Comparison of Text Mining Feature Extraction Methods Using Moderated vs Non-Moderated Blogs: An Autism Perspective

Abu Saleh Md. Tayeen
New Mexico State University
Las Cruces, New Mexico, U.S.A.
tayeen@nmsu.edu

Saleem Masadeh
New Mexico State University
Las Cruces, New Mexico, U.S.A.
saleem@nmsu.edu

Abderrahmen Mtibaa
University of Missouri Saint Louis
St. Louis, Missouri, U.S.A.
amtibaa@umsl.edu

Satyajayant Misra
New Mexico State University
Las Cruces, New Mexico, U.S.A.
misra@cs.nmsu.edu

Moumita Choudhury
Texas Tech University Health Sciences Center
Lubbock, Texas, U.S.A.
m.choudhury@ttuhsc.edu

ABSTRACT

Online social media is being widely used by social scientists to study human behavior. Researchers have explored different feature extraction (FE) and classification techniques to perform sentiment analysis, topic identification, etc. Most studies tend to evaluate FE and classification methods using only one particular class of datasets—well-defined with little/no noise or with well-defined noise. For instance, when the datasets under study have different noise characteristics, various FE and/or classification methods may fail to identify a given topic.

In this paper, we fill this gap by quantitatively comparing multiple FE methods and classifiers using three different datasets (two moderator-controlled blogs and one single-authored personal blogs) related to Autism Spectrum Disorder (ASD). Our result shows that no particular combination of FE and classifier is the best overall, but choosing the right ones can improve accuracy by over 30%.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Information extraction**.

KEYWORDS

Autism; feature evaluation; classifier; blogs; ASD.

ACM Reference Format:
Abu Saleh Md. Tayeen, Saleem Masadeh, Abderrahmen Mtibaa, Satyajayant Misra, and Moumita Choudhury. 2019. Comparison of Text Mining Feature Extraction Methods Using Moderated vs Non-Moderated Blogs: An Autism Perspective. In *9th International Digital Public Health Conference (2019) (DPH' 19), November 20–23, 2019, Marseille, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357729.3357740

ACM ISBN 978-1-4503-7208-4/19/11…$15.00
https://doi.org/10.1145/3357729.3357740

1 INTRODUCTION

Text mining is becoming an important direction of research in the era of big data analytics as the majority of today's information consists of unstructured text [6]. In recent years, many feature extraction (FE) methods have been proposed to aid the analysis of unstructured text documents [20, 26, 27]. The features are then used by classification algorithms to categorize text content. Text documents may have different types, ranging from short to long documents, and from ones where the topics are well-defined, to others with ambiguous content. Consequently, two facets tend to be missing in most studies: (i) the studies tend to choose a few (often one) FE methods and classifiers based on their popularity, and (ii) most tend to investigate documents in one of the document types described above, and the proposed solutions are fairly prescriptive to the dataset. Sometimes these solutions may fail when tested using another dataset that contains minor characteristic differences.

In this paper, our goal is to address these two missing facets by studying the problem of text mining of ASD related blogs as a case study. We choose ASD as the topic because the very recent increase in awareness has created an ever increasing large database of misleading information (documents, comments, opinions) consisting of "confusing" and/or "erroneous" articles on the internet in general and social media in particular.

Autism Spectrum Disorder (ASD) is a group of developmental disability affecting a person's social, behavioral, and communication ability [3]. The latest report from the Centers for Disease Control and Prevention (CDC) shows that one out of 68 eight-year-olds are diagnosed with ASD in the United States [44]. This rising prevalence of children diagnosed with ASD is coupled with a shortage of ASD expertise and resources. Therefore, individuals and family members affected with ASD turn to social media resources in desperation for guidance and information, sharing personal experiences, and expressing emotional relief to navigate their situations [38].

In addition to sharing information and experiences dealing with ASD in social media sources such as personal blogs, parents of autistic children write about other general topics such as their daily activities, political opinions, traveling experiences. Most of these blog post content is not directly aligned with the topic of ASD, which make them difficult to classify. We refer to such blog posts as *noisy* posts. On the other hand, there exists online communities with a dedicated topic for discussion. In these communities, the moderators do not allow people to diverge their conversations from the preset topic. As a result, posts from these communities are less noisy and binary classification of ASD content becomes easier.

We attempt to address the challenge of binary classification of posts which may not be directly aligned to a given specific topic by choosing ASD topic as a case study and experimenting with two different types of datasets–the ones that are very noisy (the unadministered, single-authored social blogs) and the ones with little noise (moderated blogs).

Contributions: In our work, we present a qualitative and quantitative comparison of different approaches of word-level feature extraction methods combined with multiple families of classifiers using *moderator-controlled* and *single-authored* datasets. We experiment with three distinct word-level feature extraction approaches: *frequency-based* which create features using the frequency estimates of words, *context-based* that captures the context of words through an underlying neural network, and *hybrid* as a combination of previous approaches. We also study the impact of three diverse families of classifiers, non-probabilistic linear, ensemble learning bagging, and ensemble learning boosting, on the performance of these feature extraction approaches. Our data-driven analysis is based on two *moderator-controlled* datasets, *LiveJ* and *RedIt*, where the moderator is responsible for the content alignment of the posts within the topics of the community (which helps reduce the noise in the dataset) and one *single-authored* personal blogs dataset, *PersB*, where bloggers write about ASD without any restrictive rules on the syntax, semantics, and subject of the post which may introduce noise in the dataset.

Our results for the binary classification problem (ASD/non-ASD) from an extensive analysis using several FE methods and classifiers on the three datasets show that classification accuracies and F1-scores vary for *moderator-controlled* and *single-authored* datasets, and the choice of the right classifier helps gain up to 30% accuracy. Our results are the first to show the impact of the (non-)presence of a moderator in the selection process of different FE methods and classifiers. We also verified our findings using classification of other topics such as *food* and *web-design*.

The rest of the paper is organized as follows. In Section 2, we review prior work of different text mining feature extraction methods and literature investigating ASD related social media content. Section 3 details the data gathering process and the characteristics of *moderator-controlled* vs *single-authored* datasets. Section 4, describes the adopted methodology to evaluate the feature extraction methods using multiple classifiers. In Section 5, we present the qualitative and quantitative comparison and discuss our findings. Finally, we conclude and present our future directions in Section 6.

