Depression Detection From Language Use

Anonimni autori

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{ana.bertic,roko.buca,tvrtko.sternak}@fer.hr

Abstract

Depression detection is a burning problem in today's age of technology. Studies show that it is possible to predict the person's mental state by examining his or her language use. On social media there is an abundance of written text by its users which can be harvested and used to identify individuals at risk of depression. In this paper, it is shown that with our methods of careful feature selection and class imbalance handling, our deep neural network performs above current state-of-the-art algorithms when trained on research collection of Reddit posts. We also proved that despite the fact that simpler models obtain promising results when using our feature extraction, deep models nevertheless exceed their performance pushing the F1 score to 0.67 from 0.64.

1. Introduction

The World Health Organization estimates that more than 300 million people worldwide are now living with depression, which is 18 percent more than 2005. Since depression is often underdiagnosed, there has always been a need to seek effective strategies for early detection, intervention and appropriate treatment of diagnosed individuals. It has been shown that machine learning algorithms can solve the problem of detecting early signs of depression from language use. Our work concerns improvement and validation of possible models, as well as finding new and untried ones that can be used for this purpose. The datataset used for training and testing our approach is a test collection for research on depression and language use (Losada & Crestani, 2016). Dataset is composed of 892 Reddit users, precisely documents containing their post histories which are binary classified in depression and control groups.

Data preprocessing in this paper consists of standard POS tagging and lemmatization techniques followed by four types of feature extractions. Three bag-of-words (BoW) models were created with an additional extracted and vectorized "special" features (Section 3.1.).

Our deep neural network (DNN) model, which is the focus of this paper, uses feature selection based on χ^2 word scores, simple oversampling techniques seen in Section 4. and additional features which were shown to reliably predict early depression signs (Wang *et al.*, 2013). DNN uses fixed length feature vectors where each vector represents an entire post history of a given user.

Additionally, we compare our DNN model to itself when various feature selection methods are used, as well as with previous work in this field and ensambles of simpler machine learning models. We show that it is possible to outperform current state-of-the-art models with ensambles which use our feature extraction techniques.

2. Related Work

Even though in the age of social media a massive amount of information can be extracted from social networks, for a long time there has been a lack of publicly available data for doing research on detecting language patterns associated with depression. Prior researches usually included analyses from microblogs like Twitter, proving that large platforms reflect the expression of depression both explicitly and implicitly.

However, in case of Twitter, the limit of 280 characters is too small to provide any meaningful context. We used a collection from open-source platform Reddit carefully constructed by (Losada & Crestani, 2016) where each subject is represented via document of a large number of posts and comments submitted during a longer period.

Numerous text mining techniques have been developed based on background knowledge of psychologycal reasearches (Wang *et al.*, 2013), some of them including sentiment analysis, linguistic rules and vocabulary construction. However, there has still not been substantially significant improvement of results in conducted related work and existing tehniques generally rely on experimenting with baseline models in hope of achieving better results which rarely exceed over 60% F1 score.

For sentiment analysis a lexicon and rule-based sentiment analysis library VADER (Gilbert, 2014) is used. It is found to work well when integrated into machine learning models and easily handles text originated in social media. VADER is also used in (Leiva Aranda, 2017) representing additional features during data preprocessing. It contains methods which retrieve percentage of negative, positive and neutral sentiment representative words and calculates overall sentence sentiment polarity.

In this paper we, use machine learning models and ensembles similar to that of (Polikar, 2006) in order to test our feature extraction independent of our modeling methods. Based on the success of our feature extraction methods we developed a DNN which showed improved performance when using those extracted feature vectors.

In our final testing, we compare our model to the best model presented in an overview of a competition (Losada *et al.*, 2017) based on the dataset we used.

3. Feature Evaluation

Corpus constructed from words used by subjects in training set consisted of more than 120,000 distinct words which are mostly typos, links and numbers that do not have any dis-

criminatory interpretation. With these observations, it was necessary that the first step in our approach to depression classification is feature evaluation and extraction.

We decided to observe results based on three approaches. Effectiveness of each approach on the validation set is shown in Table 3.

• **augmented frequency interval feature selection:** Every lemmatized word in the corpus was counted and several word intervals based on word frequency were chosen for testing. Three sizes of word vectors were used: 200, 400 and 1500. We found that models that used word vectors bigger than 200 were prone to overfitting, even with significant dropout added to neural network layers. Frequency interval that proved to work best was the one that contained words with in-corpus frequency being between 6.5% and 3%. Chosen word vector was then augmented by removing numbers and other non-words which were then replaced with words that proved to be of value in previous model selection testing.

• **term frequency-inverse document frequency feature selection:** For each word a tf-idf weight was calculated. This value represents the importance of a word in a document, considering whole corpus of the train set. Based on weights, tf-idf matrix was constructed in which each row represented a tf-idf value for each word in corpus for a single document. The matrix rows were labeled into a document of depressed and non depressed user and passed to a random forrest classifier in order to extract most important features for this classification. 247 words were extracted and manually reduced down to 200 by removing words with no semantic meaning for classification.

