AI-ML Analytics: A Comprehensive Investigation on Sentimental Analysis for Social Media Forensics Textual Data

Yashas Hariprasad[1(✉)], Suraj Lokesh[1], Nagarjun Tumkur Sharathkumar[1], Latesh Kumar KJ[1], Chance Miller[1], and Naveen Kumar Chaudhary[2]

[1] Florida International University, Miami, FL 33174, USA
{yhari001,sloke003,ntumk002,lkumarkj,cmill171}@fiu.edu
[2] National Forensics Sciences University, Gandhinagar, Gujarat, India
naveen.chaudhary@nfsu.ac.in

Abstract. Individuals spend a significant portion of their time on social media. It has become a platform for expression of feelings, sharing of ideas and connecting with other individuals using video and audio posts, textual data such as comments and descriptions and so on. Social media has a considerable impact on people's daily life. In recent time, there is an enormous growth in number of people using Twitter and Instagram to share their emotions and sentiments which represents their actual feelings. In this work, we apply Machine Learning techniques on social media data to perform a comprehensive investigation to detect the risk of depression in people. Our work can help to detect various symptoms such sadness, loneliness, detachment etc. providing an insight for forensic analysts and law enforcement agencies about the person's mental state. The experimental results show that Extra Tree Classifier performs significantly better over the other models in detecting the sentiment of people using social media data.

Keywords: Sentimental Analysis · Social Media Forensics · Machine Learning Classifiers · SMOTE · Imbalanced Data

1 Introduction

Social media has become a platform for people to share their thoughts and express their true feelings. During the pandemic, there is an increase in the time spent on social media by individuals [1]. Social media platforms such as Twitter, Facebook, Instagram and are very helpful in tracking public well-being and mental health of people with the aid of growing research and data analysis techniques [2]. Studies show that people write and post what they feel as it reduces the intensity, anxiety, and phobias [3]. There is an immense potential given that social media emphasizes on emotions and there are tweets every second, changes in the emotional state of a person can be tracked down to a granular level. Some social media platforms such as Instagram and twitter are drastically gaining the

K. Arai (Ed.): SAI 2023, LNNS 739, pp. 923–935, 2023.
https://doi.org/10.1007/978-3-031-37963-5_64

number of users [30]. There are several other platforms, for this study we choose twitter for conducting our research work. There are approximately 1.44 billion monthly users on Instagram as of August 2022 and they spend 30 min of time in a day on an average [4]. According to a study published by S. Dixon in 2022 shows that there is a 15% rise in the number of monetizable daily active users (mDAU) on Twitter amounted to approximately 237.8 Million users [5]. Figure 1 shows the rise of Twitter users from 2017 to 2022.

