Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6111–6117
Marseille, 11–16 May 2020

Searching Brazilian Twitter for signs of mental health issues

Wesley Ramos dos Santos, Amanda Maria Martins Funabashi, Ivandré Paraboni
School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
{wesley.ramos,amanda.funabashi,ivandre}@usp.br

Abstract

Depression and related mental health issues are often reflected in the language employed by the individuals who suffer from these conditions and, accordingly, research in Natural Language Processing (NLP) and related fields have developed an increasing number of studies devoted to their recognition in social media text. Some of these studies have also attempted to go beyond recognition by focusing on the early signs of these illnesses, and by analysing the users' publication history over time to potentially prevent further harm. The two kinds of study are of course overlapping, and often make use of supervised machine learning methods based on annotated corpora. However, as in many other fields, existing resources are largely devoted to English NLP, and there is little support for these studies in under resourced languages. To bridge this gap, in this paper we describe the initial steps towards building a novel resource of this kind - a corpus intended to support both the recognition of mental health issues and the temporal analysis of these illnesses - in the Brazilian Portuguese language, and initial results of a number of experiments in text classification addressing both tasks.

Keywords: text classification, mental health, depression

1. Introduction

Depression and related mental health issues are well-known challenges of modern life, and are often reflected in the language employed by the individuals who suffer from these conditions. Accordingly, research in Natural Language Processing (NLP) and related fields have developed an increasing number of studies devoted to the computational recognition of depression, anxiety, bipolar disorder, anorexia, suicidal thought and many others from social media text (usually Twitter or Reddit) (Resnik et al., 2013; Resnik et al., 2015; Jamil et al., 2017; Yates et al., 2017; Orabi et al., 2018). In addition to these, a second line of research has recently attempted to go beyond recognition by focusing on the early signs of these illnesses, and by analysing the users' publication history over time to potentially anticipate treatment and prevent further harm (Trotzek et al., 2018b; Losada et al., 2019).

The two kinds of study are of course overlapping, and usually make use of supervised machine learning methods based on annotated corpora (Coppersmith et al., 2015; Losada et al., 2017; Yates et al., 2017). As in many other fields, however, existing resources are largely devoted to English NLP, and there is little support for studies of this kind in under resourced languages.

To help bridge this gap, this paper describes the initial steps towards building a novel language resource of this kind, namely, a Brazilian Portuguese corpus of tweets written by users with a diagnosed mental health issue. The corpus - hereby called DepressBR - is intended to support both standard recognition of depression and related issues as in Coppersmith et al. (2015), and temporal analysis of these illnesses as in Losada et al. (2017). The corpus consists of a collection of tweets written by individuals who self-reported a mental health issue or the beginning of a treatment for one such condition at a specific, well-defined point in time and, as a control group, tweets written by users who did not suggest having any condition of this kind.

In its initial version, the DepressBR corpus contains 616k tweets (7.3mi words) and, although still smaller than similar resources for the English language, is presently taken as the basis of three experiments focused on tweet-level classification intended to illustrate the possible use of the data collected so far. The first two experiments address the recognition of mental health issues, and the third consists of an analysis of how much data is actually needed for classification given the goal of anticipating diagnoses as far ahead as possible.

The rest of this paper is structured as follows. Section 2 reviews existing work in depression detection in text, and related tasks. Section 3 describes our own data collection task and presents descriptive statistics of the corpus obtained so far. Section 4 reports an experiment to detect users with mental health issues based on the entire corpus data, and Section 5 addresses the same task by focusing only on messages in which users talk about themselves. Section 6 discusses the early detection of mental health issues, and Section 7 presents additional remarks and discusses the next steps in the current project.

2. Background

The present work combines many of the guidelines applied to the collection of the CLPsych-2015 shared task corpus (Coppersmith et al., 2015) whilst keeping the notion of temporal information required for early detection of mental issues as proposed in Losada and Crestani (2016). These and other related studies are briefly reviewed as follows.

