ANALYSIS OF TWITTER DATA FOR PUBLIC HEALTH SURVEILLANCE AND PRECISION DIAGNOSTICS OF AUTISM SPECTRUM DISORDER

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

MAY 2023

By

Aditi Jaiswal

Thesis Committee:

Peter Washington, Chairperson

Kim Binsted

Daniel D. Suthers

ABSTRACT

The healthcare industry is a prolific source of data, with every patient record, clinical trial, drug test, and medical research generating copious amounts of information. Consequently, the interest in using machine learning algorithms in healthcare applications has increased dramatically, with numerous breakthroughs being made. One such application is using social media to study and understand public health. With millions of users sharing their thoughts, exchanging ideas, and providing health-related information on various social media platforms, researchers and clinicians can conduct studies on diseases and associated symptoms in natural settings by establishing digital phenotypic biomarkers. Twitter is one such platform that has proven to be an exceptional source of health-related information from both public and health officials. In this study, we aim to mine data related to "autism" from " #ActuallyAutistic" tweets on Twitter. The textual differences in social media communications can help identify various behavioral symptoms, which can be used to distinguish an autistic individual from their typical peers. We were able to scrape a total of 6,469,994 tweets from approximately 70,000 individual users. We illustrate the usefulness of the dataset through simple applications such as: sentiment analysis, text classification and topic modeling. Our classifier achieved an accuracy of 73%, which is consistent with previous studies published in *Nature Digital Medicine* journal where using different modality data such as eye gazing and facial expressions, the authors achieved a similar accuracy. The collected data will also be released publicly to help accelerate scientific research in this field. This sharing of data and interdisciplinary research can enhance its analytical capacity and enable medical practitioners to make more informed decisions.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Autism spectrum disorder (ASD) is a group of developmental disorders affecting millions of individuals. According to the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5), autism delays the neurodevelopment of an individual causing physical, cognitive and behavioral changes. ASD is characterized by continuing challenges in social interaction, difficulty in communication and restricted or repetitive behaviors [1]. In 2020, the Centers for Disease Control and Prevention (CDC) reported that 1 in 36 children, aged 8 years (4% of boys and 1% of girls), were identified with autism [2]. The prevalence of autism has increased 317% over the past two decades and is shown to be highly associated with socioeconomic status.

ASD poses numerous challenges to individuals affected by the disorder as the symptom profiles change with age. Because of its complex nature its characteristics can often be mistaken for other disorders such as anxiety, obsessive compulsive disorder (OCD), and attention-deficit/hyperactivity disorder (ADHD)[3, 4]. Therefore, 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, leading to misdiagnosis or delayed treatments[5], which can put patients at risk of developing depression or even suicide [6]. The Modified Checklist for Autism in Toddlers (M-CHAT) [7] questionnaire, recommended by the American Academy of Pediatrics (AAP), was found to have a low sensitivity and positive predictive value [8]. On the other hand, the Autism Diagnostic Observation Schedule-Generic (ADOS-G) was found to be overdiagnosing the symptoms of communication problems with autism [9]. Therefore, a reliable diagnosis requires doctors to monitor the child's behavior and developmental screening questionnaires, which is a time-consuming process. During this observation, children may develop other health issues such as infections [10] and gastrointestinal problems [11].

Given all the associated problems and its social and economic burden, ASD has been the subject of multiple research involving clinical trials, reviews and epidemiological studies, as evidenced by the increasing number of academic literature. As shown in Figure 1, the number of publications including keywords like “autism”, “ autism spectrum disorder” or “ASD” has increased steadily over the last decade (extracted from PubMed). But even with the significant

scientific progress there is still a lack of understanding and awareness around it, stigmatizing the mental health disorders and developmental delays. The poorly understood etiology and heterogeneity of the ASD phenotype have also led to misinformation and misinterpretation of the disability.

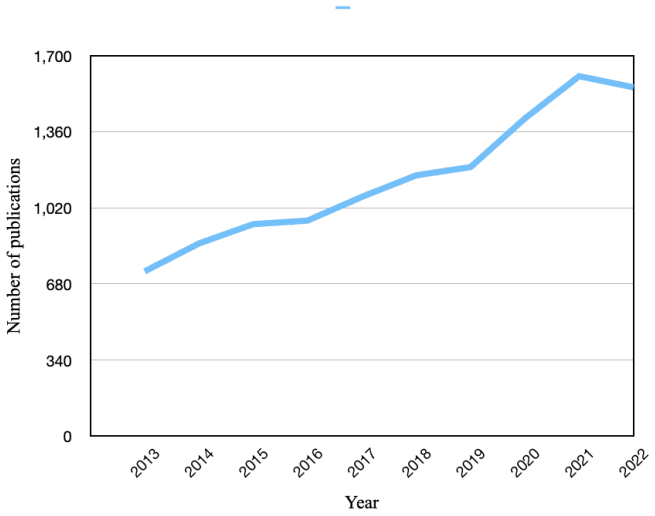


Fig. 1: The number of academic literature published in the last decade, in English, with the words "autism", " autism spectrum disorder" or "ASD" in the title or abstract as indexed by the PubMed portal

In recent years, social media has emerged as a promising tool for filling this gap by mining behavioral and observational data. Social media platforms allow users to post content about their daily lives, including their thoughts, suggestions, interests or mood on various topics, thereby facilitating social interactions. Additionally, social media has become an abundant source of health information, symptomatology, and daily struggles posted by individuals suffering from different diseases. Using digital data collected from social media, digital wearables and smartphones is referred to as "digital phenotyping". It is defined as *moment-by-moment quantification of the individual-level human phenotype in-situ using data from personal digital devices* [12].

The digital footprint of an individual can be further harnessed to study behavioral symptoms of ASD and other mental health disorders. Text based data, especially the ones obtained from social media has gained significant attention because these platforms provide a voice to individuals

struggling with ASD to express their emotions, symptoms, struggles, and thoughts. Such data belong to the category of non-clinical data, as they lack annotations from medical health professionals but can be used extensively by the research community to generate meaningful results and improve the rigor of autism research. When combined with machine learning algorithms, medical datasets have shown potential to improve the healthcare field and perform various tasks such as improving clinical trial results, medical imaging, and identifying patterns for a disease. The use of geocoded data for longitudinal tracking can also help identify changes in opinions or responses for a disease/disorders.

Twitter is a cutting-edge, microblogging platform, which has emerged as a potentially valuable source for similar data. It allows users to post tweets containing up to 280 characters and has an active monthly user base of approximately 450 million individuals [13]. With its vast amount of user-generated content, Twitter offers significant opportunities for data and text mining in various fields of application. Twitter's strength lies in its ability to capture real-time thoughts, news, conversations, and statistics, making it more suitable for collecting observational information than traditional survey-based methods. Understanding the differences in textual patterns of users, self identifying with autism or who have family members with the condition and users engaging and supporting these conversations, researchers can gain a better understanding of diagnosis and potential misdiagnosis identifiers, particularly during the early stages of autism spectrum disorder. Furthermore, studying such differences in communication patterns can enable analysis of current behavioral and emotional states, facilitating better analysis of different mental health issues.

When seeking information about autism, people usually turn to experiences shared by others, trying to learn about the different aspects that are affected by the disorder and how they can be overcome. Although the research tools using facial expressions and eye gazing for autism diagnosis [14] 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 of individuals with autism can be used to study verbal and textual differences and social interactions in naturalistic settings. In this research, we collected the data from Twitter users who self-identified as having autism and used machine learning algorithms on the corpus to build a text classifier distinguishing autistic individuals

from their typical peers. Such a corpus can be used by machine learning researchers and clinicians to understand and analyze different features associated with autism, early symptoms, specific behavioral characteristics, derive hidden patterns, propose a clinical treatment plan and also provide community support. Furthermore, we also aim to publicly release the dataset to promote interdisciplinary collaboration, gain fresh perspectives on the research problem and promote awareness and solution findings through hackathons, tutorials and public challenges.

1.1. Open sharing of data

Open sharing of data is a crucial aspect of scientific research that can significantly impact and advance the scientific community's understanding of a particular field. By making data openly available, researchers from different disciplines can use it to gain new insights, validate existing theories, and create new knowledge. There are several ways in which open sharing of data helps research:

1. Reproducibility: Open sharing of data allows researchers to reproduce and verify the findings of previous studies, which can help validate the results and conclusions. This ensures that the scientific knowledge is reliable and replicable.

2. Collaboration: Open sharing of data enables researchers from different subject matter experts and locations to collaborate on a project, contributing to a more comprehensive and interdisciplinary understanding of the research problem.

3. Cost-effective: Sharing data reduces the cost of research by enabling other researchers to use the same data for different studies, thus avoiding the need to collect new data. This becomes extremely important for academic research.

4. Accelerating discovery: Open sharing of data accelerates scientific discovery by enabling researchers to build on each other's work and generate new knowledge faster.

5. Transparency: Open sharing of data promotes transparency in research, which enhances the credibility and integrity of the scientific community. It allows other researchers to scrutinize the data, methods, and results, thus reducing the risk of fraud and error.

It is widely acknowledged that ImageNet has had a significant impact on the growth of deep learning, which has revolutionized the field of artificial intelligence (AI). First presented in 2009 [62], the aim of the research team was to expand and enhance the availability of data for training AI algorithms. This led to the creation of the largest image dataset, consisting of over 14 million images. Since then, numerous algorithms have been developed, each surpassing the previous one by producing lower error rates and advancing computer vision research.

In light of these developments, our motivation is to curate and publicly release a dataset that can serve as a standardized benchmark for natural language processing (NLP) tasks involving social media text data related to autism. Using this data, interdisciplinary research with subject matter experts and researchers from different disciplines can lead to the development of more advanced algorithms with improved prediction capabilities.

