Knowledge-Based Systems 276 (2023) 110724

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Towards finding the lost generation of autistic adults: A deep and multi-view learning approach on social media

Mahsa Khorasani [a], Mohsen Kahani [a,*], Seyed Amir Amin Yazdi [b], Mostafa Hajiaghaei-Keshteli [c]

[a] Computer Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran
[b] Department of Counseling and Educational Psychology, Ferdowsi University of Mashhad, Mashhad, Iran
[c] Department of Industrial Engineering, School of Engineering and Science, Tecnologico de Monterrey, Puebla, Mexico

ARTICLE INFO

Article history:
Received 1 October 2022
Received in revised form 27 May 2023
Accepted 9 June 2023
Available online 14 June 2023

Keywords:
Autism disorder
Mental health
Social media mining
Multi-view learning
Transformer learning

ABSTRACT

The detection of mental disorders through social media has received significant attention. With the growing prevalence of Autism Spectrum Disorder (ASD) and the inherent difficulties in diagnosing adults, researchers have attempted to identify undiagnosed adults. Previous studies have primarily concentrated on analyzing ASD characteristics rather than directly detecting ASD. The current study aims to propose a novel framework to assist in identifying the "lost generation" of ASD adults using their social media posts. Combining traditional and deep learning methods makes it possible to model complex aspects of ASD diagnostic characteristics, which have been relatively overlooked in previous studies. To accomplish this, specific formalizations for users' patterns of interest as a main ASD diagnostic characteristic are proposed first. The latent linguistic and semantic features of ASD users' postings are then modeled using deep and transformer-based language models. Finally, all these different aspects are considered together to train a detection model by employing the multi-view learning approach. The experiments show that the feature of idiosyncratic interests has more discriminative power than limited and repetitive interests. The results also indicate that the early fusion of interest-related features along with deep linguistic features outperforms the other examined feature combinations. Additionally, the proposed '$if - iuf$' fusion model demonstrates improved performance in capturing patterns of interests, compared to baselines. These findings suggest the potential application of the proposed framework towards indirectly identifying ASD users on social media, as evidenced by achieving precision and recall rates of 85% and 82% respectively on the used sampled dataset.

1. Introduction

Autism spectrum disorder (ASD) is a complex and common neurodevelopmental disorder characterized by chronic social communication and interaction deficits. The prevalence of ASD has increased considerably in recent decades, giving rise to claims of an "Autism epidemic" [1,2]. While early detection is preferred, a significant population of adults remain undiagnosed and untreated and could benefit from a diagnosis and subsequent support [3,4]. Therefore, several recent studies ([3,5–8]; Huang et al., 2022; De Broize et al., 2022; Lilley et al., 2022) have emphasized the importance of diagnosing ASD in adulthood.

The increasing awareness of autism, broadening of diagnostic criteria, and introduction of the spectrum concept in the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) have contributed to the rising number of adults diagnosed with ASD [9]. The DSM-5 recognizes that individuals with high cognitive abilities and effective coping strategies may not exhibit substantial impairments until adolescence or adulthood and may never seek clinical diagnosis or support [10]. As a result, professionals worldwide are now attempting to identify the "lost generation" of adults with ASD [11], and a 2020 review study on the diagnosis of autism in adulthood called for more research in this area [5]. Identifying people with autism can lead to personalized services, such as the systems proposed for autistic people (Mauro et al., 2022), improving their quality of life.

However, diagnosing ASD in adults is challenging, complicated, and time-consuming [6], and adults seeking a diagnosis are at high risk of psychiatric disorders and suicide ideation [12,13]. Adults face various obstacles in obtaining an autism diagnosis, and their satisfaction with the diagnostic process may vary greatly [5,14,15]. According to a large international study

* Corresponding author.
E-mail addresses: mahsakhorasani@um.ac.ir (M. Khorasani), kahani@um.ac.ir (M. Kahani), yazdi@um.ac.ir (S.A.A. Yazdi), mostafahaji@tec.mx (M. Hajiaghaei-Keshteli).

https://doi.org/10.1016/j.knosys.2023.110724
0950-7051/

with 665 participants, fear of not being believed by healthcare professionals is a significant barrier to getting an autism diagnosis in adulthood [16]. Other barriers may include anxiety about appointments, limited access to knowledgeable specialists, high costs, communication difficulties, lack of trust in professionals, stigma, and the complexity of the healthcare system [5,16].

Due to high morbidity and mortality rates in people with psychiatric disorders and a growing shortage of mental health care providers, there is an urgent need for AI to help identify high-risk individuals with mental health disorders [17]. Additionally, analyzing large amounts of Autism data using machine learning (ML) approaches can provide a comprehensive understanding and better insight into ASD diagnosis [18]. Moreover, there has been a continuous growth of research interest in applying machine learning methods to social media data for mental health [19].

Although ASD individuals have difficulties communicating, the Internet and social media may be a preferred and appropriate medium for their social communication due to the social-spatial distance it provides [20]. In fact, about 80% of adults with ASD use social media [21]. Sharing thoughts, interests, emotions, and behaviors on social media leave unwitting traces of users' traits [22,23], providing a valuable supply of mental condition information that is difficult to obtain through traditional data gathering methods, as stated in recent reviews [24,25]. These large text datasets also enable researchers to employ advanced natural language processing (NLP) techniques like transformers and contextual word embeddings, which have opened up a wealth of new opportunities in AI-based healthcare research [26,27]. Therefore, employing social media data, especially Twitter, could potentially be useful in assisting with ASD clinical diagnosis and treatment.

The current study aims to investigate the potential of employing AI and ML techniques on social media to face the problem of identifying the ASD adults without relying on active patient referral to clinical diagnosis. This is accomplished by proposing a new framework in which key ASD diagnostic characteristics introduced in the psychological literature are formalized, modeled, and incorporated into appropriate learning models. Toward addressing this problem, we constructed a dataset for studying the characteristics of ASD users on social media, as there are currently no public datasets available that include Twitter users who self-identify with ASD conditions. Then, we propose a model to analyze specific patterns of interests and language in self-reported Twitter users with autism spectrum disorders. More specifically, mathematical formalizations are proposed to model *restricted*, *repetitive*, and *idiosyncratic* patterns of interests in ASD users. ASD users' postings are also modeled and extracted using state-of-the-art language models based on the transformer method and attention mechanism.

In the end, we apply some baseline learning methods to assess the discriminative capability of the proposed feature representations in distinguishing between individuals with autism spectrum disorder and those without. The discriminative features are treated as multiple views and learned using multi-view learning methods, which have demonstrated effectiveness in Twitter-user-aspect learning models [28,29]. Additionally, we propose a new heuristic representation called '$if - iuf$' based on the concept of '$tf - idf$' measure in the field of Information Retrieval to facilitate effective fusion of user interest-related aspects, such as patterns and topics of interests.

In a nutshell, in addition to creating a dataset for investigating the problem, our work contributes in the following ways:

1. Proposing concrete mathematical formalizations and ML feature representations, referred to as $ASD - PoI$ Features, for social media users' patterns of interest. These features, based on the psychological literature (especially DSM-IV by Battle [9]), are significant diagnostic characteristics in mental disorders such as Autism, Down Syndrome, OCD, and Mental Retardation. Additionally, we propose a fusion representation called '$if - iuf$' to capture all $ASD - PoI$ and *topics of interests* (ToI) features together, leveraging the '$tf - idf$' metric from the field of information retrieval.
2. Modeling *latent linguistic* and *semantic* characteristics of users' postings, referred to as *Ling-Sem* Features, as Tweet Embedding representation using state-of-the-art language models based on transformers and recursive neural networks.
3. Ultimately, proposing a framework that employs a multi-view learning approach to jointly learn all the extracted aspects of ASD users' characteristics, to investigate the potential of using AI, ML, and NLP methods in identifying ASD users on social media platforms to assist in the detection of undiagnosed high-functioning adults with ASD.

The paper is structured as follows. Section 2 describes the related works done in automatic analysis of ASD disorder. The process of data collection and dataset creation is proposed in Section 3. The key solution is proposed in Section 4. The prevalence of suggested features in users with ASD and their ability to differentiate ASD users from non-ASD users are investigated and discussed in Section 5. In addition, the findings are compared to base models. Section 6 concludes with a review of the findings and recommendations for future studies.

2. Literature review

In the last decade, many studies have applied artificial intelligence techniques to improve researchers' understanding of mental health conditions [17], and they generally entail two categories. The first is traditional machine learning methods, which have been addressed in some current reviews [30,31]. These conventional MLmethods employ various feature extraction and classification techniques and result in good accuracy on structured data. However, they are inevitably limited by the data quality, as stated in [32], which introduces some challenges to consider when employing ML techniques in mental health applications.

The second category of studies has employed deep learning (DL) techniques as one of the most recent generations of AI technologies, which has demonstrated superior performance in assisting mental health providers, as reported in such research [33–35]. When dealing with large volumes of unstructured data, DL methods demonstrate surprising accuracy where feature extraction and classification processes are carried out intelligently and integrally.

In Autism diagnosis using AI, various types of data have been examined. One main category is medical records like the Autism Diagnostic Observation Schedule (ADOS)-Modules and the Autism Spectrum Quotient (AQ-10) dataset, which have been analyzed to identify ASD features [36–38]. Although such datasets are clinically valuable, they are not large enough to be learned by DL techniques. On the other hand, some other studies have turned to collect other types of data, such as mobile application data [39], brain MRI and brain imaging data [28,29,40], facial video and eye tracking data [41–44], and brain EEG signals [45,46], which are more suitable for ML and DL methods.

Although these approaches achieve high accuracy, this is largely due to accurate, structured, and labeled training data samples obtained manually by specialists or by using trusted medical devices and tools. The collection of these types of data through questionnaires or devices and sensors is a significant challenge and requires time and effort as well as active referrals to specialists. However, as noted in some recent research [19,47,48], many studies have recently tended to examine social media data rather than traditional strategies, which have exhibited strong potential as a large and enriching data source in the area of mental health problems.

2.1. Autism detection on the web and social media

As the use of the web and social media has grown in recent years, many studies have examined various aspects of ASD-related communities in online forums, weblogs, and social media. According to the previous research examined in this study, such studies typically follow five phases. We call them "Data Gathering", "Data Preprocessing", "Feature Engineering", "Feature Significance Analyzing", and "Disorder Detection" in the schematic workflow diagram presented in Fig. 1.

Two main goals have been followed by such studies as shown in Fig. 1, phases 4 and 5. The first one is to analyze various features to assess the significant characteristics of ASD users. We name this phase as "Feature Significance Analyzing" and there is a need to pass through the "Feature Engineering" phase. Another Goal is to Propose ML-based frameworks and models to detect ASD users and content based on conventional machine learning, conventional machine learning, or deep learning methods.

The first group examines linguistic and semantic features of ASD-related posts [49–55], topics of interests [50,53], and patterns of interactions and behaviors of those individuals [56] using conventional ML and natural language processing (NLP) methods through the "Feature Engineering" phase. However, research that employs deep learning methods to detect Autism disorder [57–59] does not pass through the "Feature Engineering" and "Feature Significance Analyzing" phases because of the nature of these techniques.

In one of the earliest studies, NLP and machine learning techniques were used to determine the psychological characteristics of autism among web users. The authors have investigated the idea that ASD people use language in ways that may indicate diverse characteristics. However, according to the research, both groups use similar languages [49]. Similarly, another study has compared the subjects and linguistic styles of online communities of people affected by ASD to those stated in standard communities. The study also has investigated latent topics of people's postings and found that this newly introduced feature has more predictive power than linguistic features in classifying blog postings [50].

Both of the previously described studies have used the Linguistic Inquiry and Word Count (LIWC) software [60], a standard dictionary containing over 2300 everyday English terms, to examine language styles. Although using dictionaries like LIWC results in acceptable accuracy, there is a lack of semantic information captured by semantic word embedding-based approaches like word2vec [61] and Bert [62], which have improved the performance of NLP approaches in recent years.

