Automated Detection of Autism Spectrum Disorder Symptoms using Text Mining and Machine Learning for Early Diagnosis

Mihaela Chistol*, Mirela Danubianu
Faculty of Electrical Engineering and Computer Science, Ştefan cel Mare University, Suceava, Romania

***Abstract*—Autism spectrum disorder (ASD) is a neurological condition whose etiology is still insufficiently understood. The heterogeneity of manifestations makes the diagnosis process difficult. Thus, many children are diagnosed too late, which leads to the loss of precious time that can be used for therapy. A viable solution could be to equip medical staff with modern technologies to detect autism in its early stages. The objective of this research was to investigate, through empirical means, how text mining and machine learning (ML) algorithms can aid in the early ASD diagnosis by identifying patterns and ASD symptoms in text data regarding children's behavior that concerned parents provided. The research involved the design of an innovative technical solution based on text mining for the identification of ASD symptoms in unstructured text data describing children's behavior and the practical implementation of the solution using Rapid Miner. The dataset was created through a controlled experiment with 44 participants, parents of children diagnosed with ASD, who answered questions about their children's (35 boys and 9 girls) behavior. Analysis of the performance of models trained with ML algorithms: Naïve Bayes, K-Nearest Neighbors, Deep Learning and Random Forest revealed that the K-Nearest Neighbors classifier outperformed the other methods, achieving the highest accuracy of 78.69%. Results obtained using text mining and ML demonstrated the feasibility of using parents' narratives to develop predictive models for autism symptoms detection. The achieved accuracy highlights the potential of text mining as an autonomous and time- and cost-effective method for early identification of ASD in children.**

***Keywords*—Text mining; machine learning; artificial intelligence; assistive technologies; Autism Spectrum Disorder; early diagnosis; screening**

I. Introduction

A. Text Mining

Advances in the information technology (IT) industry provide efficient methods for data creation, storage and processing. Global data consumption is growing at an exponential rate, and projections indicate that by 2025, the volume of data used will exceed 180 zettabytes [1]. Today, data means much more than numbers and letters—it includes images, sounds and text. King's research findings [2] indicate that 80% of the world's data is in unstructured format. This statistic highlights the importance of unstructured data processing techniques such as text mining. Text mining is a contemporary concept of computer science that contributes to solving the information crisis by combining data mining (DM), machine learning (ML) and natural language processing (NLP) techniques.

B. Medical Text Mining

The scientific community explored new horizons for the applicability of text mining and understood the utility of this technology in the medical field [3], [4], [5]. The healthcare industry collects enormous amounts of unstructured text information such as patient data, clinical test results, doctor observations and notes. These records, which are typically stored in electronic format, have the potential to enhance the standard of medical treatment by supporting physicians in making well-informed decisions. However, most of this valuable information is unused. One reason for ignoring this data is the lack of appropriate technological tools for processing the large volume of unstructured data. Contemporary advances in IT have contributed to the development of artificial intelligence (AI) algorithms that have facilitated the growth of medical text mining. The term "medical text mining" refers to methods of processing and extracting knowledge from medical text. This area of research combines ideas and techniques from linguistics and health informatics (HI). As emphasized by Dalianis [6], who analyzed the applicability of text mining in the clinical field, text mining is used for NLP, classification, clustering, information extraction (IE) and information retrieval (IR).

C. Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is defined as a complex neurological and developmental disease by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) that is characterized by difficulties with communication and social interaction [7]. The ASD prevalence is estimated at 1 in 100 children according to World Health Organization (WHO) [8]. The high incidence raises a wake-up call to improve the capacity of medical institutions to treat ASD and other developmental disabilities.

D. Challenges of Early Autism Spectrum Disorder Diagnosis

Early autism diagnosis and therapeutic interventions have been reported to be important for achieving satisfactory clinical progress [9]. Conventional approaches to ASD diagnosis involve a comprehensive clinical evaluation and developmental screening with standardized tests, and interviews with psychologists and medical specialists such as Autism Diagnostic Observation Schedule (ADOS) [10], Autism Diagnostic Interview Revised (ADI-R) [11] and Diagnostic

*Corresponding Author.

Interview for Social and Communication Disorder (DISCO) [12]. However, due to the heterogeneity of ASD, the presence of false positive results remains a challenge for clinicians [13]. In addition, clinical examinations are time-consuming and negatively influence the patient behavior because patient is not in the home environment and is surrounded by unfamiliar people and these stimuli may be triggers for crisis-value behaviors.

E. Automated Detection of Autism Spectrum Disorder Symptoms Using Text Mining and Machine Learning

New technologies can help doctors and families who suspect their child may have autism in screening and diagnosis. Text mining provides capabilities to analyze large amounts of unstructured text data related to a child's behavior, development, and interaction.

