Int J Data Sci Anal (2017) 4:209–231
DOI 10.1007/s41060-017-0073-y

REGULAR PAPER

Latent sentiment topic modelling and nonparametric discovery of online mental health-related communities

Bo Dao[1] · **Thin Nguyen**[1] · **Svetha Venkatesh**[1] · **Dinh Phung**[1]

Received: 10 November 2015 / Accepted: 30 August 2017 / Published online: 30 September 2017

Abstract Social media are an online means of interaction among individuals. People are increasingly using social media, especially online communities, to discuss health concerns and seek support. Understanding topics, sentiment, and structures of these communities informs important aspects of health-related conditions. There has been growing research interest in analysing online mental health communities; however, analysis of these communities with health concerns has been limited. This paper investigates and identifies latent meta-groups of online communities with and without mental health-related conditions including depression and autism. Large datasets from online communities were crawled. We analyse sentiment-based, psycholinguistics-based and topic-based features from blog posts made by members of these online communities. The work focuses on using nonparametric methods to infer latent topics automatically from the corpus of affective words in the blog posts. The visualization of the discovered meta-communities in their use of latent topics shows a difference between the groups. This presents evidence of the emotion-bearing difference in online mental health-related communities, suggesting a possible angle for support and intervention. The methodology might offer potential machine learning techniques for research and practice in psychiatry.

Keywords Nonparametric discovery · Latent topics · Mood and emotions · Mental health · Online communities

This paper is an extension of our earlier paper published in IEEE International Conference on Data Science and Advanced Analytics (DSAA'15) [10].

✉ Bo Dao
bo.dao@deakin.edu.au

[1] Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, VIC 3216, Australia

1 Introduction

The advantage of social media has been improving the quality of healthcare. Social media sites such as Facebook, Twitter, and Tumblr have become increasingly recognized as a promising platform for healthcare support and intervention [6,7,23]. People have been moving away from their traditional communication, now meeting up online and using social media tools to make a different communication model. Online communities have been built up as forums for individuals to share information and advice in a variety of their daily life, especially their health beings [38]. With the popularity of social networks, social media offers a low-cost sensing channel to analyse health behaviours of individuals and communities through their postings.

Mental healthcare is another area that gradually benefits from social media. By accessing social media sites (e.g. patientslikeme.com), individuals with depressive symptoms can make connection with others for sharing their experiences, finding answers to health questions, and expressing themselves with many kinds of data. In a social study [50], up to 41% of participants with depression answered that online communities help them to reconnect individuals and overcome their depressive states. Several latent patterns and factors within online depression-related communities dictate this process, including topics of discussion, posting behaviour, demographic information, relationship, interaction, and emotions [21,49,50]. Furthermore, the core symptoms of depression, such as social withdrawal and sleep disturbance, are also characteristics of autism spectrum disorders (ASD) which are associated with the most psychiatric disorder as depression. Some symptoms of ASD such as obsessionality and self-injury may be developed during an episode of depression [19,63]. Online autism communities have been studies in [46,48]. The authors investigated the

topic patterns, language styles, and affective information in these communities in comparison with general online communities. In these conditions, social media can be seen a life saver.

The content of blog posts in social media sites is more subjective than others. It is suitable for a sentiment analysis on both individuals and community contexts. For sentiment analysis, features are either manually textual or predefined terms where words or phrases are mapped to sentiment-bearing scores. Several related works in feature selection for sentiment analysis and opinion mining can be seen in studies [52,53]. In addition, other works examined in features representing user's mobile behaviours and environment to perform the demographic and location prediction [26,27,68]. However, profiling at multiple scales, from an individual to their joined community, is crucial within and across online communities. It needs the correct interpretation of features in context. Social media research has investigated global patterns of behaviours from large-scale data instead of seeking user-centric patterns [42]. It requires to build tools for inferring and utilizing both the complex and dynamic relationships between online users from individuals to their community perspectives. While questionnaire-based methods continue to dictate the research by social scientists [21,22,49], data-driven methods for online communities with mental health-related concerns are still in early stages. Understanding risk factors and latent patterns in these online mental health-related communities is an important step in many aspects of coping with mental health issues, ranging from singling out individual aspect for support (e.g. relationship between long working hour and depression) to informing preventative healthcare policy (e.g. by looking at clustering of communities spatially).

This study aims to examine patterns and formations of online mental health-related communities including depression, autism, and general online communities. Using data crawled from LiveJournal[1], we apply a Bayesian nonparametric in topic modelling to automatically infer latent topics of interest in using discrete mood tags, generic words, language styles, and the set of affective information among the communities. Moreover, mental health is intrinsically linked to emotion; hence, our next focus is to investigate the varying sentiment-based representations for communities. In addition to exploratory analysis, we demonstrate the usefulness of latent topics by further clustering these communities into meta-groups. Visualization of our clustering results indicate that these meta-groups of communities can be well separated when projecting on 2D spaces. This demonstrates the evidence of emotion-bearing differentiation in online mental health-related communities, suggesting a possible angle for support and intervention.

[1] www.livejournal.com.

The preliminary version of this work has been presented in [10]. In this work, we extended substantially on the aspect of topic models, especially hierarchical Dirichlet process. We provide comparison study to include the result for meta-communities clustering based on psycholinguistics-based community representation. In addition, we describe breakdown statistics of dataset cohorts, then presenting further analysis on sentiment sharing between autism and depression cohorts in meta-communities discovered by our algorithms. Furthermore, the difference in the use of topics of interest between these two groups is also investigated and presented in this paper.

Our contributions of this work are: (1) a novel problem on analysing both generic-based and sentiment-based topics in online mental health-related communities including those related to depression and autism symptoms; (2) a feasible application of nonparametric methods in the tasks of exploratory analysis, avoiding the problem with parametric methods when the latent patterns (e.g. topics or clusters) are unknown. In particular, to cluster communities into meta-communities automatically (is also called as hyper-communities discovery), we use a method in the theory of graphical models based on the concept of message passing between data points to automatically discover the number of clusters; and (3) that we believe to be the first to differentiate the communities based on the latent patterns (topics) of a quite comprehensive set of features: mood tags, affective information, psycholinguistics, and generic words from the blog posts.

The remaining paper is organized as follows: Sect. 2 reviews existing works on social media, mental healthcare and online community discovery. In Sect. 3, we present the description of datasets and its features, community representations, and clustering approaches for the nonparametric discovery of online communities. The experiments, analysis, and discussion on the results are provided in Sect. 4. Finally, we conclude the paper with several closing remarks and future directions in Sect. 5.

2 Related work

2.1 Social media and health

Social media are becoming "sensors" that capture and reflect user-egocentric thoughts and feelings about any happenings in their daily life. Social media platforms such as Twitter, Facebook, and Tumblr have provided a huge source of information available for studies on social media, specially healthcare [28]. It revolutionized the approaches researchers acquire their data and also the ways in which those social media data can be carried on practical use. As many of these online platforms increasingly gained more attraction from

individuals, after years, a large group of users with various records of life events is built up.

An increasing body of work has been interested in exploiting how social media can be used to infer the people health behaviours. In psychology, health behaviours shape the health and well-being of individuals, communities, and populations [62]. Moreover, [24] showed that social media can reshape healthcare in several ways (i.e. the way doctors and patients interact). By detecting changes of user-centric behaviours in social media, mental or behavioural health concerns are indicated. It is also revealed that social media data can be analysed to assess the role of sentiment, emotions, or mental status of a community [3] as well as to identify most disease-related conditions and symptoms [57,58] or mental health issues (e.g. depression or suicide) [13,14,54]. In addition, whenever posts have been made by someone, these posts can be immediately analysed to identify whether that person is an "at-risk" one in mental health or not. Then mental health support and interventions are considered to deliver self-help or proactive interventions for reducing the risk [6,7,23]. Furthermore, some work [54–56] found apparent evidences that people are progressively spending amounts of their online time to post messages about their health beings (e.g. depression) with treatment on virtual social networks such as Twitter, Facebook, and LiveJournal. Our work focus on this promising new trend of research.

2.2 Impacts of online mental health-related communities

To understand aspects of online mental health-related communities including online depression and autism communities, several research has been done for identifying characteristics of these communities [29,44,46,47,49–51]. With questionnaire-based methods, existing studies (e.g. Nimrod [49–51]) focused on investigating the content and characteristics of the discussions in online depression communities. Their findings shown that online depression communities can serve as a promising platform for exchanging the experience of living with mental health conditions (i.e. depression). These sharing experience might enhance a better understanding of people with mental health-related concerns and also encourage them during struggling over their medical conditions for a better improvement. In a related work, by investigating both language styles and topics expressed in the content of blog posts in online autism communities, the study [46] indicated that substantial differences between autism and general online communities are significantly characterized by both latent topics of discussion and psycholinguistic features. In addition, existing studies [41,44,60] analysed several informative features of blog posts including topical content, linguistic styles, and sentiment conveyance (mood) for examining online social capital and mood in two extremes (high and low) to identify distinct communities among different social capital groups of both depression and general communities. The study suggests that mining blogs have the potential to detect clinical information from online communities. It means that social media can be used as a barometer of mood in monitoring and detecting mental well-beings in online communities [39,54,60]. Echoing these above studies, we further substantiate the sentiment analysis of online mental health-related communities to support the idea that social media can be seen as an effective platform for future intervention framework. Hence, using machine learning approaches, negative thoughts expressed in posts can be detected, bringing timely help and support to those mental health-related communities.

2.3 Applied machine learning for community discovery

Community detection can be seen as an important activity for analysing social media networks in several domains. The community detection helps us to understand the formation and function of each community in the whole network. Many different approaches based on graph (or link) structures to modularity-based or model-based methods have been investigated in community detection [15,67]. Specifically, for hyper-groups discovery of online communities, link structure approaches (e.g. friendship and community membership) have been investigated in [20,30]. Furthermore, several studies [5] analysed peer interactions in online healthcare forums with the measures of the quality and homophily of discovered communities. Their findings suggest that by observing interactions on healthcare communities we can discover meaningful sub-communities among them. However, the link structures are not strong available overtime due to the dynamic nature of social media. Other alternative approaches to community discovery focused on the content messages made by members in the community to characterize online communities. These communities were represented by topical, sentiment-based, psycholinguistics-based features [43] and tagged media (e.g. in Flickr groups) [37]. The findings indicate that sentiment-based hyper-community discovery on general online communities has potential implications in mental health-related research (e.g. support or surveillance on communities with negative sentiment).

Moreover, sentiment patterns such as moods and affective words usage in blog posts have been investigated in some studies for identifying characteristics of online communities and its members [3,11,34,39,45,60]. Indeed, mood is a popular and strong form of sentiment expression, conveying a emotional state of one's mind such as being happy, sad, or anger. While social media texts are rich source of sentiment, mood sensing from these texts is an important role in monitoring and detecting mental well-being in online communities. For example, Choudhury et. al [13,14] examined emotional states of individuals in social media

to predict the levels of depression in populations. Mood used to identify mental health-related issues of communities [9,14], or to gain sentiment patterns of how people interact with others [39], or to characterize users and communities for detecting hyper-communities [40,45]. Their findings indicate that sentiment-based clustering for community discovery is potential to be explored. Any discovered sentiment-based hyper-group of online communities can be seen as social indicators for mental health, aiming support or surveillance to these meta-communities. However, none has comprehensively considered the problem of using community clustering based on sentiment.

For learning latent topics from the content of posted messages, probabilistic topic modelling approaches (e.g. probabilistic latent semantic indexing (pLSI) [25], latent Dirichlet allocation (LDA) [2], or hierarchical Dirichlet processes (HDP) [65]) have shown to be effective in discovering latent topics from the corpus of blog posts. Several studies [35,41,43,44,58,66] used the standard parametric model LDA to learn latent topics from the content of blogs and tweets in the blogosphere for their research on mental health signals in social media. Using LDA to gain latent topics, [41] found significant differences among study cohorts which are characterized by the latent topics of discussion, psycholinguistic features, and tagged moods. Many studies (e.g. [9,47]) investigated the impact of topics and language styles among users in different cohorts defined by mood tags, social connectivity, and age from the online depression community. However, the topic modelling approaches face a critical issue in determining the key parameters (e.g. the number of topics in LDA). These parameters are not always available and quite difficult to specify in advance. We address the above parametric limitation by employing the Bayesian nonparametric topic modelling in discovering latent topics from the content of blog posts made by users in online communities with and without mental health-related states.

Language styles are effective features for identifying the style of individuals in their personal blogs. In psychology studies, the Linguistic Inquiry and Word Count (LIWC) package [59] is often used to capture language styles from the words people use in writing. It offers a wide scope of features ranging from linguistics, stylistics, social, affective, cognitive, perceptual, biological, relativity, personal concerns, to spoken features. These features were found to be good indicators of depression and mental health [12,14,29,47]. Based on LIWC features (also called psycholinguistic features), existing studies on community detection (e.g. [43,45]) have investigated the language styles of individuals expressed in their blog posts on general online communities. Their findings on the LIWC-based representation suggest that LIWC features are worthy follow-up for community representation when these psycholinguistic features are cheap to obtain on blog posts.