The CLPsych-2015 corpus (Coppersmith et al., 2015) contains tweets produced by 1,746 users in the English language. The corpus was developed as a resource for the study of depression and post-traumatic stress disorder (PTSD) detection in the context of a shared task. Instances of depressed users were identified by searching for self-reported diagnoses as in 'I was just diagnosed with depression/PTSD'. The publicly available labelled dataset contains 327 depressed and 246 PTSD users, with age- and gender-matched control users.

Among the participant systems in the CLPsych-2015 shared task, the work in Resnik et al. (2015) obtained the overall best results. The work presented a number of models that combine supervised LDA, TF-IDF counts and others using SVM classifiers, with results that were consistently high for both depression and PTSD detection.

In Losada and Crestani (2016) a corpus of English texts is created by following a method similar to Coppersmith et al. (2015). The work argues against the use of Twitter data for the task by pointing out that Twitter data is generally difficult to redistribute and usually limited to up to 3,200 messages per user. Thus, the Reddit domain was chosen instead. The corpus contains 137 users in the depressed category, and further 755 in a control group.

The corpus in Losada and Crestani (2016) was taken as the basis for two shared tasks on depression detection (Losada et al., 2017; Losada et al., 2018) called eRisk, in which the computational detection of early signs of mental conditions, datasets and relevant metrics were discussed in detail. Among the participant systems in the eRisk tasks, a series of experiments in Trotzek et al. (2018a; Trotzek et al. (2018b) have attempted a large number of computational strategies. Among these, an ensemble model that combines different base predictions (including user-level linguistic meta data, bag of words, neural word embeddings, and convolutional neural networks) was found to obtain the best overall results.

In addition to the CLPsych-2015 and eRisk shared task series, an increasingly large number of studies have focused on depression detection on social media in English using similar resources. Among these, the work in Yates et al. (2017) presents a large dataset in the Reddit domain, containing 9,210 diagnosed and 107,274 control users. The work shows that a CNN model outperforms (with an average F1 score of 0.51) MNB and SVM baseline classifiers based on two feature sets: standard bag of words, and a feature-rich model comprising bag of words features encoded as sparse weighted vectors, psycholinguistics-motivated LIWC word counts (Pennebaker et al., 2001) and emotion-related lexical features.

The work in Jamil et al. (2017) presents user- and tweet-level depression classifiers based on the CLPsych-2015 corpus and on a collection of 8,753 messages produced in the context of an awareness campaign called Bell Let's Talk 2015. Messages were written by 60 Twitter users, being 30 identified as depressed and another 30 as part of a control group. Both tasks made use of SVM classifiers and, since the dataset was found to be heavily imbalanced (with 95% of tweets unrelated to depression), some of the reported experiments performed SMOTE re-sampling (Chawla et al., 2002). The models under discussion made use of a wide range of feature combinations, including bag of words, polarity, depression and LIWC word counts, community-related and sentiment features, among others.

Finally, the work in Orabi et al. (2018) addresses the issue of depression classification on Twitter by presenting a series of CNN and BI-LSTM models with optimised word embeddings. The study also makes use of training data provided by the CLPsych-2015 corpus (Coppersmith et al., 2015) and test data provided by the Bell Let's Talk corpus (Jamil et al., 2017). For the CLPsych-2015 data, a CNN model with a global max pooling layer was found to outperform a number of alternatives, including other CNN and LSTM architectures. For the Bell Let's talk data, the best alternative consisted of a model based on a CNN Multi Channel architecture.

3. Corpus

As a means to further research in the classification of depression and related mental health issues from Brazilian Portuguese text, we collected a Twitter corpus consisting of users who reported being diagnosed with a mental condition by a professional, or who reported starting treatment for one such condition. To this end, we only considered cases in which the diagnosis or treatment start is explicitly mentioned, allowing us to pinpoint a specific moment that clearly divides the user's publication history in two groups: messages written before the diagnosis/treatment event, and those written during or after the event.