Similar to the prior study [50,53] has aimed at comparing the characteristics of online Autism communities to those of others, using data from Live Journal weblog communities. In addition to analyzing topics and language style, the authors have also considered community sentiment information for the first time when analyzing blog posts. The study has aimed to develop a set of powerful Autism post-predictors. According to the report results, the Autism group's sentiment has a lower valence, suggesting that they are in a bad mood. Also, topics and language styles are good predictors of Autism posts.

In contrast to previous studies, which examined blog posts of autistic communities on the web, recent studies have examined and analyzed social media data, particularly on Twitter. They have claimed that Twitter data offers a useful proxy for identifying and detecting people with ASD. Some of such studies [51,52,54,55] have explored the use of Twitter for data-mining information related to ASD by examining a range of linguistic and semantic aspects of messages posted by individuals interested in ASD. The purpose of these studies was to investigate the possibility of data-mining Twitter messages in order to identify the topics, language, and sentiment of the tweet samples of ASD-related content shared by users through Twitter.

As in many of the related studies in this area, using LIWC and statistical analysis of word usage are the most common approaches employed for sentiment and language style analysis in this category of studies. Whereas [51,54] have focused on ASD-related post-classification, recent research has aimed to detect ASD users of a large online forum based on distinctive patterns in their post content, discussions, and interactions with other users [56]. To the best of our knowledge, the latest study is the only one aimed at detecting ASD users on the web, and it has reported a mean average precision of 37% in detecting ASD individuals.

Several researchers have employed deep learning methods to analyze and detect Autism disorders in Online Social Networks (OSNs) in the era of growing social media use and the emergence of deep learning techniques. A recent study used deep neural network models to analyze the sentiment and emotion of tweets related to autism to raise awareness about the condition's treatment and causes. Comparative analysis of this research has shown that deep neural network models are more efficient than traditional ML and NLP methods for sentiment and emotion analysis [59].

Another research [58] used deep learning techniques like convolutional neural networks and transfer learning to extract and produce patterns for face recognition. The proposed method has been employed on face images of autistic and non-autistic children through internet sources such as websites and Facebook pages. The authors of another recent study have also employed a deep learning model over a collection of Reddit social network posts from mental health communities. They have aimed to identify whether a user's post belongs to a mental state by developing independent binary classification models for a specific mental disorder, including Autism [57].

2.2. Research gap

There are, as an emerging field, a number of significant gaps that need to be addressed by future research. First, the most common mental health conditions addressed include depression, suicide risk, schizophrenia, and Alzheimer's disease. Therefore, there is a significant scope to explore whether ML can have similar accuracy in detecting and diagnosing other mental health conditions, particularly neurodevelopmental disorders or not, such as Autism which has been less considered. One of the most important research gaps in the investigated ASD papers in the literature is that most of the conducted studies have only focused on analyzing ASD-related features of users' posts and classifying ASD-related posts. In contrast, a few numbers have focused on ASD user detection, phase number 5 or "Disorder Detection" in Fig. 1.

In addition, another research gap is the absence of consideration of diagnostic ASD characteristics in the psychological literature. In this context, some recent studies have simply compared NLP features, such as word use rates, word frequency, and part of speech (POS) tags, while some others have looked into the more complicated NLP features like topics of interests, semantic aspects, and linguistic features of messages posted by individuals interested in ASD for tweet classification. To the best of the authors' knowledge, there is no quantitative study in the literature investigating primary diagnostic ASD features from the

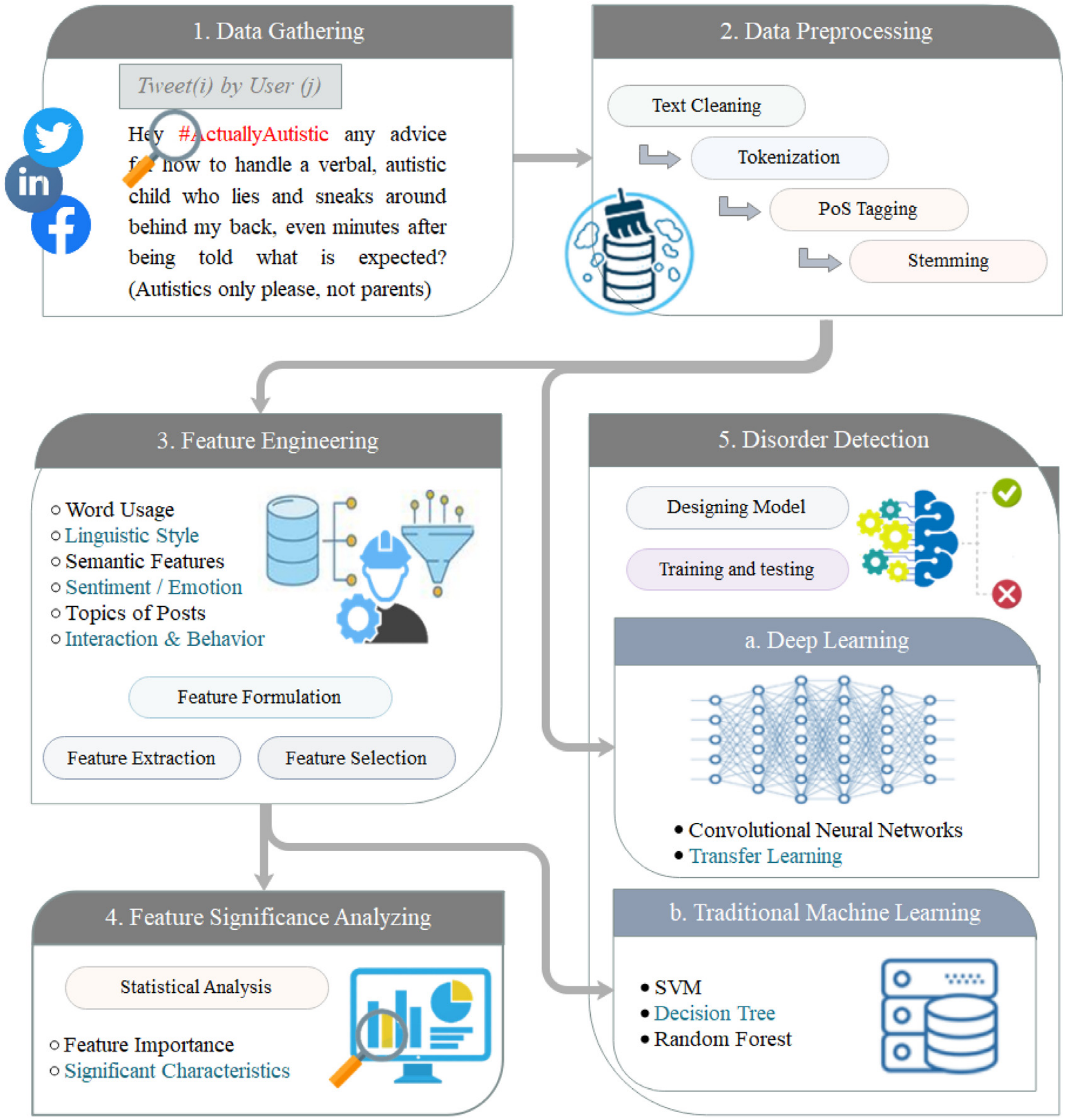


Fig. 1. Common phases of the investigated approaches in the previous studies in employing AI techniques over the web and social media data to analyze and detect Autism disorder.

machine learning perspective, such as special patterns of interests and behaviors of ASD people, as well as communication characteristics.

In the other hand, in the phase of "Disorder Detection", considering various extracted features of ASD users employing feature fusion, attention mechanism, or multi-view approaches could help improve the classification results. Furthermore, it is still possible to use high-performance deep learning methods over social media as a large enough data source to assist ASD detection research problems. Such methods could capture many hidden and unknown features of ASD users together.

Also, deep learning techniques can be used in "Feature Engineering" to model hidden linguistic and semantic features of users' posts by employing state-of-the-art language models like transformer learning which has been missed in the previous studies. A literature review published in [42,43] in the field of applying Natural language processing methods to detect mental illness. The authors investigated Transformer-based methods, which are recent popular large-scale pre-training models, and asserted that the usage of such methods in recent studies shows the potential value of BERT-based models in the application of mental illness detection.

3. Dataset creation

Currently, no appropriate datasets are available for the purpose of identifying ASD users from their tweets, as already mentioned. Therefore, we endeavored to develop an appropriate dataset using the process explained in Fig. 2. In this paper, we use Twitter's search API to identify the probable users of ASD by searching the "#ActuallyAutistic" hashtag, based on an indirect data collection approach. The authors of Wongkoblap et al. [63], as a systematic literature review paper in researching mental health disorders in the era of social media, categorized different data collection techniques into two broad direct and indirect approaches: (1) Using surveys and digital data collection tools to collect data directly from participants with their consent; and (2) collecting data from public posts on social media platforms by searching for relevant posts using regular expressions.

In tweets, the hashtag is frequently used to refer to people with autism spectrum disorders rather than activists, supporters, or parents. Then, a manual check is conducted to ensure the hashtags are used to indicate the user's condition and not for other purposes, such as allusion and addressing. As indicated in the examples of Table 1, it is obvious that *Example-Tweet 1* has not been

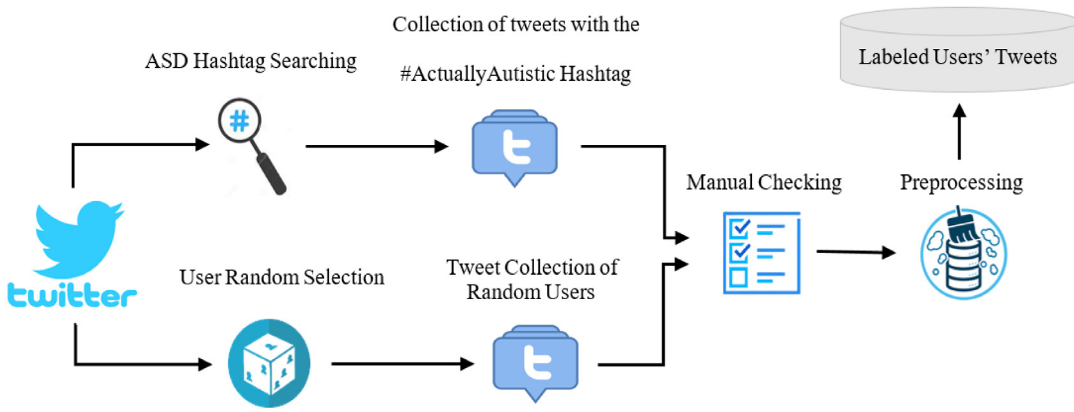


Fig. 2. Dataset development flowchart.

Table 1
Some examples of tweets with the #ActuallyAutistic hashtag.

Example tweet	Tweet content
Example-Tweet 1	Hey #ActuallyAutistic any advice for handling a verbal, autistic child who lies and sneaks around behind my back, even minutes after being told what is expected? (Autistics only, please, not parents)
Example-Tweet 2	I don't need to look at you to hear you. #ActuallyAutistic
Example-Tweet 3	I've always been autistic, but I did not personally know I was autistic until I was 29 years old, almost 30, because my diagnosis was missed for a very large chunk of my life. - Could YOU Be Autistic & Not Know?

Table 2
Dataset statistics.

User type	Number of users	Tweet count
ASD	142 (21%)	337,583 (26%)
Non-ASD	525 (79%)	929,513 (74%)
Total	667	1,267,096

sent by a person with ASD, and the #ActuallyAutistic hashtag is only used to address the issue. Nevertheless, the user's condition, associated with *Example-Tweet 2,* cannot be determined solely from this tweet. So, further user profile investigation would lead to *Example-Tweet 3*, which verifies the user's ASD condition. It should be noted that the reliability of ASD users' declarations about their conditions is presumed in this study.

