Depression Detection in Social Media Using a Psychoanalytical Technique for Feature Extraction and a Cognitive Based Classifier

Chapter · October 2020

DOI: 10.1007/978-3-030-60887-3_25

CITATIONS
0

READS
47

4 authors:

Seyed Habib Hosseini Saravani
Instituto Politécnico Nacional
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Sara Besharati
Instituto Politécnico Nacional
4 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Hiram Calvo
Instituto Politécnico Nacional
117 PUBLICATIONS **389** CITATIONS
SEE PROFILE

Alexander Gelbukh
Instituto Politécnico Nacional
556 PUBLICATIONS **6,417** CITATIONS
SEE PROFILE

Depression Detection in Social Media Using a Psychoanalytical Technique for Feature Extraction and a Cognitive Based Classifier*

Seyed Habib Hosseini-Saravani, Sara Besharati, Hiram Calvo, Alexander Gelbukh

Center for Computing Research, Instituto Politécnico Nacional
Av. JD Bátiz e/ MO de Mendizábal s/n, Nva. Ind. Vallejo, 07738, Mexico City, Mexico

{hosseinihaamed, besharatisara62}@gmail.com;
hcalvo@cic.ipn.mx, gelbukh@gelbukh.com

Abstract. Depression detection in social media is a multidisciplinary area where psychological and psychoanalytical findings can help machine learning and natural language processing techniques to detect symptoms of depression in the users of social media. In this research, using an inventory that has made systematic observations and records of the characteristic attitudes and symptoms of depressed patients, we develop a bipolar feature vector that contains features from both depressed and non-depressed classes. The inventory we use for feature extraction is composed of 21 categories of symptoms and attitudes, which are primarily clinically derived in the course of the psychoanalytic psychotherapy of depressed patients, and systematic observations and records of their characteristic attitudes and symptoms. Also, getting insight from a cognitive idea, we develop a classifier based on multinomial Naïve Bayes training algorithm with some modification. The model we develop in this research is successful in classifying the users of social media into depressed and non-depressed groups, achieving the F1 score 82.75 %.

Keywords: Depression Detection, Social Media, Natural Language Processing, Psychoanalysis, Rational Speech Act, Naïve Bayes

1. Introduction

Drastic changes in human lifestyle over the past century has led to an increase in the number of people suffering from depression, which is believed to be a "disease of modernity" [1]. It is predicted that, by 2030, one of the three leading causes of illness will be depression [2], and the epidemic of child and adolescent depression is now one the most important concerns in academia. Depression negatively affects how a person feels, thinks and acts and can bring a variety of emotional and physical problems that can decrease the person's ability to function in the society. Therefore, it is highly beneficial if depression can be detected before it hurts both the individuals and the society.

* This work was done with support of the Government of Mexico via CONACYT, SNI, CONACYT grant A1-S-47854 and grants SIP 20200811, SIP 20200859 of the Instituto Politécnico Nacional (IPN), IPN-COFAA and IPN-EDI.

Depression symptoms show themselves in different categories of human behavior and activities and can vary from mild to severe [3]. One of the most important sources that helps in detecting symptoms of depression in individuals is the language they use. In fact, language is the manifestation of how a person feels and thinks. Cognitive and linguistic analyses [4] have shown that people with depression use language differently. For example, they frequently use adjective and adverbs conveying negative emotions, such as "lonely", "sad", or "miserable", and they have a tendency to use first person singular more than second or third person, which shows that they are more focused on themselves, and less connected with others. In addition, the style in which the people with depression use language show that they frequently use absolutist words, such as "always", "nothing" or "completely" [4, 5].

