Using Linguistic Metadata for Early Depression Detection in Social Media

Andrija Perušić, Denis Kustura, Ivan Matak

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{andrija.perusic, denis.kustura, ivan.matak}@fer.hr

Abstract

The goal of this paper is to use social media network as source for predicting early signs of depression. For that purpose we process posts written by *Reddit* users. The dataset consists of 892 labeled users. Along with *tf-idf* document vectorization we use hand-crafted features in our models. Purpose of those hand-crafted features is to detect and exploit subtle signs of depression in social media text which will in turn improve evaluation result. We use three classifiers: logistic regression classifier (LR), SVM classifier and voting classifier (consisting of LR and SVM combined). Further we evaluate all three models and perform ERDE measure evaluation on all three of them. Also we present those results in the paper and show that hand-crafted features have a positive impact on evaluation.

1. Introduction

Depression negatively affects how we feel, the way we think and the way we act. It causes the feeling of sadness and a loss of interest in any activity. By WHO statistics, it affects more than 300 million people with an increase of depression of more than 18% between 2005 and 2015.

This worrying statistics impose question how to detect depression as early as possible before any significant damage is caused. One of possible, and logical, answers is to examine the area where most of the people nowadays spend lot of their time which is social media. Social media is area where human social behavior comes to the fore and where people often express their feelings and problems so that makes it ideal environment for depression detection. A lot of work has been done in the area of early depression detection through NLP and machine learning, both deep learning approach and basic model approach with feature engineering. However, even though social media is a rich source of data, detecting depression is still hard because signs of depression are subtle.

In this paper, we concentrate on finding those subtle signs of depression, by modeling existing knowledge about language use of depressed individuals into set of features for our classification models. Models are trained and tested on a dataset which consists of labeled *Reddit* posts. In following sections we present detailed explanation of dataset, additional features along with a description and evaluation of models that use them.

2. Related work

A lot of work has been done in development of quality tools for early depression detection and there are constant improvements in this area. Also a lot of psychology studies have been conducted to further explore language use of depressed individuals.

We use results from several of these studies to hand-craft informative features. In the psychology study of (Rude et al., 2004) essays written by college students were examined and it was shown that elevated use of first-person pronouns (especially word ”I”) is connected to depression. In psychological research of (Al-Mosaiwi and Johnstone, 2018) over 63 different Internet forums absolutist words were used as indicators for depression and we use this list of words.

In area of developing tools and methods for early depression detection (Losada and Crestani, 2016) presented a new dataset of *Reddit* posts collected over long period of time which stores information about depression evolution. We train and test our models on this dataset. They also introduced new evaluation metrics, *ERDE* measure, which is designed especially for detecting early signs of depression. We used this measure for evaluation.

In work of (Trotzek et al., 2017) different models (BOW models RNN, LSTM), were employed for depression detection. Also in their work they used hand-crafted features beside standard document vectorization methods to improve evaluation results. Another work that concentrates on feature engineering is (Stankevich et al., 2018) where they used stylometric and morphology features with standard document vectorization methods and showed that aditional features have an impact on improving evaluation scores.

In this paper, we will examine effectiveness of *tf-idf* document vectorization with our additional features that were made by exploiting all of previous knowledge of psychological studies on depressed people and by using *Empath* (Fast et al., 2016), text analysis tool that can generate and validate new lexical categories on demand.

3. Dataset

For training and testing we used dataset that consists of *Reddit* posts collected by (Losada and Crestani, 2016) from 892 users. It consists of 137 depressed users and 755 non-depressed users and it is split into 486 users for training set (83 depressed and 403 non-depressed) and 406 users for test set (54 depressed and 352 non-depressed).

Dataset is created as sequence of XML files, one file per user. Each XML file stores the sequence of the users' submissions (one entry per submission). For most active users there are up to 2000 submissions (1000 posts and 1000 comments) and those submissions include submission to any *subreddit* category. Every submission in XML file is represented with submission's title, text and date. Also

Table 1: Most informative semantic categories.