A control group (negative samples for machine learning classifications) is created by randomly selecting over 500 users from a corpus of 14,464 English Twitter users who are gathered to investigate the role of gender in human identity [64]. To avoid picking inactive users or bots on Twitter, we use the user IDs from this dataset. The Twitter API is used to collect Tweets from these randomly selected users over time, similar to the way we did for ASD people previously. For such users, words associated with Autism (such as "auto", "autism", and "autistic") are searched in their tweets to exclude possible ASD users. The final dataset includes 337,583 and 929,513 tweets from ASD and control users, respectively, covering the 2017–2020 interval (Table 2).

To improve the efficiency of the employed machine learning algorithms, the following steps are carried out to preprocess the text of each tweet:

(a) The Natural Language Toolkit (NLTK) is used for stop-word removal, tokenization, and stemming.
(b) Case folding, which converts all words to lowercase (a–z), is also applied to tweet text content.
(c) To prevent the classification models from being skewed toward Autism-related topics, tweets with terms related to ASD, such as "autistic" and "autism", are removed.

4. The proposed approach

The main objective of the current study is to examine some characteristics of ASD users in order to distinguish them from non-ASD control users. This problem is considered as a classification problem, a supervised machine learning method. The extracted characteristics of ASD users will be used as discriminative features for detecting ASD users. The proposed approach framework is divided into three stages, as shown in Fig. 3, named "User Interest Profile Modeling", "Feature Engineering", and "ASD User Detection".

4.1. User interest modeling

Taking one of the objectives of our work into account, which is identifying users' patterns of interest, we build a concrete representation of a user's interest profile. We first define "*User Tweets*" and "*Int*" as the main concepts.

Definition 1 (*User Tweets*)**.** *User Tweets* denoted as $\boldsymbol{TW}_u = \{tw_1, tw_2, \ldots, tw_m\}$ is a set of m Tweets, which is posted by the user u.

Definition 2 (*Int*)**.** An interest *Int* is represented as a set of discriminative words, i.e., $Int = \{w_1, w_2, \ldots, w_n\}$, where each word has a key role in the *Int* description. Collectively, $\boldsymbol{INT} = \{Int_1, Int_2 \ldots, Int_k\}$ denotes a set of k interest extracted from the collection of users' *User Tweets*.

First, we must infer all users' interests (*INT*) from a corpus of all *User Tweets*. To do so, the well-known Twitter-LDA approach [65] is employed to derive specific topics from the users' Tweets. This approach is a type of topic modeling technique tuned for Twitter content and still is a strong and powerful algorithm for topic modeling on Twitter data employed in recent health-related and social media studies [66–68]. Also, the Online Twitter LDA method has been investigated in the current research to evaluate various topic modeling methods over health-related tweets [68], and has been reported in the list of methods that achieved the best performance. Each identified topic in the Twitter-LDA is considered as interest, described by a weighted distribution of

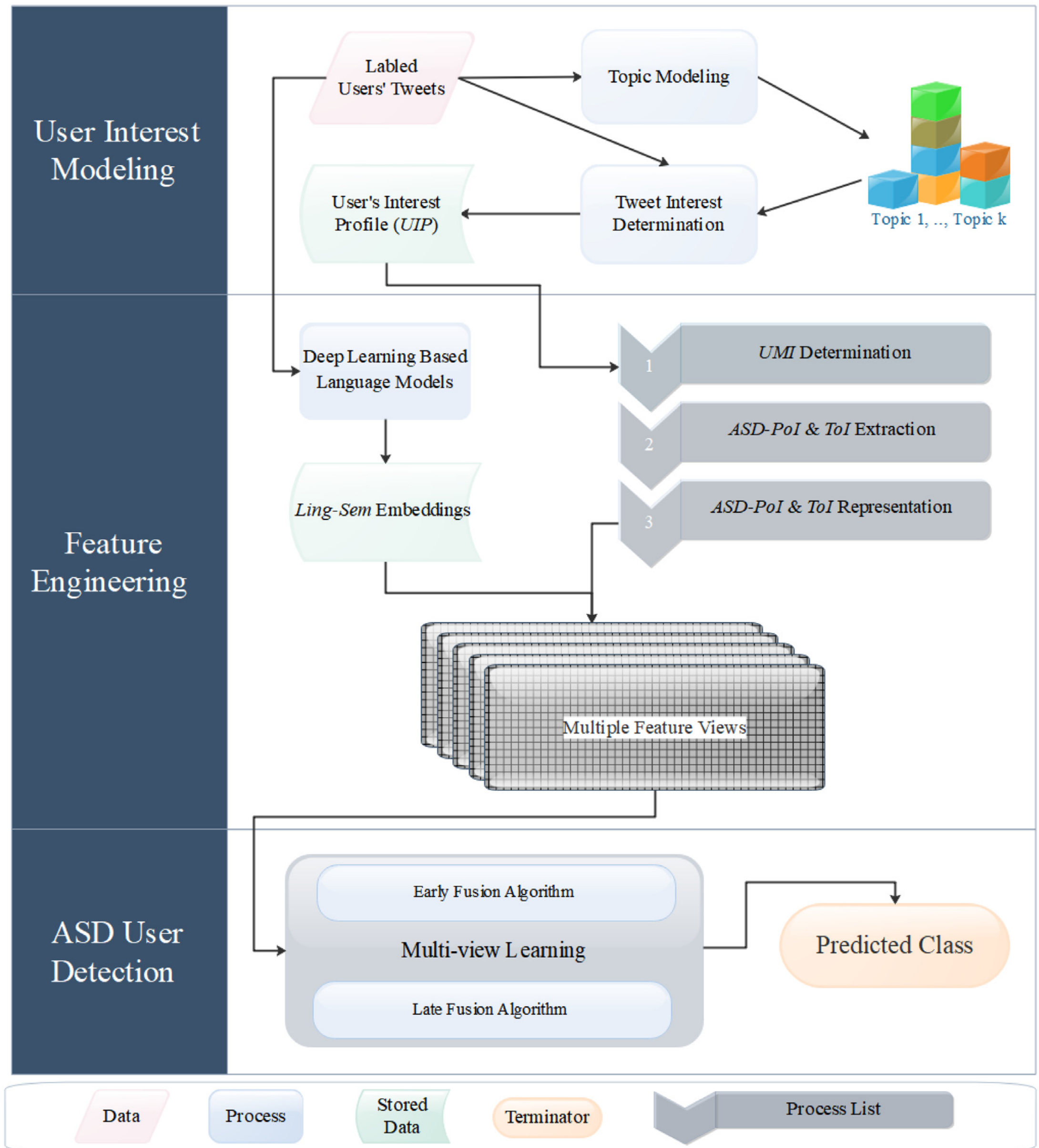


Fig. 3. Overview of the proposed framework.

words, and addressed in one or more Tweets. Because of the short length of tweets on Twitter, one of the underlying assumptions is that each Tweet has only one related interest. Some examples are shown in Table 3. Fig. 4 also shows the word clouds of some extracted interests from *User Tweets* examined in the current study.

In the next step, to identify the user's pattern of interests, the concept of *User's Interest Profile* is proposed in the study, which declares the frequency value of unique interests-ID expressed in *User Tweets*.

Definition 3 (*User's Interests Profile*)**.** Given *INT* and TW_u as the set of k interests and m Tweets of user u, respectively, the *User's Interest Profile* of user u is denoted by a vector as $\boldsymbol{UIP_u} = [f\,(Int_1)\,, f\,(Int_2)\,, \ldots, f\,(Int_k)]$, where $f(Int_i)$ is a function that computes the weight of Int_i in TW_u based on the frequency of Int_i in *User Tweets* of user u.

Table 3
Some examples of extracted users' interests.

Interest (*Int*)	Set of Words in the form of $\{w_1, w_2, \ldots, w_n\}$
Example interest 1	{Player, Football, Game, Committee, NFL, Coach, Rule, Competition, Play, Youth, Season, League, Rookie, Award, Team}
Example interest 2	{Trump, Clinton, Donald, Hillary, Debate, Bernie, Sander, Presidential, News, CBS, Campaign, GOP, Candidate, Vote, Morning, Obama, Win, President, Poll, Nominee, Democratic, Election}
Example interest 3	{Art, Paint, Large, Painting, Abstract, Acrylic, Pop, Fine, Check, Modern, Decor, Upload, Etsy, Canvas, Room, Office, Living, Watercolor, Din, YouTube, Blue, Huge, Add, Black, Minimalist, Playlist, Video, Green}

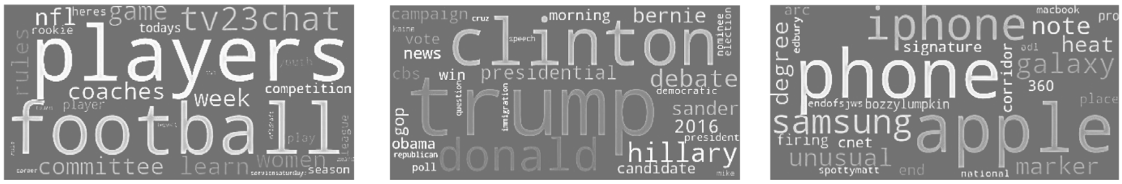


Fig. 4. Word cloud of some examples of extracted interests using twitter-LDA method.

4.2. Feature engineering

In the process of feature engineering, domain knowledge is used to reconfigure data and create "features" that optimize machine learning algorithms. In this step, some of the characteristics of individuals with autism spectrum disorders are modeled according to psychological literature. It is considered a phase of 'Feature Engineering' in which ASD users' characteristics are extracted and modeled as machine learning features. .

The DSM-IV criterion introduces the 'encompassing preoccupation with one or more stereotyped and restricted patterns of interests that are abnormal either in intensity or focus' as the best way to differentiate the interests between individuals with HF-ASD and non-ASD [9]. Also, it is reported that idiosyncratic interests are found in linguistic data of children with Autism [69–71]. In this research, the ASD users' patterns of interest, expressed in Tweets, are investigated to check whether they are different from those of the control group not. In case they are different, they could be used as discriminative features to distinguish users. Features, such as restrictedness, idiosyncrasy, and repetitiveness of users' interests are mathematically defined and extracted from *UIP* and used for user modeling in this study.

Table 4
Definitions of notations and operators in the proposed formula.

Operator/Notation symbol	Operator/Notation name	Definition
$\cdot$	Dot Product Operator	Is used in math equations to represent multiplication.
{}	Set Notation	Is used to represent the elements of a set.
ϵ	set membership Notation	Means "is an element of". So, the statement $x\epsilon A$ means that x is an element of the set A.
$\{x\epsilon A \mid P(x)\}$	Setbuilder Notation	Defines a set whose elements are precisely those $x\epsilon A$ for which $P(x)$ is true
$n(A)$	cardinal number	Shows the number of elements in a set, read "n of A" or "the number of elements in set A."

Definition 4 (*ASD-PoI*)**.** *ASD − PoI Features* denoted as ***Res***, ***Idi***, and ***Rep*** are users' *Patterns of Interests* Features, which denote restriction, idiosyncrasy, and repetitiveness of users' interests, respectively.

Based on Definition 4, three feature representations are modeled in this research to assess the degree of restriction, specificity, and intensity of users' interests. Also, some models are proposed to capture the features, such as topics of interests, as well as linguistic and semantic features of user posts contents, as these features have been highlighted as discriminative features for ASD people detection on social media in earlier studies which have been investigated in the Literature Review section. Table 4 includes operator and notation definitions that are coming in Eqs. (1) to (6).

