VOLUMEN 11. NÚMERO 3. SEP-DIC 2017. DOI: 10.7714/CNPS/11.3.201

Luciana Mariñelarena-Dondena [1] [2]
Edgardo Ferretti [2]
Manolis Maragoudakis [3]
Maximiliano Sapino [4] [2]
Marcelo Luis Errecalde [2]

Predicting Depression: a comparative study of machine learning approaches based on language usage.

Prediciendo la depresión: un estudio comparativo de distintos enfoques de aprendizaje automático basado en el análisis del lenguaje.

Prognosticando a depressão: um estudo comparativo de diferentes enfoques da aprendizagem automática baseada na análise da linguagem.

[1] Postdoctoral Fellow at the National Council of Scientific and Technical Research (CONICET) – Argentina.

[2] National University of San Luis (UNSL) - Argentina

[3] Artificial Intelligence Laboratory, University of the Aegean, Samos - Greece

[4] Ph.D. Fellow at the National Council of Scientific and Technical Research (CONICET) - Argentina

Correspondencia: Luciana Mariñelarena-Dondena, Dirección postal: Belgrano 260, San Luis, San Luis, Argentina. Correo electrónico: lucianamd.psico@gmail.com

ABSTRACT

Depression is a major public health concern and a leading cause of disability. Globally, more than 332 million people of all ages suffer from depression. Several studies in the literature show that people language usage is indicative of their psychological states. That is why, there is growing interest in the application of natural language processing techniques for predicting depression. In this work, we present a comparative study of different machine learning methods and alternative ways of representing documents to automatically detect social media users who have reported to had been diagnosed with depression. The obtained results have demonstrated that a Deep Learning approach had the superior classification performance, when combined with a Synthetic Minority Oversampling Technique to deal with the problem of class imbalances in the dataset used in our experiments. The F1 score achieved was 82.93% with an accuracy of more than 94%.

Palabras clave: Depression; Machine Learning; Deep Learning; SMOTE (Synthetic Minority Oversampling TEchnique).

RESUMEN

La depresión es uno de los mayores problemas de salud pública que constituye a su vez una de las principales causas de incapacidad. A nivel mundial, más de 332 millones de personas de todas las edades padecen este trastorno. Investigaciones previas demuestran que el lenguaje que utilizan las personas refleja su salud mental. Por tal motivo, existe un creciente interés en la aplicación de técnicas de procesamiento del lenguaje natural para predecir la depresión. En este trabajo se presenta un estudio comparativo de diferentes métodos de aprendizaje automático como así también distintas maneras de representación de los documentos con el fin de detectar automáticamente a aquellos usuarios de medios sociales que manifestaron haber sido diagnosticados previamente con depresión. Los resultados obtenidos mostraron que la performance del clasificador mejoró considerablemente cuando se aplicó un enfoque de Aprendizaje Profundo combinándolo con el algoritmo SMOTE (Synthetic Minority Oversampling TEchnique) que permite hacer frente al problema de las clases desbalanceadas alcanzando así una medida F1 del 82.93%. En síntesis, este enfoque combinado, SMOTE + Aprendizaje Profundo, predice la depresión con una exactitud de más del 94%.

Keywords: Depresión; Aprendizaje Automático; Aprendizaje Profundo; SMOTE (Synthetic Minority Oversampling TEchnique).

RESUMO

A depressão é um dos maiores problemas de saúde pública que constitui por sua vez uma das principais causas de incapacidade. A nível mundial, mais de 332 milhões de pessoas de todas as idades padecem deste transtorno. Investigações prévias demonstram que a linguagem que utilizam as pessoas refletem a sua saúde mental. Por tal motivo, existe um crescente interesse na aplicação de técnicas de processamento da linguagem natural para prognosticar a depressão. Neste trabalho se apresenta um estudo comparativo de diferentes métodos de aprendizagem automático como assim também distintas maneiras de representação dos documentos com a finalidade de detectar automaticamente àqueles usuários de meios sociais que manifestaram haver sido diagnosticados previamente com depressão. Os resultados obtidos mostraram que a performance do classificador melhorou consideravelmente quando se aplicou um enfoque de Aprendizagem Profunda combinando com o algoritmo SMOTE (Synthetic Minority Oversampling TEchnique) que permite fazer frente ao problema das classes desbalanceadas alcançando assim uma medida F1 del 82.93%. Em síntese, este enfoque combinado SMOTE + Aprendizagem Profundo prognostica a depressão com uma exatidão de mais de 94%.

