ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2021.0360

PREDICTING DEPRESSION FROM SOCIO-ECONOMICAL FACTORS

Lubnaa Abdur Rahman and Poolan Marikannan Booma
Faculty of Computing, Engineering and Technology, Asia Pacific University, Malaysia

Abstract

With the sudden arrival of the Covid19, people have been experiencing higher levels of pressure which has given rise to the number of cases of Depression, a mental illness which can further lead to suicide and contribute to global burden, and therefore, faster ways for its diagnosis need to be sought to prevent further fatalities. Machine Learning, which has been implemented with proven results across various sectors especially for the prediction of diseases, promises to be of great help to facilitate the detection of depression within patients. However, up to now, most implementation for mental health prediction are greatly built on clinical data which takes time to be generated as several tests need to be taken which is not always plausible for everyone. Subsequently, this work explores the application and evaluation of three different machine learning algorithms; ANN, CART DT and SVM, for the prompt prediction of depression based on readily available data; behavioural and social-economical characteristics, and thus provides the best algorithm which could be applied in the context. SVM produced the highest accuracy and was therefore tuned whereby the highest recorded accuracy was of 93.85%. As this work has shown promising results, it is further recommended that future works explore deeper into assessing depression severity.

Keywords:

Artificial Neural Network, Decision Trees, Depression Prediction, Socio-Behavioural Factors, Support Vector Machines

1. INTRODUCTION

With the high stakes and time criticality within the healthcare sector, there is the need to develop faster and efficient ways for disease diagnosis. Even though the high number of research on algorithms and the development of "intelligent" systems, with disease diagnosis and prediction abilities, it would have been thought that the state of health would have reached higher standards by now, but it has been observed that mental health rates are higher than ever. The unpredicted arrival of Covid19, has not only proved to be a threat to the global economy and physical health but the peaking rates of suicide rates, depression, anxiety and panic attacks show that is has accentuated the deterioration of global mental health [1].

Depression, affecting 1 in every 23 people, is a medical illness affecting how one feels, thinks and acts which beyond casual mood swings and short emotional responses of daily routine and often stems from social, biological and psychological factors or as a consequence of adverse life situations. An early diagnosis of depression can reduce the risk of being affected by Major Depressive Disorder (MDD) which is directly associated to suicide, but it can be a costly time-consuming process due to which it goes by undiagnosed in most cases.

Most past research focused efforts in implementing machine learning for its diagnosis based on either medical interaction between patients and their doctors or based on physical tests like MRI scans whereby a great deal of time goes by and when the patient is diagnosed, the depression severity might have worsened. Furthermore, currently with the health anxiety prevailing and movement restrictions among the never-ending problems of the pandemic, such previously explored methods are not viable and therefore other aspects, like tapping into readily available data like behavioural and social data need to be sought.

2. RELATED WORKS

2.1 HEALTHCARE

Researchers have explored various machine learning classifiers for predicting chronic disease such as heart disease and diabetes and it is highlighted that K-Nearest Neighbour and Support Vector Machines are robust to noisy data while Naïve Bayes is appropriate for both numerical and nominal data but needs a consequent amount of data for great accuracy [2]. Artificial Neural Network performed well at various occasions, and it was highlighted that even when the number of input features are limited, good accuracies can be obtained [3]. Decision Trees classifiers, including CART and C5.0 among others, have proved to work well on categorical data but with a overfitting as a major limitation. It has also been put forth that in the case of Heart diseases SVM, Naïve Bayes and Random Forest most of the time gave above 80% accuracy [4].

There has been various ML application for mental health prediction which were mostly focused on clinical and biological data with few implementations based on social media and survey data [5] and noted that for the former, models like Regression, SVM, BN, KNN and Decision trees were popular, and even though results varied based on data mostly, fair results were obtained. On the other hand, on easily accessible data like social media, SVM, Decision Trees, Bayesian Networks and Neural Networks were mostly used [6]. However, certain researchers have discouraged the use SVM for prediction of mental health due to low accuracy for the data use; a small brain imaging related dataset [7]. It was identified that major determinants of depression were, but not limited to, weight loss/gain, loneliness, changes in mood and feelings; moody, irritable, restlessness and feeling tense.