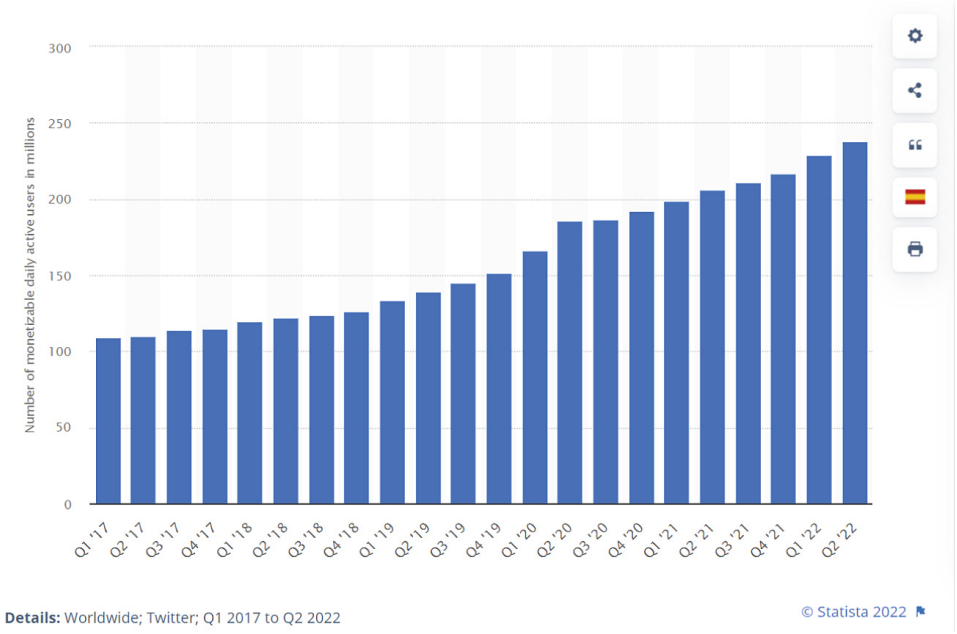

Fig. 1. Raise of Twitter Users.

Major depressive disorder is a medical illness that affects millions of people around the world. It is characterized by feelings of worthlessness, guilt, and low motivation, and it can often lead to self-injury or suicide if no measures are taken [7].

An estimated 5 percent of people around the world are exposed to depression episodes [8], and in the United States, approximately 17.3 million people 18 and older are directly affected by this disorder every year, in addition to 1.9 million children aged 3–17 [9,10].

Research has shown that people who find themselves going through a period of depression or experiencing factors that cause depression tend to experience a higher use of social media [11–13]. Since many social media platforms like Facebook, Instagram, and Twitter allow their users to express their feelings on the internet, these platforms could be used to detect early stages of depression in their users and therefore help prevent it.

In this study we make us of text data such as comments and post descriptions from Instagram, Twitter, and Redditt to investigate the and analyze the mental

state of the person. These emotions can be either positive or negative. We consider the neutral emotions as positive emotions for this work. We observed that the textual data that was obtained had class imbalance with respect to positive and negative emotions. This class imbalance possess a challenge when the model is trained and may lead to inaccurate result. Thus, we employ Synthetic Minority Oversampling Technique (SMOTE) algorithm to balance the dataset [6]

We give a comparative study using various Machine leaning models' such as Random Forest Classifier (RF), Extra Tree Classifier (ET) and Decision tree Classifier (DT). Significant results are obtained and ET performs better over the other models. We also present a comparison of our work with the existing state-of-the-art models and we can observe that our model has obtained better results.

Organization of the Paper: In Sect. 2, we discuss the related works in terms of the research we have done. In Sect. 3, we demonstrate the framework of our approach and the preliminary concepts behind this work. In Sect. 4, we present our results along with some discussion. In Sect. 5, our work is concluded.

2 Related Works

Orabi et al. [14] discusses the use of Convolutional Neural Networks (CNN's) and Recurrent Neural Networks (RNN's) to detect Twitter users with signs of depression. It compares some widely used deep learning models for depression detection on social media using two different publicly available datasets, and concludes that the CNN-based models generally outperform the RNN-based models.

Deshpande et al. [15] applies natural language processing on Twitter to detect depression on the platform's users. It makes use of a Naive Bayes Classifier and Support Vector Machine (SVM) to classify tweets as either positive or neutral based on a pre-created keyword list for detecting trigger words that could indicate depression. The Multinomial Naive Bayes outperforms the Support Vector Machine in terms of precision, recall, and accuracy, but it is still highly limited in prediction of depression through text.

Islam et al. [16] performs depression analysis on Facebook data taken from a publicly available dataset. It examines various linguistic cues in users' comments by applying supervised Machine Learning techniques, and compares four different classifiers ("Decision tree", "k-Nearest Neighbor", Support Vector Machine", and "Ensemble") to categorize user comments into a depressed or non-depressed category. The proposed method significantly improves the accuracy and classification error rate and concludes that "Decision tree" is the most accurate of the tested Machine Learning approaches to identify depression in Facebook comments.