Palavras-chave: Depressão; Aprendizagem Automática; Aprendizagem Profunda; SMOTE (Synthetic Minority Oversampling TEchnique).

Acknowledgements:

This work has been partially founded by Project PROICO P-31816, National University of San Luis (UNSL) - Argentina. We are grateful to the members from the UNSLA team at the eRisk 2017 workshop pilot task, who processed the documents in XML format (as delivered in the challenge) and make us available the parsed documents in plain text without any further preprocessing.

Introduction

Depression is the most common mental disorder. Globally, the proportion of population with depression in 2015 was estimated to be 4.4%; in other words, more than 332 million people of all ages suffer from depression. Furthermore, between 2005 and 2015 the total estimated number of people living with depression was increased by 18.4% (World Health Organization, 2017). People with depression may experience a lack of interest and pleasure in daily activities, significant weight loss or gain, insomnia or excessive sleeping, lack of energy, inability to concentrate, feelings of worthlessness or excessive guilt and recurrent thoughts of death (American Psychiatric Association, 2014).

As a matter of fact, depression can lead to suicide. Over 800.000 suicide deaths occurred every year and it is the second leading cause of death in the 15-29 years-old range; that is, every 40 seconds a person dies due to suicide somewhere in the world. In this context, suicide prevention should be a higher priority on the global public health agenda. The early risk recognition is a core component to ensure that people receive the care and social support they need (World Health Organization, 2014).

The cognitive therapy argues that people who suffer depression have negative beliefs about themselves and the world, these thoughts and beliefs are excessive and inaccurate. Thus, the aim of cognitive behavioral therapy (CBT) is to help people to change this negative and self-destructive thought patterns. Other important change that occurs during CBT is that patients learn new cognitive coping skills, and then they can use these skills for the rest of their lives. CBT is the most widely studied approach of psychotherapy; although researchers do not know how exactly it helps the patients (Anthes, 2014).

Regarding to natural language processing, James Pennebaker (2002) argues that words people use every day are a diagnostic of physical and mental health:

> “The words we use in daily life reflect who we are and the social relationships we are in. This is neither a new nor surprising insight. Language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand. Words and language, then, are the very stuff of psychology and communication. They are the medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings” (Tausczik & Pennebaker, 2010, p. 25).

For many years, psychologists have used tests or carefully designed survey questions to assess different psychological constructs. Nowadays all the information available in social media, such as Twitter and Facebook, enables novel measurement approaches applying text mining techniques (Schwartz & Ungar, 2015). In social sciences and psychology, these text mining techniques focus on two main goals: prediction and description. In the first place, predictive tasks aim at designing an automatic classifier that estimates the dependent variable, usually called label or result, depending on certain features (independent variables) extracted from the documents. Secondly, descriptive tasks try to obtain patterns that explain or summarize the underlying relationships in the data. This allows, for example, formulating new hypotheses concerning the words that people use every day (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Mariñelarena-Dondena, Errecalde & Castro Solano, 2017).

Previous research which applies the *Linguistic Inquiry and Word Count* (LIWC), an automated word counting software, shows that is possible to characterize depression through natural language use. It has been suggested that suicidal poets use more first-person pronouns (e.g., I, me, mine) and less first plural pronouns (e.g., we, ours) throughout their writing careers than non-suicidal poets (Stirman & Pennebaker, 2001). In the same way, depressed students currently use more first person singular pronouns, more negative emotion words and fewer positive emotion words in their essays in comparison to students who has never suffered this disease (Rude, Gortner & Pennebaker, 2004).

