




- ORIGINAL ARTICLE -

Using Text Classification to Estimate the Depression Level of Reddit Users

Usando Clasificación de Textos para Estimar el Nivel de Depresión de Usuarios de Reddit

Sergio G. Burdisso^{1,2} , Marcelo Errecalde¹ , and Manuel Montes-y-Gómez³ 

¹Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, San Luis, C.P. 5700, Argentina
{sburdisso, merreca}@unsl.edu.ar

²Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

³Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, Puebla, C.P. 72840, Mexico
mmontesg@inaoep.mx

Abstract

Psychologists have used tests and carefully designed survey questions, such as Beck's Depression Inventory (BDI), to identify the presence of depression and to assess its severity level. On the other hand, methods for automatic depression detection have gained increasing interest since all the information available in social media, such as Twitter and Facebook, enables novel approaches based on language use. More precisely, these methods have focused on learning to detect depressive users through their language usage. However, little effort has been put into going beyond mere detection, towards estimating users' actual clinical depression level. The present study is a first step towards that direction: we try to develop a model able to estimate Reddit's users' clinical depression level by filling in the BDI depression questionnaire on behalf of each user. To carry out his task, the model answers all 21 questions of the questionnaire using the confidence value outputted by a binary text classifier trained to detect depressed users on Reddit. Our proposal was publicly tested in the CLEF's eRisk 2019 lab obtaining the best and second-best performance among the other 13 submitted models.

Keywords: Beck's Depression Inventory, CLEF eRisk 2019, Depression Level Estimation, SS3, Text Classification.

Resumen

Los psicólogos han utilizado cuestionarios cuidadosamente diseñados, como el "Inventario de Depresión de Beck" (BDI), para identificar la presencia de depresión y evaluar su grado de severidad. Por otro lado, los métodos para automáticamente detectar depresión están ganando un creciente interés debido a la gran cantidad de información disponible en las redes sociales. Más precisamente, estos métodos se han centrado en aprender a detectar usuarios depresivos a través de su

uso del lenguaje. Sin embargo, poco esfuerzo se ha realizado en ir más allá de la mera detección, hacia la estimación del nivel de depresión clínica real de los usuarios. El presente estudio es un primer paso hacia esa dirección, en donde intentamos desarrollar un modelo capaz de estimar el nivel de depresión clínica de usuarios de Reddit completando el cuestionario de depresión BDI por cada uno de ellos. Para llevar a cabo su tarea, el modelo responde las 21 preguntas del cuestionario utilizando el valor de confianza emitido por un clasificador de texto binario entrenado para detectar usuarios depresivos en Reddit. Nuestra propuesta fue probada públicamente en el eRisk 2019 obteniendo el mejor, y segundo mejor, desempeño entre los otros 13 modelos presentados.

Palabras claves: Clasificación de textos, CLEF eRisk 2019, Estimación del nivel de depresión, Inventario de Depresión de Beck, SS3.

1 Introduction

Depression is one of the leading cause of disability and one of the major contributors to the overall global burden of disease. Globally, in 2015 it was estimated that more than 332 million people suffered from this mental illness. Additionally, between 2005 and 2015 the total estimated number of people living with depression increased by 18.4%. Depressive disorders are ranked as the single largest contributor to non-fatal health loss and in extreme cases could lead to suicide[1]. Every 40 seconds a person dies due to suicide somewhere in the world, every year over 800.000 suicide deaths occur and it is the second leading cause of death in the 15-29 years-old range[2]. In 2015, suicide was among the top 20 leading causes of worldwide death[1]. Globally, 71% of all violent deaths in women, and 50% in men, are due to suicide[2]. Along with cancer, heart disease, stroke, and diabetes, suicide is among the 10 leading causes of death in the United States, as well as in other high-income countries. Additionally, the

suicide rate increased by 3.7% from 2016 to 2017[3].

For many years, psychologists have used tests or carefully designed survey questions (such as BDI[4]) to identify the presence of depression and to assess its severity level. Nowadays, all the information available in social media, such as Twitter or Facebook, has enabled novel methods for depression detection based on machine learning techniques. Even though multiple studies have attempted to predict or analyze depression using machine learning techniques, to the best of our knowledge, Losada *et al.* carried out the first attempt to build a public dataset in which a large collection of social media users' posts leading to this disorder was made available to the public [5]. The main goal in their work was to provide the first public collection to study the relationship between depression and language usage by means of machine learning techniques. This dataset was later used in the CLEF's eRisk 2017[6] and 2018[7] public tasks on early depression detection in social media.

Machine learning models learn to characterize depression through natural language usage and obtained results have shown that, in fact, language usage can provide strong evidence in detecting depressive people. However, not much attention has been paid to measuring finer grain relationships between language usage and this disorder. For instance, little effort has been put into going beyond merely depression detection, towards *estimating people actual clinical depression level*. The present study is a first step towards that direction, we try to develop a model able to estimate Reddit¹ users clinical depression level by filling in the BDI depression questionnaire on behalf of each user. To carry out his task, the model answer all 21 questions of the questionnaire using the confidence value outputted by a binary text classifier trained to detect depressed users on Reddit. Our model was evaluated by participating in the CLEF's eRisk 2019 task, obtaining the best and second-best performance among all participating models, thus, this article also describes how our team (UNSL) approached this task.

