What about Mood Swings? Identifying Depression on Twitter with Temporal Measures of Emotions

Xuetong Chen
Loughborough University
Loughborough, United Kingdom
X.Chen5@lboro.ac.uk

Martin D. Sykora
Loughborough University
Loughborough, United Kingdom

Thomas W. Jackson
Loughborough University
Loughborough, United Kingdom

Suzanne Elayan
Loughborough University
Loughborough, United Kingdom

ABSTRACT

Depression is among the most commonly diagnosed mental disorders around the world. With the increasing popularity of online social network platforms and the advances in data science, more research efforts have been spent on understanding mental disorders through social media by analysing linguistic style, sentiment, online social networks and other activity traces. However, the role of basic emotions and their changes over time, have not yet been fully explored in extant work. In this paper, we proposed a novel approach for identifying users with or at risk of depression by incorporating measures of eight basic emotions as features from Twitter posts over time, including a temporal analysis of these features. The results showed that emotion-related expressions can reveal insights of individuals' psychological states and emotions measured from such expressions show predictive power of identifying depression on Twitter. We also demonstrated that the changes in an individual's emotions as measured over time bear additional information and can further improve the effectiveness of emotions as features, hence, improve the performance of our proposed model in this task.

ACM Reference Format:
Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about Mood Swings? Identifying Depression on Twitter with Temporal Measures of Emotions. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3191624

1 INTRODUCTION

Common mental disorders including depression, bipolar affective disorder, dementia and schizophrenia affect about 410 million people globally, among which depression alone affects around 350 million people, making it the world's fourth largest disease [1][2]. Depression, as one of the most prevalent forms of mental health problems, is associated with substantially increased morbidity and mortality [22, 24]. Major depressive disorder occurs within youth populations at comparable rates to adult populations [21], with nearly 25% of young people going to experience an episode of major depression by the age of 19 [31]. The World Health Organization estimates that depression is now the second leading cause of worldwide disability adjusted life years[3]. Despite the increasing knowledge and awareness, a considerable amount of individuals with depression remain undetected and untreated, leading to serious public health problems [18].

With the increasing engagement with social media of the public, many studies showed that social media has already been increasingly used in population health monitoring [6], and is beginning to be used for mental health applications [16]. Employing social media has been suggested to be beneficial to mental health studies, as it provides an unbiased collection of individuals' language usages and behaviours [11]. Additionally, information from social media bears the potential to complement traditional survey techniques in its ability to provide finer grained measurements of behaviour over time while dramatically expanding population sample sizes [11]. Initial evidence has been found to show that people do post about their depression and their treatments on social media [27]. And numerous studies have presented that based on the symptoms and indicators of depression, it is possible to use data mining and machine learning techniques to develop models to discover likely instances of depression on social media.

In this study, we investigated the potential of both non-temporal and temporal measures of emotions from Twitter posts over time in identifying users with depression over a control group (users who do not suffer from depression). We detected eight basic emotions (e.g. anger, fear, etc.) from each tweet as its message content attributes and calculated the strength scores based on the intensity of each expression. These attributes and strength scores were used first to calculate the average intensity of each emotion expression from each user's past tweets, and subsequently for a time series analysis of each user, creating two feature sets to build classifiers that label Twitter users as either belonging to the depression or non-depression (control) groups. The results show that by leveraging the averaged intensity of emotional expressions, our classifier was able to outperform the baseline and other prediction models in identifying users with depression. And the accuracy improved further when employing the descriptive statistics of the emotion time series as inputs to the training process. Hence, this suggests

[1] http://www.who.int/mediacentre/factsheets/fs396/en/
[2] http://www.who.int/en/

[3] http://www.who.int/mediacentre/factsheets/fs369/en/

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191624

that emotion related expressions from tweets can reveal insights of individuals' psychological states and that the changes of an individual's emotions over time bear additional information with promising potential in identifying users with depression. The main contributions of this work is, for the prediction and detection of mental health conditions on social media, we (1) employed emotions as features, (2) applied time series analysis, and (3) explored the effectiveness of temporal measures of these features for the task of identifying users who suffer from depression on social media.

2 RELATED WORKS

During the past decade, there have been an increasing number of studies investigating mental health issues using social media datasets. These studies explored features based on known symptoms and indicators of mental disorders and showed that it is possible to develop automatic detection systems for mental health problems using existing tools and methods from data mining, text mining, social network analysis and machine learning. A wide range of mental health conditions have been studied including major depressive disorder [5, 20], post traumatic stress disorder [9], ADHD and schizophrenia [25], anxiety disorder and OCD [8], borderline personality disorder and bipolar disorder [33], seasonal affective disorder [5, 7], suicide [13], eating disorders [38], sleep disorder [19], and others [8].