With respect to social media postings, De Choudhury, Counts and Horvitz (2013) have analyzed the posts on Twitter shared by subjects diagnosed with clinical depression and next developed a Support Vector Machine (SVM) classifier that can predict if a post is depression indicative (accuracy 73%). Furthermore, they suggest that the timestamp would help to predict if the post is depression indicative because one characteristic of depressed subjects is the night time Internet activity due to insomnia. Recently, other research indicated that subjects with major depressive disorder (MDD) show lower social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, heightened expression of religious thoughts and belong to highly clustered close-knit networks.

As mentioned earlier, suicide prevention is a major challenge on the global public health agenda. In this context, Poulin et al. (2014) estimated the risk of suicide –with an accuracy of 65%– by examining the clinical notes taken from a national sample of United States Veterans Administration medical records. They generated datasets of single keywords and multiword phrases and then built a 3-bin classification scheme (three matched cohorts: veterans who committed suicide, veterans who used mental health services and did not commit suicide, and veterans who did not use mental health services and did not commit suicide during the observation period). Besides, Desmet and Hoste (2013) have analyzed the emotions of suicidal notes.

Even though notable research has been published recently on the area of Text and Social Analytics, where several studies have attempted to predict or analyze depression (De Choudhury, Counts & Horvitz, 2013; De Choudhury, Gamon, Counts & Horvitz, 2013; Park, McDonald & Cha, 2013; Paul & Dredze, 2011), no one has attempted to build a dataset where a large chronological sequence of writings leading to that disorder is properly stored and analyzed (Losada & Crestani, 2016). This is mainly due to the fact that text is often extracted from social media sites, such as Twitter or Facebook, that do not allow re-distribution. Besides, in machine learning community, it is well known the importance of having publicly available datasets to foster research on a particular topic, in this case, predicting depression based on language use.

That is why, the primary objective in Losada and Crestani (2016) was to provide the first collection to study the relationship between depression and language usage by means of machine learning techniques.

In this current work, based on the dataset provided by Losada and Crestani (2016), we present a comparative study of different machine learning methods and alternative ways of representing documents in order to automatically detect users who have reported to had been diagnosed with depression.

The rest of the paper is organized as follows. Next section formally defines the problem faced in this paper and also introduces some theoretical background from text mining and machine learning research fields, that will be needed to understand the results and discussion. Then, we offer the conclusions of our work.

Methodology

In Section 2.1 we start with a formal definition of the problem faced in this paper. Section 2.2 briefly describes the dataset used in our experiments. Then, in Section 2.3, we provide some theoretical background regarding the text mining techniques used in our experimental work.

Problem Statement

In this work, the problem of predicting depression on individuals based on their language usage, is faced as a supervised classification problem. That is, we have a collection of documents from individuals that report to have been diagnosed with depression and a control group with documents belonging to individuals which are interested in depression, mainly because they could have close relatives suffering this health problem, but have had no depression. The collection used in our experiments, corresponds to the official training set of the pilot task on early risk detection of depression proposed on the CLEF eRisk 2017 workshop (http://erisk.irlab.org/).

The above-mentioned pilot task emphasizes the use of innovative early detection solutions, since this research initiative is based on the seminal work presented by Losada and Crestani (Losada & Crestani, 2016). In our case, as a first step, we focus on developing effective machine learning based approaches to detect depression, and in this work, we will leave aside the aspect of predicting depression as an early classification task; that is, we face this problem as a classical binary classification problem where the positive or target class is as expected, depressed individuals. Therefore, we will evaluate the classification methods used following the standard performance metrics like precision, recall, F1-measure and accuracy.