The present research study findings revealed that text processing algorithms can recognize ASD-specific symptoms in semi-structured screening texts and unstructured texts such as parents' narratives. This scalability holds substantial advantages for doctors, providing them with tools based on text mining technology that can analyze data and flag risk factors or markers of autism. By integrating text mining into diagnostic procedures, healthcare professionals and parents can benefit from early detection of autism symptoms, accurate analysis, and improved support for children with ASD. From this perspective, the contributions of our study are:

- The design of an innovative technical solution based on text mining to identify ASD symptoms in unstructured text data that describes the child's behavior and the practical implementation of the solution that involved a controlled experiment with 44 participants, parents of children diagnosed with ASD.
- Empirical analysis of the accuracy of models trained to identify ASD symptoms in unstructured text data with Naïve Bayes, K-Nearest Neighbors, Deep Learning, Radom Forest ML algorithms.
- Outline of implications for employing text mining in the design and development of healthcare technologies for ASD diagnosis.

F. Paper Structure

The structure of this paper is organized to provide a comprehensive understanding of the design and implementation of the automated ASD symptom detection process using text mining and ML. In the introduction section, we outline the context and highlight the importance of research for early autism diagnosis. Section II describes the methodology used to conduct the research and collect information from parents of children with ASD. Section III brings an in-depth analysis of the results obtained, emphasizing the advantages and limitations. Finally in Section IV, we present the conclusions and outline directions for future research.

II. Materials and Methods

A. Research Questions

The present study's aim was to investigate, through empirical means, how text mining and machine learning algorithms can aid in the early ASD diagnosis by detecting patterns and ASD symptoms in text data regarding children's behavior that concerned parents provided. To address the topic of interest two research questions were formulated and are presented in Table I.

TABLE I. Research Questions

ID	Research Question (RQ)
RQ_1	Can text mining be used for detection of ASD symptoms in unstructured text data describing children's behavior?
RQ_2	What are the empirical results of applying text mining and ML algorithms in terms of accuracy in correctly detecting ASD symptoms?

B. Research Participants

The research of Okoye et al. [14] emphasizes the importance of screening in early childhood, between the ages of 18 and 24 months, to achieve positive results in the therapeutic recovery process. Therefore, the premature age of the children and the reduced reading and writing abilities, at this stage of life, determined the involvement of adults in the experiment. 44 parents of children diagnosed with ASD voluntarily participated in the experiment, of which 27 came from the urban environment and 17 from the rural environment, as represented in Fig. 1.

Autism is not equally distributed between genders, with a male-to-female ratio of 4:1 [15]. Approximately the same gender ratio is also encountered in the children participating in the experiment, 35 boys and 9 girls, as indicated in Fig. 2. It is important to understand the gender distribution because symptoms in girls may manifest differently than in boys, thus leading to under diagnosis of ASD in girls [16].

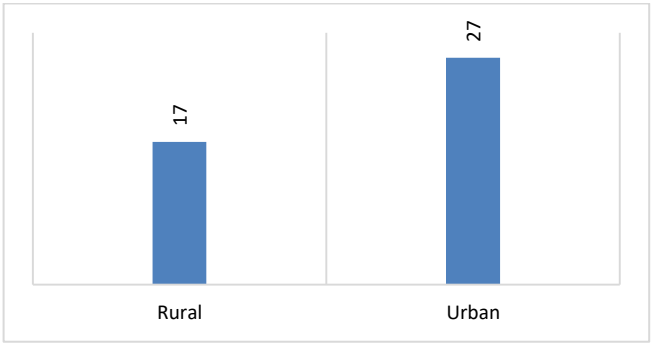

Fig. 1. The distribution of participants according to their residence.

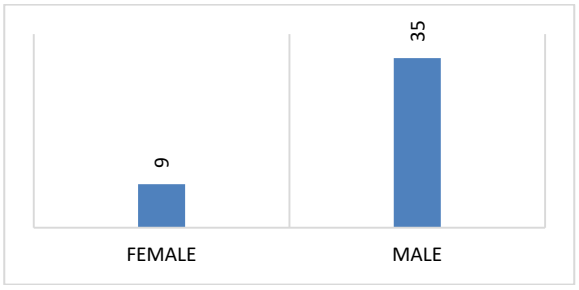

Fig. 2. The gender distribution of children with ASD participating in the experiment.

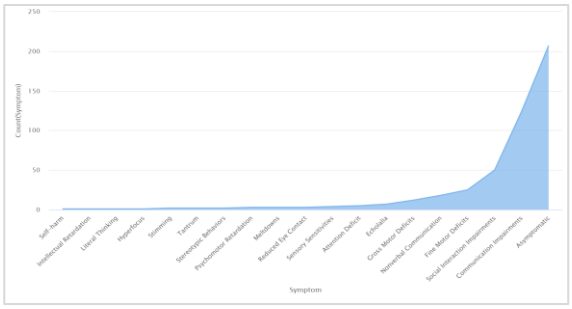

Fig. 3. The quantitative distribution of labels in the dataset.

C. Research Methodology

Following ethics committee approval, participants were invited to answer questions about their child's behavior using the web-based version of Google Forms as instrument. Questions were formulated based on the ASD diagnostic criteria described in DSM-5 [17]. The raw data collected from the participants was analyzed and labeled. The labeling scheme contains 19 labels of which 18 labels represent symptoms specific to autism and one special label "Asymptomatic" which indicates the absence of ASD symptoms. Fig. 3 highlights the quantitative distribution of labels in the dataset. It is noted that the label "Asymptomatic" is the most frequently encountered. This tendency is explained by the fact that the ASD symptom was labeled only in situations where the answer provided by the parent was sufficiently detailed and descriptive to identify that manifestation.

