Towards suicide prevention: early detection of depression on social media

Victor Leiva and Ana Freire

Department of Communications and Information Technologies
Universitat Pompeu Fabra
Carrer Tanger, 122-140, 08018 Barcelona (Spain)
victor.leiva01@estudiant.upf.edu, ana.freire@upf.edu

Abstract. The statistics presented by the World Health Organization inform that 90% of the suicides can be attributed to mental illnesses in high-income countries. Besides, previous studies concluded that people with mental illnesses tend to reveal their mental condition on social media, as a way of relief. Thus, the main objective of this work is the analysis of the messages that a user posts online, sequentially through a time period, and detect as soon as possible if this user is at risk of depression. This paper is a preliminary attempt to minimize measures that penalize the delay in detecting positive cases. Our experiments underline the importance of an exhaustive sentiment analysis and a combination of learning algorithms to detect early symptoms of depression.

Keywords: early detection, depression, social media, machine learning.

1 Introduction

Suicide has become an enormous public health problem worldwide. According to the World Health Organization (WHO), approximately 804,000 deaths occurred due to suicide in 2012, becoming the first cause of death between the ages of 15 and 44. It is worth highlighting that in high-income countries the amount of young adults (i.e., between 15 and 29 years old) who commit suicide accounts for 17.6% of the total number of deaths [13]. The statistics presented by the WHO inform that 90% of the suicides can be attributed to mental illnesses in high-income countries [12]. Besides, Minsu Park et al. [9] concluded that people with mental illnesses tend to reveal their mental condition on the social media, as a way of relief. Among all users of these social media platforms, adolescents have been pinpointed as the most frequent ones [3].

Hence, these previous studies drive us through the detection of depression on social media as a first step against suicidal behaviour.

Using a labeled dataset [7] with messages posted in *Reddit*[1], we experiment different techniques for improving previous results for early detection of depression considering time-aware classification measures.

[1] Reddit: https://www.reddit.com

2 Background

In the last few years, more research about detecting mental illnesses on social media has become available [1] [2] [4] [7] [8] [11] [14]. Twitter has been widely investigated [2] [4] [8], although other social networks have also been discussed in the literature, such as Sina [11] or Myspace [1].

Choudhury et al. [4] proposed a probabilistic model to determine which tweets could indicate depression. They extract both post-based (i.e. positive or negative effect, linguistic style) and user-based features (i.e. number of postings, number of followers/followees). After applying dimensionality reduction (Principal Component Analysis - PCA) and using Support Vector Machine (SVM) they got 74% accuracy.

A more recent work from 2016 [8] also examined Twitter with the same purpose. In this case, they use a bag of words (counts the word occurrence frequencies to quantify the content of a tweet) to vectorize the tweets. Then, they applied different machine learning algorithms (Logistic Regression, Naïve Bayes and SVM) and obtained around 80% accuracy in detecting depressive disorder. Logistic Regression has also been applied with similar objective in [1], in this case to classify Myspace post-mortem comments exhibiting (or not) emotional distress. Authors made use of Linguistic Inquire and Word Count library (LIWC)[2] to perform text analysis and get its polarity.

In the same context, sentiment analysis using LIWC was also applied to create a depression detection model for Chinese micro-blogs [11]. This model gathered 10 features from the text (again, the polarity of the text but also other innovative features such as the use of emoticons). The initial 10 features were later reduced to 5, with a corresponding degradation in the results but a high improvement regarding computational time and amount of data needed.

Lately, *Twitter* has become the most popular network among researchers. Nevertheless, a recent work by Losada et al. [7] draw our attention to *Reddit*. Authors argued that the limitation in the number of characters in a tweet reduces the context we can get about the writer. Reddit has no limitation regarding the number of characters and we can get unlimited access to previous submissions from the same user, as opposite to Twitter, that only allows to download up to 3200 tweets per user. Authors created and distributed a dataset including *Reddit* messages from different users, some of them diagnosed with depression. Such dataset was labeled indicating if the users were at risk of depression. Losada et al. noted that the most widely used classification measures (precision, recall, F-measure) are time-unaware, so they do not reward early alerts. Therefore, they proposed a new metric for early detection that penalizes the delay in detecting positive cases. They also reported the results of some preliminary baselines to detect early symptoms of depression. These baselines are quite simple and the features extracted from the text are just based on the tf-idf vectorization.

[2] LIWC: http://liwc.wpengine.com/

3 Proposal

We aim to improve the state of art of time-aware early detection of depression on social media by means of:

- Considering innovative measures that are specific for early detection.
- Extracting text features in a more accurate way than the current approaches (just based on tf-idf vectorization), by including sentiment analysis.
- Exploring the behaviour of classic and modern machine learning techniques to better predict positive cases.
- Studying if genetic algorithms can improve the previous results.