Including identifying depression from Twitter, the majority of these studies focused on the analysis of textual contents in the English language from publicly available data sources on social media. A few features were most frequently used for understanding individuals' psychological states. Linguistic patterns, often obtained by using the well-known Linguistic Inquiry and Word Count (LIWC) tool [29], were employed to extract potential signals from textual content such as first, second, third person pronouns, perceptual process related words, or positive and negative emotion words [38]. Sentiment analysis using tools like OpinionFinder [39], SentiStrength [37] and Affective Norms for English Words (ANEW) [3], were also frequently interoperated for quantifying the sentiment and emotion attributes from textual expressions [20]. Besides, emoticons and images have also been utilized for detecting positive and negative sentiments from a social media post [20]. As a part of content analysis, various types of topic modelling such as LDA and TAM were incorporated in order to extract topics, sometimes more specifically ailment related topics [28], from user generated contents [34]. Most social media platforms also provide interactive features, which allow users to follow or un-follow another user, add or remove a friend, mention, reply to a user or a post, repost, and comment. Due to these interactive features, network analysis was applied in some studies to understand users' online social activities, their relationships, and interactions with others [38]. The influences of personality, age and gender on the disclosure of mental health problems were also examined [30]. These attributes were utilised for creating a matched control group of similar age, gender according to the users included in the condition group, in order to reduce this biases for the analysis of the differences in features from the condition group compared to the control [23].

Prediction models are often used to perform this task with a selection of extracted features from the above-mentioned analysis to learn patterns from the data. To the best of our knowledge, all prediction models used supervised learning techniques, where the sample data contain labels for both the inputs as training and the outputs as true labels for evaluating the prediction performance.

In spite of these research efforts, little attention has been given to the emotional factors for understanding or analysing specific mental disorders on social media. For the task of detecting depression as well as other mental disorders, discrete basic emotions have not yet been leveraged, and their predictive power has not yet been fully examined. As many psychological phenomena occur in small time windows, *affective micropatterns* in language have recently been explored for quantifying mental health signals from Twitter [23], considering the temporal dimension for the first time in mental health research on social media. Similar to our work, the authors also suggested the study of emotions to be an important avenue for further work in this research field. Therefore, with the present work we aim to (1) expand the scope of sentiment analysis for mental disorder detection, using measurements of eight basic emotions expressed through Twitter posts as features; (2) capture the changes of individuals' emotions over time, by applying time series analysis on the measurements of emotions to produce a set of temporal features; (3) examine and demonstrate the effectiveness of these features for identifying users suffering from depression.

3 DATA

Among social media platforms, Twitter provides a unique source of big data for public health research due to the real-time nature of the content, and ease in accessing and retrieving publicly available information [35]. The increasing amount of research and technique development on Twitter related to text mining, sentiment analysis, public health surveillance and prediction highlights the significance of this stream of social media data. However, there is not yet a gold standard dataset available for mental health related research on social media, therefore data collection was performed for the purpose of this study.

3.1 Data Collection

In general, there are two broad approaches for social media data collection in the literature: (1) employing means like surveys, crowdsourcing to attract participants and collect data from their social media account directly with their consent; (2) using available Application Programming Interface (API) of social media platforms to extract and aggregate relevant data from public posts. In this study, we chose the second approach to pool public posts from Twitter.

First, in order to identify a group of users who suffer from depression. We first collected self-reported diagnosis tweets using Twitter streaming API with the regular expression "I was/have been diagnosed with depression", following the Twitter data acquisition process described in work [7]. The duration of the collection process lasted four months: from November 18th 2016 to February 15th 2017. These diagnosis tweets were not formally analysed for filtering out disingenuous statements. However, all retweets were removed, since they are often an indication of the message being a quotation of others' post, which is not originally produced by the user tweeting, hence, analysing their user profiles would be misguided. For instance, "*RT @User_screenname: I was diagnosed*

with depression before they knew what gender I was. Url_links". We considered these diagnosis statements as a qualification to obtain a list of user *screen names* of whom are likely having a genuine depression disorder, and formed the depression group of this study. In a similar manner, we collected one day (February 20th 2017) of tweets containing the keyword "the" using Twitter Streaming API to collect a group of users who can represent the general public. The obtained user *screen names* were double checked against the depression group to make sure the two groups have no overlap that may interfere with the training process later on. We then used this list of users as the control group.

Next, for individuals in the depression group, all tweets up to one year before the diagnosis tweets were posted were retrieved to capture the user generated contents, which might contain information and patterns of the user's depressive state. The past tweets of users from the control group were collected in the same manner to match the depression dataset. In this process, no private messages or user accounts were accessible to the researchers, and all collected data were publicly posted on Twitter. Users who had less than 50 tweets, or often post in non-English languages do not meet the requirements for the analysis in this study, therefore, they were filtered out. After this process, we obtained 585 and 6,596 unique and valid users with their past tweets (in average 2,000 per user) as the depression and the control group datasets, respectively. However, we only chose a random sample of 600 users from the control group to create a balanced dataset with roughly equal number of positive (depression) and negative (non-depression) instances for the later classification experiments.

This data collection method has been previously validated by various studies of depression and other mental health conditions through replication of previous findings and showing predictive power for real-world phenomena [7, 8, 10].