D. Architecture of Technical Solution for Automated Detection of Autism Spectrum Disorder Symptoms using Text Mining and Machine Learning

Diagnosing ASD is a difficult task because the etiology and factors that cause autism are unknown. In addition, the large spectrum of symptoms and the lack of an accurate medical test, such as a blood test, complicate the diagnosis process. Fig. 4 illustrates the conventional diagnosis process, which involves the human factor, represented by the doctor. The doctor analyzes patient biological parameters and applies screening tests and psychological strategies to identify ASD symptoms. The research carried out by Akinnusotu et al. [18] indicate that demanding work conditions affect the psychological state of medical personnel, having an impact on the diagnoses and the effectiveness of recommended treatments. Augmentation of medical workers with modern technologies can be a viable solution in combating burnout. Fig. 5 shows the automated ASD diagnosis process that eliminates the human factor and introduces text mining and ML algorithms to discover autism symptoms in patient data.

The architecture of technical solution for automated ASD symptoms detection in unstructured text data involves complex stages to extract relevant information and build a model capable of recognizing symptoms associated with autism. The stage is inspired by knowledge discovery in text (KDT), a documented research methodology. Fig. 6 present the technical solution implemented using RapidMiner Studio 10.2.

1) Data preprocessing: The data preprocessing is an essential stage in the identification process of ASD symptoms and requires collaboration with qualified medical professionals to draw conclusions about the health status of the participants. These conclusions should be reflected in the labels associated with the text data.

In this stage, records from the dataset that are incomplete and do not provide qualitative information must be removed.

2) Model training: The data preprocessed in the previous phase should be used to train a machine learning model capable of recognizing the ASD symptoms in text data describing the behavior of the child whose parent's suspect autism.

3) Model testing: The model testing stage consists of providing a test dataset as input to the trained model and calculating the performance indicators.

4) Results analysis: The results analysis consists in evaluating the performance of the trained model using the metrics computed in the previous stage, such as accuracy and classification error. Depending on the evaluation results, the model parameters can be adjusted and new machine learning algorithms can be explored.

The automated process of ASD symptoms detection in unstructured text data is empirical and needs to be refined iteratively by experimenting with feature representation, adjusting preprocessing steps and applied machine learning algorithms.

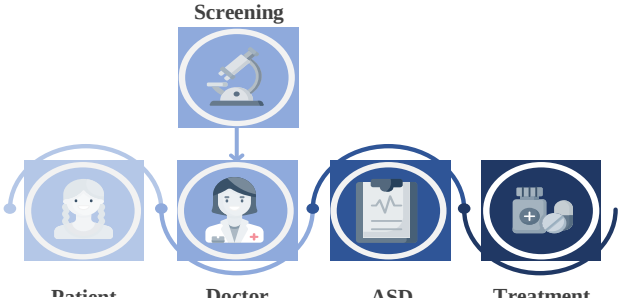

Fig. 4. Conventional ASD diagnosis process.

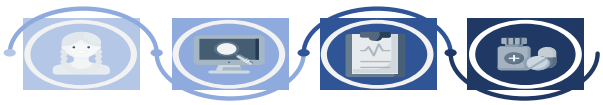

Fig. 5. Automated ASD diagnosis process.