2.2 FINANCE

In the financial sector, Machine Learning has been applied for predicting bankruptcy through algorithms of Support Vector Machines, Neural Network and Random Forest and the researchers noted that the latter gave the best results with accuracy of above 80% [8]. Having reviewed past implementations of the same problem where it was seen that other algorithm such as Logistic Regression and Naïve Bayes were additionally [9]. For classifying whether banks are stable or not based on mostly numerical data, NN and Decision Tree classifiers; CART and CHIAD were applied with the latter outperforming with an accuracy of 88.8% [10]. It was highlighted that in determining a

bank's failure, K Nearest Neighbour and NN are fairly accurate and outperformed SVM [11].

2.3 OTHER SECTORS

Flood prediction, based on numerical data, has been explored through supervised machine learning techniques of Neural Network which performed very well in case of time series, SVM that was more suited for nonlinear regression problems and Decision Tree models like CART and RF were applied with the latter giving the best results in most cases [12]. Additionally, in predicting sports results, for instance predicting a team's match result with possible predicted classes being win, lose or draw, Decision Trees, Naïve Bayes and neural network were good choices [13] and it is highlighted that SVM has good accuracy for the same matter and is good in constructing complex non-linear decision boundaries and being less prone to overfitting.

Researchers have explored machine learning algorithms for the prediction of academic performance (fail or pass) based on students' online behaviours dataset having a categorical and numerical data, whereby Decision Trees performed poorly and NN and SVM gave the best results and noted that as number of features were increased, prediction accuracy of all models increased. Furthermore, for flight delay prediction, on a variety of data involving weather, surveillance messages, flight schedule among others, through Random Forest algorithm showed promising results with an accuracy of 90.2% [14].

For predicting customer churn, whereby the predicted value is categorical with 2 class labels of 'Churn' or 'No churn', ML classification algorithms of Random Forest and Decision Trees and Gradient Boosting, an ensemble model, with the latter having the best accuracy followed by Random Forest. Researchers have explored the application of popular models used for the same problem and have highlighted that Decision Tree model C5.0, SVM and Neural Network have been popularly implemented with Neural Network and SVM performing very well (accuracy above 80%) [15].

3. MATERIALS AND METHODS

In line with the KDD data mining methodology, preferred over CRISP-DM and SEMMA, provided a more general and simple approach for the extraction of knowledge, a Machine learning workflow, Fig.1. was followed within the major steps.

- ***Problem Definition***: The first step was to establish clear definition of the problem; the classification of depressive subjects
- ***Data Collection***: The appropriate data, while ensuring that it was sufficient for the learning process, was collected from Kaggle (the data was originally obtained from a subreddit thread).
- ***Data Pre-processing***: Several tasks were applied; data exploration, missing and noisy data identification and cleaning, transformations wherever necessary and feature selection whereby the best features for the problem are identified from the data. At this stage, the bigger chunk of data was split into training and test sets.
- ***Model Training and Testing***: Supervised Learning was applied through the three different models selected namely, ANN, SVM and Decision Tree. The models are fitted and trained; and prediction applied to the test data. Cross Validation was applied for each model whereby mean accuracies were computed as well.
- ***Evaluation and Improvement***: Results were evaluated through the accuracy metrics and the AUC curve in order to identify the most suitable model followed by hyper-parameter tuning of the best model in an attempt to improve the SVM model performance by finding optimal parameter setting.