For clustering tasks, both parametric (e.g. K-means) and nonparametric clustering approaches (e.g. affinity propagation (AP) algorithm [18]) are widely applied. In clustering data, each data point from the similarity matrix is allowed to belong to a single clusters. Since the number of clusters is unknown in advance, the parametric methods are not often suitable. In addition, exemplars of clusters cannot be obtained during clustering by these approaches. Meanwhile, the nonparametric models can solve the parametric limitation when it allows an unbounded number of clusters. For example, the AP algorithm based on probabilistic graphical modelling discovers the unknown number latent clusters by passing local messages. During clustering, it also produces exemplars for each cluster as representatives for the hidden hyper-group of communities [43]. In this study, we apply the AP algorithm for detecting online meta-communities in dynamic nature of online settings.

3 Methodology

In this section, we briefly describe datasets used in our experiments in this paper. Next, we review the background of Bayesian nonparametric methods, especially the HDP model, for inferring latent patterns/topics from data corpus. We then introduce different community representations using a variety of feature sets extracted from the content of blog posts made within the community. Finally, we apply affinity propagation algorithm, a nonparametric clustering approach, to discover meta-communities, then presenting standard measures for evaluating the quality of clustering performance.

3.1 Datasets description

Data were crawled from the LiveJournal (LJ) blogging site. It is one of the world's most popular blogging sites with over 1.9 million active bloggers/users since 1999 [33]. LJ allows people to create personal blogs for maintaining their social interaction and exchanging thoughts, feelings and knowledge with others. In addition, this community publishing platform supports individuals with common interest to form the community along with their own personal blogs. Many online communities interested in health-related conditions have been created in LJ. Furthermore, LJ bloggers can either select a mood from a predefined list of 132 common moods given by LJ or enter a free text to label their posts at the time of writing. Thus, in conjunction with affective information expressed from the content of the post, the mood tags provide potential sources for sentiment analysis and for understanding affective aspects of mental health-related communities. Figure 1a shows a sample post containing a "happy" mood tagged while a "sad" mood tagged post is illustrated in Fig. 1b.

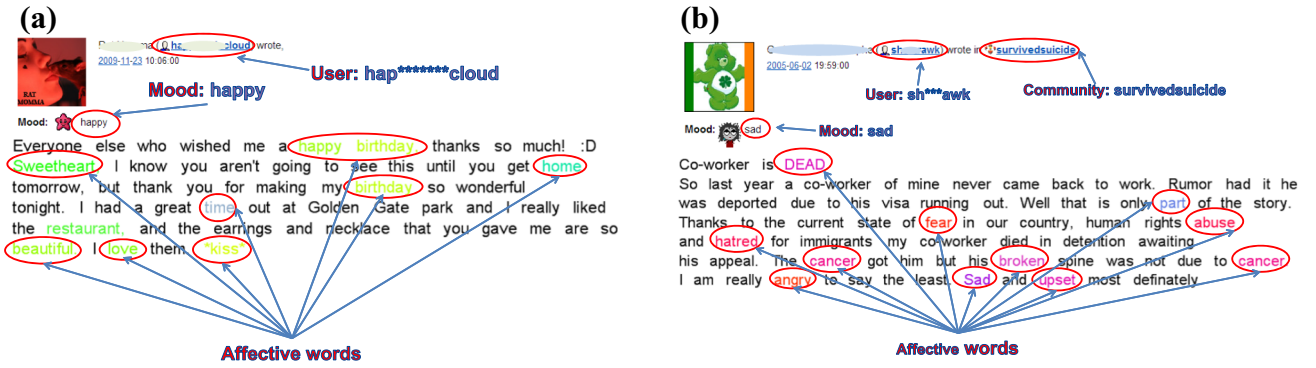

Fig. 1 (**a**) Example of a "happy" post. (**b**) Example of a post with "sad" mood. Examples of a blog post tagged either happy mood or sad mood in the LiveJournal community platform

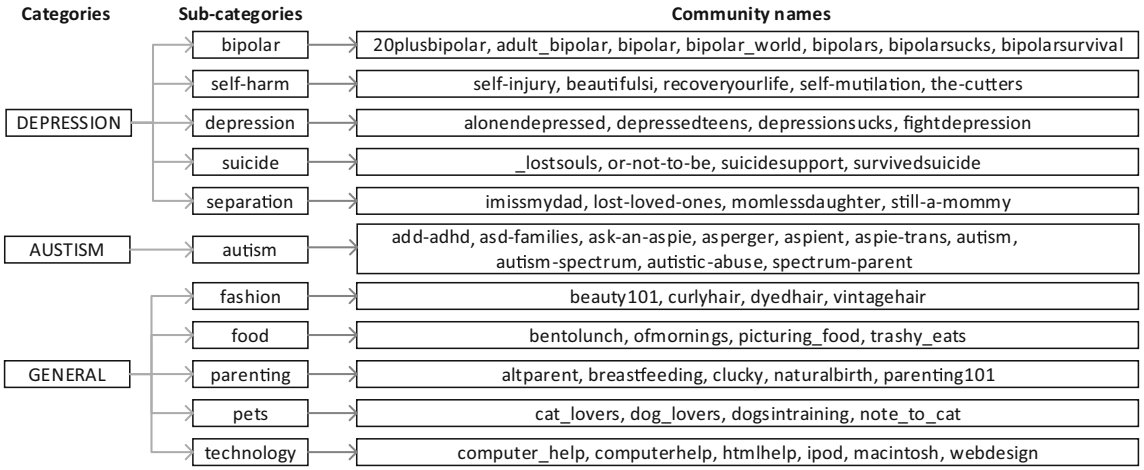

Fig. 2 Three study cohorts and its 11 LiveJournal sub-categories with selected communities used in the experiment

In this study, we aim to investigate online communities with and without mental health-related conditions. We identify a cohort of online communities who have self-description as in their profile as having any mental health-related concerns. Thus, large cohort data are further constructed for mental health-related disorders (e.g. depression and autism) as well as for general online communities who do not self-identify as having any mental-related conditions. As mentioned in Sect. 1, it is well known that people affected by autism or ASD is associated with depression. Therefore, in this study, we present an analysis on online communities with members affected by or interested in whether depression or autism (Fig. 2). A cloud visualization of moods tagged in the posts of these investigated online communities is further shown in Fig. 3a. As expected, both positive and negative valence moods are used to tag to the blog posts by individuals while blogging in online communities they have joined. The three study cohorts are described as follows:

Fig. 3 (**a**) A visualization for 132 predefined mood tags. (**b**) Tag cloud of top 150 ANEW words. Top mood tags and ANEW words used in the blog posts of communities

DEPRESSION Cohort through the functionality of "search communities by interest"[2] on LiveJournal, we searched 24 communities that interested in "depression" with at least 200 posts. Based on the name and self-description on the profile of each community, those 24 "depression" communities are categorized into one of the following five subgroups: *bipolar, depression, self-harm, suicide*, and *separation*. This cohort is considered as Depression Cohort with online depression-related communities. The briefly breakdown statistics of this group is shown in Table 1.

AUTISM Cohort similar to the selection of online communities for Depression cohort, we search for all communities interested in "autism", on LiveJournal, among the results of autism-related communities; ten online communities with largest number of posts were selected as Autism cohort for this study. A glimpse of the biography of these communities in the Autism cohort is given in Table 2.

GENERAL Cohort for diversity, we constructed a General cohort with 23 general online communities who were chosen from five common categories of communities in LiveJournal community directory[3] including *Fashion, Food, Parenting, Pets*, and *Technology* categories. In these groups, each community has at least 200 posts from their members. These selected general communities for each category are shown in Fig. 2.

Thus, the dataset for this study includes three groups (DEPRESSION, AUTISM, and GENERAL communities as can be seen at high level) and 11 subgroups (as can be seen at low level) of these 3 groups with a total of 57 online communities (24 DEPRESSION, 10 AUTISM, and 23 GENERAL communities) as shown in Fig. 2. Particularly, the study uses a large cohort of data from nearly 52,000 users with a total of 268,400 posts, including nearly 10,000 users with 38,400 posts from 24 DEPRESSION communities, nearly 2000 users with 10,000 posts in 10 AUTISM communities, and 40,000 users with 220,000 posts in 23 GENERAL online communities.

3.2 Feature extraction

In social media, the content of blog posts is more subjective than other forms of social media data, making it suitable for any analysis for both individuals and communities. We hence focus on mood, affective information, psycholinguistics, and latent topics derived from the content of blog posts as follows.

3.2.1 Mood-based features

We identify all the linguistic terms, called mood tags, tagged to the blog posts while posting. We focus on mood, a common form of sentiment, conveyed in the blog posts. Mood is a strong representation of the sentiment of someone who expresses an emotional state of feeling, such as being happy, sad, or angry. While LiveJournal provides a mechanism for people to tag a mood to their posts, this allows us to analyse the mood directly provided by users. A visualization, such as a tag cloud of these moods on our datasets, is shown in Fig. 3a. We assume that there exists a difference in the use of moods tagged to the blog posts by users in online communities. Such online communities can be grouped into a hyper-group of communities by their common mood usage.

3.2.2 Affective-based features

While mood tags may not always be available, other affective information can be extracted for any arbitrary text collection. In this study, we use the sentiment-bearing lexicon package, called Affective Norms for English Words (ANEW) [4], to measure affective information of social media as features for sentiment analysis. ANEW is a set of 1,304 sentiment-expressing English words created by the National Institute of Mental Health of the United States as a standard for research in cognition and emotion. Each ANEW word is rated in terms of three normalized well-known values: valence, arousal, and dominance. The valence and arousal values of words in the ANEW lexicon are on a scale of 1, *very unpleasant* and *least active*, to 9, *very pleasant* and *most active*. The valence value indicates the level of happiness, whereas the arousal score implies the degree of activation. The dominance values indicate the degree of control, ranging from 2.27 (helpless) to 7.88 (leader). To know more about this affective lexicon, a cloud visualization of ANEW words extracted from the study data corpus is shown in Fig. 3b.

3.2.3 Language style-based features

Another powerful feature set is psycholinguistics which people use as a language style in their writing during online communications. These psycholinguistics or language styles are extracted by the LIWC package which is a text analysis software used in psychology to measure emotion expression of writings. The LIWC package assigns English terms to one of the four high-level categories including *linguistic processes (*i.e. functional aspects of text*), psychological processes* (i.e. all social, emotional, cognitive, perceptual, biological processes and any references to time or space)*, personal concerns* (i.e. anything related to work, leisure, money and religion) and *spoken categories* (i.e. filler and agreement words) which are also further sub-categorized into a three-level hierarchy. According to several studies [59,64], with the massive social media corpora of millions of weblogs, such text corpora provide potential measurements of affective processes of psychological states, such as *positive* or

[2] http://www.livejournal.com/interests.bml.

[3] www.livejournal.com/browse/.

