IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2019, accepted November 8, 2019, date of publication November 14, 2019,
date of current version November 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2953087

A Novel Co-Training-Based Approach for the Classification of Mental Illnesses Using Social Media Posts

SUBHAN TARIQ[1], NADEEM AKHTAR[2], HUMAIRA AFZAL[3], SHAHZAD KHALID[4], MUHAMMAD RAFIQ MUFTI[5], SHAHID HUSSAIN[6], ASAD HABIB[6], AND GHUFRAN AHMAD[1]

[1]Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan
[2]Department of Computer Science and IT, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[3]Department of Computer Science, Bahauddin Zakariya University, Multan 60000, Pakistan
[4]Department of Computer Engineering, Bahria University, Islamabad 44000, Pakistan
[5]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari 61100, Pakistan
[6]Department of Computer Science, Kohat University of Science and Technology, Kohat 26000, Pakistan

Corresponding author: Shahid Hussain (shussain@comsats.edu.pk)

ABSTRACT Context: Recently, research community of certain domain showing their eagerness towards the use of social media networks to gain constructive knowledge in decision making and automation, such as aid to perform software development activities, crypto-currencies usage, network community detection and recommendation and so on. Recently, besides other domains of eHealth, the use of social media and big data analytics has become hot topic to predict the patient of mental illness involved in either depression, schizophrenia, eating disorders, anxiety or addictive behaviors. Problem: Traditional methods either need enough historic data or to keep the regular monitoring on patient activities for identification of a patient associated with a mental illness disease. Method: In order to address this issue, we propose a methodology to classify the patients associated with chronic mental illness diseases (i.e. Anxiety, Depression, Bipolar, and ADHD (Attention Deficit Hyperactivity Disorder) based on the data extracted from the Reddit, a well-known network community platform. The proposed method is employed through Co-training (type of semi-supervised learning approach) technique by incorporating the discriminative power of widely used classifiers namely Random Forrest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB). We used Reddit API to download posts and top five associated comments for construction of a feature space. Results: The experimental results indicate the effectiveness of Co-training based classification rather than the state of the art classifiers by a margin of 3% on average in par with every state of art technique. In future, the proposed method could be employed to investigate any classification problem of any domain by extracting date from the social media.

INDEX TERMS Mental disease, reddit, anxiety, depression, bipolar, ADHD.

I. INTRODUCTION

According to World Mental Health (WMH) survey report, trillions of people world-wide suffer from mental disorders. Considering the survey report statistics, mental disorders such as anxiety disorder, mood disorder, impulse control disorder, psychotic disorder, addiction disorder and personality disorder, are becoming common day by day. Both in developed and developing countries, 25% of the world's population is suffering from mental illness [1].

There are a lot of improvements in the field of medical science but still people are suffering from mental health issues, as it is considered a taboo subject. For example, Depression is the most common class of mental disorder that physically affects a person, causing severe headaches, eating disorders and deprivation of sleep. Usually the diagnosis involves report either by the person having the illness or by some close family members or friends. However, due to certain emotions and feelings the individual going through such mental health issues remain close by themselves. Furthermore, as the volume of the data related to mental health is growing so, there is need for improvement in diagnostic analysis and

The associate editor coordinating the review of this manuscript and approving it for publication was Sohail Jabbar.

classification in order to better understand the accumulation of patient's data. According to authors report [2], young people are more open to talk about their mental health issues over the social platforms. Over the years there had been a significant improvement in the field of Machine Learning (ML) in order to solve real problems or for introduction of automated system. Recently, ML is rapidly becoming popular for diagnosis of mental health issues due to the vast amount of data available on social media platforms [3].

Reddit is an American social news, web substance rating, and dialog site. Enlisted individuals submit substance to the site, for example, joins, content posts, and pictures, which are then voted up or down different individuals. Posts are composed by subject into members made boards called ''sub-reddits'', which spread an assortment of points including news, science, motion pictures, computer games, music, books, wellness, nourishment, and picture sharing. Entries with additional up-votes show up towards the highest point of their sub-reddit and, on the off chance they get enough up-votes, eventually on the site's first page.

Our approach focuses on the tools for extraction of indicators of mental illness from the words that are being used in the post by an individual. In our work we are going to target a social platform ''Reddit'' in order to scrape the posts from the clinical sub-reddits mainly for mental disorder such as Depression, Anxiety, Bipolar and ADHD. Subsequently, our model is trained for the classification of upcoming posts based on the previous learned data.