In addition to that, we also collected data from a disjoint group of users who discuss mental health issues on Twitter, but who do not suggest being diagnosed or under treatment in any way. The organisation of this data, hereby referred to as the 'diagnosed' and 'control' groups, is detailed in the next section. It should be clear, however, that there is no guarantee that the control group is free from users who might be under mental health treatment and, conversely, there is no guarantee that all users in the diagnosed group have been truly diagnosed/treated either. Although our preliminary analysis of the message history below suggests that the number of false positives is likely small, a certain amount of noise is assumed to be part of the present computational challenge.

3.1. Procedure

We created a corpus for the study of mental health issues by searching Brazilian Twitter for 'diagnosed' users. To this end, we ran a number of queries with terms that denote depression, anxiety, bipolar and panic disorders, and which are related to terms that denote diagnosis, medical treatment, or the use of antidepressant drugs (which imply a medical prescription.) The kinds of query under consideration, translated from the original Portuguese, are summarised as follows.

- *prescribed antidepressants*
- *I started taking antidepressants*
- *I started treatment + antidepressants / depression / anxiety / bipolar / panic*
- *I + diagnosed + depression / anxiety / bipolar / panic*
- *today doctor said I + depression / anxiety / bipolar / panic*

All messages that matched the queries were manually inspected and, in case that they seemed sufficiently genuine, we collected the 3,200 most recent messages written by their authors, which are therefore labelled at user level only. For each selected author, messages were examined so as to identify the specific point in time in which the diagnosis or treatment started, and to single out the subset of messages that were published prior to the event. Thus, for instance, the date referred to in ‘Last Friday I started taking antidepressants’ is marked as the time of the diagnosis/treatment event. When it was not possible to unambiguously pinpoint a date prior to the diagnosis/treatment event, or when the event occurred before the 3,200 messages publication history, the user and all his/her messages were discarded from the dataset. This was mainly the case of vague (e.g., ‘I was once diagnosed with depression’), distant (e.g., ‘Ten years ago I was treated for anxiety’) or continuous treatment (e.g., ‘Today doctor prescribed me *even more* antidepressants’) events. After filtering out unsuitable cases, all data that passed these consistency checks were taken as part of the ‘diagnosed’ group in our corpus.

Regarding the control group, we face the challenge of identifying users that were not diagnosed or treated for mental health problems at all. This is obviously complicated by the fact that not explicitly self-reporting a mental health treatment cannot be taken as evidence of not being under treatment.

Since Twitter does not support well-defined groups such as Reddit depression communities considered in Losada and Crestani (2016), we decide to spread the risk of including users with an unreported diagnose/treatment in the control group by selecting small groups from a number of categories. More specifically, we searched for users who manifested an interest in mental health issues either as (i) a general concern (e.g., by promoting the ‘Yellow September’ suicide prevention campaign), (ii) as a concern towards a particular person who suffered from a mental health issue (e.g., a friend etc.), or (iii) for being a Psychology student with a particular interest in the depression topic. The relevant queries are summarised as follows, once again adapted from the Portuguese original.

- *take care (of your) friend + yellow September*
- *help (a,your) friend + depression*
- *friend diagnosed (with) depression*
- *friend takes antidepressant*
- *Psychology student + depression*
- *lost + friend + depression*

Messages obtained in response to these queries were manually inspected to remove users who suggested (in the same or in another message) that they were under psychological treatment themselves.

3.2. Data collection results

As in Coppersmith et al. (2015), at this initial stage of our project we chose to create a relatively well-balanced dataset - which favours the comparison among machine learning methods - rather than creating a dataset that reflects the class balance that would be observed in a more realistic setting. This, according to Coppersmith et al. (2015), would require a control group 7-15 times larger than the diagnosed group.

We selected 106 diagnosed and 118 control users as discussed in the previous section. Table 1 presents descriptive statistics of the final dataset.