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

2. BACKGROUND AND PREVIOUS WORK

2.1. Social media analytics for mental health, global pandemics and event detection

In the past, research in mental health and real-time mapping of infectious disease spread has greatly benefited from digital phenotyping. Coppersmith et al [15] utilized the Linguistic Inquiry Word Count (LIWC) tool to analyze Twitter data and measure language similarities and differences for various mental health conditions. Similarly, Tausczik and Pennebaker [16] used LIWC to link word usage with social and emotional states. Hswen et al found that Twitter users who self-identify as having schizophrenia displayed elevated symptoms of depression and anxiety [17] as well as tobacco use [18]. Mowery et al used lexical features and emotions to classify depressive symptoms from the tweets [19]. These findings are particularly relevant as according to the *American Psychiatric association* and *Mental Health in America*, nearly 21% of the population experience some form of mental illness in a year [20], with half of them not receiving adequate treatment. With the increasing prevalence, untreated substance use disorders and suicide risks, social media platforms can provide behavioral intervention and inform suicide prevention and smoking cessation groups to focus on at-risk groups.

Furthermore, several published works have focused on the live infoveillance approach to measure the incidence rate of pandemics. For example, Chew et al performed sentiment analysis for “H1N1” and “swineflu” keywords using Twitter data [21]. Robinson et al [22] and Sakaki et al [23] even used the quasi-real time nature of twitter to locate the epicenter of an earthquake. These studies demonstrate the potential of social media data in real-time monitoring and tracking of disease spread and natural disasters.

2.2. Research in ASD

Several studies have utilized social media platforms, such as Twitter and YouTube, to explore the potential of data mining for ASD-related information. However, a reliable diagnostic classifier using social media data is still far from being achieved. For instance, Newton et al [24] used the LIWC tool to investigate differences in writing patterns between ASD and neurotypical bloggers on web blogs. Nguyen et al [25] used a similar corpus to investigate differences in sentiments, language styles, and topics of interest in the two groups, revealing that ASD writing often

conveys negative emotions. Schalkwyk et al [26] demonstrated that social media usage can improve friendship quality with reduced anxiety levels in adults with ASD, but this relationship was not observed among control users. Other studies have examined YouTube content to analyze public sentiments and perceptions on ASD [27] as well as the the quality, usability and understandability of uploaded ASD-related videos, which were generally found to be subpar.

2.3. Using digital wearables and smartphones

In addition to social media analytics, digital wearable technology such as Fitbits and mobile devices have also contributed to the advancement of digital phenotyping research. The ubiquity of mobile phones and the extensive time spent on them has made it easy to retrieve their behavioral and health data, allowing for the identification of digital markers that can be used for precision health. For instance, Teo et al conducted a study on wearable-derived sleep tracking data and its association with socioeconomic, demographic and lifestyle factors, which was found to be associated with cardiovascular disease risk markers [28]. Other studies have utilized mobile devices to extract sleep pattern markers indicating circadian misalignment [29] and correlating with the severity of depression [30]. User interactions with devices based on keyboard patterns, swiping or scrolling can also be useful in detecting stress [31]. Rachuri et al built a platform that utilizes the phone's built-in microphone to identify up to 14 different types of emotions [32].

2.4. The role of machine learning in behavioral and social media analytics

This study employed Natural Language Processing (NLP) techniques and machine learning (ML) algorithms to develop a text classifier. It is worth noting that the results might not be used as a diagnostic tool, but rather as an evidence of the textual differences between tweets posted by individuals with Autism Spectrum Disorder (ASD) and control groups. NLP tasks such as named entity recognition, part-of-speech tagging, text classification and topic modeling have been the subject of numerous research works. Machine learning and deep learning have provided substantial computational power, enabling the development of new methods, algorithms, and large language models using a massive corpus of data from the World Wide Web (WWW), books, clinical structured or unstructured texts, social media, and various other linguistic data. Moreover text classification is also one of the popular competing tasks over platforms like Kaggle.

Some of the previous research efforts have focused on the use of social media to extract health information. Prieto et al used bag of words (BoW) model, with Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees (DT) and Nearest Neighbour (KNN) classifiers to study eating disorders, flu and depression in Spain and Portugal using geocoded Twitter data [33]. Nadeem et al [34] combined crowdsourcing with the same approach to develop a text-based depression classifier. Signorini et al. used SVM regression-based models to measure influenza activity [35]. Similar regression-based models were also used to forecast flu and influenza rates in the USA [36], and Japan [37]. During the recent COVID-19 pandemic, Twitter served as an empirical tool for researchers to do multiple digital epidemiology surveillance, vaccination sentiment analysis using the transformer-based BERT model [38] and to evaluate its impact on mental health [39,40]. Twitter even experienced a 23% increase in daily use during COVID-19 pandemic, using which multiple studies were conducted to analyze emotional states of people [41, 42].

Numerous research works have employed computer vision models to track eye gaze and facial expressions of autistic children [43] or used mobile phones to capture videos [44] in search of specific traits to detect autism. Mythili et al [45] utilized an autism student database to classify levels of autism using social, communication and behavioral data. However, only a handful of studies have been done using social analytical tools for investigating ASD. Hswen et al [46] used Twitter data to analyze the textual patterns with obsessive-compulsive and repetitive behavioral characteristics and identified emotions such as fear, anxiety and paranoia from them. In another study [47], the authors found that ASD-related conversations are frequent on social media platforms and can provide valuable insights to clinicians and public health officials. Interestingly, they also found that ASD tweets have a higher proportion of nouns focused on related issues and struggling individuals than verbs. Corti et al [48] focused on ASD tweets in conjunction with COVID-19, leading them to discover the negative sentiments of the tweets being directed towards vaccine misinformation, poor management during the pandemic and various challenging situations faced by individuals, while the positive sentiments were aligned with topics like “family”, “community” and “therapy”. Based on the aforementioned prior works, we aim to curate a dataset, using Twitter as the source, to study various aspects of social communication that differentiate autistic people from their typical peers. We have also attempted to use the

collected data to identify the characteristics of ASD, which people use in their tweets, that could aid clinicians in preliminary diagnosis and/or tailor required treatment. The collected data would also be released publicly for the same purpose.

2.5. Studies using other social media platforms

While twitter remains one of the most popular social networking sites, these studies are not limited to it but to other platforms as well. Supervised learning [49] and topic modeling [50] has been done for early detection of depression, suicide risk detection [51] and opioids overdose risk assesment [52] using Reddit posts. Tadesse et al [53] combined LIWC+LDA+bigram features with Multilayer Perceptron (MLP) classifier to identify the lexicon of words related to depression in Reddit posts with an accuracy of 91%. Schwartz et al [54] employed a regression based model to predict the degree of depression in Facebook users. Reece et al [55] and Ricard et al [56] analyzed Instagram posts to identify depression markers, and Hassanpour et al. [57] used Instagram to study substance abuse risk. Yang et al [58] analyzed Flickr images and comments to infer emotions, while Lin et al [59] used Sina and Tencent Weibo, a Chinese microblogging website, to analyze behavior patterns and the users' social networks to detect psychological stress. In a study [60] Reddit posts were analyzed to identify themes related to applied behavioral analysis (ABA) therapy for ASD using topic modeling and LIWC. The authors observed that the users shared personal opinions and experiences on ABA therapy more than clinical information, indicating emotional interventional support. Similarly, Aspergers subreddits were found to have higher emotional and informational support scores than other average subreddit posts [61]. Overall, social media platforms have provided a valuable source of data for detecting and studying mental health conditions and risky behaviors.

3. CREATION OF THE AUTISM TWEETS DATASET

3.1. Data Collection

In recent years, social media platforms such as Twitter, Facebook, and Instagram have played a significant role in promoting social movements and campaigns, including those aimed at raising awareness about specific issues. Many of these campaigns, such as #MeToo, #BlackLivesMatter, and #StopAsianHate, have been successful in disseminating messages and influencing public opinion. In the context of the autism community, the popular hashtags until a while ago were #AutismMom or #AutismParent. These represented neurotypical parents of autistic children, whose outside perspectives have often shaped research and policy in this area. However, these advocacy groups have often overshadowed autistic adults, who have felt left out of the decision-making process. To address this issue, there has been a paradigm shift in the autism rights movement, with a focus on understanding the struggles, symptoms, and lives of actually autistic people rather than just their caregivers. As such, the present work aims to extract Twitter conversations of users self-identifying with autism, using the hashtag #ActuallyAutistic.

The process of extracting data from a website is commonly referred to as web scraping. In this study, tweets were manually collected using *snscrape* [63], a Python based library that allows for the extraction of tweets without the need for personal Twitter API keys. The library provides a powerful search functionality to help filter tweets based on various conditions, such as date-time, language, and number of likes. For this study, we obtained English tweets related to autism by using the search query *#ActuallyAutistic* and setting the since and until flags to January 1, 2014, and December 31, 2022, respectively. From these tweets we identified unique users who had keywords like “autism” OR “autistic” OR “neurodiverse” in their profile description (i.e. twitter bio) to focus on users self-identifying with ASD. It is worth noting that some users only had these keywords in their username, so we also checked usernames in addition to user bios. Finally, we extracted all the tweets from the timelines of these users to build the autism dataset, which consists of over 3 million tweets. 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 analysis.

In order to build a machine learning classifier capable of discerning semantic differences in textual patterns between ASD individuals and control groups, we collected a sample of random tweets as a comparison group. To achieve this, we formulated a search query excluding the #ActuallyAutistic 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 for users who had keywords related to autism in their profile description or usernames, or who were also present in the autism dataset, and subsequently removed them from the sample. Finally we repeated the same process of extracting all the tweets from the remaining users' timeline, which made the random dataset. Through this combined data collection method, we obtained a total of 6,469,994 tweets from approximately 70,000 individual users.

