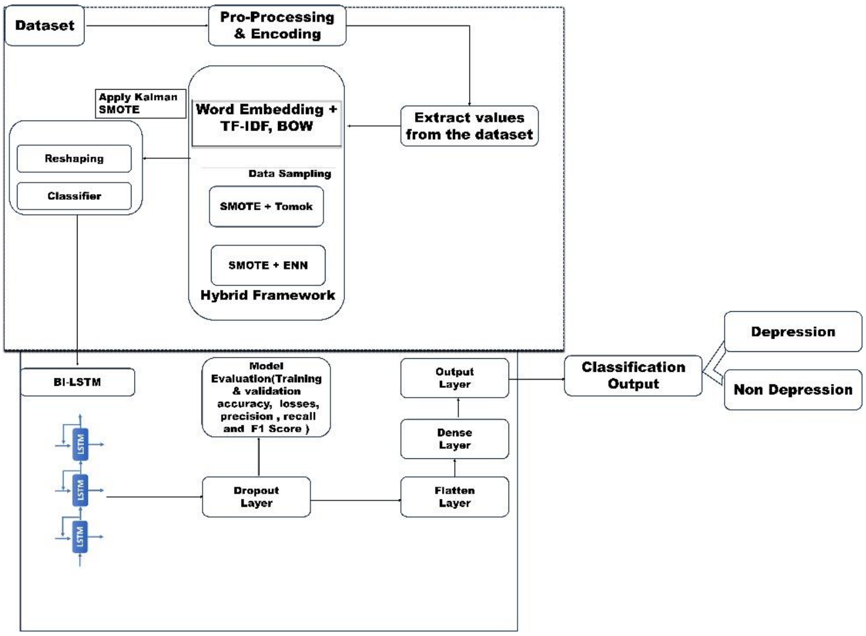


Figure 1. A novel hybrid proposed model architecture

2.2. SMOTE Tomek and ENN

SMOTE is a method [10] that produces fresh synthetic data by randomly interpolating pairs of the data's closest neighbors. The new lawsuits are not carbon copies of previously filed complaints filed by members of minority groups. SMOTE Tomek is a hybrid approach designed to remove overlapping data points for each class in the sample space. To improve the clustering of classes, Tomek linkages are utilised on the oversampled samples of the minority class generated by SMOTE. Therefore, instead of just eliminating data from the majority class, we often eliminate observations from both classes in Tomek linkages. SMOTE ENN is a hybrid methodology that eliminates a greater quantity of the findings from the data set space. The ENN methodology functions as a supplementary undersampling method by assessing the closest neighbors of each instance in the majority class. If the closest neighbors incorrectly label that specific occurrence of the majority class, it is discarded. By combining this method with oversampled data performed by SMOTE, it facilitates thorough data cleansing. In this case, misclassification by neural networks involves the removal of samples from both groups. This leads to a clearer and more succinct division of classes. We have expanded the SMOTE technique by using the Kalman filter. SMOTE's generation of synthetic data introduces noise into the dataset. We have employed the Kalman filter algorithm to exclude undesirable data samples that contain noise.

Imbalanced classes present a significant issue when it comes to the process of training a decent classifier. Because the classes are so unevenly distributed, a majority classifier would produce results that are quite accurate by identifying all of the occurrences with the class that is most prevalent. The disadvantage of this approach is that the minority class is too small to bring the other two classes down to its level, and as a result, a significant amount of information that is relevant and important is thrown away as a result. The undersampling method involves removing the portion of the training dataset that pertains to the majority class. A contrastive operation is accomplished via the use of the oversampling approach.

2.3. Text vectorization

The aim of the phase known as "text vectorization" is to convert the text into a numeric vector representation so that algorithms for learning may be easily applied to it. The term "bag of words" refers to a straightforward method of vectorization in which each piece of text included in the dataset is represented by a vector whose length is proportional to the vocabulary included in the dataset. In this particular method of encoding, a vector is populated with the number of times that each word occurs in the source text. Despite the ease with which this method may be implemented, it is common practice to find vectors to be quite lengthy and to include many zeros. In addition to that, it does not take into account the significance of the words. TF-IDF [11] is an example of a viable solution that may be found in this path.