	Average		
Category	Positive	Negative	Difference
friends	0.00846	0.00505	**0.00341**
positive_emotion	0.00899	0.00597	**0.00302**
negative_emotion	0.01224	0.00924	**0.00300**
love	0.00521	0.00259	0.00262
nervousness	0.00419	0.00159	0.00260
pain	0.00594	0.00334	0.00260
shame	0.00489	0.00236	0.00253
optimism	0.00652	0.00412	0.00240
sadness	0.00361	0.00145	0.00216
speaking	0.00910	0.00696	0.00214

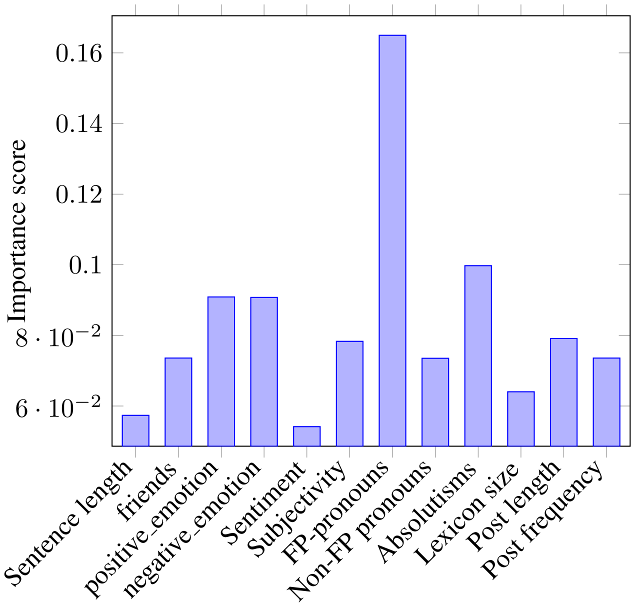

Figure 1: Feature importances.

posts in every XML file are ordered chronologically for every user. This dataset is specific because it does not only allows us to find differences in language use between depressed and non-depressed users but it also allows us to see evolution of language use of depressed users during a longer period of time.

4. Baseline experiments

The *bag-of-words* approach was taken for our NLP pipeline. For each user, all post were read sequentially and from each post the contents of tags *TITLE* and *TEXT* were concatenated. Other tags do not have information suitable for this task. The resulting document for each user was a concatenated string containing all their submission content. The data is raw and often titles and text contain noise (links, titles from a different source, etc.). We used multiple regular expressions to extract and remove this noise from data so that our final concatenated document contains only text written by a user. Finally, each cleaned document was then tokenized and stemmed with a *Snowball stemmer*. As the last step we made a *tf-idf* vectorized representation of cleaned, tokenized and stemmed document.

Classification was done by logistic regression (LR) with L1 regularization and SVM with linear kernel to asses the difference in performance. This dataset contains 755 users in control group and only 137 users labeled as depressed so it is extremely unbalanced classification problem. To address this issue and avoid building a trivial classifier that classifies everything as not depressed we adjusted misclassification costs for each class. Class weight is inversely proportional to class frequencies in the input data and is calculated by (1)

$$w_i = n_s/(n_c * c_i) \tag{1}$$

where n_s is number of samples in a dataset, n_c number of classes and c_i is number of samples labeled as class i in the dataset.

Both baseline models where optimized for optimum performance with standard grid search of two hyper-parameters: k - number of k-best features to retain, calculated by a χ^2 distribution and C - a regularization factor. Each combination was evaluated by 5-fold cross validation on training set optimizing for F1 score of minority class. Cross validation process gave following optimal values $k = 1500$ for both models and $C = 16$ for Logistic Regression and $C = 8$ for SVM.

5. Linguistic metadata features

Next step we took was attempt at modeling the existing knowledge about language use of depressed individuals as extra features that we later added to our baseline models. In addition to those features, we added several features that were based on our intuition about this classification problem. Total of 12 new features were created. In the next part of this chapter we list and explain all of new features.

Semantic feature set: In (Losada et al., 2017), analysis of emotions and topic signals with lexicons like LIWC (Linguistic Inquiry and Word Count) is advised, as it is deemed beneficial for depression detection. For out semantic features, as we call them, we used *Empath*. *Empath* is a lexicon mined from modern text on the web with 196 built-in, pre-validated lexical categories(Fast et al., 2016). Its categories are highly correlated with similar categories in LIWC. The reason we used Empath over LIWC is because it is open-source, has more potentially useful categories and is overall a bigger lexicon.

For a given text, Empath returns the count of words correlated to each of its lexical categories, normalized over total number of words. To extract categories useful for our problem, we summarized the normalized category scores for all positive and negative training examples and averaged the results separately. We then calculated the difference between the two averages for each category and picked the top 10 categories. To further optimize feature selection, we took only categories with difference score greater than difference mean average of top 10 categories. After optimization we were left with three lexical categories that show biggest distinction in language use between depressed and non-depressed individuals. Top categories are shown in Table 1.