We give more details about this proposal in the next subsections.

3.1 Feature extraction

We run our experiments using the dataset built by Losada et al. [7]. It consists of a list of users catalogued as depressive (135 of them) and a control group (752 users), each of them with a corresponding random number of messages (there are users with 10 messages up to 2000, with an average of 607 messages per user). These messages cover different topics, as the redditors are often active in different subreddits. The text collection contains messages from a wide period of time so we can observe the language evolution used by a depressive subject and a non-depressive one.

In addition to tf-idf vectorization (12,968 features), we have introduced three more features that are related to the sentiment polarization of the text. The three features correspond to the positive, neutral and negative percentages associated to each message/user. To implement this sentiment analysis we employed VADER Sentiment Analysis (Valence Aware Dictionary and sEntiment Reasoner) [5]. It consists of a lexicon and a rule-based sentiment analysis tool suitable for sentiments expressed in social media.

3.2 Learning algorithms

We have used four prediction methods. These methods are briefly described below.

- *Logistic Regression*: calculates the probability of a categorical variable (depressed / non-depressed) from a number of predicting variables.

- *Support Vector Machine*: finds a hyperplane that divides two categorical data by a clear gap that is as wide as possible.

- *K-Nearest Neighbor*: is a non-parametric method. An object is classified by a majority vote of its k neighbors. The value of each vote is weighted by the distance of the neighbors to the object.

– *Random Forest*: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

Additionally, we have built an ensemble model that uses all of the previous algorithms to determine the class of the subject. We have used a technique called *soft-voting* where the model determines the probability of one class with the mean of all the probabilities given by the algorithms.

3.3 Parameters optimization

We have followed two strategies to determine which parameters are the optimal for each algorithm. Fist, we have applied grid search to optimize the algorithms independently. Then, by means of genetic algorithms, we have optimized their contribution in the ensemble method. For the second step we have chosen genetic algorithms, as they have demonstrated to be more stable and faster than grid search for the optimization of high number of parameters [6].

3.4 Evaluation

For results evaluation, we have used the measures proposed by Losada et al. [7]: Early Risk Detection Score *ERDS*, defined as follows:

$$ERDS_o(d,k) = \begin{cases} c_{fp} & \text{if d = false positive} \\ c_{fn} & \text{if d = false negative} \\ lc_o(k) \cdot c_{tp} & \text{if d = true positive} \\ 0 & \text{if d = true negative} \end{cases} \tag{1}$$

where $c_{fp} = 0.21$, $c_{fn} = 1$, $c_{tp} = 1$ and the function $lc_o(k)$ has the next structure:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{2}$$

All the scores are normalized by the number of testing users.

We also report the traditional score functions like the precision, recall and the F1 score.

4 Experimental Setup

4.1 Dataset setup

The *Reddit* dataset is structured in XML files, one for each redditor, that contains each of her submission ordered chronologically. For each post, we have its title, its text and the publication date. We vectorize the text using tf-idf. We

remove stopwords and only select the terms that appear in 20 documents or more, following Losada et al. [7] approach.
Since this kind of vectorization leads to very sparse data we use Principal Component Analysis (PCA) in order to reduce the dimensionality of the vector to the most informative components. We apply grid search to set the dimensionality of PCA to 300 features, while initially we had 12,968 tf-idf features (see figure 1).

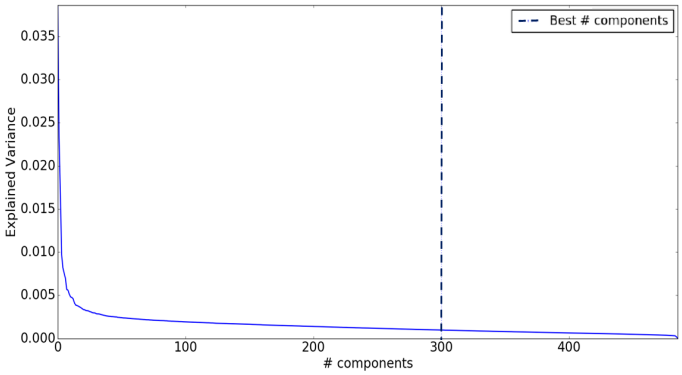

Fig. 1. Principal Component Analysis (PCA)

Both the implementation of the learning algorithms and the vectorization are implemented with scikit-learn library, version 0.18, for Python [10].

4.2 Baselines

We have used different baselines that obtained the best performance in [7]. These baselines are based on Linear Regression combined with the following strategies to emit a decision about the users:

- **First *n:*** This method consist in concatenating the first n messages from each user and making a prediction based on the result obtained from the learning model. If n is bigger than the total number of messages the method uses all the data available. This method has a delay of n messages (we experiment with n={10, 50, 100}).
- **Dynamic**: This method incrementally concatenates the texts of a user and only makes a decision if the depression language classifier outputs a confidence value above a given threshold. We experiment with probabilities 0.50 and 0.75, as they have shown the best performance in [7]. It is based on Logistic Regression with a particular set of parameters: *penalty*='l1', *solver*='liblinear', C=16 and *class_weight*={0:0.2, 1:0.8}.