A. Restrictedness of Interests
To measure the restrictedness of interests of each user, the *Res* feature is defined, indicating the number of user's main interests.
A *User's Main Interests* set (***UMI***) is the collection of interests that are more posted about. To calculate *UMI* for each user, the most frequent interest in UIP_u is considered as Int_{max} with the frequency of $f_u(Int_{max})$ which is named as *MaxU*. The coefficient $\alpha (0 \leq \alpha \leq 1)$ is used to set a threshold. All interests expressed in more than $\alpha \times MaxU$ tweets of user's Tweets are considered as the set of *User's Main Interests*.

$$\boldsymbol{UMI}_u = \left\{Int_i \epsilon INT \mid f_u(Int_i) > \alpha.max\left(UIP^N_{u,i=0}\right), 0 < i < N\right\} \tag{1}$$

$$f_u(Int_i) = UIP_{\boldsymbol{u},\boldsymbol{i}} \tag{2}$$

$$\boldsymbol{Res_u} = n(UMI_u) \tag{3}$$

B. Idiosyncrasy of Interests
The defined *Idi* is a descending vector obtained from the specificity values of the main interests. The fewer the people who mention a specific interest in their tweets, the more the specific interest is. Therefore, the specificity or idiosyncrasy of an *Int* can be defined as the inverse of the number of all users who have this *Int* in their *UIP*.

$$iuf_i = 1/n(\{u \mid Int_i \epsilon UMI_u, f_u(Int_i) > 0\}) \tag{4}$$

$$\boldsymbol{Idi_u} = \text{Descending Order}\left(\{iuf_i \mid Int_i \epsilon UMI_u\}\right) \tag{5}$$

C. Repetitiveness (Intensity) of Interests
The defined *Rep* is a normalized *UMI* descending vector, indicating the vector of a user's normalized number of tweets, corresponding to each *Int* in her *UMI*. As a result, the intensity of users' interests is represented by this vector.

$$\boldsymbol{Rep_u} = Normalized \text{ Descending Order}(\{f_u(Int_i) \mid Int_i \epsilon UMI_u, f_u(Int_i) > 0\}) \tag{6}$$

D. Topics of Interests
Another feature we attempt to investigate the predictive utility is *ToI* representing the users' topics of interest. As it is a categorical feature, we represent this feature as a concatenation vector of numbers of 1-hot vectors. The 1-hot representation model converts each of the interests existing in the *UIP* from a categorical feature to the numerical one.
E. Linguistic and Semantic Embedding
To capture latent linguistic and semantic information of Tweets, we use Bidirectional Encoder Representations from Transformers (BERT) [62] to create tweet embedding vectors (TEV). The semantic space's geometric features of created vectors are semantically and syntactically significant, implying that semantic or syntactic similar documents are likely

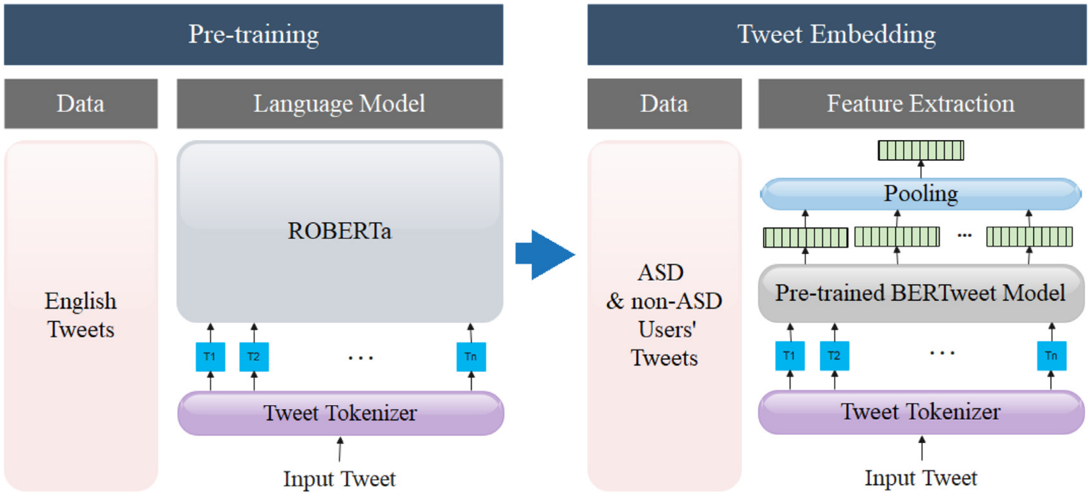


Fig. 5. Using the pretrained BerTweet model, which has created over a 80 GB corpus of 850M English Tweets using ROBERTa language model in [76], to create tweet embeddings for ASD and non-ASD users' Tweets as a feature vector for pro.

to be close. The BERT model is designed to pre-train deep bidirectional representations from the unlabeled text and has been utilized in some studies in the literature, such as [72,73] for mental health analysis and behavior modeling of social media users.

Tweets generally deviate from traditional written content such as Wikipedia and news articles Due to their typical short duration, frequent use of informal grammar, and irregular terminology, such as acronyms, typographical errors, and hashtags [74,75]. As a result, handling text analysis tasks on Tweet data using existing language models that have been trained on large-scale traditional text corpora with formal grammar and regular vocabulary may prove difficult. So, BERTweet pre-trained model, the first large-scale language model for English Tweets, was employed, which has been trained using a 80 GB corpus of 850M English Tweets and outperforms the prior state-of-the-art models on text classification [76].

To create tweet representation or tweet embeddings, each tweet is tokenized using "TweetTokenizer" from the NLTK toolkit [77] and passed through the pretrained BerTweet model, then average pooling-operation is applied on-top of the output contextualized token embeddings as showed in Fig. 5. The average pooling strategy has showed a good impact on forming a sentence representation rather than using [CLS] token as a sentence embedding technique [62] in recent sentence embedding research like [78–80].

4.3. ASD User Detection

To classify the sample users into two groups, we need to train a model based on a collection of labeled sample users. This problem is formulated concretely in Definition 5.

Definition 5 (*ASD User Detection*)**.** Having obtained *User Tweets*, the goal of *ASD User Detection* is to find users whose their *UIPs* satisfy the specific patterns of *ASD−PoI* Features, and have similar *Topics of Interests*, *linguistic* and *semantic* post content features with self-identified ASD users.

4.3.1. Multi-view learning

The input feature space in single-view learning is a d-dimensional vector. However, in this problem, we have three feature vectors as $ASD - PoI$ features, with different meanings and dimensions. In this case, the learning problem is a multi-view learning which has been cited as a successful strategy in learning models for Twitter users [28,42]. One solution to considering these different representations or views of the data is the early fusion method [81]. Based on this method, we try to convert the multi-view problem to a mono-view learning problem (Fig. 6(a)).

The first used early fusion method is to concatenate all the views before feeding them into machine learning algorithms. Besides the concatenation approach, a new heuristic representation is proposed in this study, which merges these 3 views into one N-dimensional vector. To do so, '$tf - idf$' measure in the Information Retrieval area is an inspiration resource in this study. The $tf - idf$ criterion is a statistical measure reflecting how important a word (term) is to a document within a document collection by calculating the term frequency-inverse document.

Now, each 'Interest' and 'User Tweet' has been equivalently considered as a word, and a document, respectively. So, $if_{u,i}$ is defined as the number of times that interest i occurs in UIP_u (i.e., *Rep*). In the same way, iuf_i is defined as the inverse number of users that interest i occurs in their UIP (i.e., *Idi*). Now we have

$$if - iuf_{u,i} = if_{u,i}.\ iuf_i \tag{7}$$

where '$if - iuf_{u,i}$' criterion is the new representation for ASD features of users.

The second solution for dealing with a multi-view problem is to use an ensemble model, which is also known as the *Late Fusion* (or decision-level fusion) approach, in which different models are learned separately for each of the views, and the final predicted label is determined by the majority vote of all model outputs (Fig. 6(b)). The ensemble models as a multi-view learning approach have been with reported outperforming results in recent mental illness detection studies like [82].

4.3.2. Classification model

To distinguish ASD users from non-ASD users, two different classification models have been trained. The first one is the Random Forest (RF) classifier. A random decision forest is an ensemble learning classifier that uses voting/averaging to outperform the decision trees and control over-fitting by constructing a range of decision tree classifiers on different sub-samples of the dataset. A decision tree is one of the most commonly used classifiers due to its simplicity. In each decision tree, each of

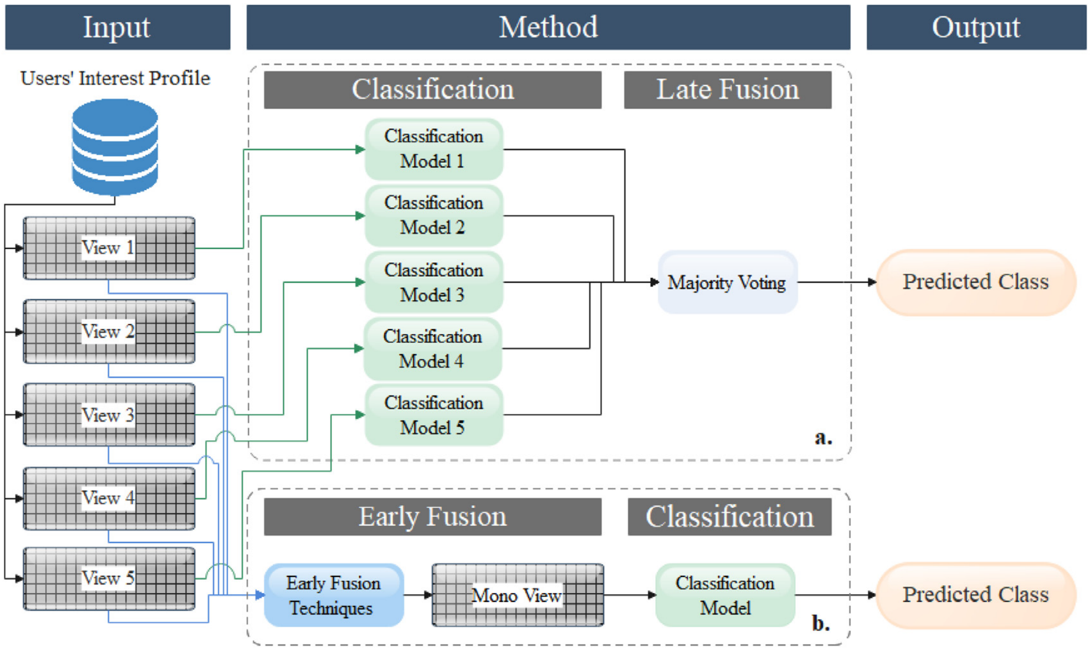


Fig. 6. Multi-view learning for ASD User Detection. **a.** Late Fusion (or decision-level fusion) approach and **b.** Early Fusion (or feature-level fusion) approach (Concatenation or heuristic fusion approach).

the internal nodes represents a "test" on one feature, whereas branches indicate the result of that test. The root-to-leaf paths represent classification rules that lead to the class labels represented by the leaf nodes. The final class label of the random forest classifier is determined by the plurality voting of all expected labels for each sample.

The second classification model is the Support Vector Machine (SVM), a supervised learning model that classifies new samples by separating them into different classes using a hyperplane. Training samples are mapped to points in a particular space in the learning phase. SVM works by maximizing the width of the gap between the two point categories relative to the hyperplane. New samples are mapped into that space and classified into a category based on which side of the gap they fall.

5. Evaluation and discussion

In this phase, a number of experiments have been conducted to evaluate the proposed approach. Python 3.6.5 programming language and NumPy, Sklearn, and Genism libraries were used to code the problems. The prepared dataset, as mentioned in Section 3 was employed for the evaluation. Each proposed feature was assessed separately, and machine learning evaluation metrics were used for this purpose, including Precision (Pre), Recall (Rec), F-measure (F-1), and Accuracy (Acc). These are well-known metrics in the machine learning field for evaluating classification models. The Precision is the percentage of truly labeled samples among retrieved samples, while the Recall is the percentage of truly retrieved samples among all related samples. Also, the F-measure metric is the harmonicmean of Precision and Recall. Finally, the Accuracy is the proportion of truepositives and truenegatives among the total number of test samples.

According to the experimental results, the RF classifier with ten estimators and depth equal to five produces the best results. In addition, due to the imbalance in the group samples, random under-sampling was used to adjust the dataset class distribution to avoid model bias toward the class with the largest number of samples. Also, the 4-fold cross-validation was used and then the whole system ran ten times and its average output, with the standard deviation of σ: 0.009, was considered the final result to achieve more reliable and stable classification results.