Social media, as one of the most important sources of individuals' thoughts, opinions and feelings [6], can carry very important information about the users [21, 22]. In the area of psychoanalytical and psychological studies, relying on social media as a behavioral health assessment tool even has many advantages over self-report methodology in behavioral surveys because social media measurement of behavior captures social activity and language expression in a naturalistic setting—in contrast with behavioral surveys, where responses are prompted by the experimenter and typically comprise recollection of health facts [6]. As a result, social media has been an important source of data for researchers [6, 7], and also there have been many works on depression detection in social media using Natural Language Processing (NLP) [8] techniques, which can help us mine the data in the social media in order to detect the signs of depression in the users of social media.

In this research, getting insight from cognitive and psychoanalytical findings, we apply modified multinomial Naïve Bayes algorithms to the area of depression detection in social media. The contributions of this research are mainly in two areas: (a) feature extraction, (b) learning method. For feature extraction, using an inventory [3] that has made systematic observations and records of the characteristic attitudes and symptoms of depressed patients, we develop a bipolar feature vector that contains features from both depressed and non-depressed classes. Then we use these features to train a classifier whose learning is based on multinomial Naïve Bayes [9] algorithm. However, we have made a modification in the training phases of our classifier in order for it to increase the importance of some of the features in certain conditions. In our model, features of each of the target classes are of more importance when they are observed in the training data of that certain class. The idea behind this modification is based on Rational Speech Act (RSA) theory, [10] which is a Bayesian based cognitive theory whose main idea is that not all features of a thing or a person have the same value for human brain when during the recognition of a thing or a person by human brain. Based on RSA theory, this characteristic of human brain can help humans have a rational inference of what others say.

In the following paragraphs, we will discuss previous work in Section 2, explain the methodology in detail in Section 3, discuss the experimental results in Section 4, and summarize our findings and talk about future work in Section 5.

2. Related Work

De Choudhury et al. [11, 12] used crowdsourcing methodology to build a large corpus of postings on Twitter that have been shared by individuals diagnosed with clinical depression. They developed a model (an SVM classifier) trained on this corpus to determine if posts could indicate depression. Their model leveraged signals of social activity, emotion, and language manifested on Twitter. For feature extraction, they proposed several features to characterize the postings in their dataset. The features could be categorized into two types: post-centric and user-centric features. Post-centric features—emotion, time, linguistic style—captured properties in the post, while the user-centric features—engagement, ego-network—characterized the behavior of the post's author. Their models could predict if a post is depression-indicative, with accuracy of more than 70% and precision of 0.82. This work demonstrated how sets of behavioral markers manifested in social media can be harnessed to predict depression-indicative postings, and thereby understand large-scale depression tendencies in populations.

Evaluating depressive symptoms using the Center for Epidemiological Studies-Depression (CES-D) scale, Sungkyu et al. [13] developed a Web application to identify depressive symptom–related features from users of Facebook as a popular social networking platform. They provided tips and facts about depression to participants and measured their responses using EmotionDiary, the Facebook application that they had developed. To identify the Facebook features related to depression, correlation analyses were performed between CES-D and participants' responses to tips and facts or Facebook social features. Last, they interviewed depressed participants (CES-D $\geq$ 25) to assess their depressive symptoms by a psychiatrist. The results of that paper showed that Facebook activities had predictive power in distinguishing depressed and nondepressed individuals.

Tsugawa et al. [14] evaluated the effectiveness of using a user's social media activities for estimating degree of depression. They used the results of a web-based questionnaire for measuring degree of depression of Twitter users. For feature extraction for estimating the presence of active depression, they extracted several features from the activity histories of Twitter users. That paper showed that (a) features obtained from user activities can be used to predict depression of users with an accuracy of 69%, (b) topics of tweets estimated with a topic model are useful features, (c) approximately two months of observation data are necessary for recognizing depression, and longer observation periods do not contribute to improving the accuracy of estimation for current depression; sometimes, longer periods worsen the accuracy.

