Leveraging Multiple Characterizations of Social Media Users for Depression Detection Using Data Fusion

Karla María Valencia-Segura(✉), Hugo Jair Escalante, and Luis Villaseñor-Pineda

Language Technologies Lab, Department of Computer Science,
Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla 72840, Mexico
{valencia.karla,hugojair,villasen}@inaoep.mx

Abstract. Depression is one of the principal mental disorders worldwide, yet very few people receive the appropriate care needed due to the difficulty involved in diagnosing it correctly. Social networks have opened the opportunity to detect those users who suffer from this disease through the analysis of their posts. In this work, we propose using three types of characterizations (demographic, emotion, and text vectorization) extracted from the users' text and a fusion method for the detection of depressive users in the social network Reddit. Considering the diversity of each of the extracted characterizations, we adopted a Gated Multimodal Unit (GMU) as a fusion method. We compare this method against traditional data fusion methods and other methods that have used the same dataset. We found the proposed method improves F1-score for the depressive class by 4% when combining these three characterizations. Showing the usefulness of characterizing user content and behavior for detecting depression and highlighting the impact that data fusion methods can have in this very relevant task.

Keywords: Depression detection · Information fusion · Social media

1 Introduction

According to the World Health Organization (WHO), depression affects around 280 million people worldwide and is one of the major causes of suicide. Nonetheless, only a tiny fraction of those who suffer from this condition receive adequate treatment [1]. One of the main challenges of depression is the difficulty of diagnosing it. This stigmatized illness prevents that people with this condition to look for medical help [2]. For this reason, we continue to search for strategies that allow their appropriate detection.

Nowadays, there is a trend on research for developing improved traditional depression diagnostic methods, as those based on interviews. Such methods rely on the use of machine learning techniques for analysing user's data. Among these, there are methods based on the observation of the depression indicators that

O. O. Vergara-Villegas et al. (Eds.): MCPR 2022, LNCS 13264, pp. 215–224, 2022.
https://doi.org/10.1007/978-3-031-07750-0_20

the people could show through written communication [3,4]. This approach has gained popularity among the language processing and machine learning fields, as it has shown that it is possible to detect depression to some extent with these type of techniques [5–7].

The adoption of methods based on the analysis of textual information for detecting depression can have a huge impact now more than ever. This mainly due to the establishment of social networks as the leading communication channel of people with the world [8]. However, the difficulty of finding useful indicators of this condition from text is not trivial. Mainly because there are many factors involved that should be taken into account when dealing with users. These include, age and gender [2], personality traits [9] and even users' interests and social context [10].

In this work, we analyze different approaches to characterize users in social media with the aim objective of identifying the different depression indicators that could be captured with these characterizations. Subsequently, these features must be fused through an adequate fusion method in which we expect to capture their redundancies and/or complementarities. Our goal is to obtain a deep understanding of the contribution of each feature type and the impact of using a sophisticated data fusion method for combining information. Specifically, we explore different fusion methods, achieving better results using a multimodal fusion method (*Gated Multimodal Unit* [11]). We experimentally evaluate in a widely used dataset for depression recognition[1]. Experimental results show the effectiveness of the adopted approach, that outperforms state-of-the-art on the same dataset. Additionally, we found that the features related to emotions have a relevant impact on the recognition of depression in social media users; also, different groups of people like gender manifest their depression differently.

2 Related Work

To detect signs of depression in social networks, we seek to characterize a user based on his or her history. Works in the area have explored different characterizations to detect signs of depression. Chen et al. [5] considered an emotion-based characterization to detect depression users on Twitter; concluding that the analysis of emotions is an essential factor in detecting depression. On other hand, Preoţiuc-Pietro et al. [6] focused on estimating demographic information (age and gender) through the analysis of posts on Twitter. Obtaining a high performance to identify users with depression.

Although there are different ways of characterizing users in social networks, there is enormous diversity among them, resulting in models that ignore the existing complementarity among characterizations. Therefore there have been some efforts to adopt information fusion to this problem. Peng et al. [12] used a model based on SVM Multi-Kernel to select the optimal kernels and combat the heterogeneity of three characterizations (text from microblogs, information on the user profile, and the behavior of the user); to identify depression users

[1] https://early.irlab.org/2018/index.html.

on Sina Weibo. On the other hand, Meng et al. [13] used different modalities (facial expressions and audio) to predict depression, applying the *linear opinion pool* method as a fusion technique. Although good results were obtained, the performance could be improved if the naturalistic vocal expressions in the audio modality is improved.