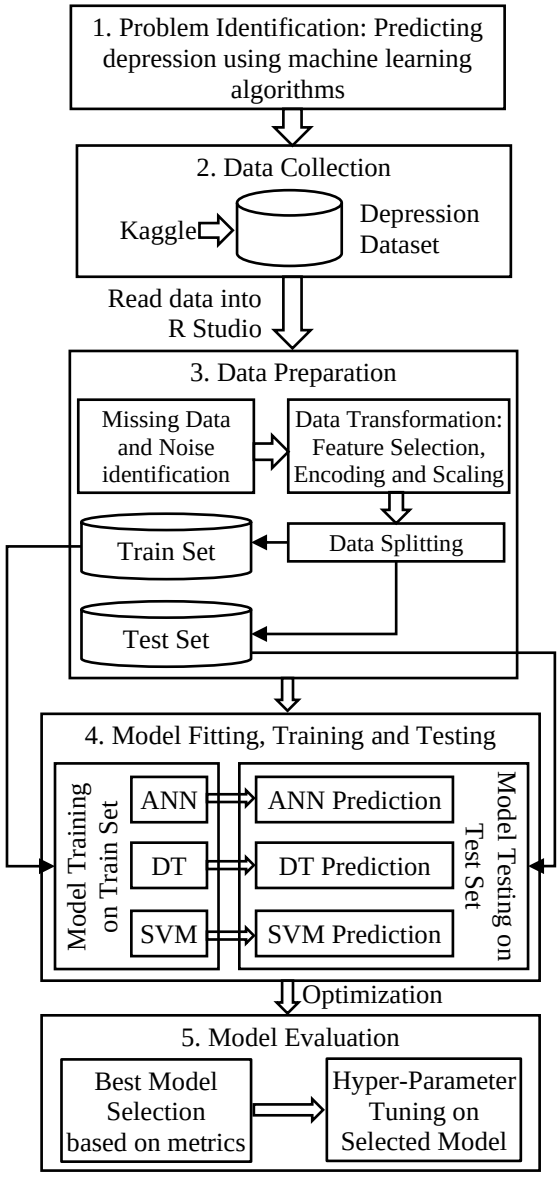

Fig.1. Machine Learning Workflow

4. RESULTS AND DISCUSSION

To solve the problem, a dataset of 19 attributes and 596 records (see Table.1 for dataset description), originating from the social media platform, Reddit, was acquired from Kaggle as it was geared towards social and behavioural attributes of subject who felt alone. Abiding to the ML workflow in Fig.1, R programming was used for pre-processing, transforming and algorithm application.

4.1 PRE-PROCESSING AND DATA ANALYSIS

Upon the initial reading of the dataset, for ease of model application, factor conversion was applied to categorical variables. This was followed by identifying the missing values, Fig.2, and the Multiple Imputation by Chained Equations (MICE) package was used as remedy. Additionally, not all variables were deemed important and were dropped leaving the data with 15 variables.

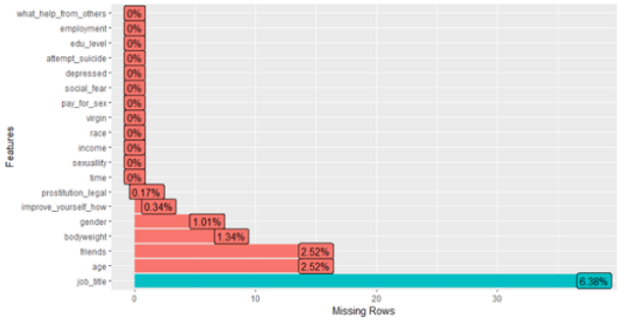

Fig.2. Missing Data Identification

Table.1. Dataset Description

Attribute	Data Type	Description
Time	Date	Time stamp at which survey is filled
Gender	Categorical	Female or Male
Sexuality	Categorical	Sexual Orientation (bi-sexual, Gay, Straight)
Age	Numerical (ordinal)	Person's age
Income	Categorical	Income Category (13 categories)
Race	Categorical	Person's race (6 races)
Bodyweight	Categorical	Person's body weight (4 weight categories)
Virgin	Categorical	If person is virgin or not (Yes/No)
Prostituion_legal	Categorical	If prostitution is legal in country or not (Yes/No)
Pay_for_sex	Categorical	If person pays for sec (Yes/No)
Friends	Numerical (ordinal)	Number of friends person has
Social_fear	Categorical	If person has social fear or not (Yes/No)
Depressed	Categorical	If person is depressed or not (Yes/No)
Attempt_suicide	Categorical	If person has attempted suicide or not (Yes/No)
Edu_level	Categorical	Person's education level (9 categories)
Employment	Categorical	Person's Job (9 categories)
What_help_from_others	Categorical	All kinds of help that person is seeking from others (51 categories)
Job_title	Categorical	Person's job title (264 categories)
Improve_yourself_how	Categorical	How person is improving themselves (61 categories)

4.2 MODEL APPLICATION AND RESULTS

For all the algorithm application, SVM, DT and ANN, the target (y) variable was set as attribute Depressed while the rest were used as independent variables. The dataset was initially split in a 75:25 ratio and cross validation was also applied for all the algorithms to verify their performance and robustness in different settings. Sample of the trained models are shown in Fig.3. and 4.

Looking at the Accuracy, it was seen that the best model was SVM with a much higher accuracy of 82.35% observed, while Decision Tree was 6% behind it and ANN showed the lowest accuracy of all with only 73.95%. On the other hand, looking at the AUC values as shown in the ROC plots, Fig.5 with a 1.5 threshold, ANN recorded the highest AUC value of 0.603 (3 d.p) which was considered as a poorly acceptable value.