In another research, Shen et al. [15] constructed well-labeled depression and non-depression dataset on Twitter, and extract six depression-related feature groups covering not only the clinical depression criteria, but also online behaviors on social media. They proposed a multimodal depressive dictionary learning model to detect the depressed users on Twitter, and analyzed a large-scale dataset on Twitter to reveal the underlying online behaviors between depressed and non-depressed users. Their proposed model—Multimodal Depressive Dictionary Learning (MDL)—achieved the best performance with 85% in F1-Measure. In another similar research [16] Shen et al. studied a problem of enhancing detection in a certain target domain with ample Twitter

data as the source domain. They first systematically analyzed the depression-related feature patterns across domains and summarized two major detection challenges, namely "isomerism" and "divergency". We further propose a cross-domain Deep Neural Network model with Feature Adaptive Transformation & Combination strategy (DNN-FATC) that transfers the relevant information across heterogeneous domains.

Based on the CLEF/eRisk 2017 pilot task, which is focused on early risk detection of depression, Stankevich et al. [17] performed a research using CLEF/eRsik dataset [18] which consists of text examples collected from messages of Reddit users. They classified users into two groups: risk case of depression and non-risk case, considering different feature sets for depression detection task among Reddit users by text messages processing. For feature extraction, they used bag-of-words, embedding and bigram models. They used support vector machines (SVM) for classification, and the best model in that research was tf·idf with morphology set as features, which achieved best results on the test data with 63% F1-score. Embedding features in that research always obtained better recall score than tf·idf but the accuracy and the precision scores were lower.

3. Methodology

The inventory we used in this research for feature extraction [3] is designed to measure the behavioral manifestations of depression and is composed of 21 categories of symptoms and attitudes that are primarily clinically derived in the course of the psychoanalytic psychotherapy of depressed patients and systematic observations and records of their characteristic attitudes and symptoms. These categories are listed in Table 1.

Table 1. The categories of symptoms and attitudes

1.	Mood	12.	Social Withdrawal
2.	Pessimism	13.	Indecisiveness
3.	Sense of failure	14.	Body Image
4.	Lack of satisfaction	15.	Work Inhibition
5.	Guilty feeling	16.	Sleep Disturbance
6.	Sense of punishment	17.	Fatigability
7.	Self-Hate	18.	Loss of Appetite
8.	Self Accusations	19.	Weight Loss
9.	Self Punitive Wishes	20.	Somatic Preoccupation
10.	Crying Spells	21.	Loss of Libido
11.	Irritability		

Each category consists of a graded series of 4 to 5 self-evaluative statements that describe a specific behavioral manifestation of depression and are ranked from 0 to 3

to reflect the range of severity of the symptom from neutral to maximal severity. The statements of three of the categories are shown in Table 2 (to see the complete statements visit [3]).

Table 2. Statement of the categories of symptoms and attitudes

Category	Statements
Sense of Failure	0 - I do not feel like a failure 1 - I feel I have failed more than the average person 2a - I feel I have accomplished very little that is worthwhile or that means any¬ thing 2b - As I look back on my life all I can see is a lot of failures 3 - I feel I am a complete failure as a person (parent, husband, wife)
Self Hate	0 - I don't feel disappointed in myself la - I am disappointed in myself lb - I don't like myself 2 - I am disgusted with myself 3 - I hate myself
Social Withdrawal	0 - I have not lost interest in other people 1 - I am less interested in other people now than I used to be 2 - I have lost most of my interest in other people and have little feeling for them 3 - I have lost all my interest in other people and don't care about them at all

To extract the features, we used the keywords of the statements in each category. First, we removed the stop words from the statements and obtained a list of words including the main verbs and the adjectives that we expect to be frequently used by the people who suffer from depression. The main idea behind our technique is that either the words from this list of words or their synonyms are frequent in the language of people with symptoms of depression; therefore, using NLTK library in python, we found the synonyms of the words in our list and added them to our list and called this list the “depressed class features” list.