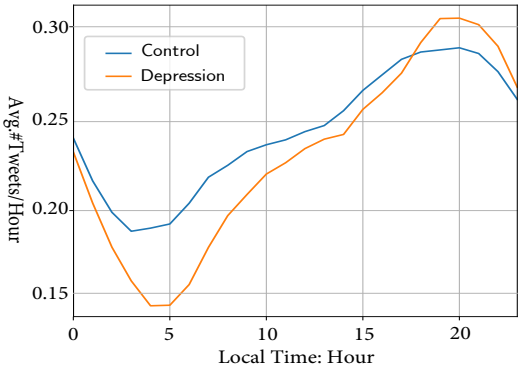

Figure 1: Mean number of tweets posted throughout the day for the depression and the control group.

Pattern of Postings

We validated the resulting datasets by examining the posting patterns of users from both groups. Figure 1 displays the daily posting distribution measured as the average number of tweets per hour (in local time) over the entire duration of the collection of users' past tweets from the depression and the control group respectively.

In general, both groups show an increase of tweet volume from 5 am to 8 pm, and a decrease after 8 pm, which align with the daily activity pattern of the general public. Around 3 am, the average number of tweets reaches the lowest for both groups, when people are most likely to be in sleep. Evenings and early nights show peaks, indicating that people tend to tweet more at the end of the day after finishing work or school. Although with a similar pattern, it can be observed that the depression group shows a higher peak in the evening (after 6pm), and a lower tweet volume during the day. This suggested that users with depression tend to have more night time online activities and postings.These patterns are in conformity with the diurnal patterns of posting presented in [12].

4 METHODOLOGY

Given the depression and the control datasets, we first detected emotions and measured the expression intensity for each emotion as strength scores from our collected tweets by applying an emotion sensor. These scores were then used to conduct a non-temporal and a temporal feature set for the task of separating users of the depression group from the control.

4.1 Measuring Emotions

Emotions are an important element of human nature, thus they have been widely studied in many research areas such as neuroscience, psychology and behavioural sciences [4]. In particular, numerous 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 [7]. However, emotion-based features have not yet been considered nor incorporated in the analysis of mental health related social media datasets. Therefore, we propose to employ discrete emotions in the task of identifying depression.

For this study, we considered measures of eight well recognised basic emotions: *Anger, Disgust, Fear, Happiness, Sadness and Surprise*, also known as Ekman's basic emotions [15], *shame*, which was tentatively included into the list of basic primary emotions, in Ekman's later work based on emerging evidence [14], and *Confusion*. Although confusion has traits that are often associated with emotions (e.g. specific distinct facial expressions), it is nevertheless, in the emotion research literature, mostly considered to be a state rather than an emotion, similar to concentration or worry [32]. Nonetheless, Rozin and Cohen (ibid.) pointed out that given its clear negative valence, confusion could arguably be considered as an emotional affect, although it is perhaps under more voluntary control than the standard emotions. They (ibid.) further suggested that confusion was very common and almost unstudied. Given the high frequency of confusion expression and the potential discriminatory value to general emotional well-being, *Confusion* was therefore included in this study. Along with the eight basic emotions, *Emotion Overall Score*, which is a sum of strength scores of all emotions, was also included as a measure of overall emotionality (i.e. emotion activation).

To detect and measure these fine-grained emotions from individuals' tweets, we employed the EMOTIVE system, an ontology (semantic model) based advanced sentiment algorithm, developed by [36]. It is a map of emotion related words and phrases including a

set of intensifiers, conjunctions, negators, interjections, and linguistic analysis rules, i.e. EMOTIVE ontology. This ontology, therefore, allows a richer semantic representation than the traditional lexicon and discovers emotions with their expression intensities as strength scores. The system first parses the text and classifies part-of-speech tags through a Natural Language Processing pipeline. Emotion related expressions are then matched by comparing the parsed words against the EMOTIVE ontology. A strength score is produced by accumulating the intensity measures of matched intensifiers for each detected emotion expression. And a strength score of zero is given to the rest of the emotions indicating there is no expression intensity of the emotion, i.e. the emotion is not expressed in the post. The EMOTIVE system was evaluated and compared against other benchmarks in [36], in which it showed a 0.962 f-measure in detecting emotions from Twitter posts, making EMOTIVE to be a highly suitable tool for our study.

4.2 Constructing Feature Set

This work aims to investigate the effectiveness of emotion based features and their temporal dimension for identifying depression from Twitter. Therefore, two types of features were created from the emotion measurements. We (1) calculated the overall intensity (strength score) of the emotions extracted from all past tweets of each user, and (2) created a time series for each emotion of every user to generate a selection of descriptive statistics for these time series. These two sets of features were subsequently employed as inputs for the task of identifying users with depression condition.

Non-temporal Feature Set

First, we aggregate individuals' all tweeting history as text documents for the emotion sensor to produce a list of emotion strength scores for each user, creating 9 emotion expression measurements as described in the previous section. Hence, for each user we obtained an emotion feature vector with 9 entries, i.e. *Emotion Overall Score, Anger, Disgust, Fear, Happiness, Sadness, Surprise, Shame, Confusion*. The resulting emotion feature vectors of all users consequently formed the non-temporal emotion feature set, which will be referred to as **EMO** for the rest of the paper.