Table 1 Breakdown statistics of DEPRESSION cohort with its communities: #members, #posts, and self-description

Sub-category	Community	#Members	#Posts	Self-description of the community
Bipolar	20plusbipolar	334	1252	"This is a bipolar community and is for anyone who is bipolar so don't let the name mislead you all bipolars are welcome"
	Adult_bipolar	307	846	"This is an active community meant for ADULTS 21+ diagnosed with bipolar disorder"
	Bipolar	484	1536	"A place for people who are bipolar and have related symptoms to discuss their issues and how being bipolar affects them"
	Bipolar_world	351	994	"A community for members of the bipolarworld forums and others with bipolar disorder"
	Bipolars	574	2551	"This is a private community for those with or otherwise affected by bipolar disorder"
	Bipolarsucks	442	1687	"Welcome! We promote free speech and focus on bipolar and it's accessories"
	Bipolarsurvival	1133	6299	"Surviving day to day. Living with bipolar"
Self-Harm	Self-injury	796	2699	"This is a community journal for self-injurers and recovering self-injurers"
	Beautifulsi	310	765	"This is a community for people who find the beauty in bleeding and burning and harming themselves"
	Recoveryourlife	486	2075	"Recoveryourlife (formerly Ruinyourlife) is a non-judgmental self-harm support site"
	Self-mutilation	702	2665	"This is a community for people that want to stop self-harming"
	The-cutters	658	2407	"We support people who cut, burn, bruise, or otherwise intentionally injure themselves"
Depression	Alonendepressed	395	1057	"This community is for anybody with any type of mental illness. Anybody bashing anyone for how they feel will be banned"
	Depressedteens	318	735	"Welcome! This is a community for depressed teenagers"
	Depressionsucks	709	1590	"This is a support community for people with any type of depression"
	Fightdepression	530	1542	"This community is meant to help those with depression"
Suicide	_Lostsouls	341	1095	"This is the community for people who feel like they cannot go on"
	Or-not-to-be	86	227	"I made this community for people interested in talking and reading about suicide"
	Suicidesupport	335	607	"This community is a support forum for people who suffer from depression and suicidal thoughts as well as those who have lost someone to suicide"

Table 1 continued

Sub-category	Community	#Members	#Posts	Self-description of the community
	Survivedsuicide	129	336	"This community is for those who have lost someone to suicide"
Separation	Imissmydad	496	1658	"This is a community for anyone whose father has died to mourn and talk about their feelings with others who have been there"
	Lost-loved-ones	200	308	"Do not stand by my grave and weep I am not there"
	Momlessdaughter	714	3094	"I'm hoping to bring together any person, especially women, with other people who lost a parent at an early age"
	Still-a-mommy	139	376	"Still a mommy... to an angel"

Table 2 Breakdown statistics of AUTISM cohort with its communities: #members, #posts, and self-description

Sub-category	Community	#Members	#Posts	Self-description of the community
Autism	Asd-families	313	283	"This is a community for people who are related to someone with ASD"
	Ask-an-aspie	306	108	"Is there an aspie in your life? Do you need help understanding the autistic point of view"
	Asperger	1868	10, 693	"This is a support community for people with asperger syndrome"
	Aspie-trans	143	102	"Transgender, transsexual, gender queer and other divergent gender identities bring with it their own special challenges. Aspergers and Autism bring different challenges"
	Aspient	104	43	"A community for people involved intimately with someone who has apsperger's or high-functioning autism. A place to vent and share"
	Autism-spectrum	158	201	"This Journal is for anyone affected by autism or other spectrum disorders"
	Autism	1506	1863	"This is a community for anyone who has been affected by autism. Discussion is welcomed and may become heated. Insults and trolling are not tolerated"
	Autistic-abuse	47	162	"This journal is a place to read and discuss public cases of abuse against autistic persons. It is meant to raise awareness of how wide spread this problem is, and encourage discussion of abuse cases"
	Bsperger	25	37	"Tired of people blaming their aspergers"
	Spectrum-parent	326	290	"Being a parent is often challenging, but parenting kids on the austism spectrum (autism, aspergers, PDD-NOS, etc.) presents special issues"

negative emotions with a high rate of words. Therefore, the LIWC with a variety of lexicon for linguistics and psychological topics is improved, enhancing the use of social media corpora for sentiment analysis.

3.2.4 Generic word-based topic features

The interaction of topic and sentiment plays a crucial role in sentiment analysis. To extract topics in the content, probabilistic approaches including PLSA, LDA, and HDP can be used. In this study, we use Bayesian nonparametric (BNP) topic modelling to automatically infer latent topics of interest in a given corpus of texts. The model learns the probabilities p (vocabulary | topic) that are used to describe a topic and assigns a topic to each word in every document. Each post can then be represented as a mixture of topics using probability p (topic | document). We present more details of the nonparametric approaches in the following Sect. 3.3.

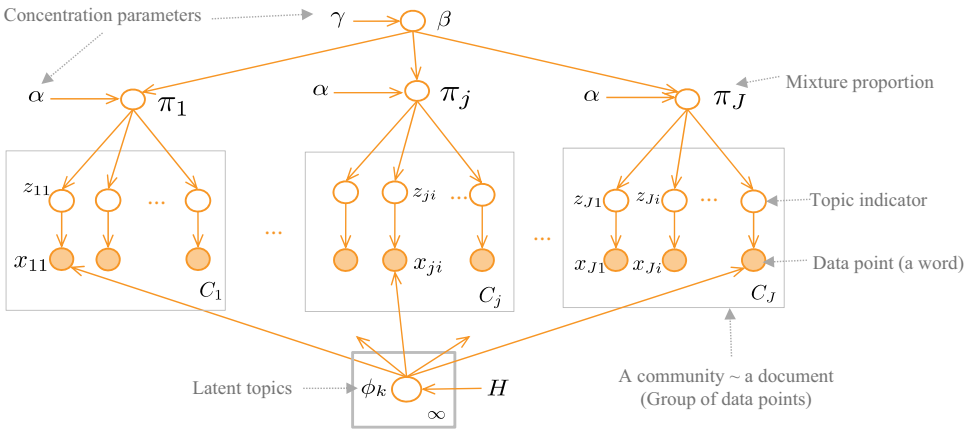

Fig. 4 Hierarchical Dirichlet processes for latent topics discovery in each community. Data points (e.g. words, moods) are shaded while unshaded nodes are latent variables. Each observed data point x_{ji} is assigned to a latent topic indicator variable z_{ji}; γ and α are the concentration parameters and H is the base measure. Each community as a group of data points has a separate mixture proportion vector π_j for each latent topic ϕ_k

3.3 Nonparametric topic modelling with hierarchical dirichlet processes

In this paper, we apply the hierarchical Dirichlet processes (HDP) model [65] to automatically infer the number K of latent topics from the corpus of blog posts among online communities in this study. The model uses a Dirichlet process (DP) [17] as the underlying nonparametric prior distribution for modelling of grouped data. HDP is a particularly attractive formalism when it posits the dependency among the group-level Dirichlet process mixture (DPM) models [1,36] by another DP.

A graphical model illustration for HDP using its stick-breaking representation is shown in Fig. 4. Specifically, let J be the number of groups of data points indexed as $j = 1, \ldots, J$ and $\{x_{j1}, \ldots, x_{jN_j}\}$ be N_j exchangeable observations associated with the group j. These observations are assumed to be exchangeable within the group j. Under HDP model, each group j is equipped with a random group-specific mixture distribution G_j which is statistically connected with other mixture distributions via another DP sharing the same base probability measure G_0:

$$G_j \mid \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0), \; for \; j = 1, \ldots, J$$

The base measure G_0 is also a random probability measure distributed according to another DP:

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

This generative process further suggests that G_j (s) are exchangeable at the group level which admits the following stick-breaking representation for HDP:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma), \quad \phi_k \overset{\text{iid}}{\sim} H, \quad k = 1, 2, \ldots$

When linked together by G_0, G_j as shown in [65], we have the following form:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \text{ where } \boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha, \boldsymbol{\beta})$$

From the definition of the HDP, it is clear that all random mixture component G_j's, G_0, and H share the same support Θ across groups. The local atom in the group j can be drawn as $\theta_{ji} \overset{\text{iid}}{\sim} G_j$, and the observation (e.g. data point) is generated as following process $x_{ji} \sim F(\theta_{ji})$.

Using a similar scheme described earlier for the DPM, inference in HDP can be carried out under a collapsed Gibbs sampling whose details can be found in [65]. For this inference task, we run Gibbs sampling for 1000 iterations with a burn-in period of 50 samples and the concentration parameters γ and α are also resampled by following a gamma distribution as described in [65]. Using our Gibbs inference codes in MATLAB and C#, HDP automatically yielded latent topics for mood tags, generic words, and affective words in the blog posts made in the communities. Then such communities based on different community representations are clustered into meta-communities (or hyper-groups of communities). In HDP, the data are organized as documents in which each document is a bag-of-words, generated from underlying latent topics. Following this, we treat each community as one document, each mood tag, affective, or generic word can be seen as one word in the model. In the following subsection, we present the different community representations based on each type of latent topic features which are automatically inferred by HDP from each corpus of "content" in the blog posts of the communities.

3.4 Community representation

Online social network has its own definition of an online community [8]. Since the visual representation of a social network can be more complex in a traditional graph model,

in the literature, this definition focuses on some explicitly or implicitly properties of the host network platform. Thus, online communities have been recognized in a variety of shapes and sizes. Each online community (e.g. a blogging community in LiveJournal) is defined by the scope of properties (i.e. topics) aiming to the host and among other things (i.e. members, interests, posts, and comments). It might have a number of interesting features. Given above definition of online community, in this paper, we aim to group similarity communities that somehow share the same or related common properties. This section presents different representations of online communities based on different interesting features for the meta-communities discovery.

3.4.1 Latent mood-based community representation

In this section, we introduce community representation based on latent mood-based topics called LMCR. Assuming that there is a difference in moods tagged to posts among online communities, such communities can be grouped into meta-communities by the similarity of their tagged moods. Indeed, LiveJournal provides a set of 132 predefined moods for their members to tag to their posts while blogging. From the data cohorts, a tag cloud visualization of these moods tagged in the blog posts is shown in Fig. 3a.

Denoted by $\mathscr{M} = \{happy, sad, \ldots\}$ the set of moods, where $|\mathscr{M}| = 132$ is the total number of moods predefined by LiveJournal. Let $\mathbf{m}_j = \{m_{j1}, m_{j2}, \ldots, m_{jn_j}\}$ be the set of mood tags in community j where n_j is the total number of posts made by all members of the community j, and $m_{ji} \in \mathscr{M}$ is a mood tagged to the blog post i in the community j. In the HDP model, for inferring latent mood-based topics, we treat each community as one document while each mood tag is considered as one word in the model. Thus, with J online communities, the corpus is built with J documents aggregated from all J communities $\mathscr{C} = \cup_{j=1}^{J} \mathbf{m}_j$. We run the HDP model over the corpus $\mathscr{C}$ with J documents to automatically learn K latent topics. Each community j can be represented by θ_j which is a K-dimensional vector, where the k^{th} element is a mixture proportion π_j of topic k for the community j. These mixtures are used to perform the mood-based community clustering. Moreover, the latent topics extracted from the corpus of mood tags shall be referred to as *mood topics* formally defined as follows.

Definition 1 Let $\mathscr{M}$ be a collection of 132 predefined mood tags defined by LiveJournal, and a mood topic is a discrete distribution over the set $\mathscr{M}$.

3.4.2 Mood usage-based community representation

This representation is called MUCR. Using the notation in Sect. 3.4.1, let $\mathbf{m}_j = \{m_{j1}, m_{j2}, \ldots, m_{j132}\}$ be the 132-dimensional mood usage vector as mood usage-based representation for the community j, where the element m_{jk} is the total number of times the mood $k^{th} \in \mathscr{M}$ was tagged within the community j. Each 132-dimensional mood usage vector $\mathbf{m}_j$ is normalized by the total number of blog posts with a mood tag in the community to unity, so that $\sum_{i=1}^{132} m_{ji} = 1$. For J communities, a mood usage matrix $T \in \mathbb{R}^{J \times 132}$ is used as input to perform the mood usage-based community clustering task.

When mood tags are unavailable or not used in any blog posts within the community, the sentiment-based community representations can be constructed by using emotion or affective information in the content of blog posts in the following subsections.

3.4.3 Latent ANEW-based community representation

In this representation called LACR, online communities can be represented by HDP latent topics on affective words in blog posts made by members of each community. We use the ANEW lexicon to identify affective words in blog posts made by members of each community.

In each community j, let $\mathbf{a}_j = \{a_{j1}, a_{j2}, \ldots, a_{jn_j}\}$ be the set of ANEW feature vectors where n_j is the total number of blog posts in this community and each element a_{ji} is a 1034-dimensional ANEW feature vector whose kth element is the number of times the kth ANEW word occurs in the content of the blog post ith of the community j. Each community is considered as one document while each element a_{jk} is as one word in the HDP model. Thus, with J communities, a matrix $C \in \mathbb{R}^{(J \times n_j) \times 1034}$ is used as inputs to the HDP model, where n_j is the total number of blog posts in the jth community. If π_j denotes the topic mixture proportion for the community jth, the jth community can be represented by a vector θ_j which is a K-dimensional vector of the topic mixture proportions, where the kth element of the vector θ_j represents the mixture proportion π_j of the topic k for the community j. These mixtures are used to perform the latent ANEW-based community clustering task. Furthermore, latent topics are extracted from the corpus of ANEW words shall be referred to as *ANEW topics* formally defined as follows.

Definition 2 Let A be a set of 1, 034 sentiment-conveying English words known as ANEW, and an ANEW topic is a discrete distribution over the set A.

Examples of these ANEW topics are shown in Fig. 5.