This paper is an extended version of our preliminary work presented in *XXV Congreso Argentino de Ciencias de la Computación*[8]. The main changes are an improvement in the overall writing of the article, where several sections have been rewritten and the content in general has been better polished, including the addition of a new figure (Figure 2), that we believe help to make the paper clearer and more easy-to-read; a more precise formulation of the Equation (2) is also given; submitted results have been added to Figure 3 to make the analysis easier for the reader; and finally, a whole new section has been included in which the models submitted by the other research teams are described and compared against ours. This paper is organized as follows. First, in Section 2, we describe the eRisk 2019 task and introduce the evaluation metrics. Then,

¹<https://www.reddit.com/>

we introduce the approach we used to carry out the task of estimating the depression level in Section 3 and the evaluation results are presented and discussed in Section 4. In Section 5 models submitted by the other research teams are described and a brief comparison against ours is given. Finally, the main conclusions derived from this study are summarized in Section 6, along with suggestions for possible future work.

2 Measuring the Severity of Depression

As it is described in more detail in the CLEF's eRisk 2019 overview [9], the lab was divided into three different tasks, T1, T2 and T3, being only T3 related to depression. T3 task consisted of estimating the level of depression from a thread of Reddit's user posts. For each user, participating models were given the history of postings and they had to automatically fill in a standard depression questionnaire of user's behalf. More precisely, this questionnaire was the psychologists-well-known Beck's Depression Inventory (BDI)[4] depression questionnaire. The BDI is a 21-question, self-report rating inventory that measures characteristic attitudes and symptoms of depression. Each question has 4 possible answers, numbered from 0 to 3, and is useful to assess the presence of feelings like sadness, pessimism, loss of energy, etc. For example, the first two questions are the following:

Question 1. Sadness:

0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

Question 2. Pessimism:

0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

Thus, participating models had to estimate the user's response to each individual question based on the content generated on Reddit by the corresponding user. Therefore, this task aimed at exploring the viability of automatically estimating the clinical severity of depression and their multiple symptoms.

It is worth mentioning that for this task, no training data was provided by the organizers. The test set used to evaluate the performance of all participants was built by asking Reddit users to manually fill in the BDI questionnaire and then collecting their answers along with their history of Reddit posts. These posts were collected right after the user filled in the BDI questionnaire. Thus, these users' questionnaires were then used as the ground truth to assess the quality of the questionnaires filled in by the participating models,

Table 1: Summary of the test data

No. of users	20
No. of posts	10,941
Avg. No. of posts per user	547
Avg. No. of days from first to last posts	881.2
Avg. No. of words per posts	46.4

automatically. The details of this test set are presented in Table 1.

2.1 Evaluation Metrics

In order to evaluate and measure the performance of the models in relation to the quality of questionnaires generated by them, four metrics were used:

- **Hit Rate (HR).** This measure computes the ratio of cases where the model’s questionnaire has exactly the same answer as the user’s real questionnaire. For example, a questionnaire with 5 matches (out of the 21 total questions) gets an HR equal to $\frac{5}{21}$.
- **Closeness Rate (CR).** This measure takes into account that each answer of the BDI questionnaire represent an ordinal scale. For example, imagine that the real user answered option 0. A system, S1, whose answer was option 3 should be penalized more than a system S2 whose answer was 1. For each question i , the absolute difference (ad) between the real and predicted answer (e.g. $|0 - 3| = 3$ and $|0 - 1| = 1$ for S1 and S2, respectively) is computed and next, the absolute difference is normalized as follows: $CR_i = \frac{3-ad_i}{3}$.² Finally, the CR_i for each question is averaged to obtain the final overall effectiveness score, i.e. $CR = \frac{1}{21} \sum_{i=1}^{21} CR_i$.
- **Difference between Overall Depression Levels (DODL).** The previous measures assess the systems’ ability to answer each question. This measure, instead, does not look at question-level hits or differences but computes the overall depression level (i.e. sum of all the answers) for the real and estimated questionnaire and next, the absolute difference ($ad_{overall}$) between these two scores is computed. In the BDI, depression level is an integer between 0 and 63 and, therefore, the final DODL is normalized between 0 and 1 as follows: $DODL = \frac{63-ad_{overall}}{63}$.
- **Depression Category Hit Rate (DCHR).** In the psychological domain, it is customary to associate BDI depression levels with the following categories: *minimal* (depression levels 0-9); *mild* (depression levels 10-18); *moderate* (depression levels 19-29); *severe* (depression levels 30-63).

²Note that this 3 here is equal to the maximum possible answer.

This measure consists of computing the fraction of cases where the automated questionnaire led to a depression category that is equivalent to the depression category obtained from the real questionnaire.

Finally, for the first three measures, results were reported using the average over all the users and were referred to as *AHR*, *ACR* and *ADODL*.

3 Our approach

To carry out this task, we trained a binary text classifier to detect depressed users and then we use its confidence values to estimate the user’s clinical depression level by completing the BDI questionnaire. We decided to use the SS3 text classifier which was firstly introduced by Burdisso *et al.* and that has obtained state-of-the-art performance on early depression detection tasks[10, 11]. Thus, Section 3.1 briefly introduces the SS3 classifier, and then Section 3.2 describes how questionnaires were actually filled in by our model.

