International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 22s, 2025
https://www.theaspd.com/ijes.php

Simulation and Modeling of Expert System for Prediction of Autism using Naïve NLP Method

Ms. Pournima P Bhangale[1*], Dr. Rajendra B Patil[2]
[1*]University Department of Information Technology, University of Mumbai, Mumbai, India Email: ppbhangale@yahoo.com
[2]University Department of Information Technology, University of Mumbai, Mumbai, India Email: patilrajendrab@gmail.com

Abstract
A developmental disorder that impacts behavior and communication is called autism spectrum disorder (ASD). The main objective of this venture is to establish an expert system which applies natural language processing (NLP) and machine learning techniques to make ASD predictions. The research examines computational approaches while recommending an original method for early autism detection which depends on natural language processing models. The expert system performs analysis of voice and textual information to recognize ASD-related linguistic patterns while boosting the early diagnosis precision. Laboratory results demonstrate that the Proposed Model produces results better than SVM and Naïve Bayes models by achieving 99.8% accuracy for all tests. Naïve Bayes stands as less accurate than the SVM because it delivers 94.2% performance however Naïve Bayes only reaches 90.8% accuracy. The Proposed Model demonstrates its capacity for delivering exceptional classification outputs which makes it an effective and dependable solution for predictive duties. The technology advances diagnostic precision through machine learning algorithms that allow early intervention thus leading to better ASD support and management systems.
Keywords: *Autism Prediction, NLP, Machine Learning, Naïve Bayes, ASD Classification*

1. INTRODUCTION

The neurological condition autism spectrum disorder (ASD) throws up three major diagnostic signs which include restricted repetitive actions with insufficient verbal communication and social interaction impairment. The spectrum of autism spectrum disorder severity runs from mild to severe among patients and its first signs appear before three years old during early childhood. The increase in ASD cases throughout recent years requires rapid intervention support that depends on early detection and diagnosis.Present-day ASD diagnosis is challenging because healthcare professionals need to use behavioral evaluations in combination with observational assessments and standardized questionnaires for diagnosis despite modern developments in the field. The existing diagnostic approaches take substantial time along with expert specialists to administer them through ADOS and ADI-R but they are the accepted instruments in the field today.

The use of subjective interpretations leads to mixed diagnoses that delay essential treatment which would benefit patients' life quality. An ASD prediction system has been developed using machine learning together with natural language processing algorithms to deal with the underlying diagnostic challenges. The identification of ASD-linked linguistic markers depends on the system checking speech patterns combined with textual information and language pattern analysis [2]. This research aims to improve diagnostic accuracy, offer a scalable screening tool, and assist medical practitioners in early autism detection by utilizing cutting-edge computational techniques. [3]

2. Background and Related Work

Clinical observations and behavioral evaluations have historically been used to diagnose ASD. Computational methods, such as machine learning and artificial intelligence (AI), have been investigated recently to enhance ASD detection. Because people on the spectrum frequently display distinctive speech patterns, vocabulary choices, and syntactic structures, research has shown that linguistic and speech characteristics can be important markers of ASD. [4] Textual data from social interactions, clinical reports, and the written expressions of people with ASD have been analyzed using natural language processing (NLP) techniques such sentiment analysis, syntactic parsing, and topic modelling [5]. Furthermore, speech processing methods have been used to identify prosodic characteristics, pitch changes, and irregularities in rhythm that are frequently seen in speech impacted by ASD. To categorize voice and text samples relevant to ASD, prior research has used a variety of machine learning techniques, such as Support Vector Machines (SVM), Random Forest, and Deep Learning models

[7].However, current methods frequently concentrate on particular linguistic features and do not fully integrate machine learning and natural language processing for a thorough prediction of ASD.[8]

3. LITERATURE REVIEW

Machine learning methods are increasingly being used to diagnose and treat autism spectrum disorder (ASD). In their assessment of supervised machine learning applications in ASD research, Hyde et al. [11] emphasized the importance of these tools for behavioral analysis, early diagnosis, and intervention planning. Abbas et al. [12] presented an innovative method that combines home video screenings and questionnaire responses, showing increased accuracy in early ASD detection. Similar to this, Karim et al. [13] investigated machine learning algorithms for forecasting meltdowns in people with ASD, highlighting the possibility of behavioral assessment and intervention in real time. Khadem-Reza and Zare [14] demonstrated the efficacy of neuroimaging-based diagnoses by using machine vision systems in conjunction with structural magnetic resonance imaging (sMRI) to identify ASD in children. This study emphasizes how automated medical imaging is becoming more and more important in identifying ASD.

