#ActuallyAutistic Twitter dataset for precision diagnosis of Autism Spectrum Disorder (ASD)

Aditi Jaiswal[1], Dr Peter Washington[1]

[1] Department of Information and Computer Sciences, University of Hawaii at Manoa

Corresponding author: Aditi Jaiswal (ajaiswal@hawaii.edu)

Abstract

The increasing usage of social media platforms has given rise to an unprecedented surge in user-generated content with millions of users sharing their thoughts, experiences, and health-related information. Because of this social media has turned out to be a useful means to study and understand public health. Twitter is one such platform that has proven to be a valuable source of such information for both public and health officials. We present a novel dataset consisting of 6,515,470 tweets collected from users self identifying with autism using "#ActuallyAutistic" and a control group. The dataset also has supporting information such as posting dates, follower count, geographical location, and interaction metrics. We illustrate the utility of the dataset through common Natural Language Processing (NLP) applications such as sentiment analysis, tweet and user classification, and topic modeling. The textual differences in social media communications can help researchers and clinicians to conduct symptomatology studies, in natural settings, by establishing effective biomarkers to distinguish an autistic individual from their typical peers. For better accessibility, reusability and new research insights, we have released the dataset publicly.

Keywords

Social media analysis, Twitter, autism spectrum disorder, sentiment analysis, public health

Background and Summary

Autism spectrum disorder (ASD) is a developmental delay causing physical, cognitive, and behavioral changes and affecting millions of individuals. A core complexity of ASD lies in its symptom profile changing with age, often leading to the misattribution of its characteristics to other conditions such as anxiety, obsessive-compulsive disorder (OCD), and attention-deficit/hyperactivity disorder (ADHD)[1,2]. Therefore, an early diagnosis is crucial to provide appropriate treatment and improve the efficacy of screening tools. However, there are limitations on the availability of standard tests[3], leading to misdiagnosis or delayed treatments[4], which can place patients at risk of developing depression or suicidal tendencies[5]. Social media has turned out to be a useful means for real-time public health monitoring. Such non-clinical data holds considerable potential for the research community to extract meaningful insights through a less intrusive approach and improve the rigor of ASD analytics research. The digital footprint of an individual can be analyzed to study behavioral symptoms of ASD and other mental health disorders[6].

Twitter, a popular microblogging platform, has emerged as a particularly valuable source for data. The platform allows users to post tweets containing up to 280 characters and has an active monthly user base of approximately 450 million individuals[7]. Twitter's strength lies in its ability to capture real-time thoughts, news, conversations, and statistics, making it more suitable for collecting observational data than traditional survey-based methods. Research in mental health such as identifying depression and mood changes[8-14], and real-time mapping of natural disasters[15,16] or infectious disease spread and its effect on emotional health[17-24] has greatly benefited from digital phenotyping.

ASD has been the subject of multiple clinical trials, reviews, and epidemiological studies conducted using behavioral features such as eye gaze[25], prosody[26], asynchronous body movement[27], facial expressions [28,29], mobile phone data[30-33] or even electroencephalogram (EEG)[34]. However, only a handful of studies have used social analytical tools[35-38] , especially using Twitter[39,40,41] for investigating ASD. In addition, other social networking sites such as Reddit[42-45] , Facebook[46], Instagram[47,48], Flickr[49] and Sina Weibo[50] have also provided a valuable source of data for detecting and studying mental health conditions, substance abuse and risky behaviors. Using these prior works as inspiration, we curated a novel large scale Twitter dataset to study various aspects of social communication that differentiate autistic people from their neurotypical peers. Figure 1 provides an overview of the steps taken to curate the dataset.