	Diagnosed	Control	Overall
users	106	118	224
words	3,404,156	3,958,011	7,362,167
messages	284,341	332,028	616,369
words/msgs	11.97	11.92	11.94
vocabulary	206,826	243,327	386,202

Table 1: Descriptive statistics for the entire corpus after removal of undetermined language messages.

Although messages are consistently short in both classes (i.e., on average less than 12 words per tweet), we notice that the diagnosed group produced fewer messages, and has a smaller vocabulary if compared to the control group. Both vocabularies are however still large since at this stage we have not performed any pre-processing or feature selection.

Table 2 presents descriptive statistics regarding the subset of messages prior to the diagnosis/treatment self-report made by diagnosed users.

Metrics	Mean	Min.	Max.
Number of messages	1,671	48	3,144
Number of words	19,188	489	46,757
Time span (days)	536	1	2,840

Table 2: Descriptive statistics for messages prior to the self-reported diagnosis or treatment.

The data includes three users whose elapsed time between the start of his/her Twitter history and diagnosis is shorter than one week, and one user whose elapsed time is only one day, which is of course inadequate for prediction purposes. For most users, however, the actual time span is considerably large, with about one third of them providing at least one year of data prior to diagnosis, with an average of 1.4 years (1.67k tweets) of data history.

Figure 1 illustrates the distribution of users according to the number of days of data history prior to diagnosis/treatment.

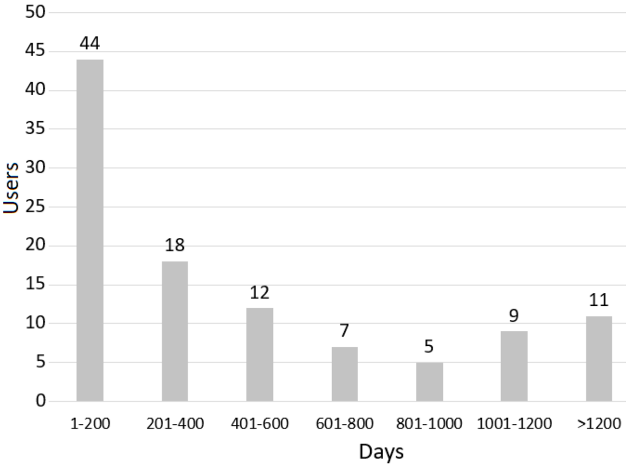

Figure 1: Users distribution according to number of days of data history prior to diagnosis/treatment.

4. Experiment 1: Detecting depression-related tweets

As a means to illustrate the computational challenge of detecting diagnosed Twitter users, we ran a simple experiment involving a number of tweet-level classification models. In doing so, our goal is to produce initial results to be taken as reference for more developed studies based on the present dataset.

4.1. Models

The experiment considers three standard text classifiers - based on psycholinguistics-motivated features, TF-IDF counts and averaging word embeddings - and a majority class baseline system. These are summarised in Table 3 and further discussed below.

Model	Method	Features
LR.LIWC	Log.regression	LIWC word counts
LR.Tfidf	Log.regression	K-best TF-IDF counts
MLP.WordEmb	Multilayer perc.	Avg. word embeddings
Baseline	Majority class	na

Table 3: Models

The *LR.LIWC* model takes as an input the 64-feature set provided by the Brazilian Portuguese version of the LIWC dictionary (Pennebaker et al., 2001) discussed in Balage Filho et al. (2013). LIWC provides word categories based on affective (positive and negative emotions, anger etc.), cognitive (e.g., insight, causation etc.) and perceptual processes (e.g., see, hear etc.), among others, and it is a popular knowledge source for sentiment analysis and author profiling models in general (dos Santos et al., 2017; Silva and Paraboni, 2018b; Silva and Paraboni, 2018a).

LR.LIWC is built by counting the number of words belonging to each LIWC category. When a word belongs to more than one categories simultaneously (e.g., 'he' is both a pronoun and a masculine word), all corresponding counts are updated. After scanning the entire document, counts are divided by the total number of words, therefore creating 64-feature vectors with values within the 0..1 range. The model uses logistic regression with balanced class weights, L2 penalty and a lbfgs solver.