Temporal Feature Set

In order to capture the hidden information carried by emotion expressions over time, we conducted time series analysis on the strength scores of these expressions. We first obtained 9 emotion strength scores for each tweet using the emotion sensor. To each strength score generated by the emotion sensor, we assigned a time stamp as the posting time of the tweet, from which the score was calculated from. Therefore, from each user we obtain a sequence of measurements as a signal over time for each emotion expression. These sequences were then used to create nine emotion time series (i.e. *Emotion Overall Score, Anger, Disgust, Fear, Happiness, Sadness, Surprise, Shame, Confusion*) for every user. Each time series was created by the summation of the accumulated expression strength scores with the same date. This is due to the reason that most users do not tweet several times a day and only around 10% of the tweets contain emotion expressions. Hence, using one day as the unit of the emotion time series is more suitable than using the actual tweet time and is able to capture a more continuous signal. Note that a score of zero was assigned to the date when no tweet was found, or when no emotion expression was detected.

Next, a selection of descriptive statistics were calculated as temporal features for each time series. Given a time series $X_1, X_2, \ldots, X_n$ these statistics are defined as follows:

- **Mean**: the average measure of an emotion signal over the entire period of analysis.
- **Standard Deviation**: measures the variation of an emotion signal over the entire period of analysis.
- **Entropy**: measures the amount of regularity and the uncertainty of fluctuations over an emotion signal.

$$EN = -\sum_{t=1}^{n} X_t \log X_t$$

- **Mean Momentum**: Momentum is the change in an m-day simple moving average (SMA) between two days, with a scale factor m+1, defined as:

$$MTM = (m+1) \times (SMA_{day_{i-1}} - SAM_{day_i})$$

where $i \in [m+1, n]$, and a simple moving average (SMA) is the unweighted mean of the previous m data, given as:

$$SMA = \frac{1}{n} \sum_{t=1}^{m} X_t$$

Hence, momentum measures the changing rate of the simple moving average of the time series signal of an m-day time window, and the mean momentum, consequently, is the average of these rates, which measures an over all trend of the emotion signals, calculated as:

$$\overline{MTM} = \frac{MTM}{n-(m-1)}$$

Where $m = 14$ (days), due to the reason that a diagnosis of major depressive episode requires the patient to have over a two-week period in experiencing a number of symptoms including changes in moods (i.e. emotions in our context), given by [2].

- **Mean Differencing**: Differencing is a transformation applied to time-series data in order to eliminate trend and seasonality. Due to the same reason, we decided $m = 14(days)$ and calculated the second order differencing of the simple moving average of the m-day time windows to capture the emotion differences in between every two-week period, given as

$$DIF^* = (SMA_{day_i} - SMA_{day_{i-1}}) - (SMA_{day_{i-1}} - SMA_{day_{i-2}})$$

where $i \in [m+2, n]$. $\sigma(DIF)$, hence, is the the standard deviation of the transformed differencing series, which measures the trending rate of an emotion signal.

Consequently, these five statistics were used as the temporal features for each emotion expression, resulting a temporal emotion feature vector with 45 entries (9×5) for each user. These feature vectors were then used to form the temporal emotion feature set and we will refer to it as **EMO_TS** throughout the rest of the paper.

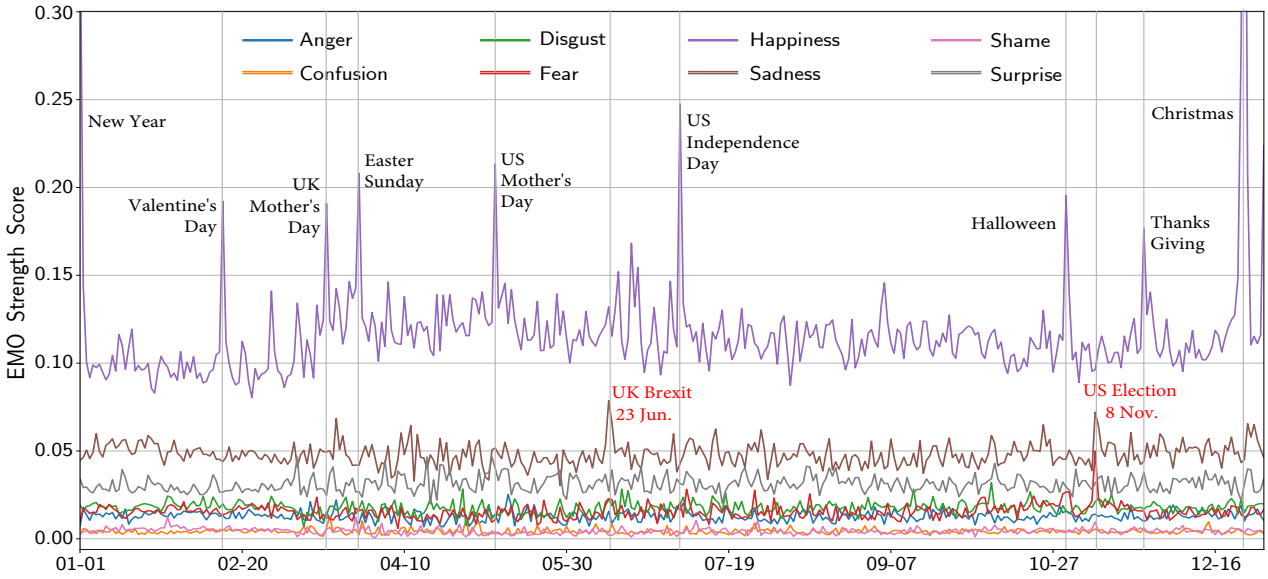

Figure 2: Public emotions form tweets posted in 2016.