Dataset and Preprocessing

Our dataset consists of 486 articles gathering submissions (post and comments) from members of the Reddit community (https://www.reddit.com/). Reddit is an open-source platform where community members, so called *redditors*, can submit content (posts, comments, or direct links), vote submissions, and the content entries are organized by areas of interests (subreddits). From among the 486 articles, each one containing the post and comments of one redditor over a period of time, there are 83 positive samples (target class) and 403 negative samples (control group / negative class). As it can be observed, the training set is not balanced; that is, there is not the same number of samples in both classes.

The official test dataset of the competition is to be delivered in ten chunks from the beginning of February to the beginning of April. This is done this way because, as mentioned above, the challenge mainly targets early risk prediction methods. Moreover, the ground-truth results will not be delivered until the end of April 2017. In this way, to evaluate our classification approaches we have built a validation set from the training set following a random split of approximately 80-20; that is, 80 percent of the documents from the original training set will be used for training purposes and the remaining 20 percent will be used as test documents to evaluate the classifiers' performance. This split was performed so that the average number of post per user and the average numbers of words per user in the validation set were similar to that of the training set. Hence, the final split resulted in 63 positive documents and 288 negative documents in the training set and 20 positive documents and 115 negative documents in the validation set. From now on, on the rest of the paper, the validation set will be referred as test set.

Text Mining Techniques

In the domains of data mining and machine learning, classification is a process where one uses a set of annotated data as training and anticipates on learning a model that would not only accurately predict the label of the instances it accepted as input, but on the contrary, to be able to generalize its prediction functionalities over, new, previously unseen data. One of the most common difficulties experts face during classification is the selection of an appropriate set of data characteristics that will be used to formulate a so-called "vector", i.e. a set of the selected attributes with a certain value per each attribute that characterizes each instance.

For the task at hand, textual information is among the most common data resource since it can be easily found in Web 2.0 frameworks, such as web portals, blogs, social media, discussion fora, etc. Therefore, the need to find appropriate attributes to transform plain text into a wellstructured, unified vector is usually fulfilled without having a human annotator dictating a pre-defined set of words of interest such as (“depression”, “anxiety”, “stress”, etc.). These words may be straightforwardly associated with the class labeled “Depression” but it is quite uncommon that most documents will certainly contain these exact words and not implying the notion of depression with other tokens that are more difficult to be pre-associated with the class by any human.

Consequently, most of studies on text mining deal with the problem of feature representation using a simpler method, i.e. the inclusion of each word appearing in a collection of documents and then implement a method for weighting words that appear often within a single document but also appear often in the whole document collection with lower weights than terms that are frequent in a document and not so frequent in the others. The method, also known as “tf-idf” is very popular and in fact it is the basis for any text mining problem. The inclusion of every token is called as “Bag-of-words” or simply “BoW” feature selection.

Note however that this approach usually generates a vector space of enormous size, since each token is considered as a vector dimension. The resulting problem, named as “curse of dimensionality” is that the training examples usually contain a very small subset of the words that are present in the whole collection; therefore, most features have zeros as weights for these documents. In essence, the problem of classification becomes too sparse and even the most sophisticated algorithms face objective difficulties in learning what separates the instances of each class labels. Finally, an additional burden for our task is the imbalance of training data, meaning that the documents with class “Depression” are very few compared to the other class label, approximately in an analogy of 1:6. Even if it sounds logical that in real-life situations, most documents would not have any signs of depression, this phenomenon, called as “class imbalance” results in a biased prediction towards the majority class label.