Other works have been proposed to analyze textual information using the same dataset we used herein. Some of these implement fusion techniques, as in [14] where use multiple features such as linguistic metadata at the user level, bag of words, neural word embeddings, and Convolutional Neural Networks (CNN); and an ensemble was implemented as a fusion method for the final prediction. In [15] used various classification techniques (Ada Boost, Random Forest, and Recurrent Neural Network (RNN)), using a bag of words and metamaps features; obtaining the best F1-score using Random Forest. Finally, in [16] two models were built, one for the extraction of 58 features, where 18 were chosen because the combination of these provided the best results, extracted from the user's text. The second model corresponds to vectorization using doc2vec, and a voting ensemble determines if the user is depressed or not.

The previous review shows that there are many methods out there trying to detect depression from social media posts. However, these works use complex models with various features and sophisticated approaches to text representations. Instead, this work proposes a simple model that analyses and evaluates different representations (emotions, demographics, and thematic information) extracted from the user. Please note that even when fusion methods have been used to approach this problem, they usually use different modalities to represent the user, but in this work, all characterizations associated with the user were extracted from the same modality. With this in mind, we aim to show that the relevance of each feature and the use of an adequate fusion method lead to a model that improves the prediction of depressed users on Reddit.

3 Fusion of Multiple Characterizations of Users

The aim of this study is to evaluate the usefulness of a depression detection model that leverages on characterizations of social media users. On the other hand, motivated by the tentative thematic and style differences in the use of language between genders, a demographic attribute characterization was used. Finally, text vectorization is another characterization that captures thematic content. A working hypothesis of this work is that the combination of different characterizations would result in better recognition performance. In doing so, we propose the usage of *Gated Multimodal Units* (GMU) based network. The remainder of this section elaborates on the characterizations and the fusion method that was adopted.

3.1 Feature Extraction

Pre-processing: From the data collection, the pre-processing process involved removing unnecessary information, such as special characters, numbers, URLs, and punctuation marks. Once the user's text was pre-processed, four features were chosen to represent the users, which are described below:

1. **Sentiment Analysis:** This characterization was carried out using two different approaches. The first, with the help of an *NRC Emotion Lexicon* (EmoLex)[2], which is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, joy, and disgust). The second approach identifies four emotions (sadness, fear, anger, and joy) using a CNN[3].
2. **Gender:** For this characterization, a lexical resource was used that is available in the *World Well-Being Project* (WWBP)[4]. The prediction, as well as the other characterizations, is made at the post level. So that they accumulate the weights of each one and the final sum of each post represents the result of the prediction in the user's history, the gender is indicated with the sign obtained from the sum, a positive result indicates a female user and a negative result indicates a male user.
3. **Thematic:** Finally, a vectorization of the users' histories was made using a weighted TF-IDF, which assigns weights to the words to discriminate between classes (depressive and non-depressive).

The emotion features were include based on the results from Chen et al. [5], who show that emotion-based features have a positive impact on the detection of depression. On the other hand, profile information such as demographic attributes has related to depressive indicators. Therefore, it was considered to include gender. Finally, a thematic analysis provides a structured, systematic approach to understanding sentiment makes it possible to spot patterns.

3.2 Information Fusion with Gated Multimodal Units

The information obtained by the different characterizations should be feed to a predictive model that will learn to discriminate depressive from non-depressive users. Since the considered characterizations capture different aspects from the user, a model that exploits the complementariness and redundancy of characterizations is expected to result in better performance. While there are many methods to fuse information from multiple modalities for classification purposes, there is not yet an established method of proven performance, apart from the late fusion and early fusion methods. After an extensive literature search, we found a promising model that could be used for this purpose: a GMU based network. This model has the advantage of finding the best representation between late

[2] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.
[3] https://github.com/lukasgarbas/nlp-text-emotion.
[4] https://wwbp.org/lexica.html.

and early fusion by combining attributes of different modalities and determining the relevance of each modality in the model to the predictive phase.

The neural architecture to integrate a GMU started with the various features of the text were extracted. From this, these characterizations were normalized and they were passed through the GMU module (Arevalo et al. [11]) describe this module with more detail. Once the GMU module is obtained, it goes through a fully connected hidden layer, a dropout of 0.2, and finally an output layer, with sigmoid activation, to carry out the classification of the depression. The hyper-parameters established for this network were optimized using the Kerashypetune library[5] in the training data set.

4 Experiments

This section details the evaluation of this model and other data fusion methods base line for the depression detection task and also, result from a comparison of various methods is conducted to find out which performed the best.