3.2. Data Labeling

In order to train a supervised machine learning model, it is necessary to have labeled data where each data point, in this case, tweet text, is associated with a corresponding class. In this study, we have manually labeled tweets posted by individuals with Autism Spectrum Disorder (ASD) as belonging to the "autism" category, which has been assigned a class label of 1. All other tweets have been labeled as belonging to the "random" category, which has been assigned a class label of 0. The data labeling process was carried out by the researchers themselves using their domain knowledge, and therefore, falls under the category of weakly supervised learning. 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 address these challenges by using noisy or partially accurate sources to label the data, which can be more efficient than manual labeling. In cases where there is a shortage of labeled data, weak supervision techniques such as crowdsourcing, transfer learning, or heuristic rules can be used to generate weak labels, which can then be used to train a model.

4. EXPLORATORY DATA ANALYSIS

4.1. Methods

Data analysis and visualization is a crucial step in machine learning research to gain insights from the data. It can help in identifying missing values, understanding data distributions, visualizing trends and choosing the right set of features. We first started by observing the distribution of missing values in the data that can help understand how and where the missing values are before imputing or dropping them. Missingno is a python-based library that can be used to visualize missing values in a dataset through matrix, bar charts, and heatmaps.

Plotting monthly and yearly distribution of tweet counts can help identify trends over time and understand how issues discussed on Twitter have changed with people's perception and histograms of character and word counts in tweets can provide insights into the length of the tweets and whether there are any significant differences between the two groups. Word clouds, a text visualization technique, can help identify trending topics and prominent words in the corpus. These insights can help researchers to discover trending topics and choose the feature set and algorithm accordingly.

4.1.1. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a crucial NLP task that involves identifying and extracting subjective information from textual data to determine its overall sentiment. It has been extensively researched in academic literature and is commonly studied using two approaches: machine learning and lexical. While the former approach tends to have better performance with larger datasets, it requires a model to be trained using previously labeled data to predict sentiments of new unseen data. The latter approach, however, uses a dictionary of sentiments to map words or sentences to a category (positive, negative, or neutral) or a numeric range of emotional intensity of sentiments.

For this work, we utilized the Valence Aware Dictionary for sEntiment Reasoning (VADER), a lexical based approach *specifically attuned to sentiments expressed in social media or microblogs like context* [71] to analyze the sentiments of the curated dataset. At the word level, VADER uses a lexicon of words to determine the sentiment of each individual word, with their

polarities score ranging from -4 (most negative) to +4 (most positive). At the sentence level, it applies a set of rules and heuristics on the sentiment scores of the individual words to determine the overall sentiment of the sentence. After determining the sentiment score for each sentence, VADER returns a dictionary containing four values: neg, neu, pos, and compound, each representing the negative, neutral, positive, and overall (normalized) sentiment scores of the sentence, respectively. The compound score ranges from -1 (most negative) to +1 (most positive), which is obtained by normalizing the sum of the individual sentence scores. For this study, the analysis was conducted to compare the sentiments of tweets posted by individuals with ASD to those of the control group in two scenarios: one including tweets with profanity and the other excluding it.

4.2. Results

It was observed that the location column of the dataset had the most number of missing values. This is likely due to the fact that Twitter users have the freedom to enter any location they desire on their profiles. Analysis revealed that the majority of users did not enter their actual location and those who did had inconsistent location entries. Of the top 20 location values found, the majority were variations of "United Kingdom", including "UK", "London, England", "England, United Kingdom", "South East, England", while others were less informative strings such as "Picnic party" and "My parent's basement".

Further analysis of the yearly distribution of tweets revealed an increase in conversations about Autism Spectrum Disorder (ASD) over the years, as shown in Fig 2. Autistic individuals appear to be more comfortable with social media interactions, which provides them with multiple employment opportunities and serves as an effective platform for educating people about developmental delays and sharing behavioral symptoms that may be helpful to others.

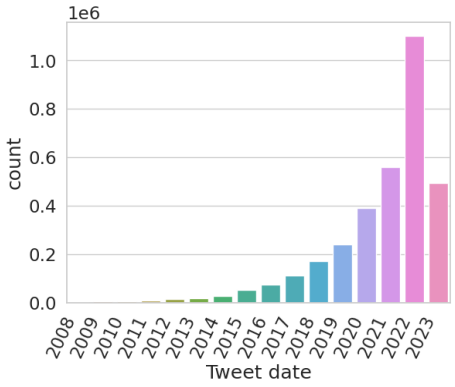


Fig. 2: Timely distribution of tweets posted by ASD users. Note that the count for 2023 is incomplete as it only includes data, from timelines of the users, up to and including February 2023.

The sentiments of most of the original autism tweets (with profanity) were found to be positive, followed by negative and neutral sentiments. A similar trend was observed in the cleaned autism dataset as well, although with a slight increase in positive sentiment and a decrease in negativity. The original control group dataset, on the other hand, had the majority of the tweets with positive sentiments followed closely by neutral sentiments and some negative sentiment tweets. However, upon removing profanity from the dataset, the distribution was dominated by neutral sentiment tweets and a nearly equal percentage of positive but fewer negative tweets.

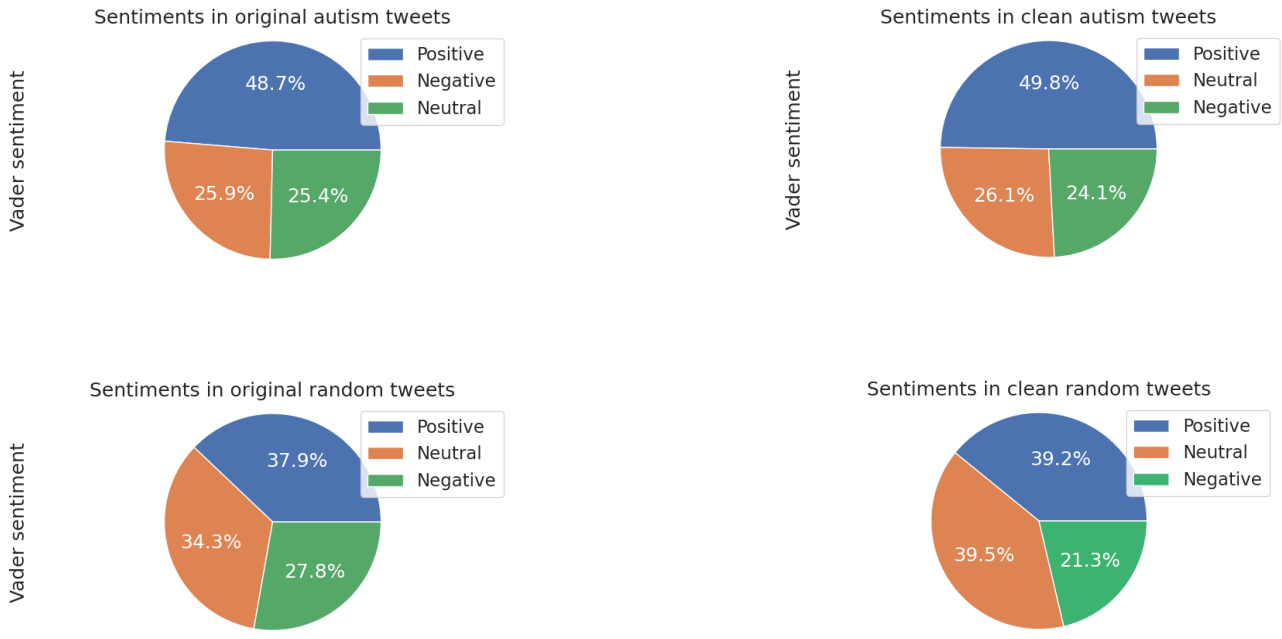


Fig. 3: Distribution of sentiments in the autism and control group dataset

This was supported by another interesting observation, where the autistic individuals tend to use more characters or words in their content as compared to the control group, shown in Fig 4. The histograms of the tweet word counts of both the groups follow similar distributions, but with a substantial difference in their means. This is clearly indicative of the differences in textual patterns between the two groups and a similar pattern was observed in the word count distribution of each sentiment as well (Fig 5).

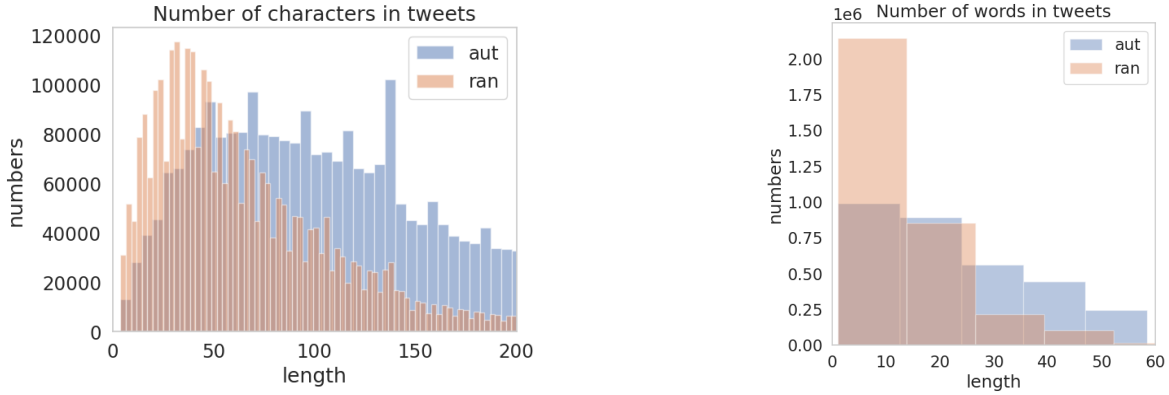
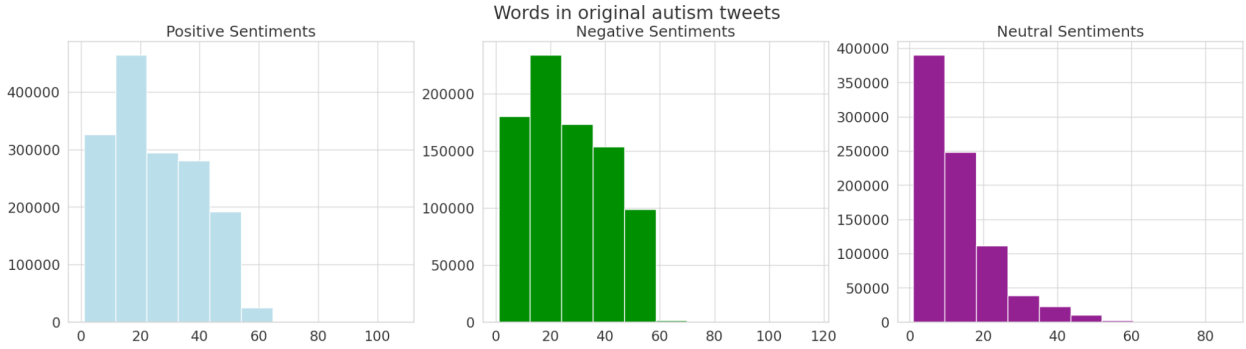