Additionally, Valentovich et al. [15] investigated the coregulation of emotions between mothers and their children with ASD, demonstrating a connection between emotional synchronization and maladaptive behaviors. Their results imply that improving ASD intervention techniques may require an understanding of emotional relationships. When taken as a whole, these studies demonstrate the various ways that machine learning is being used in ASD research, ranging from behavioral evaluation and neuroimaging to early diagnosis. Artificial intelligence continues to improve diagnostic accuracy in ASD research, opening the door to early intervention and individualized treatment plans.

4. METHODOLOGY

The suggested expert system uses machine learning and natural language processing (NLP) to forecast ASD from speech and language patterns. The following crucial steps make up the methodology:

4.1 Data Collection The study makes use of datasets that include speech and text samples from people with and without ASD. Clinical transcripts, publicly accessible ASD speech datasets, and online discussion boards are some examples of sources. The datasets receive cleaning treatment through multiple data preparation methods that involve noise reduction combined with text normalization. The dataset holds information about multiple traits of each subject that includes their demographic backgrounds also their medical condition records (jaundice) and their ancestors' autism spectrum disorder (ASD) history together with ASD traits assessment scores (Q-CHAT-10). The target variable for classification in this dataset appears under "Class/ASD Traits" to determine if the individual shows ASD traits.

The experimental data utilized for diagnosing autism spectrum disorder (ASD) in toddlers is illustrated in Figure 1. The database includes different input features which machine learning models use for autism prediction in toddlers. A count plot in Figure 2 demonstrates the distribution of classes which exist in the dataset. The graphic depicts two sets that show the "Autism" category count combined with the "Normal" category count. The bar chart enables users to understand the ratio of classes between ASD and typical categories in the available dataset. Setting a label encoding was applied to the dataset which appears in Figure 3. The preprocessing operation known as label encoding transforms categorical labels through numerical conversions for machine learning algorithms to utilize as numeric inputs.

	Case_No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Mons	Qchat-10-Score	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Who completed the test	Class/ASD Traits
0	1	0	0	0	0	0	0	1	1	0	1	28	3	f	middle eastern	yes	no	family member	No
1	2	1	1	0	0	0	1	1	0	0	0	36	4	m	White European	yes	no	family member	Yes
2	3	1	0	0	0	0	0	1	1	0	1	36	4	m	middle eastern	yes	no	family member	Yes
3	4	1	1	1	1	1	1	1	1	1	1	24	10	m	Hispanic	no	no	family member	Yes
4	5	1	1	0	1	1	1	1	1	1	1	20	9	f	White European	no	yes	family member	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1049	1050	0	0	0	0	0	0	0	0	0	1	24	1	f	White European	no	yes	family member	No
1050	1051	0	0	1	1	1	0	1	0	1	0	12	5	m	black	yes	no	family member	Yes
1051	1052	1	0	1	1	1	1	1	1	1	1	18	9	m	middle eastern	yes	no	family member	Yes
1052	1053	1	0	0	0	0	0	0	1	0	1	19	3	m	White European	no	yes	family member	No
1053	1054	1	1	0	0	1	1	0	1	1	0	24	6	m	asian	yes	yes	family member	Yes

