Tweeting Your Mental Health: Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions

Xuetong Chen
Loughborough University
X.Chen5@lboro.ac.uk

Martin D. Sykora
Loughborough University
M.D.Sykora@lboro.ac.uk

Thomas W. Jackson
Loughborough University
T.W.Jackson@lboro.ac.uk

Suzanne Elayan
Loughborough University
S.Elayan2@lboro.ac.uk

Fehmidah Munir
Loughborough University
F.Munir@lboro.ac.uk

Abstract

Applying simple natural language processing methods on social media data have shown to be able to reveal insights of specific mental disorders. However, few studies have employed fine-grained sentiment or emotion related analysis approaches in the detection of mental health conditions from social media messages. This work, for the first time, employed fine-grained emotions as features and examined five popular machine learning classifiers in the task of identifying users with self-reported mental health conditions (i.e. Bipolar, Depression, PTSD, and SAD) from the general public. We demonstrated that the support vector machines and the random forests classifiers with emotion-based features and combined features showed promising improvements to the performance on this task.

1. Introduction

Mental health problems are the fifth greatest global burden of disease and a leading cause of disability worldwide [1, 2]. Based on the information provided by the World Health Organization [3], common mental disorders including depression, bipolar affective disorder, dementia and schizophrenia affect about 410 million people globally, among which depression alone affects about 350 million people, making it the world's fourth most common disease [4]. Mental disorders could lead to self-harm, even suicide, which is a leading cause of death among teenagers and adults under 34 years old [5, 6]. In 2016, the WHO suicide statistics [3] showed that suicide contributed to more than 800,000 deaths every year. Bloom et al. pointed out that the estimated economic cost of mental illness was 2.5 trillion dollars in 2010, and would double by 2030 [7]. Hence, it is clear that the scale of the global impact of mental illness is substantial. Research institutions, governments and health organizations have performed numerous studies in a concentrated effort to reduce the overall mental health burden. But most existing studies heavily rely on small, often homogeneous samples of individuals, who may not necessarily be representative of the larger population. Moreover, traditional studies are usually based on surveys, depending on retrospective self-reports about moods and observations about health. This kind of traditional approaches are significantly ineffective [8] because they require repeated assessments and observations of individuals' behavior over a long period of time in order to collect useful levels of data of a patient's experiences. Also the measurements often suffer from large temporal gaps, which limits the capability of tracking and identifying risk factors that may be associated with mental illness, or developing effective intervention programs for agencies [8].

As social media becomes a central part of our daily life, user generated content and posts on social media have shown great potential in revealing sentiments, as well as emotional and behavioral patterns of users. This stream of data is "real-time", continuously generated, often capturing relatively fleeting, in-situ users' personal states, and yet publicly available. Due to these unique characteristics, social media data has been used in a variety of research areas with tools like natural language processing, sentiment analysis and machine learning. Based on the work of Conway and O'Connor [9], social media has already been increasingly used in population health monitoring, and is beginning to be used for mental health applications [10]. Furthermore, De Choudhury [8] suggests that mental health studies would benefit from employing social media, as it provides an unbiased collection of individuals' language usages and behaviors. De Choudhury also highlights that information from social media bears the potential to complement traditional survey techniques in its ability to provide finer

URI: http://hdl.handle.net/10125/50309
ISBN: 978-0-9981331-1-9

grained measurements of behavior over time while dramatically expanding population sample sizes [8]. Park et al. presented initial evidence showing that people do post about their depression and their treatments on social media [11]. According to Oxman et al., linguistic analysis can be used to classify patients who suffer from depression and paranoia [12]. In other words, analyzing individuals' language patterns in social media postings could help detect mental health conditions.

Recent mental health studies mainly focused on linguistic patterns using simple NLP methods on social media data to reveal insights of specific mental health disorders [13], such as post-traumatic stress disorder (PTSD) [14], seasonal affective disorder (SAD) [15], and depression [16]. Coppersmith et al. found that features including frequencies of first and third person pronouns, anger words, varied negative emotions as well as related patterns of language have a strong link to Twitter users with mental disorders [14]. Yet, these research efforts are still in their infancy and very few studies have incorporated sentiment analysis approaches. Fine-grained emotion analysis has not been considered although it is known that emotions play an important role in psychology and mental health domain.

From a theoretical viewpoint, emotions have been conceptualized in both dimensional (e.g. valence, arousal and motivation) and discrete (e.g. anger, sadness, happiness) perspectives [17]. Ekman's discrete model of emotion [18] consists of "sadness, happiness, anger, fear, disgust and surprise" and has been used in systems that recognize these emotional states [19]. Negative sentiments such as anger, fear and sadness are common in those with mental health conditions such as depression and bipolar [20]. Overall, negative emotions are considered to be a core feature of many mental disorders [20].