Fig. 4: Histograms of tweet character and word counts of the two groups.



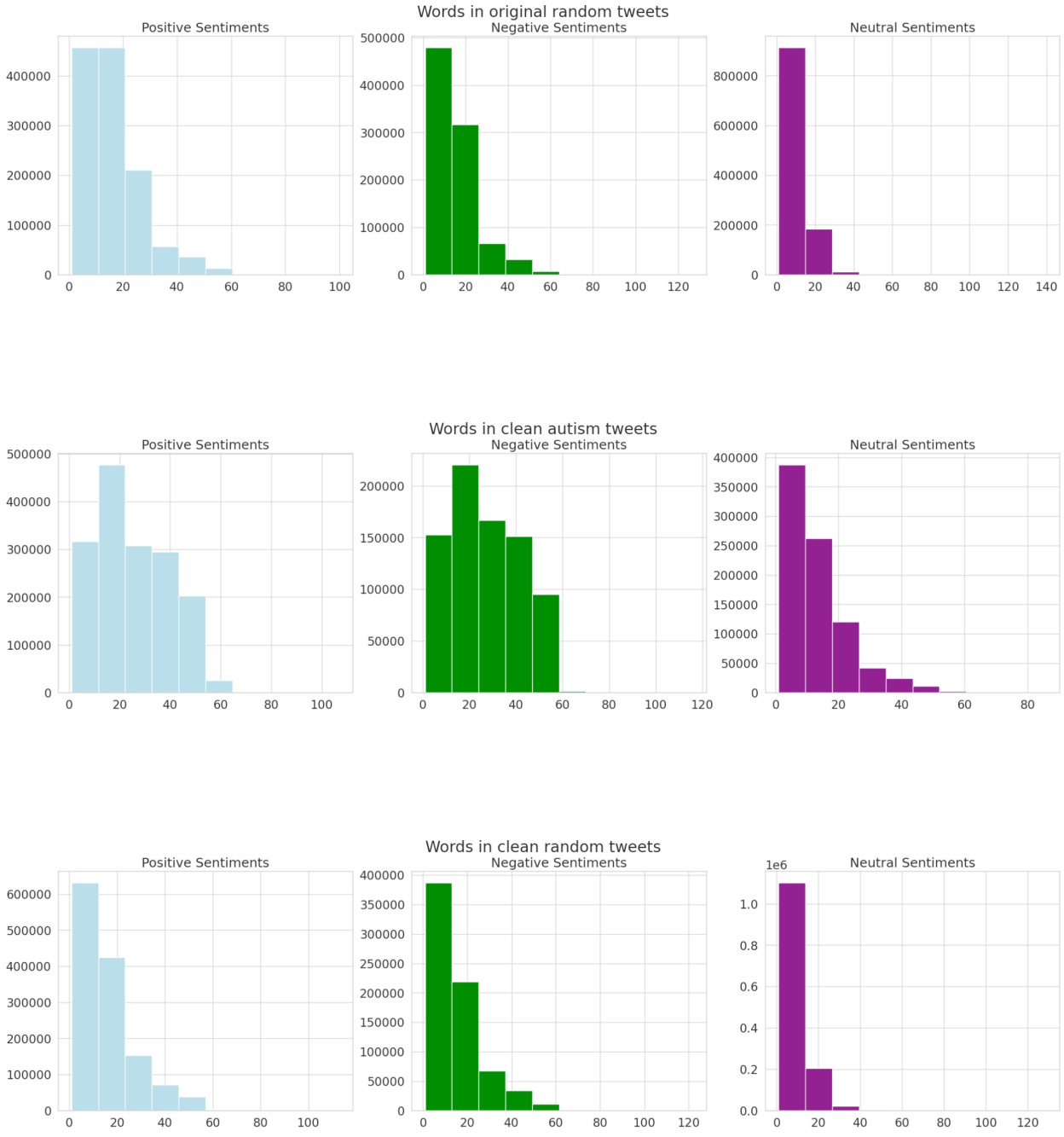


Fig. 5: Distribution of word counts in positive, negative and neutral sentiments in autism and control group dataset.

We also visualized the positive and negative sentiments in the tweets using a word cloud (Fig 6). This word cloud plot was followed by studying a few randomly sampled tweets from both positive and negative sentiment categories to analyze the contribution of words in predicting

emotional intensity. The positive tweets were found to be associated with sharing support, providing helpful resources, individuals acknowledging and embracing their identity, and showcasing their artistic or creative side. In contrast, the negative sentiment tweets were linked to struggles, feelings of dejection experienced on a particular day, or negative attitudes encountered from others.

Fig. 6: Word cloud plots of the positive and negative sentiments

5. TEXT CLASSIFICATION

5.1. Data Preprocessing

Working with unstructured, raw Twitter data is challenging because the conversational text has too much noisy information such as punctuation, abbreviations, emojis and other stray characters. Thus, before using such data for training a model, it is necessary to perform data cleaning and preprocessing, which is an essential step in any natural language processing (NLP) task. NLP makes human language understandable to computer programs and python provides a package called Natural Language Toolkit or NLTK for the same. The preprocessing usually involves:

- Removing profanity: One significant issue in this context is the use of profane language in tweets, where individuals often express their emotions using profanity, such as cursing or swear words. However, the socially offensive use of a language could also be related to hatred or harassment on social networks and must be avoided. To address this issue, we utilized the better-profanity Python library [65], which is designed to flag inappropriate words using string comparison and mask them using special characters (the default setting uses "*"). In our study, we applied this library to clean our entire dataset and remove any profane words from user tweets.

- Tokenization: breaking down a text into smaller units such as words or sentences to build a vocabulary.

- Lower case conversion: the computer treats each case differently and that could increase the data dimension so doing this would reduce redundancy and complexity of the corpus.

- Stop word removal: the very common words such as “a”, “an”, “in”, “is”, “for” etc add noise to the features. Because of their high frequency they can affect model complexity and performance as well, so it is better to remove these words to keep only important words to train the model.

- Stemming: oftentimes a word is used in many different contexts and formats such as “run”, “running”, “ran”, where the base or root word is “run”. So, stemming removes the suffix and prefix from a word to keep only the core word and ignores its usage.

- Lemmatization: stemming a word sometimes leaves just a fragment of words which may or may not make sense so lemmatization too reduces a word to its base form but gives a complete English word. Eg: the word “studies” would be stemmed to “studi” but lemmatized to “study”.

Other than this NLTK also provides methods like Part of Speech (POS) tagging, which labels the words from a text according to their part of speech i.e. nouns, pronouns, verbs, adjectives and named entity recognition (NER) to extract date, names, time, locations etc from noun phrases.

5.2. Text representation

Real-world textual data is typically unstructured and messy, rendering it unfit for direct use with machine learning algorithms that require well-defined, fixed-length data. Additionally, to be understood and processed by machines, the input text must be converted into numerical representations through a process known as feature extraction or encoding. Some of the popular text encoding approaches, used in this research work, are discussed below:

5.2.1. Bag of words (BoW)

One of the most commonly used and simple approaches for feature extraction from text is the Bag of Words (BoW). This technique uses a vocabulary of words and their frequency of occurrence in the document to generate a mapping of word vectors representing each sentence. Here, any information about the structure or order of words is ignored and only the word count is used as a feature. However, this approach could lead to a sparse vector representation where more common words, such as "the," may receive higher weights and hide the less frequent, informative words. To address this issue, n-grams, which are sequences of n-tokens, are used to group words together and create a vocabulary.

5.2.2. Term frequency-inverse document frequency

To prevent more frequent words outweighing the less frequent ones, a solution is to rescale the frequency of words by their count in all the documents, thereby penalizing the scores. This

brings the rare words, containing more information, into highlight. An easy mathematical way to understand is:

Term frequency (TF) = $\frac{number\ of\ repeating\ words\ in\ a\ document}{total\ number\ of\ words\ in\ a\ document}$

and, inverse document frequency, IDF (w, D) = $\log\left(\frac{number\ of\ documents}{number\ of\ documents\ where\ the\ word\ 'w'\ appear}\right)$

where, |D| = corpus size or number of documents.
Finally, TF-IDF is the product of two aforementioned frequencies = TF * IDF.

While the BoW model has shown promising results in text classification and language modeling, on a smaller corpus, it has limitations as well. The vocabulary used to build the model should be chosen carefully to avoid complexity and sparsity, which can significantly increase computational time. Furthermore, since BoW ignores the order of words, the resulting document may lose its semantics, which can lead to the model being trained differently.