Abdi et al. [17] proposes an improved deep learning-based method to apply sentiment analysis to user reviews. It makes use of a Recurrent Neural Network (RNN) to classify user's opinions on given products or services as either positive or negative based on pre-processed textual reviews. The proposed model outperformed other state-of-the-art machine learning-based models for

sentiment analysis and considered strategies to minimize existing problems with other similar models, like contextual polarity and sentiment shifter.

Chiong et al. [18] investigates several text pre-processing and textual-based featuring methods in addition to machine learning methods to detect depression in user-generated social media texts, particularly those without any specific keywords that would indicate depression. The proposed approach in this model is able to effectively detect depression without the need for specific keywords in the training dataset and when unrelated datasets are used to train the model. The findings of the current study indicate that the proposed detection techniques need improvement, and the training data has to be balanced. Data balancing has not been addressed by any of the existing state-of-the-art techniques.

3 Framework

Our framework presents the process that we have followed for this work. We have considered 2 datasets for this study; one is the Distress Analysis Interview Corpus (DAIC) dataset [19] and the other one is Twitter and Reddit Sentimental analysis Dataset from Kaggle Repository [20]. The data pre-processing was performed by applying a vectorizer on the textual data. We present a comparative analysis of three different Machine Learning Classifiers applied on the datasets, namely, RF, ET and DT. We observed that the data is imbalanced, which means that the samples of one class is significantly more than the other. This will lead to an improper training to the model. In order to balance the data, we employ SMOTE. Figure 2 shows the detailed flow of our work.

3.1 Data Pre-Processing

Raw DAIC textual data was obtained which contains the data from interviews related to anxiety, depression and stress disorder and so on. The responses of the interview were particularly extracted using python scripts. Responses that contained less than 3 words were dropped as they did not support in sentimental analysis. After cleaning the data, the following three steps were applied [21]. 1) Lemmatization- to extract a root-word of a given world and it performs morphological analysis. 2) Stopwords - These are the words that do not influence sentimental analysis, such words are dropped. 3) Stemming - This process is similar to Lemmatization and is also used to extract the root-words, but stemming process, removes letter one by one until the root-word is obtained. It works in most cases. Once the pre-processing is complete, we perform Sentimental analysis.

3.2 Sentimental Analysis

After obtaining the clean data, which is suitable for analysis, a score is assigned to the data using a module called Sentiwordnet. It is a resource for opinion mining.