Pronoun frequency: There is confirmed correlation of elevated use of first person pronouns and depression (Karmen et al., 2015). In contrast, there is also correlated reduced usage of non-first person pronouns. This reflects the phenomenon that depressed individuals are often focused on themselves and much less on others. We constructed two features. One that represents first-person pronoun frequency and the other for non-first person pronoun frequency for each user. Frequency is calculated as total number of pronouns divided by word count of most frequent word in a document.

Absolutisms frequency: Absolutist thinking is considered a cognitive distortion by most cognitive therapies for anxiety and depression and they track the severity of affective disorder more faithfully than negative emotion words (Al-Mosaiwi and Johnstone, 2018). We use a list of absolutist words from this paper, augmented by additional words we added, to devise a feature in a same way as pronoun frequency - number of absolutist words divided by word count of most frequent word in a document.

Sentiment and subjectivity: Negative sentiment is also often sign of depression or anxiety (Al-Mosaiwi and Johnstone, 2018). We use a learned model from *TextBlob* library to extract sentiment and subjectivity information. Model returns polarity $[-1.0, 1.0]$ and subjectivity $[0.0, 1.0]$ score for each sentence. We use this to devise two features as polarity (sentiment) and subjectivity averaged across sentences of a document.

Lexicon size: This feature is calculated as number of different words divided by total number of words in a document. This feature is simplifed version of a set of features proposed in (Trotzek et al., 2017) where they use measures for text readability (namely Gunning Fog Index, Flesch Reading Ease, Linsear Write Formula, New Dale-Chall Readability). Those measures store information about language complexity for an individual, which can be connected to depression.

Other: We created 3 additional metadata features on intuitive presumption they hold information that can improve our models. Their impact was tested in evaluation process. These features are average sentence length, average post length and post frequency. Average post and sentence length are calculated on a word basis and post frequency is calculated as number of minutes between first and last post divided by number of posts in a dataset for a given user.

All of the additional features were standardized by removing the mean and scaling to unit variance. The idea was to make a more sophisticated model by adding these new features to existing 1500 *tf-idf* baseline features and evaluate their performance.

6. Evaluation

As a first test we did not optimize feature selection. All new features were added to both baseline models and only regularization factor was optimized. It is important to note that all hyper-parameter optimization in our tests was done by 5-fold cross validation on a training set, optimizing for F1 score of minority class. This way the unbalanced nature of dataset was addressed. As we predicted, models had poor overall performance because there was a lot of noise in the dataset and subsequently, our features. For logistic regression the cross-validation score was worse by more that 3% and for SVM 0.02%. Nevertheless, we tested these models with optimized regularization factor on a test set and got the result shown in Table 2.

It should be noted that, even though the overall F1 score is worse than baseline, for LR with all features we got the biggest recall on test set. This is arguably very important trait because it is better to have false positives that can be later labeled as non depressed that completely miss depressed cases.

The next step was feature selection. For 12 new features, there are 2^{12} possible combinations and it was too resource intensive to cross-validate all combinations as a hyper-parameter of a model. Thus we used tree-based feature selection method. Tree-based estimators can be used to compute feature importances, which in turn can be used to discard irrelevant features. We used extremely randomized trees classifier with 500 trees in the forest and fitted it with only the new features (without *tf-idf* features) and labels of the training set. Calculated feature importances are shown in Figure 1. The threshold for labeling feature as important was mean average of all importances and this approach selected following features, sorted by level of importance: *FP pronouns, absolutisms, positive_emotion* and *negative_emotion*.

We added the optimal feature subset to baseline features and again cross-validated the regularization factor. We ended up with following values: $C_{LR} = 22$, $C_{SVM} = 10$. In case of logistic regression cross validation score showed 1% improvement over baseline, and for SVM, there was a slight improvement of 0.1%. Results on a test set show improvement over baseline models with roughly 2% and 1.5% rise in precision for LR and SVM respectively while retaining recall score.

With an effort to further improve we combined two models into soft voting ensemble which predicts the class label based on the argmax of the sums of the predicted probabilities of each model. Ensemble gave the best best precision score on a test set.

We have also conducted significance tests, where we used paired t-test on F1 score, at 95% of significance level and results showed that our SVM with optimum feature subset outperforms both baselines and our LR model. However, it did not outperform our voting classifier. On other side, our LR model failed to outperform any of other models on 95% significance level. Even though our voting classifier showed improvement on evaluation over all models presented in the work it has not succeeded to outperform nor baselines nor SVM on 95% significance level.

Table 2: Model results on test set.