3.1 The SS3 text classifier

As it is described in more detail in [10], SS3 first builds a dictionary of words for each category during the training phase, in which the frequency of each word is stored. Then, using those word frequencies, and during the classification stage, it calculates a value for each word using a function $gv(w, c)$ to value words in relation to categories. gv takes a word w and a category c and outputs a number in the interval $[0,1]$ representing the “importance” that w is believed to have in c . For instance, suppose categories are $C = \{food, music, health, sports\}$, then, after training, SS3 would learn to assign values like:

$$\begin{aligned} gv('sushi', food) &= 0.85; & gv('the', food) &= 0; \\ gv('sushi', music) &= 0.09; & gv('the', music) &= 0; \\ gv('sushi', health) &= 0.50; & gv('the', health) &= 0; \\ gv('sushi', sports) &= 0.02; & gv('the', sports) &= 0; \end{aligned}$$

Additionally, a vectorial version of gv is defined as:

$$\vec{gv}(w) = \langle gv(w, c_0), gv(w, c_1), \dots, gv(w, c_k) \rangle$$

where $c_i \in C$ (the set of all the categories). That is, \vec{gv} is only applied to a word and it outputs a vector in which each component is the gv of that word for each category c_i . For instance, following the above example, we have:

$$\begin{aligned} gv('sushi') &= \langle 0.85, 0.09, 0.5, 0.02 \rangle; \\ gv('the') &= \langle 0, 0, 0, 0 \rangle; \end{aligned}$$

The vector $\vec{gv}(w)$ is called the “confidence vector of w ”. Note that each category c_i is assigned a fixed

position in $\vec{g}\hat{v}$. For instance, in the example above $\langle 0.85, 0.09, 0.5, 0.02 \rangle$ is the *confidence vector* of the word “sushi” and the first position corresponds to *food*, the second to *music*, and so on. For those readers interested in how the *gv* function is actually computed, we highly recommend to read the SS3 original paper[10], since its equations are not given here to keep this paper shorter and simpler.

The SS3’s classification process can be thought of as a 2-phase process. In the first phase the input document is split into multiple blocks (e.g. paragraphs), then each block is in turn repeatedly divided into smaller units (e.g. sentences, words). Thus, the previously “flat” document is transformed into a hierarchy of blocks. In the second phase, the *gv* function is applied to each word to obtain the “level 0” *confidence vectors*, which then are reduced to level 1 *confidence vectors* by means of a level 0 *summary operator*, \oplus_0 . This reduction process is recursively propagated up to higher-level blocks, using higher-level *summary operators*, \oplus_j , until a single *confidence vector*, \vec{d} , is generated for the whole input. Finally, the actual classification is performed based on the values of this single *confidence vector*, \vec{d} , using some policy—for example, selecting the category with the higher *confidence value*. Note that using these *confidence vectors* in the hierarchy of blocks, it is quite straightforward for SS3 to visually justify the classification if different blocks of the input are colored in relations to their values³. This is quite relevant when it comes to health-care systems, because specialists should be able to manually analyze classified users and this type of visual tools could be really helpful.

We used the *addition* as the *summary operators* for generating the *confidence vectors* for all the levels, i.e. $\oplus_j = \text{addition}$ for all j , which simplified the classification process to the summation of all words’ $\vec{g}\hat{v}$ vectors read so far, in symbols, for every user s :

$$\vec{d}_s = \sum_{w \in WH_s} \vec{g}\hat{v}(w) \quad (1)$$

where WH_s is the user’s writing history. Note that for this task, \vec{d}_s was a vector $\langle d_{pos}, d_{neg} \rangle$ with only two components, one for the “depressed” class, d_{pos} , and the other for the “non-depressed” class, d_{neg} . In Burdisso *et al.* [10], early classification of users was carried out by analyzing how this *confidence vector* changed over time (i.e. as more posts were processed).

3.2 Filling in the BDI questionnaires

The scenario addressed in this article is considerably more difficult than standard depression detection scenarios since, unlike our previous experience in pre-

³As can be seeing on the live demos available at <http://tworld.io/ss3> in which interested readers can try out the SS3 classifier online.

vious CLEF’s eRisk challenges and previous work related to early depression detection[12][13][10], the problem addressed here goes beyond a mere binary “yes-or-no” classification problem. Trying to estimate the real clinical depression level of users is a problem that involves multiple decisions, one for each one of the 21 questions of the BDI questionnaire. In addition, another aspect that makes this task even harder was the fact that, as mentioned in Section 2, no training data was released by the eRisk organizers for this task.

Therefore, we decided to train the SS3 text classifier using the dataset for the eRisk 2018 depression detection task[7], using the same hyperparameters configuration ($\lambda = \rho = 1$ and $\sigma = 0.455$) that we used to address that task in our previous work [10]. However, the main problem was then deciding how to turn this binary classifier, trained for a simple “yes-or-no” decision, into a classifier capable of estimating the clinical depression level by filling BDI questionnaires. Our approach to address this problem was to use the SS3’s *confidence vector*, \vec{d} in Equation (1), to try to infer the BDI depression level between 0 and 63. To achieve this, first, we converted the *confidence vector* into a single *confidence value* (*cv*) normalized between 0 and 1, by applying the following equation:

$$cv = \begin{cases} \frac{d_{pos} - d_{neg}}{d_{pos}} & d_{pos} > d_{neg}, \\ 0 & otherwise \end{cases} \quad (2)$$

Where d_{pos}, d_{neg} are the confidence values for the positive and negative class from the \vec{d}_s confidence vector defined in Equation (1). Then, after SS3 classified an user, the obtained *cv* value was mapped into a region/category c , one for each BDI depression category ($c \in \{0, 1, 2, 3\}$). This was carried out by the following equation:

$$c = \lfloor cv \times 4 \rfloor \quad (3)$$

And finally, the user’s depression level was predicted by mapping the percentage of *cv* left inside the predicted c region to its corresponding BDI depression level range (e.g. $(0.5, 0.75] \rightarrow [19, 29]$ for $c = 2 =$ “moderate depression”) by computing the following:

$$dep_level = min_c + \lfloor (max_c - min_c + 1) \times (cv \times 4 - c) \rfloor \quad (4)$$

Where min_c and max_c are the lower and upper bound for category c , respectively (e.g. 19 and 29 for “moderate depression” category).