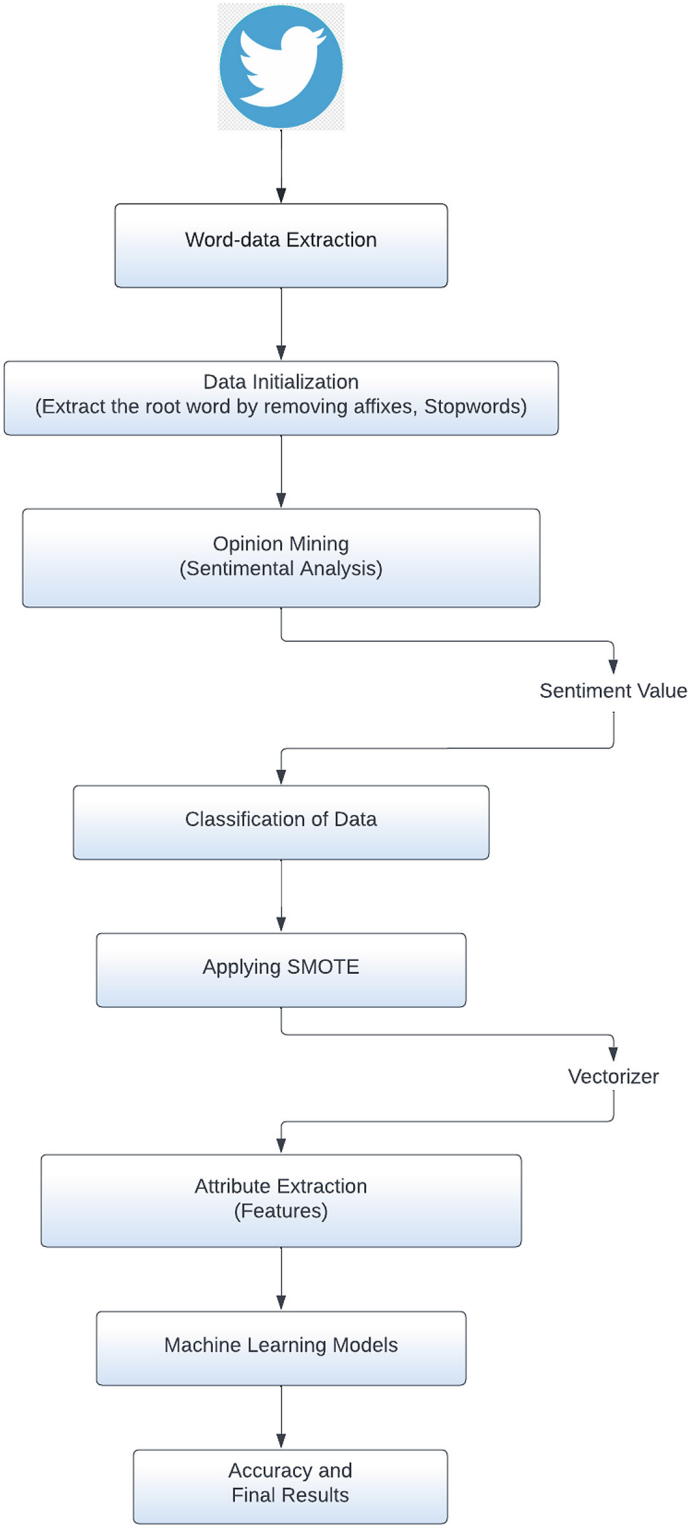

Fig. 2. Work Flow of the Model.

This module assigns a unique value ranging from 0 to 1 to each word in a word-network. Based on the average scores considered for the whole sentence, we can categorize them into Positive, Negative and Neutral. We considered the Positive and Neutral values as Positive and a class label 1 was assigned, and for Negative, a class label of 0 was assigned. After this process of labeling the data, the textual data was converted to vectors using a method called as Term Frequency-Inverse Document Frequency (TFIDF) Vectorizer [22]. This process essentially considers the frequency of words in a sentence and converts the sentences into vectors based on the frequency.

3.3 Data Balancing

We observed that the data is not balanced, which means that the number of samples of one class is significantly more than the other. This will lead to an improper training of the Machine Leaning model and may lead to over fitting [23]. In order to balance the data, we have used SMOTE algorithm which considers the nearest neighboring datapoints and uses k Nearest Neighbor (KNN) algorithm to balance the data. The first step is to determine the distance that a sample is from its nearest neighbor. The distance is then multiplied by a random value between 0 and 1 and added to the sample. A random point is chosen to be on the line segment connecting the two given samples.

3.4 Random Forest Classifier

A Random Forest Classifier (RFC) is a type of ensemble-based Machine Learning algorithm for data classification. It consists of a series of several individual decision trees that act as an ensemble to provide an output based on the classification of most independent trees. It provides several advantages over other classifiers, like high accuracy when observing several datasets and the ability to handle several input variables without any variable deletion [24].

3.5 Extra Tree Classifier

An Extra Tree Classifier (ETC) is a type of classifier similar to the Random Forest Classifier that is also popular for classifying data. It differs from the Random Forest Classifier in that it uses the entire original data sample while the Random Forest Classifier uses bootstrap replicas, and that the selection of cut points to split nodes is done randomly instead of by choosing the optimum split like the Random Forest Classifier does [25].

3.6 Decision Tree Classifier

A Decision Tree Classifier is another Machine Learning-based method for data classification. It creates a model similar to a tree diagram consisting of a root node, branches, internal nodes, and leaf nodes, and it is used to classify data based on a number of set rules and the analysis of data features [26].