The *LR.Tfidf* model takes as an input TF-IDF vectors computed from the input document, and subsequently reduced to the k=15,000 best features with the aid of univariate feature selection using ANOVA f-value as a score function. The model also uses logistic regression with balanced class weights, L2 penalty and the lbfgs solver.

Finally, the *MLP.WordEmb* model takes as an input a document representation built from self-trained TF-IDF-weighted average word embeddings of size 100, which in turn are computed using Word2vec Skipgram (Mikolov et al., 2013). Document vectors are created by computing the weighted average of individual word embedding vectors multiplied by the individual TF-IDF scores of each word. In doing so, only word embeddings corresponding to the k=15,000 best features are considered. As in the previous *LR.Tfidf* model, features are selected with the aid of univariate feature selection using ANOVA f-value as a score function. *MLP.WordEmb* uses a multi-layer perceptron (MLP) classifier with one hidden layer of 25 neurons, rectified linear unit (ReLU) as an activation function, and Adam as a solver.

All models take as an input the entire set of messages written by every individual. In the case of the diagnosed class, however, the message denoting the diagnosis or treatment event was removed to prevent the classifiers from finding a trivial solution (e.g., by focusing on expressions of the kind 'I was diagnosed' etc.) For evaluation purposes, a random 80:20 stratified train-test split was performed.

4.2. Results

Table 4 summarises results for diagnosed versus control classification obtained by each of the models introduced in the previous section, and overall classification results (i.e., obtained by taking both classes into account.)

From these results we notice that recognising diagnosed users is considerably more challenging than recognising those in the control group, and that *LR.Tfidf* outperforms the alternatives for both classes.

5. Experiment 2: Focusing on self reports

Using the entire set of messages written by an individual to detect mental health issues may be problematic given that many - or possibly most - messages may be neutral with respect to his/her feelings or thoughts. To shed light on this issue, we envisaged a variation of the previous experiment in which we focus on self reports only. More specifically, we applied a simple heuristics to retain only those messages containing the first person 'I' pronoun from the corpus. In doing so, the original set of 616,369 messages (284,341 written by diagnosed and 332,028 written by control users) was reduced to 118,309 messages (55,648 written by diagnosed and 62,661 written by control users.)

5.1. Models

The experiment considers the same classifiers in the previous Table 3 in Section 4.1., namely, *LR.LIWC*, *LR.Tfidf*, *MLP.WordEmb* and the majority class baseline.

5.2. Results

Table 5 summarises results for diagnosed versus control classification obtained by each of the models discussed in the previous section when using only messages that included the 'I' pronoun, and overall classification results.

Once again, the use of TF-IDF counts with univariate feature selection outperforms the alternatives. However, we notice that focusing on self report messages, although greatly simplifying the classification task, has decreased overall results if compared to those obtained in the previous experiment, which were based on the entire dataset. Clearly, messages that do not contain the 'I' pronoun are relevant for the detection of mental health issues in text as well. These may include, for instance, messages containing possessive pronouns (e.g., 'mine') and others presently not accounted for (Paraboni and de Lima, 1998; Cuevas and Paraboni, 2008).

	Diagnosed			Control			Overall		
Model	P	R	F	P	R	F	P	R	F
Majority	0.00	0.00	0.00	0.54	1.00	0.70	0.29	0.54	0.38
LR.LIWC	0.48	0.53	0.50	0.56	0.51	0.53	0.52	0.52	0.52
MLP.WordEmb	0.56	0.52	0.54	0.62	0.65	0.63	0.59	0.59	0.59
LR.Tfidf	0.68	0.63	**0.65**	0.70	0.75	**0.72**	0.69	0.69	**0.69**

Table 4: . Classification of Twitter users with diagnosed mental health issues based on the entire set of messages written by each individual. Best F1 scores for each class are highlighted.