2.4. Classifier

BI-LSTM [12], [13] is a design of artificial recurrent neural networks applied to DL. It is only able to handle a single piece of data at a time, but it can process a whole series of data in (1)-(8):

$$F(t) = \sigma\left(W_f \cdot [H_{t-1}, X_t] + b_f\right) \quad (1)$$

$$I(t) = \sigma(W_i \cdot [H_{t-1}, X_t]) + b_i) \quad (2)$$

$$\tilde{C}(t) = tanh(W_c \cdot [H_{t-1}, X_t] + b_c) \quad (3)$$

$$C(t) = f_t * C_{t-1} + I_t * \tilde{C} \quad (4)$$

$$O(t) = \sigma(W_o \cdot [H_{t-1}, X_t] + b_o) \quad (5)$$

now, input weight is W_f, W_i , W_c, and W_c , bias is b_f, b_i, b_c, and b_o, t is time state, $t - 1$ is prior time state, X is input; H is output, and C is cell status.

$$H(t) = O_t * tanh(C) \quad (6)$$

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \quad (7)$$

$$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (8)$$

3. RESULTS AND DISCUSSION

In the process of predicting and diagnosing depression, a variety of methods have been used in order to circumvent the challenge posed by class-unbalanced datasets. Sentiment_tweets3 benchmark, a publically accessible dataset on Gargmanas, “Sentimental Analysis,” Kaggle, Oct. 01, 2021. Available: https://www.kaggle.com/code/gargmanas/sentimental-analysis/input was taken into account for this investigation.We have carried out the trials in three phases in this suggested work. The depression identification procedure is carried out in Figure 2(a), using the balanced datasets. In the SMOTE Tomek and ENN approach, oversampling was used to address the problem of the training dataset having an uneven distribution of classes, as shown in Figure 2(b).

The effectiveness of the various classifiers covered in the previous section is evaluated and compared. The scikit-learn library and Matplotlib was used to analyze the classification techniques. This is a prevalent problem in data on depression, which may indicate that people from more privileged backgrounds do not experience depression while those from more disadvantaged groups do. A wide variety of data-level and model-level methodologies were examined to deal with the class imbalance issue [14]. The works demonstrated how the authors' class imbalance management strategies enhanced their model performance and their ability to accurately predict and identify depression. The articles were found to have used the benchmark depression dataset, maybe because it was specifically designed for assessing depression. Based on rectifying this imbalance, the method presents the SMOTE. This technique ensures that the total number of samples is the same. Many publications that were examined used more single sampling strategies [15], especially the most prevalent method that SMOTE used. In each of the different instances, investigators also found that a significant percentage of patients did not suffer from depression. As a result, in addition to building binary classifiers, researchers also created binary classifiers that predict the depression post. Nevertheless, there are drawbacks associated with using SMOTE. It has the potential to generate occurrences in loud and overlapping areas far from safe regions. This can result in examples that do not accurately reflect the minority class and can harm categorisation performance.

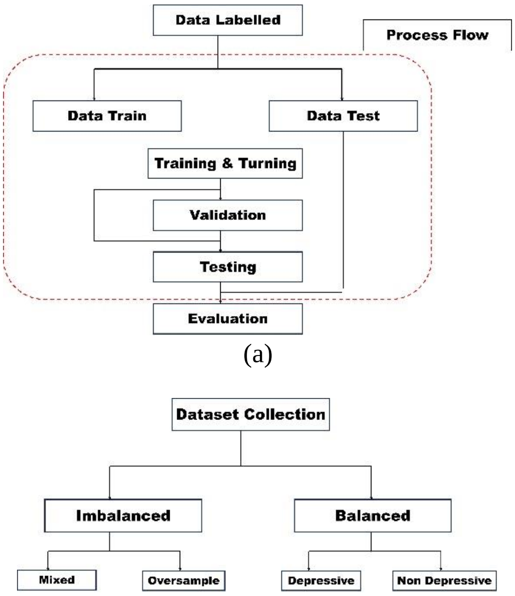
The graphical depiction of the evaluation of training and validation accuracy and loss through the unbalanced dataset is shown in Figure 3. This measurement was performed on the different learning models. Evaluating the DL models that BI-LSTM constructed [16], use a split ratio of 80:20 for the train and test data, and investigate the effectiveness of all models concerning the accuracy (Acc) and weighted average scores for precision (p), recall (r), and F-1 in (9)-(12):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$p = TP/(TP+FP) \quad (10)$$

$$r = TP/(TP+FN) \quad (11)$$

$$F-1-Score = \frac{2*Precision*Recall}{Precision*Recall} \quad (12)$$



(a)