According to cognitive theories of emotion, cognitive appraisals determine if an emotion is experienced and which emotion is experienced [21]. Emotions are therefore seen as a response to a specific situation (internal or external) or as a person-situation transaction [22, 23]. In addition, the Differential Emotions theory [24] suggests that emotions are motivational and organize perception, cognition and behavior, to help us adapt and cope with the environment. Discrete emotions therefore serve us with biological functions. For example, fear functions to solve the problem of immediate danger by urging us to flee [25], and sadness facilitates the adaptation to loss [22]. Hence, emotions have consequences on health. Studies have demonstrated that higher activation of positive emotion is associated with increased life satisfaction and a longer life span; and higher activation of negative emotion is associated with increased mortality and morbidity [26]. Although positive emotions are common in those who suffer from bipolar, these emotions are abnormally intense and the intensity of emotion seems to be an important aspect that influences mental health [27]. Intense negative emotions are not only experienced in many mental health conditions including PTSD and depression, giving rise to feelings of being "out of control," but can also lead to the development of these conditions [28]. In many cases of depression, when intense negative emotions occur, there is numbing of these emotions, especially grief, fear, anger and shame [28]. And the numbing of emotions usually leads to a build-up of emotional tension, which in turn, can result in even more intense emotions [28].

In this paper, we replicated the work of Coppersmith et al. [15] on a new and more recent dataset collected in a similar fashion, and extended this work by employing fine-grained emotions as features for the first time in the task of identifying users with mental health conditions from users without. Additionally, we explored a broader set of machine learning classification algorithms and different combinations of features for a thorough comparison of the performances on this task.

2. Data

In order to identify users with specific mental health conditions from Twitter, we first collected tweets with self-reported diagnosis using the regular expression "I was/have been diagnosed with *condition*" [15] with Twitter streaming API (Application Programming Interface). *Condition* is one of the four selected mental health conditions, which are bipolar disorder, depression, post-traumatic stress disorder (PTSD) and seasonal affective disorder (SAD). The duration of the collection process lasted four months: from November 18th 2016 to February 15th 2017. Although, disingenuous statements from these self-reported diagnosis tweets were not formally analyzed, retweets were removed, since they are often an indication of the message being a quotation of others which likely to be a joke. For instance:
"*RT @user_screenname: Me: yeah, I was officially diagnosed with PTSD. Classmate: Wtf are you talking about? You weren't in any wars. Me: picture_url* ".

Using this approach, we obtained diagnosis tweets from more than 100 unique users for each condition, except for the SAD which had only 15 users. Due to the sparsity of SAD samples, we additionally searched the query "I was/have been diagnosed with SAD/ S.A.D./seasonal affective disorder" on Twitter, manually browsed through and examined all resulting tweets, hence selecting a list of 84 users with genuine SAD diagnosis tweets. The users who posted these diagnosis tweets, i.e. who self-reported being diagnosed with a targeted mental health condition, were considered as candidates to form the the four condition groups, which are the the bipolar, depression, PTSD and SAD groups according to their tweeted condition.

To select a sample of users representing the general population who do not suffer from mental illnesses, we collected one day (February 20th 2017) of tweets containing the keyword "the" using Twitter streaming API, and considered the users who posted these tweets as candidates of the control group. The control group was designed to contain data generated by normal Twitter users without any mental disorders in order to provide a comparison to reveal the differences and abnormality of the condition groups. Therefore, candidates from the control group were double checked against the candidates from the condition groups to make sure the control and the condition groups have no overlap that may interfere with the training process later on. It is worth noting that users who suffer from mental illnesses but did not post about their diagnosis might exist in the control group. It is also possible that some users from the control group have mental disorder symptoms or the actual condition but remain undetected and untreated. These users could add noise to the control group data and weaken the classification to some extent. However, considered that the self-selective and self-imposed representation of users is a significant feature of online social media, these noises are hardly avoidable and thus an ineluctable limitation of the usage of live user generated data.

Next, for obtained candidates of each group (four condition groups and one control group) up to 3200 past tweets were retrieved using the public Twitter search API. In this process, no private messages or protected user accounts were accessed by the researcher, and all collected tweets was publicly posted on Twitter. Users who had less than 25 tweets, or used non-English languages did not fit for the requirement of this study and were excluded. Hence, we managed to collect tweets (in average 1.5k/users) from 438, 585, 265, 84 and 6596 unique and valid users for respectively the bipolar, depression, PTSD, SAD, and control groups. These tweets of each group formed the dataset for this study.

3. Methodology

This work aims to explore the effectiveness of emotion based features and examine the performances of different machine learning classifiers for separating users with self-reported diagnosis from users without any mental health conditions (control users). We first trained a log-linear classifier using two feature sets, which are the LIWC (Linguistic Inquiry Word Count) language analytic feature set and the Pattern of Life (POF) feature set as presented in the work of Coppersmith et al. [15]. Then we extended this experiment by employing an additional set of fine-grained emotional features, followed by various combinations of the three feature sets. In addition to the Log-Linear (LR) classifier, our study employed four popular machine learning classifiers, which are the Support Vector Machines (SVM), the Naive Bayesian (NB), the Decision Trees (DT) and the Random Forests (RF).