In order to calculate the $ASD - PoI$ features and train the learning model with these features, the optimal value of the parameter $\boldsymbol{\alpha}$ must first be determined, leading to the best model outputs. In separate trials, the degree of predictive power for each feature was assessed using different values of the parameter $\boldsymbol{\alpha}$ in Eq. (1). Figs. 7, 8, and 9 depict the final results. As observed in Figs. 7 and 8, increasing the value of the parameter α does not influence the discrimination power degree of the *Rep* feature, but lessens the effect of the *Res* feature in the target groups' distinction. Fig. 9 illustrates that increasing the value of the parameter α has varying effects depending on the learning models used. After conducting these experiments, the optimal α values for each feature and learning algorithm were independently picked based on the resultsdisplayed in Figs. 7 to 9.

Table 5 shows the evaluation outcomes of trained learning models for each $ASD - PoI$ features. Using the proposed features, more than 65 percent of the total accuracy of the system can be achieved in distinguishing between user classes. Due to the capability of all three features to classify users, they were all used in the feature fusion process to feed into the final learning model. However, the feature *Idi* produces better results for detecting users rather than the *Res* and *Rep* features.

To illustrate the differences in $ASD - PoI$ features between ASD and non-ASD users, a comparison of the best evaluation results has been presented in Table 6. The results show that the performance and success rate of the *Res* feature in classifying the samples of both groups are almost identical. However, in the case of the two features *Idi* and *Rep*, there is a significant difference in the precision and recall measures.. The feature *Rep*, which expresses the amount of repetition and intensity of interests, has been able to be almost 17% more successful in recalling samples from the ASD group than in recalling non-ASD samples. This shows that the repetition and intensity of interests in non-ASD users do not follow a specific pattern. While in ASD users, this pattern is more specific and stronger. On the other hand, the *Idi* feature, which indicates the presence of special interests in

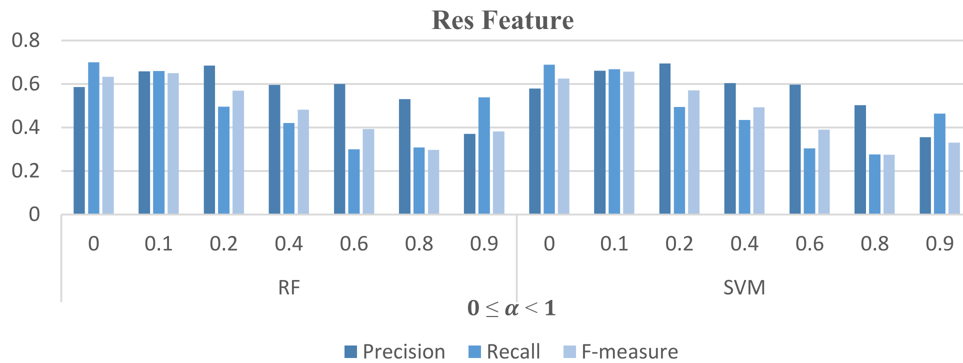


Fig. 7. Evaluation results of classification models based on Res feature for $0 \leq \alpha < 1$.

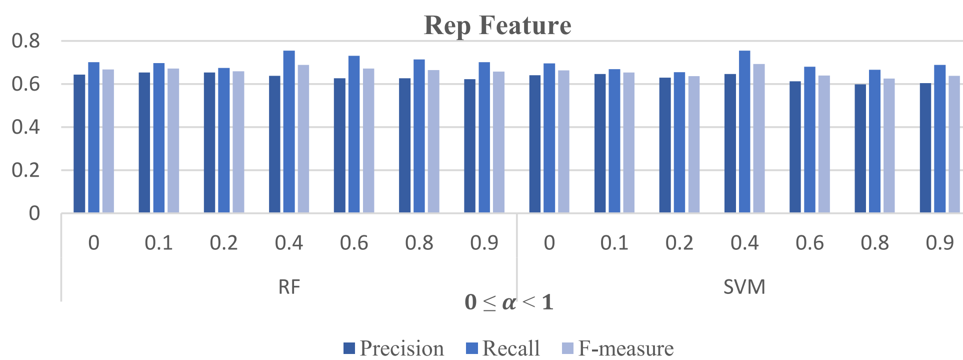


Fig. 8. Evaluation results of classification models based on Rep feature for $0 \leq \alpha < 1$.

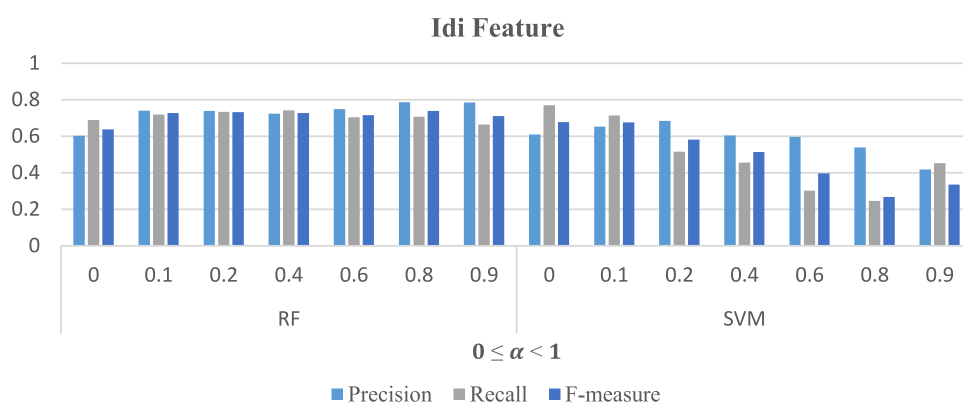


Fig. 9. Evaluation results of classification models based on Idi feature for $0 \leq \alpha < 1$.

autistic people, has been successful in recalling non-ASD users at a rate of about 80%, and this indicates the existence of a pattern of general interests among non-ASD users. Furthermore, the ranking of the success rate of each of the $ASD - PoI$ features in the measures of precision and recall in identifying users from both groups is similar, with *Idi* and then *Rep* outperforming the *Res* feature, respectively.

The 3-view early and late fusion approaches, as described in Section 4.2, are used to differentiate users based on $ASD - PoI$ features, and the results are shown separately for both groups in Table 7. In the Early Fusion approach with RF and SVM learning models, the best α values for *Idi*, *Rep*, and *Res* features are $\alpha = 0.8, 0, 0.1$, and $\alpha = 0, 0.4, 0.1$, respectively. Also, in the Late Fusion approach, the best α values for *Idi*, *Rep*, and *Res* features are $\alpha = 0.8, 0.4, 0.1$, respectively. The results reported in Table 7 show that the ensemble method works about seven to twelve percent better in precision measure in learning these three views of $ASD - PoI$ feature; though still, the *Idi* measure alone has yielded better performance. Comparing the results of Tables 5 and 7 revealed that only Recall and F-measure evaluation metrics in predicting ASD class samples have been improved due to applying the ensemble method to all three features of $ASD - PoI$ as a 3-view representation.

Another feature we attempted to investigate the predictive utility is *ToI* feature, which represents users' topics of interests. The experimental results reported in Table 8 reveal that *ToI* is a powerful feature in discriminating ASD and non-ASD users. These findings suggest that employing *ToI* as the fourth view in the planned multi-view learning model will boost the likelihood of growing model performance. Also, the results show that the *Idi* feature, as opposed to the *Rep*, is more successful in precision for the detection of non-ASD group and recall for the autistic group. And for this reason, as a result of combining these features in the case of 4-views in Table 9, the performance of the model has improved considerably in both of these evaluation criteria compared to before.

Table 5
The evaluation results of RF and SVM models learned by the *ASD − PoI* features based on ML evaluation metrics.

Classifier	*ASD-PoI* features	$0 \Longleftarrow \alpha < 1$	Overall measurement (%)				Measurement on ASD class (%)		
			Acc	Pre	Rec	F-1	Pre	Rec	F-1
Random forest	Idiosyncrasy (*Idi*)	0.8	**75.2**	**76.3**	**75.2**	**75**	**78.6**	70.7	**73.8**
	Repetitiveness (*Rep*)	0	65.6	66.4	65.6	65.5	64.4	70.1	66.8
	Restrictedness (*Res*)	0.1	65.5	66.5	65.4	65.1	65.8	65.9	65
SVM	Idiosyncrasy (*Idi*)	0	63.8	65.4	63.8	63	61	**77**	67.8
	Repetitiveness (*Rep*)	0.4	66.8	67.9	66.8	66.4	64.7	75.5	69.3
	Restrictedness (*Res*)	0.1	65.9	66.9	65.9	65.7	66	66.8	65.6

Table 6
The comparison of the best evaluation results of *ASD − PoI* features in distinguishing ASD from non-ASD users.

Classifier	*ASD-PoI* features	$0 \Longleftarrow \alpha < 1$	Measurement on ASD class (%)			Measurement on non-ASD class (%)		
			Pre	Rec	F-1	Pre	Rec	F-1
Random forest	Idiosyncrasy (*Idi*)	0.8	**78.6**	70.7	**73.8**	**73.2**	**79.7**	**75.8**
SVM	Repetitive-ness (*Rep*)	0.4	64.7	**75.5**	69.3	69.9	58.2	63.5
SVM	Restricted-ness (*Res*)	0.1	66	66.8	65.6	67.8	65.1	65.8

Table 7
Best evaluation results of the 3-view (*ASD − PoI*) learning models based on different fusion approaches and optimum values for $\boldsymbol{\alpha}$.

Fusion approach	Classifier	Overall measurement (%)			Measurement on ASD class (%)			Measurement on Non-ASD class (%)			
		Acc	Pre	Rec	F-1	Pre	Rec	F-1	Pre	Rec	F-1
Early fusion (Concate-nation approach)	Random Forest	68.3	69	68.1	68	67	723	69.1	70.1	64.2	66.5
	SVM	64.1	65.7	64.2	63.6	61.4	76.3	67.7	68.8	52	58.8
Late fusion	Ensemble of Classifiers	**73.9**	**74.6**	**73.7**	**73.6**	**74**	**74.3**	**73.7**	**74.1**	**73.4**	**73.3**

Table 8
The results of ML evaluation metrics on RF and SVM algorithms learned based on users' topics of interests ToI feature.

	Classifier	Overall measurement (%)				Measurement on ASD class (%)			Measurement on Non-ASD class (%)		
		Acc	Pre	Rec	F-1	Pre	Rec	F-1	Pre	Rec	F-1
ToI feature (User Topics of Interests)	Random Forest	66.2	67	66	65.8	67.4	62.9	64.4	65.6	**69.5**	66.9
	SVM	**73**	**75.1**	**72.9**	**72.4**	**68.5**	**85.4**	**75.7**	**80.8**	60.7	**68.8**

Table 9 shows the results of evaluation metrics for the 4-views model. As reported in this table, adding *ToI* to *ASD − PoI* views, using concatenation early fusion and late fusion approach, does not influence the performance of the proposed model. Also, the findings depict that both early and late fusion approaches yield nearly identical results. However, the proposed heuristic fusion '*if − iuf*' outperforms the other employed algorithms. It is concluded from comparing the results of Tables 7, 9, which shows an increase on the total accuracy and precision on ASD-class of late fusion approach, which outperformed concatenation early fusion, from 73.9 to 81.1 and 74 to 88.9, respectively. The reason is that feature vector representation affects performance of a classification model. Using the proposed '*if − iuf*' in *ASD − PoI* and ToI feature fusion instead of concatenation fusion, the precision level has likely increased due to the fused vector's inclusion of knowledge of both topics of interest and topics that users are not interested to.