4.3 Prediction Framework

In order to examine the predictive power of emotions, we leveraged the non-temporal and the temporal feature sets for a binary classification task to predict the self-reported diagnosis of depression. We also replicated the LIWC feature set as described in [7] on our own dataset to provide a comparison of approaches on the same task with the same data. Note that the LIWC was not the best performing feature set for separating users with depression from the control group in the original work [7]. However, the reason we chose this feature set was that it is (1) easily replicable (2) with all features extracted from a well-known and widely used software, LIWC [29], which provides reliable unbiased measures, and (3) has been validated and proven to be effective for mental disorder detections by several studies [7, 8, 12].

We first experimented the task with the EMO, LIWC, and the combined feature sets, EMO+LIWC. A range of popular machine learning classifiers was incorporated in order to find the the most suitable one for further exploration. These classifiers are the Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayesian (NB), Decision Trees (DT) and the Random Forests (RF). We then performed the same classification task using the temporal feature sets, EMO_TS, with the best performing classifiers from the previous experimentation to discover the effects of the temporal dimension of the emotions. 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 performances of this task were evaluated using the classification accuracy through leave-one-out cross validation, precision, recall and F-score.

5 RESULTS

We first validated the ability of EMOTIVE to capture various emotion aspects of the public. To do so, we aggregate all users from the control group and calculated the mean emotion strength scores for every day in 2016. The resulting expression intensity distributions of the the eight emotions over the year are shown in Figure 2. Despite the sparsity of emotions expressed through Twitter, the figure shows that the EMOTIVE has successfully identified the public emotional response to festivals and important events happened in 2016 as annotated on the graph. *Happiness* peaks with several well known festivals in English speaking countries, accompanied by a decrease of *Sadness*. For festivals that involve giving and receiving gifts, a raise of the *Surprise* signal can be observed shortly before and or after the day, such as Valentine's Day, Mother's Day and Christmas. Two major events have also been captured by the EMOTIVE, on 23 June and 8 November when the Brexit and the US Presidential Election took place. Both events caused a significant increase of *Sadness*. Besides, peaks of *Disgust* can be noticed in the next couple days of the Brexit. And the US Election results seems to have caused a considerable raise of *Fear*.

In Figure 4, we compared each of these emotion signals (from the control group coloured in blue) against the depression group (in orange). As we can see, except *Happiness* and *Surprise*, all emotion signals have more and higher peaks from the depression group than from the control group, which suggests that users who suffer from depression express these emotions more frequently and intensely. An overall higher expression intensity from the depression group can be observed in *Sadness*, *Disgust* and *Fear*. For the depression group, Mother's and Father's Day seem to trigger a stronger signal in *Shame*. Studies have shown that feelings of shame and guilt are factors associated with depression [17]. While guilt is a feeling of doing wrong, *Shame* is a feeling of being wrong. Hence, *Shame*, in contrast to guilt, elicits rumination, which then leads to depression [1, 26]. Overall, an increase of negative emotions including *Anger*, *Disgust*, *Fear*, *Sadness*, and *Shame* can be observed from the depression group compare to the control group. These insights are in line

with previous findings that depressed users use negative emotion words and anger words on Twitter more frequently, while there are no notable differences in positive emotion word usage [7, 12, 27].

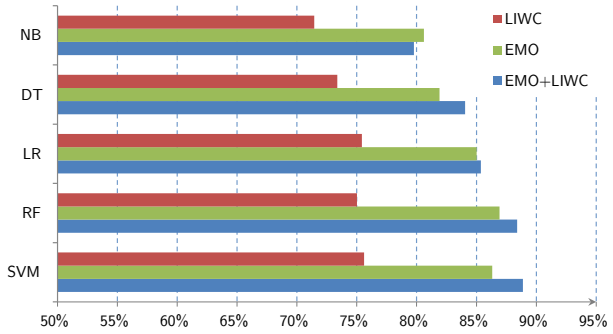

Figure 3: Prediction accuracy in the task of separating users with depression from controls, evaluated by leave-one-out cross validation.

5.1 Predicting with Emotions

Following our proposed framework, the initial prediction experiment was conducted as a binary classification task in separating users with self-reported depression from the control users. This experiment was designed to examine the predictive power of the non-temporal emotion features, and to discover a suitable machine learning classifier for further experimentation. The selected machine learning classifiers are the LR, SVM, NB, DT and RF, for each classifier parameters was optimised with random grid search CV using the training data.

Figure 3 presents the accuracy results of a leave-one-out cross validation of each classier on this prediction task using three feature sets, the EMO, LIWC, and EMO+LIWC. Among all experiment results, a precision of 0.719 was obtained from the depression detection performance using the LR classifier with the LIWC feature set, which is slightly higher than the result reported in [7] (0.68). This could be due to the differences between datasets and feature normalisation methods, since the data preprocessing steps were not described in the work of [7]. However, the difference of the performances is not significant with the same LIWC features produced from different datasets. Hence the replicated LIWC feature set is valid and its effect in performance could be considered as a valid baseline for comparison with our proposed approach.