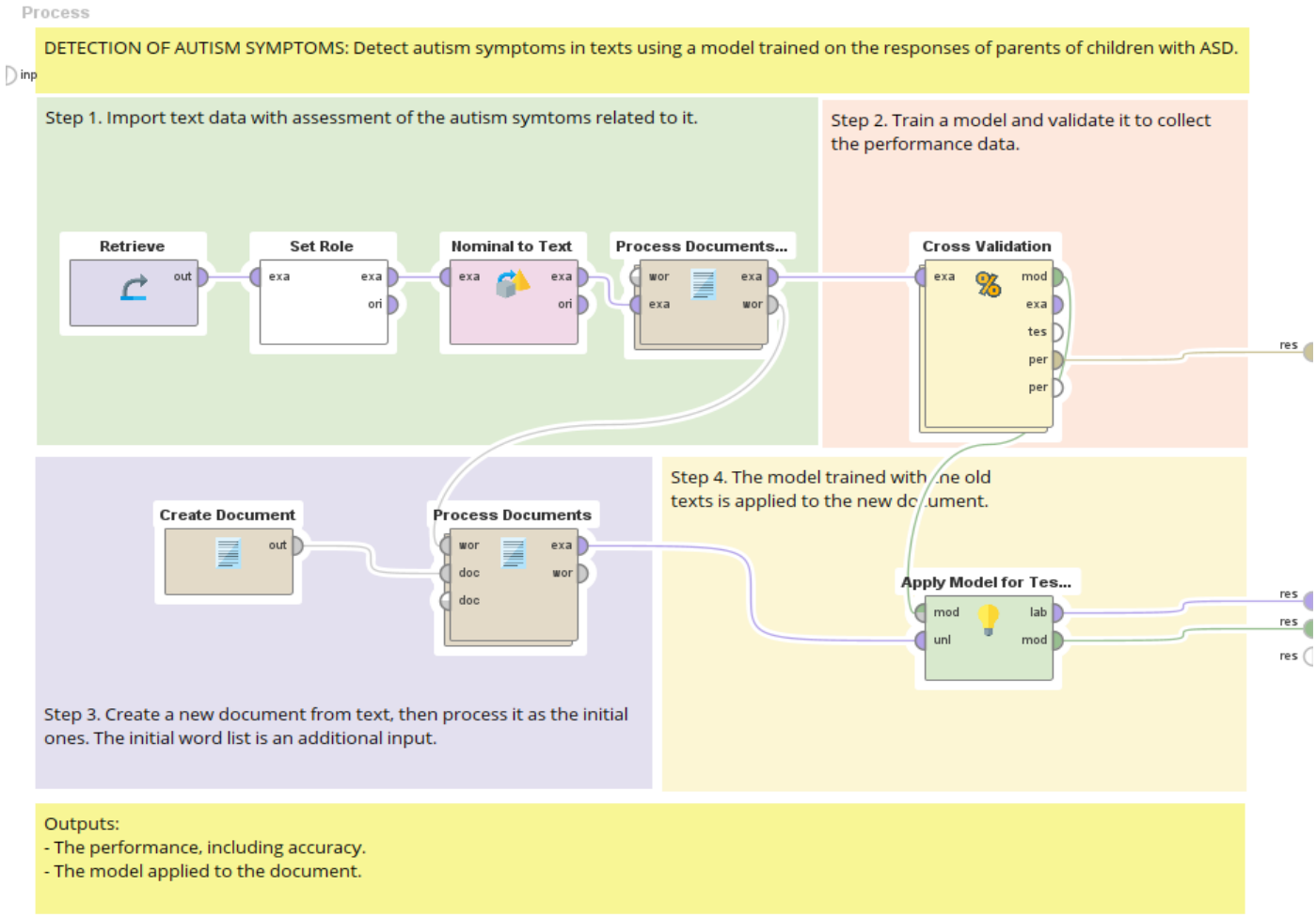

Fig. 6. Automated ASD symptoms detection process.

E. Text Mining

In text mining, the most influential step is text preprocessing, as it prepares the data for mining. Text preprocessing involves cleaning and transforming raw text data into a suitable format for analysis. In the present context, the Text Processing package was used to preprocess the dataset, with a focus on removing words from parents' responses to the questions that did not make a significant contribution to meaning. This approach aimed not only to improve the quality of the extracted information, but also to reduce the size of the vocabulary, thus contributing to the efficient management of the computational complexity of the dataset. Several methods were applied during text preprocessing: capitalization, tokenization, filter stopwords, stemming, n-Grams and term weighting.

1) Capitalization: Bringing all words to a standardized form, such as converting all letters to uppercase or lowercase, to ensure consistency in textual analysis. Capitalization helps reduce the amount of distinct information that ML algorithms have to process. In ADS symptoms detection process the text data was converted a to lowercase (see Fig. 7).

Example

Question: *Describe your concerns about your child's behavior.*
Response: *He does not interact with the other children.*

After capitalization is applied, the text appears as follows:

Question: *describe your concerns about your child's behavior.*
Response: *he does not interact with the other children.*

Fig. 7. Example of applying capitalization to a record from the dataset.

2) Tokenization: Breaking down the text into meaningful element, known as tokens, such as words or phrases, to facilitate manipulation and subsequent analysis [19]. Tokenization is language dependent and involves the removal of punctuation marks from the text. The example in Fig. 8 shows the result of applying tokenization to a record from the dataset. As can be noticed this technique does not take into account words composed by the use of hyphens as "non-verbal". For this reason, engineers must pay attention when using the tokenization technique.

Example

Question: *describe your concerns about your child's behavior.*
Response: *he is non-verbal.*

After tokenization is applied, the text appears as follows:

Question: *{„describe"; „your"; „concerns"; „about"; „your"; „child"; „s"; „behavior"}*
Response: *{„he"; „is"; „non"; „verbal"}*

Fig. 8. Example of applying tokenization to a record from the dataset.

3) Filter stopwords: Exclusion of frequently used or rare words called stopwords to reduce the vocabulary size and focus on key words. Stopwords usually are prepositions, conjunctions, auxiliary verbs and pronouns that do not bring significant contribution to the definition of information.

In ADS symptoms detection process was used the Filter Stopwords operator which has a built-in list of stop words for the English language and contains words such as "at", "etc", "if", "or", etc. The Fig. 9 demonstrates the result of filtering stopwords from a parent's response describing concerns about the child's behavior. Words such as "am", "the", "of", "that", "about", "he", "has", "when", "not" and "always" have been removed from the original text, keeping the meaning intact.