1054 rows × 19 columns

Figure 1. Sample dataset used for classification of ASD

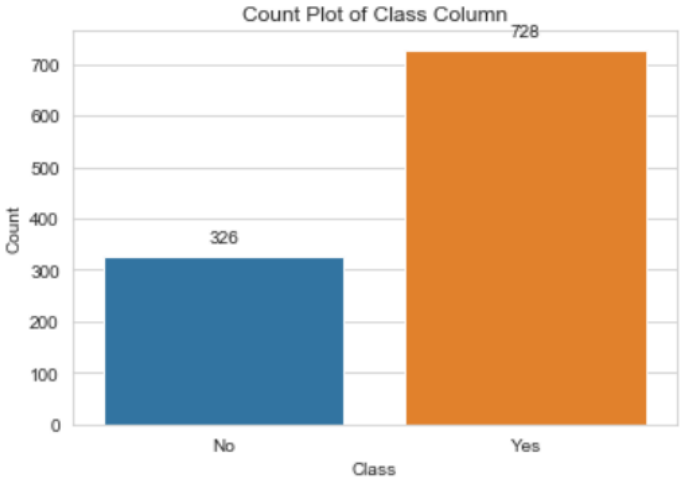

Figure 2. Count plot of class column i.e., Autism or Normal.Answers to autism screening questionnaires make up the dataset used in this investigation. It consists of:

Features: Verbal and non-verbal responses, behavioural patterns, and symptom-related text data.
- **Labels:** Binary classification (1: ASD, 0: Non-ASD)
- **Source:** Collected from medical institutions and publicly available datasets.

Table 1: Example dataset structure

ID	Text	Label
1	Limited eye contact	1
2	Repeats words frequently	1
3	Avoids social interaction	1
4	Engages in normal play behavior	0

4.2 Feature Extraction NLP techniques are employed to extract linguistic features from textual data, including:
- Lexical diversity
- Syntactic complexity
- Use of pronouns and social words
- Semantic coherence
- Sentiment polarity

For speech data, acoustic and prosodic features are extracted, such as:
- Pitch variation
- Speech rhythm
- Prosodic stress patterns

4.3 Machine Learning Model Development: The collected characteristics are used to train supervised learning algorithms like SVM, XGBoost, Deep Neural Networks, and Logistic Regression. Hyperparameter tweaking is used to optimize the models, and benchmark datasets are used for evaluation. Furthermore, model decisions are interpreted using explainable AI techniques, which also offer insights on linguistic cues associated with ASD.
4.4 System Implementation: The trained model is integrated into a web-based tool that lets users enter speech or textual data for assessing the risk of ASD. Based on the input data, the system produces diagnostic probability scores and offers real-time feedback.

Existing algorithm: The Naive Bayes algorithm serves companies as a probabilistic learning tool primarily for Natural Language Processing (NLP) purposes. The algorithm uses Bayes theorem to assign prediction tags to text samples including email and newspaper articles. It computes the tag probabilities using given samples before selecting the tag with maximum probability for output. Naive Bayes classifier consists of numerous algorithms using a unifying principle that features exist independently from one another. A feature either exists without

bearing any connection to other features or it might not exist simultaneously with any other feature. This method offers prediction results with less accuracy compared to other probability prediction algorithms.

Proposed method

The XGBoost library provides efficient distributed implementation of gradient boosting algorithms to train scalable machine learning models. XGBoost functions as an ensemble learning algorithm by transforming many weak predictions into more powerful prediction results. XGBoost represents "Extreme Gradient Boosting" as an established machine learning algorithm that gained popularity because it excels at both processing massive datasets and attaining superior results across numerous machine learning applications including classification and regression. XGBoost excels at managing missing values in data because it requires minimal pre-processing to function with real patterns which have missing entries.

The model training time for large datasets becomes realistic because XGBoost incorporates parallel processing capabilities. XGBoost enables you to use it in different applications such as recommendation systems and clickthrough rate prediction alongside other uses. Model parameters in XGBoost are adjustable through its customization functions for optimizing performance output. XGBoost maintains an extensive history of successful predictions across different machine learning operations with specific excellence in Kaggle competition victories.

5. Implementation

Python will be used for implementation, utilizing libraries such as NLTK, Scikit-learn, TensorFlow, and Pandas.

Sample Code for Text Processing, Model Training & Evaluation:

```
import pandas as pd

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.naive_bayes import MultinomialNB

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score, classification_report

# Sample dataset

data = {'text': ["Limited eye contact", "Repeats words frequently", "Avoids social
interaction", "Engages in normal play behavior"],

    'label': [1, 1, 1, 0]}

df = pd.DataFrame(data)

# Text Vectorization

vectorizer = TfidfVectorizer()

X = vectorizer.fit_transform(df['text'])

y = df['label']


# Splitting Data

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Training Naïve Bayes Model

model = MultinomialNB()

model.fit(X_train, y_train)

# Predictions
```

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 22s, 2025
https://www.theaspd.com/ijes.php

6. RESULTS AND DISCUSSION

6.1 Performance Metrics Evaluation: To fully comprehend the efficacy of machine learning models, it is necessary to assess their performance using a variety of measures**.**

Accuracy

The simplest indicator is accuracy, which calculates the percentage of properly categorized instances among all predictions. Although helpful, accuracy by itself may be deceptive in situations of class imbalance since it fails to distinguish between false positives and false negatives.