Feature Selection Process

Given the need to transform each document into a unified representation scheme and simultaneously attempt to retain as much useful information as possible, the following pre-processing steps were carried out:

a) **Tokenization:** All stop-words were removed and all letters were transformed to lowercase.
b) **Part-of-speech (PoS) tagging:** Based on the logic that the essence and meaning of a text lies mostly on its nouns, but often adjectives, adverbs and verbs bear semantic information, we experimented with retaining only such tags. The PENN treebank set was used to instruct the tagger of which part-of-speech elements to keep.
c) **Stemming:** The Porter implementation of stemming was incorporated, in which the lemma of each word is kept by removing all of its suffix variations. For example, tokens such as “says”, “said”, “saying”, etc. are all merged into “say”.
d) **Creation of the document-term matrix:** The rows of the matrix represent the documents and the columns correspond to the distinct number of tokens in the whole document collection, upon applying the second and third step. Each matrix cell bears the value of the *tf-idf* weight of each single term, calculated using the following formula:

$$tf - idf(term) = frequency(term) \cdot log(\frac{m}{N(term)} + 1)$$

Figure 1.

Where:

· *m*: is the total number of terms,
· *N* (*term*): is a function that returns the number of documents in which the term appears.

e) **Generation of n-grams of terms:** In certain cases of text analysis, words are not independently appearing within a document but are regularly affected by their neighboring context. According to several studies, the n-gram technique produces a much larger feature space but can also hold important information which is lost when one merely considers unigram BoW.

f) **Dimensionality reduction:** Despite the aforementioned steps that aim towards keeping the essential pieces of information, the feature space remains at high levels of dimensionality and problems such as sparsity can deteriorate the performance of even the most sophisticated algorithms. Therefore, by borrowing ideas from the domain of mathematics and more specifically of linear algebra, the Singular Value Decomposition (SVD) approach was applied. SVD is a procedure that reduces dimensionality by performing the transformation of initial, potentially correlated variables to a fully new dataset with linearly uncorrelated variables, singular vectors. Each singular vector models feature variance in a descending order, meaning that the first singular vector bears the greatest impact compared to the second, the second holds greater impact than the third and so forth. According to researchers in feature reduction, a rather small number of singular vectors is needed to model the largest part of a rather massive dataset, without losing too important information (Kalman, 1996; Platt, 2000; Potha & Maragoudakis, 2014).

Deep Learning

A deep neural network topology is comprised of an input layer, one or more hidden layers, and an output layer. The input layer matches the features space, so that there are as many input neurons as attributes. The output layer is a classification layer to match the output space, a binominal set of class labels for our case. All layers are composed of neurons, the basic units of such a model. In the traditional feed-forward architecture, each neuron in the previous layer *i* is fully connected with all neurons in the subsequent layer *i+1* though directed arcs, each representing a certain weight. Also, each neuron in a non-output layer of the net has a bias unit, serving as an activation threshold. As such, each neuron receives a weighted combination *a* of the *ni* outputs of the neurons in the previous layer *i* as input.

$$a = \sum_{j=1}^{ni} w_j x_j + b$$

Figure 2.

Where *wj* is the weight of each output *xj* and *b* represents the bias.

By selecting an activation function, α is adapted to minimize a loss function such as cross-entropy, which suits well the classification purposes. During this optimization process, we make use of two advanced methods via H_2O, an open source library for Deep Learning. Initially, dropout is used ––a modern form of regularization in which each neuron suppresses its activation with a certain dropout probability during forward propagation for a given training instance. As such, instead of one network architecture, efficiently 2^N architectures are trained, with *N* denoting the examples. The resulting network denotes an ensemble of an exponentially large number of accumulated models. This regularization method helps to avoid over-fitting and improves generalization as stated earlier. Consequently, we exploit an optimization function provided by H_2O called ADADELTA (Candel, LeDell, Parmar & Arora, 2016; Zeiler, 2012), combining momentum learning and rate annealing, two important parameters during training. Momentum rate results in avoiding local minima and annealing overcomes the so-called "optimum skipping" in the search space during optimization (Zeiler, 2012).

Dealing with the Class Imbalance Problem

A common situation in numerous real-world applications, such as the one at hand, is that class imbalance problems often deteriorate the performance of traditional classifiers. The main reason is that in certain

VOLUMEN 11. NÚMERO 3. SEP-DIC 2017. DOI: 10.7714/CNPS/11.3.201

domains, having annotated examples from both classes in a similar analogy is unrealistic. Such domains include suicide note analysis, spam email detection, cyber-bullying identification, etc.