For all algorithms, the EMO feature set appeared to be more effective than the LIWC, improving the accuracy by nearly 10%, which indicated that emotion related features are more relevant and straight forward in capturing depressive patterns than basic linguistic features. The combined feature set EMO+LIWC achieved the best prediction performance, by a slight advantage in accuracy (in average 1.63%) than the EMO feature set, except for the NB classifier. This suggests that there might be an overlap of the information contained in the two feature sets, and more is captured by the EMO feature set. In all cases, by employing the emotion features (EMO) the prediction models outperformed the chance classification baseline (50%), and all prediction accuracies reached above 80%, which also outperformed the prediction models in [8], precision of 0.48 at 10% false alarm for depression specifically, and [12], mean prediction accuracy of 68.42%. Across classifiers, the performances appear to be consistent for each feature set. The SVM classifiers achieved the highest classification accuracy on the LIWC (75.61%) and EMO+LIWC (88.88%), while the RF classifier achieved the best performance on the EMO (87.27%) feature set. However, the differences of the performances of these two classifiers were minor. And considered that the RF is a non-linear classifier while the SVM is linear, we decided to experiment with both classifiers for the following prediction task with the temporal feature sets.

5.2 Predicting with Temporal Features

We have discovered that emotions carry predictive information that can reveal users' depression state. In this section, we investigated whether the temporal features can provide additional information and improve the prediction model even further. Again, we performed the binary classification task to separate users with self-reported depression from the control users with the temporal feature set (EMO_TS) using the best performing linear (SVM) and non-linear (RF) classifier from the previous experiment, with optimised parameters.

Support Vector Machines					
Feature Set	**75/25 Acc.**	**Prec.**	**Recall**	**F**	**LOO.CV**
EMO	86.90%	0.864	0.830	0.844	86.32%
EMO_TS	79.67%	0.855	0.721	0.739	80.62%
EMO+EMO_TS	88.52%	0.910	0.851	0.869	88.87%
LIWC	77.62%	0.774	0.730	0.741	75.61%
LIWC+EMO	87.62%	0.887	0.835	0.853	88.88%
LIWC+EMO_TS	87.56%	0.846	0.850	0.848	85.47%
ALL	89.71%	0.917	0.871	0.886	89.83%

Random Forests					
Feature Set	**75/25 Acc.**	**Prec.**	**Recall**	**F**	**LOO.CV**
EMO	87.38%	0.869	0.860	0.864	87.27%
EMO_TS	89.71%	0.908	0.865	0.881	89.77%
EMO+EMO_TS	92.82%	0.935	0.904	0.917	91.81%
LIWC	75.95%	0.733	0.713	0.720	75.56%
LIWC+EMO	89.52%	0.905	0.852	0.872	87.51%
LIWC+EMO_TS	92.82%	0.935	0.908	0.920	90.31%
ALL	93.06%	0.944	0.901	0.918	92.17%

Table 1: Performance with different feature sets for predicting depression or non-depression classes of Users. Measures 75/25 split classification accuracy, precision, recall, f-score, leave-one-out cross validation.

RF classifier performed the best with *n_estimator=10* (the number of estimated trees in the forest), and *max_depth = 6* (the maximum tree depth) on our feature set. And the SVM classifier performed the best with a RBF kernel function. Table 1 presents the performance of the classification using different feature sets with evaluation measures of respectively the SVM and the RF classifier.

For the SVM classifier, using the temporal feature set EMO_TS did not produce better performance than EMO with approximately 6% reduce in accuracy, but still outperformed the LIWC feature set by 5% with the leave-one-out cross validation. For the RF classifier