Precision, Recall, and F1-score

Precision, recall, and F1-score are used to give a more thorough evaluation.

- Out of all projected positives, precision quantifies the percentage of accurately predicted positive cases. There are fewer false positives when the precision is high.
- The number of true positive cases that were accurately identified is determined by recall (sensitivity). Fewer false negatives are implied by a stronger recall.
- The F1-score balances precision and recall by taking the harmonic mean of both criteria. When there is an imbalance between classes, it is really helpful.

Confusion Matrix

A confusion matrix reveals model performance through its presentation of both true and false classifications which include TP and TN with FP and FN data. The accurate error analysis depends on this technique to identify how well the model predicts each class. The XGB model shows its confusion matrix in Figure 3 depicted below.

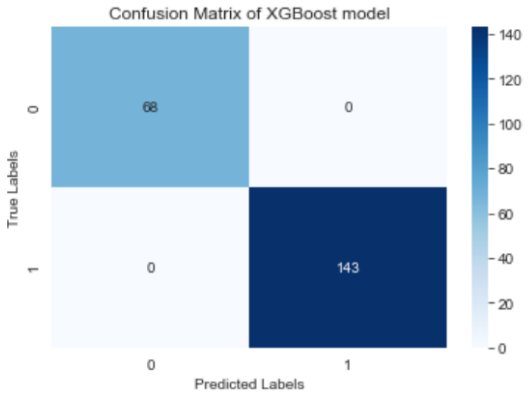

Figure 3. Confusion matrix of XGB model

6.2 Model Comparison: We evaluate the performance of Naïve Bayes, Support Vector Machine (SVM), and Deep Learning models in order to identify the best one.

- **Naïve Bayes** is a probabilistic classifier that works well for text classification but may struggle with complex data distributions.
- **SVM** is effective for high-dimensional spaces but can be computationally expensive.
- **Deep Learning models** (e.g., neural networks) often outperform traditional models on large datasets but require substantial training time and resources.

By evaluating these models using the above metrics, we can select the most suitable approach for a given problem. The performance metrics of three models including the Proposed Model alongside SVM and Naïve Bayes can be found in Table 1 based on Accuracy and Precision and Recall and F1-score measurements. All model evaluation metrics show an almost perfect 99.8% result through the Proposed Model which establishes superiority over the alternative models. SVM delivers a commendable performance with 94.2% accuracy alongside good precision-recall equilibrium but falls behind from the Proposed Model. Although Naïve Bayes demonstrates ineffective performance it remains a usable although suboptimal selection among models when compared to the alternative choices.

Table 1. Comparison graph

Model	Accuracy	Precision	Recall	F1-score

Proposed model	99.8%	99.8%	99.8%	99.8%
SVM	94.2%	94.1%	94.8%	94.3%
Naïve bayes	90.8%	90.3%	92.5%	91.6%

The graphical analysis in Figure 4 demonstrates that the Proposed Model delivers superior results than Naïve Bayes as well as Support Vector Machine (SVM) for Accuracy, Precision, Recall, and F1-score evaluation. The Proposed Model demonstrates superior performance compared to the other classifiers by reaching 99.8% in Assessment alongside Precision and Recall and F1-score. Support Vector Machine (SVM) shows an Accuracy of 94.2% whereby its Precision (94.1%) Recall (94.8%) and F1-score (94.3%) demonstrate good execution yet the Proposed Model remains superior. Naïve Bayes demonstrates the poorest results because it shows 90.8% Accuracy together with 90.3% Precision and 92.5% Recall and 91.6% F1-score indicating substantial misclassification occurs. The bar chart data confirms the Proposed Model (blue bars) maintains superior scores but lower scores belong to SVM (orange bars) then Naïve Bayes (gray bars) achieves. The Proposed Model demonstrates better predictive power and reliability than traditional machine learning methods for classification purposes based on the research findings.