There are two trends for dealing with such domains. The former assumes that some examples of the majority are usually found quite distant from the center point of the cluster of all examples of majority class, hence they could be discarded without losing important information about the boundaries that separate the two classes. This method is commonly referred to as "sub-sampling" of the majority class. The latter deals with the examples from the minority class and aims at increasing their number by generating data using a slight modification of their initial distributions in order not to be identical to the original samples of the minority class but also not to be far apart from them.

Therefore, according to current state-of-the-art in the field of class imbalance, a hybrid strategy was followed that aligns with the aforementioned trends. When dealing with the minority class, the SMOTE (Synthetic Minority Oversampling TEchnique) algorithm was applied. Subsequently, in order to perform under-sampling of the majority class, we have incorporated the Tomek links metric, as the means to detect noisy and borderline examples (Chawla, Bowyer, Hall & Kegelmeyer, 2002). SMOTE is constructed upon the notion that each example of the minority class can form the basis for synthetically re-generating neighboring examples using its k-NN (Nearest Neighbors) graph instead of the traditional method of randomized sampling with replacement. The Tomek link method can be considered as the means for guided under-sampling, where the noisy observations from the majority class are removed. Note that the term "noisy" refers to examples that are not densely located within the center of their corresponding task (Athanasiou & Maragoudakis, 2017).

$$Precision = \frac{tp}{tp + fp}$$

Figure 3.

Evaluation Criteria and Performance

For classification tasks, the terms true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) compare the results of the classifier under test with trusted external judgments (Table 1). The terms positive and negative refer to the classifier's prediction and the terms true and false refer to whether that prediction corresponds to the external judgment. In a binary classification scenario as the one faced in this work, the predictions made by the classifier are usually interpreted as a confusion matrix like the one shown below:

Table 1. Confusion Matrix Scheme.

	Trusted external judgments	
Classifier's prediction	*true positives* (tp)	*false positives* (fp)
	false negatives (fn)	*true negatives* (tn)

Then, several performance metrics are usually defined based on *tp*, *fp*, *fn* and *tn*. The more popular ones, which have been also used in our experimental work are:

Precision reflects how accurate is the classifier from among the documents which have been predicted as positives. Conversely, *recall* aims at measuring how many real positive documents have been effectively classified as such. *Accuracy* measures how accurate is the classifier predicting documents –whether positive or negative– in their real classes. Finally, *F1 score* is approximately the average of precision and recall when they are close, and is more generally the harmonic mean. This measure aims at weighting the overall performance of a classifier in terms of its precision and recall.

$$Recall = \frac{tp}{tp + fn}$$

Figure 4.

$$F1 - score = 2\frac{precision \cdot recall}{precision + recall}$$

Figure 5.

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

Figure 6.

Results and Discussion

In this section, we present the different results obtained in our experimental study. Each row of the Table 2 corresponds to a particular method (classifier + document model) and the columns report standard performance metrics of the classification task.

Method 1 uses as document model the PoS of all words, retaining only nouns, adjectives and adverbs. Given the high dimensionality of the document models obtained, then dimensionality reduction is performed using the SVD technique and the 100 most significant components are finally those composing the document models used as input by the classifier. Then, a deep learning approach is trained using Tanh as rectifier. First row of Table 2 reports the results achieved with this method. As it can be observed, the accuracy of the classifier is quite good (85.93%) since almost all negative samples were correctly classified and the same occurs with the positive samples. As there are only four positive samples incorrectly classified (fn column), a high recall value (80%) is expected. However, given that the test set is unbalanced, the fifteen false positive samples –that represent 13% of the true negative samples– severely affect the classifier's precision (51.61%) since they approximately represent half of the samples predicted as positives, and hence the classifier performs a little better than random guessing. Finally, the F1 measure for this method results in a low value of approximately 63%, since the low precision achieved reduces the good recall value obtained.