4.1 Data Collection

For the evaluation of the proposed method we used the data set provided by the CLEF 2018 forum (eRisk2018), Table 1 reports the statistics of the data set. This data set is made up of the users' publication history and published time. For this work, only the text was considered.

Table 1. Statistics of the train and test dataset.

	Train		Test	
	Depressed	No-depressed	Depressed	No-depressed
Subjects	135	752	79	741
Number of post	49557	481837	40655	504523

4.2 Data Fusion Baselines

We considered traditional data fusion and neural networks methods as baselines: 1) *Concatenation*: Different works have concluded that a simple concatenation of representation could be good [11]. In this case, we concatenate the four characterizations to train a simple SVM with linear Kernel, 2) *Features-Union* [17]: This method concatenates results of multiple objects to create a single representation; the principal difference with concatenation is that this method assign the same weight to each modality and it is useful to combine several feature extraction mechanisms into a single. 3) *Multi-kernel:* has shown a high performance between heterogeneous data. We performed experiments with two types of Multi Kernel

[5] https://github.com/cerlymarco/keras-hypetune.

Learning with SVM: 3.1) *MLK (Average)* [18]– It's a simple wrapper defining the combination as the average base kernels– and 3.2) *MKL (GRAM)* [19]– Gradient-based RAdius-Margin optimization, this method focuses on finding the combination of the kernel that simultaneously maximizes the margin between classes while minimizing the resulting kernel's radius. Finally 4) *EmbraceNet:* This method guarantees compatibility with any learning model and deals properly with cross-modal information [20]. Moreover, this model has already been compared with several neural network fusion techniques (Late, Early, Intermediate, Compact multilinear pooling, and Multimodal autoencoders) obtaining better performance.

4.3 Performance Metrics and Results

Different measurement criteria are shown here to evaluate how well our model performed, Precision, Recall, and F1-score of the depressive class. It's important to note that in this work, we are focused on improving the F1-score of the depressive class. Various experiments were performed to evaluate the proposed method. The first evaluates the performance of each feature individually, the second evaluates the complementarity and diversity of the combination of the features, the third experiment has the objective of compare the proposed method against three data fusion methods baseline, and the last one has the intention of evaluating the proposed method against the works present in the erisk2018 forum.

As we mention before, we explored the individual performance of each of the previously described features using a Support Vector Machine as a classification method. Table 2 shows the evaluation metrics on the depressive class of the 4 characterizations extracted from the users' posts on Reddit.

Table 2. Performance in the depressive class for each characterization.

Characterizations	Number of attributes	Precision	Recall	F1-score
Gender	1	0.32	**0.62**	0.42
Thematic	5233	**0.76**	0.43	0.55
Emo-Lex	8	0.60	0.58	**0.59**
CNN-Emotions	4	0.26	0.56	0.35

Once these results were obtained, we calculated the Coincident Failure Diversity (CFD) value [21], to evaluate the diversity and complementarity between each fusion of the four characterizations. This measure is used to determine the chance that members of the same system make mistakes coincidentally.

When $CFD = 0$, it indicates that the faults are the same for all the characteristics; therefore, there is no diversity. On the contrary, a $CFD = 1$, indicates that all the faults are unique. As can be seen in Table 3, combining these four characteristics, the highest CFD is obtained, which is equal to 0.87. This result indicates that, by choosing a good fusion method, a model with better performance can be calculated.

Table 3. Coincident Failure Diversity analysis of each of the combinations of the 4 characterizations.

Characterization	CFD	Characterization	CFD
Emo-Lex, Thematic	**0.65**	CNN-Emotions, Emo-Lex, Gender	0.36
Emo-Lex, CNN-Emotions	0.49	CNN-Emotions, Emo-Lex, Thematic	**0.83**
CNN-Emotions, Thematic	0.54	CNN-Emotions, Thematic, Gender	0.82
Emo-Lex, Gender	0.15	Thematic, Emo-Lex, Gender	0.61
CNN-Emotions, Gender	0.21		
Thematic, Gender	0.48	All Characterizations	**0.87**

To properly evaluate the GMU module, we consider using four traditional data fusion methods and a neural network fusion model described above. All the methods presented here were trained and tested with the same erisk2018 dataset. For this experiment, we expected to confirm the hypothesis presented in this work, and also we expect to determine the effectiveness of the use of GMU over the traditional fusion methods.

Table 4. Comparison of the various fusion methods and GMU.