Our propose methodology is employed by leveraging the abilities of C-training algorithm. It is a semi-supervised learning method which needs two views of the data. Further, it work on assumption which are related to two different set of features give complementary information about the instances. Normally, two views are created which worked conditionally independent which mean that the two feature space of each instance is conditionally independent for the given class. Moreover, each view is adequate to predict a class correctly. In Co-training technique, firstly each classifier is learned separately for each view by using the labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data. In our work we are going to use Co-training in order to improve the classification accuracy of the state-of-the-art classification models used in the target research domain.

The rest of the paper is organized as follows: section 2 presents an overview of the approaches used for the classification of data. Section 3 discusses about Co-training. Section 4 describes the proposed methodology. In section 5, we describe our results and made discussion. Finally, in section 6, we present the conclusion of proposed study.

II. RELATED WORK

Most studies that used natural language in order to predict mental illnesses focus on the words that overlap with the symptoms of that mental illness. In [4], [5], the author predicted future mental illness based on the posts from an individual's post on Reddit, by gathering the posts from clinical sub-reddits and then classifying them to the corresponding mental illness. After gathering the posts, clustering was applied on those posts to find the markers of mental illnesses present in their everyday spoken language. Finally they took posts from the same users before they posted in clinical sub-reddits. To some extent, they showed that it is possible to predict the future mental illness of an individual, on the basis of their posts on social media.

Reference [6] Predicted the severity of depression by using the text present in the status update of an individual on Facebook. They trained their model on the basis of the posts of depressed Individual and gave a score for depression with respect to certain posts. In reference [7], authors used self-identifying statements such as ''I was depressed'' from twitter to train the model and then predicted the depression on the basis of extracted statements. Reference [8] predicted mental health and well-being based on the clinical sub-reddits on Reddit, an individual is subscribed to using a logistic regression model. In study [9], the authors predicted depression and PTSD from the individual's posts of twitter, which had particular language markers for mental illnesses. The result of all these studies shows that we can predict mental illnesses from posts, posted on social network platforms, containing the markers for those mental illnesses.

In study [10], the authors have used semi-supervised learning to train the text data. They improved the classification of text data using Support Vector Machine (SVM) on labeled and unlabeled data, the unlabeled data was labeled with the help of expert opinion but since unlabeled data is in huge amount so they proposed a methodology which selects a batch of informative data to be labeled and then that data is labeled by experts. It took less effort and maintains the accuracy of the model. In [11], the authors showed four state-of-the-art feature selection models, then applied most commonly used models for the text classification: Nearest Neighbor (NN), Naïve Bayes (NB), Decision Tree (DT), Neural Networks and SVM. According to their survey SVM works best in most cases for text classification.

In study [12], authors used Deep Beliefs Networks (DBN) and Soft-max Regression. As the text data contains sparse high dimensionality feature matrices, so to solve this issue the authors used DBN for extraction of features, followed by a Soft-max Regression for classification of the data from learned feature space. Furthermore a Limited-memory, Broyden–Fletcher–Goldfarb–Shanno Algorithm was used for optimizing the parameter of the system model. Their results were better than SVM and NN.

In study [13], authors predicted the origin of death from autopsy reports, which were in the form of text. They gathered the autopsy reports, belonging to 8 different reasons of death, applied preprocessing, feature extraction, feature selection and created a feature space containing 43 features per report. Subsequently, authors have used different state-of-the-art text classifications models such as SVM, NN, NB, and Random

Forest (RF). Finally, authors reported the SVM as an outperformed classifier as compared to others.

In study [14], authors have applied, Hybrid Association Classification (HAC) with NB for classification of text data. HAC performed very well with the text classification. However, the main issue was a large number of classification procedures and pruning techniques that might remove vital information for classification. Consequently, the authors combined HAC with NB in order to reduce the number of classification rules, they used several datasets for classification and proved that HAC produce less classification rules and gives stable efficacy in terms of Accuracy and F-measure.

In study [15], authors have conducted a study to compares the performance of different approaches for automatic classification of text. Authors used 5 lexicon based algorithms and 5 machine learning algorithms on the target datasets. The dataset comprised of 41 major social media platforms with various sample size and different languages. For market based research, SVM and LIWC (Linguistic Inquiry and Word Count) out-performed every other algorithm. For human intuition, RF or NB performed better than others. Authors concluded that NB or RF are not so far back in the market based research, so they can be modified in order to perform well on the dataset.

Given the above studies, we can conclude that SVM, NB and Random Forest perform better for classification of text. It mainly depends on the characteristics of a Dataset which is under study. Most of the studies used simple SVM or SVM with some modification in order to get better classification accuracy. Similarly, for RF it can be inferred that it is very good for multi-class classification and NB also works well with the dataset of small sample sizes. Hence, the Machine learning approaches outperform the lexicon-based algorithms. Furthermore, it can be noticed that feature extraction and feature selection techniques can help improve the classification accuracy.