4 Results and Discussion

In recent times, the usage of social media by individuals has drastically increased. From teenagers to elderly people, social media is a platform to share thoughts, ideas and feelings either in the form of text, pictures or videos. We considered the textual data from these social media channels to perform our study. People's comments or opinions are analyzed using the concept of sentimental analysis specifically on Twitter data.

Machine learning models were applied to classify the data and various metrics such as Accuracy (ACC), Recall, F1-Score (F1) and Precision have been used to demonstrate the results. We also make use of Area under the curve (AUC) [27] for Receiver Operating Characteristic (ROC) to verify the overfitting issue that is caused due to imbalanced data.

Table 1 shows the datasets and the number of observations of each class label for the raw dataset. Number of 1's shows the total number of samples having class label 1 and Number of 0's shows the total number of samples having class label 0.

Table 1. Dataset Description (Raw Data).

Dataset	Number of 1's	Number of 0's
DAIC dataset (D1)	4255	2793
Twitter and Reddit Sentimental analysis Dataset (D2)	72249	35509

We can clearly observe that there is an imbalance and if we train the model with an imbalanced data, there will be an overfitting issue.

Table 2 portrays the results that we have obtained when three different machine leaning classifiers were applied on the raw imbalanced data.

Table 3 shows the datasets and the number of observations of each class label for the balanced data.

Table 4 shows the results on the dataset after applying SMOTE.

There is a slight decrease in accuracy after SMOTE is applied, we can infer that the previous accuracy that was obtained on imbalanced data was not precise due to overfitting.

In order to represent the AUC ROC curve graphically, we have plotted the graphs in Fig. 3, 4, 5 and 6 as shown below. Figure 3 and 4 portrays ROC curve for the classifiers on Raw (imbalanced) data for Dataset 1 and 2, respectively. Figure 5 and 6 shows the ROC curve for the classifiers on the balanced data for Dataset 1 and 2, respectively.

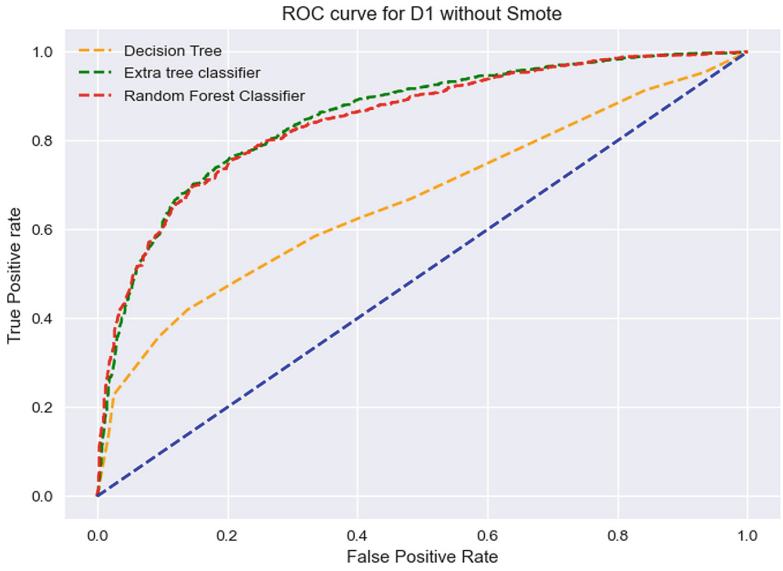

Fig. 3. ROC Curve with Dataset-1 without Smote.