	Diagnosed			Control			Overall		
Model	P	R	F	P	R	F	P	R	F
Majority	0.00	0.00	0.00	0.54	1.00	**0.70**	0.29	0.54	0.38
LR.LIWC	0.50	0.53	0.51	0.56	0.52	0.54	0.53	0.53	0.53
MLP.WordEmb	0.53	0.43	0.48	0.57	0.66	0.61	0.55	0.56	0.55
LR.Tfidf	0.62	0.61	**0.61**	0.66	0.67	0.66	0.64	0.64	**0.64**

Table 5: . Classification of Twitter users with diagnosed mental health issues based on self reports only. Best F1 scores for each class are highlighted.

6. Experiment 3: Early detection of mental health issues

Finally, as a means to illustrate the potential for early detection of mental health issues in the current corpus, we envisaged an experiment to classify diagnosed users based on tweet sets of different sizes, all of which selected from their earliest publications. In doing so, our goal is to assess the amount of data required for identifying users who will develop a mental condition, and how early (i.e., before the actual diagnosis/treatment event) identification is possible.

6.1. Models

We created five classification tasks by varying the amount of data that each model had access to. For the diagnosed class, we selected the earliest 10, 50, 100, 200 or 500 tweets of each user. As a control group, we randomly selected similarly sized sets of tweets from every user in the control portion of the corpus data. The resulting class distribution is summarised in Table 6.

Tweets	Diagnosed	Control
Earliest 10	1,060	1,180
Earliest 50	5,296	5,900
Earliest 100	10,442	11,800
Earliest 200	20,339	23,600
Earliest 500	47,988	59,000

Table 6: Class distribution for different portions of the users' earliest tweets.

For all datasets, we used the *LR.Tfidf* and *LR.LIWC* classifiers discussed in the previous section.

6.2. Results

Table 7 presents weighted F1 score obtained for the diagnosed class based on each dataset, and the average number of days by which diagnosis could in principle be anticipated. This is computed as the number of days between the date of the actual diagnosis/treatment and the date of the latest tweet that the classifier had access to.

Tweets	Avg. Days	LR.LIWC	LR.Tfidf
Earliest 10	497	0.51	**0.55**
Earliest 50	474	0.48	**0.58**
Earliest 100	457	0.49	**0.58**
Earliest 200	415	0.49	**0.61**
Earliest 500	339	0.50	**0.64**

Table 7: Weighted F1 scores for the diagnosed group, and average number of days of diagnosis anticipation based on different portions of the users' Twitter history. Best F1 scores for each dataset are highlighted.

From these results, a number of observations are warranted. First, we notice that using more data improves results for *LR.Tfidf*, whereas *LR.LIWC* results remain relatively stable. Second, *LR.Tfidf* makes potentially useful predictions even from a very small dataset (e.g., when using only the ten earliest publications of each user.) Moreover, even when the longest (500 tweets) data history is considered, diagnosis could still be anticipated in almost one year, on average.

7. Final Remarks

This paper described the collection of the DepressBR corpus of messages written by individuals who self-reported a mental health condition or the start of a treatment for one such condition at a specific, well-defined point in time, and a control group of tweets written by users who did not explicitly report any condition of this kind. The corpus is intended as a resource for the automated recognition of mental health conditions in the Brazilian Twitter domain, with a particular focus on the issue of early detection (i.e., prior to the moment in each the diagnosis or treatment actually started.)

As a means to illustrate the use of the data collected so far, three experiments in tweet-level classification were carried out: the recognition of mental health issues from either the entire set of messages, or from the subset of messages containing first person pronouns only, and an analysis of how much data was actually needed for classification given the goal of anticipating the diagnoses as far ahead as possible. Preliminary results are in our view encouraging - in particular, we notice that even with a relatively small number of messages it is in principle possible to anticipate a mental health issue in several weeks or months - and pave the way for more focused research on this computational task.

As future work, we intend to expand the corpus by identifying more diagnosed users and, in particular, we are aware of the need for expanding the control group to model more realistic scenarios. We also intend to investigate more refined message selection strategies as a means to identifying relevant information sources for user-level classification. We notice also that the availability of a larger dataset should enable us to make use of more recent deep learning methods in the present tasks.