3.1 LIWC and POF Features

Linguistic Inquiry Word Count (LIWC) is a computational tool for analyzing pieces of writing [29]. It has been demonstrated that the function and emotion words people use provide important psychological cues to their thought processes, emotional states, intentions, and motivations [30]. LIWC was used on individuals' past tweets in order to produce an LIWC feature set, which is formed by some of the LIWC categories directly (*Swear, Anger, PosEmo, NegEmo, Anx*) and combined pronoun classes *Pro1 (I and We)*, *Pro2 (SheHe)* and *Pro3 (They)*.

The Pattern of liFe (POF) feature set used in this study is formed by several measurements of user's activities. User posting-based activity features include daily *Tweet rate*, *Proportion of tweets with @mentions*, *Number of @mentions*, *Number of self-@mentions*, *Number of unique users @mentioned*, and *Number of users @mentioned more than 3 times*. Life analytic features include proportion of tweets that show evidence of *Insomnia or sleep disturbance*, *Exercise*, *Positive sentiment* and *Negative sentiment*.

More details of these features can be found in [15].

3.2 Emotion-based Features

Real-world problems cannot be simply explained or tackled with only positive-negative classification [19], not to mention identifying mental health conditions. Emotions are an important element of human nature, and thus they have been widely studied in neuroscience, psychology and behavior sciences [31]. In particular, many psychological studies examined the correlation between emotions, eating disorders, and other health issues. More recently, psychologists have also been exploring such signals from social media [15]. However, emotion-based features have not yet been considered nor incorporated in the analysis of mental health related social media datasets. Therefore, for the first time, we propose to employ fine-grained emotions into this task.

EMOTIVE, an ontology (semantic model) based advanced sentiment algorithm, developed by Sykora et al.

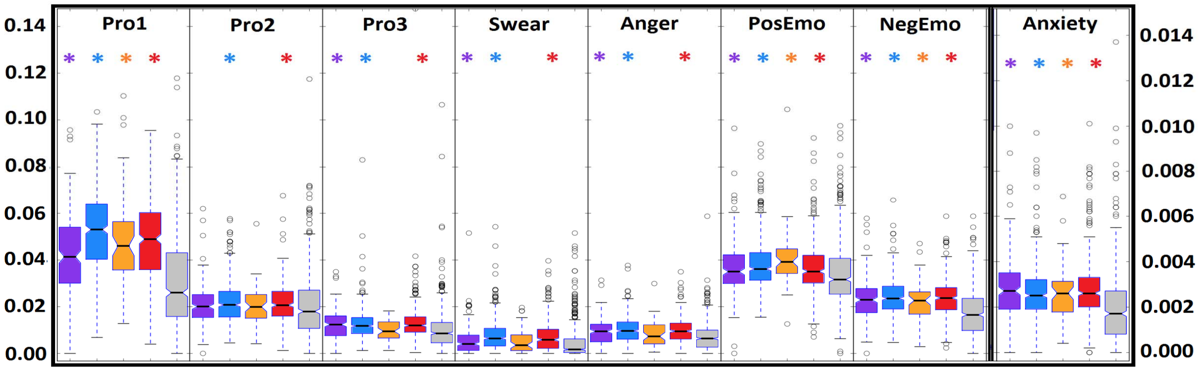


Figure 1: Box plot of proportion of individual's tweets (y-axis) matching various LIWC categories.

[32], is used in this work to detect the dominant fine-grained emotions from individuals' tweets. The EMOTIVE feature set is formed of the proportion of 9 emotion expressions (emotion scores) extracted from each user. These emotion attributes include 8 basic cross-cultural emotions, *Anger*, *Confusion*, *Disgust*, *Fear*, *Happiness*, *Sadness*, *Shame*, *Surprise*, and an emotion *Overall Score* which is a sum of all emotion scores, indicating an overall emotionality (i.e. Emotion Activation).

3.3 Experimental Setup

This work proposes to explore novel features and feature combinations, and to utilize a wider range of machine learning classification algorithms for identifying mental health conditions. Therefore, each of the three aforementioned feature sets (*LIWC, POF, EMOTIVE*) and their combinations (*LIWC+POF, LIWC+EMOTIVE, POF+EMOTIVE, LIWC+POF+EMOTIVE*) are used as inputs of a classification algorithm that separates users with a condition from those without.

The experiments were set to be four separate binary classification tasks for each of the selected mental health conditions against the control. For these classification tasks, we analyzed with four classifiers, which are the SVM, DT, RF and NB classifiers, in addition to the LR classifier also used in [15]. The combination of feature sets was made by concatenating the feature vectors from each set for every user. Z-score normalization ($z = \frac{x-\mu}{\sigma}$) was applied on all feature sets before the training and classification process. The classification performances were evaluated through leave-one-out cross validation. We then compared the classification accuracies on each of the seven feature sets across the four conditions and across the five classifiers.

4. Results

In this section, first, we validated the data collection method for this study by replicating previous findings with the features extracted by LIWC. Then, five classifiers were built to distinguish users with reported mental health condition diagnosis from the control users using various features including a novel emotion feature set. The performance achieved by different classifiers with different combinations of feature sets were evaluated and compared. Finally, statistical correlations were applied on each group for a more in depth analysis of the relationships among features of each condition.

