International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Crowd-Based Profiling : A Framework to Detect Psychological Disorders in Social Media Users

A.Sharmila Agnal, Akshay Kannan V, Durga S, Nishanth Kumar. S, Dheeraj. R

Abstract- Psychological disorders are presently striking a large number of population from various civilization, society, occupation and different locations around the world. The main obstacle of psychological disorders is the difficulty to detect on people suffering from these disorders, hence resulting in introducing a worrying amount of undetectable cases and false detection issue. Our methodology aims at constructing detective models to identify psychological disorders among online social media users. These detective models are attainable by engaging a basic data collection process formulated as crowd based profiling, which assists us to collect accurate and more efficient data set of people from various categories. Our experiment proposes that obtaining specific English language patterns and socializing attributes from data sets paves the way to deal with advanced experiments on psychological disorders.

Index Terms: Crowd based profiling, Data sets, Online social media, Psychological disorders detection, Sentiment analysis.

I. INTRODUCTION

People who suffer from psychological disorders seem to have minimal contact with the people who are personally known to them. This makes them express their thoughts, feelings through online social media. Twitter is commonly used by everyone in the world as it allows them to commune their ideas and views to the public. People suffering from psychological disorders find Twitter as the perfect platform for them as it has various community groups [1] where they can discuss their problem and the difficulties they are going through and from which they believe they could get help from. By sharing information regarding the problems they face each day, they provide enormous content subliminally, and with the behaviour, their stability could also be measured. By using this information as input we could construct a model to detect Psychological disorders.

Revised Manuscript Received on 30 March 2019.
* Correspondence Author
A Sharmila Agnal*, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
Akshay Kannan V, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
Durga S, Department of CSE, SRM Institute of Science and Technology, Tamil Nadu, India.
Nishanth Kumar S, Department of CSE, SRM Institute of Science and Technology, Vellore, Tamil Nadu, India
Dheeraj R, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

The collecting of untapped data is referred to as crowd based profiling which is a practical data collection method used to gather data and to develop an efficient set of semantic and behavioural patterns [3].This type of detective models might help to construct an advance mechanism to reduce the numbers in self-slaughter, web addiction and other major depressions which are to be existing in people affected by psychological disorders. There occurs a challenging factor in applying online media to extract patterns regarding mental stress as it is impossible for a machine to understand sarcasm, emoticons, abbreviations, etc. Thus experts accounts which are retrieved at the time of account gathering is used to conversate with professionals for pieces of advice on online social media crowdsourcing. It is important to devise a convenient data collection model to extract specific language patterns from user data so that it works accurately in a methodical manner to analyze unique language patterns. Utilizing some of the related and previous work, we systemize a group of features as attributes to construct the detective models we proposed.

II. RELATED WORK

Social Network Mental Disorder Detection [1] SNMDD model introduced data mining features to three types of SNMDDs [1] [4] Cyber Relationship CR obsession which comprises the obsession with social media surfing to converse and share private information to the state where online relationships became more important than friends and family circles Web addiction which comprises obsessive online social gaming and gambling which affects ones career Information Overload IO [1] includes obsessive scanning of user status tweets posts which leads minimal in person interaction and huge decrease in labour rate There are two main challenges which are said to exist in the design of SMNDD A mingled manual methodology and keyword matching data collection technique are implemented to effectively collect data from patients and regular users which is termed as Mental Illness Detection and Analysis via Social Media MIDAS For the collection of patient's data social media groups have been created manually which are related to mental disorders Using these portals followers list the self volunteering users are also being selected Finally after getting the nal list of patients their tweets are retrieved The preprocessing work considers only the English language keywords from the tweet ignoring other language terms