Other possible lines of investigation include using computational models of personality recognition (Ramos et al., 2018; dos Santos et al., 2020), author profiling (Hsieh et al., 2018) and moral stance classification (dos Santos and Paraboni, 2019) to aid the detection of mental health issues in text, possibly combined with ensemble methods for text classification (Custódio and Paraboni, 2018).

Acknowledgements

This work received support from the University of São Paulo grant # 668/2018.

8. Bibliographical References

Balage Filho, P. P., Aluísio, S. M., and Pardo, T. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology - STIL*, pages 215–219, Fortaleza, Brazil.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Coppersmith, G., Dredze, M., Harman, C., Kristy, H., and Mitchell, M. (2015). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the Second Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, USA. Association for Computational Linguistics.

Cuevas, R. R. M. and Paraboni, I. (2008). A machine learning approach to Portuguese pronoun resolution. In *IBERAMIA-2008, Lecture Notes in Artificial Intelligence 5290*, pages 262–271, Lisboa, Portugal. Springer-Verlag.

Custódio, J. E. and Paraboni, I. (2018). EACH-USP ensemble cross-domain authorship attribution. In *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, Avignon, France.

dos Santos, W. R. and Paraboni, I. (2019). Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In *Recents Advances in Natural Language Processing (RANLP-2019)*, pages 1069–1075, Varna, Bulgaria.

dos Santos, V. G., Paraboni, I., and Silva, B. B. C. (2017). Big five personality recognition from multiple text genres. In *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, pages 29–37, Prague, Czech Republic. Springer-Verlag.

dos Santos, W. R., Ramos, R. M. S., and Paraboni, I. (2020). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.

Hsieh, F. C., Dias, R. F. S., and Paraboni, I. (2018). Author profiling from facebook corpora. In *11th International Conference on Language Resources and Evaluation (LREC-2018)*, pages 2566–2570, Miyazaki, Japan. ELRA.

Jamil, Z., Inkpen, D., Buddhitha, P., and White, K. (2017). Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver. Association for Computational Linguistics.

Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, Cham. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2017). eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In *Lecture Notes in Computer Science vol 10456*, pages 346–360, Cham. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2018). Overview of eRisk: Early Risk Prediction on the Internet. In *Lecture Notes in Computer Science vol 11018*, pages 343–361, Cham. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. In *Lecture Notes in Computer Science vol 11696*.

Mikolov, T., Wen-tau, S., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proc. of NAACL-HLT-2013*, pages 746–751, Atlanta, USA. Association for Computational Linguistics.

Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. (2018). Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, USA. Association for Computational Linguistics.

Paraboni, I. and de Lima, V. L. S. (1998). Possessive pronominal anaphor resolution in Portuguese written texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1010–1014. Association for Computational Linguistics.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., and Dias, R. F. S. (2018). Building a corpus for personality-dependent natural language understanding and generation. In *11th International Conference on Language Resources and Evaluation (LREC-*

2018), pages 1138–1145, Miyazaki, Japan. ELRA.

Resnik, P., Garron, A., and Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, USA. Association for Computational Linguistics.

Resnik, P., Armstrong, W., Claudino, L., and Nguyen, T. (2015). The university of Maryland CLPsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, Denver, USA. Association for Computational Linguistics.

Silva, B. B. C. and Paraboni, I. (2018a). Learning personality traits from Facebook text. *IEEE Latin America Transactions*, 16(4):1256–1262.

Silva, B. B. C. and Paraboni, I. (2018b). Personality recognition from Facebook text. In *13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS vol. 11122*, pages 107–114, Canela. Springer-Verlag.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2018a). Early detection of depression based on linguistic metadata augmented classifiers revisited. In *International Conference of the Cross-Language Evaluation Forum for European Languages CLEF-2018*, pages 191–202, Avignon. Springer.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2018b). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.