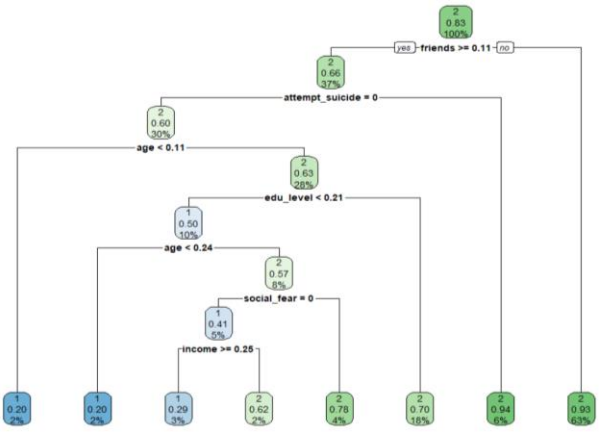

Fig.3. Decision Tree

Under nested cross validated conditions with a k-fold and with three different splitting ratios of 60:40, 70:30 and 80:20, the models SVM, DT and ANN all showed slight improvements in terms of accuracy values, whereby SVM recorded a highest accuracy of 86.13%. It was further noted that the mean AUC values showed improvement only for Decision Tree which recorded 0.64 (see Table 2).

The Accuracies and AUCs created a certain ambiguity since ANN was seen to have the lowest accuracy value but outperformed based on the AUC value while SVM which had the highest accuracy, showed the lowest AUC value in the case of the simple model implementation. It was highlighted that AUC is, in certain cases, a misleading metric [16].Therefore, it was deemed fit to look at the confusion matrix of each model to check the False Positives and the False Negatives since these are highly undesirable in the healthcare/clinical setting for the simple reason that False positives mean that the model is telling that healthy people are ill which can lead to undesired health anxiety and False Negatives meant that the model shows that those who are affected by the disease are actually healthy which can lead to death in extreme case.

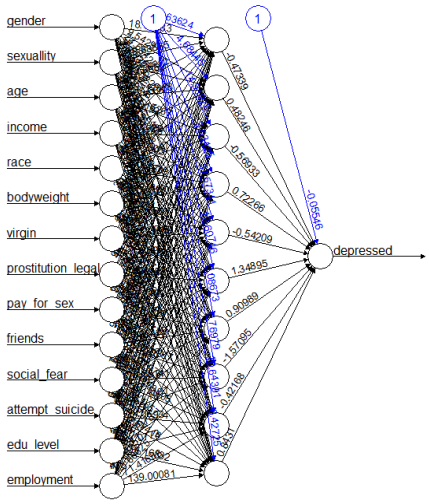

Fig.4. Neural Network Plot

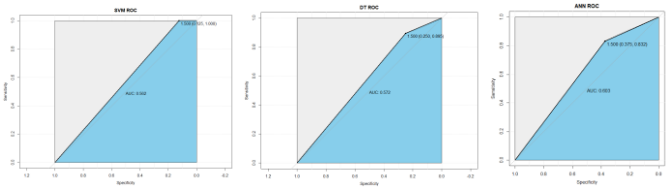

Fig.5. AUCROC Plots

Table.2. Model Summary

Model	Mean Accuracy (under Cross Validation)	Highest Accuracy Recorded
ANN	78.11%	81.56%
DT (CART)	79.94%	83.61%
SVM	82.93%	86.13%
Tuned SVM	88.43%	93.85%

From resulting confusion matrix, it was noted that SVM had 21 FNs which meant that a percentage of 17% of people are said not to have depression while they actually have and a 0 FP which means that none of the non-depressed subjects are being classified as depressed. The Decision Tree model showed consequent FP and TP which is highly undesirable in a clinical setting and was therefore disregarded. Lastly, NN was seen to have the highest AUC in simple settings, but when the numbers of FNs and FPs were looked at, they are noted to be high with, respectively, 13% and 12%.

Having reviewed the Accuracies and AUC values for simple models and CV models and FP and FN for the simple models, it was decided to use the model with the highest accuracy, and which showed a better trade off in sense of FPs and FNs, in this case SVM as the best model. Even though the AUC value remained low in all settings for SVM, and the FNs was quite high, it is noted that depression is not a deadly disease on its own, but it can, indeed, lead to suicide which then causes death [17]. However, it is not like cardiovascular diseases or Cancers which require might immediate attention and action like cardiovascular diseases and therefore, it is considered as tolerable, up to a limit, for the model to have a FN of 17% since misclassified subjects are not, ideally, on the verge of death. Furthermore, it is noted that the FP rate was 0% which is good.