In the Early Fusion approach (Concatenation approach) with RF and SVM learning models, the best α values for *Idi*, *Rep*, and *Res* features are $\alpha = 0.8, 0, 0.1$, and $\alpha = 0, 0.4, 0.1$, respectively. Also, in the Early Fusion approach (proposed '*if − iuf*' approach) with RF and SVM learning models, the best α values for *Idi*, *Rep*, and *Res* features are $\alpha = 0.8, 0.4, 0.1$, and $\alpha = 0.8, 0.4, 0.1$, respectively. Besides, in the Late Fusion approach, the best α values for *Idi*, *Rep*, and *Res* feature $\alpha = 0.8, 0.4, 0.1$. Though, Fig. 10 shows the distribution of user interests, it is revealed that both groups' users use all of the derived interests in their tweets. Furthermore, there is no distinguishable difference in the

Table 9
The best evaluation results of the classification models based on the 4-view fusion approaches and optimum values for α.

Classifier	*ASD-PoI* and *ToI* Features	Overall Measurement (%)				Measurement on ASD Class (%)			Measurement on Non-ASD Class (%)		
		Acc	Pre	Rec	F-1	Pre	Rec	F-1	Pre	Rec	F-1
Early Fusion (Concatenation Approach)	Random Forest	69.1	69.9	69	68.9	68.3	72.1	69.7	70.7	66	67.6
	SVM	63.7	65.6	63.7	62.7	60.8	78.2	68	69.5	49.2	56.9
Early Fusion (*if-iuf* Approach)	Random Forest	**81.1**	**82.9**	**81.1**	**80.9**	**88.9**	**71.4**	**78.8**	**76**	**90.9**	**82.6**
	SVM	80.6	82.2	80.5	80.3	87.7	**71.4**	78.2	75.9	89.8	82
Late Fusion	Ensemble of Classifiers	74.5	75	74.4	74.4	74.7	74.4	74.2	74.3	74.7	74.1

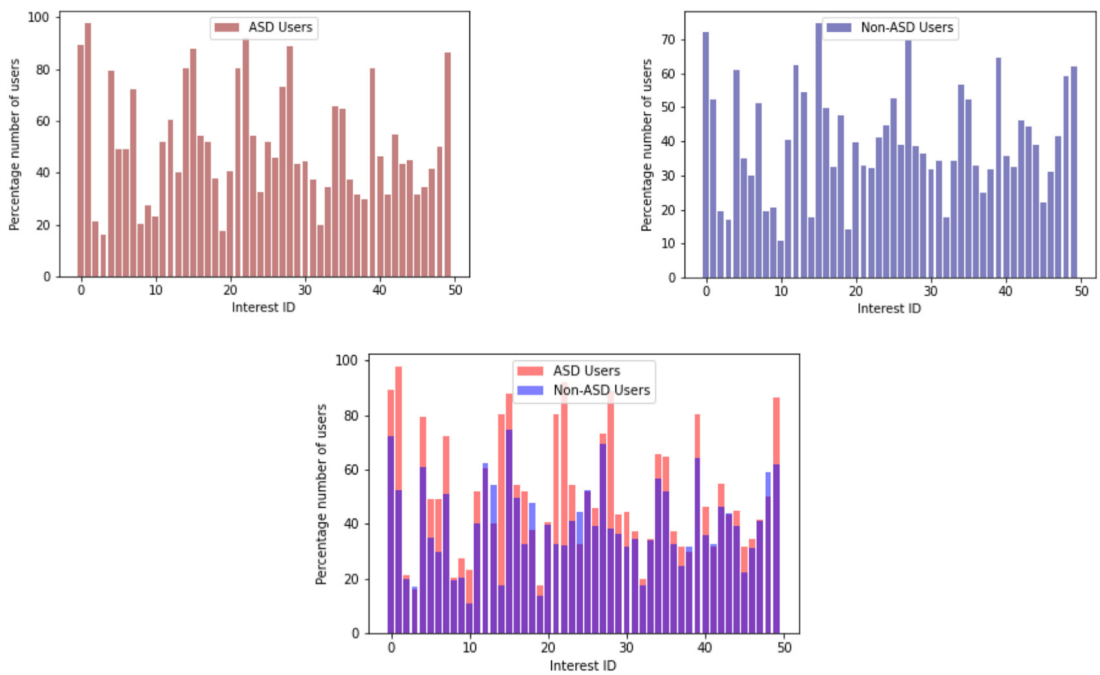


Fig. 10. The distribution of topics of interest among ASD and non-ASD users. The horizontal axis represents Interest-ID and the vertical one is for the number of users have 1 or more Tweets expressed that Interest.

distribution of interests among user groups. Since it produces high-quality results and has fewer dimensions than the concatenation vector mode, this representation can be employed as a powerful final representation to detect ASD users.

The paper also experiments to determine whether the tweet embedding representation covers the knowledge gained from the interest-related characteristics ($ASD - PoI$ and ToI). An ensemble learning based on both of the proposed interest-related 4-views and *Ling-Sem* embeddings is employed using the Late Fusion method described in Section 4. As shown in Table 10, the proposed 5-views learning model, which incorporates all suggested ASD characteristics in this study, outperforms all n-views learning models across all evaluation measures, except for the suggested '$if - iuf$' model which is nearly identical in all evaluation metrics.

5.1. Comparing results with the base methods

Since no previous study has assessed the characteristics of autism on social media (e.g., Twitter) in order to detect individuals with ASD, there is no baseline approach to compare. As a result, we employ two base language models to model the users' Tweets and compare the performance of the suggested method. Also, a topic of interest that has been reported as a good predictors of Autism posts in the previous research [50,53] was considered a base comparing feature.

a. N-gram Language Model
The word n-gram language model, in which each n-gram is a sequence of N tokens from a text corpus, is one of the foundation and base models that has been used in ASD related research [83–85]. This model also, has been used as a base model in a recent research. Many speeches and natural language processing tasks make effective use of N-gram models. In this study, n-gram frequencies are calculated for each collection of *User Tweets* and compared to user categorization.
As shown in Table 11, the *2-gram* language model is capable of user group detection with an overall precision score of 69.7 and a recall score of 60.1 in this study. According to the obtained results, it could be implied that these models demonstrate acceptable effectiveness of the word usage feature, which is captured in the *n-gram* language model, but the addition of other suggested ASD attributes improves the outcomes of the proposed method of this study. Also, experiments on the parameter n in *n-gram* model show that the *2-gram* model outperforms the other methods. The

Table 10
Comparing the results of the proposed interest-related views and their fusion methods, and Ling-Sem Embedding view.

Applied method	Overall measurement (%)				Measurement on ASD class (%)			Measurement on non-ASD class (%)		
	Acc	Pre	Rec	F-1	Pre	Rec	F-1	Pre	Rec	F-1
Best 3-Views: *ASD-PoI*/ Late Fusion	73.9	74.6	73.7	73.6	74	74.3	73.7	74.1	73.4	73.3
Best 4-Views: *if-iuf* / Early Fusion	81.1	82.9	81.1	80.9	**88.9**	71.4	78.8	76	**90.9**	82.6
Ling-Sem Embedding (Bert)	83.7	84.1	83.6	83.6	84.3	83	83.4	83.1	84.4	83.5
Best 5-Views: (*tf-iuf, Ling-Sem*)/ Early Fusion	**84.3**	**84.7**	84.2	**84.2**	85.6	82.5	**83.8**	**83.1**	86.1	**84.4**

Table 11
Comparing results of the best-proposed features/approaches.

Applied Method		Overall Measurement (%)				Measurement on ASD Class (%)			Measurement on non-ASD Class (%)		
		Acc	Pre	Rec	F-1	Pre	Rec	F-1	Pre	Rec	F-1
Best *ASD-PoI* Feature: ***Idi***		75.2	76.3	75.2	75	78.6	70.7	73.8	73.2	79.7	75.8
Best 3-Views: *ASD-PoI*/ Late Fusion		73.9	74.6	73.7	73.6	74	74.3	73.7	74.1	73.4	73.3
Best 4-Views: *if-iuf* / Early Fusion		81.1	82.9	81.1	80.9	**88.9**	71.4	78.8	76	**90.9**	82.6
Ling-Sem Embedding (Bert)		83.7	84.1	83.6	83.6	84.3	83	83.4	83.1	84.4	83.5
Best 5-Views: (*tf-iuf, Ling-Sem*)/ Early Fusion		**84.3**	**84.7**	84.2	**84.2**	85.6	82.5	**83.8**	**83.1**	86.1	**84.4**
Base Feature/ Method	*ToI* (Topics of Interests)	73	75.1	72.9	72.4	68.5	**85.4**	75.7	80.8	60.7	68.8
	TEV (Linguistic & Semantic with Doc2vec)	72.7	73.8	72.7	72.5	70.5	78.7	73.9	76.3	66.8	70.6
	2-gram Model	69.8	69.7	60.1	64.3	71.8	79.6	75.4	69.7	60.1	64.3
	3-gram Model	62.2	52.5	**89.1**	65	80	35.2	46.6	52.5	89.1	65

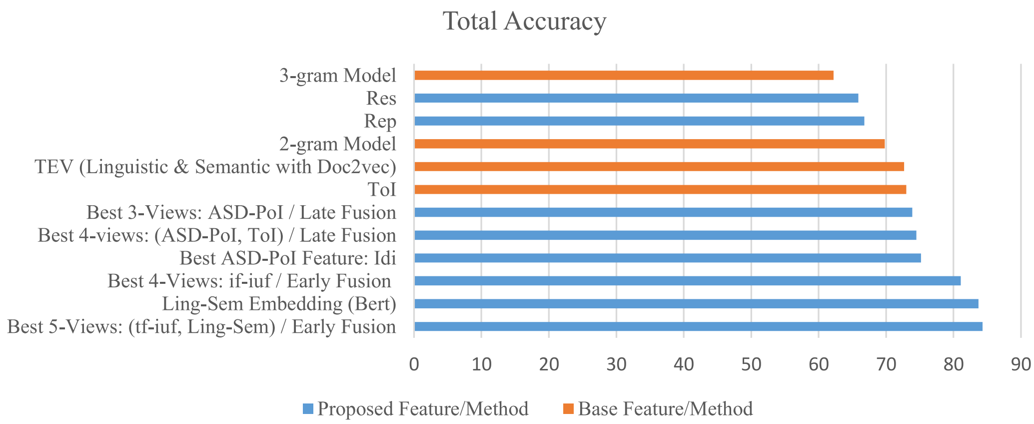
greater the value of n, the poorer the model's performance in *F-1* score.

b. Document Embedding Language Model

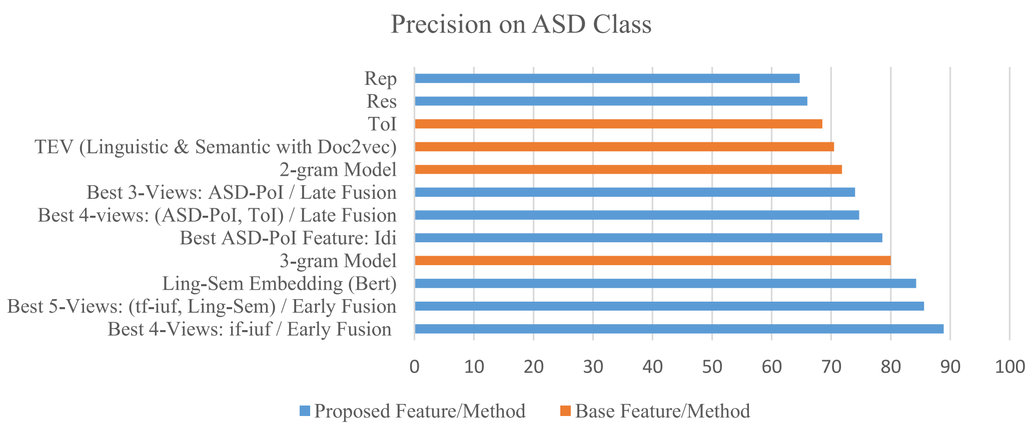
The proposed method consists of providing a computation model to build feature representation for some ASD characteristics of users' patterns of interest to detect ASD users. Reportedly, since there is no similar study in the literature that used such features to compare the obtained results in this study with, a well-known featureless model named Doc2vec [86] was used for comparison with the performance of the suggested method.