III. CO-TRAINING

Previously, it has been reported that labeled and structured data is less available rather than unlabeled and unstructured data. In 1998, Avrim Blum and Tom Mitchell present a concept for a new type of learning. They gave the idea that if we train two weak classifiers on the less available labeled data and later try to label the unlabeled data based on the predictions from those weak classifier, then we get samples from both weak classifiers having high confidence. Subsequently, we place these samples in labeled dataset and after labeling them to train the weak classifiers again on the newly labeled dataset to improve the performance of those weak classifiers. Co-training only works if both of the views (i.e., labeled data and unlabeled data) are conditionally independent (i.e., the two feature space of each instance are conditionally independent of the given class) and each view is adequate (i.e., an instance's class can be accurately predicted with respect to view).

In our work, we use the Co-training algorithm in order to enhance the classification accuracy. In study [16], authors present an algorithm for Co-training, which work as follow.

Pseudo code for Co-training Technique

***Given*:**

- $\boldsymbol{L}$ = *labeled training data*
- $\boldsymbol{U}$ = *unlabeled data*
- $\boldsymbol{U'}$ = *a pool of examples by choosing* $\boldsymbol{u}$ *examples at random from* $\boldsymbol{U}$

***Loop for k iterations*:**

- *Use* $\boldsymbol{L}$ *for classifier's training* $\boldsymbol{h_1}$ *which considers only* $\boldsymbol{x_1}$ *portion of* $\boldsymbol{x}$
- *Use* $\boldsymbol{L}$ *for classifier's training* $\boldsymbol{h_2}$ *which considers only* $\boldsymbol{x_2}$ *portion of* $\boldsymbol{x}$
- *Use* $\boldsymbol{h_1}$ *to label* $\boldsymbol{p}$ *positive and* $\boldsymbol{n}$ *negative examples from* $\boldsymbol{U'}$
- *Use* $\boldsymbol{h_2}$ *to label* $\boldsymbol{p}$ *positive and* $\boldsymbol{n}$ *negative examples from* $\boldsymbol{U'}$
- *Randomly chose* $\boldsymbol{2p+2n}$ *examples from* $\boldsymbol{U}$ *to replenish* $\boldsymbol{U'}$

IV. METHODOLOGY

The main objective of proposed methodology is to classify the mental illnesses, based on the training data from the posts on clinical sub-reddits. We employed the proposed methodology with four of the sub-reddits for training of our model. The four sub-reddits include ADHD, Depression, Bipolar and Anxiety. We construct three different models by leveraging the discriminative power of SVM, NB and RF for the employment of target classification problem. Firstly, we construct model by leveraging the Co-training technique with base level classifiers. Secondly, we compare constructed models with base level classifiers. The words of different clinical sub-reddits will be sufficient enough for accurate classification [17]–[19]. In each experiment, we divided the dataset in 80-20 percent split for training and testing data respectively. Model fitted on training data was generalized on never seen before test data. Fig. 1 depict the graphical view of our model. The description of each phase of proposed approach is as follows.

A. DATA ACQUISITION

The first phase of the proposed methodology is data acquisition. We use python (API) for Reddit, PRAW in order to download the top 1000 posts from each of the following sub-reddits such as r/Depression, r/Anxiety, r/ADHD and r/Bipolar. Some of the posts were deleted by the users in the sub-reddit, so the overall number of posts count goes up to 3922. Furthermore, we also download the top 5 comments per post and save them in a separate file.

B. DATA PREPROCESSING

The second phase of the proposed methodology is data preprocessing. Data preprocessing employs several activities, such as:

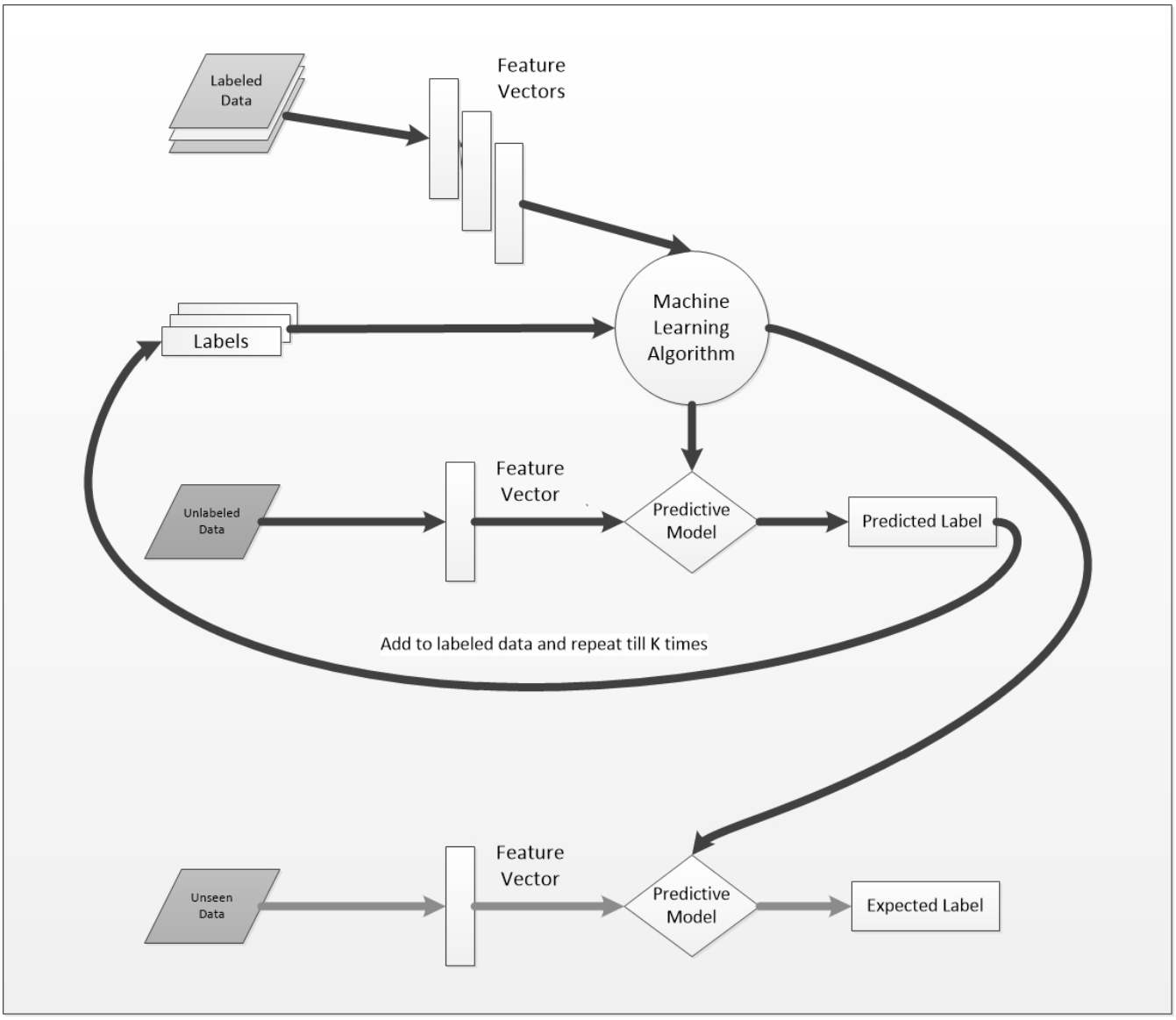

FIGURE 1. Graphical overview of proposed model.

- First activity is the removal of the blank rows or null data from the dataset.
- Second activity is changing all the text data to lower case, it is necessary because of the uncertainty that the same two words can be considered different. For example the word 'dog' and word 'Dog' will be considered two different words.
- Third activity is tokenization. It is a process of creating words from sentences where each sentence in dataset will be broken down into words.
- Fourth activity is removal of stop words. In this process we remove those words which are very commonly used, carry no information and have a large number of frequency such as; is, the, are, etc., we remove these words because they can negatively affect the creation of representative feature vectors and yield to reduce the classification accuracy.
- Fifth activity is Stemming or Lemmatization. In this process, root for each word is generated and words with same conceptual meanings are grouped together. This is done to reduce the number of words.
- The final activity of data processing is the removal of non-alphabetic characters.