Method 2 maintains the same document model used for method 1 but changes the classifier used to Gradient Boosting Machines –using the value 0.25 as learning rate– (Natekin & Knoll, 2013). As it can be observed in the second row of Table 2, all the performance metrics reported for this classifier are worse than for the deep learning approach of method 1. Even though not reported in this table, other classifiers were evaluated like SVM (Boser, Guyon & Vapnik, 1992) and Random Forests (Breiman, 2001), and similar results were achieved but none improve the deep learning classifier of method 1.

For the two methods mentioned above, we can notice the poor performance regarding to the minority class label, i.e. "Yes". On the contrary, method 1 seems quite robust in predicting the negative class. Therefore, additional measures are needed apart from the vector representation method and the classification method. The main idea of the methods to follow; that is methods 3 to 5, is to transform either the feature space or help the deep learning classifier by adding artificial instances of the minority class and removing noisy examples from the majority one.

In this respect, method 3, uses BoW as document model and applies the same SVD reduction technique than method 1. That is, we retain the 100 most significant components but in this case, they could also include other PoS components like pronouns, verbs, etc. When compared method 1 versus method 3, we can see that using the reduced BoW document models highly reduces the number of false positive samples, thus increasing the classifier's precision on approximately 25%; even though the positive samples correctly predicted (tp column) also suffered a little decrease. The increase in the negative samples correctly predicted (tn column) justify the increase of the accuracy achieved by the classifier (91.11%). However, the recall achieved decrease 20% but as precision increase was higher than the recall decrease,

Table 2. Comparative Performance Metrics.

	tp	fp	tn	fn	Precision	Recall	F1	Accuracy
Method 1	16	15	100	4	51.61%	80.00%	62.75%	85.93%
Method 2	14	18	97	6	43.75%	70.00%	53.85%	82.22%
Method 3	12	4	111	8	75.00%	60.00%	66.67%	91.11%
Method 4	13	6	109	7	68.42%	65.00%	66.67%	90.37%
Method 5	**17**	**3**	**111**	**4**	**85.00%**	**80.95%**	**82.93%**	**94.81%**

the F1 measure for this method performs a little better than for method 1. It is worth remembering that the F1 score represents the harmonic mean of precision and recall, and thus reports an overall performance viewpoint of the classifier.

Method 4 uses as document model words 3-grams instead of plain words and again the SVD technique for dimensionality reduction is applied. In this case, the F1 score reported coincides with the one achieved by method 3 (66.67%) but in this latter case the precision decrease approximately 6% and recall increase 5%. Considering that the accuracy for method 4 also decreased a bit with respect to the one reported for method 3, the last-referred method should be preferred.

The first four rows of Table 2, that report the results of the methods discussed above, vary the classifiers used as well as the document models designed but all were trained with the unbalanced dataset described in the section regarding Dataset and Preprocessing. In this respect, the last row of Table 2 reports the performance of the same deep learning approach than for methods 1, 3 and 4 but the SMOTE method was applied to deal with data imbalance, which plays with under and over sampling of negative and positive class respectively. This row, highlighted in bold, reports the best results obtained in our experiments. The F1 score achieved is almost 83%, that shows that high compromise achieved between precision (85%) and recall (81%). This fact also impacts on the accuracy (approximately 95%) obtained with this method.

Conclusions

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction (LeCun, Bengio & Hinton, 2015). These methods have dramatically improved the state of the art of data science in several fields of applications. In this work, we have shown its adequacy for depression detection in social media.

Even when confronted with other state-of-the-art classifiers, such as Support Vector Machines (SVM), Random Forests and Gradient Boosting Machines (GBM), Deep Learning demonstrated superior performance. Moreover, it was shown the importance of dealing with the problem of the existing class imbalance in the dataset used, given that by applying SMOTE (Synthetic Minority Oversampling TEchnique), the performance of the deep learning approach used increases its F1 score from 66.67% to 82.93%. Besides, the accuracy of the combined approach viz. SMOTE + Deep Learning can predict depression with an accuracy of more than 94%.