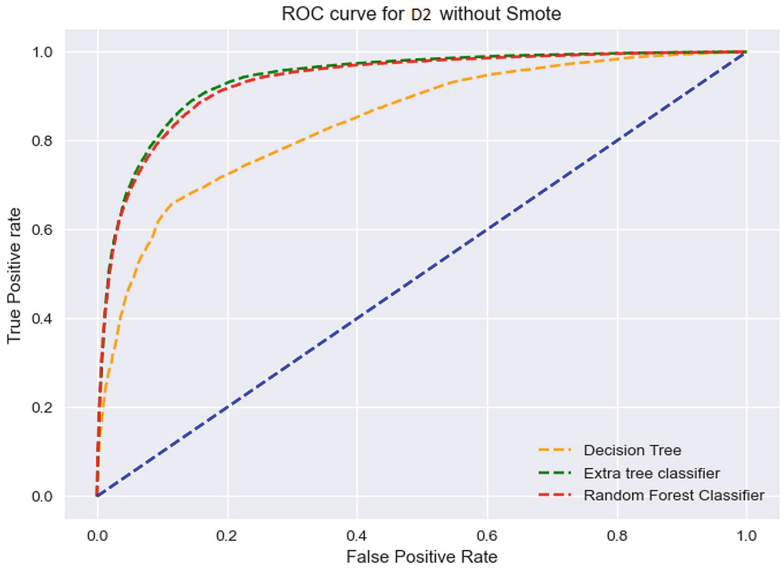

Fig. 4. ROC Curve with Dataset-2 without Smote.

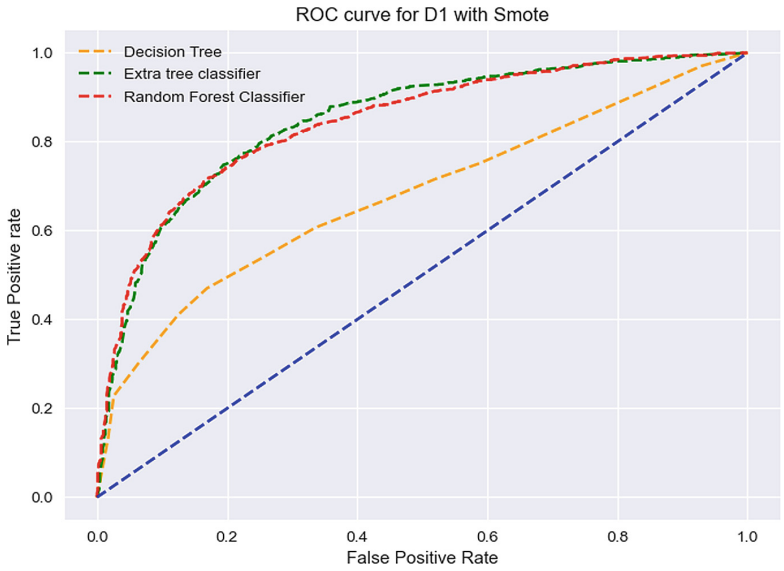

Fig. 5. ROC Curve with Dataset-1 with Smote.

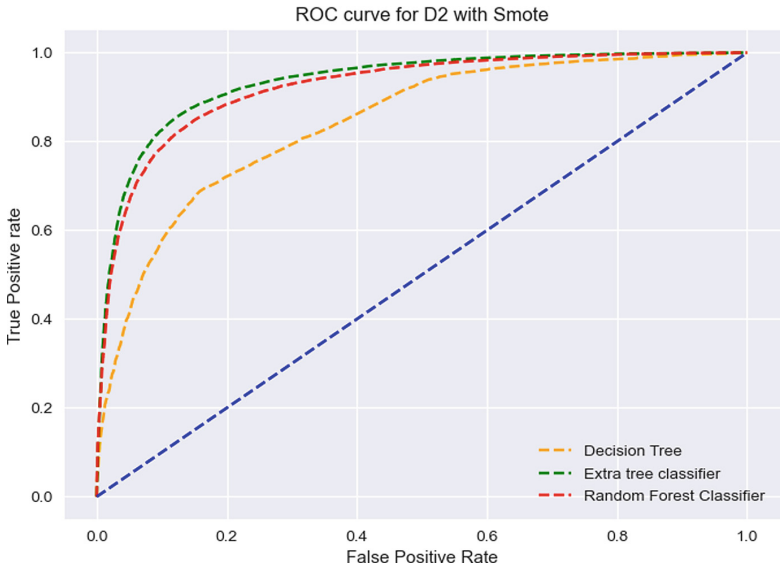

Fig. 6. ROC Curve with Dataset2 with Smote.