2 RELATED WORK

In this section, we first provide a literature review of the state-of-the-art feature extraction methods which represent the building blocks for any text mining research. Then, we list the most related research studies leveraging online social media content to improve access to ASD information and assess the social support within the ASD community.

2.1 Feature Extraction from Text

In recent years, many feature extraction (FE) methods have been proposed to aid the analysis of unstructured information such as text documents [1, 20]. Well-known FE methods include; methods to extract syntactic features for language identification [11], semantic features for word sense disambiguation [33], lexical features for sentiment analysis [45], and bag-of-words based features for document classification [18].

Since this work involves classification task, we mainly focus on word-level feature extraction (features involving word statistics, dependency between adjacent words) methods. Current state-of-the-art word-level feature extraction methods can be roughly categorized into three main approaches: *frequency-based*, *context-based*, and *hybrid*.

The *frequency-based* approach constructs features by measuring the frequency estimates of words with different granularities. In this approach, proposed methods use: (*i*) the number of occurrences of words within a document, such as Bag-of-Words [19]; (*ii*) weighted n-grams based on their discriminative scores to reduce the impact of noise introduced by common words, such as Term Frequency Inverse Document Frequency (TfIdf) [5]; (*iii*) frequency of words within predefined categories (e.g. psychological, emotional), such as Linguistic Inquiry and Word Count (LIWC) [28, 35], Affective Norms for English Words (ANEW) [8, 32], and Empath [13]; and (*iv*) the probability of the co-occurrence of different words in a document, such as Latent Dirichlet Allocation (LDA) topics [7, 29].

The *context-based* approach captures the context of words through an underlying neural network. Word2Vec-Average [24] retrieves embeddings per words in the corpus exploiting the Word2Vec model [26] and creates the final feature vector for documents by averaging such embeddings. On the other hand, Doc2Vec [23] converts each document into a continuous distributed feature vector.

In the *hybrid* approach, frequency-based methods are used to weight the feature vectors derived from the context-based feature approach. For instance, few methods utilize TfIdf weights to provide higher importance to Word2Vec feature vectors [21, 48], and others employ clustering techniques to categorize different word embeddings [2, 10].

2.2 Social Media and ASD

Various studies have investigated social media content relevant to ASD in order to (i) analyze social interactions [37, 40–42], (ii) identify ASD topics [4], and (iii) classify ASD content [5, 29, 30]. Researchers have focused on inspecting the structure of ASD-related Twitter social network (e.g. modularity and degree of separation) [40–42], and have categorized different types of social support messages (e.g. informational and emotional) exchanged between parents and caregivers of children with ASD in two Facebook groups [37]. Another study investigates the persistence of given ASD-related topics over time using Twitter [4].

Many FE methods have been proposed to classify social media content, such as tweets and blog posts into ASD-related content [5, 29, 30]. Most relevant to our work, Nguyen *et. al* have analyzed content of posts from 10 ASD and 20 non-ASD related communities from Live Journal using multiple *frequency-based* feature extraction methods, and proposed a combination, Topics+LIWC+ANEW, to predict whether a post is from an ASD related community or not [29, 30]. They found that the joint feature achieves the highest accuracy with around 93% using a corpus of 20,000 posts.

However, most of these studies did not consider (i) a large spectrum of feature extraction methods including *context-based* and *hybrid* approach; (ii) assessment of the performance of these methods with multiple classifiers, such as Random Forest [9], Ada Boost [15], XG Boost [12]; (iii) the impact of various sources that may have different post lengths and non-availability of moderators who monitor the alignment of content with a predefined topic (a major cause of diversity); and (iv) experimenting with many datasets and validating binary classification with multiple topics. *This paper fills these gaps by presenting a comparative study of these feature extraction methods and different classifiers using ASD related blogs as a principle case study.*

3 DATA

In this section, we introduce the datasets, their collection strategy and annotation process, as well as the statistics and distinct characteristics of these datasets.

3.1 Data Collection

We implement web crawlers to gather three different datasets: two from *moderator-controlled* blogs and one from *single-authored* personal blogs. Our *moderator-controlled* datasets, *LiveJ* and *RedIt*, consist of posts gathered, during January 2018, from communities in the Live Journal[1] and Reddit[2] websites respectively. In these websites, moderators administer their respective blogs, filtering-out posts irrelevant to their communities' topic of interest (e.g., *autism, fashion, pet*). We use this regulation process as a convenient alternative approach to annotate blog posts automatically without the need for labor intensive manual annotation. In the period between January and March 2017, we also collected posts from ASD-related personal blogs[3], each of these blogs is written by a single author which we refer to as *PersB* dataset.

Annotation: We annotate posts into ASD-relevant content, which we refer to as "A", and content irrelevant to topic ASD, denoted by "$\bar{A}$". By ASD-relevant content, we refer to text containing reference to ASD information, daily experiences, curing information related to ASD in general.

In the *moderator-controlled* datasets, we assume that the posts are self-labeled based on the community topic. For instance, posts from the community "asperger" in *LiveJ* are annotated[4] as "A" and posts from communities like "cat-lovers" or "htmlhelp" are annotated as "$\bar{A}$".

While our *single-authored* dataset, *PersB*, consists of blogs related to ASD, the author tends to occasionally post personal experiences making drawing the line between ASD and non-ASD related content a difficult task, which motivates us to seek annotation from ASD scientists, who we refer to as experts ($e \in E$), as well as others ($o \in \bar{E}$).

[1]https://www.livejournal.com/
[2]https://www.reddit.com/
[3]Blogs are selected from the lists in https://anautismobserver.wordpress.com and [38]
[4]We used annotation and label interchangeably.