4.1 Data Collection Validation

Using the same validation process as [15], the proportion of tweets that score positively on the selected LIWC categories from each group is presented in the box plot shown in Figure 1. Each box represents the distribution of one feature, and each color represents one of the condition and control groups where the features were extracted, which are bipolar in red, depression in blue, PTSD in purple, SAD in orange, and control in gray. *Anxiety* has a separate y-axis, on the right, due to its sparsity. The language features show a similar proportion and distribution for each group compared to the result from the original work [15]. As the differences that reach statistical significance from the control group are noted with asterisks, it can be observed that our dataset is consistent and also seems to show more features that have statistically significant deviations from the control users. Furthermore, the depression group shows significant differences from the control group for all eight LIWC features. This result replicated previous findings that for depressed users significant increases are expected in *NegEmo*, *Anger*, *Pro1* and *Pro3* compared to control users [11, 33, 34]. Hence, these findings validate the data collection method and the resulting dataset used in this study.

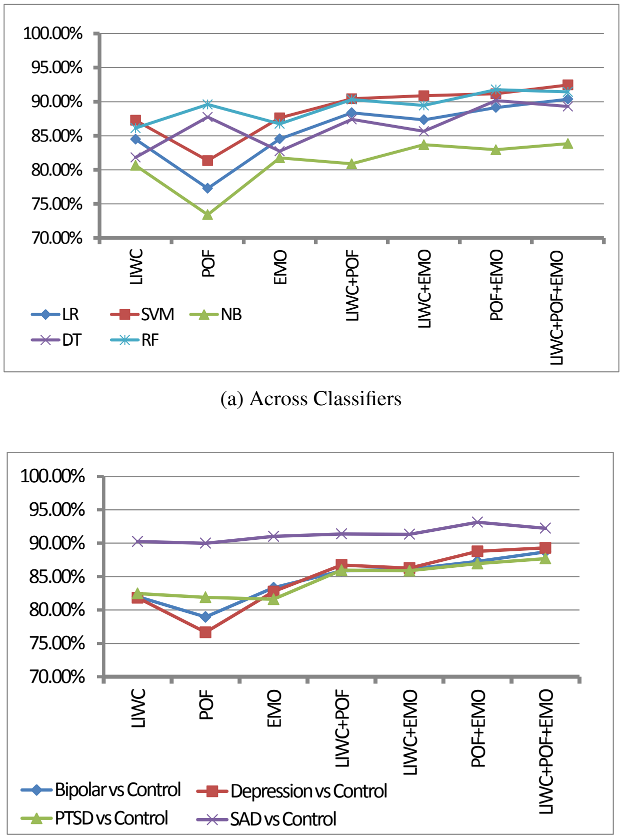
Bipolar vs Control					
Feature Set	**Acc.**	**Prec.**	**Rcl.**	**F**	**AUC**
LIWC	85.35%	0.836	0.809	0.820	91.16%
POF	74.89%	0.712	0.645	0.656	79.35%
EMOTIVE	86.68%	0.861	0.816	0.833	91.64%
LIWC+POF	89.46%	0.883	0.864	0.873	94.31%
LIWC+EMO	90.37%	0.894	0.875	0.884	94.71%
POF+EMO	89.26%	0.892	0.850	0.867	93.40%
ALL	91.91%	0.909	0.897	0.903	95.72%
Depression vs Control					
Feature Set	**Acc.**	**Prec.**	**Rcl.**	**F**	**AUC**
LIWC	85.25%	0.842	0.828	0.834	91.25%
POF	74.60%	0.722	0.699	0.706	79.57%
EMOTIVE	85.01%	0.844	0.819	0.829	90.97%
LIWC+POF	88.99%	0.882	0.874	0.877	93.60%
LIWC+EMO	88.22%	0.876	0.861	0.868	93.65%
POF+EMO	89.59%	0.895	0.872	0.882	93.53%
ALL	91.08%	0.908	0.893	0.900	94.96%
PTSD vs Control					
Feature Set	**Acc.**	**Prec.**	**Rcl.**	**F**	**AUC**
LIWC	84.26%	0.804	0.752	0.771	88.44%
POF	79.54%	0.731	0.672	0.690	83.22%
EMOTIVE	81.90%	0.780	0.694	0.719	85.42%
LIWC+POF	87.56%	0.843	0.814	0.827	91.69%
LIWC+EMO	87.84%	0.855	0.807	0.827	91.65%
POF+EMO	89.44%	0.884	0.824	0.848	91.56%
ALL	90.10%	0.879	0.850	0.863	93.93%
SAD vs Control					
Feature Set	**Acc.**	**Prec.**	**Rcl.**	**F**	**AUC**
LIWC	90.27%	0.787	0.639	0.678	90.56%
POF	89.19%	0.729	0.638	0.667	90.93%
EMOTIVE	92.30%	0.849	0.723	0.768	91.49%
LIWC+POF	93.78%	0.874	0.794	0.827	94.44%
LIWC+EMO	91.49%	0.809	0.718	0.753	92.74%
POF+EMO	94.32%	0.887	0.812	0.844	95.02%
ALL	94.86%	0.903	0.826	0.859	95.82%

Table 1: Model performances of various feature sets.