Table 2. Results on Raw Imbalanced Data.

Decision Tree					
Dataset	Recall	F1	Precision	AUC	ACC
D1	0.66930266	0.671474937	0.67366136	0.770364211	0.608340499
D2	0.878908209	0.839183445	0.802894345	0.881634102	0.773375327
Extra Tree Classifiers					
Dataset	Recall	F1	Precision	AUC	ACC
D1	0.825305536	0.816500711	0.807881773	0.868829252	0.778159931
D2	0.92501254	0.916640779	0.908419195	0.94194013	0.886814207
Random Forest Classifiers					
Dataset	Recall	F1	Precision	AUC	ACC
D1	0.825305536	0.813031161	0.801116539	0.865446635	0.77300086
D2	0.923382378	0.911382483	0.899690478	0.937309044	0.879193498

Table 3. Dataset Description (Balanced Data).

Dataset	Number of 1's	Number of 0's
D1	2864	2864
D2	48325	48325

Table 4. Results on Balanced Data.

Decision Tree					
Dataset	Recall	F1	Precision	AUC	ACC
D1	0.470165349	0.593735815	0.80541872	0.79621852	0.61521926
D2	0.799197459	0.820196041	0.84232786	0.88830856	0.76426422
Extra Tree Classifiers					
Dataset	Recall	F1	Precision	AUC	ACC
D1	0.827462257	0.818343406	0.80942335	0.87003352	0.78030954
D2	0.932118375	0.906467217	0.88219005	0.92998821	0.87058857
Random Forest Classifiers					
Dataset	Recall	F1	Precision	AUC	ACC
D1	0.780733285	0.802660754	0.82585551	0.8688576	0.77042132
D2	0.917405116	0.896165939	0.87588794	0.92442977	0.85697815

Table 5. Comparison of our Results with Existing Models (D1 Dataset).

Model	F1	ACC
"Detecting Depression with Audio/Text Sequence Modeling of Interviews" [28]	0.77	NA
"Detecting depression using vocal, facial and semantic communication cues" [29]	0.76	NA
"Learning Approaches for Detecting Early-Stage Depression using Text" [21]	NA	77.47%
Our Model (Extra tree classifiers)	**0.82**	**87.11%**

In Table 5, we present a comparison of our work with the existing models for D1 Dataset. We can observe that our model (Extra Tree Classifier) performs better in terms of F1 Score and Accuracy as compared to the existing work.

5 Conclusion

In this paper, we apply Machine Learning techniques to social media data to detect risk of depression in social media users. We compare three different popularly used Machine Learning methods for data classification: Random Forest Classifier, Extra Tree Classifier, and Decision Tree Classifier. Two datasets were considered, and the data was pre-processed by applying a vectorizer on the textual data from the datasets. Sentiment analysis was performed on the pre-processed data by assigning a score to it using Sentiwordnet. These scores are used to classify data as either positive, negative, or neutral, after which the data is converted to vectors using a Term Frequency-Inverse Document Frequency (TFIDF) Vectorizer. Data was later balanced using SMOTE in order to avoid improper training of the model and overfitting. The tree previously discussed classifiers were applied to the data resulting from the previous process and evaluated based on Accuracy, Recall, F1-Score, and Precision. Results show that the Extra Trees Classifier (ETC) generally outperforms both the Decision Tree Classifier (DTC) and the Random Forest Classifier (RFC) and also in comparison with the existing state-of-the-art models within the parameters of our study.

In the future, this sentiment analysis model for detection of depression risk in social media users could be applied to audio and video data. Furthermore, the use of Neural Networks for depression detection could be further investigated. In the era of social media, where people share their thoughts and feelings, the development of sentiment analysis could have the capability to help people who are experiencing mental disorders, and even potentially save lives by detecting these disorders with anticipation on social media platforms.