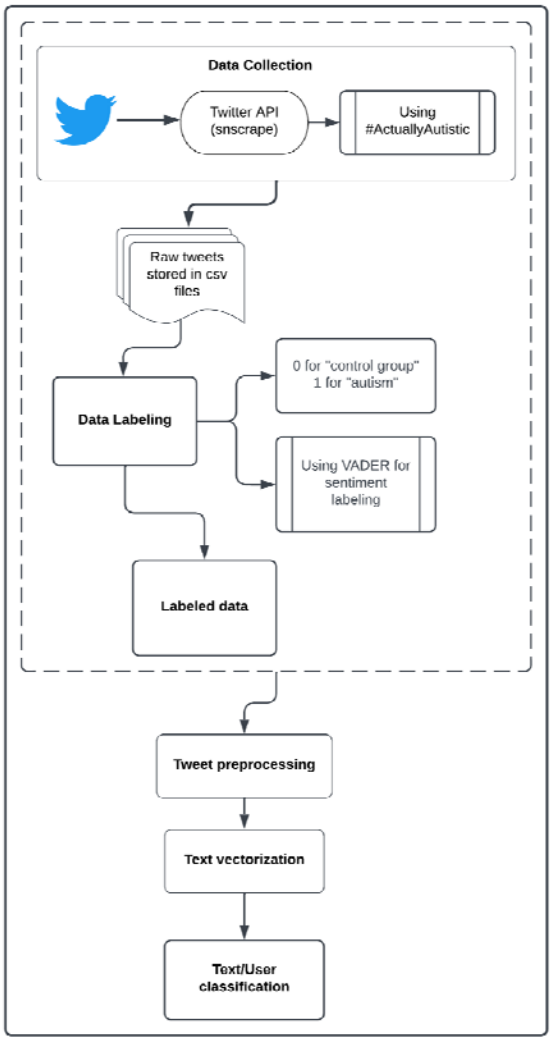

Figure 1. Pipeline for creation of the novel Twitter autism dataset.

The profound shift in society's reliance on social media for information, in contrast to traditional news sources, coupled with the enormity of data generated, has brought about an increased focus on the use of NLP for text analytics. Although research tools using facial expressions[6,51-60] and eye gazing for phenotyping ASD[61,62] are consistently reliable, there is currently a lack of standardization and preciseness in tools to measure deficits in social interaction. Therefore, linguistic and behavioral markers extracted from Twitter conversations can be used to study

textual differences and social interactions in naturalistic settings. This dataset, which we release to the public, can be used by researchers and clinicians to understand and analyze textual different features associated with ASD, enabling the research community to build precision health tools to identify and monitor ASD and its early symptoms, identify specific behavioral characteristics, derive hidden patterns, propose a clinical treatment plan, and provide community support. Furthermore, sharing the data can promote interdisciplinary collaboration to gain fresh perspectives on the research problem, and promote awareness and solution finding through hackathons, tutorials, and public challenges.

Methods

Data Collection

In recent years, hashtags such as #MeToo, #BlackLivesMatter, and #StopAsianHate has played a significant role in promoting social movements and campaigns, including those aimed at raising awareness about specific societal issues. Within the ASD community, the popular hashtags were #AutismMom and #AutismParent, representing neurotypical parents of autistic children whose outside perspectives have often shaped research and policy in this area. However, these advocacy groups often overshadow autistic adults, creating a gap in the decision-making process. To address this issue, a paradigm shift occurred in the autism rights movement through the hashtag '*#ActuallyAutistic*', focusing on understanding the experiences, challenges, and lives of individuals on the autism spectrum rather than solely focusing on caregivers. We extracted Twitter conversations of users self-identifying with ASD using this hashtag to study the differences in linguistic patterns of autistic people.

We collected tweets using *snscrape*[63], a Python based library allowing social media scraping without the need for personal Twitter API keys which provides a powerful search functionality to help filter tweets based on various conditions, such as date-time, language or location. We specifically targeted English tweets using the search query '*#ActuallyAutistic*' ranging from January 1, 2014 to December 31, 2022. From these tweets, we identified unique users who had keywords like "autism" OR "autistic" OR "neurodiverse" in their profile description (i.e., Twitter bio), signifying self-identification with ASD. It is worth noting that some users only had these keywords in their username, so we also examined usernames and their tweet contents in addition to user bios. Finally, we extracted all the tweets from the timelines of these users to construct the autism dataset, which consists of 3,137,952 tweets from 17,323 individuals. Associated metadata such as username, account created, friends count, date of tweets posted, and location (if the user had mentioned in their profile) were also extracted that could be used for statistical or network analysis.