3.4.4 ANEW usage-based community representation

This approach is called AUCR. Using the notation in the previous Sect. 3.4.3, each community j can be represented by a 1, 034-dimensional ANEW usage vector $\mathbf{a}_j$ whose kth

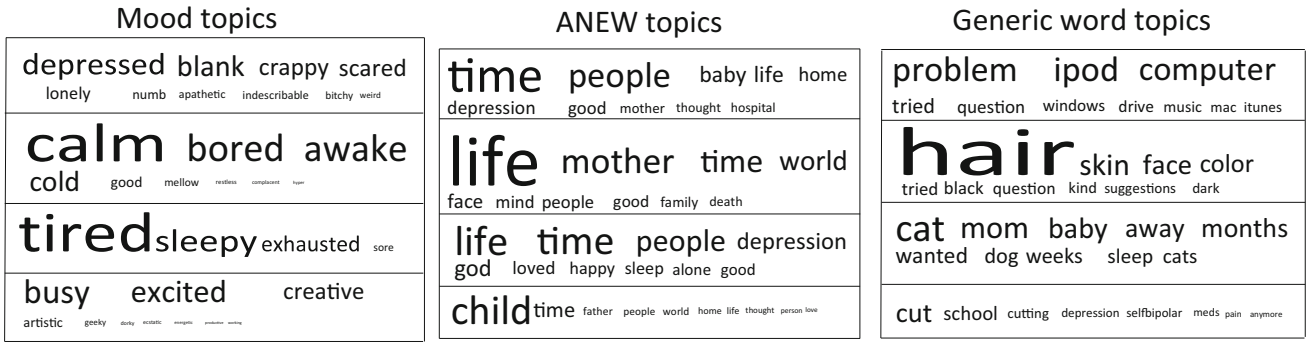

Fig. 5 Word cloud of some latent topics, in terms of mood-based topics (left), ANEW-based topics (middle), and generic word-based topics (right) inferred from the HDP model

element is the number of times of the kth ANEW term used in the content of blog posts made by users of the community j. The ANEW usage vector $\mathbf{a}_j$ is normalized to unity, so that $\sum_{i=1}^{1034} a_{ji} = 1$. With J communities, a matrix $T \in \mathbb{R}^{J\times 1034}$ which each row is the vector $\mathbf{a}_j$ is used to perform the ANEW usage-based meta-community discovering task.

3.4.5 Latent ANEW-based over community representation

Another method for community representation is based on latent ANEW-based topics across each community called LAoCR. For each community, all ANEW words from all blog posts are combined to form the corpus to input to the HDP model. Each community is treated as one document while each ANEW word is considered as a word in the model. For J online communities, the corpus consists of J documents containing ANEW words. We also ran the HDP model over this corpus of J documents to obtain K latent ANEW-based topics automatically. When π_j denotes the topic mixture proportion for the community jth, a vector θ_j can represent for the community j. The vector θ_j is a K-dimensional vector of the topic mixture proportions, where K is the number of latent topics and the kth element represents the mixture proportion of the topic k for the community jth. These mixtures θ_j are used to perform the community clustering task.

3.4.6 Latent generic word-based community representation

Assuming that similar online communities discuss a similar mix of topics, each community can be represented by the latent topics of what members of a community talk about. This community representation is called LGWCR. Let $\mathbf{x}_j = \{x_{1j}, x_{2j}, \ldots, x_{n_j j}\}$ be the set of blog posts in the jth community where n_j is the total number of blog posts in this community. We run the HDP model over a corpus of J documents aggregated over all communities $\mathscr{C} = \cup_{j=1}^{J} \mathbf{x}_j$, wherein each community is considered as one document while each generic word in the content of all posts of each community in the model [65]. Let θ_{ij} be the topic mixture for the post x_{ij}, then $\pi_j = \frac{1}{n_j}\sum \theta_{ij}$ denotes the topic mixture for the jth community. It is a K-dimensional vector whose the kth element represents the mixture proportion of the topic kth for the community jth. Such latent topics inferred by HDP from the corpus of generic words shall be referred to as *generic word topics* formally defined as follows.

Definition 3 Let V be a vocabulary of generic words in every post, and a generic word topic is a discrete distribution over the vocabulary V.

3.4.7 Psycholinguistics-based community representation

Another community representation is based on psycholinguistic features of blog posts made by users in each community. This representation, called LIWCR, can be seen a combination of pure topical and sentiment-based representations. The psycholinguistic or language features are extracted and classified by the LIWC package. The LIWC consists of 68 English terms assigned to one of the four high-level categories: linguistic processes, psychological processes, personal concerns, and spoken categories, which are also further sub-categorized into a three-level hierarchy. It also organizes terms into psychological meaningful categories based on words and language style reflect most domain of cognitive and emotional processes involved in interaction during communication. For each community, all 68 LIWC features are applied to build a 68-dimensional vector. It can be represented as a psycholinguistics-based community representation for the community clustering. We average all LIWC values on the number of posts over the number of members in the community.

3.5 Community clustering and evaluation

Understanding the formation of online mental health-related communities is important to comprehend aspects of mental health disorders in the online settings. We aim to group online

communities with and without mental health-related conditions that rely on a variety of features including sentiment, affective information, generic words, and psycholinguistic (or language) features. Nonparametric clustering approach is applied to automatically discover the unknown number of meta-communities. Then, to evaluate the performance of community clustering algorithm, we use four popular and standard metrics given in Sect. 3.5.2.

3.5.1 Nonparametric clustering

In this paper, we use affinity propagation (AP) [18] algorithm, a nonparametric clustering method, to automatically discover the number of clusters of online communities as well as the cluster exemplars. In our setting, this is crucial since the number of meta-communities cannot be easy to know in advance. The algorithm requires the pairwise similarities between data points that are representative of communities. In our case, we define similarity measures between two communities based on the latent sentiment (mood tags or ANEW lexicon), generic topical, or psycholinguistic (LIWC) features in these communities. The Jensen–Shannon (JS) divergence [based on negative Kullback–Leibler (KL) divergence] [16] is used to compute the topic proportion similarities between communities. As defined in Sect. 3.4, each community is represented as a proper probability mass function over topics (HDP latent topics), mood usage, or affective word usage. With topic modelling, x and y are two probability measures, representing the mixing proportions of topics, the JS divergence is calculated as follows:

$$JS(\boldsymbol{x}||\boldsymbol{y}) = \frac{1}{2}KL(\boldsymbol{x}||z) + \frac{1}{2}KL(\boldsymbol{y}||z) \tag{1}$$

where $z = \frac{1}{2}(\boldsymbol{x} + \boldsymbol{y})$ and $KL(\boldsymbol{x}||y) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i}$ is the KL distance. The pairwise similarity between two topic proportions (data points) is computed by $D(x, y) = e^{-JS(x,y)}$.

3.5.2 Evaluation of clustering

The clustering performance is judged by using four standard metrics of clustering quality as follows: *cluster purity* (CP), the *normalized mutual information* (NMI), the *rand index* (RI) and the *F-measure* (F1-score) [32]. In particular, CP is a transparent evaluation measure, while NMI can be an information theoretic interpretation of clustering. The RI punishes both types of errors: false-positive and false-negative decisions during clustering. In addition, the F-measure supports differential weighting of these two types of errors. More details are given in [32]. The higher the metric which always a number between 0 and 1, the better the clustering quality.

4 Experiments and Analysis

In this section, we perform experiments and analysis on nonparametric discovery of online meta-communities. We first use the HDP model to automatically learn the latent topics for mood tags, affective information, psycholinguistics, and generic words in the posts made by users in the communities with different community representations (except MUCR, AUCR, and LIWCR because of their different representation approaches, as shown in Sects. 3.4.2, 3.4.4, and 3.4.7, respectively). The community representations are discussed in Sect. 3.4 and are summarized in Table 3. In addition, Table 4 shows the number of HDP latent topics yielded from the model for each type of features. In particular, we emphasize on both the mood-based and ANEW-based topics for which, unlike usual notion of generic word topics, they represent the sentiment conveyance in the context. We *then* use the topic mixture proportions inferred by the HDP model as input to the AP algorithm for categorizing meta-groups of communities. The nonparametric clustering methods were described in Sect. 3.5. We evaluate the clustering performance at two levels of ground truth including 3 categories and 11 subcategories which are shown in Fig. 2.

Furthermore, examples of tag cloud visualization for different types of latent topics are shown in Fig. 5. In the visualization, mood-based topics are reasonably informative by grouping moods with similar valance value together (e.g. "*tired*", "*sleepy*", "*exhausted*"). For generic word topics, the topic {*cut, school, cutting, depression, self, bipolar, …*} is related to mental health-related conditions such as self-harm, depression and bipolar. In addition, ANEW-based topics also provide informative topics related to affective information through the discussion within online mental health-related communities. It is potential input features for clustering distinct groups of similarity communities. We present the results for the discovery of meta-communities as follows.

4.1 Meta-communities Discovery

In this section, we present the results for the meta-community discovery on the different community representations, including mood-based (i.e. LMCR and MUCR), affective word-based (i.e. LACR, AUCR, and LAoCR), language style-based (LIWCR), and generic word-based (LGWCR) representations. Regarding the results in Table 4, for each community representation, the number of clusters is called the number of meta-communities. In addition, for clustering evaluation, we report all four metrics as presented in Sect. 3.5.2, namely CP, NMI, RI, and F-score. The clustering performance based on different community representation for two levels of the ground truth of 3 groups and of 11 subgroups is shown in Tables 5 and 6, respectively. This two levels of ground truth are mentioned earlier in Fig. 2.

Table 3 A summary of community representations

Representation	Description	Sections
LMCR	By mixture proportions of latent mood-based topics	3.4.1
MUCR	By mood usage-based features for each community	3.4.2
LACR	By using affective-based topic mixture proportions.	3.4.3
AUCR	By affective information usage in each community	3.4.4
LAoCR	By latent affective-based topic mixtures over community	3.4.5
LGWCR	By topic mixtures of latent generic word-based topics	3.4.6
LIWCR	By mixture proportions of psycholinguistics-based topics	3.4.7

Table 4 Clustering results with different community representations

Community representation	#HDP Topics	#Meta-communities
Latent mood-based (LMCR)	15	11
Mood usage-based (MUCR)		6
Latent ANEW-based (LACR)	17	14
Latent ANEW-based (LAoCR)	29	8
ANEW usage-based (AUCR)		12
Latent generic word-based (LGWCR)	41	8
Psycholinguistics-based (LIWCR)		6

Table 5 Clustering performance in comparison with the ground truth of 3 groups

	Community representation						
	Word-based	Mood-based		Affective-based			Psycholinguistics-based
	LGWCR (%)	LMCR (%)	MUCR (%)	LACR (%)	AUCR (%)	LAoCR (%)	LIWCR (%)
Purity	79	**84**	82	**84**	74	75	**84**
NMI	38	**45**	47	**45**	36	30	44
Rand index	67	68	71	65	63	65	71
F-score	35	34	45	33	33	23	47

Bold values indicate the best performance of clustering on proposed community representations

Table 6 Clustering performance in comparison with the ground truth of 11 subgroups

	Community representation						
	Word-based	Mood-based		Affective-based			Psycholinguistics-based
	LGWCR (%)	LMCR (%)	MUCR (%)	LACR (%)	AUCR (%)	LAoCR (%)	LIWCR (%)
Purity	56	68	53	**74**	47	53	47
NMI	60	69	69	**73**	47	56	56
Rand index	86	88	85	85	83	88	82
F-score	39	46	45	37	28	29	33

Bold values indicate the best performance of clustering on proposed community representations

For the clustering performance on the ground truth of 3 groups, Table 5 shows that both mood tags and affective words are in fact the most informative sources for identifying online meta-communities. The best performance is obtained on user-annotated mood-based representations, achieving 84% CP and 45% NMI for LMCR; 82% CP and 47% NMI for MUCR. More interesting, without any user-annotated mood tags, only LACR can be considered an effective and efficient way for large-scale sentiment analysis on online mental health-related communities. The results for the LACR representation are 84% CP and 45% NMI. Similar to the results for MUCR and LACR, the clustering performance on LIWCR also achieved the same high value at 84% CP and 44% NMI. However, for AUCR, LAoCR, and LGWCR representations,

the clustering performance was poor compared to MUCR, LMCR, and LACR representations.

At a finer level of the clustering evaluation on the ground truth of the 11 subgroups, the best clustering performance is achieved on the LACR representation. Indeed, Table 6 shows that the clustering performance on the LACR approach achieves 74% CP and 73% NMI, while the metrics for the mood-based representation (LMCR) are the second highest with 68% CP and 69% NMI. Similar to the clustering result on the 3 groups, the metrics for AUCR, LAoCR, and LGWCR representations are around 50% CP and 50% NMI which is still lower than for LACR, LMCR, and MUCR representations.

Furthermore, we investigate the distance between online communities in the usage of latent topics for mood tags, affective words, generic words, and language styles to understand how latent topics shape the community clustering. Particularly, we use a recent advanced tool in machine learning called t-distributed stochastic neighbour embedding (t-SNE) [31] to project each representation for online communities into 2D spaces. The t-SNE tool is considered as a new method for the visualization of large datasets [61]. As shown in Figs. 13a, 14a, and 15a, it is interesting to see that the visualization of online communities of 3 groups (DEPRESSION, AUTISM, and GENERAL) are well separated in using mood-based and ANEW-based representation for such communities. The data points represented for the communities in each group are almost separated from other groups in 2D spaces. However, the visualization shows the present of some autism-related and depression-related communities in the area of the GENERAL group. Moreover, we note that most communities of AUTISM group are often appeared closely to those in the DEPRESSION group on different aspects of community representation. In short, the separation between mental health-related and general groups on social media is visible, confirming the difference in the usage latent topics of either moods, affective information or generic words between three groups. In the finer level of 11 sub-categories, all Figs. 13b, 14b, and 15b show the distance in the interest of latent topics between communities of such sub-categories. The results indicate the informative aspects of sentiment analysis on online mental health-related communities to separate them. Therefore, the results of detecting online meta-communities based on different community representations including sentiment-based [e.g. mood-based (LMCR, MUCR), ANEW-based (LACR)], language style-based, and generic word-based (LGWCR) representations are analysed in more detail as follows.