4.3 Experimentation steps

Our experiments are organized in several consecutive steps.

First, we report the results we obtained with the baselines proposed by Losada et al. [7], as the dataset they released is slightly different from the one they used to publish the results. Note that these baselines are build upon Linear Regression with tf-idf vectorization. From this analysis we select the best baseline for the next steps. We vary the parameter n of the strategy *First n* (i.e. [10, 50, 100] instead of [100, 500]), in order to select an even more demanding baseline that emits a decision earlier and so to avoid using all the messages from some users.

Then, we compare the best baseline with the proposed algorithms (see Section 3.2: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF) and the ensemble voting algorithm (Ensemble).

Then, we apply the Principal Component Analysis (PCA) to see the effect of dimensionality reduction and we add new features resulting from VADER Sentiment Analysis.

Finally, we report the behaviour obtained with the hybrid model that uses a genetic algorithm (GA) to optimize the ensemble method algorithm described in section 3.2. In the genetic algorithm we can tune, basically, five parameters: the number of individuals, the number of generations, the tolerance, the birth/death rate and the mutation rate. In our experiments we have set the *ERDS* as our tolerance with a stopping value of 5%, the number of individuals is set to 100, the maximum number of generation is 40 and the birth/death rate corresponds to 20%. For the mutation rate we have considered three different strategies. The first one maintains a constant rate over time. Therefore, we try to avoid getting stuck in local minima of the score function. However, as a drawback, if the rate is too high, the algorithm may take infinite steps to converge to the optimal values. In our experiment with the mutation rate constant, we have tuned it to a value of 30%. For the other two strategies, we have implemented two functions that decrease over generations. We wanted to keep a high value of the mutation rate in the initial generations to diversify the individuals. At the same time, we wanted a low mutation rate in further generations in order to facilitate the convergence of the algorithm. The functions selected are an Asymptotic and a Gompertz function, with the following expressions:

$$Mutation\,rate\,functions \begin{cases} Asymptotic(x) = \frac{1}{x} \\ Gompertz(x) = 1 - 0.9e^{-6e^{-0.15x}} \end{cases} \tag{3}$$

The comparison between the two functions can be found in Figure 2. We observe that the Asymptotic function decreases more rapidly than the Gompertz function, which stays at higher values during more generations.

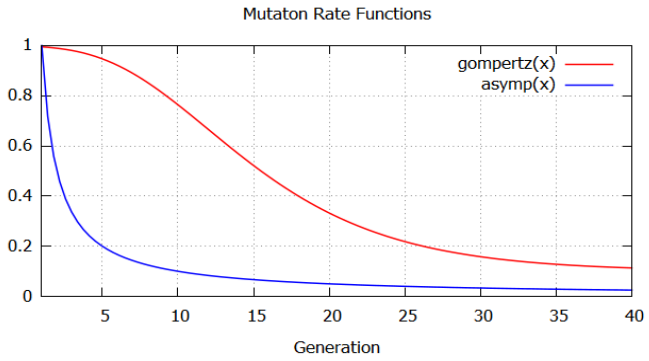

Fig. 2. Mutation rate functions.

5 Results

5.1 Baselines

The evaluation of the proposed baselines (using Linear Regression) can be found in Table 1. The results obtained from the *Dynamic* strategies obtain good results for most of the scores. The *Dynamic* strategy with probability 50% gets the best values for the early risk measures, as well as for the recall. Thus, we select linear regression with Dynamic 50% as the baseline for the next experiments.

	ERDE5	**ERDE50**	**P**	**R**	**F1**
First 10	0.10	0.09	0.63	0.29	0.39
First 50	0.08	0.08	0.57	0.46	0.51
First 100	0.07	0.07	0.55	0.56	0.55
Dynamic (prob. 50%)	**0.06**	**0.05**	0.41	**0.77**	0.54
Dynamic (prob. 75%)	0.07	0.07	0.61	0.54	**0.57**

Table 1. Baselines (Linear Regression)

5.2 Learning algorithms

We report the results obtained after running the learning algorithms described in section 3.2, including the dimensionality reduction (PCA) and the new features based on VADER sentiment analysis (see Table 2). LR (Linear Regression) corresponds to the baseline Dynamic (prob. 50%). We have used a grid search over a list of possible parameters combinations to select those used to train the learning algorithms.