C. FEATURE EXTRACTION / SELECTION

Third phase of the proposed methodology is the feature extraction and selection. This step involves three activities i.e. Feature extraction using TF-IDF words factorization, feature selection and 80-20 split for training and testing. We used Term Frequency-Inverse Document Frequency (TF-IDF) to extract the features from the pre-processed data. This process can be split into two parts. In the first part the Term Frequencies (TF) are calculated, which is the number of times a term is used in a document. The output of this

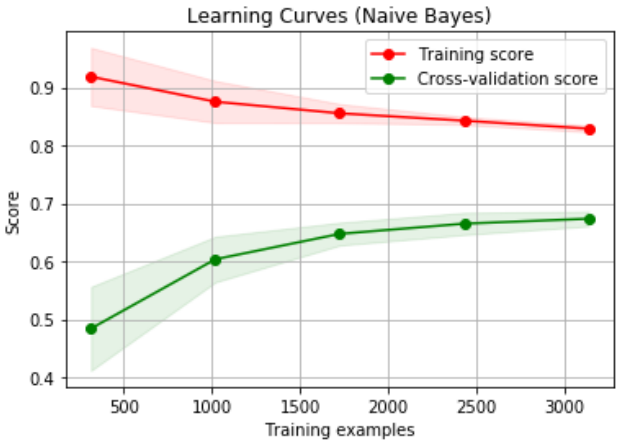

FIGURE 2. Learning curve of NB.

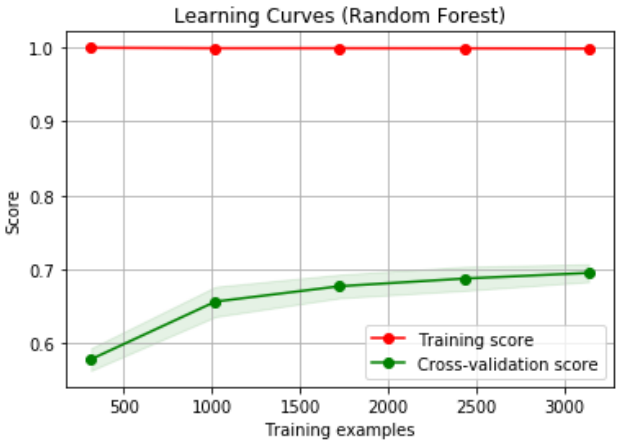

FIGURE 3. Learning curve of RF.

activity is a weight of each word, proportional to its number of occurrences. The second part is Inverse Document Frequency (IDF), which is opposite of TF. There are some terms which tend to occur more often than the others, and these terms wrongly emphasize the document leading to lowering the classification accuracy. So, the IDF reduce the weight of the terms which occur frequently. The output of TF-IDF is vectors containing the weights of the words.

After the feature extraction, next activity is feature selection. For feature selection there are two well-recognized techniques namely pruning and clustering [20]–[22]. For our dataset we use pruning and chi-squared method. In accordance with our dataset, the chi-squared technique outperforms the pruning technique. In chi-squared technique, Chi-square test is employed. For categorical features in a dataset, we calculate the Chi-square between each of the features and the target. Moreover, we select the required number of features with the best Chi-square scores. It determines the association between two categorical variables of the sample which would lead to reflect their real association in the population. The Chi-squared score is computed through

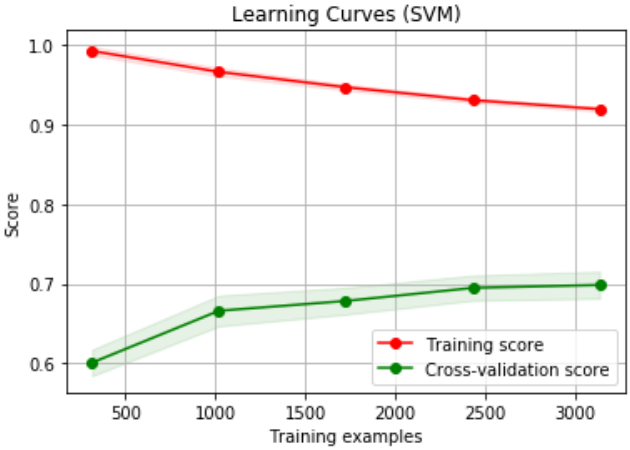

FIGURE 4. Learning curve of SVM.

TABLE 1. Performance evaluation of Co-training based SVM.

Class	Performance Measures		
	Precision	Recall	F-measure
Anxiety	0.87	0.81	0.84
ADHD	0.66	0.68	0.67
Depression	0.85	0.55	0.67
Bipolar	0.60	0.82	0.70

TABLE 2. Performance evaluation of Co-training based NB.

Class	Performance Measures		
	Precision	Recall	F-measure
Anxiety	0.85	0.81	0.83
ADHD	0.64	0.68	0.66
Depression	0.70	0.69	0.70
Bipolar	0.71	0.71	0.71

TABLE 3. Performance evaluation of Co-training based RF.