4.1.1 Mood-based Meta-communities

In this section, we present the discovery of meta-communities based on two mood-based community representations, namely LMCR and MUCR. For LMCR, based on the similarity communities on the preference of latent mood topics, 11 meta-communities are clustered by the AP algorithm shown in Table 7. In addition, Fig. 10 illustrates the proportion of mood-based topics being used in each meta-community. As shown in Fig. 6, of these 11 meta-communities, there are eight and five 100% pure meta-communities with regard to the ground truth for both the 3 groups and the 11 subgroups, respectively. Mood-based community clustering reveals differences in these 100% pure meta-communities. For example, of the three 100% pure meta-groups those communities in the DEPRESSION category (Separation in LJ categories), one of which is the only 100% pure meta-communities in the clustering task with the 11 LJ categories. Regarding the GENERAL category, one of 100% pure meta-communities, cluster *no. 11*, includes all communities: {*computer_help, computerhelp, htmlhelp, ipod, macintosh, webdesign*} from the Technology category (Fig. 7). In addition, as shown in Fig. 13c, the visualization of distance in the interest of the HDP latent mood topics among the 11 meta-communities shows a visible separation between the clusters. In the visualization, three 100% pure meta-groups those communities in DEPRESSION category are separated from others. For MUCR, the community clustering task yielded 6 meta-communities as shown in Fig. 8. Of these 6 meta-communities, there are two 100% pure meta-communities (in the ground truth of the 3 groups) and only one 100% pure meta-communities (in the ground truth of the 11 subgroups) in the clustering task.

Table 7 Mood-based topics meta-communities by LMCR

No.	Community members
1	Autism, clucky, dogsintraining, naturalbirth
2	Bentolunch, picturing_food
3	_Lostsouls, 20plusbipolar, alonendepressed, beautifulsi, bipolar, bipolar_world, bipolars, bipolarsucks, depressedteens, fightdepression
4	Altparent, asperger, breastfeeding, parenting101
5	Beauty101, curlyhair
6	Cat_lovers, dog_lovers, note_to_cat
7	Imissmydad, momlessdaughter
8	Add-adhd, adult_bipolar, asd-families, ask-an-aspie, aspient, aspie-trans, autism-spectrum, autistic-abuse, lost-loved-ones, ofmornings, or-not-to-be, spectrum-parent, still-a-mommy, suicidesupport, survivedsuicide
9	Bipolarsurvival, depressionsucks, recoveryourlife, self-injury, self-mutilation, the-cutters
10	Dyedhair, trashy_eats, vintagehair
11	Computer_help, computerhelp, htmlhelp, ipod, macintosh, webdesign

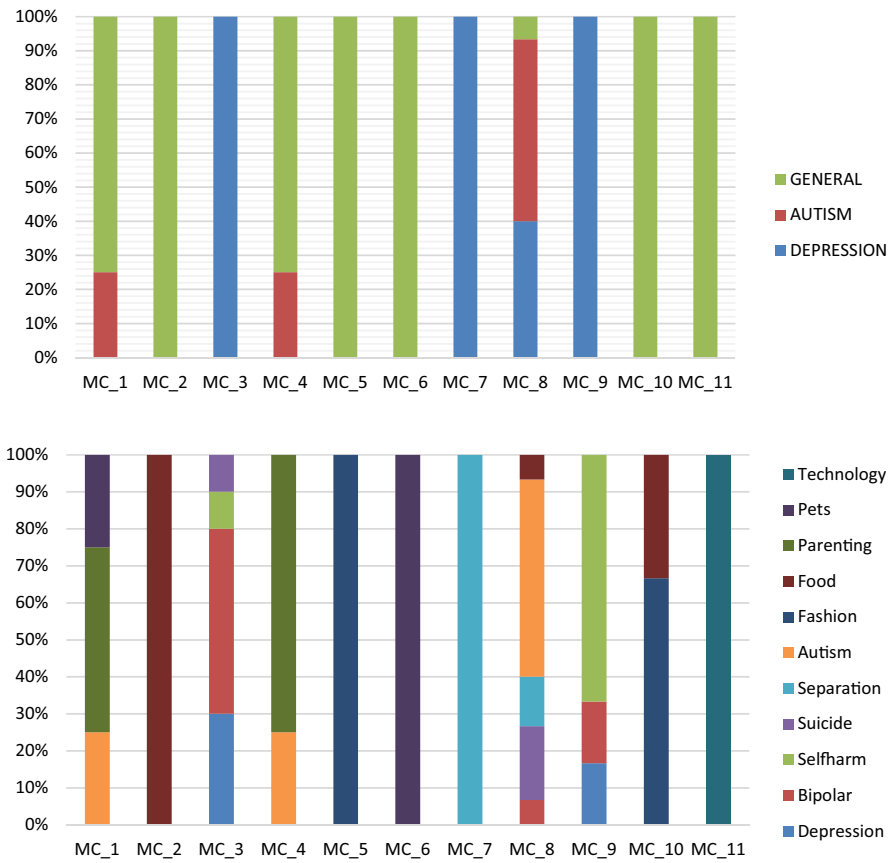

Fig. 6 Mood-based meta-communities by LMCR on 3 and 11 categories; multicoloured meta-community is less pure

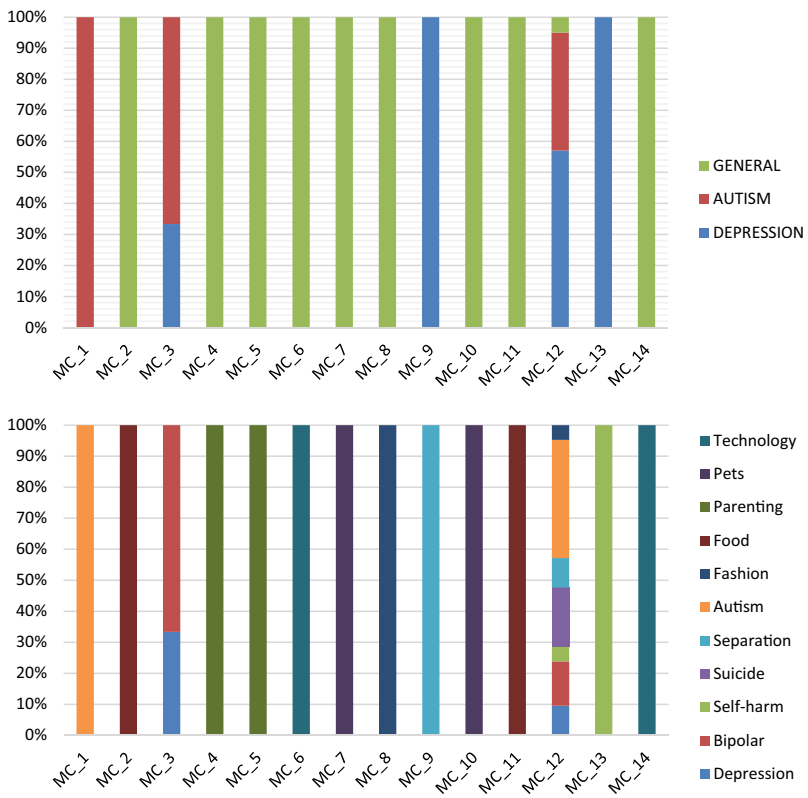

Fig. 7 ANEW-based meta-communities by LACR on 3 and 11 categories; multicoloured meta-community is less pure

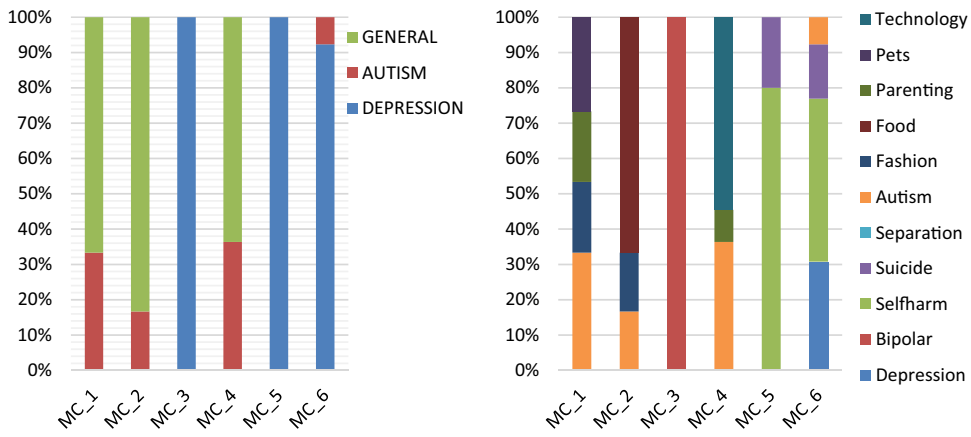

Fig. 8 Mood-based meta-communities by MUCR on 3 and 11 categories; multicoloured meta-community is less pure

4.1.2 ANEW-based meta-communities

We present the clustering results of the three ANEW-based community representations consisting of LACR, AUCR, and LAoCR. Of three representations, we focus on the clustering results of the LACR representation because its clustering performance is higher than the other two ANEW-based representations (AUCR and LAoCR). Clustering on the LACR yielded 14 meta-communities, as shown in Table 8. Interestingly, in Fig. 7, of these 14 meta-communities, there are twelve 100% pure meta-communities regarding the ground truth of both the 3 groups and the 11 subgroups, including one belonging to AUTISM cohort, two meta-groups from the DEPRESSION cohort, and nine from the LJ categories. Moreover, two other clusters (*Nos. 3* and *12*) consist of almost all mental health-related communities in both *the* DEPRESSION and AUTISM cohorts. Significantly, of 100% pure clusters, two (*Nos. 2* and *9*) are found in both the LACR and LMCR representations, such as {*bentolunch, picturing_food*} and {*imissmydad, momlessdaughter*}. In addition, one discovered meta-community, {*computer_help, computerhelp, htmlhelp, ipod, macintosh, webdesign*} in the LMCR is further devided into two meta-communities ({*computerhelp, ipod, macintosh*} and {*computer_help, htmlhelp, webdesign*}) in LACR. Some meta-communities in this clustering are found as subgroups in mood-based clustering. Moreover, three 100% pure meta-communities (*Nos. 1, 9,* and *13*) consist of 100% pure mental health-related communities in either autism, separation or self-harm categories. As shown in Fig. 6, the results of clustering performance of this representation are the highest results in both cluster purity and NMI.

For a better understanding of how latent ANEW topics shape meta-communities, we visualize the distance in the usage of the latent affective-based topics of these meta-communities. Figure 14c shows a visible separation between meta-communities. Particularly, clusters of those communities in both the depression or autism cohorts, namely clusters *Nos. 1, 3, 9,* and *13*, are further separated in the visualization.

Table 8 ANEW-based meta-communities by LACR

no.	Community member
1	Asperger, autism
2	Bentolunch, picturing_food
3	Bipolar, bipolars, bipolarsucks, bipolarsurvival, depressionsucks, fightdepression
4	Breastfeeding, parenting101
5	Altparent, clucky, naturalbirth
6	Computerhelp, ipod, macintosh
7	Dog_lovers, dogsintraining
8	Beauty101, curlyhair, dyedhair
9	Imissmydad, momlessdaughter
10	Cat_lovers, note_to_cat
11	Ofmornings, trashy_eats
12	_Lostsouls, 20plusbipolar, add-adhd, adult_bipolar, alonendepressed, asd-families, ask-an-aspie, aspient, aspie-trans, autism-spectrum, autistic-abuse, beautifulsi, bipolar_world, depressedteens, lost-loved-ones, or-not-to-be, spectrum-parent, still-a-mommy, suicidesupport, survivedsuicide, vintagehair
13	Recoveryourlife, self-injury, self-mutilation, the-cutters
14	Computerhelp, htmlhelp, webdesign

This confirm that there are differences in the usage latent topics of affective information in the content of the posts between discovered meta-communities.