In order to clarify the above process, we will illustrate it with the example shown in Figure 1. First, SS3 processed the entire writing history computing the *confidence value* (given by Equation (2)) and then, the final *cv* value (0.941) was used to predict the depression category, “severe depression” ($c = 3$), by using the Equation (3). Finally, the depression level was

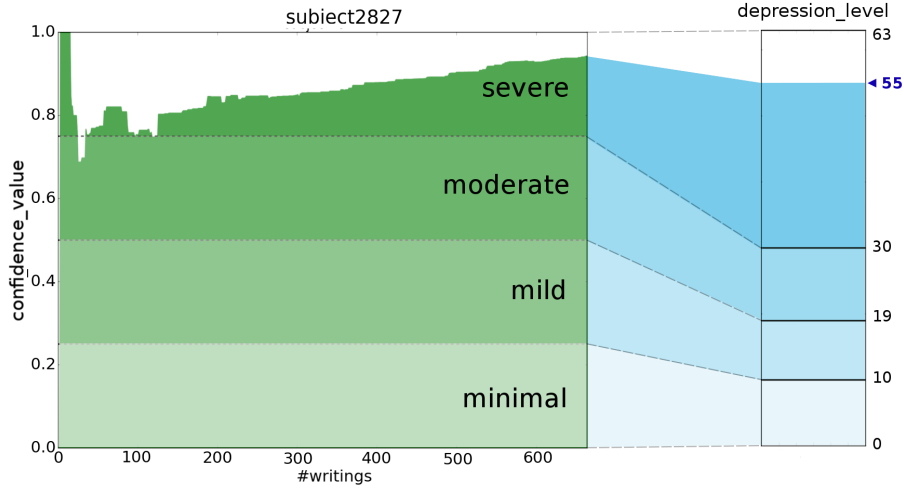


Figure 1: Diagram of the *dep_level* computation process for user referred to as “subject 2827”. As reader can notice, after processing all the user’s posts (writings), the final *confidence_value* (0.941) was mapped into his/her corresponding *dep_level* (55).

computed by the mapping given by Equation (4), as follows:

$$\begin{aligned}
 dep_level &= 30 + [(63 - 30 + 1) \times (0.941 \times 4 - 3)] \\
 &= 30 + [34 \times (3.764 - 3)] \\
 &= 30 + [34 \times 0.764] \\
 &= 30 + 25 = \mathbf{55}
 \end{aligned}
 \tag{5}$$

At this point we have transformed the output of SS3 from a 2-dimensional vector, \vec{d} , into a BDI depression level (a value from 0 to 63). However we have not covered yet how to actually answer the 21 questions in the BDI questionnaire using this *depression level*. Regardless the method, we decided that for all those users whose *dep_level* was less or equal to 0, all the BDI questions were answered with 0. For the other users we applied different methods since every participating team was allowed to use up to five different models (called “runs”) to carry out the task. Thus, we use five different methods to accomplish this task, as described below:

- *UNSLA*: using the predicted *dep_level* our model filled the questionnaires answering the answer number $\lfloor \frac{dep_level}{21} \rfloor$ on each question. If this division had a remainder, the remainder points were randomly scatter so that the sum of all the answers always matched the predicted depression level given by SS3.
- *UNSLB*: this time, only the predicted category, *c*, was used. Our model filled the questionnaire randomly in such a way that the final depression level always matched the predicted category, *c*. Compared to the following three ones, these two models were the ones with the worst performance.

- *UNSLC*: this model, and the following, were more question-centered. Once again, as in UNSLA, our model filled the questionnaires answering the expected number derived from the predicted depression level ($\lfloor \frac{dep_level}{21} \rfloor$). But this time, answering this number only on questions for which a “textual hint” for a possible answer was found in the user’s writings, and randomly and uniformly answered between 0 and $\lceil \frac{dep_level}{21} \rceil$ otherwise. To find this “textual hint”, our model split the user’s writings into sentences and searched for the co-occurrence of the words “I” or “my” with at least one word matching a regular expression specially crafted for each question.⁴ This method obtained the best AHR (41.43%) and the second-best DCHR (40%).
- *UNSLD*: the same as the previous one, but not using the “textual hints”, i.e. always answering every question randomly and uniformly between 0 and $\lceil \frac{dep_level}{21} \rceil$. This model was mainly used only with the goal of measuring the actual impact of using these “textual hints” to decide which questions should be answered with the expected answer ($\lfloor \frac{dep_level}{21} \rfloor$).
- *UNSL E*: the same as previous one, but this time not using a uniform distribution. More precisely, from the overall depression level predicted by SS3, once again the expected answer was computed ($\lfloor \frac{dep_level}{21} \rfloor$) and, depending on the value of the expected answer, actual answers were given using the probability distributions shown in Figure 2. Unlike in UNSLD, where uniform distribution was used, here we manually crafted these

⁴e.g. “(sad)|(unhappy)” for question 1, “(future)|(work out)” for question 2, “fail\w*” for question 3, “(pleasure)|(enjoy)” for question 4, etc.