(b)

Figure 2. Flow process of depression identification procedure; (a) flow process and (b) type of data set

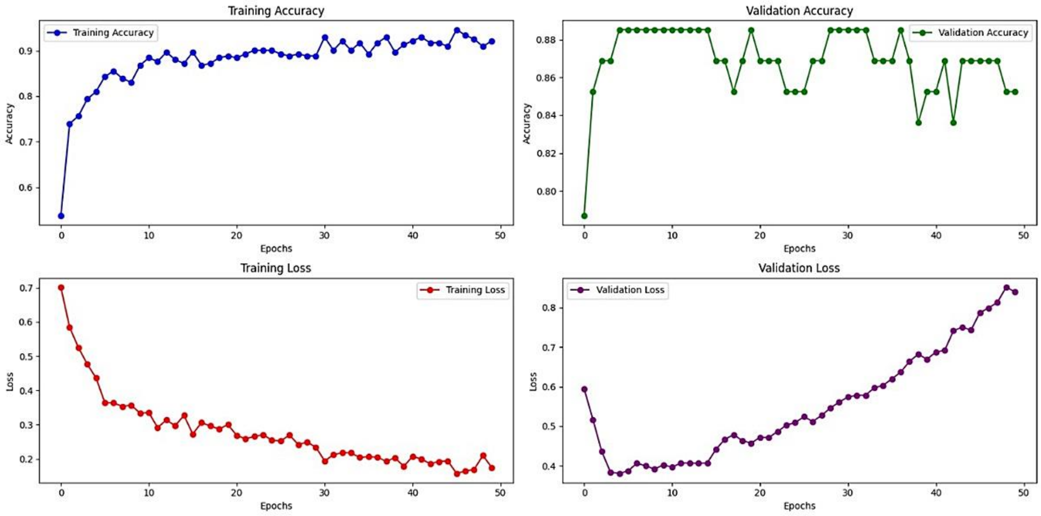


Figure 3. The accuracy and loss value of imbalanced dataset using SMOTE Tomek and ENN

3.1. Comparison results

In this work, a comparison of many different ML approaches [17] for the diagnosis of depressive symptoms is offered. Experiments are carried out with the SMOTE Tomek and ENN balanced dataset, and both TF-IDF and BoW features [18] are used as feature descriptors. The effectiveness of ML models has significantly increased as a result of their training using features derived from the SMOTE Tomek and ENN oversampled dataset that are TF-IDF. Oversampling results in a larger dataset, which in turn results in an increase in the total number of features used to train models. The fact that its recall score is greater than LR, however, demonstrates that it performs more effectively. The KNN models had the worst performance of all the classifiers, with an accuracy of just 0.64, while the RF and SVC models [19] have excellent performance, with accuracies of 0.90 and 0.91, respectively and the comparison methodologies. The depression comment categorization task involves testing the effectiveness of the hybrid BI-LSTM SMOTE Tomek ENN, together with ML models. Text categorization challenges often exhibit superior performance when using DL methods [20]. The proposed method surpasses all other baseline models in its ability to diagnose depression. The suggested model outperforms SVM in accuracy, achieving a 93% accuracy rate. Additionally, it significantly enhances precision, recall, and F1 measure. Compared to random over sampling (ROS) and neighbourhood cleaning rule (NCL), which exhibit lower accuracy and F1 scores, the suggested model's balanced data management considerably enhances performance metrics. The accuracy of the SMOTE ENN and SMOTE Tomek models is approximately 75%, indicating that the proposed Hybrid BI-LSTM technique is substantially more successful. The suggested model has demonstrated exceptional performance across all criteria, solidifying its position as the most dependable and resilient option for detecting depression in healthcare settings. Consequently, we compare evaluate the outcomes of using SMOTE on an SVM classifier [10], as well as the SMOTE Tomek and SMOTE ENN approaches [21] and its show in Figure 4.