• χ^2 **feature selection:** After extraction of BoW representation for each user in train set best 150 words were determined using Chi2 algorithm based on χ^2 statistic (Liu & Setiono, 1995). The limit of 150 words was chosen based on observation of words that appeared when the limit increased, most of those words were typos or numbers and did not represent any objective semantic meaning. χ^2-based feature selection produced models with best generalization properties especially when augmented with additional features and for that reason was used by our DNN model.

3.1. Special Data

Inspired by the work of (Wang *et al.*, 2013), we did a similar analysis of the corpus in order to find additional features with discriminatory value.

Much like previous works, we found that the use of emoticons was a powerful predictor of somebody's mental state. We managed to recreate the general gist of the (Wang *et al.*, 2013) and also found that use of personal pronouns in either singular or plural form can be used as another feature for detecting early signs of depression.

There were some other factors used which can be found in Table 1. Even though we ran our statistical evaluation on both the train set and test set and found some differences, we modeled our special feature vectors based only on the evaluation of the train data. Additionally, it is important to point out that some models had more trouble than others with using this extra feature vector. Our final DNN presented in this paper does not use all nine features, but instead uses the best four features based on the train data evaluation. We chose specifically four features based on both the empirical experimentation and the obvious gap between the fifth and sixth most relevant feature. We did not use the 'emoticon use' feature since that knowledge is already contained in other used emoticon statistics.

4. Oversampling Method

To remedy the problem of severe class imbalances in used dataset (ratio of depressed/control groups is around 1/5), we used various oversampling methods.

In our baseline model we used Synthetic Minority Oversampling Technique (SMOTE) proposed in (Chawla *et al.*, 2002). SMOTE generates synthetic minority class instances between existing minority instances in euclidean space. Final experiment presented in Table 4 shows that this approach significantly improves models F1 score on the test set.

Since DNN uses an extra special feature vector which contains different diferent type of data and additonally expand the dimensionality of the input vector we resorted to using other, simpler, oversampling technique. Used techinque artifically increases weights of depressed subjects by copying positive training examples five times in order to equalize the dataset imbalance.

5. Classifier Model

This paper presents two different and tested improvements to previous solutions, especially in regards to (Losada & Crestani, 2016) since our proposition shares the most similarity in approaching the problem of detecting early signs of depression.

We propose a user-level representation of data based on BoW models for the first step of the model improvement, we do not use special features in this step. We then apply models and ensambles similar to those used in previous works. To make it more fair, we separately train those machine learning models in order to achieve the best possible results on the feature vector types we use.

For the second step of the model improvement, we present the best DNN we constructed for this problem. The testing showed that the feature complexity can be much better described with the complex function DNN provides than with simpler machine learning algorithms. The final DNN achieves 0.67 F1 score compared to 0.65 of our ensamble. However our calculations show that this improvement is not statistically significant. Both the DNN and the ensamble work best when using the χ^2-based BoW vectors.

5.1. Baseline Classifier

After evaluation of the dataset and its size, we concluded that the best way to tackle the problem of depression classification is by using an ensamble of simple classifiers. The ensamble consists of five individually fine tuned classifiers on train set. In Table 2 we justify the usage of each classifier in the final ensamble by observing that the F1 score on 5

Table 1: Special features average statistics on train set (features used in our model are bolded)

Special feature (average)	Depressed users	Non-depressed users
emoticon use	26.7952	13.87096
positive emoticon use	20.1084	10.8064
negative emoticon use	6.6506	3.0074
sentence length	17.8904	20.3568
first person pronoun use	0.04573	0.024899
plural first person pronoun use	0.00266	0.00308
posting time	23:39:09	22:47:41
number of posts between 00:00 and 06:00	30.45%	27.46%
number of posts between 09:00 and 16:00	17.39%	21.06%

fold cross validation with χ^2 feature selection of ensamble is highest when all five classifiers are used.

Models used in ensamble and their hiperparameters were:

• **Logistic regression:** Regularization was done with L2 penalty, and regularization strength λ was set to 20. We did not use oversampling when validating the classifier so class weight parameter was set to be directly proportional to class frequencies in the input data ('balanced').

• **Random forest**

• **SVM:** C-support vector classifier with rbf kernel. Penalty parameter of the error term was set to 0.35 and class weight parameter was also set to 'balanced' like in logistic regression.

• **Ridge classifier:** Regularization strength α was set to 2.5

• **Ada boost:** Number of classifiers was set to 150 and learning rate to 1.