Fusion method	Precision	Recall	F1-score
SVM (Concatenation)	0.93	0.48	0.63
SVM Features-Union	0.61	0.58	0.59
MKL-Average	**1.0**	0.46	0.63
MKL-Gram	**1.0**	0.42	0.59
EmbraceNet	0.59	0.72	0.66
GMU	0.58	**0.87**	**0.70**

As shown in Table 4, the GMU module achieves the best F1-score of 0.70 in depressive class, indicating that it learns more efficiently diversity between the four characterizations compared to the other fusion methods. Additionally, we compared the results obtained in this work against the works presented in the eRisk2018 evaluation forum. The results can be observed in Table 5.

Table 5. Results of the depressive class vs. the top three places in eRisk 2018.

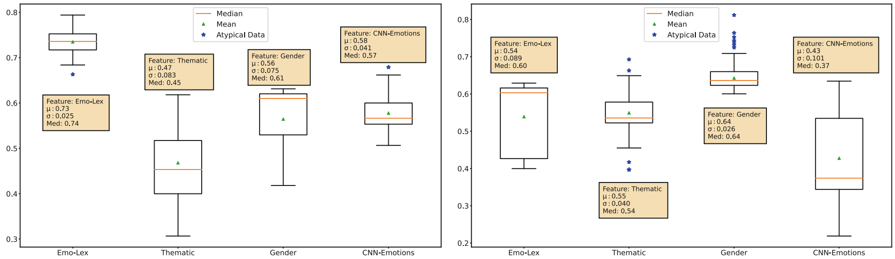
Model	Precision	Recall	F1-score
FHDO-BCSGB [14]	**0.64**	0.65	0.64
Random Forest [15]	0.63	0.64	0.63
LIIRB [16]	0.38	0.67	0.48
GMU Model	0.58	**0.87**	**0.70**

The following can be highlighted from the results obtained: (i) Implementing a GMU in a simple neural architecture outperform traditional fusion techniques, which indicates that fusion of the multiple features at a deeper level is relevant for detecting depression; (ii) The approach outperformed the top ranked methods from eRisk 2018. It is essential to note that the participants tested different complex models with a wide range of characteristics with traditional fusion methods such as late and early fusion. At the same time, the one presented here was only based on four characterizations and a GMU module as a fusion method.

5 Analysis of the Proposed Model

Figure 1 illustrates how GMU weighted the relevance of each feature according to each observation, showing the standard deviation. We inspected the z_i gates of the GMU module, averaging the activations of each class (depressive and non-depressive) in the entire test set.

In Fig. 1a we can observe the standard deviation of depressive class where thematic characterization has the higher standard deviation of 0.087; this indicates that with this characterization, we have higher variability in the activation's relevance. Conversely, Emo-Lex has a lower standard deviation of 0.025, which means that the relevance activation in this characterization remained more uniform than the rest. On the other hand, for the non-Depressive class, Fig. 1b shows a higher standard derivation in CNN-Emotions with 0.101 and the lower in for Gender with 0.024, however we have atypical data for gender. In this case, the mean value makes it appear that the data values are higher than they really are.



(a) Gates Activation's of Depressive Class(b) Gates Activation's of Non-Depressive Class

Fig. 1. Standard deviation, median, and mean of GMU unit activations for each class across the entire test set.

In general, Emo-lex and Gender for depressive and non-depressive classes, respectively, were the most relevant characteristics according to the GMU module. This result is expected because depressed users tend to make more emotional posts than non-depressed users, and several studies confirm a difference in depression signs between women and men.

GMU Error Analysis: Based on the results obtained, we evaluated and determined why some users were misclassified. In the case of the depressive class, ten users were misclassified, and in the case of the non-depressive class, fifty users were misclassified. The main reasons for these errors in both classes was due to the amount of information. Either because of very short history of posts (45.9 posts per history average in misclassified depressive users; 373.4 post in misclassified non-depressive users) or very large histories (with averages of 2407 for misclassified depressive users and, 14150 for misclassified non-depressive users). In the latter case, having records that span a long time is ineffective because the users could have received some treatment. Hence, the signs of depression are not so clear for the model.