Acknowledgments. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-21-1-0264. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. https://www.economicsobservatory.com/what-do-social-media-reveal-about-our-emotions-during-the-covid-19-crisis
2. McDool, E., Powell, P., Roberts, J., Taylor, K.: The internet and children's psychological wellbeing. J. Health Econ. **69**, 102274 (2020)
3. https://socialsciences.nature.com/posts/42357-measuring-the-effects-of-expressing-your-feeling-on-social-media
4. https://thesmallbusinessblog.net/instagram-statistics/
5. https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/#:~:text=In%20the%20last%20reported%20quarter,the%20second%20quarter%20of%202021
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Torres, F.: What Is Depression? Psychiatry.org - what is depression?, October 2020. https://psychiatry.org/patients-families/depression/what-is-depression#section_1
8. Erzen, E., Çikrikci, Ö.: The effect of loneliness on depression: a meta-analysis. Int. J. Soc. Psychiatry **64**(5), 427–435 (2018)
9. National Institute of Mental Health "Major Depression" (2017)
10. Centers for Disease Control "Data and Statistics on Children's Mental Health" (2018)
11. Caplan, S.E.: Preference for online social interaction: a theory of problematic internet use and psychosocial well-being. Commun. Res. **30**(6), 625–48 (2003). https://doi.org/10.1177/0093650203257842
12. Caplan, S.E., Andrew, C.H.: Online social interaction, psychosocial well-being, and problematic Internet use. Internet Addict. Handb Guide Eval. Treat. 35–53 (2007)
13. https://www.sciencedirect.com/science/article/pii/S0747563213004093?via%3Dihub#b0045
14. Orabi, Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 88–97 (2018)
15. Deshpande, M., Vignesh, R.: Depression detection using emotion artificial intelligence. In: 2017 International Conference on Intelligent Sustainable Systems (ICISS), pp. 858–862. IEEE (2017)
16. Islam, M., et al.: Depression detection from social network data using machine learning techniques. Health Inf. Sci. Syst. **6**(1), 1–12 (2018)
17. Abdi, A., Shamsuddin, S.M., Hasan, S., Piran, J.: Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. Inf. Process. Manag. **56**(4), 1245–1259 (2019)
18. Chiong, R., Budhi, G.S., Dhakal, S., Chiong, F.: A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Comput. Biol. Med. **135**, 104499 (2021)
19. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. University of Southern California Los Angeles (2014)
20. https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset
21. Suhas, G.H., Suraj, L., Varun, J., Veda, D.V., Jayanna, H.S.: Machine learning approaches for detecting early-stage depression using text. In: 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 106–110 (2021). https://doi.org/10.1109/ICEECCOT52851.2021.9707950

22. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
23. Thejas, G.S., Hariprasad, Y., Iyengar, S.S., Sunitha, N.R., Badrinath, P., Chennupati, S.: An extension of synthetic minority oversampling technique based on Kalman filter for imbalanced datasets. Mach. Learn. Appl. **8**, 100267 (2022)
24. Thaseen, S., Kumar, C.A.: An analysis of supervised tree based classifiers for intrusion detection system. In: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, pp. 294–299 (2013). https://doi.org/10.1109/ICPRIME.2013.6496489
25. Pierre, G., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
26. Panigrahi, R., Borah, S.: Classification and analysis of Facebook metrics dataset using supervised classifiers (2019)
27. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**(7), 1145–1159 (1997)
28. Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/Text sequence modeling of interviews. In: Interspeech, pp. 1716–1720 (2018)
29. Williamson, J.R., et al.:Detecting depression using vocal, facial and semantic communication cues. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 11–18 (2016)
30. Hariprasad, Y., Latesh Kumar, K.J., Suraj, L., Iyengar, S.S.: Boundary-based fake face anomaly detection in videos using recurrent neural networks. In: Arai, K. (eds.) Intelligent Systems and Applications. IntelliSys 2022. LNCS, vol 543, pp. 155–169. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-16078-3_9