Example

Question: *{„describe"; „your"; „concerns"; „about"; „your"; „child"; „s"; „behavior"}*
Response: *{„i"; „am"; „worried"; „about"; „the"; „fact"; „that"; „he"; „has"; „various"; „fears"; „that"; „he"; „gets"; „angry"; „quickly"; „when"; „things"; „are"; „not"; „like"; „hi"; „wants"; „the"; „fact"; „that"; „he"; „is"; „not"; „always"; „careful"; „that"; „he"; „is"; „not"; „aware"; „of"; „the"; „danger"}*

After stopwords removal, the text appears as follows:

Question: *{„describe"; „concerns"; „child"; „s"; „behavior"}*
Response: *{„i"; „worried"; „fact"; „various"; „fears"; „gets"; „angry"; „quickly"; „things"; „hi"; „wants"; „fact"; „careful"; „aware"; „danger"}*

Fig. 9. Example of applying filter stopwords to a record from the dataset.

4) Stemming: In linguistic morphology stemming is the normalization process of reducing inflected, derived words to their basic form, the root. The root is the part of a word that is common to all its inflected variants. The process of stemming involves removing prefixes or suffixes. To achieve this we used Porter's stemming algorithm. Porter's stemming algorithm removes suffixes from a word, from the English language, to obtain its root [20]. The algorithm consists in marking the consonants in the word with the letter C and the vowels with the letter V. Thus all words can be represented by the Eq. (1).

$$[C]VCVC \dots [V] \quad (1)$$

5) n-Grams: n-Grams is a text preprocessing method mainly used for feature extraction. An n-Gram is a series of consecutive tokens of length *n* used to capture contextual information and improve the understanding of the text. Fig. 10 shows the result of applying n-Gram to a record from the dataset.

Example

Question: *{„child"; „make"; „sentenc"; „word"}*
Response: *{„child"; „form"; „multi"; „word"; „sentenc"; „help"}*

After n-Grams is applied, the text appears as follows:

Question: *{„child"; „child_make"; „make"; „make_sentenc"; „sentenc"; „sentenc_word"; „word";}*

Fig. 10. Example of applying n-Grams to a record from the dataset.

6) Term weighting: Weighting is the process by which the importance of term in the text dataset is quantified. Each term is associated with a value called a weight, which symbolizes how indispensable it is for the text mining process.

In the automated ADS symptoms detection process, Term Frequency - Inverse Document Frequency (TF-IDF) was used to determine the weights. TF-IDF is a statistical measure intended to quantify the importance of a word in a document or corpus. TF-IDF aims to reduce the influence of common words on the model [21].

F. Machine Learning

ML is a subset of AI that focuses on developing systems capable of learning and making decisions without being explicitly programmed. The automated detection of ASD symptoms in text data involved the exploration of four ML algorithms suitable for training a classification model:

- Naïve Bayes (NB) - The NB classifier is a supervised learning algorithm based on Bayes theorem. The algorithm calculates the probability of each class (labels) and then chooses the class with the most likely probability [22].
- K-Nearest Neighbors (k-NN) - The k-NN is a non-parametric algorithm that operates by finding the K nearest neighbors to a given data point based on a metric such as Euclidean distance. The class or value of the data point is then determined by the average of its K neighbors [23].
- Deep Learning (DL) - The DL algorithm is based on a multi-layer artificial neural network that is trained with stochastic gradient descent using back-propagation. The network can contain a large number of hidden layers consisting of neurons with tanh, rectifier and maxout activation functions [24].
- Radom Forest (RF) - The RF is a ML algorithm that uses an ensemble of decision trees to make predictions. Each decision tree is trained on a different data subset, and the predictions of all trees are averaged to produce the final prediction [25].

In order to identify the algorithm that produces the best performing model for automated detection of ASD symptoms a cross-validation was implemented. Cross-validation is a technique used in ML for assessing and comparing learning

algorithms by dividing data into two segments: one used to train a model and the other used to validate the model [26].

The performance evaluation involved calculation of the metrics: accuracy, classification error, Cohen's kappa coefficient, and weighted mean recall. To determine these metrics the confusion matrix was used. Confusion matrix is a performance measurement where the columns represent the predicted values and the rows represent the actual values. According to Kulkarni et al. [27] confusion matrix contains the following elements:

- True Positive (TP): Instances where the model correctly predicted the positive class.
- True Negative (TN): Instances where the model correctly predicted the negative class.
- False positive (FP): Instances where the model incorrectly predicted the positive class.
- False Negative (FN): Instances where the model incorrectly predicted the negative class.

Using TP, TP, FP, FN, and Observed Accuracy (OA) and Expected Accuracy (EA) the follow metrics were determined:

- Accuracy - The accuracy represents the percentage of correct predictions out of the total predictions and is calculated using (2).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

- Classification error - The classification error represents the percentage of incorrect predictions and is calculated using (3).

$$\text{Classification error} = (FP + FN)/(TP + FP + FN + TN) \quad (3)$$

- Cohen's kappa coefficient – The kappa coefficient is a robust statistical metric that measures the observed versus expected accuracy and is calculated using (4).

$$Cohe's\ kappa\ coefficient = (OA - EA)/(1 - EA) \quad (4)$$

- Weighted mean recall – The weighted mean recall is calculated by taking the weighted mean of negative recall in Eq. (5) and positive recall in Eq. (6).