6 Conclusions

With the development of the Internet, social networks provide a new approach to identifying those users who present indicators of depression. To do this, we propose a GMU-based depression recognition model. First, we analyzed and extracted characteristics with greater diversity and complementarity to represent users in social networks. Then we integrated a GMU module to fuse these characterizations to improve the identification of depressed users. We compared the performance of GMU with other fusion methods such as SVM, MLK (Average), MKL (GRAM), Features-Union and EmbraceNet, to classify the detection of depressive users, where GMU showed a better performance compared to this fusion methods. In addition, we compared the results obtained in this work against the works presented in the eRisk2018 evaluation forum, showing that our model exceeded the F1-score brought by the first place in the forum. It should be noted that the simplicity and capacity of this model contrast with the first places obtained in the evaluation forum of erisk 2018. Finally, as far as we know, the work presented in this study is the first work to recognize depression that uses a GMU, and we obtained an F1-score of 70%. We expect to replicate this work in another depression collection data; we also intend to extract and integrate other relevant features such as personality as future work.

Acknowledgments. This work was supported by CONACyT/México (scholarship 782579 and grant CB-2015-01-257383). Authors thank CONACYT for the computational resources provided by the Deep Learning Platform for Language Technologies.

References

1. Skaik, R., Inkpen, D.: Using social media for mental health surveillance: A review. ACM Comput. Surv. **53**(6), 1–31 (2020)
2. “Depressión”. World Health Organization (2020). https://www.who.int/es/news-room/fact-sheets/detail/depression
3. Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., Wolf, M.: First-person pronoun use in spoken language as a predictor of future depressive symptoms: preliminary evidence from a clinical sample of depressed patients. Clin. Psychol. Psychotherapy **24**(2), 384–391 (2017)

4. Bucci, W., Freedman, N.: The language of depression. Bull. Menninger Clin. **45**(4), 334 (1981)
5. Chen, X., Sykora, M., Jackson, T., Elayan, S., Munir, F.: Tweeting your mental health: an exploration of different classifiers and features with emotional signals in identifying mental health conditions. In: Proceedings of 51st HICSS Conference (2018)
6. Preoţiuc-Pietro, D., et al..: The role of personality, age, and gender in tweeting about mental illness. In: Proceedings of the 2nd workshop on CLPsych, pp. 21–30 (2015)
7. Aragón, M.E., López-Monroy, A.P., González-Gurrola, L.C., Montes, M.: Detecting depression in social media using fine-grained emotions. Proc. NAACL-HLT **1**, 1481–1486 (2019)
8. Baruah, T.D., Kanta, K., State, H.: Effectiveness of social media as a tool of communication and its potential for technology enabled connections: a micro-level study. Int. J. Sci. Res. Publ. **2**(5) (2012)
9. Kendler, K.S., Gatz, M., Gardner, C.O., Pedersen, N.L.: Personality and major depression: a swedish longitudinal, population-based twin study. Arch. Gen. Psychiatry **63**(10), 1113–1120 (2006)
10. Jackson, P.B., Williams, D.R.: Culture, race/ethnicity, and depression. In: Women and Depression: A Handbook for the Social, Behavioral, and Biomedical Sciences, pp. 328–59. Cambridge University Press, New York (2006)
11. Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated multimodal networks. Neural Comput. Appl. **32**(14), 10209–10228 (2019). https://doi.org/10.1007/s00521-019-04559-1
12. Peng, Z., Hu, Q., Dang, J.: Multi-kernel SVM based depression recognition using social media data. Int. J. Mach. Learn. Cyb. **10**(1), 43–57 (2019)
13. Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (2013)
14. Trotzek, M., Koitka, S., Friedrich, C.M.: Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In: Proceedings of the 9th CLEF Association Conference (2018)
15. Paul, S., Jandhyala, S.K., Basu, T.: Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In: Proceedings of the 9th CLEF Association Conference (2018)
16. Ramiandrisoa, F., Mothe, J., Benamara, F., Moriceau, V.: Irit at e-risk 2018. In: Proceedings of the 9th CLEF Association Conference (2018)
17. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Lea. Res. **12**, 2825–2830 (2011)
18. Lauriola, I., Aiolli, F.: Mklpy: a python-based framework for multiple kernel learning. CoRR, vol. abs/2007.09982 (2020)
19. Lauriola, I., Polato, M., Aiolli, F.: Radius-margin ratio optimization for dot-product boolean kernel learning. In: Lintas, A., Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) ICANN 2017. LNCS, vol. 10614, pp. 183–191. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68612-7_21
20. Choi, J.-H., Lee, J.-S.: Embracenet: a robust deep learning architecture for multimodal classification. Inf. Fus. **51**, 259–270 (2019)
21. Wang, W.: Some fundamental issues in ensemble methods. In: International Joint Conference on Neural Networks, pp. 2243–2250, IEEE (2008)