4.2 Classification

In order to explore the ability of identifying users with a mental health condition from the control users while using different features, we first trained a binary log-linear (LR) classifier for separating each condition against the control performing with leave-one-out cross validation as in the work of Coppersmith et al. [15]. Table 1 displays the performance on each of the binary classification task for predicting depression or non-depression classes of Users using different feature sets. Measures of classification accuracy include leave-one-out cross validation (Acc.), precision (Prec.), recall (Rcl.), f-score (F), and the area under curve (AUC).

As can be observed, when using a single feature set, the EMOTIVE feature set shows better performance than both the LIWC and the POF feature sets, except for PTSD where the LIWC feature set achieved the best performance. For all four conditions, the classification performances are improved further when using combined feature sets, with the best performance achieved when combining all three feature sets. These improvements suggest that the emotion-based features provide information from a more abstract emotional aspect and more relevant compared to the LIWC and POF, which can efficiently reveal differences between users who suffer from a mental disorder and users who do not.



(a) Across Classifiers

(b) Across Conditions

Figure 2: Average classification accuracy.

We then conducted the same binary classification tasks with four additional classifiers, which are the SVM (with RBF kernel), NB, DT (max_depth=6), and RF (n_estimator=10). The averaged classification accuracies while using each of the 7 different feature sets were calculated (a) across classifiers and (b) across conditions as shown in Figure 2. As can be inferred from Figure 2a, all classifiers show an increase in accuracy when more feature sets are utilised. The best performance overall was achieved by the SVM and the RF classifiers. The inconsistency of performance among classifiers when using the POF feature set only, could indicate that this set of features is not linearly separable, since the two nonlinear classifiers (DT, RF) showed notably better performance on this feature set than the linear classifiers (SVM, LR, NB). One possible way to build a more robust classifier for all feature sets could be to leverage the best performing classifier for each feature set and aggregate their decisions [35].

The averaged performances for each condition is displayed in Figure 2b. For all four conditions, an increase of classification accuracy can also be observed along the x-axis. The PTSD group seems to be less sensitive to the EMOTIVE feature set compared to other condition groups. The POF feature set alone appears to be less effective for the bipolar and the depression groups (can also be referred in Table 1). These results indicate that pattern of life measurements are the least relevant features to these two conditions. In other words, the differences in POF features between the bipolar or depression and control groups are less significant. However, there is always a considerable increase in accuracy, for all conditions, whenever a feature set is combined with the POF feature set. Either the LIWC or the EMOTIVE when combined with the POF feature set performs better than combined with each other (LIWC+EMO). This suggests that the EMOTIVE and the LIWC feature sets have more overlap of the information they capture in comparison to the overlap with the POF feature set. Higher steady classification accuracy achieved by the SAD group could be explained by the less noise (false self-reported diagnosis tweets) contained in the dataset for this group due to its manual data collection method, which highlights the importance of cleaning and preprocessing the diagnosis tweets at the very early stage.

Further studies could use deep learning approaches such as word2vec [36, 37] and autoencoders [38] to filter out the disingenuous diagnosis statements collected from Twitter. Feasible sarcasm, irony and humour detection methods could also be incorporated into this task. Moreover, the increase in classification accuracy when using more features motivates us to perform a more in depth user profile analysis. Applying topic modelling on individuals' tweets is also worth exploring to reveal content-based information related to mental disorders.

4.3 Feature Analysis

In order to analyze the relationship between features, Pearson's statistical correlations were extracted separately from each of the control and condition groups, as presented in Figure 3 and Figure 4. Each number from 1 to 27 along the two axis represents one of the overall 27 features from the three feature sets used in this study, as shown in Table 2. As can be referred to the color bar, the shade of the color indicate the strength and polarity of the correlations from 100% positive to 40% negative as shown beside the matrices. The range of correlations, $[-0.4, 1]$, was decided based on the resulting maximum and minimum values of all groups to make the color shade most representative.