abbreviations etc Thus users who have very less number of posts or tweets are also ignored MIDAS [2] is concentrated on two important types of features which are semantic and behavioral Text Frequency TF [2] is used to capture the frequent and illustrative words used by the patients The pattern of Life Features PLF [2] let slip the emotional patterns and behavioural traits of the user by measuring polarity [13] scores regarding emotions interaction via social media To utilize multi source learning in SNMDD [1] one basic method is to directly interconnect the attribute of each person's data which is collected from different social networks as a large vector This technique frequently misses the mutual relationship of a attribute in different online social networks and introduce intrusion Thus a tensor [4] techniques have been used in great numbers to model multiple data origins because a tensor [4] can naturally constitute multi source data The latest technology SNMDD based Tensor model STM [4] is presented which allows incorporating the characteristics of SNMDs Furnished with a new tensor model semi supervised learning has been constructed to categorize each user by utilizing Transductive Support Vending Machine TSVM [4] Screening tests are conducted for people of a certain category who has a greater chance of getting affected by psychological disorders [9] Subjects are adjusted into identical age and gender proportion for a less biased analysis [11] Few methods exploit both manually labelled data and noisy labelled data for training In these methods a model called Emoticon Smoothed Language Model ESLAM [3] has been used to continuously club these two kinds of data ESLAM method is compared to the completely supervised Language Model LM to check whether the smoothing with emoticons is impactful or not Under all the evaluations the ESLAM performs profitably in every case more than the completely supervised LM [6] This indicates the truth that the emoticon data do have some impactful and more accurate information and ESLAM can efficiently utilize it to achieve greater performance [8] Detailed emotions provide evidence that further explains a user's behaviour online [6] The system is only used for studying and analysing emoticons used in social media but does not have extended applications Members in a society own standards that make them extraordinarily effectual in opening out their views to others These exceptional individuals drive trends in support of the majority of ordinary people [7] They are merely described as being switched on appreciated and well associated [9] [10] With the help of these works we propose to develop a simple and basic methodology to detect two particular psychological disorder by collecting crowd based data on one hand and acquiring the attributes of patients on other and comparing them to produce needful results

III. PROPOSED METHODOLOGY

This work aims to build a framework for detecting psychological disorders in social media users. We pursue to accomplish our complete method through :

1) Collection of Data

2) Cleaning and preprocessing of Data

3) Extracting Features

Building Detection Models

A. Psychological Profiling

To acquire random-sampled users, a collection of user accounts from Twitter was initially collected. This was done by using a Twitter Streaming package on R and by randomly sampling random accounts. Then to collect tweets we download each set of chosen accounts using the 'TwitteR' package on R. And for the gathering of patient's data and experts data, we utilize a five-step approach that merges manual work and also keywords matching technique, to make the psychological profiling of data.

1) Initially, we manually collect data through a package in R, using one of the social media groups where abundant data for mental disorder is available. A social media group is a common portal where a great number of potential patients and people are available to collect as a resource [12]. This propagates easy data collection. Sometimes there are dedicated groups where related people from clinics, support groups or even doctors are available. For example, there is a portal called @HealingFromBPD [1] that is a viable candidate for social media group. This is because the account shares information on psychological and therapeutic information regarding psychological illnesses. It has a following of over thousands of users. To use the social media groups efficiently we can search Twitter manually using associated disorder as a keyword. There are no additional limitations for selecting one of these accounts. But as a cautionary method, a number of spam accounts with similar profiling were weeded out for quality data collection. These accounts were manually reviewed to confirm if there were entities that believed to be a trustable social media group.

2) Once enough data is collected through these portals, we use the 'TwitteR' package to acquire the follower's list of social media groups. The collected accounts then become the main crowd from where we choose patient and expert data into their respective categories.

3) The interest group in these collected data is taken as self-volunteering users, who are categorized by the information in their bio description. We consider self-volunteering also as a part of data collection process.

4) Once these accounts are identified and collected, we label them manually into three categories *Patient*, a known patient who is affected from any form of psychological disorder, *Expert*, a professional in the field of psychology, including psychiatrist, analyst, and primary care providers (PCP), *Non-related*, a user who is neither of the above

5) Finally, the tweets and posts of the accounts from the final list are obtained by the 'TwitteR' package in R language.

B. Preprocessing

After filtering the information, we apply Sentiment study and Emotion categorization to acquire both the sentiment contrast and emotion depicted by every user posts. To acquire the sentiment information of tweets, we get the aid from R package called CRAN, which is available to download. The sentiment tool arranges the attribute of tweets into three contrast groups positive, negative and neutral.

C. Obtaining the Features

Term Frequency (TF) [3] is the number of times a particular word is referenced. To get this data using the R program, TF-IDF (Term Frequency – inverse document frequency) [3] Feature is used. This feature captures the recurrent and typical words that patients use. The term recurrence is the recurrence of word sequences appear in a crowd of tweets posted by each Twitter user.

'Quanteda' feature is considered to have the psychological terms recurrently used by patients. 'Quanteda' is a simplified category of TF-IDF package, where only the words related to psychology are considered (e.g., stress, feeling, sensation and dysthmia).