5.2.3. Word embeddings using Word2Vec

The shortcomings of BoW can be addressed by using a dense word-vector representation or word embeddings, which can also capture the meaning of the sentence in a large corpus. In this method, each point in the embedding space represents a word, which learns and moves around in the space to refer to a target word. The resulting vector representation allows similar words to locally cluster together in the space and have similar embeddings. This use of word embeddings has allowed deep learning to take over text data and has given breakthrough performance in tasks such as machine translation. A team of researchers led by Tomas Mikolov from Google [66] developed a vector space for word representations, known as Word2Vec. Python provides support for an open-source library called Gensim, which enables working with word embeddings with a focus on topic modeling. Word2Vec is also supported in Gensim to learn word embeddings from a corpus. Prior to training the model, each sentence must be tokenized after which multiple queries can be made from the resulting word vectors.

5.3. Machine learning algorithms

In this section, we discuss various machine learning algorithms that were employed to solve the binary classification problem in our work, which involved identifying whether a given user tweet belonged to an autistic individual or the control group.

5.3.1. Logistic regression

Logistic regression is a supervised learning algorithm that utilizes a linear equation with independent real value input, the output of which is then passed through a probabilistic sigmoid function to generate a number between 0 and 1. These probabilistic predictions are then labeled into one of the two classes using a threshold value or decision boundary. The mathematical representation of the sigmoid function is:

$$p(x) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 * x)}}$$

where x is the input value, p(x) is the probabilistic output, β0 and β1 are the bias and coefficient terms respectively.

5.3.2. Support Vector Machines

Another popular supervised learning algorithm used for text classification is Support Vector Machine (SVM). In this algorithm, each input is plotted in an n-dimensional space, and the objective is to find an optimal hyperplane that can differentiate the data points into separate classes with maximum between-class distance. The dimension 'n' of the hyperplane is dependent on the number of features. If the input is not linearly separable, SVM creates a new variable using "kernel," which maps the low-dimensional input to a high-dimensional space to make it easily separable.

5.3.3. Naive Bayes

Naive Bayes is a probabilistic algorithm that assumes equal contributions and independence of the features to predict the class. It is based on Bayes' theorem, which states that given the conditional probabilities and some prior knowledge, the probability of the hypothesis can be determined. Mathematically this can be written as:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

where, P(A | B) is the posterior i.e. probability of hypothesis given the event occurs, P(B | A) denotes likelihood i.e. probability that the hypothesis is true, P(A) is prior and P(B) is marginal probability.

For the classification task, the target vector (class labels) substitute event A, and the data matrix (features) substitute event B in the Bayes theorem equation. Since the features are assumed to be conditionally independent given the class 'y', we get:

$$P(y \mid X) = P(y \mid x_1, x_2, \ldots, x_d) = \frac{P(y) \prod P(xi \mid y)}{P(x1)P(x2)P(x3)\ldots P(xd)}$$

Here, P(y | X) is the probability of observing class 'y' given the feature vector X. The feature vector has the dimension 'd' i.e. X = (x_1, x_2,.... x_d), which denotes the number of features/variables of the sample. The probability of each class is estimated from the training data, and the conditional probability of each feature given the class is calculated as the frequency of the feature in the training examples of that class. This results in a probability distribution over the possible classes for a given input text, and the class with the highest probability is chosen as the prediction. Since the denominator remains constant, the above equation can be simplified as:

P(y | x1, x2, ..., xn) = P(y) * P(x1 | y) * P(x2 | y) * ... * P(xn | y)

Now, for all the possible values of the class 'y' the probability of a given document will be the output with maximum probability i.e.

$$y = \text{argmax } P(y) \prod P(xi \mid y)$$

5.3.4. Gradient Boosting

Gradient Boosting is an ensemble learning method that iteratively trains a sequence of weak learners to minimize the error of the previous learners. The weak learners are usually decision trees, and the gradient boosting algorithm fits them sequentially to the negative gradient of the loss function of the previous tree. This approach can improve the overall model performance by reducing the bias and variance of the model. eXtreme Gradient Boosting (XGBoost) is an

extension of the gradient boosting algorithm that incorporates additional regularization techniques and can handle sparse data efficiently.

5.3.5. Training a machine learning model

Once a machine learning model has been trained on a dataset, it is essential to evaluate its performance to ensure its reliability and robustness when exposed to new, unseen data. However, when the entire dataset is used for training, assessing the model's performance on new data becomes challenging. To address this issue, it is standard practice to split the dataset into two or three sections, depending on its size. The majority of the data is used for training, enabling the model to learn the underlying patterns, referred to as the training set. A smaller subset, known as the validation set, is reserved for evaluating the model's performance during training. This step facilitates tuning and optimizing hyperparameters of the model, improving its performance for classification tasks. Finally, the model's performance is evaluated on a completely separate, unseen test dataset to determine its generalization ability to new, real-world data.

Cross-validation is another technique used in machine learning to assess model performance. This technique involves dividing the training data into several subsets, with one subset used for evaluation while the remaining subsets are used for training. This process is repeated multiple times, with each subset serving as a validation set once. Finally, the results of each validation step are averaged to obtain a more robust estimate of the model's performance.

5.3.6. Evaluation metrics

The selection of an appropriate metric for evaluating the performance of a machine learning model is task-dependent and application-specific. For instance, classification problems typically employ metrics such as accuracy, precision, F1 score, and recall, while regression problems rely on mean squared error (MSE), mean absolute error (MAE), or root mean square error (RMSE). Given that the present study pertains to a classification problem, we will utilize the former ones. Most of the classification metrics usually rely on *confusion matrix*, an NxN matrix (N being the number of classes), which uses each test sample prediction to integrate information. Each prediction is associated with one of the four outcomes:

- True Positives (TP): prediction and real output are both true
- True Negatives (TN): prediction and real output are both false
- False Positives (FP): predicted output is true but the real output is false
- False Negatives (FN): predicted output is false but the real output is true

Based on this, *accuracy* quantifies the number of times the classifier provides correct results, as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is a widely used metric, but it may be misleading in cases of imbalanced samples in the data, and can introduce bias towards the majority class samples. However, since our data is nearly balanced, as discussed in a later section, it is acceptable to use accuracy as a performance metric.

Precision and *Recall* are commonly utilized in medical data, where the former measures the number of correctly classified samples, and the latter measures the proportion of positive samples that were correctly identified.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{and Recall} = \frac{TP}{TP + FN}$$

Lastly, the F1 score combines precision and recall using their harmonic mean, thereby mitigating the influence of large outliers:

$$\text{F1 score} = 2 * \frac{precision * recall}{precision + recall}$$

5.4. Methods

The main objective of this study was to develop a classification model that could differentiate between tweets posted by individuals with Autism Spectrum Disorder (ASD) and those belonging to a control group. To accomplish this, we first identified unique usernames from both the ASD and control datasets, and split each set into a 85:15 ratio for training and testing. This was done to avoid data leakage, which could occur if a user's tweets were split between training

and testing username datasets and cause the model to overfit to learn the semantic patterns specific to an individual user. The training and testing text datasets were then constructed from the preprocessed tweets associated with each user. 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.

We initially used a simpler Bag-of-Words (BoW) model transformed by Term Frequency-Inverse Document Frequency (TF-IDF) to construct the text representation, using 25,000 features. The BoW representation produced an n x m matrix, where n is the number of documents in the corpus and m is the number of features. We then compared the performance of various machine learning algorithms, including Support Vector Machines, Naive Bayes, Logistic Regression, and XGBoost gradient boosting, using 5-fold cross-validation and F1 score as the evaluation metric. We also measured the time taken for each algorithm during model training.

After identifying the most effective and efficient algorithm, we trained the model again on the entire training dataset and evaluated its performance on the testing dataset by constructing a confusion matrix. As our corpus was large, we also trained a word2vec model using the same algorithm, with a vector size of 500, to generate word embeddings for improved feature representation.

5.5. Results

For the four algorithms used to train the data using cross-validation, the obtained F1 scores were 0.615, 0.598, 0.615, and 0.624 for SVM, Naive Bayes, logistic regression and XGBoost respectively. While the scores were similar, logistic regression was chosen as the best predictor due to its superior performance and shorter training time. Subsequently, the logistic regression model was trained on the complete training dataset, and an accuracy of 63% was obtained. The logistic regression model was then trained on the word2vec features, and a better accuracy of 73% was obtained.

	precision	recall	f1-score	support
0	0.62	0.69	0.66	423853
1	0.63	0.56	0.59	400259
accuracy			0.63	824112
macro avg	0.63	0.62	0.62	824112
weighted avg	0.63	0.63	0.62	824112

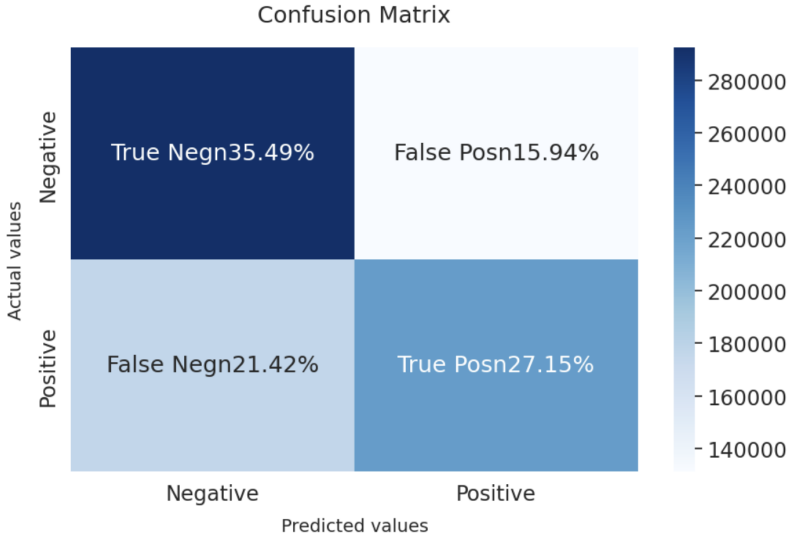


Fig. 7: Confusion-matrix of BoW + LR model, which gave an accuracy of 63%

	precision	recall	f1-score	support
0	0.73	0.75	0.74	424834
1	0.72	0.70	0.71	399278
accuracy			0.73	824112
macro avg	0.73	0.72	0.72	824112
weighted avg	0.73	0.73	0.73	824112