To build a tweet classifier between individuals with ASD and their neurotypical peers, we collected a sample of random tweets as a part of a control group. To achieve this, we formulated a search query excluding the hashtag i.e. "*-#ActuallyAutistic*", using the advanced query searching operators and methods provided by Dr. Igor Brigadir[64]. However, this approach carries the risk of data leakage, whereby users who have not posted any autism-related content may possess autism-related keywords in their profile description or username. To avoid this, we screened users who had any such keywords in their profile description or usernames, or who were also present in the autism dataset, and subsequently removed them from the sample. As the main objective of curating a control group dataset was to have tweets different from those posted

by ASD users, and given that there are millions of tweets posted everyday, we collected 1,000 tweets per day between the same time period to build the control group dataset. Through this, we obtained 3,377,518 tweets from 171,273 individual users.

Data Labeling

To effectively train a supervised machine learning model, it is necessary to have labeled data, where each data point is associated with a corresponding class. We manually annotated the tweets posted by individuals with ASD as belonging to the class "autism", assigned label 1. All other tweets were labeled as belonging to the class "control group" or label 0. It is important to note that obtaining ground-truth labels can be a costly and time-consuming process, and performance of the machine learning model is often found to decrease with a decrease in labeled data. Weak supervision approaches leverage partially accurate or noisy sources for annotations, which can be more efficient than manual labeling.

Below is a small sample of tweets in our dataset and their corresponding labels:

- Sadly as #autistic #adhd #audhd we have limited choices in life and interm of profession and other life circumstances. I wish we have friendly #Neurodivergent environment in work place, where we have a sensory room too for #ActuallyAutistic #actuallyadhd . (Labeled as belonging to class “autism”)
- Celtic are now only 3 trophies behind the ‘World’s most successful team’.

  And they are 9 pts ahead for another

  Let that sink in

  This time next year they could be level on trophies won

  History being destroyed by this board; players (Labeled as belonging to class “random”)

This study has been approved by the University of Hawaii Institutional Review Board (UH IRB) under an expedited review procedure and the user information has been deidentified.

Data Preprocessing

Working with raw, unstructured Twitter data is challenging because the conversational text has too many noisy elements such as punctuation, abbreviations, emojis and other stray characters. Thus, before using such data for model training, it is necessary to clean and preprocess the data, which is an essential step for any NLP task. We started by removing the usage of any profane language in the tweets such as cursing or swear words using a Python library called *better-profanity*[65], which is designed to flag inappropriate words using string comparison and mask them using special characters (the default setting uses "*"). We then tokenized the text into words, removed any non-alphanumeric characters, hyperlinks, user mentions, and HTML tags, and converted the word tokens into lower case to avoid any confusion and data redundancy. We then removed stop words to avoid adding noise and complexity to the features with no meaningful information. To further simplify the input space and normalize the vocabulary, we applied stemming and lemmatization. We also removed any hashtags or a list of keywords related to ASD such as ‘actuallyautistic’, ‘autism’, ‘autistic’, ‘autismacceptance’, ‘autismawareness’, ‘askingautistics’, ‘askingautistic’, ‘neurodiversity’, ‘neurodivergent’, ‘allautistics’, ‘adhd’, ‘mentalhealth’, ‘asd’, ‘diagnosis’, ‘autistics’, ‘autismpride’, ‘autismspeaks’, which could introduce bias and lead to model overfitting.

Text classification

To build a tweet classifier, we first identified unique users from both the ASD and control datasets and split them into a 85:15 ratio for training and testing. This was done to avoid data leakage, which could occur if any user's tweets were split between the training and testing sets, causing the model to overfit by learning the semantic patterns specific to an individual user. The tweets were preprocessed as defined in the previous section and formed the training and test dataset. The categorical labels, representing whether a tweet belonged to an ASD or control user, were used for model training and evaluation. The training dataset was further divided into 85% training data and 15% validation data, which was used to fine-tune the model and adjust hyperparameters.

For text to numeric vectorization, we used term frequency-inverse document frequency (TF-IDF) as well as a predefined word2vec embedding method. We started by training TF-IDF word vectors using various classical ML algorithms: Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), and XGBoost gradient boosting (XGB), using 5-fold cross-validation and accuracy as the primary evaluation metric to identify the best classification method. We then trained the word2vec model using the best identified algorithm for better feature representation capturing the semantic and syntactic similarity of words and measured a more comprehensive list of evaluation metrics.