We create a web-based survey to help participants annotate a set of 400 randomly chosen posts from each blog. An annotator, x, is assigned a fixed number of posts to read and provide a score, $s_{p_i}(x) \in \{0, 1, 2, 3, 4, 5\}$, for a given post, p_i, where the score 0 implies the post is irrelevant to ASD, score 5 represents unequivocal relevance, and a score in between implies higher association to ASD for higher values. Then, we accumulate the scores of all annotators and compute a weighted average score, s_{p_i} as shown in Equation 1:

$$s_{p_i} = \begin{cases} \alpha \dfrac{\sum\limits_{x \in E} s_{p_i}(x)}{|E|} + (1-\alpha) \dfrac{\sum\limits_{x \in \bar{E}} s_{p_i}(x)}{|\bar{E}|} & \text{, if } |E| > 0 \\ \dfrac{\sum\limits_{x \in \bar{E}} s_{p_i}(x)}{|\bar{E}|} & \text{, if } E = \emptyset \end{cases} \tag{1}$$

For every post, we assign at least three annotators. In order to give more importance to the score(s) from expert(s), we choose $\alpha = 0.6$. We then compute Fleiss' inter-rater reliability[14] metric using the scores obtained from three annotators. We found $\kappa = 0.65$, which means there exists substantial agreement among our three annotators. For each post p_i, we assign a label y_i such that:

$$y_i = \begin{cases} A, & \text{if } 2 \le s_{p_i} \le 5 \\ \bar{A}, & \text{if } 0 \le s_{p_i} \le 1 \end{cases} \tag{2}$$

The number of communities and posts in *RedIt* are less than those of *LiveJ*. Moreover, the length of posts in most *RedIt*'s communities is smaller than the minimum *post length* (at least 10 words per post) we used in the experiments. In order to balance our annotation sets, we choose a subset of all posts in *LiveJ*, such that posts are chosen from a random *LiveJ* community without replacement until we reach balanced annotation sets.

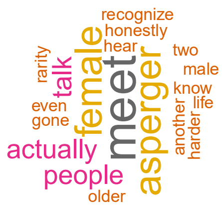
Table 1: The number of labeled posts per dataset where Total refers to the number of posts used for training in each dataset

Dataset	No. of A Posts	No. of $\bar{A}$ Posts	Total
LiveJ	4414	4377	8791
RedIt	4044	4034	8078
PersB	222	178	6140

3.2 Characteristics of Data

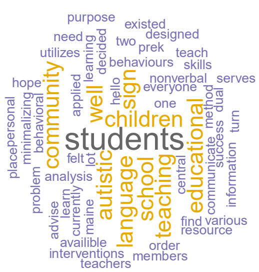
Our *LiveJ* and *RedIt* datasets contain a total of 19,924 and 10,810 posts across 22 and 18 communities respectively. In *LiveJ* , 24.35% of the posts are from the 12 ASD related communities such as *autism-spectrum* and *asperger*, and 75.65% of the posts are from 10 non-ASD related communities, such as *cat-lovers* and *htmlhelp*. In *RedIt* dataset, 48.87% of the posts are from the 8 ASD related communities, such as *asd* and *aspergers*, and 51.13% of the posts are from 10 non-ASD related communities, such as *hair* and *html*. The *PersB* dataset, has 6,287 posts from 8 different single-authored blogs such as *Autism Day By Day* and *Adventure in Autism*. We present in Tables 2, 3, and 4 the blog/community names, the number of posts, average post length (number of words per post), and the top 5 keywords from each blog/community of *LiveJ*, *RedIt*, and *PersB* datasets respectively.

We identify two main characteristics that distinguish these datasets: average *post length* for each blog/community, and *content alignment* of the posts.



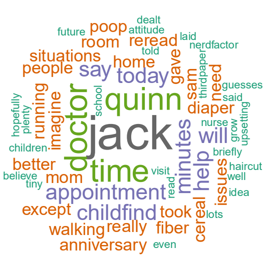
(a) A post from *aspergirls* community in *RedIt* labeled as A

(b) A post from *cats* community in *RedIt* labeled as $\bar{A}$



(c) A post from *teaching-autism* community in *LiveJ* labeled as A

(d) A post from *webdesign* community in *LiveJ* labeled as $\bar{A}$



(e) A post from *stimeyland* blog in *PersB* labeled as A

(f) A post from *stimeyland* blog in *PersB* labeled as $\bar{A}$

Figure 1: Highlighting the high content alignment of the posts in *moderator-controlled* datasets, (a)-(d), and the misalignment of sample posts the *single-authored* dataset, (e) and (f).

3.3 Post Length

We find that the average post length of the three datasets varies considerably. For instance, the average length of posts in *moderator-controlled* blogs, *LiveJ* and *RedIt*, are 74.1% (152.9) and 72.3% (162.66) shorter than the average post length in *PersB* (590) respectively. In Reddit and Live Journal, most of the bloggers tend to write shorter posts, often in form of questions to seek responses. On the other hand, in *single-authored* blogs, the average post length is 590 words per post, which indicates that the authors are accustomed to posting long articles by writing mostly their personal experiences, opinions, and information.

3.4 Content Alignment

Content alignment can be interpreted as the discursiveness or (in)coherence of the post content to a given topic such as *ASD*. We define a post to be aligned to a community, c, if its content's topic is within the scope of the community goals specified by the moderator of c. The *moderator-controlled* datasets consist of mainly posts that are aligned with the goal of the community. We highlight the high content alignment in *moderator-controlled* dataset, as shown in Figures 1a-1d. For instance, in Figure 1d, the post from *webdesign* community consists of a request for the best web programming language to study.

To demonstrate the content alignment accross all posts with a given blog, we extracted keywords from all posts of each blog and community by applying TextRank [25], a graph-based text summarizing algorithm implemented in the Gensim[5] Python library. As shown in Table 2, the *webdesign* community of *LiveJ* dataset includes the top keywords "site", "web", and, "design", that align to the goal of community, which is a community of web programmers.

We assume that *single-authored* blogs consist of one community in which its goal topic is defined by the blog title (or authors blog narrative). However, in *single-authored* blogs, writers share their personal thoughts and stories which may involve many topics resulting in multiple misaligned posts. For instance, in Figures 1e and 1f, we highlight two daily experiences from *Stimeyland* blog where the mother details the first visit of her autistic child, Jack, to a doctor (Figure 1e) and a visit to stadium to watch Jack playing hockey (Figure 1f). The one shown in Figure 1f is reporting a usual routine activity of the child instead of mentioning any ASD related experience (a case of misalignment). Moreover, the *Stimeyland* blog consists of keywords such as "kid", "time", and "day" (as shown in Table 4) which indicates that the author uses his blog to write about diverse topics such as his daily activities.