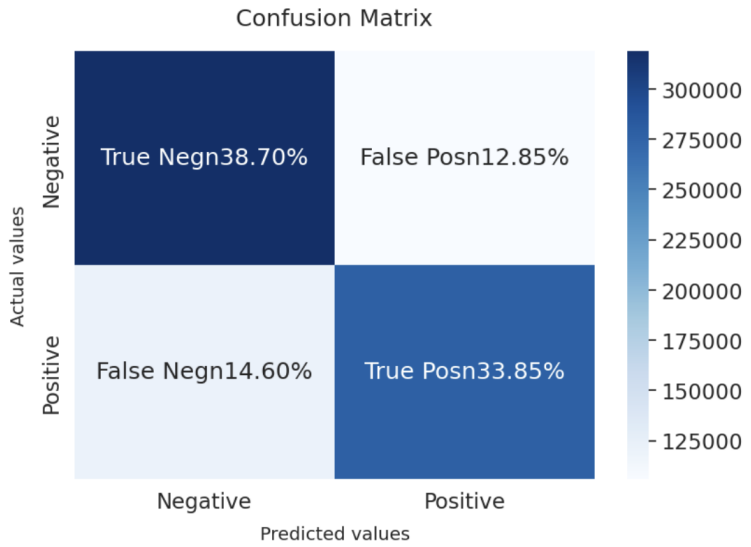


Fig. 8: Confusion matrix of word2vec + LR model giving an accuracy of 73%. Note that the model has improved in predicting fewer incorrect classes.

The results of the study 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.

6. EXPLAINABILITY AND TOPIC MODELING

6.1. Explainability

The field of machine learning and deep learning has revolutionized the data-driven world with its continuous and promising advances. However, the inability of models to help humans understand or interpret their black box nature still limits their potential. Despite becoming more accurate, robust, and autonomous, it is essential for machine learning systems to gain human trust, particularly in critical domains such as healthcare, national security, or judiciary. Such applications require algorithms to explain their predictions with a rationale and characterize their strengths and weaknesses while maintaining a high level of accuracy. The lack of transparency in a model can make humans lose control over decision support agents, introduce bias, and raise ethical or legal concerns. To address these challenges, the concept of explainable artificial intelligence (XAI) was introduced, which involves *a set of processes and methods that allow human users to comprehend and trust the results and output created by machine learning algorithms* [67].

In 2016 [67], two techniques were introduced to introduce trust in machine learning models: Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). This research work specifically uses LIME, so this section will describe the intuition behind it. LIME aims to use interpretable explanations that are understandable to humans over feature representations used by the classifier. As the name suggests, it uses a local surrogate model to explain the classifier around the sample that is being predicted. The model-agnostic nature of LIME utilizes the weights of the perturbed neighborhood points of the sample to interpret the associated predictions. For text data, perturbation can be introduced by the presence or absence of a word and how each word contributes to the class prediction of a document.

6.2. Topic modeling

Topic modeling is an unsupervised learning technique that is used to identify semantic structures in the documents to find hidden topics and cluster the related expressions and word groupings. In this work we used the Top2Vec algorithm, which clusters semantically similar words in the same topic by making use of the spatial proximity of these words. To avoid sparse vector

representation, Top2Vec uses uniform manifold approximation and projection (UMAP) dimensional reduction, followed by a hierarchical density-based spatial clustering of applications with noise (HDBSCAN) clustering. This reserves the local and global structure of the vector space and finally the topic vectors are extracted from the centroids of the dense vectors.

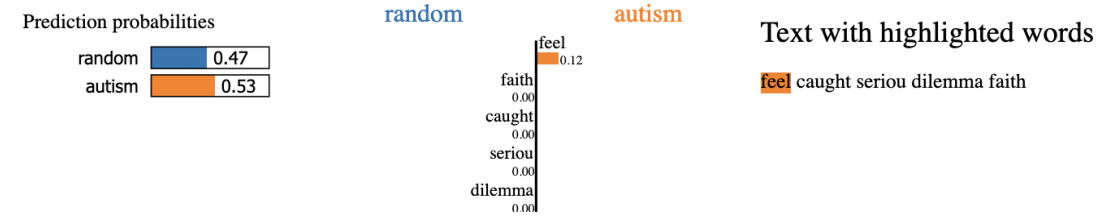
6.3. Methods

To get the general idea of the black-box model's decision making process, the LIME explainer was used with TF-IDF feature vectors and the classifier linked together into a pipeline. A surrogate model is trained on the neighborhood instances of the selected few individual test examples, using the same machine learning algorithm as the original model. This returns feature importance scores for each feature in the instance by analyzing the surrogate model's coefficients as an explanation for the predictions allowing us to understand which words or features are distinguishing the two groups. For topic modeling, the Top2Vec algorithm was utilized to extract word and sentence embeddings using a pretrained embedding model known as Universal Sentence Encoder.

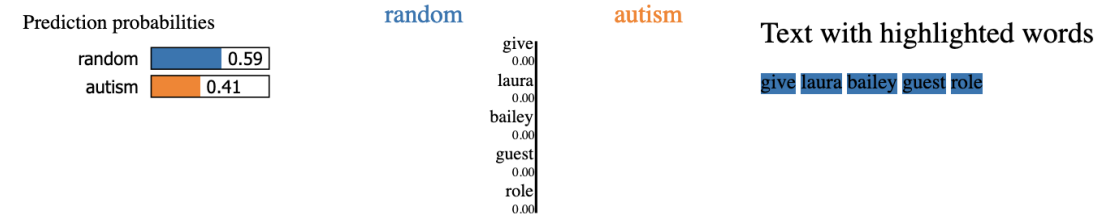
6.4. Results

For this we used LIME explanations with the BoW model and the analysis revealed that words such as "brain," "think," "feel," and "need" were associated with tweets related to autism, indicating that the BoW model was still able to capture meaningful features related to autism. This finding was further supported by the word cloud analysis, which was constructed using the 100 most frequent words from the ASD and control groups, as well as the complete dataset and the frequency of a word is encoded by its corresponding font size on the plot. The top most frequent words used in the ASD tweets were more emotion-driven, with words related to "kids" and "school" being more prevalent. This finding aligns with previous research that has highlighted the importance of early childhood intervention and thus it is reasonable for the parents or autistic individuals themselves to seek help or raise awareness and support for ASD. Conversely, the control group tweets were characterized by daily activities and verb-based patterns, reflecting a more typical language use.

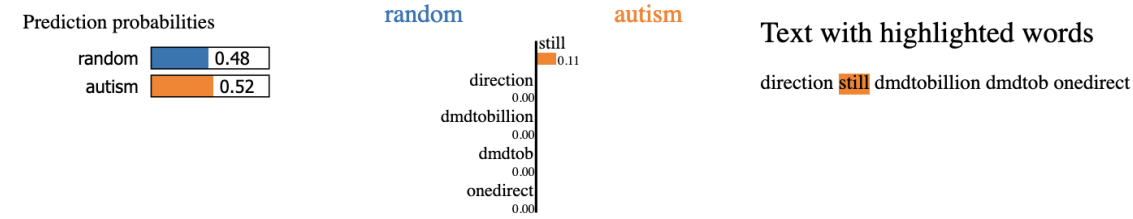
Feeling caught between serious dilemmas and faith .
True label: autism



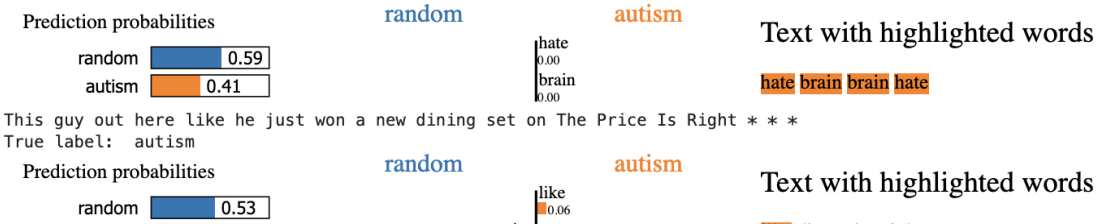
Give Laura Bailey a guest role in this too .
True label: autism



DIRECTIONERS STILL HERE 💞 #DMDTo1Billion #dmdTo1B @onedirection https://t.co/gESS7gBnks
True label: random



I hate my brain , and my brain hates me .
True label: autism



This guy out here like he just won a new dining set on The Price Is Right * * *
True label: autism

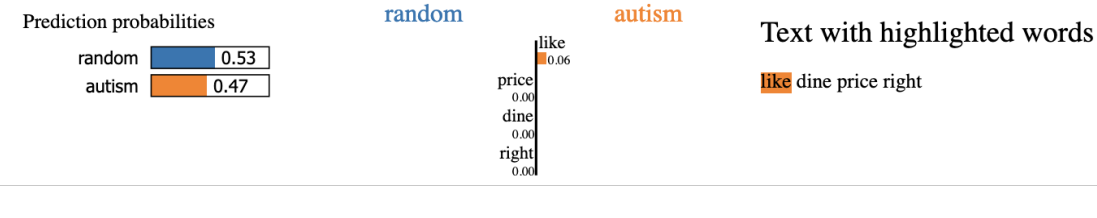


Fig. 9: LIME text explainer. Note the highlighted words emphasizing the prediction for class “autism”