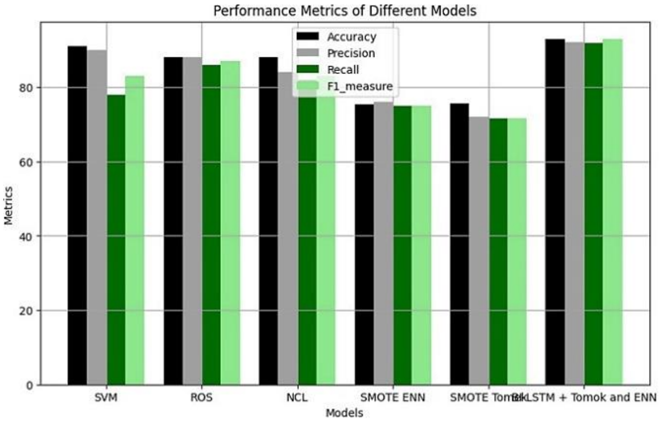


Figure 4. The BI-LSTM+SMOTE Tomek and ENN classification the other baseline models

The findings obtained with SMOTE Tomek and ENN are shown in Table 2. In compared to the performance of models that use the undersampling approach, the outcome of methods that use the oversampling method is substantial. We choose the area under the curve (AUC) as the comparison metric since it is a well-accepted and standardized measure that is not influenced by overfitting. The AUC of our model is most similar to the highest performing model, SMOTE Tomek ENN. Figure 5 illustrates a visual illustration of the evaluation measurement for ROC curve using an imbalanced dataset. In the process of random under-sampling, data are removed at random from the majority class in order to achieve sample parity between the majority and minority classes [22]. As a direct consequence of this, both the amount of data and the number of features in the feature set have become much smaller. In addition, the technique of deleting data that is used in random under-sampling [23] might result in the loss of many critical records, which can have an impact on the way models are trained and thus impair the overall performance of the model. However, oversampling increases data by creating new records. These new data help generate a vast feature set, which improves learning models. The overall outcome of the suggested techniques is higher than that of the separate approach because it combines the results of two techniques that have performed very well such as SMOTE [24], [25] with BI-LSTM.

Table 2. Comparison of previous results and proposed model for depression detection

Models	Accuracy	Precision	Recall	F1 measure
SVM [10]	91	90	78	83
ROS [20]	88	88.1	86	87
NCL [20]	88.1	84	82	83
SMOTE ENN [20]	75.46	76	75	75
SMOTE Tomek [20]	75.55	72	71.5	71.6
Proposed hybrid BI-LSTM+SMOTE Tomek ENN	93	92	91.9	93