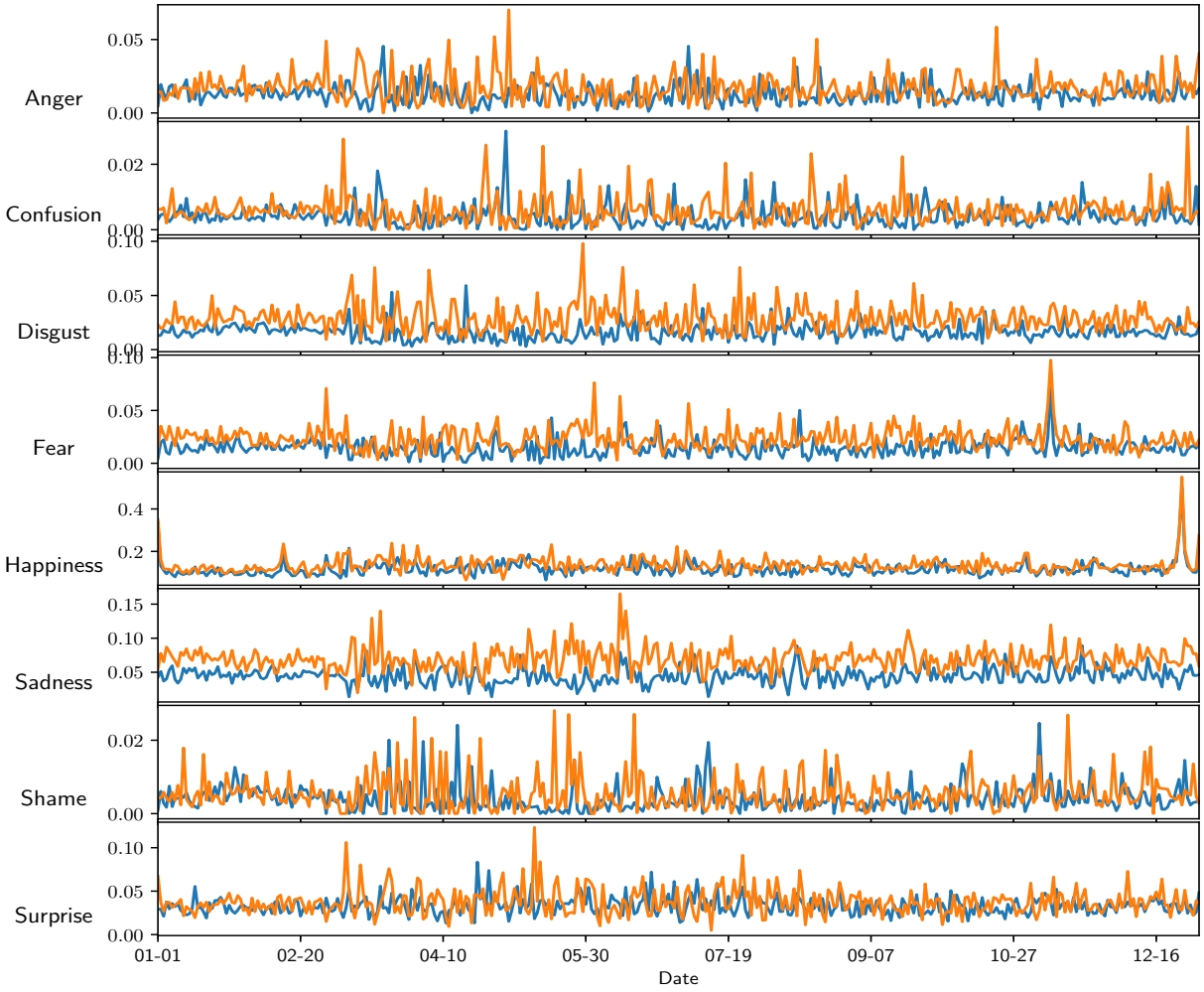

Figure 4: Averaged emotion intensity in 2016 for the depression (in orange) and control (in blue) group.

on the other hand, EMO_TS increased 2% accuracy compared to EMO, and nearly 10% compared to the SVM with EMO_TS. The best performance resulted by the temporal emotion features alone (EMO_TS) was achieved by the RF classifier with 89.77% accuracy of a leave-one-out cross validation. This might indicates that the RF classifier is able to leverage more information from the temporal features and perform better with greater number of features.

For both classifiers, when combined with EMO_TS, the combined feature sets appeared to be more effective and improve the classification accuracy, and the best performances were achieved when using the combination of all three feature sets 89.83% and 92.17% for the SVM and the RF classifiers respectively. The accuracies of using emotions improved by 8% for the SVM and 3%for the RF classifier by adding the temporal measures to the non-temporal emotion features (i.e. using EMO+EMO_TS). These improvements show that the temporal measures of emotions can capture more information in addition to the average intensities of emotions and that time series analysis and temporal measures of emotions over time can improve the effectiveness of emotion features on this prediction task. To extend this work, it would be of great interest to measure other types of features over time, such as linguistic style, pattern of life, and online social activities, to capture more meaningful patterns that can provide more detailed insights into depression as well as other types of mental disorders from social media.

6 CONCLUSION

This work is the first to employ fine-grained emotions for identifying mental disorders, and to implement time series analysis for the detection of mental health condition on social media. It demonstrated the potential of using discrete emotions as features and their temporal measurements for predicting depression in individuals. We first extracted emotions with their expression intensities as strength scores to create emotion features. A time series analysis was applied to the emotion strength scores over time and produced a selection of descriptive statistics as temporal features. By incorporating the emotion features, our model outperformed the baseline and other prediction models in [7, 8, 12] with 87.27% classification

accuracy using only emotion features (EMO). Moreover, by employing a variety of temporal features of emotions over time the prediction accuracy improved even further, achieved 89.77% using the temporal features (EMO_TS), and 91.81% when our proposed emotion (non-temporal) and temporal feature sets were combined (EMO+EMO_TS). In summary, these results showed that basic emotions provide considerable insights in identifying twitter users who suffer from depression. Besides, additional information can be discovered by analysing these features over time. After learning the traces and patterns of depressed users from these features, the trained classifiers can be easily applied for detecting Twitter users with depression who did not post about their conditions and users who are at risk of depression. However, more training, testing data and in depth evaluation are required before any experimentation in real life. These findings provide a roadmap for future work in the research of mental health on social media. There is still significant work required to discover the value and meaning of these emotions and their temporal features in terms of psychological understanding and intervention. However, fine-grained emotions and the dynamics of their temporal variables is certainly worth more in depth exploration for our future studies.