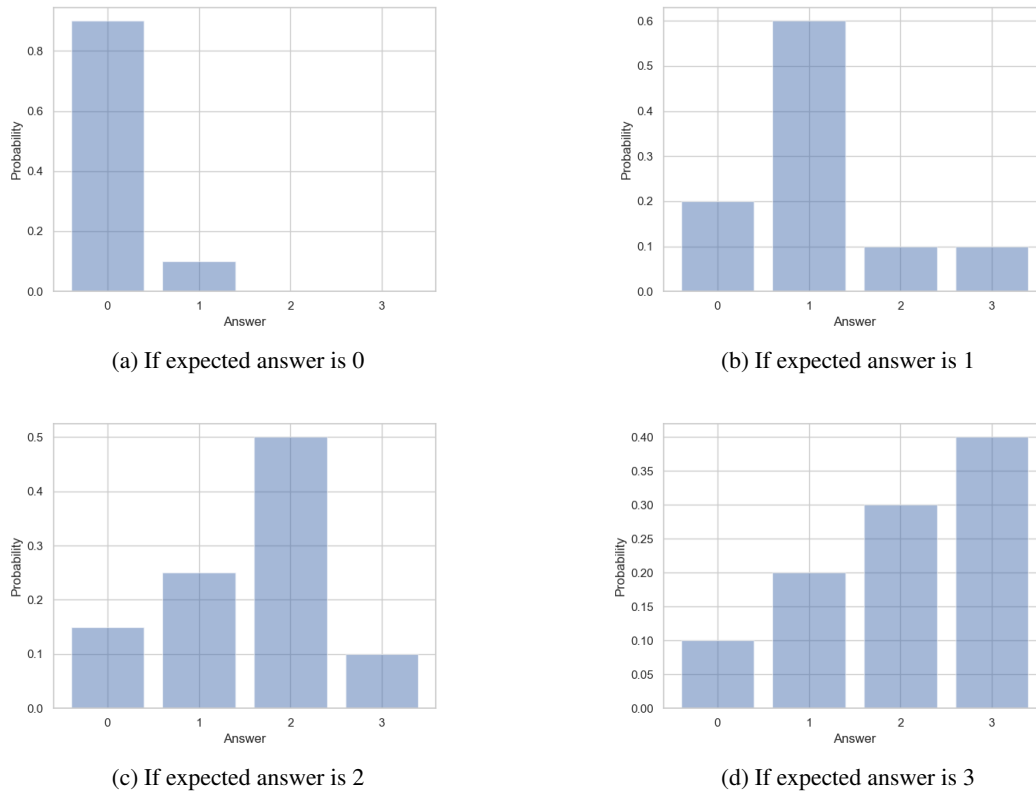


Figure 2: Discrete probability distribution for each possible expected answer.

probability distributions so that the expected answer is always more likely to be selected over the other ones. This model obtained the best ACR (71.27%) and the second-best AHR (40.71%) and ADODL (80.48%, best was only 0.54% above).

4 Evaluation Results

The task's results are shown in Table 2. As mentioned above, we obtained the best AHR (41.43%) and ACR (71.27%), and the second-best ADODL (80.48%) and DCHR (40%), best DCHR and ADODL were obtained by CAMH[14]. However, since most of our models' answers are stochastically generated, it implies that all of these measures are also stochastically generated.⁵ The natural question in cases like this is "How do we know these results properly represent our models performance and we did not obtained them just by pure chance?". In order to clarify this, once the eRisk finished and the golden truth was released, we run each model 1000 times and calculated the values for AHR, ACR, ADODL and DCHR each time.⁶ After this process finished, we ended up with a sample of 1000 values for each measure and model, which we then used to produce the box plots shown in Figure 3.

⁵Only ADODL and DCHR for UNSLA and DHR for UNSLB are deterministically determined by *depression_Level* and *c*.

⁶Just as if we had participated 1000 times in this task.

Analyzing the box plots one can notice that, in fact, when we participated we had a little bit of bad luck, specially for UNSLE's ADODL, since one can see in Figure 3c that the actual value we obtained (80.84%) is almost a lower bound outlier. Another important thing that can be seen in Figure 3a is that the use of "textual hints", in UNSLC, really improved the Average Hit Rate (AHR) but did not had an impact on the other measures (as seen in the other figures). In Figure 3c we can see that UNSLE was considerably the best method to estimate the overall depression level since its ADODL takes values within a range that is quite above the others. Additionally, another important aspect is that, looking at the range of values each method takes, for the different measures, in Figure 3, we can see that the obtained values would be among the best ones, even in the worst cases (compared against the other participant's).

5 The other teams' models

From the other 7 participating research teams, only three submitted a paper describing their models to the conference, namely, BioInfo@UAVR [15], BiTeM [16], and CAMH [14].