	ERDE5	ERDE50	P	R	F1
LR	0.06	**0.05**	0.41	**0.77**	0.54
LR + PCA + VADER	0.06	0.06	**0.49**	0.67	**0.57**
Support Vector Machine	0.07	0.07	*0.58*	**0.58**	**0.58**
SVM + PCA + VADER	0.07	0.07	0.58	0.56	0.57
KNN	0.07	0.07	0.23	0.88	0.37
KNN + PCA + VADER	**0.06**	**0.05**	**0.38**	0.77	**0.51**
RF	**0.08**	0.08	**0.22**	0.83	**0.35**
RF + PCA + VADER	0.09	0.08	0.19	***0.96***	0.31
Ensemble	0.06	**0.05**	0.53	**0.69**	***0.60***
Ensemble + PCA + VADER	0.06	0.06	0.53	0.67	0.59
GA (Constant) + PCA + VADER	0.08	0.07	**0.54**	0.51	0.52
GA (Asymptotic) + PCA + VADER	0.07	0.07	0.52	0.58	0.55
GA (Gompertz) + PCA + VADER	0.07	0.06	0.53	0.60	0.56
GA (Gompertz) + PCA	0.06	0.06	0.53	0.67	**0.59**
GA (Gompertz) + VADER	***0.05***	***0.05***	0.45	**0.77**	0.57

Table 2. Proposed learning algorithms with the Dynamic strategy (prob. 50%). LR is the baseline. In bold, the best result for each algorithm with and without PCA+VADER. In italic, the best result for each measure.

Curiously, there are not significant differences after applying PCA and VADER, and the results highly depend on the learning algorithm. For instance, regarding LR, the combination PCA+VADER improves precision and F-measure but increases the error in case of ERDE50 and decreases recall. However, in case of KNN, after applying PCA+VADER all measures experiment some improvement, except for recall, that gets a value of 0.77 (the same as the baseline).

Although dimensionality reduction not always gets better results, it can still be applied to gain computational resources [11]. Regarding sentiment analysis, we will explore more advanced methods that also classify the text into numerous categories such as depression, happiness, sadness and so on.

Regarding the learning methods, some of them can get better results than the baseline, as we can see after applying genetic algorithms for parameter optimization with VADER (last row in the table). We run the Genetic Algorithm with the Gompertz function (highest performance) considering PCA and VADER separately. The use of the Genetic Algorithm along with the three text polarity features (VADER) performs 16,7% better than the baseline, in terms of ERDE5, while it maintains the value of ERDE50 in 0.05. Precision and F1 are also improved, and Recall is maintained at 77%.

6 Conclusions

This paper has investigated how to better detect early risk of depression in social media, by optimizing time-aware classification measures: ERDE5 and ERDE50. We have applied different learning algorithms, and combinations of them. Other

techniques such as dimensionality reduction and text polarity have been studied. We have provided further evidence for the benefit of applying genetic algorithms and text polarity (16,7% improvement regarding the baseline). Future work will concentrate on applying more accurate text sentiment classification to get a better representation of the input features.

Acknowledgements. This work was supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

1. Jed R Brubaker, Funda Kivran-Swaine, Lee Taber, and Gillian R Hayes. Grief-stricken in a crowd: The language of bereavement and distress in social media. In *In Proc. of ICWSM*, 2012.
2. Sowles S Connolly S Rosas C Bharadwaj M Bierut LJ Cavazos-Rehg PA, Krauss MJ. A content analysis of depression-related tweets. *Comput Human Behav. 1;54:351-357*, 2016.
3. Elizabeth Nick Nina C. Martin Kathryn M. Roeder Keneisha Sinclair-McBride Tawny Spinelli David A. Cole, Rachel L. Zelkowitz. Longitudinal and incremental relation of cybervictimization to negative self-cognitions and depressive symptoms in young adolescents. *Journal of Abnormal Child Psychology, pp 1-12.*, 2016.
4. Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.
5. C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Web and Social Media*, 2014.
6. Gary Wills Iwan Syarif, Adam Prugel-Bennett. Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA, Vol.14, No.4, pp. 1502 1509*, 2016.
7. David E. Losada and Fabio Crestani. A test collection for research on depression and language use. In *LNCS*, pages 28–39. Springer International Publishing, 2016.
8. Moin Nadeem. Identifying depression on twitter. *arXiv preprint:1607.07384*, 2016.
9. Minsu Park, Chiyoung Cha, and Meeyoung Cha. Depressive moods of users portrayed in twitter. *In Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics*, pages 1–8, 2012.
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
11. Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 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, 2013.
12. WHO. Online at http://www.euro.who.int. Last access: 06. May 2017.
13. WHO. *Preventing Suicide: A Global Imperative.* 2015.
14. Srinivasan P Yang C. Life satisfaction and the pursuit of happiness on twitter. *PLoS One,16;11(3):e0150881*, 2016.