$$Negative\ recall = TN/(TN + FP) \quad (5)$$

$$Positive\ recall = TP/(TP + FN) \quad (6)$$

III. Results

A. Results for RQ_1

Each child with ASD has a unique behavior pattern and level of severity. Some children exhibit signs of autism in early childhood, while others may develop typically in the first few months or years of life, but then suddenly become withdrawn, aggressive, or lose previously acquired linguistic abilities [28]. Fig. 11 presents most common ASD symptoms displayed by children of study participants. The process of automated detection of ASD symptoms in unstructured text data, representing parents' responses to questions about the behavior of their children, was empirical and involved iterative refinement and experimentation with different feature representations and ML algorithms. The results of the performance metrics analysis revealed that the model trained using the k-NN algorithm produces a high accuracy of 78.69% and is feasible for ASD symptoms detection. As we progressed in exploring text mining technology in the ASD field, we identified several important implications for future research:

- The efficacy of the text mining process is contingent upon the size of the data. Identifying ASD symptoms proved to be more challenging in the concise responses provided by parents.
- The text mining technique is negatively influenced by figures of speech such as metaphors, irony, and euphemism and can introduce ambiguity in the interpretation of language. Future research should focus on addressing this limitation by improving algorithms so that they handle figures of speech more effectively.
- The limited diversity of the dataset requires caution in generalizing findings. Future research should include more diverse linguistic expressions and cultural contexts to increase the model's applicability.

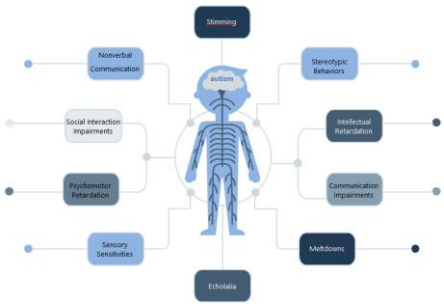

Fig. 11. Most common ASD symptoms exhibited by children of study participants.

The automated detection of ASD symptoms creates opportunities for future research that can further develop healthcare technologies using text mining and ML for autism diagnosis. Automating the symptom identification process provides an efficient alternative to manually reviewing texts written by parents in questionnaires, messages and videos recorded during medical examinations. This efficiency is important for widespread implementation.

B. Results for RQ_2

The results obtained from the computation of the accuracy are presented in Fig. 12. The k-NN algorithm outperformed other methods, achieving the highest accuracy of 78.69%. In contrast, the RF algorithm yielded less favorable outcomes, with an accuracy rate of only 54.25%.

The outcomes derived from the calculation of the classification error are depicted in Fig. 13. Analysis of the classification error revealed that k-NN had the lowest error rate at 21.31%, while RF registered the highest classification error of 45.75%.

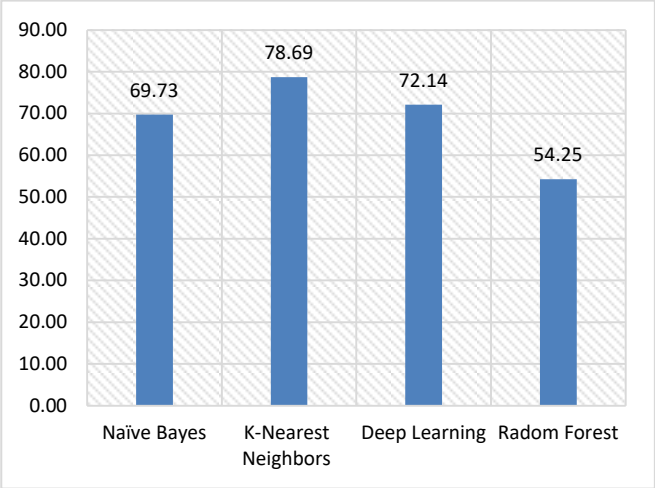

Fig. 12. The accuracy generated by models trained for the automated detection of ASD symptoms.

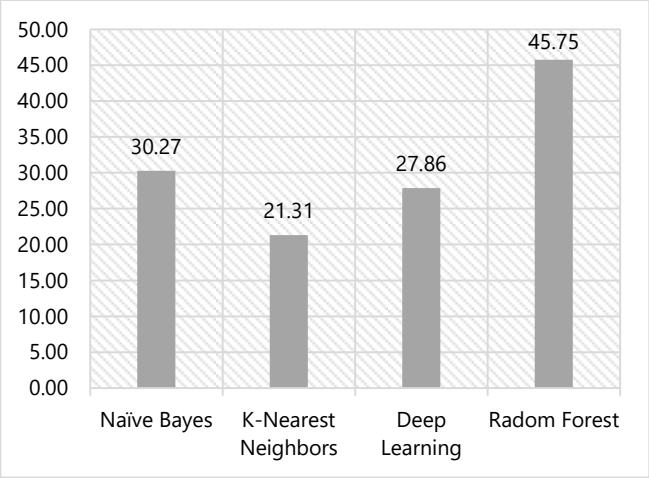

Fig. 13. The classification error generated by models trained for the automated detection of ASD symptoms.