Finally, we conclude that social media posts stand for a new type of lens in the understanding of depression that complements traditional tests. In addition, we hope that this paper highlights the potential of the text mining techniques as a psychological assessment tool.

Received: 28/03/2017
Accepted: 25/09/2017

VOLUMEN 11. NÚMERO 3. SEP-DIC 2017. DOI: 10.7714/CNPS/11.3.201

REFERENCIAS

Anthes, E. (2014). A change of mind. *Nature, 515*, 185-187. [Special issue: Depression]

Asociación Americana de Psiquiatría [American Psychiatric Association] (2014). *Manual diagnóstico y estadístico de los trastornos mentales (DSM-5)* [*Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*]. Arlington, VA: Asociación Americana de Psiquiatría.

Athanasiou, V. & Maragoudakis, M. (2017). A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources Are Not Plentiful: A Case Study for Modern Greek. *Algorithms, 10*(34). doi:10.3390/a10010034

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Candel, A., LeDell, E., Parmar, V. & Arora, A. (2016). *Deep Learning with H2O*. Mountain View: H2O.ai.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

De Choudhury, M., Counts, S. & Horvitz, E. (2013). Social Media as a Measurement Tool of Depression in Populations. *Proceedings of the 5th Annual ACM Web Science Conference*, 47-56.

De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. (2013). Predicting Depression via Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Desmet, B. & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*, 6351–6358.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37-54.

Kalman, D. (1996). A Singularly Valuable Decomposition: The SVD of a Matrix. *The College Mathematics Journal, 27*, 2-23.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature, 521*, 436-444. doi:10.1038/nature14539

Losada, D. E. & Crestani, F. (2016). A Test Collection for Research on Depression and Language Use. In N. Fuhr et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2016. Lecture Notes in Computer Science* (pp 28-39). Vol. 9822. Cham: Springer.

Mariñelarena-Dondena, L.; Errecalde, M. L. & Castro Solano, A. (2017). Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. *Revista Argentina de Ciencias del Comportamiento, 9*(2), 65-76.

Natekin, A. & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*. doi: 10.3389/fnbot.2013.00021

Park, M., McDonald, D. W. & Cha, M. (2013). Perception Differences between the Depressed and Non-Depressed Users in Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 476-485.

Paul, M. J. & Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 265-272.

Pennebaker, J. W. (2002). What our words can say about us: Toward a broader language psychology. *Psychological Science Agenda, 15*, 8-9.

Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparison to regularized like-lihood methods. In A. Smola, P. Bartlett, B. Scholkopf & D. Schuurmans (Eds.), *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press.

Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., et al. (2014). Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. *PLoS ONE, 9*(1), e85733. doi:10.1371/journal.pone.0085733

Potha, N. & Maragoudakis, M. (2014). Cyberbullying Detection using Time Series Modeling. *IEEE International Conference on Data Mining Workshop*. doi:10.1109/ICDMW.2014.170

Rude, S., Gortner, E. M. & Pennebaker, J. (2004). Language Use of Depressed and Depression-Vulnerable College Students. *Cognition and Emotion, 18*(8), 1121-1133.

Schwartz, H. A. & Ungar, L. H. (2015). Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science, 659*, 78-94.

Stirman, S. W. & Pennebaker, J. W. (2001). Word Use in the Poetry of Suicidal and Non-Suicidal Poets. *Psychosomatic Medicine, 63*(4), 517-522.

Tausczik, Y. R. & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

World Health Organization (2014). *Preventing suicide. A global imperative. Executive summary.* Luxembourg: World Health Organization.

World Health Organization (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. Geneva: World Health Organization.

Zeiler, M. D. (2012). *ADADELTA: An Adaptive Learning Rate Method*. arXiv:1212.5701