4.1.3 Generic word-based meta-communities

For the LGWCR, the clustering yielded 8 meta-communities as listed in Table 9. Of these discovered meta-communities, there are three 100% pure meta-communiteies, namely cluster *No. 6* (those communities in the DEPRESSION group), clusters *No. 7*, and *No. 8* (those communities in the *General* category) for the ground truth of both the 3 groups and 11 sub-categories. Only the cluster *No. 8* has 100% purity with respect to the topical ground truth, including all communi-

Table 9 Generic word-based topics meta-communities

no.	Community member
1	Altparent, asperger, cat_lovers, dog_lovers, note_to_cat, parenting101
2	Add-adhd, asd-families, ask-an-aspie, aspient, aspie-trans, autism-spectrum, autistic-abuse, lost-loved-ones, ofmornings, or-not-to-be, spectrum-parent, still-a-mommy, suicidesupport, survivedsuicide, vintagehair
3	_Lostsouls, 20plusbipolar, adult_bipolar, alonendepressed, autism, beautifulsi, bipolar_world, depressedteens, trashy_eats
4	Bipolarsurvival, breastfeeding, clucky, dogsintraining, momlessdaughter, naturalbirth
5	Beauty101, curlyhair, dyedhair
6	Bipolar, bipolars, bipolarsucks, depressionsucks, fightdepression, imissmydad, recoveryourlife, self-injury, self-mutilation, the-cutters
7	Bentolunch, picturing_food
8	Computer_help, computerhelp, htmlhelp, ipod, macintosh,webdesign

ties in the Technology from the 11 LiveJournal categories. As shown in the results, we observed that 90% of communities of two meta-communities (*Nos. 2* and *3*) are from mental health-related groups, namely DEPRESSION and AUTISM categories. Again, the visualization of the distance in the use of the HDP generic word topics among all 8 discovered meta-communities shows a further separation between these meta-communities, as shown in Fig. 15c. Moreover, the separation between three mental health-related meta-communities (i.e. *Nos. 2*, *3,* and *6*) is also visible.

4.1.4 Psycholinguistics-based meta-communities

Community clustering based on the psycholinguistics-based community representation yielded 6 meta-communities. As shown in Fig. 9, for the ground truth of the three groups, two meta-communities are 100% pure clusters (*Nos. 2* and *4*) with the same communities in the GENERAL category. For the ground truth of the 11 sub-categories, there is only one 100% pure meta-communities cluster (*No. 2*) with the same communities of the Food category. The top three LIWC features for all meta-communities are linguistic features such as *wc* (i.e. word count - length of post), *dict* (i.e. dictionary words), and *funct* (i.e. total function words), appearing in these meta-communities to help associate them together. Furthermore, of top ten LIWC categories shared in discovered meta-communities, both *cogmech* (i.e. cognitive processes) and *relativ* (i.e. relativity) features contribute significantly in the grouping process. Even though the number of 100% pure meta-communities is low, the clustering performance of this representation is quite high in comparison with other representations, such as LMCR and LACR, in the ground truth of the 3 groups.

4.2 Discussion on discovered meta-communities

In this section, we discuss the performance of the non-parametric discovery of online mental health-related communities from the different community representations. Sentiment-based representations such as LMCR, MUCR, and LACR are the best choices for the task of meta-community discovery. The clustering performance of these representations is higher than the results of other community representations (i.e. AUCR, LAoCR, and LGWCR). Particularly, the results illustrate that the LACR is highly recommended to be used for assessing sentiment aspect of online mental health-related communities. In contrast to expectations, other ANEW-based community representations do not perform well for the task of hyper-community (meta-communities) detection. Even though the generic word-based representation is considered as a suitable choice of the method for hyper-community detection of general community categories [43], in this study, the results for the LGWCR clustering are not better than some sentiment-based representations (e.g. LMCR or LACR). Moreover, the existing related study by Nguyen et al. [45] identified hyper-communities of 100 general online communities with the best measures on

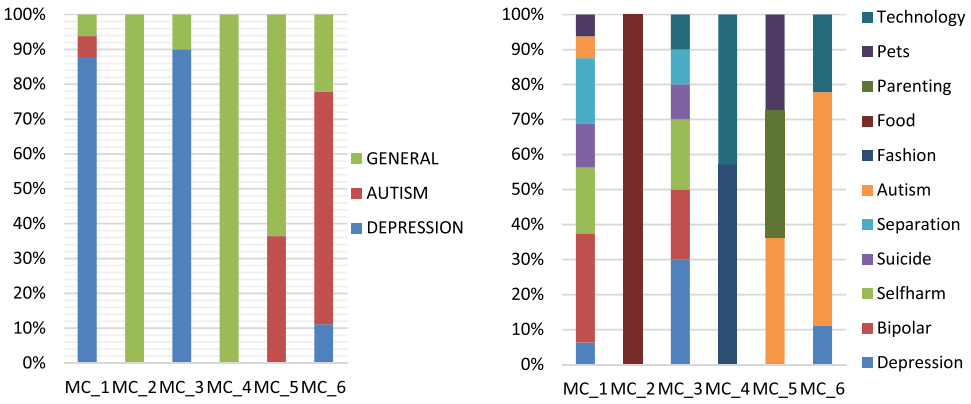

Fig. 9 LIWC-based meta-communities on 3 and 11 categories; multicoloured meta-community is less pure

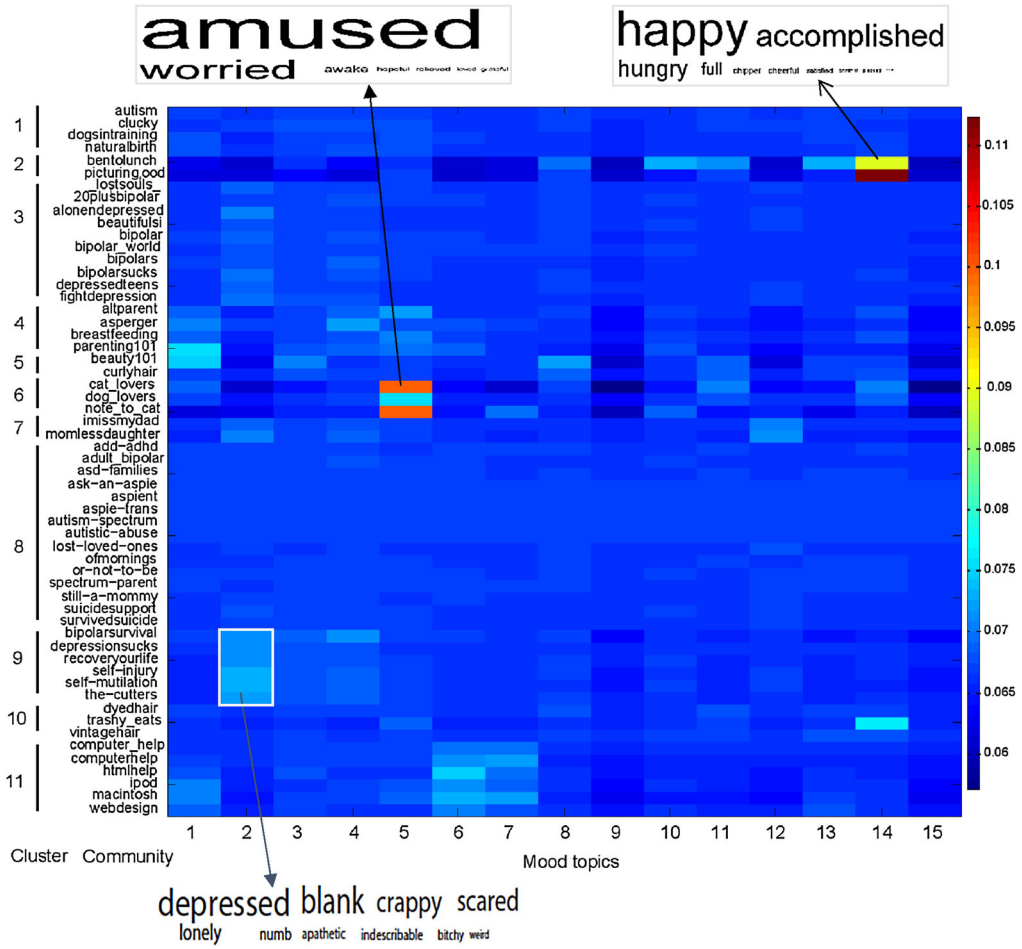

Fig. 10 Proportion of mood-based topics being used in each meta-communities (by LMCR)

cluster purity and NMI for HDP latent topic-based community representation versus others. Thus, our results are more meaningful when the better cluster performances are from latent HDP topics for the sentiment-based community representations (LMCR and LACR).

According to the meta-communities discovered using the LMCR representation, we investigate the shared emotions or feelings between the Depression and Autism communities in the same clustered meta-community. Figure 10 shows latent mood tag topics being used among the 11 discovered clusters (meta-communities). Cluster *No. 2,* consisting of *bentoluch* and *picturing_food* from GENERAL group, focused on using high positive valence moods such as *happy*, *accomplished and hungry* (topic 14). Likewise, cluster *No. 9,* including communities in the DEPRESSION group expressed negative emotional moods such as *depressed*, *blank*, and *scared* (topic 2). Only cluster *No. 8* is a mix group of communities from DEPRESSION, AUTISM and GENERAL groups. Details of shared feelings (mood) between the AUTISM's communities and DEPRESSION's communities in this cluster *No. 8* are found in Fig. 11. Obviously, tag cloud of top 20 moods usage in this cluster shows all negative emotion (low valence moods) such as *curious, confused, sad,* and, *tired,* as shown at the bottom of Fig. 11. In particular, three latent mood-based topics (topicS 2, 4, and 12) are dominant topics being discussed in this cluster (of The AUTISM and DEPRESSION communities). The DEPRESSION community in this meta-group tagged more moods in these three latent topics than the AUTISM community. Meanwhile, the only GENERAL's community in this cluster interested in other topics.

From the discovered meta-communities in the LACR clustering, cluster *No. 12* of the AUTISM and DEPRESSION communities strongly focused on 5 of 17 HDP affective word-based topics, namely as *3, 6, 9, 11,* and *14*. Figure 12 details on the latent mood-based topics shared by the Autism- and Depression-related communities in the cluster *No. 12*. Furthermore, DEPRESSION communities shows a stronger contribution on the shared latent topics than AUTISM communities when they were clustered into the same meta-community *No. 12* in the LACR representation.

Of the psycholinguistics-based meta-communities, as shown in Fig. 9, three meta-communities (Clusters *Nos. 1, 5, and 6*) include a mix of mental health-related communities

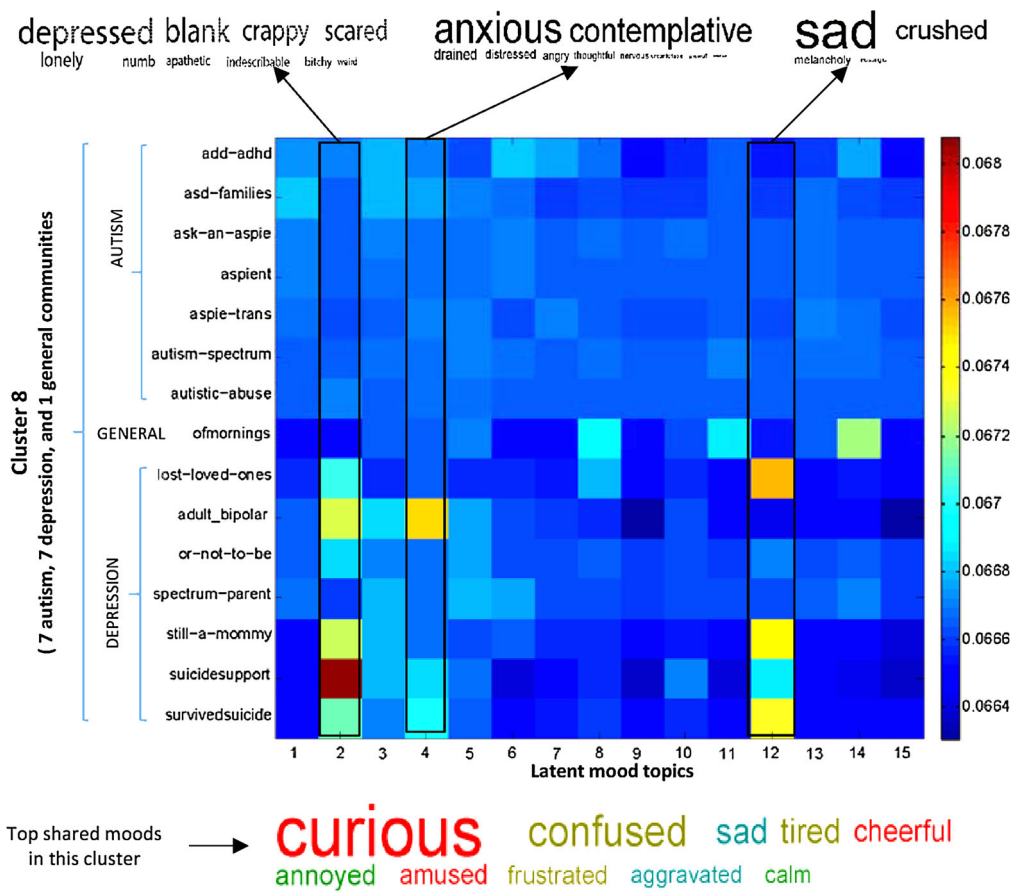

Fig. 11 Proportion of mood-based topics being used in each community of meta-group (Cluster No. 8 in LMCR)—shared moods between AUTISM and DEPRESSION groups