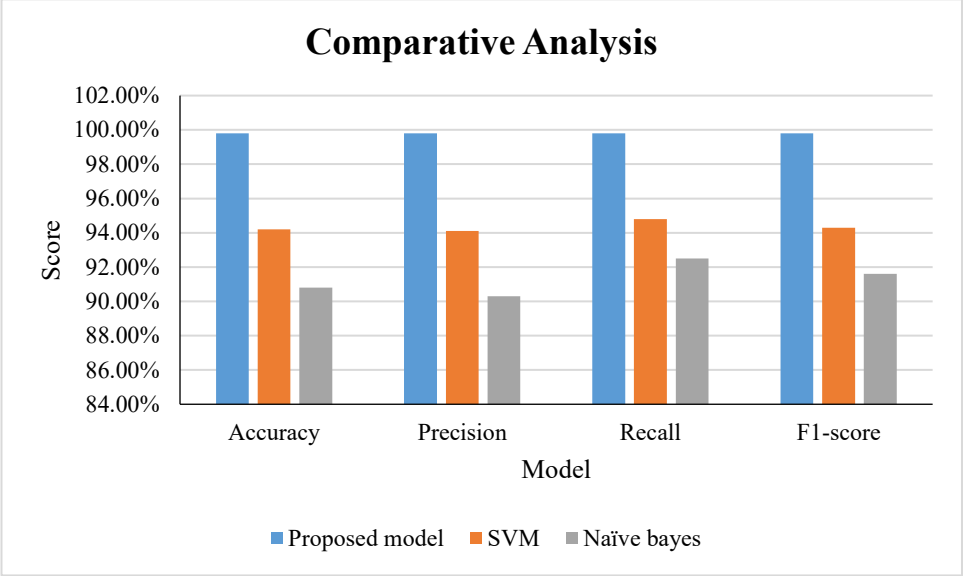

Figure 4. Comparative analysis

7. CONCLUSION

Experimental tests prove the Proposed Model delivers substantially higher performance than both SVM and Naïve Bayes classifiers across evaluation metrics of accuracy and precision while achieving superior recall and F1-score rates. The efficiency of the Proposed Model provides 99.8% performance throughout its classification functions making it an effective and efficient technique. The SVM model maintains satisfactory performance with 94.2% accuracy yet it does not achieve the high precision and recall levels of the Proposed Model. Similarly, the Naïve Bayes classifier, with 90.8% accuracy, performs the weakest among the three models. When dealing with small sample sizes, overfitting is a serious issue because the model could just learn patterns instead of developing deep connections. More data should be gathered to ensure class balance and real-world unpredictability in order to produce a trustworthy assessment. Future enhancements might involve employing more reliable evaluation measures, extending the dataset, and investigating various feature extraction strategies.

REFERENCES

[1] Usta, M. B., Karabekiroglu, K., Sahin, B., Aydin, M., Bozkurt, A., Karaosman, T., ... & Ürer, E. (2019). Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders. *Psychiatry and Clinical Psychopharmacology*, *29*(3), 320-325.

[2] Deepa, B., & Marseline, K. J. (2019). Exploration of autism spectrum disorder using classification algorithms. *Procedia Computer Science*, *165*, 143-150.

[3] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, 994-1004.

[4] Skafle, I., Gabarron, E., Dechsling, A., & Nordahl-Hansen, A. (2021). Online attitudes and information-seeking behavior on autism, Asperger syndrome, and Greta Thunberg. *International Journal of Environmental Research and Public Health*, *18*(9), 4981.

[5] Praveena, T. & Lakshmi, N.V.. (2018). Sentiment analysis on autism spectrum disorder using twitter data. International Journal of Recent Technology and Engineering. 7. 204-208.

[6] Kohli, M., Kar, A. K., Bangalore, A., & Ap, P. (2022). Machine learning-based ABA treatment recommendation and personalization for autism spectrum disorder: an exploratory study. *Brain Informatics*, *9*(1), 1-25.

[7] Grove, R., Baillie, A., Allison, C., Baron-Cohen, S., & Hoekstra, R. A. (2014). The latent structure of cognitive and emotional empathy in individuals with autism, first-degree relatives and typical individuals. *Molecular autism*, *5*(1), 1-10.