The objective of our topic modeling analysis was to investigate whether there exist specific themes that are frequently discussed in relation to autism. Using the autism dataset, multiple topics were discovered, however, the majority of them revolved around ASD or related behavioral and emotional symptoms, such as "hyperactivity", "tourette", "fidgeting", "schizophrenia", "jitters", "fear", "anxiety", "trembling", and "overwhelmed", as well as disability-related topics, such as "wheelchair", "paralysis", "bedridden", "crippled", and "psychosomatic". Interestingly, a considerable number of documents were also focused on "vaccine", "therapy", "misdiagnosis", and "cats". These could be explained by the fact that autism is frequently misdiagnosed or diagnosed late, leading to worsening of symptoms. Consequently, individuals experiencing such symptoms seek therapy, support, and self-help from others. Furthermore, misinformation about autism being caused by vaccinations can negatively impact the lives of those affected by autism. However, given the timeframe in which the dataset was collected, it is also possible that these tweets are related to COVID-19 vaccines. Lastly, multiple studies [68, 69, 70] have found that autistic children are more at ease with cats, as they are less intrusive, do not maintain prolonged eye contact, and can help to relieve stress and understand emotional cues.

On the contrary, it was difficult to extract specific topics from control group Twitter conversations as they were quite random. Most of the topics were related to internet personalities, different exclamations used in daily conversations, discussions around specific days of the week or special days such as birthdays or anniversaries. Surprisingly some topics were found to be related to animals, in general, as opposed to just cats which was observed in autistic user conversations. Some of these posts also displayed usage of emotional words suggesting that pets or animals may provide therapeutic benefits.

DISCUSSION

The work presented in this study demonstrates the potential of using data-mining techniques to learn about autism and related topics from social media platforms such as Twitter. The 73% achieved accuracy in the study shows that there are significant semantic differences in the messages posted by individuals with and without ASD. This finding, along with previous studies published in *Nature Digital Medicine* and *PubMed journal* using computer vision models [14, 72], suggests that social phenotypical behavior could be used to support effective ASD screening strategies and facilitate early detection. Using machine-learning tools, Drimalla et al [14] *detected individuals with ASD with an accuracy of 73%, sensitivity of 67%, and specificity of 79%, based on their facial expressions and vocal characteristics alone and found that the performance was equal to clinical expert ratings*. The collected dataset in our study has valuable information related to various topics of discussion, which could help public health officials, policymakers, and clinicians in decision-making. By observing the behavioral symptoms of individuals in a non-clinical setting, clinicians can reduce doctor-patient meeting time and improve the rigor of autism research. Overall, this study highlights the potential of using social media data for ASD research and its potential impact on healthcare, which could be improved by combining 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.

However there are certain limitations to consider in this work as well. While the study focused on individuals who self-identified as autistic, there is no clinical validation for their autism diagnosis. Furthermore, there is a possibility of data leakage, where the identified users may not be autistic but instead could be family members, parents, caregivers, advocacy organizations, or researchers belonging to a different study population. The use of VADER to label the emotional intensity and sentiments of the tweets can be relatively inaccurate than human labels, whose sentiments tend to get affected by their surroundings, politics and other factors, thus making it difficult to provide reliable labels. VADER uses dictionary mapping to label the social media posts that may not be generalizable to different topics being discussed. Additionally, the 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 population of Twitter users.

This study presents several opportunities for future research. The rapid advancements in deep learning techniques offer potential for improved results using recurrent neural networks (RNNs) or convolutional neural networks (CNNs) applied to the large corpus of data curated in this study. In future we plan to use pre-trained models such as BERT for text classification, topic modeling, and feature extraction and integrate it with additional modality data, such as audio, video, and clinical texts, which could potentially enhance the reliability and accuracy of ASD diagnosis. 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, gender analysis using crowdsourcing or other metadata analysis techniques may hold promise for future investigations.

REFERENCES

[1] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th edition. Arlington, VA: American Psychiatric Association; 2013. ISBN:0890425558

[2] 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, 2020. MMWR Surveill Summ 2023;72(No. SS-2):1–14. DOI: http://dx.doi.org/10.15585/mmwr.ss7202a1

[3] Cath DC, Ran N, Smit JH, van Balkom AJ, Comijs HC. Symptom overlap between autism spectrum disorder, generalized social anxiety disorder and obsessive-compulsive disorder in adults: a preliminary case-controlled study. Psychopathology 2008;41(2):101-10. PMID:18033980

[4] Zandt F, Prior M, Kyrios M. Repetitive behavior in children with high functioning autism and obsessive compulsive disorder. J Autism Dev Disord 2007 Feb;37(2):251-9. PMID:16865546

[5] Lord C, Risi S, DiLavore PS, et al. Autism from 2 to 9 years of age. Arch Gen Psychiatry 2006;63:694–701.doi:10.1001/archpsyc.63.6.694.pmid:
http://www.ncbi.nlm.nih.gov/pubmed/16754843

[6] https://doi.org/10.53053/WRSX5772

[7] Lipkin PH, Macias MM, Council on children with disabilities, section on developmental and behavioral pediatrics. Promoting optimal development: identifying infants and young children with developmental disorders through developmental surveillance and screening. Pediatrics 2020;145. doi: doi:10.1542/peds.2019-3449

[8] Guthrie W, Wallis K, Bennett A, et al. Accuracy of autism screening in a large pediatric network. Pediatrics 2019; 144. doi:10.1542/peds.2018-3963 pmid: http://www.ncbi.nlm.nih.gov/pubmed/31562252

[9] Bishop DV, Norbury CF. Exploring the borderlands of autistic disorder and specific language impairment: a study using standardized diagnostic instruments. J Child Psychol Psychiatry 2002 Oct;43(7):917-29. PMID:12405479

[10] Rosen NJ, Yoshida CK, Croen LA. Infection in the first 2 years of life and autism spectrum disorders. Pediatrics 2007; 119:e61–9. doi:10.1542/peds.2006-1788 pmid: http://www.ncbi.nlm.nih.gov/pubmed/17200260

[11] Chaidez V, Hansen RL, Hertz-Picciotto I. Gastrointestinal problems in children with autism, developmental delays or typical development. J Autism Dev Disord 2014;44:1117–27. doi:10.1007/s10803-013-1973-x pmid: http://www.ncbi.nlm.nih.gov/pubmed/24193577

[12] Torous J, Kiang M, Lorme J, Onnela JP (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health (in press).

[13] https://thesmallbusinessblog.net/twitter-statistics/

[14] 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

[15] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

[16] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1):24–54.

[17] Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Online Communication about Depression and Anxiety among Twitter Users with Schizophrenia: Preliminary Findings to Inform a Digital Phenotype Using Social Media. Psychiatr Q 2018 Dec;89(3):569-580. [doi: 10.1007/s11126-017-9559-y]

[18] Hswen Y, Naslund JA, Chandrashekar P, Siegel R, Brownstein JS, Hawkins JB. Exploring online communication about cigarette smoking among Twitter users who self-identify as having schizophrenia. Psychiatry Res 2017 Dec;257:479-484 [FREE Full text] [doi: 10.1016/j.psychres.2017.08.002]

[19] Danielle Mowery and Craig Bryan and Mike Conway. Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health. https://arxiv.org/abs/1701.08229

[20] https://mhanational.org/

[21] Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One. 2010 Nov 29;5(11):e14118. doi: 10.1371/journal.pone.0014118. PMID: 21124761; PMCID: PMC2993925.

[22] B. Robinson, R. Power, and M. Cameron. An evidence based earthquake detector using Twitter. In Proc. Workshop on Language Processing and Crisis Information, pages 1–9, 2013.

[23] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proc. International Conference on World Wide Web , pages 851–860, 2010.

[24] A. Taylor Newton, Adam D.I. Kramer, and Daniel N. McIntosh. 2009. Autism online: a comparison of word usage in bloggers with and without autism spectrum disorders. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). Association for Computing Machinery, New York, NY, USA, 463–466. https://doi.org/10.1145/1518701.1518775

[25] Nguyen, T., Duong, T., Phung, D., Venkatesh, S. (2014). Affective, Linguistic and Topic Patterns in Online Autism Communities. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds) Web Information Systems Engineering – WISE 2014. WISE 2014. Lecture Notes in Computer Science, vol 8787. Springer, Cham. https://doi.org/10.1007/978-3-319-11746-1_35

[26] van Schalkwyk GI, Marin CE, Ortiz M, Rolison M, Qayyum Z, McPartland JC, Lebowitz ER, Volkmar FR, Silverman WK. Social Media Use, Friendship Quality, and the Moderating Role of Anxiety in Adolescents with Autism Spectrum Disorder. J Autism Dev Disord. 2017 Sep;47(9):2805-2813. doi: 10.1007/s10803-017-3201-6. PMID: 28616856; PMCID: PMC6688174.

[27] Bakombo S, Ewalefo P, Konkle ATM. The Influence of Social Media on the Perception of Autism Spectrum Disorders: Content Analysis of Public Discourse on YouTube Videos. Int J Environ Res Public Health. 2023 Feb 13;20(4):3246. doi: 10.3390/ijerph20043246. PMID: 36833941; PMCID: PMC9961260.

[28] Teo JX, Davila S, Yang C, Hii AA, Pua CJ, Yap J, Tan SY, Sahlén A, Chin CW, Teh BT, Rozen SG, Cook SA, Yeo KK, Tan P, Lim WK. Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging. Commun Biol. 2019 Oct 4;2:361. doi: 10.1038/s42003-019-0605-1. PMID: 31602410; PMCID: PMC6778117.