REFERENCES

[1] Barbara Alexander, Chris R Brewin, Simon Vearnals, Geoffrey Wolff, and Julian Leff. 1999. An investigation of shame and guilt in a depressed sample. *Psychology and Psychotherapy: Theory, Research and Practice* 72, 3 (1999), 323–338.

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

[3] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Technical report C-1, the center for research in psychophysiology, University of Florida.

[4] Lea Canales and Patricio Martínez-Barco. 2014. Emotion Detection from text: A Survey. *Processing in the 5th Information Systems Research Working Days (JISIC 2014)* (2014), 37.

[5] Xuetong Chen, Martin Sykora, Thomas Jackson, Suzanne Elayan, and Fehmidah Munir. 2018. Tweeting Your Mental Health: an Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[6] Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology* 9 (2016), 77–82.

[7] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.

[8] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. *NAACL HLT 2015* (2015), 1.

[9] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter.. In *CLPsych@ HLT-NAACL*. 31–39.

[10] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring Post Traumatic Stress Disorder in Twitter.. In *ICWSM*.

[11] Munmun De Choudhury. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd International Workshop on Socially-aware Multimedia*. ACM, 49–52.

[12] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media.. In *ICWSM*. 2.

[13] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2098–2110.

[14] Paul Ekman and Daniel Cordaro. 2011. What is meant by calling emotions basic. *Emotion Review* 3, 4 (2011), 364–370.

[15] Paul Ed Ekman and Richard J Davidson. 1994. *The nature of emotion: Fundamental questions*. Oxford University Press.

[16] Oliver Gruebner, Martin Sykora, Sarah R Lowe, Ketan Shankardass, Ludovic Trinquart, Tom Jackson, SV Subramanian, and Sandro Galea. 2016. Mental health surveillance after the terrorist attacks in Paris. *The Lancet* 387, 10034 (2016), 2195–2196.

[17] David W Harder, Lisa Cutler, and Liesl Rockart. 1992. Assessment of shame and guilt and their relationships to psychopathology. *Journal of personality assessment* 59, 3 (1992), 584–604.

[18] Thomas K Houston, Lisa A Cooper, Hong Thi Vu, Joel Kahn, Janice Toser, and Daniel E Ford. 2001. Screening the public for depression through the Internet. *Psychiatric services* 52, 3 (2001), 362–367.

[19] Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1501–1510.

[20] Keumhee Kang, Chanhee Yoon, and Eun Yi Kim. 2016. Identifying depressive users in Twitter using multimodal analysis. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*. IEEE, 231–238.

[21] Ronald C Kessler, Howard Birnbaum, Evelyn Bromet, Irving Hwang, Nancy Sampson, and Victoria Shahly. 2010. Age differences in major depression: results from the National Comorbidity Survey Replication (NCS-R). *Psychological medicine* 40, 2 (2010), 225–237.

[22] Ronald C Kessler, Evelyn J Bromet, Peter de Jonge, Victoria Shahly, and Marsha Wilcox. 2017. The Burden of Depressive Illness. *Public Health Perspectives on Depressive Disorders* (2017), 40.

[23] Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language. *CLPsych 2017* (2017), 85.

[24] Colin D Mathers and Dejan Loncar. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine* 3, 11 (2006), e442.

[25] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the Language of Schizophrenia in Social Media.. In *CLPsych@ HLT-NAACL*. 11–20.

[26] Ulrich Orth, Matthias Berking, and Simone Burkhardt. 2006. Self-conscious emotions and depression: Rumination explains why shame but not guilt is maladaptive. *Personality and social psychology bulletin* 32, 12 (2006), 1608–1619.

[27] Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*. 1–8.

[28] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsm* 20 (2011), 265–272.

[29] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

[30] Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age and gender in tweeting about mental illnesses. In *NAACL HLT*, Vol. 2015. 21.

[31] Paul Rohde, Peter M Lewinsohn, Daniel N Klein, John R Seeley, and Jeff M Gau. 2013. Key characteristics of major depressive disorder occurring in childhood, adolescence, emerging adulthood, and adulthood. *Clinical Psychological Science* 1, 1 (2013), 41–53.

[32] Paul Rozin and Adam B Cohen. 2003. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion* 3, 1 (2003), 68.

[33] Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo, and Yi-Shin Chen. 2016. MIDAS: Mental illness detection and analysis via social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 1418–1421.

[34] H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, et al. 2016. Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*. 516–527.

[35] Lauren Sinnenberg, Alison M Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M Merchant. 2017. Twitter as a tool for health research: a systematic review. *American Journal of Public Health (ajph)* (2017).

[36] Martin D Sykora, Thomas Jackson, Ann O'Brien, and Suzanne Elayan. 2013. Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *International Journal on Computer Science and Information Systems* 8, 2 (2013), 106–118.

[37] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology* 61, 12 (2010), 2544–2558.

[38] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 91–100.

[39] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*. Association for Computational Linguistics, 34–35.