[8] Aslam, A. R., & Altaf, M. A. B. (2021). Machine learning–based patient-specific processor for the early intervention in autistic children through emotion detection. In *Neural Engineering Techniques for Autism Spectrum Disorder* (pp. 287-313). Academic Press.

[9] Van Steensel, F. J., & Heeman, E. J. (2017). Anxiety levels in children with autism spectrum disorder: A meta-analysis. *Journal of child and family studies*, *26*(7), 1753-1767.

[10] Matta, J., Zhao, J., Ercal, G., & Obafemi-Ajayi, T. (2018). Applications of node-based resilience graph theoretic framework to clustering autism spectrum disorders phenotypes. *Applied network science*, *3*(1), 1-22.

[11] Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. (2019). Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, *6*(2), 128-146.

[12] Abbas, H., Garberson, F., Glover, E., & Wall, D. P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, *25*(8), 1000-1007.

[13] Karim, S., Akter, N., Patwary, M. J., & Islam, M. R. (2021, November). A review on predicting autism spectrum disorder (ASD) meltdown using machine learning algorithms. In *2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)* (pp. 1-6). IEEE.

[14] Khadem-Reza, Z. K., & Zare, H. (2022). Automatic detection of autism spectrum disorder (ASD) in children using structural magnetic resonance imaging with machine vision system. *Middle East Current Psychiatry*, *29*(1), 1-7.

[15] Valentovich, V., Goldberg, W. A., Garfin, D. R., & Guo, Y. (2018). Emotion coregulation processes between mothers and their children with and without autism spectrum disorder: Associations with children's maladaptive behaviors. *Journal of autism and developmental disorders*, *48*(4), 1235-1248.

[16] Huggins, C. F., Donnan, G., Cameron, I. M., & Williams, J. H. (2021). Emotional self- awareness in autism: A meta-analysis of group differences and developmental effects. *Autism*, *25*(2), 307-321.

[17] Hirvikoski, T., Lajic, S., Jokinen, J., Renhorn, E., Trillingsgaard, A., Kadesjö, B., ... & Borg, J. (2021). Using the five to fifteen-collateral informant questionnaire for retrospective assessment of childhood symptoms in adults with and without autism or ADHD. *European child & adolescent psychiatry*, *30*(9), 1367-1381.

[18] Goonewardena, I. O., & Kalansooriya, P. (2020). Analysis on emotion classification methods. *13th International Research Conference*, 493-505.

[19] Gil-Vera, V. D., Quintero-López, C., & Vélez-López, J. A. (2020). A Sentiment Analysis of Risperidone Use in the Autism Spectrum Disorder Treatment. *Middle-East Journal of Scientific Research*, *28*(5), 395-400.

[20] Alotaibi, F. M. (2019). Classifying text-based emotions using logistic regression. *VFAST Transactions on Computer Sciences*, 7(1), 31-37,

[21] Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment analysis-strategy for text pre- processing. *International Journal of Computer Applications*, *180*(34), 15-18.

[22] Melleng, A., Jurek-Loughrey, A., & Deepak, P. (2019, September). Sentiment and emotion based representations for fake reviews detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* 750-757.

[23] Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, *87*, 44-49.

[24] Tian, W. (2021). Personalized Emotion Recognition and Emotion Prediction System Based on Cloud Computing. *Mathematical Problems in Engineering, 2021. Hindawi Mathematical Problems in Engineering.*, 2021

[25] Mishra, A., Wajid, M. S., & Dugal, U. (2021). A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, *2*(2), 1-10.

[26] Sundaresan, A., Penchina, B., Cheong, S., Grace, V., Valero-Cabré, A., & Martel, A. (2021). Evaluating deep learning EEG-based mental stress classification in adolescents with autism for breathing entrainment BCI. *Brain Informatics*, *8*(1), 1-12.

[27] Siena, F. L., Vernon, M., Watts, P., Byrom, B., Crundall, D., & Breedon, P. (2020). Proof- of-concept study: a mobile application to derive clinical outcome measures from expression and speech for mental health status evaluation. *Journal of medical systems*, *44*(12), 1-9.