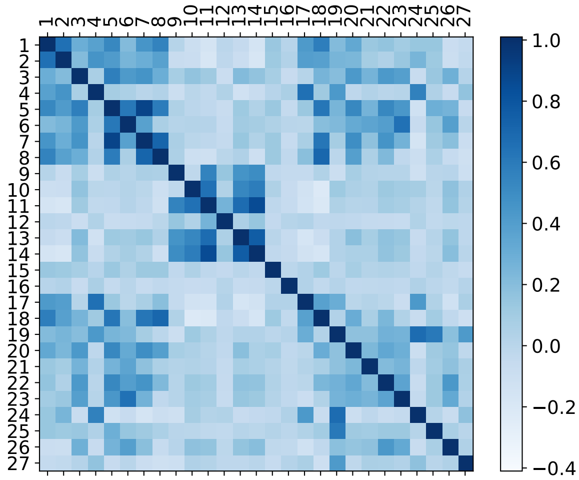
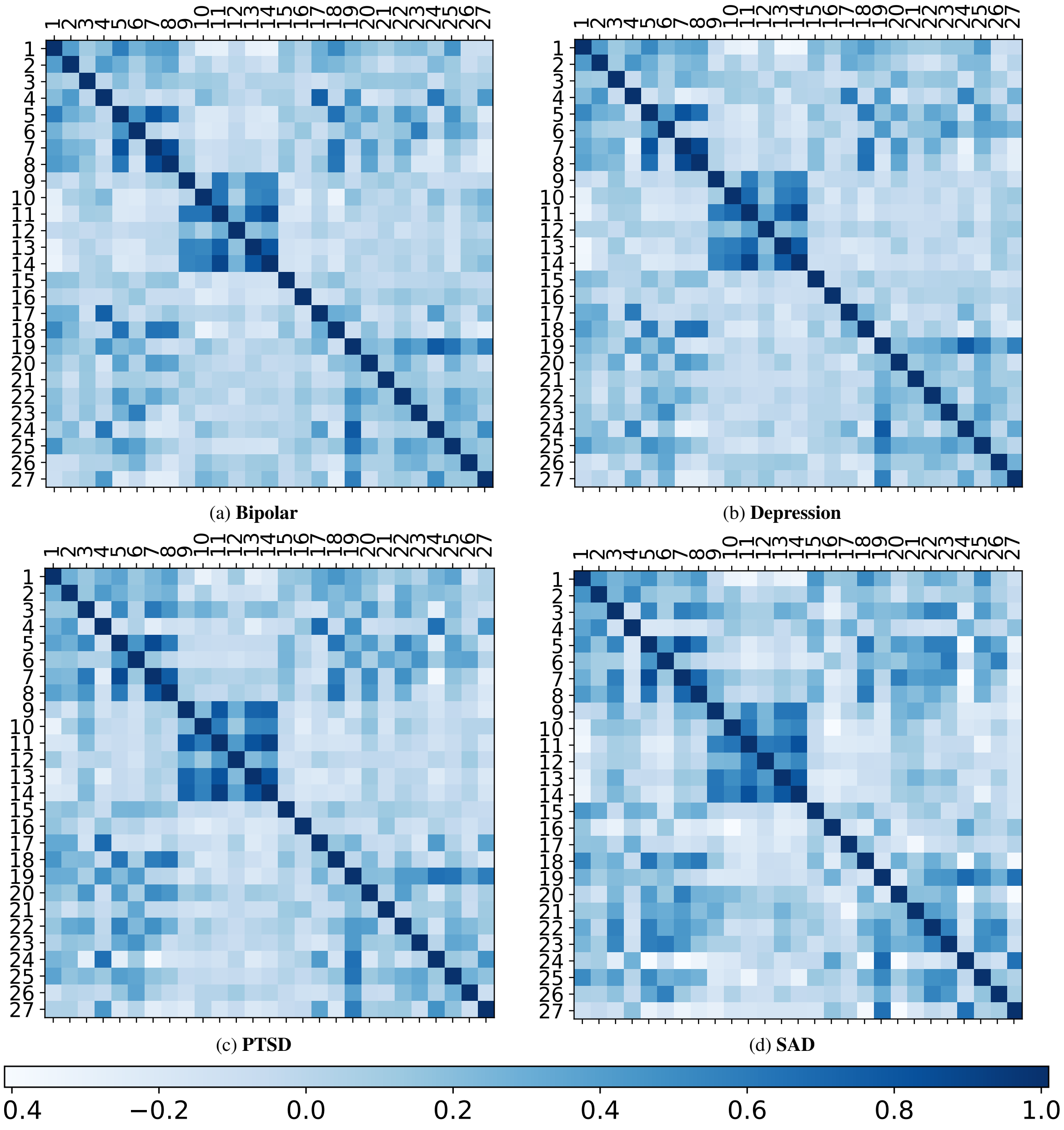


Figure 3: Feature correlation matrix for control group.

Inx	Feature	Inx	Feature
1	Pro1	2	Pro2
3	Pro3	4	pos.emo
5	neg.emo	6	anxiety
7	LIWC.anger	8	swear
9	tweet rate (daily)	10	@ propn.
11	@ count	12	self-@ count
13	unique @ count	14	frqnt. @ count
15	insomnia propn.	16	exercise propn.
17	pos.sentimt. propn.	18	neg. sentimt. propn.
19	Overall.score	20	EMOTIVE.Anger
21	Confusion	22	Disgust
23	Fear	24	Happiness
25	Sadness	26	Shame
27	Surprise		

Table 2: Feature reference

Except for SAD (Figure 4d), positive emotions (*pos.emo*) positively correlated with positive sentiment (*pos.sentimt.propn*); and with *Happiness*. For all conditions, negative sentiment (*neg.sentimt.propn*) positively correlated with negative emotions (*neg.emo*). Both of these negative features positively correlated with *LIWC.anger* and *swear*; and negative emotions (*neg.emo*) also positively correlated with *anxiety*, *Disgust* and *Fear*. *Fear* also appeared to have a positive correlation with *anxiety* but the relationship was stronger



(a) **Bipolar**

(b) **Depression**

(c) **PTSD**

(d) **SAD**

Figure 4: Feature correlation matrices for condition groups.

for the controls than for the mental health conditions. Overall, the POF features show little correlations with features from the other two feature sets, while the LIWC and EMOTIVE features can be observed to have notable correlations among each other. These correlations seem to indicate that the relationships of various emotions and sentiments are consistent among users who do or do not suffer from a mental heath condition.