Class	Performance Measures		
	Precision	Recall	F-measure
Anxiety	0.85	0.81	0.83
ADHD	0.64	0.68	0.66
Depression	0.70	0.69	0.70
Bipolar	0.71	0.71	0.71

equation 1 as follow.

$$X^2 = \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency} \quad (1)$$

After the selection of features, we divide the dataset into 80-20 split namely 80% data for training and 20% data for testing.

D. CLASSIFICATION

Fourth phase of the proposed methodology is named as classification which is employed through using machine learning techniques for the grouping of the dataset. We use

TABLE 4. Comparison of proposed approach with state of the art classifiers.

Class	Classifiers with and without Co-training					
	SVM		NB		RF	
	With Co-training	Without Co-training	With Co-training	Without Co-training	With Co-training	Without Co-training
Anxiety	**0.84**	0.76	**0.83**	0.81	**0.83**	0.74
ADHD	**0.67**	0.67	**0.66**	0.58	**0.66**	0.65
Depression	**0.67**	0.67	**0.70**	0.65	**0.70**	0.70
Bipolar	**0.70**	0.70	**0.71**	0.66	**0.71**	0.71

the pre-built model for SVM, RF and NB by tuning the parameters. Subsequently, three more models were created, which based on the Co-training models of SVM, RF and NB. For Co-training purpose, we created a dataset which contains both: posts and comments. Since comments are not as reliable as posts, we did not label comments. In other words comments were used as unlabeled data. After the creation of dataset. We divided the feature set of that dataset in two half in terms of creating 2 different views, which are necessary for Co-training. Two models were created of SVM, RF and NB trained on labeled data (i.e. posts), same for the both views. Since the label data is less in comparison to unlabeled data, we call these models weak classifiers. After creation and training of these weak classifiers, we took a batch of unlabeled examples from the dataset and request these weak classifiers for prediction. We took two examples of each class on which those weak classifiers were most confident. Subsequently, these examples are added to labeled dataset and trained those weak classifiers again on this newly created labeled dataset. This process is repeated several times. After each iteration, we took predictions from all models and compare them.

V. RESULTS

We performed several experiments in order to evaluate the effectiveness of proposed method. The evaluation was performed by employing the cross validation method. It use k as 10, 20 percent of the data as test data.

As shown in Fig. 2, the training accuracy is 82 percent while the test accuracy is around 68 percent. Hence proving that it is the worst performing model of the three (i.e., NB, SVM and RF).

Similarly, Fig. 3 show the learning curves of RF model that is trained on the same dataset, under the same conditions. Resultantly, RF over-fits on the training data, but still it gives better performance on the testing data (70 percent) as compared to NB.

Fig. 4 show the learning curves of SVM model that is trained on the same dataset, under the same conditions. SVM performance on testing data is similar to that of RF but training performance of RF remain better as compared to SVM.

For Co-training versions of these 3 models SVM perform better as compared to NB and RF. Furthermore, there is no graph to show their learning curves because of the function used to produce these graphs did not worked with the Co-training version hence, a theoretical explanation using classification report is provided for these models.

From the Table 1, we can observe the effectiveness of proposed approach (Co-training based SVM) in terms of class-wise classification of comments. In terms of F-measures, we observe the performance of proposed approach for the classification of Anxiety post and its related comments (F-measure = 0.84) as compared to ADHD (F-measure = 0.67), Depression (F-measure = 0.67), and Bipolar (F-measure = 0.70).

From the Table 2, we can observe the effectiveness of proposed approach (Co-training based NB) in terms of class-wise classification of comments. In terms of F-measures, we observe the performance of proposed approach for the classification of Anxiety post and its related comments (F-measure = 0.83) as compared to ADHD (F-measure = 0.66), Depression (F-measure = 0.70), and Bipolar (F-measure = 0.71).

From the Table 3, we can observe the effectiveness of proposed approach (Co-training based RF) in terms of class-wise classification of comments. In terms of F-measures, we observe the performance of proposed approach for the classification of Anxiety post and its related comments (F-measure = 0.83) as compared to ADHD (F-measure = 0.66), Depression (F-measure = 0.70), and Bipolar (F-measure = 0.71).

Moreover, in order to benchmark the performance of proposed approach (Co-training with base level classifier), we assess its performance with state of the art base level classifier that is the performance of SVM is assessed with Co-training approach with SVM. The results are shown in Table 4 in terms of F-measure as follow.

The results of Table 4 indicate the effectiveness of proposed approach (Co-training with base level classifiers) in terms of F-measure. We observe significant improvement in the classification of Anxiety posts for all types of Co-training techniques such as SVM with Co-training (F-measure = 0.84), NB with Co-training (F-measure = 0.83), and RF with Co-training (F-measure = 0.83). However, in case of classification of ADHD, Depression, and Bipolar posts, the performance of SVM (with and without Co-training) remain same which might be existence of feature label noise.