4 METHODS

In this section, we highlight the methodology we have adopted to quantitatively assess the performance of different feature extraction methods and classifiers using the previously described datasets. Our machine learning pipeline that takes text documents $\{p_i\}_{i \in 1..K}$ (blog posts) as input and classifies each post into $\hat{y}_i = \{A, \bar{A}\}$, consists of two distinct phases: feature extraction and classification, as shown in Figure 2.

[5] https://radimrehurek.com/gensim/

Table 2: Statistics of the gathered *LiveJ* communities organized as ASD and non-ASD related

Category	Community	No. Posts	Mean Post Length	Top 5 keywords
ASD	speechpathology	3098	127	school, program, work, working, slp
	teaching-autism	113	156	work, autism, working, child, kid
	spectrum-parent	197	220	son, kid, autism, parent, thing
	special-parents	179	204	child, thing, son, school, special
	autistic-abuse	147	57	autistic, child, usa, abuse, http
	autism-spectrum	118	195	child, autism, article, son, thing
	aspient	49	237	asperger, people, time, talk, community
	aspie-trans	46	147	community, gender, thing, asperger, death
	asperger	289	181	people, autism, thing, autistic, feel, person
	ask-an-aspie	65	233	thing, school, people, son, friend
	asd-families	199	194	school, son, autism, child, brother
	add-adhd	352	178	medication, feel, thing, day, time
Non-ASD	bentolunch	1694	73	bentos, bento, lunch, carrot, cheese
	naturalbirth	571	189	birth, baby, week, midwife, hospital
	parenting101	375	199	time, kid, baby, day, night, child
	dyedhair	2212	91	hair, color, dye, blonde, bleach
	cat-lovers	719	161	cat, vet, time, food, kitty
	dog-lovers	286	154	dog, time, vet, day, food
	curlyhair	370	110	hair, product, curly, curl, curls
	trashy-eats	643	95	cheese, food, sauce, bacon, eat
	webdesign	482	116	page, site, web, website, design
	htmlhelp	7720	47	code, journal, entry, link, page

Table 3: Statistics of the gathered *RedIt* communities organized as ASD and non-ASD related

Category	Community	No. Posts	Mean Post Length	Top 5 Keywords
ASD	autistic	468	173	autistic, people, autistics, autism, thing
	autism	646	170	autism, thing, feel, people, autistic
	aspergirls	924	212	feel, thing, people, time, feeling
	aspergers	904	169	people, time, feel, thing, social
	anxiety	944	187	anxiety, feel, feeling, time, day
	asd	64	229	autism, people, research, time, spectrum
	adhd_anxiety	409	216	anxiety, feel, feeling, day, work
	adhd	924	216	time, feel, work, adhd, thing
Non-ASD	cats	73	67	cat, vet, day, time, kitten
	dogs	764	242	dog, time, day, puppy, training
	fashionreps	367	94	https, haul, ship, size, rep
	fashionsouls	189	61	knight, armor, fashion, weapon, gauntlet
	food_pantry	783	174	food, list, week, work, month
	hair	557	112	hair, color, shampoo, dye, bleach
	html	923	87	html, page, code, https, image
	parenting	891	234	time, kid, thing, school, parent
	pets	410	165	cat, dog, pet, vet, food
	webdev	570	120	site, work, web, page, development

Table 4: Statistics of the eight gathered *PersB* blogs

Blog Title	No. Posts	Mean Post Length	Top 5 keywords
Adventure in Autism	1246	704	autism, vaccine, child, vaccination, vaccinate
Autism From A Father's Point Of View	478	655	child, autism, people, time, thing
Autism Day By Day	773	488	autism, child, nicky, kid, time
Crazy Girl In An Aspie World	286	1011	time, people, thing, aspie, life
Emma's Hope Book	936	545	emma, emmas, thing, child, people
Faith, Hope, And Love With Autism	285	449	philip, people, day, life, autistic
Mom-NOS	736	429	bud, time, day, child, thing
Stimeyland	1547	440	kid, time, jack, day, thing

4.1 Feature Extraction

The feature extraction phase consists of two steps: preprocessing and transformation. In the preprocessing step, we prepare the input posts that need to be converted into feature vectors through (i) cleaning to eliminate non-ASCII characters, extra white spaces, and urls from the posts; (ii) tokenization to transform words into lowercase tokens; and (iii) removing stop words, such as "a", "the", "of", "what", and "because". This step transforms a given original post p_i into a pre-processed post p_i'.

In the transformation step, we convert the pre-processed post p_i' to a numerical vector called feature vector $X_{p_i} = [v_{i1}..v_{ij}..v_{iN}]$, where N is the total number of features and v_{ij} is the j^{th} feature of the i^{th} post p_i.

Borrowing from our related work in Section 2, we consider multiple feature extraction methods into three approaches: *frequency-based*, *context-based*, and *hybrid*.

4.1.1 Frequency-based Approach: From this approach, we choose to compare the following feature extraction methods in our experiments: Topics, LIWC, ANEW, Empath and Topics+LIWC+ANEW. ***Topics:*** To characterize text, topics were found to have good expressive power[31]. This method applies the LDA algorithm [7] to extract topics from posts to construct features. In this method, the j^{th} element of the feature vector X_{p_i} is the probability of topic j in post p_i. Following the commonly accepted criterion, we set the total number of features, $N = 10 \log |V|$, where V is the vocabulary size of the dataset [30]. As a result, for datasets *LiveJ* and *RedIt*, $N = 44$ and for *PersB* dataset $N = 47$.