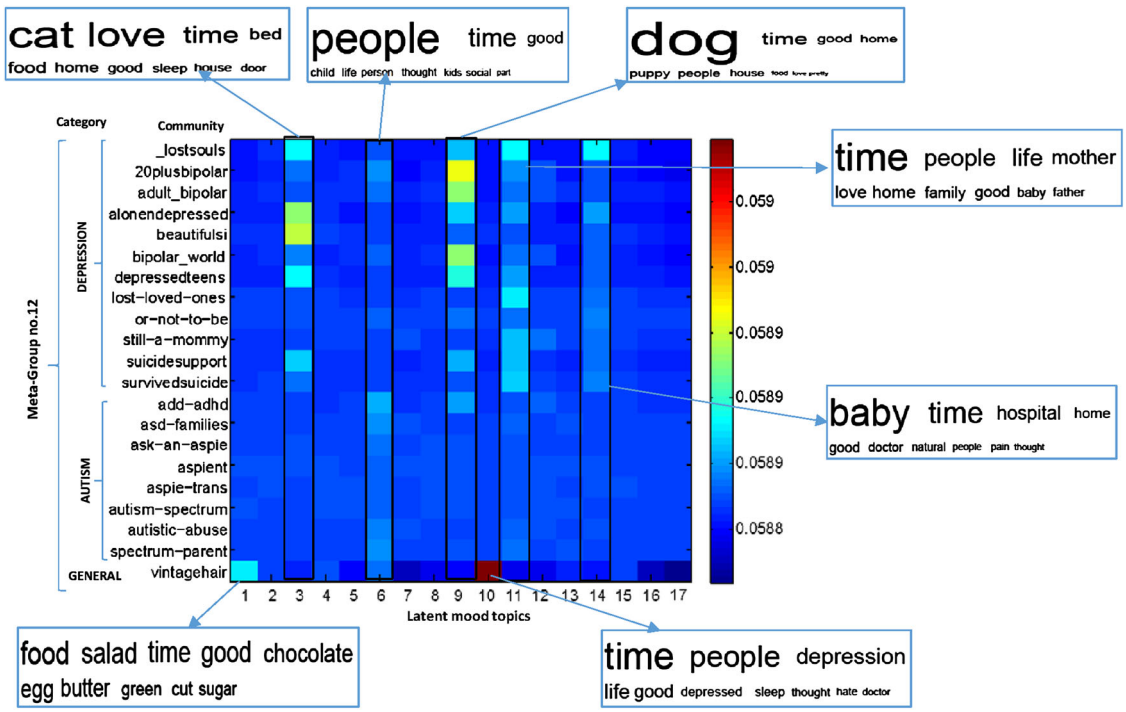

Fig. 12 Proportion of ANEW-based topics being shared between AUTISM and DEPRESSION communities in the meta-group *No. 12* by LACR

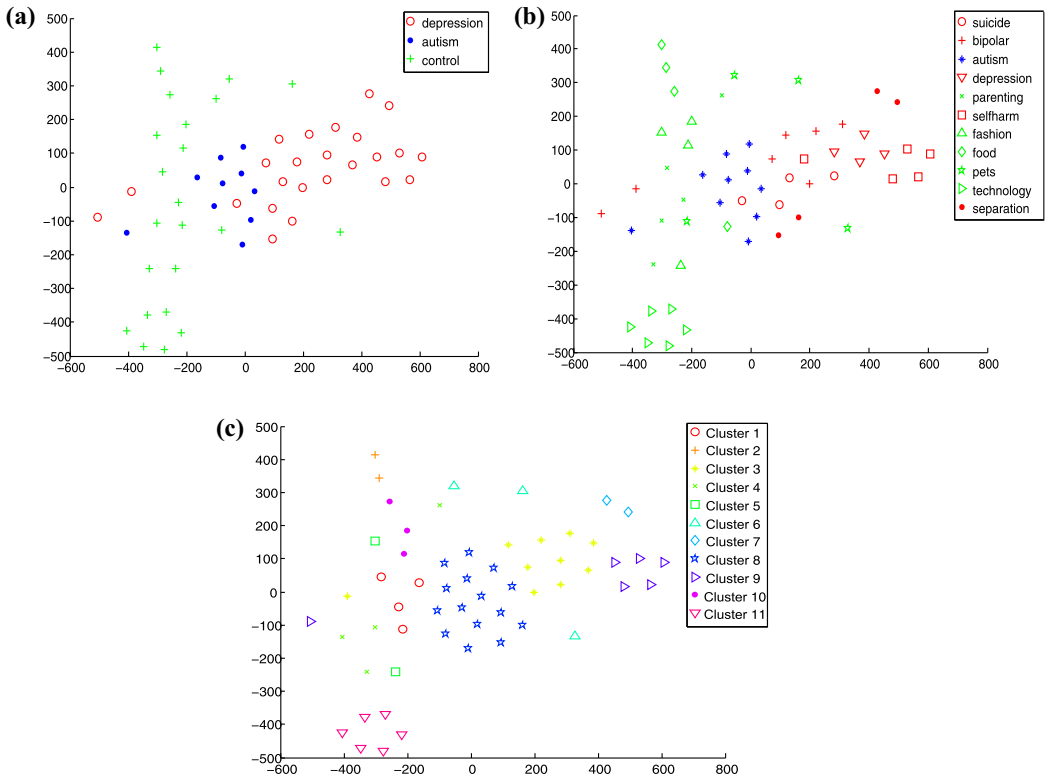

Fig. 13 (**a**) Visualization of 3 groups. (**b**) Visualization of 11subgroups. (**c**) Visualization of 11 clusters (from AP). Visualization of the distance in the interest of HDP mood topics among communities and meta-communities (visualized by t-SNE)

and others. The significant difference between these 3 meta-groups with the first 4 meta-groups in this clustering is that the *social* process (e.g. mate, they, talk, child) including *family* (e.g. daughter, wife, aunt), *friend* (e.g. friends, neighbour) and *human* (e.g. adult, boy, baby) sub-categories appear in the top 10 LIWC features above average by meta-communities.

4.3 Topics of interest in autism and depression cohorts

In this section, the difference in the topics of interest between the Autism and Depression communities is investigated. As given in Sect. 3.5.1, we identify that many meta-clusters consist of a mix of Autism and Depression communities. The relationship between these two groups can be seen in the visualizations in Figs. 13, 14, and 15. Furthermore, using the HDP latent topics of interests for the two groups, the difference in the use of the topics of interest between these two groups is defined as $diff = \left(\frac{mean_{asd} - mean_{dep}}{mean_{asd}}\right)$, for AUTISM versus DEPRESSION, where $mean_{asd}$ is the average of topic mixture proportion over communities of the AUTISM cohort, and $mean_{dep}$ is the average of topic mixture proportion over the DEPRESSION cohort. If $diff > 0$, the latent topic is more interested in the autism than the depression community and vice versa. As expected, autism-related topics are used more by the AUTISM community than the DEPRESSION community. As shown in Table 10, *autism, school,* and *read* were three top dominant topics for the AUTISM community, reflecting the difficulties facing individuals with autism in relation to education due to their issues with social communication and interactions (i.e. *social* is also a dominant topic in this group). These findings are quite similar to [48] in relation to differences in the use of topics between the *Autism* and *Control (General)* communities. Family-related topics (i.e. *mom, dad, mother*, and *family*) with negative emotions (i.e. *away*, *gone, died*,), mental-related concerns topics (i.e. *depression, bipolar, meds*, and *doctor*), and self-injured topics (i.e. *cut, cutting, hurt, pain,* and *suicide*) are dominant topics for the *Depression* community.

5 Conclusion

This paper analysed on online communities with and without mental disorders using a variety of features from blog post corpus to discover meta-communities. We used the HDP algorithm to infer latent topics from the corpus which was built from mood, affective words, language styles, and generic words in the blog posts made by users in online communities. We applied a nonparametric clustering algorithm to discover significant meta-communities among online

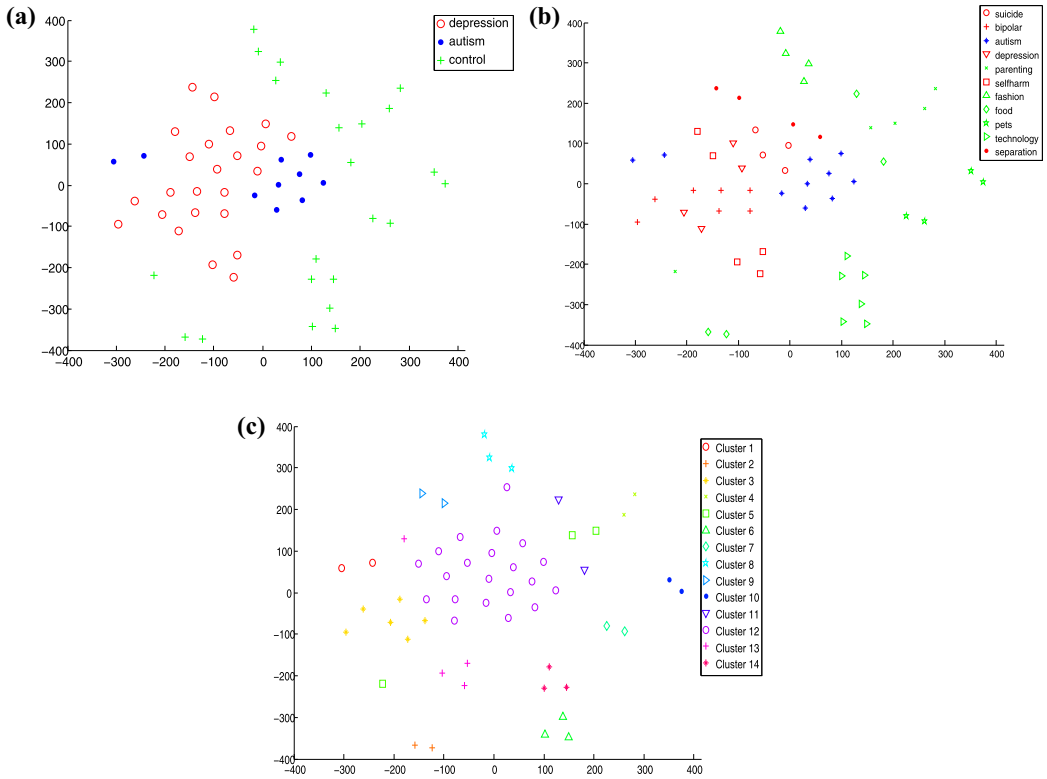

Fig. 14 (**a**) Visualization of 3 groups. (**b**) Visualization of 11 subgroups. (**c**) Visualization of 14 clusters (from AP). Visualization of the distance in the interest of HDP ANEW topics among communities and meta-communities (visualized by t-SNE)

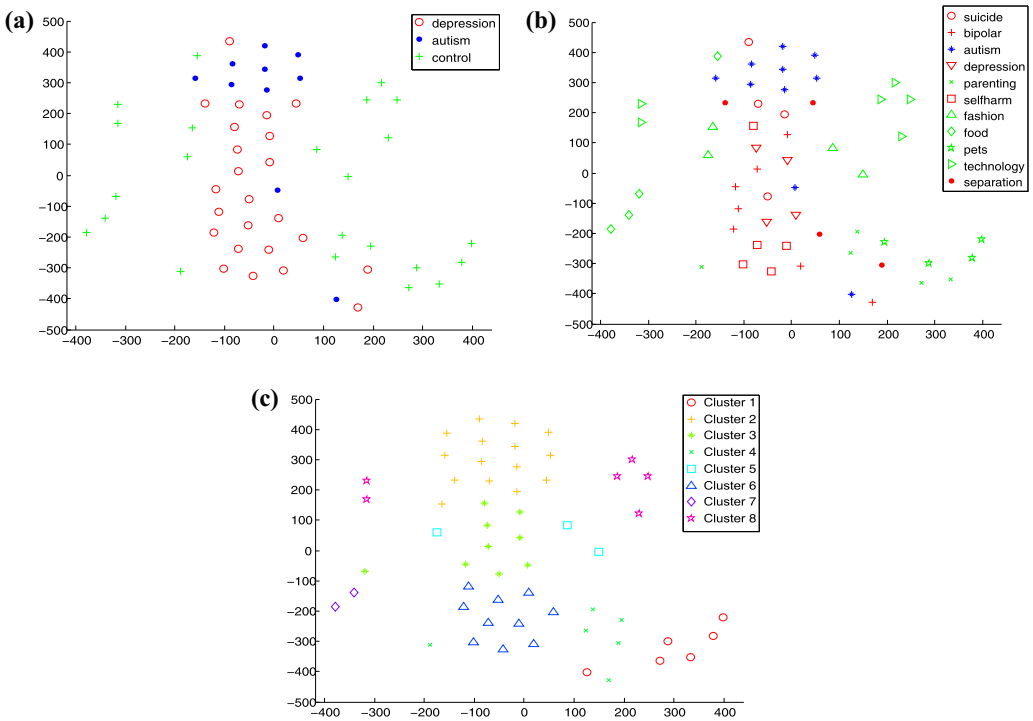

Fig. 15 (**a**) Visualization of 3 groups. (**b**) Visualization of 11 subgroups. (**c**) Visualization of 8 clusters (from AP). Visualization of the distance in the interest of HDP generic word topics among communities and meta-groups (visualized by t-SNE)