Table 2: ensamble f1 scores on validation with leaving one classifier out on each validation

Ensamble	F1 validation score
w/o logistic regression	0.57
w/o random forest	0.61
w/o SVM	0.70
w/o ridge	0.70
w/o ada boost	0.59
full ensamble	0.72

5.2. Deep Neural Network

DNN presented in this paper is a fairly simple model. It has four hidden layers which use the ReLU activation function. With all data already preprocessed and loaded into RAM, it takes only a couple of seconds to converge when run on a standard Intel i5 processor. Model has probabilistic output as defined by two neurons with softmax activation function in the last layer. To determine the class a datapoint belongs to we simply use an argmax function.

We tuned all hyperparameters by doing an extensive search. 10-fold validation setup was used to determine the best models. During optimization, we only used F1 score as the accuracy score is not very useful due to the inability to detect overfitting on the imbalanced dataset. Even though we used SMOTE oversampling while training the baseline, we found that a simple artificial data weighing worked better for the DNN.

Several different approaches were attempted and compared. The bare DNN that uses only the main BoW vector represented dataset performs similarly to the ensamble with a F1 score of 0.63. In order to additionally improve our models we implemented additional features (as described in the section 3.1.) to be used by the DNN. Best DNN constructed uses only the top four most significant features from the special feature vector. DNN performs better when the special feature vector is concatenated to the BoW vector and fed to the network, rather when the cascade of classifiers is used. This shows that there is an implicit interaction between χ^2-based BoW model and the special features which the ensamble of classifiers cannot fully encompass.

6. Experiments

Table 3: Full ensamble F1 scores on validation with preprocessed data using feature selection methods

Feature extraction method	F1 validation score
w/o feature selection	0.69
w/ afi feature selection	0.7
w/ tf-idf feature selection	0.75
w/ χ^2 feature selection	0.72

In our final experiment we compare F1 scores of our proposed models to those from previous works on the standard test set. Models that we compare ours to are the proposed baseline from (Losada & Crestani, 2016), model proposed in (Leiva Aranda, 2017) and the current state-of-the-art FHDOA model proposed in (Losada *et al.*, 2017). We show in Table 4 that both of our models outperform all other models F1 score-wise (0.65 and 0.67 compared to 0.64 achieved by FHDOA). We also point out that our neural network with

Table 4: Final results on test sets comparing to (Leiva Aranda, 2017),(Losada & Crestani, 2016) and (Losada *et al.*, 2017)

Model	p	r	f1
Genetic Algorithm + VADER	0.45	**0.77**	0.57
Logistic regression	0.64	0.59	0.62
FHDOA	0.61	0.67	0.64
Ensamble (afi)	**0.76**	0.42	0.55
Ensamble (tf-idf)	0.73	0.46	0.57
Ensamble (χ^2)	0.75	0.56	0.65
Ensamble (χ^2) w/o oversampling	0.49	0.74	0.37
Bare neural networks (χ^2)	0.65	0.61	0.63
Augmented neural networks (χ^2)	0.65	0.68	**0.67**

augmented feature vectors has the most balanced precision-recall score.

Without using special features our ensamble performs better than our neural network. This is probably due to relatively small train set size. However, when we add special features the ensamble suddenly performs worse while the neural network significantly improves its F1 score. We conclude that this type of augmented feature vector contains a high level of feature interactivity that is better suited for deeper models.

7. Future work

Various improvements could be done to boost the presented model. With the help of psychological and deeper statistical research, more special features could be added to the system. It would also be beneficial to include sentiment polarity as an additional feature. More complex models could be created that are compatible with a more fine-grained way of data representation instead of the used coarse user-level one.

Given that our models are easily overfitted on the available data it is probable that simply adding more data would automatically improve models' performance.

8. Conclusion

In this work we experimented with both machine learning and deep learning models in hope to design a model that is effective at recognizing signs of depression in Reddit users based on their post histories. Starting with ensembles of simple classifiers, we combined three different feature selection methods and studied the results. Models that use χ^2 feature selection proved to be the most successful in this task, surpassing other state-of-the-art models' F1 score by a significant margin. Additionally, we focus on one particular DNN model which reliably outperforms ensambles aforementioned.

The model is augmented with additional text-mined features which were shown to contain significant linguistic markers in related works. Our results have proven that the classification of depressed users is possible with deep models as well as with ensambles of simple models, both of which outperform current state-of-the-art techniques when using the feature extraction presented in this paper.

9. Acknowledgments

We would like to use this opportunity to thank our professor and mentor, Jan Šnajder for the invaluable knowledge of this field that he has provided us. We would also like to thank Mladen Karan for suggesting the χ^2 feature evaluation and for showing us how to perform statistical significance testing.

References

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.*

Víctor Leiva Aranda. 2017. Towards suicide prevention: early detection of depression on social media.

Huan Liu and Rudy Setiono. 1995. Chi2: Feature selection and discretization of numeric attributes. In *Tools with artificial intelligence, 1995. proceedings., seventh international conference on*, pages 388–391. IEEE.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations.

Robi Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45.

Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213. Springer.