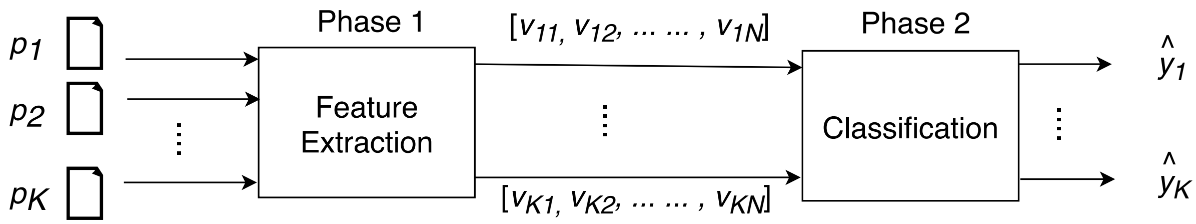


Figure 2: Our two phase machine learning methodology to classify a given post p_i into ASD (A) vs non-ASD ($\bar{A}$) related content respectively

LIWC: Linguistic features have been proven to be great in sentiment classification and depression detection[31, 39]. To capture the language style, LIWC [35] is a popular software package which classify words from text files into different linguistic and psychological categories using an internal dictionary. For this method, the j-th element of the feature vector X_{p_i} is the percentage of words belonging to category j in post p_i. Based on the number of categories in LIWC, we set $N = 64$ [35].

ANEW: Affective information is another important aspect to characterize text content. Researchers have found that pleasantness and activation are useful to conceptualize the emotion expressed in posts in online depression communities[29]. ANEW [8] is a lexicon of 1030 English words rated in three dimensions: valence, arousal, and dominance. It is often used to capture the conveyed affective information in the text. We generate the feature vector X_{p_i} according to the ANEW lexicon size $N = 1030$, where the j^{th} element of the vector corresponds to the frequency of j^{th} ANEW word in post p_i.

Topics + LIWC + ANEW: This method is a combination of Topics, LIWC, and ANEW [30]. To form the feature vector X_{p_i}, we concatenate the features from these methods resulting in $N = 1138$ for *LiveJ* and *RedIt* datasets and $N = 1141$ for *PersB* dataset.

Empath: Empath [13] is a text analysis tool similar to the LIWC software. But Empath has introduced more modern and useful categories like violence, tourism, social media, which don't exist in the lexicon of LIWC. We add a custom category *autism* in addition to the 193 built-in pre-validated categories[6] in the Empath tool which makes the number of features, $N = 194$ for all three datasets. Thus, the j-th element of the feature vector, X_{p_i} corresponds to the raw counts of words belonging to j-th category, normalized over all words in the post p_i.

4.1.2 Context-based Approach: We choose Doc2Vec as a sample feature extraction method from this approach.

Doc2Vec: We apply the Gensim [36] implementation of *PV-DBOW* model to learn the feature vectors of posts [23]. We empirically set the hyper-parameters: window to five and the number of epochs to ten, for all three datasets. Also, we choose $N = 50$ for *LiveJ* and *RedIt*, and $N = 100$ for *PersB* proportional to the Words/Posts in each dataset.

[6]List of categories and words in Empath: https://github.com/Ejhfast/empath-client

4.1.3 Hybrid Approach: We consider the following feature extraction methods from this approach in our experiments.

Word2Vec-Cluster: Researchers [2, 10] have used clustering of word vectors as features to improve the effectiveness of sentiment analysis. We assume that posts from each community of *LiveJ* and *RedIt* dataset and each blog of *PersB* dataset contains latent themes which can aid in the classification task. In order to capture these latent themes, we obtain feature vectors by clustering word vectors. These word vectors are learned from a dataset utilizing the Gensim [36] implementation of *CBOW* model [26]. We set N equal to the number of clusters and the j^{th} element of X_{p_i} is the normalized counts of words belonging to j^{th} cluster in post p_i. Since we have 22 and 18 communities in *LiveJ* and *RedIt* datasets respectively and 8 blogs in *PersB* dataset, in this method, we set $N = 22$ and $N = 18$ for *LiveJ* and *RedIt* datasets respectively, and $N = 8$ for *PersB* dataset.

TfIdf-Word2Vec: The word embeddings derived from word2vec method have the capability to capture linguistic regularities and patterns in text[23]. Gabrilovich *et al.*[17] has showed that meaning of text can be represented as a weighted vector of words. This is why, researchers have utilized both word2vec and TfIdf to produce features for text [47, 48]. This method constructs the feature vector as follows: $X_{p_i} = \sum_{w \in p_i} \text{TfIdf}(w, p_i) \times v_w$, where v_w is the word vector for word w learned from a dataset using the *CBOW* model. In our experiments, for all datasets we empirically set $N = 50$ which corresponds to the dimension parameter in this model.

4.2 Classification

In order to evaluate the impact of choosing different classifiers on the performance of a given feature extraction method, we test five different classifiers belonging to three main families: one non-probabilistic linear–Support Vector Machine (*SVM*) [46], one ensemble learning with bagging technique–Random Forest (*RF*) [9], and three ensemble learning with boosting technique–Ada boost (*ADB*) [15], Gradient Boost (*GDB*) [16], and XGBoost (*XGB*) [12]. All the classifiers were implemented using the scikit-learn[7] package [34] in Python.

All results shown in the next section are measured with the five classifiers mentioned above. We use the randomized grid search method to find the optimal hyper-parameters of the classifiers [43].

[7]http://scikit-learn.org/stable/

Also, we perform 10-fold stratified cross-validation for the *moderator-controlled* datasets (*LiveJ* and *RedIt*) [22]. In general, cross validation is mainly used to assess how well the model prediction will work on unseen and does avoid the problem of over-fitting. Since the *single-authored* dataset is small, we apply leave-one-out cross-validation for *PersB* during parameter tuning [22].

5 RESULTS

In this section, we describe the metrics used to evaluate the different feature extraction methods and classifiers presented in the previous section and highlight the best combination to use based on the dataset characteristic.

5.1 Evaluation Metrics

We evaluate the performance of different feature extraction methods over various classifiers, using four metrics: accuracy, precision, recall, and F1-score. We define accuracy, *Acc*, as the percentage of posts that are correctly classified as *ASD* (A) and *non-ASD* ($\bar{A}$) as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{3}$$

where TP and FP are the number of posts correctly classified and misclassified as A respectively, and TN and FN are the number of posts correctly classified and misclassified as $\bar{A}$ respectively.