Table 10 Differences in the topics of interest of the AUTISM and DEPRESSION groups

Diff.	Word cloud
0.15	autism school read problem different child social autistic post parents idea children community job asperger question reading kids diagnosed group
0.01	anymore care joined boyfriend suicide worse loves wanted community sucks break soul pretend cause eyes kill blame scream crying hold
-0.1	school end started wrong start real reason point far head live journal mind world wants goes moment comes feelings change
-0.21	cut cutting stop self hurt pain anymore blood cuts away scars tried suicide deep arm started emo girl ass ive
-0.24	depression bipolar meds sleep disorder doctor mood medication manic weeks worse anxiety thoughts mental started past therapist diagnosed control posted
-0.25	mom dad mother away wanted family remember gone died miss care months knew father house heart cry left crying death

communities with different community representations. We also investigated the shared emotion/feeling between online autism-related and depression-related communities when they are clustered into the same meta-communities. Furthermore, the visualization of online discovered meta-communities in the use of latent topics shows a separation between the groups. This is evidence of sentiment-bearing differentiation in online mental health-related communities, suggesting a possible angle for building interventions that can bring help and support in mental healthcare for this online susceptible communities.

Acknowledgements This work is partially supported by the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning.

References

1. Antoniak, C.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. **2**(6), 1152–1174 (1974)
2. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: Proceedings of the 5th AAAI International Conference on Weblogs and Social Media (ICWSM'11), pp. 450–453 (2011)
4. Bradley, M., Lang, P.: Affective norms for english words (ANEW): instruction manual and affective ratings. Technical report C-1, the center for research in sychophysiology, University of Florida, Gainesville (1999)
5. Chomutare, T., Årsand, E., Fernandez-Luque, L., Lauritzen, J., Hartvigsen, G., et al.: Inferring community structure in healthcare forums. Methods Inf. Med. **52**(2), 160–167 (2013)
6. Christensen, H., Petrie, K.: Information technology as the key to accelerating advances in mental health care. Aust. N. Z. J. Psychiatry **47**(2), 114–116 (2013)
7. Christensen, H., Griffiths, K.M., Farrer, L.: Adherence in internet interventions for anxiety and depression: systematic review. J. Med. Int. Res. **11**(2), e13 (2009)
8. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Stat. Anal. Data Min. ASA Data Sci. J. **4**(5), 512–546 (2011)
9. Dao, B., Nguyen, T., Phung, D., Venkatesh, S.: Effect of mood, social connectivity and age in online depression community via topic and linguistic analysis. In: Proceedings of the 15th International Conference on Web Information System Engineering (WISE'14) pp. 398–407. Springer(2014)
10. Dao, B., Nguyen, T., Phung, D., Venkatesh, S.: Nonparametric discovery of online mental health-related communities. In: Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics pp. 1–10 . IEEE (2015)
11. De Choudhury, M.: You're happy, i'm happy: Diffusion of mood expression on twitter. In: Proceedings of Human–Computer Interaction (HCI) Korea (HCIK'15) pp. 169–179, ACM, HCI KORIA (2014)
12. De Choudhury, M., Counts, S., Horvitz, E.: Predicting postpartum changes in emotion and behavior via social media. In: Proceedings of the 2013 ACM Annual Conference on Human Factors in Computing Systems (CHI 2013) pp 3267–3276. ACM (2013a)
13. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Proceedings of the 5th ACM International Conference on Web Science (WebSci 2013) pp. 47–56. ACM (2013b)
14. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of The 7th AAAI International Conference on Weblogs and Social Media (ICWSM 2013) pp 1–10. AAAI (2013c)
15. Duan, L., Street, W.N., Liu, Y., Lu, H.: Community detection in graphs through correlation. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14) pp. 1376–1385. ACM (2014)
16. Endres, D., Schindelin, J.: A new metric for probability distributions. IEEE Trans. Inf. Theory **49**(7), 1858–1860 (2003)
17. Ferguson, T.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**(2), 209–230 (1973)
18. Frey, B., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007). doi:10.1126/science.1136800
19. Ghaziuddin, M., Ghaziuddin, N., Greden, J.: Depression in persons with autism: implications for research and clinical care. J. Autism Dev. Disord. **32**(4), 299–306 (2002)
20. Girvan, M., Newman, M.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
21. Griffiths, K.M., Calear, A.L., Banfield, M.: Systematic review on internet support groups (ISGs) and depression (1): Do ISGs reduce depressive symptoms? J. Med. Internet Res. **11**(3), e40 (2009a)
22. Griffiths, K.M., Calear, A.L., Banfield, M., Tam, A.: Systematic review on internet support groups (ISGs) and depression (2): What is known about depression ISGs? J. Med. Internet Res. **11**(3), e41 (2009b)
23. Griffiths, K.M., Mackinnon, A.J., Crisp, D.A., Christensen, H., Bennett, K., Farrer, L.: The effectiveness of an online support group for members of the community with depression: a randomised controlled trial. PLoS ONE **7**(12), e53,244 (2012)
24. Hawn, C.: Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. Health Aff. **28**(2), 361–368 (2009)
25. HofmannFrey, T.: Probabilistic latent semantic indexing. In: Proc. of ACM SIGIR Int. Conf. on Research and Development in Information Retrieval, ACM, pp 50–57 (1999)
26. Huang, C.M., Ying, J.J.C., Tseng, V.S.: Mining users behaviors and environments for semantic place prediction. In: Nokia Mobile Data Challenge 2012 Workshop, pp. 1–6. (2012)
27. Huang, C.M., Ying, J.J.C., Tseng, V.S., Zhou, Z.H.: Location semantics prediction for living analytics by mining smartphone data. In: 2014 International Conference on Data Science and Advanced Analytics pp. 527–533 (DSAA). IEEE (2014)

28. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Bus. Horizons **53**(1), 59–68 (2010)
29. Kumar, M., Dredze, M., Coppersmith, G., De Choudhury, M.: Detecting changes in suicide content manifested in social media following celebrity suicides. In: Proceedings of 26th ACM Conference on Hypertext and Social Media (HT'15) pp. 85–94. ACM (2015)
30. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: Proceedings of Conference on Knowledge discovery and data mining pp. 611–617. ACM (2006)
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(2008), 2579–2605 (2008)
32. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
33. Marwick, A.: Livejournal users: passionate, prolific and private. http://www.livejournalinc.com/LJ_Research_Report.pdf. Accessed 28 Sept 2015 (2008)
34. Mishne, G.: Experiments with mood classification in blog posts. In: Proceedings of ACM Workshop on Stylistic Analysis of Text for Information Access pp. 321–327. ACM (2005)
35. Mitchell, M., Hollingshead, K., Coppersmith, G.: Quantifying the language of schizophrenia in social media. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality pp. 11–20. ACL (2015)
36. Neal, R.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Gr. Stat. **9**(2), 249–265 (2000)
37. Negoescu, R., Adams, B., Phung, D., Venkatesh, S., Gatica-Perez, D.: Flickr hypergroups. In: Proceedings of ACM International Conference on Multimedia pp. 813–816 (2009)
38. Neuhauser, L., Kreps, G.L.: Rethinking communication in the e-health era. J. Health Psychol. **8**(1), 7–23 (2003)
39. Nguyen, T.: Mood patterns and affective lexicon access in weblogs. In: Proceedings of the ACL 2010 Student Research Workshop, Association for Computational Linguistics pp. 43–48 (2010)
40. Nguyen, T., Phung, D., Adams, B., Tran, T., Venkatesh, S.: Hypercommunity detection in the blogosphere. In: Proceedings of second ACM SIGMM Workshop on Social media pp. 21–26. ACM (2010)
41. Nguyen, T., Phung, D., Adams, B., Venkatesh, S.: Prediction of age, sentiment, and connectivity from social media text. In: Proceedings of International Conference on Web Information System Engineering (WISE'11) pp. 227–240, Springer (2011)
42. Nguyen, T., Phung, D., Adams, B., Venkatesh, S.: Event extraction using behaviors of sentiment signals and burst structure in social media. Knowl. Inf. Syst. (KAIS) **37**(2), 279–304 (2012a)
43. Nguyen, T., Phung, D., Adams, B., Venkatesh, S.: A sentiment-aware approach to community formation in social media. In: Proceedings of the 6th AAAI International Conference on Weblogs and Social Media (ICWSM'12) pp. 527–530. AAAI (2012b)
44. Nguyen, T., Dao, B., Phung, D., Venkatesh, S., Berk, M.: Online social capital: Mood, topical and psycholinguistic analysis. In: Proc. of the 7th AAAI Int. Conf. on Weblogs and Social Media (ICWSM'13) pp. 449–456. AAAI (2013a)
45. Nguyen, T., Phung, D., Adams, B., Venkatesh, S.: Mood sensing from social media texts and its applications. Know. Inf. Syst. **39**(3), 667–702 (2013b)
46. Nguyen, T., Phung, D., Venkatesh, S.: Analysis of psycholinguistic processes and topics in online autism communities. In: Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME'13) pp. 1–6. IEEE (2013c)
47. Nguyen, T., Phung, D., Dao, B., Venkatesh, S., Berk, M.: Affective and content analysis of online depression communities. IEEE Trans. Affect. Comput. **5**(3), 217–226 (2014)
48. Nguyen, T., Duong, T., Phung, D., Venkatesh, S.: Autism blogs: expressed emotion, language styles and concerns in personal and community settings. IEEE Trans. Affect. Comput. **6**(3), 312–323 (2015)
49. Nimrod, G.: From knowledge to hope: online depression communities. Intl. J. Disabil. Human Dev. **11**(1), 23–30 (2012a)
50. Nimrod, G.: Online depression communities: members' interests and perceived benefits. Health Commun. **28**(5), 425–434 (2012b)
51. Nimrod, G.: Challenging the internet paradox: online depression communities and well-being. Intl. J. Internet Sci. **8**(1), 30–48 (2013)
52. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
53. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of ACL Conference on Empirical Methods in Natural Language Processing pp. 79–86. ACL (2002)
54. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in Twitter. In: Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics pp. 1–8. ACM (2012)
55. Park, M., McDonald, D.W., Cha, M.: Perception differences between the depressed and non-depressed users in twitter. In: AAAI The 7th International Conference on Weblogs and Social Media (ICWSM 2013) pp 476–485. AAAI (2013a)
56. Park, S., Lee, S.W., Kwak, J., Cha, M., Jeong b Bumseok2: Activities on facebook reveal the depressive state of users. J. Med. Internet Res. **15**(10), e217 (2013b)
57. Paul, M., Dredze, M.: You are what you tweet: analyzing twitter for public health. In: Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM'11) pp. 266–272 , AAAI (2011)
58. Paul, M., Dredze, M.: Discovering health topics in social media using topic models. PloS one **9**(8), e103,408 (2014)
59. Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., Booth, R.: The Development and Psychometric Oroperties of LIWC2007. LIWC Inc, Austin, Texas (2007)
60. Phung, D., Gupta, S.K., Nguyen, T., Venkatesh, S.: Connectivity, online social capital and mood: a bayesian nonparametric analysis. IEEE Trans. Multimed. **15**(6), 1316–1325 (2013)
61. Platzer, A.: Visualization of SNPs with t-SNE. PLoS ONE **8**(2), e56,883 (2013)
62. Short, S.E., Mollborn, S.: Social determinants and health behaviors: conceptual frames and empirical advances. Curr. Opin. Psychol. **5**, 78–84 (2015)
63. Stewart, M.E., Barnard, L., Pearson, J., Hasan, R., OBrien, G.: Presentation of depression in autism and asperger syndrome a review. Autism **10**(1), 103–116 (2006)
64. Tausczik, Y., Pennebaker, J.: The psychological meaning of words: Liwc and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
65. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006)
66. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15) pp. 3187–3196. ACM (2015)
67. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Know. Inf. Syst. **42**(1), 181–213 (2015)
68. Ying, J.J.C., Chang, Y.J., Huang, C.M., Tseng, V.S.: Demographic prediction based on users mobile behaviors. In: Mobile Data Challenge 2012 (by Nokia) Workshop, pp 1–6 (2012)