User profile classification

To further evaluate the effectiveness of the curated dataset, we performed user classification using all the tweets from each user’s timeline. We started by identifying all the unique users who

had posted at least 5 tweets and divided them into an 80:20 training to test ratio in order to avoid model overfitting as users with only one tweet might not be representative of the population. The preprocessed tweets from each user were then grouped together to form an individual document. For model training, we used an attention-based bidirectional long-short term memory (Bi-LSTM) model vectorized with a randomly initialized, self-trained embedding layer. As the tweets vary in their lengths and raw text cannot be directly represented as dense vectors like images, we used padding and an extra "unknown" token during tokenization to achieve the fixed length input and represent any unseen tokens. As a potential avenue for follow up work, we encourage researchers to try different model architectures and pre-trained word embeddings to improve the model performance and learn new information.

Results

Data Records

The dataset is available at: https://figshare.com/s/7ce063c3d713adab8764 and is presented in two comma-separated value (csv) files, one collected from 17,323 autistic individuals with 3,137,952 tweets and the other collected from 171,273 control group users consisting of 3,377,518 tweets. Both the datasets have columns: User ID (a unique value assigned to each Twitter account), profile description (a short summary of the account posted by the user), account created (datetime when the account was created), friends count (number of accounts the user follows) , followers count (number of account the user is being followed by), tweet date (datetime when the tweet was posted), tweet ID (a unique ID assigned to each tweet), language in which the tweet was posted, tweet text (original tweet), a list of hashtags present in each tweet, location (specified in the user profile, if the user provided one), number of replies (number of

times the tweet has been replied to), number of retweets (number of times the tweet was retweeted), number of likes the tweet got, and source from where the tweet was posted (web, mobile device or app).

Technical Validation

For tweet and user classification, the selection of an appropriate metric for evaluating the performance of a machine learning model is task-dependent and application-specific. Given that the present study pertains to a classification problem, we utilized accuracy, F1 score and AUC-ROC score as evaluation metrics. The results obtained from our tweet and user classifier is summarized below:

Table 1. Summary of results obtained for tweet classification from TF-IDF vectorization to identify best algorithm based on accuracy

Word vectorization method	**Model**	**Validation set accuracy**
TF-IDF	SVM	0.615
	Naive Bayes	0.598
	Logistic Regression	0.63
	XGBoost	0.624

Table 2. Summary of results obtained for tweet classification from word2vec model using best identified algorithm

Word vectorization method	**Model**	**Metric performance on test set**
Word2Vec	Logistic Regression	Accuracy: 0.73
		F1 score: 0.71
		AUC score: 0.728

While the TF-IDF vectorization yielded similar accuracy using different ML algorithms for tweet classification, logistic regression was chosen as the best predictor due to its superior performance and shorter training time. The results of the word2vec model were found to be consistent with the semantic similarities of the words. For instance, the word "autism" was found to have a higher cosine similarity to words such as "Aspergers", "neuroatypical", and "autism spectrum condition". This suggests that the word2vec model was able to capture the semantic relationships between the words.

Table 3. Summary of results obtained for user classification

User profile classification		
Word vectorization method	**Model**	**Metric performance on test set**
Keras Embedding	Attention + BiLSTM	Accuracy: 0.87
		F1 score: 0.805
		AUC score: 0.78

Although there is a class imbalance in the number of users found in ASD and control group dataset, the attention-based LSTM model still seems to make good predictions yielding an F1 score of 0.7 and 0.9 on the "autism" and "control group" classes respectively as well as an AUC score of 0.78. The results for topic modeling and sentiment analysis as obtained from the tweets using Top2Vec algorithm and VADER respectively are discussed in the Supplementary File 1.