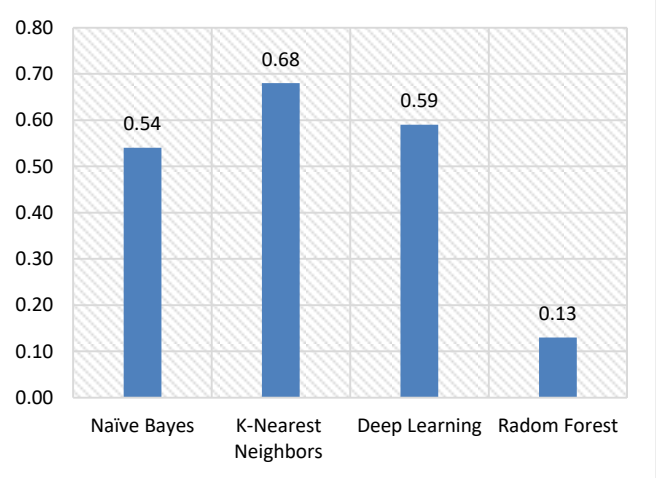

Fig. 14. The Cohen's kappa coefficient generated by models trained for the automated detection of ASD symptoms.

The results obtained after computing Cohen's kappa coefficient are presented in Fig. 14. This coefficient is useful in distinguishing correct predictions that occur by chance and a value below 0.40 is considered unsatisfactory. k-NN demonstrates the stronger association with a kappa coefficient of 0.68.

The results obtained from the computation of the weighted mean recall are presented in Fig. 15 and contribute to the conclusion that k-NN and DL are algorithms that produce more effective models for automated ASD symptoms detection.

Fig. 16 presents the result of testing the model trained using k-NN on a text data that was labeled with the "Communication Impairments" symptom.

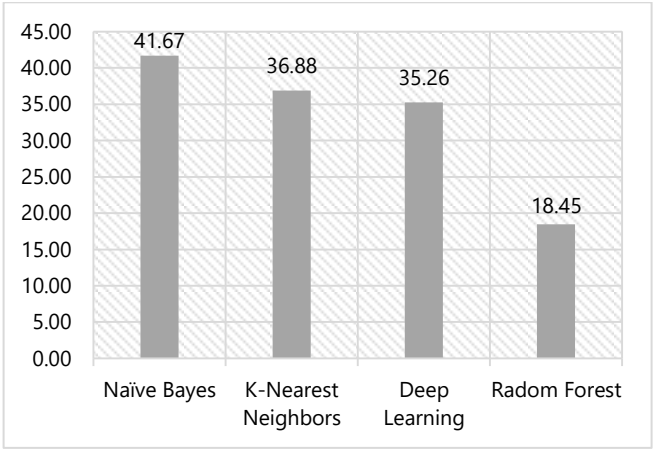

Fig. 15. The weighted mean recall generated by models trained for the automated detection of ASD symptoms.

Example

Question: *Do you think that the way the child speaks is similar to children of his age?*
Response: *The way the child speaks is not similar to that of children of his age because the other children say more words and know more things, and my child barely pronounces some vowels.*

Label*: Communication Impairments*

The output provided by the model trained using the k-NN algorithm is:

Fig. 16. Example of testing the model trained with the k-NN algorithm.

It can be observed from the example that through the application of text mining, the ASD symptom "Communication Impairments" was correctly identified. This is impressive, considering that the text included words and expressions that might have posed a challenge for the trained model, such as "know more" or "say more words", which could be associated with positive behavior and the absence of autism markers. This promising result may have significant implications for text mining integration into applications for autism detection.

IV. Conclusions

Text mining is an efficient IT concept for extracting knowledge from text data. The current study explored text mining techniques and methods in a practical way and focused on analyzing text data provided by 44 parents of children

diagnosed with ASD, trying to identify linguistic patterns and indicators that may contribute to the detection of ASD symptoms. The data collected from the participants underwent labeling, using a scheme that comprises 19 labels. Of these, 18 correspond to ASD symptoms, while the remaining label is designated as "Asymptomatic". The dataset was employed to train four predictive models using ML algorithms, including NB, k-NN, DL and RF.

Results obtained through text mining and ML demonstrated the feasibility of using parents' narratives to develop predictive models for autism symptoms detection. The achieved accuracy of 78,69% highlights the potential of text mining as an autonomous and time- and cost-effective method for the early identification of ASD in children. However, it is important to mention that the ambiguous nature of language can pose challenges in the exploration process and for this reason a representative and diverse training dataset must be employed.

Future research could address the identified limitations and develop healthcare technologies based on the process of detection of ASD symptoms in unstructured text data designed in this study.

ACKNOWLEDGMENT

We would like to express our gratitude to the parents of children with ASD involved in this research for sharing their personal experiences.

REFERENCES

[1] P. Taylor, "The amount of data created, consumed, and stored 2010-2020, with forecasts to 2025." [Online]. Available: https://www.statista.com/statistics/871513/%20worldwide-data-created.

[2] T. King, "80 Percent of Your Data Will Be Unstructured in Five Years," Data Management Solutions Review.

[3] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, "Text mining in healthcare. Applications and opportunities," J Healthc Inf Manag, vol. 22, no. 3, pp. 52–56, 2008.