On the other hand, psychoanalytical studies tell us that the people who are not suffering from depression not only do not use the words in "depressed class features" list frequently but also might use the words with the opposite meaning. For example, a depressed person might show the signs of depression through producing utterances like "I hate other people"; however, a person who is not depressed is more willing to use sentences like "I love my friends". Therefore, to distinguish depressed people from non-depressed ones, the antonyms of the words in "depressed class features" list were also important because the occurrence of these words could reduce the probability that a person is depressed. As a result, we created a second list including the antonyms of the words in "depressed class features" list and called it the "non-depressed class features" list.

Finally, our ultimate feature list was the combination of "depressed class features" and "non-depressed class features" lists. Based on this feature extraction method, which we call a bipolar feature extraction, we developed a classifier with two target classes. In addition, we develop another model based on word frequency as the feature extraction in order to compare the results of our bipolar feature extraction with this model.

Using the features we obtained, we apply a Bayesian classifier whose training is based on multinomial Naïve Bayes algorithm to classify each of the users in the test data into either depressed or non-depressed class.

3.1. Training and test data set

We use a subset of the CLEF/eRsik dataset [18], which was provided as the part of the pilot task of early risk detection of depression and consists of text examples collected from messages of Reddit users. The dataset consists of 15% of positive examples (risk case of depression) and 85% of negative examples (non-risk case), and each example of the dataset contains all text messages—between 10 to 2000 messages—of the user for a certain period of time with different time intervals. The dataset that we had access to consists of 340 examples (40 positive samples and 300 negative samples). To make a balance between the two classes, we used 25 positive examples and 50 negative examples as the training data, and allocated 15 positive and 25 negative examples for the test data set. Note that all 40 positive examples were used in this work.

3.2. Classification method

To classify the users, we use a classifier based on Naïve Bayes algorithm, which is a kind of a frequently used supervised learning method that examines all its training input and applies Bayes theorem with the "Naïve" assumption of conditional independence between features given the value of the class variable [19]. Equation 1 below shows Bayes theorem, where C stands for class variable and x_1 through x_n are dependent feature vectors:

$$P(C \mid x_1, \cdots x_n) = \frac{P(C)P(x_1, \ldots, x_n|C)}{p(x_1, \ldots, x_n)}. \quad (1)$$

There are different kinds of Naïve Bayes classifiers based on their training and classification algorithms and their *attitude* toward the data distribution. The classifier we used is based on a multinomial Naïve Bayes algorithm which is implemented to the data that is multinomially distributed and is one of the Bayes variants that is usually used in text classification [19]. The distribution is parametrized by vectors $\theta y = (\theta y_1, \ldots, \theta y_n)$ for each class y, where *n* is the number of features—the size of the vocabulary—and θy_i is the probability $P(x_i|y)$ of feature *i* appearing in a sample belonging to class y. The parameter $\hat{\theta} y_i$ is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta} y_i = \frac{N_{y_i} + \alpha}{N_y + \alpha n}. \quad (2)$$

In the equation above, $Ny_i = \sum_{x \in T} x_i$ is the number of times feature *i* appears in a sample of class *y* in the training set T, and $Ny = \sum_{i=1}^{n} N_{y_i}$ is the total count of all features for class *y*. The smoothing priors $\alpha \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing [20].

Our modified classifier

The classifier we developed in this research is a modification of Naïve Bayes algorithm; however, because of the feature extraction method we used in this research, the conditional value of some of the features in the feature vector gets doubled in certain conditions. Based on the bipolar nature of the features we extract for the two classes in this research—depressed and non-depressed people—, we modified our classifier in a way that the words in the depressed class feature list and the words in the non-depressed class feature list get a double importance when they appear in classes depressed and non-depressed respectively. Therefore, in this model:

$$\hat{\theta} y_i = \begin{cases} \frac{N_{y_i} + \alpha}{N_y + \alpha n} \times 2 & \text{if } N_{y_i} \text{ only in feature list of class } y \\ \frac{N_{y_i} + \alpha}{N_y + \alpha n} & \text{else.} \end{cases} \quad (3)$$