4.3 MODEL OPTIMIZATION: HYPER-PARAMETER TUNING

The SVM model was tuned using tune function whereby three different kernels, that can be viewed as algorithm class used in analysing patterns, were supplied and the function also attempts to find the best epsilon and cost values where it was seen that the cost was 1 with a degree of 3. The model performance was plotted, in Fig.6, whereby it was seen that the data near the soft margin decreases with increasing cost and whereby the similarity was high in the beginning and then decreases as the soft margin is constructed.

The accuracy of the tuned model was of 83.19% and a slightly better AUC value of 0.58 was observed. Since only a minor increase in the model's performance was observed in the above, the tuning process was then cross validated (10-fold and 3 different splits). A much better mean accuracy of 88.43 % was observed while the maximum accuracy, of 93.85%, was observed for a 70:30 split, and the lowest accuracy was of 83% which is still considered fair.

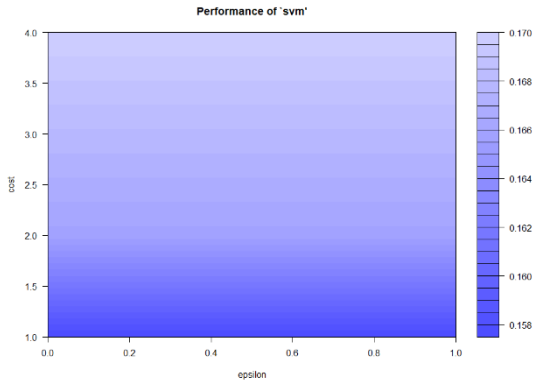

Fig.6. Model Performance Plot

Overall looking at the accuracies produced by SVM under different conditions, it can be said that, indeed, SVM shows great promise as it has delivered in the current scenario where a desired accuracy of 80% was being looked at. To further add on, even though less common, a similar depression classification which looked at non-clinical data has shown a highest accuracy of only 70% that was reached. Additionally, an error rate of 7% is acceptable as noted that depression on its own is not deadly even though it can lead to suicide attempt which is then chained to mortality.

5. CONCLUSION

This study has put forth a all-inclusive research and implementation of the usage of commonly available data like socio-behavioural characteristics of people for identifying depression. Such an application is even more relevant in today's setting where the Covid19 pandemic has exerted much undesired pressure on everyone. The results have shown that such an application has great potential and should be regarded as an alternative to the traditional methods of using imaging and clinical

which are more costly and time consuming. Thorough research showed that ANN, SVM and DT were popular for binary classification, more so in healthcare and were therefore looked at.

Certain results like the novel pattern found between depression and the fact that higher rates were observed in countries where prostitution was legal, followed by the misidentification of gender's linkage to depression were unexpected results. Additionally, the rather poor performance of Artificial Neural Network, especially whereby decision trees outperformed the latter, at first came as a surprise but upon further analysis did seem rationally plausible due to the fact that Neural network is an algorithm which indeed requires a much larger amount of training data that what was used as input in this project. Support Vector Machines showed very good results especially under cross validation, but it was noted that SVM took take time to run especially during the tuning process [18].

During this study, as CART Decision Tree did not deliver the required results, and since time did not permit, it is recommended that future works look into the application of other decision tree models like C4.5 and even the Random Forest Algorithm. Additionally, regarding model application different settings of the hidden layer could be tried out to verify the performance of Neural Network with a much bigger dataset.

It is noted that depression is a big umbrella and comes in varying degrees from mild to severe depression and that there are different types of depression as well such as Major Depressive Disorder (MDD), manic depression/bipolar disorder, psychosis, perinatal depression amongst various others. In this current work, none of these have been explored. Therefore, it also is recommended that future works delve into these as well as it has been put forth that everyone is affected in one way or another by depression be it passive or active and it is the higher severity degree / certain types that lead to the biggest number of fatalities and therefore need to be identified with priority [19]. Lastly, as the practical implications in terms of socio-ethical issues were looked at, it is to be noted that future works not only should look at better data which could be verified but also subjects whose data are to be mined need to be made fully aware.