Model	Precision		Recall		F1	
	Minority	Overall	Minority	Overall	Minority	Overall
Baseline						
Logistic Regression	0.507	0.893	0.704	0.870	0.589	0.878
SVM	0.520	0.897	0.722	0.874	0.605	0.883
All additional features						
Logistic Regression	0.462	0.895	**0.778**	0.850	0.579	0.865
SVM	0.524	0.884	0.611	0.874	0.564	0.878
Optimized feature subset						
Logistic Regression	0.528	0.896	0.704	0.877	0.603	0.884
SVM	0.534	0.899	0.722	0.879	0.614	0.887
Soft voting (LR + SVM)	**0.607**	0.898	0.630	0.897	0.618	**0.897**

Table 3: Comparison of ERDE results for our final models. * - (Losada and Crestani, 2016)

Model	P	R	F1	$ERDE_5$	$ERDE_{50}$
LR	0.39	0.78	0.52	11.82%	7.50%
SVM	0.45	0.69	**0.54**	11.21%	7.14%
Soft voting	0.49	0.76	**0.59**	10.95%	6.83%
LR^*	0.40	0.78	0.53	6.00%	5.30%

6.1. ERDE(Early risk detection error) measure

In addition to standard evaluation measures, which we used to assess the system's output with respect to golden truth judgments, we employ the ERDE (early risk detection error) metric proposed in (Losada and Crestani, 2016) and the goal is to minimize its value. This is a time-aware measure which detects system's performance considering how early it gives its predictions and how correct they are. The decision delay is measured by counting the number of distinct textual items seen before giving the final output. Considering detection of risk cases is the main problem, the ERDE measure associates greater costs to true positives and false negatives than to true negatives and false positives.

In domains like this one late detection of positive cases has severe consequences, therefore ERDE scoring function multiplies the cost of true positives by a latency factor. This implements the idea that late detection of risk cases is equivalent to not detecting them at all. The latency factor is increasing with the number of seen items. To control the delay at which the cost grows more quickly, ERDE measure is parameterised by parameter o. To test out system we used two versions of the proposed metric: $ERDE_5$ and $ERDE_{50}$ which were used in previous work. It should be noted that no latency cost is introduced for true negatives because non-risk cases would not demand early intervention.

In order to process the stream of texts written by each user with our developed model, we utilized a dynamic method proposed in (Losada and Crestani, 2016) which incrementally builds a representation of each user, taking one of his posts at a time, and only makes a positive decision if our classifier outputs a confidence value above an arbitrarily initialized threshold of 0.5. If the stream of user's text posts gets exhausted the method outputs a negative decision.

The results of our ERDE tests compared to ERDE results of (Losada and Crestani, 2016) are shown in Table 3. We should state that our results are directly comparable only with this original results because other results are from CLEF 2017[1] where researchers had to process the documents in weekly chunks and it was not possible to submit predictions before processing a complete chunk.

Comparing our results to those in (Losada and Crestani, 2016) it can be seen that our model yields slightly worse ERDE scores but a better F1 score. We infer the better ERDE score is a result of higher recall and lower cost for false positives in the ERDE scoring function. The overall system performance should be observed as a joint of ERDE and F1 measures.

7. Conclusion

In this paper, we explore and extract various possible features hidden in one's social media posts in an attempt to improve results on the task of early detection of depression risk cases. We create several models, using several word-, sentence- and document level features, which we train and test on CLEF 2017 eRisk pilot task dataset. Proposed models improved performance after introducing the crafted features.

In future work, we would like to further explore possible features underlying a textual document which could lead to depression detection. Also, incorporating link analysis for many hyperlinks we removed from the dataset using regular expressions could be useful but should be carefully approached because of possible digression from original data.

[1] `http://clef2017.clef-initiative.eu/`

References

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, pages 1–14, January.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162. Association for Computational Linguistics.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4647–4657, New York, NY, USA. ACM.

Christian Karmen, Robert C. Hsiung, and Thomas Wetter. 2015. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer Methods and Programs in Biomedicine*, 120(1):27 – 36.

D. Losada and F. Crestani. 2016. A test collection for research on depression and language use. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*, pages 28–39, Evora, Portugal, September.

David E. Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 346–360, Cham. Springer International Publishing.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin, and Ivan Smirnov. 2018. Feature engineering for depression detection in social media. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, pages 426–431, Funchal, Madeira, Portugal. INSTICC, SciTePress.

Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2017. Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, volume CEUR-WS 1866, Dublin, Ireland, September.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *CoRR*, abs/1709.01848, October.