4. Experimental Results

To have the ratio of correctly predicted positive observations to the total predicted positive observations, and also the ratio of correctly predicted positive observations to the all observations in actual class, we used the metrics precision and recall

respectively. In addition, to take both false positives and false negatives into account, we used F1 Score, which is the weighted average of precision and recall. Equations 4, 5, and 6 show the formulae for the calculation of precision, recall, and F1 Score respectively:

$$\text{Precision} = \frac{\text{TP}}{TP+FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{F1 Score} = \frac{2\times(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{6}$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) refer to the correctly predicted depressed users, correctly predicted non-depressed users, incorrectly predicted depressed users, and incorrectly predicted non-depressed respectively. Tables 3, and 4 show the confusion matrices of the modified Bayes model and basic Bayes model, both with bipolar feature vector, and Tables 5 and 6 show the confusion matrices of modified and basic Bayes models, both with 1000 most frequent words as the feature vector.

Table 3. Confusion matrix of the modified Bayes model with bipolar feature vector

	Positive	Negative
Positive	12	3
Negative	2	23

Table 4. Confusion matrix of the basic Bayes model with bipolar feature vector

	Positive	Negative
Positive	14	1
Negative	6	19

Table 5. Confusion matrix of modified Bayes model with 1000 most frequent words as the feature vector

	Positive	Negative
Positive	10	5
Negative	1	24

Table 6. Confusion matrix of basic Bayes model with 1000 most frequent words as the feature vector

	Positive	Negative
Positive	9	6
Negative	1	24

Table 7. Comparison of the results obtained from our models

	Precision %	**Recall %**	**F1 Score %**
Modified – bipolar	80.00	85.71	82.75
Basic – bipolar	93.33	70.00	79.99
Modified – frequency	66.66	90.90	76.91
Basic – frequency	60.00	90.00	72.00

As Table 7 shows, considering F1 score, the models that used bipolar feature vectors have a much better performance than the ones with term frequency as feature extraction method. Also, it can be seen in Figure 1 that the models developed by the basic Bayes classifiers were biased to one class, but our modified Bayes classifiers were able to make a balance between the two classes; consequently, they improved the F1 score of the models based on basic Bayes classifiers. However, the performance of our modified model was much more significant for the models with bipolar feature extraction.

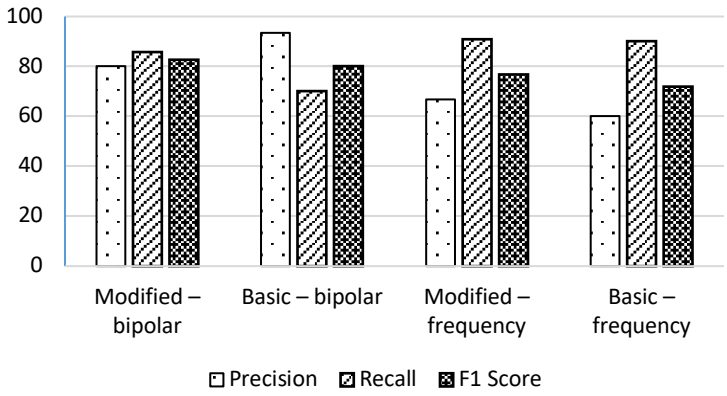

Fig. 1. Comparison of the results of different models developed in this research

5. Conclusions and Future Work

We introduced a new method for feature extraction using findings in the realms of Psychoanalysis and Psychology. We called our feature extraction method a *bipolar feature extraction* since the final feature vector we obtained from this method consists of two groups of features that are clearly opposite to each other in meaning. In addition, getting insight from a cognitive idea, we developed a modified Bayesian classifier, which improved the performance of the basic Bayesian classifiers, especially when it was used with a bipolar feature vector. The results obtained from this research showed an achievement in classifying social media users into depressed and non-depressed classes when we used our bipolar feature vector instead of word frequency. Also, the modified classifier we developed was successful in improving the performance of our models reaching the F1 score 82.75 %.