REFERENCES

[1] S. Pappa, V. Ntella, T. Giannakas, V. Giannakoulis, E. Papoutsi and P. Katsaounou, "Prevalence of Depression, Anxiety, and Insomnia among Healthcare Workers during the COVID-19 Pandemic: A Systematic Review and Meta-Analysis", *Brain, Behavior, and Immunity*, Vol. 88, pp. 901-907, 2020.

[2] N. Kumar and S. Khatri, "Implementing WEKA for Medical Data Classification and Early Disease Prediction", *Proceedings of International Conference on Computational Intelligence and Communication Technology*, pp. 1-6, 2017.

[3] A. Shatte, D. Hutchinson and S. Teague, "Machine Learning in Mental Health: A Scoping Review of Methods and Applications", *Psychological Medicine*, Vol. 49, No. 9, pp. 1426-1448, 2019.

[4] L. Tennenhouse, R. Marrie, C. Bernstein and L. Lix, "Machine-Learning Models for Depression and Anxiety in Individuals with Immune-Mediated Inflammatory Disease", *Journal of Psychosomatic Research*, Vol. 134, pp. 1-13, 2020.

[5] T. Horigome, "Evaluating the Severity of Depressive Symptoms using Upper Body Motion captured by RGB-Depth Sensors and Machine Learning in a Clinical Interview Setting: A Preliminary Study", *Comprehensive Psychiatry*, Vol. 98, pp. 1-15, 2020.

[6] L. Cui, "Symptomatology Differences of Major Depression in Psychiatric Versus General Hospitals: A Machine Learning Approach", *Journal of Affective Disorders*, Vol. 260, pp. 349-360, 2020.

[7] F. Barboza, H. Kimura and E. Altman, "Machine Learning Models and Bankruptcy Prediction", *Expert Systems with Applications*, Vol. 83, pp. 405-417, 2017.

[8] Y. Qu, P. Quan, M. Lei and Y. Shi, "Review of Bankruptcy Prediction using Machine Learning and Deep Learning Techniques", *Procedia Computer Science*, Vol. 162, pp. 895-899, 2019.

[9] H.A. Abdou, W.M. Abdallah, J. Mulkeen, C.G. Ntim and Y. Wang, "Prediction of Financial Strength Ratings using Machine Learning and Conventional Techniques", *Investment Management and Financial Innovations*, Vol. 14, No. 4, pp. 194-211, 2017.

[10] H. Le and J. Viviani, "Predicting Bank Failure: An Improvement by Implementing a Machine-Learning Approach to Classical Financial Ratios", *Research in International Business and Finance*, Vol. 44, pp. 16-25, 2018.

[11] A. Mosavi, P. Ozturk and K. Chau, "Flood Prediction using Machine Learning Models: Literature Review", *Water*, Vol. 10, No. 11, pp. 1-16, 2018.

[12] R. Bunker and F. Thabtah, "A Machine Learning Framework for Sport Result Prediction", *Applied Computing and Informatics*, Vol. 15, No. 1, pp. 27-33, 2019.

[13] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning", *IEEE Transactions on Vehicular Technology*, Vol. 69, No. 1, pp. 140-150, 2020.

[14] A. Ahmad, A. Jafar and K. Aljoumaa, "Customer Churn Prediction in Telecom using Machine Learning in Big Data Platform", *Journal of Big Data*, Vol. 6, No. 1, pp. 1-18, 2019.

[15] S. Khodabandehlou and M. Zivari Rahman, "Comparison of Supervised Machine Learning Techniques for Customer Churn Prediction based on Analysis of Customer Behavior", *Journal of Systems and Information Technology*, Vol. 19, No. 12, pp. 65-93, 2017.

[16] J. Muschelli, "ROC and AUC with a Binary Predictor: A Potentially Misleading Metric", *Journal of Classification*, Vol. 37, No. 3, pp. 696-708, 2019.

[17] M. Pompili, "Critical Appraisal of Major Depression with Suicidal Ideation", *Annals of General Psychiatry*, Vol. 18, No. 1, pp. 1-14, 2019.

[18] K. Korovkinas, P. Danėnas and G. Garšva, "Accuracy and Training Speed Trade-Off in Sentiment Analysis Tasks", *Proceedings of International Conference on Information and Software Technologies*, pp. 227-239, 2018.

[19] M. Fleury, G. Grenier, L. Farand and F. Ferland, "Reasons for Emergency Department use among Patients with Mental

Disorders", *Psychiatric Quarterly*, Vol. 90, No. 4, pp. 703-716, 2019.