D. Pattern Analysis (PA)

Emotional patterns and behavioural tendencies of a user is predicted by measuring emotional contrast, sentiment and social well being. We merge four various types of features as follows:

1) Emotional Tallying: To measure the quantity of the difference in emotional score between a regular user and patients, using the 'Psych' Package in R. It is used to categorize each tweet into one of 8 identified emotions.

2) Age and Gender: As information regarding the age and gender are not provided openly, we adopted the metadata feature using R package. The distribution of age with respect to the number of people affected are analysed as shown in Fig 1. To predict gender and age of the user, we use lexica. This feature is important and inevitable like other feature.

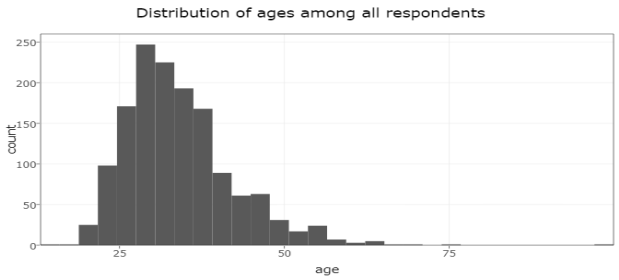

Fig 1: Distribution of ages among all respondents

3) Contrast Features: Using the Twitter package and Quanteda Package, each tweet is grouped as having eithe a positive, negative or neutral attitude. To acquire the attributes of each user, the contrast is changed into five various values which are Positive Quotient, Negative Quotient, Positive Correspondence, Negative Correspondence, Overturn Quotient which helps us in providing information regarding the mental stability of the users.

4) Social Features: For conciseness, features are modeled to attain a user's interactivity with others on the online social network and how constantly they commit on Twitter. The four social features depicted are Tweet recurrence, Mention Quotient, Mention recurrence, Distinct mentions.

IV. RESULTS AND DISCUSSION

Social media groups regarding Bipolar and BPDs were manually collected and begin to download thousands of followers for each community groups [12]. Accounts relevant and matching to each psychological disorder cases were selected manually and grouped to three categories as discussed above which is shown in Table I. Random samples were collected using the Twitter REST API. Expert's accounts are utilized in selection bias test [7]. The random samples take the negative class in the final datasets.

Table I : The cumulated number of accounts, tweets and tweets per user for various categories of users

Category	Users	Tweets	Tweets/User
Random	559	800103	1431
Bipolar Patients	256	339877	1328
BPD Patients	200	230024	1150
Combined	34	55012	1618
Bipolar Experts	11	13765	1251
BPD Experts	12	20024	1669

The performance of both the cases, Borderline Personality Disorder (BPD) and Bipolar Disorder are compared as shown in Fig 2 and Fig 3. Each arc correlates to a model analyzed on a distinct group of features (LIWC [2], TF-IDF [2], Pattern Analysis) which

are described above. The y-axis represents the quota of sensitivity and the x-axis represent the quota of false alarms.

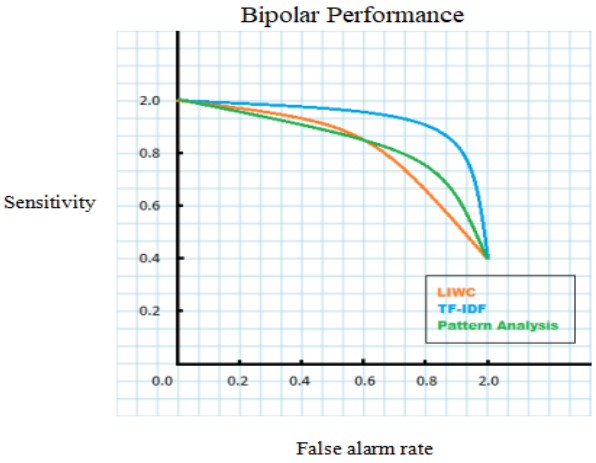

Fig 2 : Execution of the Bipolar model using a unique group of features (LIWC, TF-IDF, Pattern Analysis)