The BioInfo@UAVR [15] team submitted only one model. Their approach for addressing this task was a rule-based one. Each rule was modelled with respect to several behavioral and psycholinguistics patterns that

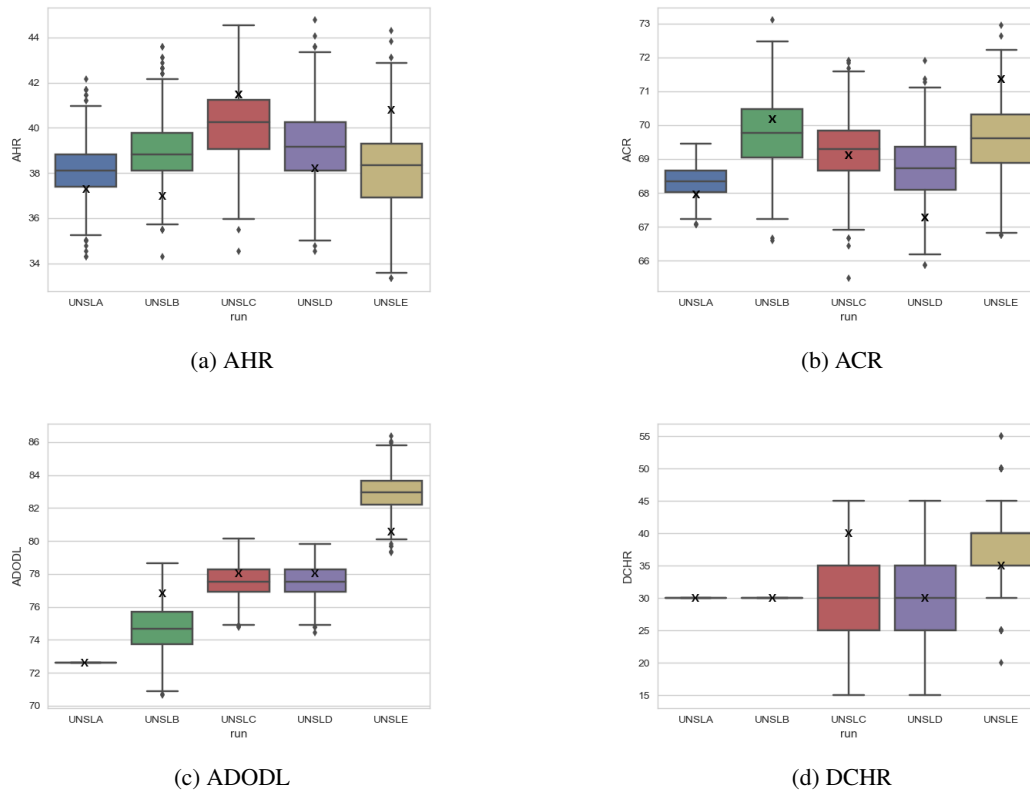


Figure 3: Box plots for each measure and run. Results submitted to eRisk 2019 are marked with an “x”.

Table 2: Results for eRisk 2019 Task 3.

Model	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	34.05%	66.43%	77.70%	25%
BiTeM	32.14%	62.62%	72.62%	25%
CAMH unsupervised	23.81%	57.06%	81.03%	45%
CAMH supervised (SVM - LIWC)	35.95%	66.59%	75.48%	25%
CAMH supervised (GPT - 180 features)	35.47%	68.33%	75.63%	20%
CAMH supervised (GPT - 768 features)	36.43%	67.22%	72.30%	20%
CAMH supervised (GPT - 948 features)	36.91%	69.13%	75.63%	15%
Fazl	22.38%	56.27%	72.78%	5%
Illinois	22.62%	56.19%	66.35%	40%
ISIKol multiSimilarity-5000-Dtac-Qtac	29.76%	57.94%	74.13%	25%
ISIKol-bm25-1.2-0.75-5000-Dtac-Qtac	29.76%	57.06%	72.78%	25%
ISIKol-lm-d-1.0-5000-Dtac-Qtac	30.00%	57.94%	73.02%	15%
Kimberly	38.33%	64.44%	66.19%	20%
UNSLA	37.38%	67.94%	72.86%	30%
UNSLB	36.93%	70.16%	76.83%	30%
UNSLC	41.43%	69.13%	78.02%	40%
UNSLD	38.10%	67.22%	78.02%	30%
UNSLE	40.71%	71.27%	80.48%	35%
Random (avg 1000 repetitions)	23.98%	58.55%	77.78%	33.55%

are known to be associated with the state of depression. Namely, they represent user’s text as a feature vector composed of a collection of the following 6 different types of features: depression, guilt, appetite, anxiety, fatigue, and sleep. For instance, “depression” features were made using the average polarity of user’s writ-

ings, the frequency of self-related words (e.g. I, myself, mine), absolutist words (e.g. absolutely, never, completely), anti-depressants listed by WebMD⁷, words

⁷<https://www.webmd.com/depression/guide/depression-medications-antidepressants>

related to mental disorders (e.g. depression, bipolar, psychotic). Then, they divided the 21 questions of the BDI questionnaire into 6 groups, one for each feature type. All questions belonging to a given group were scored with the same answer. The score was calculated, for each user, using the frequency of the features considered for each group with respect to its frequency over the entire dataset. These scores were then mapped to an actual answer (i.e. a [0-3] score) using manually predefined thresholds extracted from the frequency histograms.⁸

The BiTeM [16] team also submitted only one model. Their approach for addressing this task was using an ensemble of the following 3 different models:

- **Word polarity:** this model used word polarity to classify users as depressed and then to associate posts to relevant BDI dimensions. They made use of the Multi Perspective Question Answering (MPQA) subjectivity lexicon[17].⁹ For analyzing the posts with the BDI dimensions, they created a MPQA lexicon that provides single-word cues for each dimension. Additionally, they expanded this list of cues with the following resources: WordNet[18] to find synonyms, a sexual desires vocabulary¹⁰ and the F.E.A.S.T.'s Eating Disorders Glossary.¹¹ The annotation process of assigning BDI dimensions to each of the cues was done by three team members. Finally, questionnaire answers were calculated by normalizing the tag counts for each BDI dimension lexicon into a [0-3] score.
- **Mutual information:** for this model, a training dataset was created from Reddit, containing 107,129 posts. Using mutual information[19] they extracted the top-200 most relevant tokens for each category (depressed or not). Representing posts by those 200-most-relevant tokens, a logistic regression classifier was trained to classify posts into depressive or not. Then, keywords for each BDI question were created using WordNet and used to tag the positively classified posts. If a post was tagged as related to a BDI question, then the given answer was always 2, otherwise it was answered with 0 (i.e. answers were scored using a binary approach).
- **Semantic Similarity:** using pre-trained GloVe word embeddings¹² [20], a representation of each user post is generated by averaging the embeddings of all the words in the post. The same

process is performed over the BDI questionnaire to represent each answer, i.e., the embeddings of all the words in each questionnaire answer is also averaged. Finally, each BDI question was answered with the answer that was “semantically” closest to any of the user’s posts. This was carried out by computing the similarity of each user post and each questionnaire answer using the *cosine similarity*.

The CAMH [14] team submitted four supervised models and one unsupervised model. The unsupervised model was similar to the “Semantic Similarity” model explained above, but instead of using the averaged GloVe embeddings of all the post’s words to construct each post-level vector, here the GPT-1 (Generative Pre-trained Transformer version 1) [21] language model was used to generate them. The pre-trained GPT model was fine-tuned on the provided text from the 20 users (3 epochs). To predict the user’s response to each of questions, the *cosine similarity* of each of the possible responses to all of the user’s post was computed, and the answer that had the highest *cosine similarity* was picked. Therefore, like with the BiTeM’s “Semantic Similarity” model, this is a nearest neighbor approach that asks which possible response to the question is closest to what the user has previously written, but this time, in the space of the GPT-1 features. The supervised models were trained using a private dataset which was the result of one the team members’ previous work, it is a collection of BDI and several other questionnaires filled out by 236 undergraduate Psychology students (197 females, 39 males). The four supervised models were the following:

- **“SVM - LIWC”:** for each BDI question, a Support Vector Machine (with linear kernel and L2 regularization) to predict the corresponding answer. Before being fed to the SVMs, each user’s post (or training document) was converted into a “LIWC feature vector”. These features were extracted with the Linguistic Inquiry Word Count (LIWC) tool [22], which calculated the proportions of words from each post belonging to different categories.¹³
- **“GPT - 180 features”:** the same fine-tuned GPT-1 language model used for the unsupervised model was used here to represent each user by a “relationship feature vector”. Then, *auto-sklearn*¹⁴ was used to learn a classifier for each question. This “relationship feature vector” was created by concatenating two vectors, one holding the minimum euclidian distance between each of the possible 90 answers¹⁵ in the BDI questionnaire

⁸e.g. a score lower than 0.3 would be converted to 0, a score in the range (0.3,0.5) to 1, a score in the range [0.5,1) to 2, and anything over 1 to 3.

⁹http://mpqa.cs.pitt.edu/#subj_lexicon

¹⁰<https://www.macmillandictionary.com/thesaurus-category/british/feeling-sexual-excite-ment-or-desire>

¹¹<http://glossary.feast-ed.org/>

¹²<https://nlp.stanford.edu/projects/glove/>

¹³All available LIWC categories were used, resulting in 70 features per post

¹⁴<https://www.automl.org/automl/auto-sklearn/>

¹⁵19 questions x 4 possible answers + 2 questions x 7 possible answers each

and the user's posts, and the other, the maximum correlation between them. This resulted in a single vector of 180 relationship features.

- “GPT - 768 features”: This model is the same as the previous one, but instead of representing each user by a “relationship feature vector”, it uses an “average feature vector”. To construct this feature vector, first a post-level GPT-1 features vector is created for each user's post, and then all these post-level vectors are averaged to obtain a single average feature. As a result, each user's writing history was summarized into a single vector of size 768.¹⁶
- “GPT - 948 features”: this model is a variation of the previous one, it uses a “combined feature vector” to represent users. This vector is obtained by concatenating the relationship and average feature vectors used by the previous models, resulting in a single vector of size 948.

From the obtained results (see Table 2) one can see that, despite their complexity, none of these models performed better than ours, except for the CAMH's unsupervised model which had the best ADODL and DCHR but performed worst than random for the other two metrics, AHR and ACR. In contrast to all the models described in this section, which make use of embeddings, hand-crafted features, ensemble mechanisms, etc., our model is clearer and simpler since it only accumulates evidence word by word,¹⁷ which is then directly used to infer the overall depression level. In addition, the obtained performance by our 5 models were quite consistent across all 4 metrics, that is, our 5 models performed better than random for almost all 4 matrices and they obtained the best AHR and ACR values.

6 Conclusions and Future Work

In previous scenarios, machine learning models have shown that, in fact, language usage can provide strong evidence in detecting depressive people, since these models have to learn to characterize depression through language use. The work presented in this article is a first step towards measuring finer grain relationships between these both aspects, namely, we studied how the language usage could be connected with the *severity level* of depression. We tested our proposal by participating in the eRisk 2019 T3 task. Obtained results were promising and showed us that there could be a strong, and somewhat direct, relationship within these both aspects —i.e. it could really be a relationship between how users write, what words they use, and the actual depression level they have.

¹⁶Which is the size of GPT-1 vectors.

¹⁷Only two positive numbers are computed, one for depressed and the other for non-depressed.