However, there are a number of differences between the mental health conditions and the controls. For example, *Surprise* negatively correlated with a number of negative emotions (e.g. *EMOTIVE.Anger*), and positively correlated with *Happiness* across the four mental health conditions. There were no significant correlations between *Surprise* and other features for the controls. Whilst, first person pronouns “I” and “we” (*Pro1*) positively correlated with a number of negative and positive emotion features such as negative (*neg.sentimt.propn*) and positive sentiment (*pos.sentimt.propn*), *swear* and *LIWC.anger* across all conditions; it only positively correlated with *Sadness* for the condition groups and with *insomnia* for SAD. Furthermore, *Pro1* negatively correlated with the proportion, frequency and/or counts of *mentions* for those with a mental health condition, but not for the controls. These results suggest that users who suffer from a mental health condition are less likely to talk about themselves, but when they do, they are more likely to use words that indicate negative emotions. However, there are few significant correlations between *mentions* and negative emotions which may reflect the numbing of emotions [28]. Perhaps further insight into the dialogue structure is required, beyond simple emotion feature analysis.

Despite SAD being considered as a specifier for major depression or bipolar [20], there were a number of dissimilarities between SAD and these mental health conditions specifically and with all groups generally. This may reflect differences in the language used by the SAD group to refer to contexts and emotions. It may also reflect possible differences in the symptoms of SAD [39]. The dissimilarities between SAD and the other groups are consistent with the results of other social media research while using Twitter data [15].

For further development, using EMOTIVE to map the linguistics of emotional features over time could identify how mental health conditions fluctuate or change over time, and the context in which the emotions are generated. These emotion-based features could also be used to identify patterns of rumination (being preoccupied with the same situation giving rise to negative emotions) that typically occurs in depression. Emotion measures on social media could be a useful tool for examining the impact of interventions in how they change cognitions which in turn impact the generation and experience of negative emotions. For example, the most successful current depression treatment, cognitive-behavioral therapy [40], proposes that changes in cognition will lead to improvement of other symptoms of the disorder including negative emotions. Therefore, emotion detection systems like EMOTIVE could be used as a part of an intervention. In therapeutic interventions, a key goal is for those with a mental health condition to become aware of their emotions [41]. Increased emotional awareness is considered to be therapeutic as individuals are helped to make sense of what their emotion is telling them and to identify the goal, need, or concern that it is organizing them to attain [41]. Online emotion surveillance thus might be an innovative way to work with clients on how they express emotions through their usage of social media, which would worth exploring.

5. Conclusion

This work demonstrated again the relevance of the LIWC language features and Pattern of Life measurements for separating users with self-reported diagnosis from the control users for bipolar disorders, depression, PTSD and SAD, using a log-linear classifier, reported in prior work. For the first time, fine-grained emotions were employed as features for identifying mental health conditions from online social network. The high classification accuracy achieved by leveraging emotion-based features show that emotion expressions encode critical information about the mental states of Twitter users. We also presented that the best performance was reached when the emotion-based features, linguistic features and pattern of life measurements are combined. The various experiments performed in our study suggest that for the task of identifying mental health conditions, choosing suitable classification models for different feature sets, e.g DT and RF classifier for POF features, and suitable features for different conditions, e.g. EMOTIVE features for bipolar and depression, are necessary. Note that many mental health conditions have comorbidity, hence, distinguishing between these conditions is yet to be addressed in future works. Finally, further development of both features and classification techniques would be necessary in order to more accurately identify users who suffer from mental health conditions on social media.

6. References

[1] A. J. Ferrari, R. E. Norman, G. Freedman, A. J. Baxter, J. E. Pirkis, M. G. Harris, A. Page, E. Carnahan, L. Degenhardt, T. Vos, *et al.*, “The burden attributable to mental and substance use disorders as risk factors for suicide: findings from the global burden of disease study 2010,” *PLoS One*, vol. 9, no. 4, p. e91936, 2014.

[2] H. A. Whiteford, L. Degenhardt, J. Rehm, A. J. Baxter, A. J. Ferrari, H. E. Erskine, F. J. Charlson, R. E. Norman, A. D. Flaxman, N. Johns, *et al.*, “Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010,” *The Lancet*, vol. 382, no. 9904, pp. 1575–1586, 2013.

[3] WHO, “Mental disorders: Fact sheet.” World Health Organisation, Apr. 2016.

[4] WHO, “World health organisation,” 2016.

[5] CDC, “Suicide: Facts at a glance.” Centre for Disease Control and Prevention 1-800-CDC-INFO (232-4636), 2015.