Discussions

We demonstrate the potential of using data mining techniques to learn about ASD and related topics from social media platforms such as Twitter and its ability to transform healthcare. The 73% accuracy achieved in the tweet classification and 87% in user classification shows that there are significant semantic differences in the messages posted by individuals with and without ASD. This finding, along with previous studies using computer vision models[61,66] suggests that social phenotypical behavior could be used to support effective ASD screening strategies and facilitate early detection. Our dataset can facilitate valuable insights into various topics of discussion and the associated sentiments, carrying significant implications for public health decision-making, policy formulation, and clinical practice. The ability to observe behavioral symptoms in a non-clinical environment could lead to streamlined interactions between clinicians and patients, consequently enhancing both autism research and healthcare efficiency. We encourage researchers to benchmark the performances on this dataset and enhance the results using various techniques, or even combine our findings with research works similar to aforementioned works to build a multi-modal analytical tool. Such multimodal digital phenotyping methods have the potential to improve grading quality of clinical tools and shift healthcare from a reactive, disease-based model to a proactive, prevention-based model.

There are certain limitations to consider as well. While this study focused on individuals who self-identified as autistic, there is no clinical validation for their diagnosis. This is where annotations from clinical experts or crowdsourcing can help. Furthermore, there is a possibility of data leakage, where the identified users may not be autistic but instead could be family members, parents, caregivers, or advocacy organizations belonging to a different study population and still using the hashtags. Moreover, using VADER to label the emotional intensity and sentiments of the tweets can be relatively inaccurate compared to human labels, whose sentiments tend to get affected by their surroundings, politics and other factors, thus making it difficult to provide reliable labels. Additionally, this study only considered the English language, potentially missing out on information from other countries or languages that could aid the model in making better predictions. This also raises concerns of the lack of diversity in the data, where only English-speaking users from higher socio-economic groups or younger adults are represented in the dataset, as they comprise a larger portion of Twitter users.

This study presents several opportunities for future research, such ase using pre-trained large language models like BERT and GPT for text classification, topic modeling, and feature extraction. Another interesting avenue is the integration with additional data modalities such as audio and video which could also be mined from social media. In addition, incorporating auxiliary information to textual features may further improve the effectiveness of machine learning models. Lastly, as the Centers for Disease Control and Prevention has reported that boys are four times more likely to receive an ASD diagnosis than girls[67], gender analysis using

crowdsourcing or other metadata analysis techniques may also hold promise for future investigations.

Code Availability

We used standard Python packages such as Natural Language Toolkit (NLTK), matplotlib, and numpy for data preprocessing and analysis as well as tensorflow and scikit-learn for classification. The reference code used for tweet and user profile classification is available on GitHub (https://github.com/jaiswal-aditi/Twitter-Autism-Data.git)

Acknowledgements

The technical support and advanced computing resources from University of Hawaii Information Technology Services – Cyberinfrastructure, funded in part by the National Science Foundation CC* awards # 2201428 and # 2232862, are gratefully acknowledged.

Author Information

Authors and Affiliations

Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, Hawaii

Aditi Jaiswal, Dr Peter Washington

Contributions

Aditi Jaiswal - data collection, data analysis, manuscript writing - original draft

Dr Peter Washington - conceptualization, supervision, manuscript reviewing and editing

Corresponding authors

Aditi Jaiswal (ajaiswal@hawaii.edu), Dr Peter Washington (pyw@hawaii.edu)

Ethics Declaration

Competing interests

The authors declare no competing interests.

Supplementary Information

Supplementary File 1

References

1. Cath, D. C., Ran, N., Smit, J. H., van Balkom, A. J. L. M., & Comijs, H. C. (2007). Symptom Overlap between Autism Spectrum Disorder, Generalized Social Anxiety Disorder and Obsessive-Compulsive Disorder in Adults: A Preliminary Case-Controlled Study. *Psychopathology*, *41*(2), 101–110. https://doi.org/10.1159/000111555

2. Zandt, F., Prior, M., & Kyrios, M. (2006). Repetitive Behaviour in Children with High Functioning Autism and Obsessive Compulsive Disorder. *Journal of Autism and Developmental Disorders*, *37*(2), 251–259. https://doi.org/10.1007/s10803-006-0158-2

3. Ning, M., Daniels, J., Schwartz, J., Dunlap, K., Washington, P., Kalantarian, H., Du, M., & Wall, D. P. (2019). Identification and Quantification of Gaps in Access to Autism Resources in the United States: An Infodemiological Study. *Journal of Medical Internet Research*, *21*(7), e13094. https://doi.org/10.2196/13094

4. Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of General Psychiatry*, *63*(6), 694–701. https://doi.org/10.1001/archpsyc.63.6.694

5. Weinstock, C. P. (2019, July 31). *The deep emotional ties between depression and autism*. Spectrum | Autism Research News. https://www.spectrumnews.org/features/deep-dive/the-deep-emotional-ties-between-depression-and-autism/

6. Washington, P., & Wall, D. P. (2023). A Review of and Roadmap for Data Science and Machine Learning for the Neuropsychiatric Phenotype of Autism. *Annual Review of Biomedical Data Science*, *6*, 211–228. https://doi.org/10.1146/annurev-biodatasci-020722-125454

7. Campbell, S. How Many People Use Twitter in 2023? (Twitter Statistics). The Small Business Blog https://thesmallbusinessblog.net/twitter-statistics (2023)

8. Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. https://doi.org/10.3115/v1/w15-1201

9. Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

10. Hswen, Y., Naslund, J. A., Brownstein, J. S., & Hawkins, J. B. (2018). Online Communication about Depression and Anxiety among Twitter Users with Schizophrenia: Preliminary Findings to Inform a Digital Phenotype Using Social Media. *Psychiatric Quarterly*, *89*(3), 569–580. https://doi.org/10.1007/s11126-017-9559-y

11. Mowery, D., Bryan, C., & Conway, M. (2017). Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health. ArXiv.org. https://doi.org/10.48550/arXiv.1701.08229

12. De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2470654.2466447

13. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432

14. Nadeem, M. (2016). Identifying Depression on Twitter. *ArXiv:1607.07384 [Cs, Stat.*] https://arxiv.org/abs/1607.07384

15. Robinson, B., Power, R., & Cameron, M. (2013). An Evidence Based Earthquake Detector using Twitter. In *Proceedings of the Workshop on Language Processing and Crisis Information 2013*, pages 1–9, Nagoya, Japan. Asian Federation of Natural Language Processing.

16. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, pages 851–860. https://doi.org/10.1145/1772690.1772777

17. Chew, C., & Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, *5*(11), e14118. https://doi.org/10.1371/journal.pone.0014118

18. Prieto, V. M., Matos, S., Álvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A Good Place to Detect Health Conditions. *PLoS ONE*, *9*(1), e86191. https://doi.org/10.1371/journal.pone.0086191

19. Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, *6*(5), e19467. https://doi.org/10.1371/journal.pone.0019467

20. Kim, E. H.-J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science*, *42*(6), 763–781. https://doi.org/10.1177/0165551515608733

21. Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*. https://doi.org/10.1145/1964858.1964874

22. Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, Scotland, UK.. Association for Computational Linguistics. https://aclanthology.org/D11-1145/

23. Ye J. (2020). Pediatric Mental and Behavioral Health in the Period of Quarantine and Social Distancing With COVID-19. *JMIR pediatrics and parenting*, *3*(2), e19867. https://doi.org/10.2196/19867

24. Gupta, V., Jain, N., Katariya, P., Kumar, A., Mohan, S., Ahmadian, A., & Ferrara, M. (2021). An Emotion Care Model using Multimodal Textual Analysis on COVID-19. *Chaos, solitons, and fractals*, *144*, 110708. https://doi.org/10.1016/j.chaos.2021.110708

25. Washington, P., et al. Data-Driven Diagnostics and the Potential of Mobile Artificial Intelligence for Digital Therapeutic Phenotyping in Computational Psychiatry. *Biological psychiatry. Cognitive neuroscience and neuroimaging*, *5*(8), 759–769. https://doi.org/10.1016/j.bpsc.2019.11.015

26. Chi, Nathan A., et al. Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study. *JMIR pediatrics and parenting*, *5*(2), e35406. https://doi.org/10.2196/35406

27. Lakkapragada, Anish, et al. The Classification of Abnormal Hand Movement to Aid in Autism Detection: Machine Learning Study. *JMIR Biomedical Engineering*, *7*(1), e33771. https://doi.org/10.2196/33771

28. Wu, C., et al. Machine Learning Based Autism Spectrum Disorder Detection from Videos. *Healthcom. International Conference on e-Health Networking, Applications and Services*, *2020*, 10.1109/healthcom49281.2021.9398924. https://doi.org/10.1109/healthcom49281.2021.9398924