Precision, Pr is the proportion of posts that has been correctly classified as A. Recall, Rc is the fraction of A posts predicted over the total number of A posts in the corpus. We have:

$$Pr = \frac{TP}{TP + FP}, \quad Rc = \frac{TP}{TP + FN}$$

To evaluate precision and recall in conjunction, we use the single measure $F1$-score, $F1$, which is the harmonic mean of precision and recall as shown in the following equation:

$$F1 = 2 \times \frac{Pr \times Rc}{Pr + Rc}. \tag{4}$$

5.2 Comparison of Feature Extraction Methods

In this subsection, we chose the well-known SVM classifier to compare the performance of different feature extraction methods. Table 5, presents the classification accuracies (*Acc*), precisions (Pr), recalls (Rc), and F1-scores ($F1$) using the FE methods grouped into frequency-based, context-based, and hybrid; using the two *moderator-controlled* datasets, *LiveJ* and *RedIt*, and the *single-authored* dataset *PersB*.

The hybrid feature extraction methods achieve the best accuracy, precision, and F1 scores. Within the hybrid approach, TfIdf-Word2Vec method outperforms for *moderator-controlled* datasets and Word2Vec-Cluster method outperforms for *single-authored* dataset. TfIdf-Word2Vec method achieves up to 91.23% accuracy, 91.45% precision, 92.35% recall, and 91.16% F1-score for *moderator-controlled* datasets (*LiveJ* and *RedIt*), and Word2Vec-Cluster obtains 82.0% accuracy, 87.5% precision and 82.94% F1-score for *single-authored* dataset (*PersB*). We believe that this is due to the combination of the discriminative mechanism used by TfIdf (as mentioned in Section 2) and the context-based method Word2Vec, which helps capturing the relation between words (semantics). This phenomenon is also verified with Word2Vec-Cluster that combines frequency and context-based methods.

Even though frequency-based feature approaches such as the Topics method gain high recall compared to hybrid approaches for all three datasets, the Topics method performs worse with regards to precision. For example, for *PersB* dataset SVM classifier achieves 100% recall value, but it obtains only 55.50% precision. That means, in addition to correctly classifying all ASD related posts, this method also misclassified many non-ASD related posts. Because frequency-based features only look at the word statistics (not context relation between words), these features are very sensitive to noise (out-of-context words) which may result in high misclassification occurrences.

From the results shown in Table 5, we found that indeed the combination of context and frequency-based approaches help improve the accuracy and F1-score performance by 34% and 45% respectively from the baseline (i.e., the performance gain of TfIdf-Word2Vec compared to Topics in *RedIt*), and is the best performing feature extraction approach.

5.3 Generalizing to Other Topics and Domains

Thus far, we have performed analysis using only one topic, *ASD*. Now, we investigate expanding our analysis to other topics, such as **food** or **web-design**.

We took 2123 posts of the *bentolunch* and *trashy-eats* communities from *LiveJ* datatset to represent the **food** domain. To build the negative samples for classifying **food** topic, we took 2181 posts from five other communities: *naturalbirth*, *cat-lovers*, *dog-lovers*, *curly-hair*, *webdesign* of the same dataset. To represent the **web-design** domain, we choose 1296 posts from *html* and *webdev* communities of *RedIt* dataset. To produce the negative samples for classifying **web-design** topic, we selected 1305 posts from *lunch*, *hair*, *pets*, *fashoinreps*, *fashoinsouls*, *parentingfails* communities.

Similar to our previous experiment, we use the SVM classifier and the same feature extraction methods using both domains **food** and **web-design** from *LiveJ* and *RedIt*, datasets respectively.

From the results summarized in Table 6, we find that the hybrid feature extraction methods achieve the best accuracy, precision, recall and F1 scores. TfIdf-Word2Vec method performs 97.84% accuracy, 98.25% precision, 97.36% recall, 97.84% F1-score when classifying *food* against all other subjects. While TfIdf feature extraction method uses mechanism to reduce the noise of common words in a corpus as mentioned in Section 2, it is oblivious of capturing the inter-relation between words. However, combining TfIdf with semantic-based method Word2Vec helps capturing the missing semantics to achieve the aforementioned high performance. While Empath method achieves 92.75% accuracy, it has lower recall 89.73% compared to the hybrid feature types.

Similar findings are also verified for the *web* topic domain as shown in Table 6. For example, TfIdf-Word2Vec method attained 97.38% accuracy, 97.90% precision, 96.92% recall and 97.38% F1 scores. We also observe that TfIdf-Word2Vec method boosts the accuracy and F1-score performance by 46.21% and 59.11% respectively

Table 5: Performance results of all feature extraction methods on *ASD* domain; where *Acc*, *Pr*, *Rc* and *F*1 are the best average accuracy, precision, recall, and F1-scores respectively, using the SVM classifier

Feature Type	Feature Approach*	*LiveJ*				*RedIt*				*PersB*			
		Acc	*Pr*	*Rc*	*F1*	*Acc*	*Pr*	*Rc*	*F1*	*Acc*	*Pr*	*Rc*	*F1*
Frequency	Topics	66.06	60.89	95.0	62.10	53.38	51.84	98.02	41.59	55.50	55.50	100.0	71.38
	LIWC	84.42	85.54	84.81	84.23	80.50	80.97	81.97	80.21	71.50	73.28	76.58	74.89
	ANEW	78.08	74.39	88.97	77.49	78.06	82.99	71.98	77.87	67.25	71.56	68.02	69.75
	Topics+LIWC+ANEW	86.05	86.05	87.91	85.89	83.24	83.51	84.82	82.98	73.25	73.86	80.18	76.89
	Empath	85.98	87.94	86.03	85.69	83.06	84.73	83.31	82.73	55.50	55.50	100.0	71.38
Context	Doc2Vec	83.52	83.77	86.23	82.48	83.01	86.02	84.26	81.82	80.75	83.11	81.98	82.54
Hybrid	Word2Vec-Cluster	89.51	91.13	88.82	89.40	85.60	87.64	84.13	85.51	**82.00**	87.50	78.83	**82.94**
	TfIdf-Word2Vec	**91.23**	91.45	92.35	**91.16**	**87.27**	89.43	87.34	**86.94**	79.50	83.98	77.93	80.84