However, the performance of NB and RF (with Co-training) remain better as compared to base level classifiers.

VI. THREATS TO VALIDITY

In this study, we also observe some threats. The first threat is related to generalization of results. We have reported the results with limited number of datasets and number of classifiers. We can consider more case studies and include classifiers to benchmark the classifier's performance. The second threat is related to use of classifier with their default parameters. However, the effectiveness of proposed approach can be altered by tuning the parameters.

VII. CONCLUSION

The experimental results of Co-training technique based propose approach are promising which indicate its effectiveness as compared to the use of state of the art classifiers in terms of classification of posts with respect their influential features. We performed several experiments to classify the posts and their associated comments related to four mental issues such as Anxiety, ADHD, Depression and Bipolar. We mined date from the Reddit platform where community related posts are published. We used an API to extract posts and associated comments and performed experiments by using SVM, NB, and RF classifiers. The experimental results indicate that SVM, NB, and RF outperformed with Co-training technique as compared to their individual use in terms of Precision, Recall, and F-measure. In future, we will employ the proposed approach in terms of classification of posts of other domains according to interest of research community.

REFERENCES

[1] *The Global Burden of Disease 2004*, WHO, Geneva, Switzerland, 2008, p. 146.

[2] J. A. Naslund, K. A. Aschbrenner, G. J. McHugo, J. Unützer, L. A. Marsch, and S. J. Bartels, "Exploring opportunities to support mental health care using social media: A survey of social media users with mental illness," *Early Interv. Psychiatry*, vol. 13, no. 3, pp. 405–413, 2019.

[3] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, 2019.

[4] R. Thorstad and P. Wolff, "Predicting future mental illness from social media: A big-data approach," *Behav. Res. Methods*, vol. 51, no. 4, pp. 1586–1600, 2019.

[5] J. Dabbs, D. M. Crow, M. R. Mehl, J. W. Pennebaker, and J. H. Price, "The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations," *Behav. Res. Methods, Instrum., Comput.*, vol. 33, no. 4, pp. 517–523, 2011.

[6] H. A. Schwartz, J. Eichstaedt, M. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar, "Towards assessing changes in degree of depression through facebook," in *Proc. Proc. Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, 2014, pp. 118–125.

[7] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, "Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, vol. 1, 2015, pp. 99–107.

[8] S. Bagroy, P. Kumaraguru, and M. De Choudhury, "A social media based index of mental well-being in college campuses," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 1634–1646.

[9] G. C. Qntfy, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Depression PTSD Twitter*, 2015, pp. 1–10.

[10] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using SVM for text classification," *Int. J. Automat. Comput.*, vol. 15, no. 3, pp. 290–298, 2018.

[11] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review," *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3797–3816, 2019.

[12] M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, and R. Guan, "Text classification based on deep belief network and softmax regression," *Neural Comput. Appl.*, vol. 29, no. 1, pp. 61–70, 2018.

[13] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, and K. Shaikh, "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study," *J. Forensic Legal Med.*, vol. 57, pp. 41–50, Jul. 2018.

[14] W. Hadi, Q. A. Al-Radaideh, and S. Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification," *Appl. Soft Comput.*, vol. 69, pp. 344–356, Aug. 2018.

[15] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Marketing*, vol. 36, no. 1, pp. 20–38, 2019.

[16] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with Co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, Jul. 1998, pp. 92–100.

[17] S. Hussain, J. Keung, M. K. Sohail, M. Ilahi, and A. A. Khan, "Automated framework for classification and selection of software design patterns," *Appl. Soft Comput.*, vol. 75, pp. 1–20, Feb. 2019.

[18] S. Hussain, J. Keung, and A. A. Khan, "Software design patterns classification and selection using text categorization approach," *Appl. Soft Comput.*, vol. 58, pp. 225–244, Sep. 2017.

[19] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, and R. Dutta, "Corrigendum: Characterisation of mental health conditions in social media using informed deep learning," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 46813.

[20] A. K. Uysal, "On two-stage feature selection methods for text classification," *IEEE Access*, vol. 6, pp. 43233–43251, 2018.

[21] S. Hussain, "A methodology to predict the instable classes," in *Proc. 32nd ACM Symp. Appl. Comput. (SAC)*, Apr. 2017, pp. 1307–1308.

[22] S. Hussain, H. Afzal, M. Mufti, M. Imran, A. Ali, and B. Ahmad, "Mining version history to predict the class instability," *PLoS ONE*, vol. 14, no. 9, 2019, Art. no. e0221780.