[6] MHF, “Mental health foundation: Suicide.” Mental Health Foundation, 2016.

[7] D. E. Bloom, E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. R. Bloom, S. Fathima, A. B. Feigl, T. Gaziano, A. Hamandi, M. Mowafi, *et al.*, “The global economic burden of noncommunicable diseases,” tech. rep., Program on the Global Demography of Aging, 2012.

[8] M. De Choudhury, “Role of social media in tackling challenges in mental health,” in *Proceedings of the 2nd International Workshop on Socially-aware Multimedia*,

pp. 49–52, ACM, 2013.

[9] M. Conway and D. O'Connor, "Social media, big data, and mental health: current advances and ethical implications," *Current opinion in psychology*, vol. 9, pp. 77–82, 2016.

[10] O. Gruebner, M. Sykora, S. R. Lowe, K. Shankardass, L. Trinquart, T. Jackson, S. Subramanian, and S. Galea, "Mental health surveillance after the terrorist attacks in Paris," *The Lancet*, vol. 387, no. 10034, pp. 2195–2196, 2016.

[11] M. Park, C. Cha, and M. Cha, "Depressive moods of users portrayed in Twitter," in *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, pp. 1–8, 2012.

[12] T. E. Oxman, S. D. Rosenberg, and G. J. Tucker, "The language of paranoia.," *The American journal of psychiatry*, 1982.

[13] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," *NAACL HLT 2015*, p. 1, 2015.

[14] G. Coppersmith, C. Harman, and M. Dredze, "Measuring Post Traumatic Stress Disorder in Twitter.," in *ICWSM*, 2014.

[15] G. C. M. D. C. Harman, "Quantifying mental health signals in Twitter," *ACL 2014*, p. 51, 2014.

[16] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 47–56, ACM, 2013.

[17] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and emotion*, vol. 23, no. 2, pp. 209–237, 2009.

[18] P. Ekman, "Are there basic emotions?," *Psychological Review*, vol. 99, pp. 550–553, 1992.

[19] M. Sykora, T. Jackson, A. OBrien, S. Elayan, and A. von Lunen, "Twitter based analysis of public, fine-grained emotional reactions to significant events," *ECSM 2014 University of Brighton Brighton, UK 10-11 July 2014*, p. 540, 2014.

[20] APA, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[21] R. S. Lazarus and S. Folkman, *Stress, Appraisal, and Coping*. Springer Publishing Company, 1984.

[22] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press on Demand, 1991.

[23] K. R. Scherer, K. R. Scherer, and P. Ekman, "On the nature and function of emotion: A component process approach," *Approaches to emotion*, vol. 2293, p. 317, 1984.

[24] C. E. Izard, "Emotions and facial expressions: A perspective from differential emotions theory," *The psychology of facial expression*, pp. 57–77, 1997.

[25] J. R. Spoor and J. R. Kelly, "The evolutionary significance of affect in groups: Communication and group bonding," *Group processes & intergroup relations*, vol. 7, no. 4, pp. 398–412, 2004.

[26] N. S. Consedine and J. T. Moskowitz, "The role of discrete emotions in health outcomes: a critical review," *Applied and Preventive Psychology*, vol. 12, no. 2, pp. 59–75, 2007.

[27] R. J. Larsen and E. Diener, "Affect intensity as an individual difference characteristic: A review," *Journal of Research in personality*, vol. 21, no. 1, pp. 1–39, 1987.

[28] T. Scheff, "Emotions and depression, finding and facing intense emotions.." https://www.psychologytoday.com/blog/lets-connect/201107/emotions-and-depression-0, July 2011.

[29] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, p. 2001, 2001.

[30] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[31] L. Canales and P. Martínez-Barco, "Emotion detection from text: A survey," *Processing in the 5th Information Systems Research Working Days (JISIC 2014)*, p. 37, 2014.

[32] M. D. Sykora, T. Jackson, A. O'Brien, and S. Elayan, "Emotive ontology: Extracting fine-grained emotions from terse, informal messages," *International Journal on Computer Science and Information Systems*, vol. 8, no. 2, pp. 106–118, 2013.

[33] C. Chung and J. W. Pennebaker, "The psychological functions of function words," *Social communication*, pp. 343–359, 2007.

[34] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media.," in *ICWSM*, p. 2, 2013.

[35] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[38] N. Al Moubayed, T. Breckon, P. Matthews, and A. S. McGough, "Sms spam filtering using probabilistic topic modelling and stacked denoising autoencoder," in *International Conference on Artificial Neural Networks*, pp. 423–430, Springer, 2016.

[39] S. Harrison, "Emotional climates: Ritual, seasonality and affective disorders," *Journal of the Royal Anthropological Institute*, vol. 10, no. 3, pp. 583–602, 2004.

[40] A. T. Beck, *Depression: Clinical, experimental, and theoretical aspects*. University of Pennsylvania Press, 1967.

[41] L. S. Greenberg and A. Pascual-Leone, "Emotion in psychotherapy: A practice-friendly research review," *Journal of clinical psychology*, vol. 62, no. 5, pp. 611–630, 2006.