* All methods are cited in sub-section 4.1

Table 6: Performance results of all feature extraction methods on *food* and *web* domains; where *Acc*, and *F*1 are the best average accuracy, and F1-scores respectively, using the SVM classifier

Feature Type	Feature Approach	*LiveJ*		*RedIt*	
		Acc	*F1*	*Acc*	*F1*
Frequency	Topics	66.53	65.27	51.17	38.27
	Empath	92.75	92.73	92.99	92.98
Context	Doc2Vec	86.35	85.89	78.53	77.59
Hybrid	Word2Vec-Cluster	95.21	95.21	93.88	93.86
	TfIdf-Word2Vec	**97.84**	**97.84**	**97.38**	**97.38**

from the baseline method Topics in *RedIt*. The result of these experiments substantiate that combination of context and frequency-based approaches are the best binary classification method for topic alignment in general as verifies using two blog sources and three different topic domains.

5.4 Impact of Classifiers on Performance

We have previously showed the performance of multiple FE methods using only one classifier. In this subsection, we discuss the performance of the FE methods with different classifiers (*RF*, *ADB*, *GDB*, *XGB*, and *SVM*) and highlight the gain of various combinations of classifiers and FE methods using our datasets. In Table 7, we present the results given by the classifier that provides the best accuracy (*Acc*) when combined with a given FE method. We represent the accuracy in the following format $[R]^{[C]}$, where C is the classifier that achieves the best accuracy R.

The results indicate that the choice of a classifier can have a major effect on the performance, achieving up to 30% accuracy improvement. For example, for the *PersB* dataset, Empath when combined with *SVM* achieves only 55.50% accuracy, but when combined with *RF* it gets highest 85.25% accuracy and 86.74% F-1 score for *PersB* dataset. Though Empath did not gain highest precision compared to 87.50% achieved by Word2Vec-Cluster, it obtained highest recall 86.94% for *PersB* dataset. We believe that this is due to *RF* using only a subset of the total number of features to exclude the non-discriminant ones in the classification task; these non-discriminant features may occur often in datasets with misaligned content such as *PersB*. However, *SVM*, as a result of always using the full set of features (including non-discriminant ones), generates many misclassified posts which justifies the high recall (100%) and high FP numbers as in Table 5.

We also find that no single combination of feature extraction method and classifier works the best for all datasets. Similar to the performance achieved using the *SVM* classifier (Table 5), the *hybrid* methods outperformed other methods using different classifiers for *moderator-controlled* datasets. For example, the best accuracies 92.80% and 89.33%, the best precision 92.41% and 90.48%, and the best F1-scores 92.75% and 89.01% for *LiveJ* and *RedIt* datasets respectively, are achieved by TfIdf-Word2Vec using the *XGB* classifier, which represents a slight improvement over using the *SVM* classifier (no more than 2.07% for accuracy and F1-score). However, unlike what we showed in Table 5, for *PersB*, the *frequency-based* Empath feature method achieves the best accuracy (85.25%) and F1 score (86.74%) among all feature methods.

We conjecture that the *single-authored* (*PersB*) dataset has lengthy posts (words per post), which will make the distribution of features smoother compared to the distribution in non-lengthy posts in *moderator-controlled* dataset. When discriminative features are not well represented in short posts, then that results in misclassifications.

6 CONCLUSION

This paper has presented a quantitative and qualitative comparison study of multiple text mining feature extraction approaches combined with different classifier families. We have adopted a data-driven approach to compare the classification accuracy and F1-score using *moderator-controlled* versus *single-authored* datasets which have major differences in post length and content alignment. We have used ASD related blogs for both classes to perform the analysis. We have also demonstrated that those feature extraction approaches can be used for classifying other topics such as food or web-design as well. The experimental results revealed that the frequency-based feature extraction approach using Random Forest classifier performs the best for *single-authored* dataset, while the hybrid feature extraction approach combined with the XGBoost classifier outperforms all other combinations for classification in

Table 7: Performance results of all feature extraction methods. Here, Accuracy $[R]^{[C]}$ stands for the best average accuracy, R from the best classifier, C.

Feature Type	Feature Approach*	*LiveJ*				*RedIt*				*PersB*			
		Acc	*Pr*	*Rc*	*F1*	*Acc*	*Pr*	*Rc*	*F1*	*Acc*	*Pr*	*Rc*	*F1*
Frequency	Topics	77.44^{XGB}	76.91	80.78	77.04	65.81^{RF}	65.83	70.18	65.29	65.75^{ADB}	64.41	85.59	73.50
	LIWC	84.42^{SVM}	85.54	84.81	84.23	80.50^{SVM}	80.97	81.97	80.21	75.00^{RF}	74.40	83.78	78.81
	ANEW	78.12^{RF}	75.06	86.64	77.76	78.06^{SVM}	82.99	71.98	77.87	74.75^{RF}	74.69	82.43	78.37
	Topics+LIWC+ANEW	86.81^{XGB}	87.15	87.78	86.67	83.24^{SVM}	83.51	84.82	82.98	73.75^{XGB}	75.11	78.83	76.92
	Empath	88.19^{XGB}	88.94	89.36	88.03	85.14^{XGB}	85.86	86.47	84.85	$\mathbf{85.25}^{RF}$	86.55	86.94	**86.74**
Context	Doc2Vec	85.01^{RF}	83.13	91.67	83.82	83.70^{XGB}	84.20	87.72	82.54	80.75^{SVM}	83.11	81.98	82.54
Hybrid	Word2Vec-Cluster	89.86^{XGB}	90.17	90.72	89.78	85.53^{XGB}	86.10	86.38	85.39	82.00^{SVM}	87.50	78.83	82.94
	TfIdf-Word2Vec	$\mathbf{92.80}^{XGB}$	92.41	94.29	**92.75**	$\mathbf{89.33}^{XGB}$	90.48	90.46	**89.01**	82.25^{GDB}	84.79	82.88	83.83

* All methods are cited in sub-section 4.1

moderator-controlled datasets. Moreover, the choice of the classifier has a major impact on the performance of different feature extraction methods, and can achieve up to 30% performance gain.