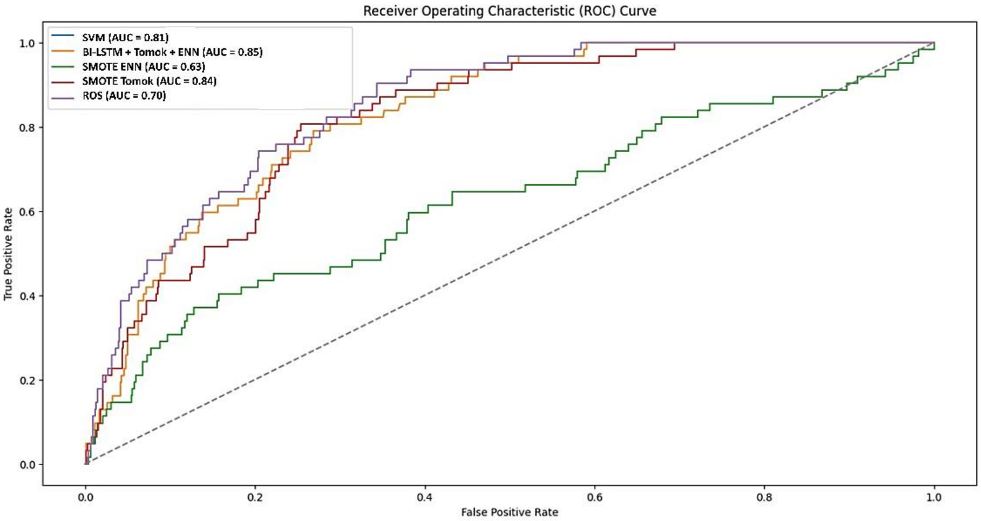


Figure 5. ROC curve of imbalanced dataset using SMOTE Tomek ENN

4. CONCLUSION

This research presents a number of methods for psychological analysis that may be used to identify depressive symptoms. In order to identify symptoms of depression, the researchers utilized SMOTE Tomek and ENN oversampling and an BI-LSTM classifier. To get a more even distribution of values throughout the dataset, researchers combined the SMOTE Tomek and ENN oversampling. Researchers applied the strategy to the baselines of many different neural networks. The findings demonstrated that using both the oversampling and undersampling approaches allowed for the issue of unbalanced data to be resolved.

The suggested model effectively addresses the issue of class imbalance in depression identification within healthcare by merging a BI-LSTM network with SMOTE-Tomek and ENN resampling techniques. The BI-LSTM is highly effective at capturing contextual information. When combined with SMOTE Tomek and ENN, it leads to a cleaner and more balanced dataset. The results suggest that this model surpasses previous basic models, exhibiting exceptional performance in identifying symptoms of depression independent of the distribution of data, thereby demonstrating its efficacy in healthcare applications. For future work, some hybrid algorithm's that will address with more accuracy for depression detection. Integrating undersampling and oversampling with SVM and hybrid attention GRU method might be giving good accuracy for imbalanced dataset as well as the balanced dataset with SMOTE techniques.

REFERENCES

[1] T. R. Shaha *et al.*, "Feature group partitioning: an approach for depression severity prediction with class balancing using machine learning algorithms," BMC Medical Research Methodology, vol. 24, no. 1, Jun. 2024, doi: 10.1186/s12874-024-02249-8.

[2] A. Rajendran, C. Zhang, and M. Abdul-Mageed, "UBC-NLP at SemEval-2019 task 6: ensemble learning of offensive content with enhanced training data," *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, doi: 10.18653/v1/s19-2136.

[3] S. Dowlagar and R. Mamidi, "DepressionOne@LT-EDI-ACL2022: using machine learning with SMOTE and random undersampling to detect signs of depression on social media text," *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, May 2022, pp. 301–305, doi: 10.18653/v1/2022.ltedi-1.45.

[4] A. Tripathi, R. Chakraborty and S. K. Kopparapu, "A novel adaptive minority oversampling technique for improved classification in data imbalanced scenarios," *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 10650-10657, doi: 10.1109/ICPR48806.2021.9413002.

[5] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, no. 7, pp. 4903–4923, Jan. 2023, doi: 10.1007/s10994-022-06296-4.

[6] Z. Jiang, T. Pan, C. Zhang, and J. Yang, "A new oversampling method based on the classification contribution degree," *Symmetry*, vol. 13, no. 2, p. 194, 2021, doi: 10.3390/sym13020194.

[7] R. M. Pereira, Y. M. Costa, and C. N. Silla Jr, "Toward hierarchical classification of imbalanced data using random resampling algorithms," *Information Sciences*, vol. 578, pp. 344-363, 2021, doi: 10.1016/j.ins.2021.07.033.

[8] R. Nareshkumar and K. Nimala, "Interactive deep neural network for aspect-level sentiment analysis," *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Jan. 2023, doi: 10.1109/iceconf57129.2023.10083812.

[9] Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, "Depression detection in social media comments data using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 987–996, Apr. 2023, doi: 10.11591/eei.v12i2.4182.

[10] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on imbalanced text features

for toxic comments classification using RVVC model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/access.2021.3083638.

[11] V. Nurcahyawati and Z. Mustaffa, "Improving sentiment reviews classification performance using support vector machine-fuzzy matching algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1817–1824, Jun. 2023, doi: 10.11591/eei.v12i3.4830.

[12] R. Bayoumi, M. Alfonse, M. Roushdy, and A.-B. M. Salem, "Text-to-image generation based on AttnDM-GAN and DMAttn-GAN: applications and challenges," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 1180–1188, Apr. 2023, doi: 10.11591/eei.v12i2.4199.