The doc2vec model, which captures latent linguistic and semantic information of words and documents, has been used in many previous mental disorder studies in twitter such as [68,87] and was used to create Tweet embedding vectors (TEV) in this study as a base method. This model uses the word2vec model [61,88] to create a document that embeds with adding an extra feature for the document ID. The semantic space's geometric features of created vectors are semantically and syntactically significant, implying that semantic or syntactic similar documents are likely to be close.

As shown in Table 11, the proposed framework in this study could be a proper strategy for recognizing ASD users with the precision of 88.9%. In comparison, the Tweet2vec model has a precision of 70.5% for detecting ASD users as a featureless base solution. Also, Fig. 11 compares the best learning model performance on each proposed feature and their various combination as multi-view features. The comparison of the results shows that although each of the proposed features can identify autistic users with more than 60% on accuracy and precision, combining them with proposed multi-view approaches leads to a better efficiency of up to more than 80% on such metrics.



(a)



(b)

Fig. 11. Comparison the results of the best proposed features and learning models with the baselines in automatic detection of ASD users. (a) Comparison on the accuracy measure. (b) Comparison on the precision measure on ASD class.

5.2. Threats to validity

While our study demonstrates the feasibility of using social media data to identify adults with ASD, it is important to acknowledge the limitations and potential threats to the validity of our findings. The first limitation is the number of positive samples in our dataset. Although, the limitation of the small number of data samples is seen in many research conducted in the field of mental disorders and especially the field of autism identification [43,45,46], it may affect the generalizability of our results. Additionally, the prevalence of ASD in the general population is approximately 1 in 44, and the real-world data may exhibit greater imbalances than the curated dataset, potentially leading to performance degradation.

Furthermore, our study is limited by the variability of language use by different people on social media, which may affect the performance of our proposed model in real-world scenarios. To reduce the effect of the different language characteristics on our model, we have used the BERT language model, which is independent of language and can learn about the structures and semantic of different languages. However, we acknowledge that our dataset is monolingual, and we have recommended that interested researchers examine the effect of language differences in future studies.

6. Conclusion and future work

As one of the most prevalent neurodevelopmental disorders worldwide, autism spectrum disorder (ASD) is a complicated neurodevelopmental disorder. Many individuals with ASD may go undetected due to a lack of qualified clinical evaluations or their families' ignorance because ASD was not well understood in the past. In such situations, automatic diagnosis methods based on machine learning and artificial intelligence using social media could potentially offer valuable assistance, as social media provides a rich resource of human discourse. In this study, we present a framework to help in identifying Twitter users with autism spectrum disorders, based on cutting-edge deep learning and traditional machine learning techniques. Additionally, we explore the specific pattern of interests among ASD users, a key diagnostic characteristic, which has received limited attention in previous studies.

Based on the findings of the study, artificial intelligence and machine learning show promise to assist in detection of high-functioning autism spectrum disorder users on social media. It was observed that the pattern of idiosyncratic, repetitive, and limited interests, introduced as a main diagnostic characteristic of ASD individuals in the autism diagnostic guidelines, is also prevalent among Twitter users with ASD in the sampled dataset. However, the feature of idiosyncratic interests exhibits greater discriminatory power than other investigated interest-related characteristics. Compared to previous studies, this research expands on existing findings by focusing on deep representations of linguistic and semantic characteristics of social media contents posted by individuals with autism spectrum disorders.

The results also indicate that using interest-related features along with deep linguistic and semantic features yielded better performance than employing each of the proposed features individually. Moreover, it was demonstrated how each of the proposed fusion methods for integrating all of the different views of ASD characteristics affects the performance of the classification model to distinguish between ASD users from non-ASD users. The proposed formalized features were evaluated on a manually gathered and verified dataset of self-reported ASD users on Twitter. Additionally, various perspectives were explored, and suggestions for improving the ASD detection process were provided.

Based on the obtained results, the contributions and outcomes of this research suggest the potential promise to support future scientific and empirical advancements in the development of decision support systems (DSSs) and Health Recommender Services (HRS) for identifying adults with high-functioning Autism, enabling broader screening. Moreover, the specific pattern of interests modeled in this study is also a key diagnostic characteristic for other mental disorders such as "Down Syndrome", "OCD", and "Mental Retardation", as mentioned in the psychological literature [89]. Consequently, we hope that the proposed formalizations would facilitate future studies to develop automated methods for detecting such disorders.

Autism spectrum disorders may affect individuals ranging from those who are extremely retarded to those who are brilliant. We may state that the sample data used in this study include normal to high cognitive function people, as they are Twitter users. In this group of ASD individuals, the indicators of communication and social interaction problems are more important than interest-related indicators. Therefore, future work is needed to examine the other two dimensions of the Autism diagnosis. Also, there are other suggestions for interested researchers as future studies:

- Modeling and investigating communication and social interaction of ASD users on social networks
- Investigating the performance of proposed formalization for specific patterns of interest to detect people with other mental conditions such as Down Syndrome, OCD, and Mental Retardation
- Examining the utility of proposed features and methods in detecting disorders, such as depression, anxiety, and bipolar disorder, in which patients' interest patterns become affected
- Considering the time of postings as a possible significant feature using time series methods or attention mechanisms to model changes in interests and behavioral patterns of users with mental disorders
- Examining the impact of crises like the Covid pandemic on patterns of interests and behaviors of ASD users

Despite the limitations discussed in the Threats to Validity section, our study provides a promising proof-of-concept approach for identifying adults with ASD, but further research is necessary to address the potential threats to the validity of our results. In particular, further studies should aim to collect larger and more representative datasets, which include individuals with ASD with varying levels of functioning, socio-demographic characteristics, and language use patterns. Additionally, future studies could aim to develop more advanced machine learning algorithms that can effectively account for the variability of language use on social media.

In conclusion, while our study investigate the potential capacity of using social media data to identify adults with ASD, it is important to acknowledge the limitations and potential threats to the validity of our approach. We hope our findings will inspire further research in this area and contribute to developing more effective and robust approaches for identifying adults with ASD using social media data.

Ethical approval

All procedures in this study were carried out in line with the institution's and/or national research committee's ethical standards, the 1964 Helsinki Declaration, and its later revisions or comparable ethical standards.

CRediT authorship contribution statement

Mahsa Khorasani: Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing - original draft. **Mohsen Kahani:** Validation, Supervision, Methodology, Conceptualization, Writing - review & editing. **Seyed Amir Amin Yazdi:** Supervision, Data curation, Conceptualization, Writing - review & editing. **Mostafa Hajiaghaei-Keshteli:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] C. Lord, et al., The lancet commission on the future of care and clinical research in autism, Lancet (2022) http://dx.doi.org/10.1016/S0140-6736(21)01541-5.
[2] J. Zeidan, et al., Global prevalence of autism: A systematic review update, Autism Res. (2022) http://dx.doi.org/10.1002/aur.2696.
[3] G. Russell, et al., Time trends in autism diagnosis over 20 years: a UK population-based cohort study, J. Child Psychol. Psychiatry 63 (6) (2022) http://dx.doi.org/10.1111/jcpp.13505.
[4] M.M. Hughes, et al., Adolescents with autism spectrum disorder: Diagnostic patterns, co-occurring conditions, and transition planning, J. Adolesc. Health. (2023) 1–8, http://dx.doi.org/10.1016/j.jadohealth.2022.12.010.
[5] Y. Huang, et al., Diagnosis of autism in adulthood: A scoping review, Autism (2020) http://dx.doi.org/10.1177/1362361320903128.
[6] L. Fusar-Poli, et al., Missed diagnoses and misdiagnoses of adults with autism spectrum disorder, Eur. Arch. Psychiatry Clin. Neurosci. 272 (2) (2022) http://dx.doi.org/10.1007/s00406-020-01189-w.
[7] T. Hirota, B.H. King, Autism spectrum disorder: A review, JAMA 329 (2) (2023) 15.
[8] Y. Huang, et al., I've spent my whole life striving to be normal: Internalized stigma and perceived impact of diagnosis in autistic adults, in: Autism in Adulthood, 2023, http://dx.doi.org/10.1089/aut.2022.0066.
[9] D.E. Battle, Diagnostic and statistical manual of mental disorders (DSM), in: CoDAS, 2013, http://dx.doi.org/10.4135/9781412956321.n79.
[10] F.G. Happé, et al., Demographic and cognitive profile of individuals seeking a diagnosis of autism spectrum disorder in adulthood, J. Autism Dev. Disord. 46 (11) (2016) http://dx.doi.org/10.1007/s10803-016-2886-2.
[11] M.C. Lai, S. Baron-Cohen, Identifying the lost generation of adults with autism spectrum conditions, Lancet Psychiatry (2015) http://dx.doi.org/10.1016/S2215-0366(15)00277-1.
[12] S. Cassidy, et al., Risk markers for suicidality in autistic adults, Mol. Autism 9 (1) (2018) http://dx.doi.org/10.1186/s13229-018-0226-4.
[13] J. Nyrenius, et al., Prevalence of autism spectrum disorders in adult outpatient psychiatry, J. Autism Dev. Disord. (2022) http://dx.doi.org/10.1007/s10803-021-05411-z.
[14] L. van der Meer, K. Evans, The autism diagnostic process in New Zealand: final report, 2021.
[15] K. Evans, et al., A survey of autistic adults from New Zealand on the autism diagnostic process during adolescence and adulthood, J. Autism Dev. Disord. 52 (2) (2022) http://dx.doi.org/10.1007/s10803-021-04983-0.
[16] L.F. Lewis, A mixed methods study of barriers to formal diagnosis of autism spectrum disorder in adults, J. Autism Dev. Disord. 47 (8) (2017) http://dx.doi.org/10.1007/s10803-017-3168-3.
[17] E.E. Lee, et al., Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom, Biol. Psychiatry: Cogn. Neurosci. Neuroimaging (2021) http://dx.doi.org/10.1016/j.bpsc.2021.02.001.

[18] C.M. Parlett-Pelleriti, et al., Applications of unsupervised machine learning in autism spectrum disorder research: a review, Rev. J. Autism Dev. Disord. (2022) http://dx.doi.org/10.1007/s40489-021-00299-y.

[19] J. Kim, D. Lee, E. Park, Machine learning for mental health in social media: Bibliometric study, J. Med. Internet Res. 23 (3) (2021) http://dx.doi.org/10.2196/24870.

[20] D.M. Ward, K.E. Dill-Shackleford, M.O. Mazurek, Social media use and happiness in adults with autism spectrum disorder, Cyberpsychology Behav. Soc. Netw. 21 (3) (2018) http://dx.doi.org/10.1089/cyber.2017.0331.

[21] M.O. Mazurek, Social media use among adults with autism spectrum disorders, Comput. Hum. Behav. 29 (4) (2013) http://dx.doi.org/10.1016/j.chb.2013.02.004.

[22] V. Kulkarni, et al., Latent human traits in the language of social media: An open-vocabulary approach, PLoS One 13 (11) (2018) http://dx.doi.org/10.1371/journal.pone.0201703.

[23] S. Ghosh, T. Anwar, Depression intensity estimation via social media: A deep learning approach, IEEE Trans. Comput. Soc. Syst. 8 (6) (2021) http://dx.doi.org/10.1109/TCSS.2021.3084154.

[24] A. Gupta, R. Katarya, Social media based surveillance systems for healthcare using machine learning: A systematic review, J. Biomed. Inform. (2020) http://dx.doi.org/10.1016/j.jbi.2020.103500.

[25] E.A. Ríssola, D.E. Losada, F. Crestani, A survey of computational methods for online mental state assessment on social media, ACM Trans. Comput. Healthc. 2 (2) (2021) http://dx.doi.org/10.1145/3437259.

[26] P. López-úbeda, et al., Article how successful is transfer learning for detecting anorexia on social media?, Appl. Sci. (Switzerland) 11 (4) (2021) http://dx.doi.org/10.3390/app11041838.

[27] P. Rajpurkar, et al., AI in health and medicine, Nat. Med. (2022) http://dx.doi.org/10.1038/s41591-021-01614-0.