For future work, we intend to get insight from findings in cognitive and psychoanalytical studies to predict the severity of depression in users of social media.

References

1. Hidaka, B.H.: Depression as a disease of modernity: explanations for increasing prevalence. Journal of affective disorders 140, no. 3 (2012): 205-214.
2. Mathers, C. D., and Loncar, D.: Projections of global mortality and burden of disease from 2002 to 2030. In: PLoS medicine 3, no. 11 (2006): e442.
3. Beck, A. T., Ward, C. H., Mendelson, M., Mock J., Erbaugh, J.: An inventory for measuring depression. Archives of general Psychiatry 4, no. 6 (1961): 561-571.
4. Al-Mosaiwi, M., Johnstone, T.: In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science 6, no. 4 (2018): 529-542.
5. Al-Mosaiwi, M.: (2020) People with depression use language differently – here is how to spot it. The Conversation. https://theconversation.com/people-with-depression-use-language-differently-heres-how-to-spot-it-90877 (accessed July 24, 2020).
6. Paul, M. J., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: Fifth International AAAI Conference on Weblogs and Social Media. 2011.
7. Sadeque, F., Xu, D., Bethard S.: Measuring the latency of depression detection in social media. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 495-503. 2018.
8. Manning, C,. Schutze, H.: Foundations of statistical natural language processing. MIT press, 1999.
9. Kibriya, A. M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive bayes for text categorization revisited. In: Australasian Joint Conference on Artificial Intelligence, pp. 488-499. Springer, Berlin, Heidelberg, 2004.

10. Monroe, W., Potts C.: Learning in the rational speech acts model. arXiv preprint arXiv:1510.06807 (2015).
11. De Choudhury, M, Gamon M., Counts S., Horvitz E.: Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media. (2013).
12. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 47-56. 2013.
13. Park, S., Lee, S.W., Kwak, J., Cha, M., Jeong, B. Activities on Facebook reveal the depressive state of users. Journal of medical Internet research 15, no. 10 (2013): e217.
14. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 3187-3196. (2015).
15. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T., Zhu, W.: Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In IJCAI, pp. 3838-3844. (2017).
16. Shen, T., Jia, J., Shen, G., Feng, F., He, X., Luan, H., Tang, J., Tiropanis, T., Chua, T.S., Hall, W.: Cross-domain depression detection via harvesting social media. International Joint Conferences on Artificial Intelligence, (2018).
17. Stankevich, M., Isakov, V., Devyatkin, D., Smirnov, I. Feature Engineering for Depression Detection in Social Media. In: ICPRAM, pp. 426-431. (2018).
18. Losada, D. E., Crestani, F.: A test collection for research on depression and language use. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 28-39. Springer, Cham, (2016).
19. scikit-learn. 1.9. Naïve Bayes. scikit-learn https://scikit-learn.org/stable/modules/Naïve_bayes.html (accessed May 5, 2020).
20. Franco-Penya, H. H., Sanchez, L. M. Tuning Bayes Baseline for dialect detection. In Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), pp. 227–234, (2016).
21. Lytras, M.D., Mathkour, H.I., Abdalla, H., Yáñez-Márquez, Cornelio & Ordóñez de Pablos, P. Guest Editorial: The Social Media in Academia and Education: Research R-evolutions and a Paradox: Advanced Next Generation Social Learning Innovation, Journal of Universal Computer Science (J.UCS), Vol. 20, No. 15, pp. 1987-1994, (2014).
22. Moreno-Moreno, Prudenciano & Yáñez-Márquez, Cornelio. The New Informatics Technologies in Education Debate, Communications in Computer and Information Science, CCIS Vol. 19, Springer-Verlag Berlin Heidelberg, pp. 291-296, (2008).