[13] R. Nareshkumar, G. Suseela, K. Nimala, and G. Niranjana, "Feasibility and necessity of affective computing in emotion sensing of drivers for improved road safety," *Advances in Computational Intelligence and Robotics*, pp. 94–115, Sep. 2022, doi: 10.4018/978-1-6684-3843-5.ch007.

[14] İ. Abaci and K. Yildiz, "SMOTE vs. KNNOR: An evaluation of oversampling techniques in machine learning," *Gümüşhane University Journal of Science and Technology*, vol. 13, no. 3, pp. 767-779, Jun. 2023, doi: 10.17714/gumusfenbil.1253513.

[15] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Information Sciences*, vol. 565, pp. 438–455, Jul. 2021, doi: 10.1016/j.ins.2021.03.041.

[16] R. Nareshkumar and K. Nimala, "An enhanced BERT model for depression detection on social media posts," *Artificial Intelligence: Theory and Applications*, pp. 53–64, 2024, doi: 10.1007/978-981-99-8479-4_5.

[17] M. Mustaqim, B. Warsito, and B. Surarso, "Combination of synthetic minority oversampling technique (SMOTE) and backpropagation neural network to contraceptive IUD prediction," *Media Statistika*, vol. 13, no. 1, pp. 36–46, Jun. 2020, doi: 10.14710/medstat.13.1.36-46.

[18] M. S. Hammoodi and A. Al-Azawei, "A proposed approach to discover nearest users on social media networks based on users' profiles and preferences," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 2464-2473, Aug. 2023, doi: 10.11591/beei.v12i4.4436.

[19] R. Salas-Zárate, G. Alor-Hernández, M. del P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, and J. L. Sánchez-Cervantes, "Detecting depression signs on social media: a systematic literature review," *Healthcare*, vol. 10, no. 2, p. 291, Feb. 2022, doi: 10.3390/healthcare10020291.

[20] U. Ependi, A. F. Rochim, and A. Wibowo "A hybrid sampling approach for improving the classification of imbalanced data using ROS and NCL methods," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 3, pp. 345–361, Jun. 2023, doi: 10.22266/ijies2023.0630.28.

[21] M. T. Islam and H. A. Mustafa, "Multi-layer hybrid (MLH) balancing technique: a combined approach to remove data imbalance," *Data and Knowledge Engineering*, vol. 143, p. 102105, Jan. 2023, doi: 10.1016/j.datak.2022.102105.

[22] K. Zhu, M. Yin, D. Zhu, X. Zhang, C. Gao, and J. Jiang, "SCGRU: a general approach for identifying multiple classes of self-admitted technical debt with text generation oversampling," *Journal of Systems and Software*, vol. 195, p. 111514, Jan. 2023, doi: 10.1016/j.jss.2022.111514.

[23] M. Umer *et al.*, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognition Letters*, vol. 150, pp. 250–257, Oct. 2021, doi: 10.1016/j.patrec.2021.07.009.

[24] N. K. Mishra and P. K. Singh, "Feature construction and smote-based imbalance handling for multi-label learning," *Information Sciences*, vol. 563, pp. 342–357, Jul. 2021, doi: 10.1016/j.ins.2021.03.001.

[25] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: a feature-weighted oversampling approach for imbalanced classification," *Pattern Recognition*, vol. 124, p. 108511, Apr. 2022, doi: 10.1016/j.patcog.2021.108511.

BIOGRAPHIES OF AUTHORS

Nareshkumar Raveendhran received the M.E. and B.E. (Computer Science and Engineering) from Anna University, Chennai, India in 2008 and 2013, respectively. He is currently a research scholar at Department of Networking and Communication, SRM Institute of Science and Technology, Kattankulathur, Chennai. His research includes machine learning, data mining, natural language processing, text mining, and deep learning analytics. He can be contacted at email: nr7061@srmist.edu.in.

Nimala Krishnan received the P.hD. degree in Computer Science from SRM University, Kattankulathur, Chennai, India. She is a Professor of Department of Networking and Communication, SRM Institute of Science and Technology, Kattankulathur, Chennai since 1998. She has published over 30 papers in international journals and conferences. Her research includes machine learning, data mining, natural language processing, and text. She can be contacted at email: nimalak@srmist.edu.in.