SUBHAN TARIQ is currently pursuing the M.S. degree in computer science with COMSATS University Islamabad. His research interests include topic modeling, machine learning, and text mining.

NADEEM AKHTAR received the M.S. degree in computer science from IUP, University of South Brittany, Vannes, France, and the Ph.D. degree from the VALORIA Research Laboratory, University of South Brittany, Vannes. He has published several articles in reputed conferences and journals. His major research interests include formal software engineering, software architecture, and multiagent robotics.

HUMAIRA AFZAL received the M.Sc. degree in computer engineering from the Centre for Advanced Studies in Engineering (CASE), Islamabad, Pakistan, in 2010, and the Ph.D. degree in computer science from the School of Electrical Engineering and Computer Science, University of Bradford, U.K., in August 2014. She is currently a Lecturer with the Institute of Computing, Bahauddin Zakariya University, Multan, Pakistan. Her research interests include MAC protocol design for cognitive radio networks, performance modeling, queuing theory, and network security.

SHAHZAD KHALID received the B.Sc. degree in computer systems engineering from the GIK Institute of Engineering Sciences and Technology, Topi, in 2000, the M.Sc. degree in software engineering from the National University of Science and Technology, Rawalpindi, in 2003, and the Ph.D. degree in motion data mining and machine learning from The University of Manchester, U.K., in 2009. He is currently a Professor with the Department of Computer Engineering, Bahria University, Islamabad. He is also the Director (ORIC) of the Head Office of Bahria University. He is also headed by the research group Computer Vision and Pattern Recognition Research Center. He is also a supervisor of 17 M.S. students and eight Ph.D. students. He has completed his 76 publications so far including eight ISI indexed and 68 journal articles and 43 conference papers. He has completed nine research projects and published five book chapters and one complete book. His research interests include computer vision and pattern recognition. He has received 18 different awards, including the Best Researcher and the Best Teacher from Bahria University and the Higher Education Commission, Pakistan, and other organizations.

MUHAMMAD RAFIQ MUFTI received the M.Sc. degree in computer science from Bahauddin Zakariya University, Multan, Pakistan, in 1994, the M.Sc. degree in computer engineering from the Centre for Advanced Studies in Engineering (CASE), Islamabad, in 2007, and the Ph.D. degree in electronic engineering from Mohammad Ali Jinnah University (MAJU), Islamabad, in 2012. He is currently a Faculty Member of the COMSATS Institute of Information Technology, Vehari, Pakistan. His research interests include sliding mode control, fractional control, neural networks, cognitive radio networks, and network security.

SHAHID HUSSAIN received the M.Sc. degree in computer science from Gomal University, Dera Ismail Khan, Pakistan, the M.S. degree in software engineering from the City University of Science and Information Technology (CUSIT), Peshawar, Pakistan, and the Ph.D. degree from the City University of Hong Kong. He has published certain articles in reputed conferences and journals. His research interests include the software design patterns and metrics, text mining, empirical studies, and software defect prediction. He was the Leading Guest Editor of a special issue namely Knowledge Discovery for Software Development in IET Software.

ASAD HABIB received the Dr.Eng. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan. He is currently the Director of the Institute of Information Technology (IIT), Kohat University of Science and Technology, Kohat, Pakistan. His research interests include data science, natural language processing, computational modeling, software engineering, knowledge-based organizational analytics, prediction, and recommender systems.

GHUFRAN AHMAD received the Ph.D. degree from the Department of Computer Science, Mohammad Ali Jinnah University (renamed to Capital University of Science and Technology), Islamabad, Pakistan, in 2013, and the Ph.D. degree from the Department of Computer Science and Digital Technology, Faculty of Engineering and Environment, Northumbria University, Newcastle Upon Tyne, U.K., in 2015. He was a Visiting Scholar with the CReWMaN Laboratory, Department of Computer Science and Engineering, The University of Texas at Arlington, from 2008 to 2009. He is currently serving as an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Pakistan. He contributed in two book chapters published by Elsevier and IET. His research interests include the IoT, wireless sensor networks, and wireless body area networks. He currently serves as an Associate Editor for IEEE Access, *Ad Hoc & Sensor Wireless Networks* (AHSWN), *Journal of Sensors* (Hindawi), and *Wireless Communications and Mobile Computing* (Hindawi). He was leading special issues, as a Guest Editor of the *International Journal of Distributed Sensor Networks* (IJDSN). He is currently leading special issue at *Journal of Sensors* (Hindawi) and *Wireless Communication and Mobile Computing* (Hindawi).