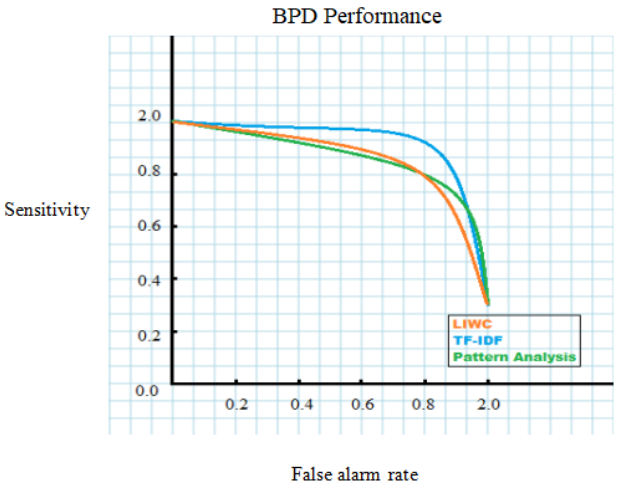

Fig 3 : Execution of the BPD model using a unique group of features (LIWC, TF-IDF, Pattern Analysis)

The average for each case is shown in Table II. It is given that the TF-IDF [2] model produced the greatest average of 94% for both the Bipolar and BPD cases. The Pattern Analysis feature has a lower average than the TF-IDF [2] feature but it is moderately better than the LIWC [2] feature.

Table II : The average performance measures of the group of features (LIWC, TF-IDF, Pattern Analysis)

	LIWC	TF-IDF	Pattern Analysis
Bipolar Disorder	0.89	0.94	0.90
BPD	0.90	0.94	0.91

V. CONCLUSION

In summary a basic data collection mechanism Crowd based profiling is proposed to collect patient and regular users datasets Thereafter an own semantic and habitat features are gathered and adopted for the case of psychological disorder detection It is concluded that to produce satisfying results a combinational methodology of mechanical and automatic effort is needed. The mechanism we use make provision for more advanced research and experiments on psychological disorders using other features such as Linear Regression Support Vector Machines SVM etc

REFERENCES

1. Hong-Han Shuai, Chih-Ya Shen, De-Nan Yang, Yi-Feng Carol Lan and Wang-Chein Lee "A Comprehensive study on Social Network Disorders Detection via Online Social Media Mining" IEEE Transactions on knowledge and data engineering, vol 30, 2018.
2. Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo and Yin-Shin Chen "MIDAS - Mental Illness Detection and Analysis via Social Media" International conference on advances in social networks analysis and mining (ASONAM), 2016
3. Kun-Lin Liu, Wu-Jun LI and Miny Guo "Emoticon Smoothed Lanuage Models for Twitter Sentiment Analysis" Twenty sixth AAAI Conference on Artificial Intelligence, 2012
4. Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chein Lee and Phlips S .Yu "Mining Online Social Data for Detecting Social Network Mental Disorders" Proc. Int. Conf. World Wide Web, 2016
5. M. Cha,H. Haddadi, F. Benevenuto, and K. P.Gummand, "Measuring user influence on Twitter : The million follower fallacy", Proc. Int. AAAI Conf. Weblogs Social Media, 2010
6. E. Saravia, C. Argueta, and Y.-S. Chen. "Emoviz: Mining the world's in-terest through emotion analysis". IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.
7. G. Coppersmith, M. Dredze, and C. Harman. "Quantifying mental health signals in twitter" In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014.
8. C. Argueta, E. Saravia, and Y.S. Chen."Unsupervised graph based patterns extraction for emotion classification" In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.
9. 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), 2012.
10. G. A. C. C. T. Harman and M. H. Dredze. "Measuring post traumatic stress disorder in twitter" In ICWSM, 2014.
11. 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.
12. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. "Predicting depression via social media" In ICWSM, 2013.
13. A. Go, R. Bhayani, and L. Huang. "Twitter sentiment classification using distant supervision" CS224N Project Report, Stanford, 2009.

AUTHORS PROFILE

A. Sharmila Agnal, Received M.E degree from Anna University in the year 2011. Her research interests includes data mining. She is an Assistant Professor (O.G) for the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram Campus.

Akshay Kannan V, pursuing B Tech degree in the department of Computer Science and Technology, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and data analytics.

Durga S, pursuing B Tech degree in the department of Computer Science and Technology, Ramapuram Campus, graduating in the year 2019. Her research interests include Internet of Things, Embedded systems and data mining.

Nishanth Kumar S, pursuing B Tech degree in the department pd Computer Science and Technology, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and data analytics

Dheeraj R, pursuing B Tech degree in the department of Computer Science and Technology, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and data analytics.