[29] Abdullah S, Matthews M, Murnane EL, Gay G, Choudhury T. 2014. Towards circadian computing: “Early to bed and early to rise” makes some of us unhealthy and sleep deprived Proc. UbiComp ‘14: 2014 ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Seattle, WA, pp. 673–84. New York: Assoc. Comput. Mach.

[30] Wang R, Chen FL, Chen Z, Li TX, Farari G, et al. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones Proc. UbiComp ‘14: 2014 ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Seattle, WA, pp. 3–14. New York: Assoc. Comput. Mach

[31] Ciman M, Wac K, Gaggi O. 2015. Assessing stress through human-smartphone interaction analysis Pervasive Health ‘15: Proc. 9th Int. Conf. Pervasive Comput. Technol. Healthc., Istanbul.Brussels: Inst. Comput. Sci. Social-Inform. Telecom. Eng. http://ieeexplore.ieee.org/document/7349382/

[32] Rachuri KK, Musolesi M, Mascolo C, Rentfrow PJ, Longworth C, Aucinas A. 2010. Emotion sense: a mobile phones based adaptive platform for experimental social psychology research Proc. UbiComp ‘10: 2010 ACM Conf. Ubiquitous Comput., Copenhagen, Denmark, pp. 281–90. New York: Assoc. Comput. Mach.

[33] Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. PLoS One. 2014 Jan 29;9(1):e86191. doi: 10.1371/journal.pone.0086191. PMID: 24489699; PMCID: PMC3906034.

[34] M. Nadeem. (2016). ‘‘Identifying depression on twitter.’’ [Online]. Available: https://arxiv.org/abs/1607.07384

[35] Signorini A, Segre AM, Polgreen PM. 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. 2011 May 4;6(5):e19467. doi: 10.1371/journal.pone.0019467. PMID: 21573238; PMCID: PMC3087759.

[36] Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). Association for Computing Machinery, New York, NY, USA, 115–122. https://doi.org/10.1145/1964858.1964874

[37] Aramaki, E., Maskawa, S., & Morita, M. (2011, July). Twitter catches the flu: detecting influenza epidemics using Twitter. In Proceedings of the 2011 Conference on empirical methods in natural language processing (pp. 1568-1576).

[38] Ye J, Hai J, Wang Z, Wei C, Song J. Leveraging natural language processing and geospatial time series model to analyze COVID-19 vaccination sentiment dynamics on Tweets. JAMIA Open. 2023 Apr 12;6(2):ooad023. doi: 10.1093/jamiaopen/ooad023. PMID: 37063408; PMCID: PMC10097455.

[39] Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. J Med Internet Res. 2020 Oct 12;22(10):e22635. doi: 10.2196/22635. PMID: 32936777; PMCID: PMC7575341.

[40] Ye J. Pediatric Mental and Behavioral Health in the Period of Quarantine and Social Distancing With COVID-19. JMIR Pediatr Parent. 2020 Jul 28;3(2):e19867. doi: 10.2196/19867. PMID: 32634105; PMCID: PMC7389340.

[41] Gupta V, Jain N, Katariya P, et al. An emotion care model using multimodal textual analysis on COVID-19. Chaos Solitons Fractals. 2021;144:110708. doi: 10.1016/j.chaos.2021.110708.

[42] Dhingra S, Arora R, Katariya P, Kumar A, Gupta V, Jain N. Understanding Emotional Health Sustainability Amidst COVID-19 Imposed Lockdown. In: Agrawal R, Mittal M, Goyal LM, editors. Sustainability Measures for COVID-19 Pandemic. Singapore: Springer; 2021.

[43] Wu C, Liaqat S, Helvaci H, Cheung SS, Chuah CN, Ozonoff S, Young G. Machine Learning Based Autism Spectrum Disorder Detection from Videos. Healthcom. 2021 Mar;2020:10.1109/healthcom49281.2021.9398924. doi: 10.1109/healthcom49281.2021.9398924. Epub 2021 Apr 14. PMID: 34693405; PMCID: PMC8528233.

[44] Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, et al. (2018) 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

[45] M. S. Mythili, and AR Mohamed Shanavas. (2014) “A study on Autism spectrum disorders using classification techniques.” International Journal of Soft Computing and Engineering (IJSCE), 4: 88-91.

[46] Hswen Y, Gopaluni A, Brownstein JS, Hawkins JB. Using Twitter to Detect Psychological Characteristics of Self-Identified Persons With Autism Spectrum Disorder: A Feasibility Study. JMIR Mhealth Uhealth. 2019 Feb 12;7(2):e12264. doi: 10.2196/12264. PMID: 30747718; PMCID: PMC6390184.

[47] Beykikhoshk A, Arandjelović O, Phung D, Venkatesh S, Caelli T. Using Twitter to learn about the autism community. Soc Netw Anal Min. 2015 Jun 2;5(1) doi: 10.1007/s13278-015-0261-5.

[48] Corti L, Zanetti M, Tricella G, Bonati M. Social media analysis of Twitter tweets related to ASD in 2019-2020, with particular attention to COVID-19: topic modelling and sentiment analysis. J Big Data. 2022;9(1):113. doi: 10.1186/s40537-022-00666-4. Epub 2022 Nov 25. PMID: 36465137; PMCID: PMC9702597.

[49] Briand, A., Almeida, H., Meurs, MJ. (2018). Analysis of Social Media Posts for Early Detection of Mental Health Conditions. In: Bagheri, E., Cheung, J. (eds) Advances in Artificial Intelligence. Canadian AI 2018. Lecture Notes in Computer Science(), vol 10832. Springer, Cham. https://doi.org/10.1007/978-3-319-89656-4_11

[50] Maupomé, Diego and Marie-Jean Meurs. “Using Topic Extraction on Social Media Content for the Early Detection of Depression.” Conference and Labs of the Evaluation Forum (2018).

[51] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. 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.

[52] Sanjana Garg, Jordan Taylor, Mai El Sherief, Erin Kasson, Talayeh Aledavood, Raven Riordan, Nina Kaiser, Patricia Cavazos-Rehg, Munmun De Choudhury, Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on Reddit, Internet Interventions, Volume 26, 2021, 100467, ISSN 2214-7829, https://doi.org/10.1016/j.invent.2021.100467

[53] Tadesse, Michael M. et al. “Detection of Depression-Related Posts in Reddit Social Media Forum.” IEEE Access 7 (2019): 44883-44893.

[54] H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. 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, Baltimore, Maryland, USA. Association for Computational Linguistics.

[55] Reece, A.G., Danforth, C.M. Instagram photos reveal predictive markers of depression. EPJ Data Sci. 6, 15 (2017). https://doi.org/10.1140/epjds/s13688-017-0110-z

[56] Ricard B, Marsch L, Crosier B, Hassanpour S. Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An Analytical Study on Instagram. J Med Internet Res 2018;20(12):e11817. URL: https://www.jmir.org/2018/12/e11817 DOI: 10.2196/11817

[57] Hassanpour, S., Tomita, N., DeLise, T. et al. Identifying substance use risk based on deep neural networks and Instagram social media data. Neuropsychopharmacol 44, 487–494 (2019). https://doi.org/10.1038/s41386-018-0247-x

[58] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. 2014. How do your friends on social media disclose your emotions? In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14). AAAI Press, 306–312.

[59] 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.

[60] Bellon-Harn ML, Boyd RL, Manchaiah V. Applied Behavior Analysis as Treatment for Autism Spectrum Disorders: Topic Modeling and Linguistic Analysis of Reddit Posts. Front

Rehabil Sci. 2022 Jan 5;2:682533. doi: 10.3389/fresc.2021.682533. PMID: 36188818; PMCID: PMC9397756.

[61] Thin, N., Hung, N., Venkatesh, S., Phung, D. (2017). Estimating Support Scores of Autism Communities in Large-Scale Web Information Systems. In: , et al. Web Information Systems Engineering – WISE 2017. WISE 2017. Lecture Notes in Computer Science(), vol 10569. Springer, Cham. https://doi.org/10.1007/978-3-319-68783-4_24

[62] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[63] https://github.com/JustAnotherArchivist/snscrape

[64] https://github.com/igorbrigadir/twitter-advanced-search

[65] https://github.com/snguyenthanh/better_profanity

[66] Tomas Mikolov and Kai Chen and Greg Corrado and Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

[67] Marco Tulio Ribeiro and Sameer Singh and Carlos Guestrin (2016). “Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.48550/arXiv.1602.04938

[68] Carlisle GK, Johnson RA, Wang Z, Bibbo J, Cheak-Zamora N, Lyons LA. Exploratory Study of Cat Adoption in Families of Children with Autism: Impact on Children's Social Skills and Anxiety. J Pediatr Nurs. 2021 May-Jun;58:28-35. doi: 10.1016/j.pedn.2020.11.011. Epub 2020 Dec 6. PMID: 33290937.

[69]
http://cvm.missouri.edu/cats-may-help-increase-empathy-decrease-anxiety-for-kids-with-autism/

[70] Hart Lynette A., Thigpen Abigail P., Willits Neil H., Lyons Leslie A., Hertz-Picciotto Irva, Hart Benjamin L. Affectionate Interactions of Cats with Children Having Autism Spectrum Disorder. (2018) https://www.frontiersin.org/articles/10.3389/fvets.2018.00039

[71] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

[72] Alcañiz M, Chicchi-Giglioli IA, Carrasco-Ribelles LA, Marín-Morales J, Minissi ME, Teruel-García G, Sirera M, Abad L. 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 Res. 2022 Jan;15(1):131-145. doi: 10.1002/aur.2636. Epub 2021 Nov 22. PMID: 34811930.