Due to the lengthy process of diagnosing Autism Spectrum Disorder (ASD) and the lack of specialized clinics, online social media is becoming an important source for advice and support for people with ASD and their families. While treating children with ASD in the clinics we have experienced that parents/families of children living with ASD often seek not only medical and communication experts, but also a community where people are going through similar experiences. Meeting other families, virtually or face-to-face, often gives them the inspiration to deal with the unique challenges of parenting a child with ASD. Clinicians also direct the families to online social blogs that intends to connect, educate, inspire and empower anyone that deals with ASD. Despite a preponderance of information, there is no way to weed out useful posts from noisy, ineffective ones (particularly in the *single-authored* blogs). We attempt to address this by studying the classification problem per post (ASD/non-ASD) and assessing the viability of the different feature extraction approaches in the ASD context.

Our future work includes: (i) extending the binary classification of the text document (blog post) to a multi-class classification (i.e, a sub-classification of the ASD class into symptoms, treatments, awareness, and support); (ii) experiment with more complex feature extraction methods and classifiers; and (iii) building prediction model, which takes properties of text document corpus to estimate the optimal feature extraction method and classifier candidates for the predefined settings.

7 ACKNOWLEDGEMENT

Research supported by US NSF awards #1800088; #1719342; #1345232, #1914635, EPSCoR Cooperative agreement OIA-1757207. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the federal government.

REFERENCES

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919* (2017).

[2] Eissa M Alshari, Azreen Azman, Shyamala Doraisamy, Norwati Mustapha, and Mustafa Alkeshr. 2017. Improvement of Sentiment Analysis Based on Clustering of Word2Vec Features. In *Database and Expert Systems Applications (DEXA), 2017 28th International Workshop on*. IEEE, 123–126.

[3] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[4] Adham Beykikhoshk, Ognjen Arandjelović, Dinh Phung, and Svetha Venkatesh. 2015. Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, 1354–1361.

[5] Adham Beykikhoshk, Ognjen Arandjelović, Dinh Phung, Svetha Venkatesh, and Terry Caelli. 2015. Using Twitter to learn about the autism community. *Social Network Analysis and Mining* 5, 1 (2015), 22.

[6] Mekkin Bjarnadottir. 2014. Why text analytics is so important in search. https://www.techradar.com/news/world-of-tech/management/why-text-analytics-is-so-important-in-search-1247983

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[8] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Citeseer.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[10] Andrei M Butnaru and Radu Tudor Ionescu. 2017. From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings. *Procedia Computer Science* 112 (2017), 1783–1792.

[11] Serhiy Bykh and Detmar Meurers. 2014. Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 1962–1973.

[12] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[13] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.

[14] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[15] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[16] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[17] Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis.. In *IJcAI*, Vol. 7. 1606–1611.

[18] Thorsten Joachims. 1996. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Technical Report. Carnegie-mellon univ pittsburgh pa dept of computer science.

[19] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.

[20] Vineet John. 2017. A Survey of Neural Network Techniques for Feature Extraction from Text. *arXiv preprint arXiv:1704.08531* (2017).

[21] Edilson Anselmo Corrźa Jśnior, Vanessa Queiroz Marinho, and Leandro Borges dos Santos. 2017. NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 611–615.

[22] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.

[23] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[24] Haixia Liu. 2017. Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177* (2017).

[25] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[28] A Taylor Newton, Adam DI 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*. ACM, 463–466.

[29] Thin Nguyen, Thi Duong, Dinh Phung, and Svetha Venkatesh. 2014. Affective, linguistic and topic patterns in online autism communities. In *International Conference on Web Information Systems Engineering*. Springer, 474–488.

[30] Thin Nguyen, Thi Duong, Svetha Venkatesh, and Dinh Phung. 2015. Autism blogs: Expressed emotion, language styles and concerns in personal and community settings. *IEEE Transactions on Affective Computing* 6, 3 (2015), 312–323.

[31] Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2011. Prediction of age, sentiment, and connectivity from social media text. In *International Conference on Web Information Systems Engineering*. Springer, 227–240.

[32] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing* 5, 3 (2014), 217–226.

[33] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 241–257.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[35] James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. LIWC2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net* (2007).

[36] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[37] Siti Hajar Mohd Roffeei, Noorhidawati Abdullah, and Siti Khairatul Razifah Basar. 2015. Seeking social support on Facebook for children with Autism Spectrum Disorders (ASDs). *International journal of medical informatics* 84, 5 (2015), 375–385.

[38] B. Romero and M. Choudhury. 2006. Social media use in families with autism spectrum disorders. In *American Speech-Language-Hearing Association (ASHA) Annual Convention, Philadelphia, PA*.

[39] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 8 (2004), 1121–1133.

[40] Amit Saha and Nitin Agarwal. 2015. Demonstrating social support from autism bloggers community on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, 1053–1056.

[41] Amit Saha and Nitin Agarwal. 2015. Insight into Social Support of Autism Blogger Community in Microblogging Platform. In *2015 AAAI Spring Symposium Series*.

[42] Amit Saha and Nitin Agarwal. 2016. Emotional Resiliency of Families Dealing with Autism in Social Media. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS-Science and Technology Publications, Lda, 377–382.

[43] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.

[44] Statistics-CDC 2018. Data and Statistics| ASD| CDC. Retrieved April 10, 2018 from https://www.cdc.gov/ncbddd/autism/data.html

[45] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.

[46] Vladimir Vapnik. 1998. *Statistical learning theory. 1998*. Wiley, New York.

[47] Zhi-Tong Yang and Jun Zheng. 2016. Research on Chinese text classification based on Word2vec. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 1166–1170.

[48] Wei Zhu, Wei Zhang, Guo-Zheng Li, Chong He, and Lei Zhang. 2016. A study of damp-heat syndrome classification using Word2vec and TF-IDF. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 1415–1420.