[4] P. Nitiéma, "Artificial Intelligence in Medicine: Text Mining of Health Care Workers' Opinions," J Med Internet Res, vol. 25, p. e41138, Jan. 2023, doi: 10.2196/41138.

[5] I. Hendrickx, T. Voets, P. Van Dyk, and R. B. Kool, "Using Text Mining Techniques to Identify Health Care Providers With Patient Safety Problems: Exploratory Study," J Med Internet Res, vol. 23, no. 7, p. e19064, Jul. 2021, doi: 10.2196/19064.

[6] H. Dalianis, Clinical Text Mining. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-78503-5.

[7] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. American Psychiatric Association, 2013. doi: 10.1176/appi.books.9780890425596.

[8] "Autism," World Health Organization (WHO).

[9] B. Reichow, K. Hume, E. E. Barton, and B. A. Boyd, "Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD)," Cochrane Database of Systematic Reviews, vol. 2018, no. 10, Art. no. 10, May 2018, doi: 10.1002/14651858.CD009260.pub3.

[10] C. Lord et al., "The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism," Journal of Autism and Developmental Disorders, vol. 30, no. 3, pp. 205–223, Jun. 2000, doi: 10.1023/A:1005592401947.

[11] S. H. Kim, V. Hus, and C. Lord, "Autism Diagnostic Interview-Revised," in Encyclopedia of Autism Spectrum Disorders, F. R. Volkmar, Ed., New York, NY: Springer New York, 2013, pp. 345–349. doi: 10.1007/978-1-4419-1698-3_894.

[12] L. Wing, S. R. Leekam, S. J. Libby, J. Gould, and M. Larcombe, "The Diagnostic Interview for Social and Communication Disorders: background, inter-rater reliability and clinical use," Child Psychology Psychiatry, vol. 43, no. 3, pp. 307–325, Mar. 2002, doi: 10.1111/1469-7610.00023.

[13] M. Briguglio et al., "A Machine Learning Approach to the Diagnosis of Autism Spectrum Disorder and Multi-Systemic Developmental Disorder Based on Retrospective Data and ADOS-2 Score," Brain Sciences, vol. 13, no. 6, p. 883, May 2023, doi: 10.3390/brainsci13060883.

[14] C. Okoye et al., "Early Diagnosis of Autism Spectrum Disorder: A Review and Analysis of the Risks and Benefits," Cureus, Aug. 2023, doi: 10.7759/cureus.43226.

[15] R. Loomes, L. Hull, and W. P. L. Mandy, "What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis," Journal of the American Academy of Child & Adolescent Psychiatry, vol. 56, no. 6, pp. 466–474, Jun. 2017, doi: 10.1016/j.jaac.2017.03.013.

[16] A. B. Ratto et al., "What About the Girls? Sex-Based Differences in Autistic Traits and Adaptive Skills," J Autism Dev Disord, vol. 48, no. 5, pp. 1698–1711, May 2018, doi: 10.1007/s10803-017-3413-9.

[17] American Psychiatric Association and American Psychiatric Association, Eds., Diagnostic and statistical manual of mental disorders: DSM-5, 5th ed. Washington, D.C: American Psychiatric Association, 2013.

[18] O. Akinnusotu, A. Bhatti, C. A. Doubeni, and M. Williams, "Supporting Mental Health and Psychological Resilience Among the Health Care Workforce: Gaps in the Evidence and Urgency for Action," Ann Fam Med, vol. 21, no. Suppl 2, pp. S100–S102, Feb. 2023, doi: 10.1370/afm.2933.

[19] T. A. Mat, A. Lajis, and H. Nasir, "Text Data Preparation in RapidMiner for Short Free Text Answer in Assisted Assessment," in 2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), Songkla, Thailand: IEEE, Nov. 2018, pp. 1–4. doi: 10.1109/ICSIMA.2018.8688806.

[20] V. Mallawaarachchi, "Poter stemming algorithm - basic intro."

[21] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia: IEEE, Oct. 2014, pp. 1–4. doi: 10.1109/ICITEED.2014.7007894.

[22] F.-J. Yang, "An Implementation of Naive Bayes Classifier," in 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA: IEEE, Dec. 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.

[23] "K-Nearest Neighbor(KNN) Algorithm." [Online]. Available: https://www.geeksforgeeks.org/k-nearest-neighbours/.

[24] "RapidMiner Documentation." [Online]. Available: https://docs.rapidminer.com.

[25] S. Wang, C. Aggarwal, and H. Liu, "Random-Forest-Inspired Neural Networks," ACM Trans. Intell. Syst. Technol., vol. 9, no. 6, pp. 1–25, Nov. 2018, doi: 10.1145/3232230.

[26] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in Encyclopedia of Database Systems, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 532–538. doi: 10.1007/978-0-387-39940-9_565.

[27] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," in Data Democracy, Elsevier, 2020, pp. 83–106. doi: 10.1016/B978-0-12-818366-3.00005-8.

[28] "Autism spectrum disorder." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/autism-spectrum-disorder/symptoms-causes/syc-20352928.