[28] X. Chen, et al., Multi-view learning with distinguishable feature fusion for rumor detection, Knowl.-Based Syst. (2022) 108085, http://dx.doi.org/10.1016/j.knosys.2021.108085.

[29] Y. Chen, et al., Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification, in: IEEE Transactions on Neural Networks and Learning Systems, 2022, http://dx.doi.org/10.1109/TNNLS.2022.3154755.

[30] R.A. Rahman, et al., Application of machine learning methods in mental health detection: A systematic review, IEEE Access 8 (2020) http://dx.doi.org/10.1109/ACCESS.2020.3029154.

[31] A. Thieme, D. Belgrave, G. Doherty, Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems, ACM Trans. Comput.-Hum. Interact. (2020) http://dx.doi.org/10.1145/3398069.

[32] A.B.R. Shatte, D.M. Hutchinson, S.J. Teague, Machine learning in mental health: A scoping review of methods and applications, Psychol. Med. (2019) http://dx.doi.org/10.1017/S0033291719000151.

[33] T. Abuhmed, S. El-Sappagh, J.M. Alonso, Robust hybrid deep learning models for Alzheimer's progression detection, Knowl.-Based Syst. 213 (2021) http://dx.doi.org/10.1016/j.knosys.2020.106688.

[34] A. Le Glaz, et al., Machine learning and natural language processing in mental health: Systematic review, J. Med. Internet Res. (2021) http://dx.doi.org/10.2196/15708.

[35] A.M. Alvi, et al., A deep learning based framework for diagnosis of mild cognitive impairment, Knowl.-Based Syst. 248 (2022) 108815, http://dx.doi.org/10.1016/j.knosys.2022.108815.

[36] N. Goel, et al., Modified grasshopper optimization algorithm for detection of autism spectrum disorder, Phys. Commun. 41 (2020) http://dx.doi.org/10.1016/j.phycom.2020.101115.

[37] C. Küpper, et al., Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning, Sci. Rep. 10 (1) (2020) http://dx.doi.org/10.1038/s41598-020-61607-w.

[38] A.N. Optimal, et al., Article an optimal metaheuristic based feature selection with deep learning model for autism spectrum, 12, 2021, pp. 1–7.

[39] K.S. Oma, et al., A machine learning approach to predict autism spectrum disorder, in: 2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019, 2019, http://dx.doi.org/10.1109/ECACE.2019.8679454.

[40] Z.A. Huang, et al., Identifying autism spectrum disorder from resting-state fMRI using deep belief network, IEEE Trans. Neural Netw. Learn. Syst. 32 (7) (2021) http://dx.doi.org/10.1109/TNNLS.2020.3007943.

[41] V. Yaneva, et al., Detecting high-functioning autism in adults using eye tracking and machine learning, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (6) (2020) http://dx.doi.org/10.1109/TNSRE.2020.2991675.

[42] C. Zhang, et al., Deep partial multi-view learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2022) http://dx.doi.org/10.1109/TPAMI.2020.3037734.

[43] N. Zhang, et al., Discriminative few shot learning of facial dynamics in interview videos for autism trait classification, IEEE Trans. Affect. Comput. 3045 (c) (2022) 1, http://dx.doi.org/10.1109/taffc.2022.3178946.

[44] T. Zhang, et al., Natural language processing applied to mental illness detection: a narrative review, npj Digit. Med. 5 (1) (2022) 1–13, http://dx.doi.org/10.1038/s41746-022-00589-7.

[45] M. Baygin, et al., Automated ASD detection using hybrid deep lightweight features extracted from EEG signals, Comput. Biol. Med. 134 (2021) http://dx.doi.org/10.1016/j.compbiomed.2021.104548.

[46] T. Wadhera, Brain network topology unraveling epilepsy and ASD association: Automated EEG-based diagnostic model, Expert Syst. Appl. 186 (2021) http://dx.doi.org/10.1016/j.eswa.2021.115762.

[47] Q. Khubaib Ahmed, et al., Deception detection on social media: A source-based perspective, Knowl.-Based Syst. (2022) http://dx.doi.org/10.1016/j.knosys.2022.109649.

[48] S.J. Teague, et al., Methods and applications of social media monitoring of mental health during disasters: Scoping review, JMIR Ment. Health (2022) http://dx.doi.org/10.2196/33058.

[49] T. Newton, A.D.I. Kramer, D.N. McIntosh, Autism online: A comparison of word usage in bloggers with and without autism spectrum disorders, in: Conference on Human Factors in Computing Systems - Proceedings, 2009, http://dx.doi.org/10.1145/1518701.1518775.

[50] T. Nguyen, D. Phung, S. Venkatesh, Analysis of psycholinguistic processes and topics in online autism communities, in: Proceedings - IEEE International Conference on Multimedia and Expo, 2013, http://dx.doi.org/10.1109/ICME.2013.6607615.

[51] A. Beykikhoshk, et al., Data-mining twitter and the autism spectrum disorder: A pilot study, in: ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2014, http://dx.doi.org/10.1109/ASONAM.2014.6921609.

[52] A. Beykikhoshk, et al., Using Twitter to learn about the autism community, Soc. Netw. Anal. Min. 5 (1) (2015) http://dx.doi.org/10.1007/s13278-015-0261-5.

[53] T. Nguyen, et al., Affective, linguistic and topic patterns in online autism communities, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8787, 2014, http://dx.doi.org/10.1007/978-3-319-11746-1_35.

[54] T. Nguyen, et al., Autism blogs: Expressed emotion, language styles and concerns in personal and community settings, IEEE Trans. Affect. Comput. 6 (3) (2015) 312–323, http://dx.doi.org/10.1109/TAFFC.2015.2400912.

[55] M.L. Bellon-Harn, J. Ni, V. Manchaiah, Twitter usage about autism spectrum disorder, Autism 24 (7) (2020) http://dx.doi.org/10.1177/1362361320923173.

[56] Y. Hswen, et al., Using twitter to detect psychological characteristics of self-identified persons with autism spectrum disorder: a feasibility study, JMIR mHealth uHealth 7 (2) (2019) http://dx.doi.org/10.2196/12264.

[57] J. Kim, et al., A deep learning model for detecting mental illness from user content on social media, Sci. Rep. 10 (1) (2020) 1–7, http://dx.doi.org/10.1038/s41598-020-68764-y.

[58] M.S. Alzahrani, Classification and detection of autism spectrum disorder, 2022, 2022.

[59] T.L.P.V.M. Lakshmi, Predicting sentiment and emotion on ASD using machine learning and deep neural networks based on twitter data, in: Intelligent Sustainable Systems, in: Lecture Notes in Networks and Systems, vol. 333, Springer, Singapore, 2022, http://dx.doi.org/10.1007/978-981-16-6309-3.

[60] C.K. Chung, J.W. Pennebaker, Linguistic inquiry and word count (LIWC), Appl. Nat. Lang. Process. 2015 (2013) 206–229, http://dx.doi.org/10.4018/978-1-60960-741-8.ch012.

[61] Tomas Mikolov, G.Z. Yih Wen-tau, Linguistic regularities in continuous space word representations - Microsoft research, in: Hlt-Naacl, 2013.

[62] J. Devlin, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[63] A. Wongkoblap, M.A. Vadillo, V. Curcin, Researching mental health disorders in the era of social media: Systematic review, J. Med. Internet Res. (2017) http://dx.doi.org/10.2196/jmir.7215.

[64] D. Bamman, GEnder in T Witter: S Tyles, Stances, and Social Networks, Technical Report 1210.4567, 2012, arXiv. Available at: https://arxiv.org/vc/arxiv/papers/1210/1210.4567v1.pdf.

[65] W.X. Zhao, et al., Comparing twitter and traditional media using topic models, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, http://dx.doi.org/10.1007/978-3-642-20161-5_34.

[66] M. Duan, Q. Li, L. Xiao, Topic-extended emotional conversation generation model based on joint decoding, IEEE Access 9 (2021) http://dx.doi.org/10.1109/ACCESS.2021.3090435.

[67] M. Hossen, et al., Classification of Social Media Users Based on Temporal Behaviors and Interests, in: Lecture Notes in Networks and Systems, 2021, http://dx.doi.org/10.1007/978-981-16-1089-9_72.
[68] J.A. Lossio-Ventura, et al., Evaluation of clustering and topic modeling methods over health-related tweets and emails, Artif. Intell. Med. 117 (2021) http://dx.doi.org/10.1016/j.artmed.2021.102096.
[69] L.G. Anthony, et al., Interests in high-functioning autism are more intense, interfering, and idiosyncratic than those in neurotypical development, Dev. Psychopathol. 25 (3) (2013) http://dx.doi.org/10.1017/S0954579413000072.
[70] M. Rouhizadeh, et al., Detecting linguistic idiosyncratic interests in autism using distributional semantic models, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal To Clinical Reality, 2014, pp. 46–50, Available at: http://www.aclweb.org/anthology/W/W14/W14-3206.
[71] M. Rouhizadeh, et al., Measuring idiosyncratic interests in children with autism, 2015, pp. 212–217.
[72] S. Jarang, D. Joshi, V.S. Deshpande, Behaviour analysis using word embedding machine learning on social media, in: Proceedings - 2019 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019, 2019, http://dx.doi.org/10.1109/ICCUBEA47591.2019.9129273.
[73] D. Joshi, D.M. Patwardhan, An analysis of mental health of social media users using unsupervised approach, Comput. Hum. Behav. Rep. 2 (2020) http://dx.doi.org/10.1016/j.chbr.2020.100036.
[74] J. Eisenstein, What to do about bad language on the internet, in: NAACL HLT 2013-2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, 2013.
[75] B. Han, P. Cook, T. Baldwin, Lexical normalization for social media text, ACM Trans. Intell. Syst. Technol. 4 (1) (2013) http://dx.doi.org/10.1145/2414425.2414430.
[76] D.Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English Tweets, 2020, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.2.
[77] S. Bird, E. Loper, E. Klein, Natural Language ToolKit (NLTK) Book, O'Reilly Media Inc, 2009.
[78] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2019, http://dx.doi.org/10.18653/v1/d19-1410.
[79] H. Choi, et al., Evaluation of BERT and albert sentence embedding performance on downstream NLP tasks, in: Proceedings - International Conference on Pattern Recognition, 2020, http://dx.doi.org/10.1109/ICPR48806.2021.9412102.
[80] S. Zhao, et al., AP-BERT: enhanced pre-trained model through average pooling, Appl. Intell. (2022) http://dx.doi.org/10.1007/s10489-022-03190-3.
[81] C.G.M. Snoek, M. Worring, A.W.M. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th ACM International Conference on Multimedia, MM 2005, 2005, http://dx.doi.org/10.1145/1101149.1101236.
[82] L. Ansari, et al., Ensemble hybrid learning methods for automated depression detection, in: IEEE Transactions on Computational Social Systems, 2022, http://dx.doi.org/10.1109/TCSS.2022.3154442.
[83] E.T. Prud'hommeaux, et al., Classification of atypical language in autism, in: Acl Hlt 2011, No. June, 2011, p. 88.
[84] J. Yuan, et al., Autism spectrum disorder detection from semi-structured and unstructured medical data, Eurasip J. Bioinform. Syst. Biol. 2017 (1) (2016) http://dx.doi.org/10.1186/s13637-017-0057-1.
[85] J.A. Pereira, et al., PictoBERT: Transformers for next pictogram prediction[Formula presented], Expert Syst. Appl. 202 (April) (2022) 117231, http://dx.doi.org/10.1016/j.eswa.2022.117231.
[86] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: 31st International Conference on Machine Learning, ICML 2014, 2014.
[87] E. Kadkhoda, et al., Bipolar disorder detection over social media, Inform. Med. Unlocked (2022) http://dx.doi.org/10.1016/j.imu.2022.101042.
[88] T. Mikolov, et al., Efficient estimation of word representations in vector space, in: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013.
[89] M. Langen, et al., The neurobiology of repetitive behavior: ...and men, Neurosci. Biobehav. Rev. (2011) http://dx.doi.org/10.1016/j.neubiorev.2010.02.005.