29. Washington, Peter, et al. Improved Digital Therapy for Developmental Pediatrics Using Domain-Specific Artificial Intelligence: Machine Learning Study. *JMIR pediatrics and parenting*, *5*(2), e26760. https://doi.org/10.2196/26760

30. Tariq, Q.,et al. Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS medicine*, *15*(11), e1002705. https://doi.org/10.1371/journal.pmed.1002705

31. Varma, Maya, et al. Identification of Social Engagement Indicators Associated With Autism Spectrum Disorder Using a Game-Based Mobile App: Comparative Study of Gaze Fixation and

Visual Scanning Methods. *Journal of medical Internet research*, *24*(2), e31830. https://doi.org/10.2196/31830

32. Banerjee, A., et al. Training and Profiling a Pediatric Facial Expression Classifier for Children on Mobile Devices: Machine Learning Study. *JMIR formative research*, *7*, e39917. https://doi.org/10.2196/39917

33. Anzulewicz, A., Sobota, K., & Delafield-Butt, J. T. (2016). Toward the Autism Motor Signature: Gesture patterns during smart tablet gameplay identify children with autism. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep31107

34. Alotaibi, N., & Maharatna, K. (2021). Classification of Autism Spectrum Disorder From EEG-Based Functional Brain Connectivity Analysis. *Neural computation*, *33*(7), 1914–1941. https://doi.org/10.1162/neco_a_01394

35. Newton, A. T., Kramer, A. D. I., & McIntosh, D. N. (2009). Autism online. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/1518701.1518775

36. Nguyen, T., Duong, T., Phung, D., & Venkatesh, S. (2014). Affective, Linguistic and Topic Patterns in Online Autism Communities. *Web Information Systems Engineering – WISE 2014*, 474–488. https://doi.org/10.1007/978-3-319-11746-1_35

37. van Schalkwyk, G. I., et al. Social Media Use, Friendship Quality, and the Moderating Role of Anxiety in Adolescents with Autism Spectrum Disorder. *Journal of autism and developmental disorders*, *47*(9), 2805–2813. https://doi.org/10.1007/s10803-017-3201-6

38. Bakombo, S., Ewalefo, P., & Konkle, A. T. M. (2023). The Influence of Social Media on the Perception of Autism Spectrum Disorders: Content Analysis of Public Discourse on YouTube Videos. *International journal of environmental research and public health*, *20*(4), 3246. https://doi.org/10.3390/ijerph20043246

39. Hswen, Y., Gopaluni, A., Brownstein, J. S., & Hawkins, J. B. (2019). Using Twitter to Detect Psychological Characteristics of Self-Identified Persons With Autism Spectrum Disorder: A Feasibility Study. *JMIR mHealth and uHealth*, *7*(2), e12264. https://doi.org/10.2196/12264

40. Corti, L., Zanetti, M., Tricella, G., & Bonati, M. (2022). Social media analysis of Twitter tweets related to ASD in 2019-2020, with particular attention to COVID-19: topic modelling and sentiment analysis. *Journal of big data*, *9*(1), 113. https://doi.org/10.1186/s40537-022-00666-4

41. Beykikhoshk, A., Arandjelović, O., Phung, D., Venkatesh, S., & Caelli, T. (2015). Using Twitter to learn about the autism community. *Social Network Analysis and Mining*, *5*(1). https://doi.org/10.1007/s13278-015-0261-5

42. Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018, June 1). Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0603

43. Garg, S., et al. Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on Reddit. *Internet interventions*, *26*, 100467. https://doi.org/10.1016/j.invent.2021.100467

44. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of Depression-Related Posts in Reddit Social Media Forum. *IEEE Access*, *7*, 44883–44893. https://doi.org/10.1109/access.2019.2909180.