Finally, since all the methods we used are based on the depression level predicted by SS3, results also showed us that SS3 correctly inferred the depression level (calculated by Equation (4)) from the textual evidence accumulated while processing the user's writings, i.e. SS3 correctly valued words in relation to each category (depressed and non-depressed).

Overall, these experiments indicate that it is possible to automatically extract some depression-related evidence from social media activity but we are still far from a really effective depression screening tool. ADODL and, particularly, DCHR metric showed that the models, although effective at answering some depression-related questions, do not perform well at estimating the overall level of depression of the users. For example, the best performing model gets the depression category right for only 45% of the users which indicates that, although better than random, there is still much room for improvement. However, most models are clearly better than random in terms of AHR and ACR. This suggests that the models do a reasonable job at getting answers right. This suggests that the analysis of the user posts is useful at extracting some signals or symptoms related to depression. For future work, we will try to get access to a bigger test set since more data is needed to draw better and more robust conclusions. Furthermore, since it might be interesting to analyze the results from a psychological point of view, and it is outside our area of expertise, we are looking forward to being able to work interdisciplinary with mental health professionals.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

Authorship statements are formatted with the names of authors first, followed by the CRediT¹⁸ roles.

Sergio G. Burdisso: Conceptualization, Methodology, Software, Investigation, Writing. **Marcelo Errecalde:** Methodology, Validation, Supervision, Writing. **Manuel Montes-y-Gómez:** Methodology, Validation, Supervision.

References

- [1] World Health Organization, *Depression and other common mental disorders: global health estimates*. WHO, 2017.
- [2] World Health Organization, *Preventing suicide: a global imperative*. WHO, 2014.
- [3] National Center for Health Statistics, “Mortality in the United States, 2017.” <https://www.cdc.gov/nchs/products/databriefs/db328.htm>, 2019. [Online; accessed 13-April-2019].
- [4] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, “An inventory for measuring depression,”

¹⁸<https://casrai.org/credit/>

- Archives of general psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.
- [5] D. E. Losada and F. Crestani, “A test collection for research on depression and language use,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 28–39, Springer, 2016.
- [6] D. E. Losada, F. Crestani, and J. Parapar, “erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 346–360, Springer, 2017.
- [7] D. E. Losada, F. Crestani, and J. Parapar, “Overview of erisk: Early risk prediction on the internet,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 343–361, Springer, 2018.
- [8] S. G. Burdisso, M. Errecalde, and M. M. y Gómez, “Towards measuring the severity of depression in social media via text classification,” in *Actas del XXV Congreso Argentino de Ciencias de la Computación (CACIC 2019)*, pp. 577–588, 2019.
- [9] D. E. Losada, F. Crestani, and J. Parapar, “Overview of eRisk 2019: Early Risk Prediction on the Internet,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019*, (Lugano, Switzerland), Springer International Publishing, 2019.
- [10] S. G. Burdisso, M. Errecalde, and M. M. y Gómez, “A text classification framework for simple and effective early depression detection over social media streams,” *Expert Systems with Applications*, vol. 133, pp. 182–197, 2019.
- [11] S. G. Burdisso, M. Errecalde, and M. Montes-y Gómez, “ τ -SS3: A text classifier with dynamic n-grams for early risk detection over text streams,” *Pattern Recognition Letters*, vol. 138, pp. 130–137, 2020.
- [12] D. G. Funez, M. J. G. Ucelay, M. P. Villegas, S. G. Burdisso, L. C. Cagnina, M. Montes-y Gómez, and M. L. Errecalde, “UNSL’s participation at erisk 2018 lab,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Association, CLEF 2018*, (Avignon, France), Springer International Publishing, 2018.
- [13] M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. G. Ucelay, and L. C. Cagnina, “Temporal variation of terms as concept space for early risk prediction,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017*, (Dublin, Ireland), Springer International Publishing, 2017.
- [14] P. Abed-Esfahani, D. Howard, M. Maslej, S. Patel, V. Mann, S. Goegan, and L. French, “Transfer learning for depression: Early detection and severity prediction from social media postings,” in *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, (Lugano, Switzerland), 2019.
- [15] A. Trifan and J. L. Oliveira, “Bioinfo@ uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders,” in *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, (Lugano, Switzerland), 2019.
- [16] P. van Rijen, D. Teodoro, N. Naderi, L. Mottin, J. Knafou, M. Jeffryes, and P. Ruch, “A data-driven approach for measuring the severity of the signs of depression using reddit posts,” in *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, (Lugano, Switzerland), 2019.
- [17] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 347–354, 2005.
- [18] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.
- [19] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical review E*, vol. 69, no. 6, p. 066138, 2004.
- [20] K. Ethayarajh, “Unsupervised random walk sentence embeddings: A strong but simple baseline,” in *Proceedings of The Third Workshop on Representation Learning for NLP*, pp. 91–100, 2018.
- [21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf*, 2018.
- [22] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, “The development and psychometric properties of LIWC2007,” 2007. LIWC, Austin, Texas.

Citation: S. Burdisso, M. Errecalde and M. Montes-y-Gómez. *Using Text Classification to Estimate the Depression Level of Reddit Users*. Journal of Computer Science & Technology, vol. 21, no. 1, pp. 1–10, 2021.

DOI: 10.24215/16666038.21.e01

Received: March 17, 2020 **Accepted:** March 19, 2021.

Copyright: This article is distributed under the terms of the Creative Commons License CC-BY-NC.