45. Bellon-Harn, M. L., Boyd, R. L., & Manchaiah, V. (2022). Applied Behavior Analysis as Treatment for Autism Spectrum Disorders: Topic Modeling and Linguistic Analysis of Reddit Posts. *Frontiers in rehabilitation sciences*, *2*, 682533. https://doi.org/10.3389/fresc.2021.682533

46. Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014, June 1). Towards Assessing Changes in Degree of Depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125 ACLWeb; Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3214

47. Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1). https://doi.org/10.1140/epjds/s13688-017-0110-z

48. Hassanpour, S., Tomita, N., DeLise, T., Crosier, B., & Marsch, L. A. (2019). Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, *44*(3), 487–494. https://doi.org/10.1038/s41386-018-0247-x

49. Yang, Y., et al. How Do Your Friends on Social Media Disclose Your Emotions?. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(1). https://doi.org/10.1609/aaai.v28i1.8740

50. Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., & Feng, L. (2014). User-level psychological stress detection from social media using deep neural network. *Proceedings of the 22nd ACM International Conference on Multimedia*. https://doi.org/10.1145/2647868.2654945

51. Daniels, J., Schwartz, J.N., Voss, C. *et al.* Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. *npj Digital Med* **1**, 32 (2018). https://doi.org/10.1038/s41746-018-0035-3

52. Kalantarian, H., Washington, P., Schwartz, J., Daniels, J., Haber, N., & Wall, D. P. (2018). Guess What? *Journal of Healthcare Informatics Research*, *3*(1), 43–66. https://doi.org/10.1007/s41666-018-0034-9

53. Kalantarian, H., Jedoui, K., Washington, P., & Wall, D. P. (2020). A Mobile Game for Automatic Emotion-Labeling of Images. *IEEE Transactions on Games*, *12*(2), 213–218. https://doi.org/10.1109/TG.2018.2877325

54. Kalantarian, H., et al. Labeling images with facial emotion and the potential for pediatric healthcare. *Artificial intelligence in medicine*, *98*, 77–86. https://doi.org/10.1016/j.artmed.2019.06.004

55. Kalantarian, H., et al. The Performance of Emotion Classifiers for Children With Parent-Reported Autism: Quantitative Feasibility Study. *JMIR mental health*, *7*(4), e13174. https://doi.org/10.2196/13174

56. Haik Kalantarian, Washington, P., Schwartz, J., Daniels, J., Haber, N., & Wall, D. P. (2018). *A Gamified Mobile System for Crowdsourcing Video for Autism Research*. https://doi.org/10.1109/ichi.2018.00052

57. Kline, A., et al. Superpower Glass. *GetMobile*, *23*(2), 35–38. https://doi.org/10.1145/3372300.3372308

58. Washington, P. et. al. SuperpowerGlass: A Wearable Aid for the At-Home Therapy of Children with Autism. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–22. https://doi.org/10.1145/3130977

59. Voss, C., Winograd, T., Wall, D., Washington, P., Haber, N., Kline, A., Daniels, J., Fazel, A., De, T., McCarthy, B., & Feinstein, C. (2016). Superpower glass. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*. https://doi.org/10.1145/2968219.2968310

60. Voss, C., et al. Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder: A Randomized Clinical Trial. *JAMA pediatrics*, *173*(5), 446–454. https://doi.org/10.1001/jamapediatrics.2019.0285

61. Drimalla, H., Scheffer, T., Landwehr, N. et al. Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (SIT). npj Digit. Med. 3, 25 (2020). https://doi.org/10.1038/s41746-020-0227-5

62. Ahmed, I. A., et al. Eye Tracking-Based Diagnosis and Early Detection of Autism Spectrum Disorder Using Machine Learning and Deep Learning Techniques. *Electronics*, *11*(4), 530. https://doi.org/10.3390/electronics11040530

63. JustAnotherArchivist. *JustAnotherArchivist/snscrape*. GitHub. https://github.com/JustAnotherArchivist/snscrape

64. Brigadir, I. *Advanced Search on Twitter*. GitHub. https://github.com/igorbrigadir/twitter-advanced-search

65. Nguyen, S. T. *better_profanity*. GitHub. https://github.com/snguyenthanh/better_profanity

66. Alcañiz, M., et al. Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: A proof of concept for diagnosis. *Autism research : official journal of the International Society for Autism Research*, *15*(1), 131–145. https://doi.org/10.1002/aur.2636

67. Maenner MJ, Warren Z, Williams AR, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. *Morbidity and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002)*, *70*(11), 1–16. https://doi.org/10.15585/mmwr.ss7011a1