Noname manuscript No.
(will be inserted by the editor)

Evaluating and Improving Lexical Resources for Detecting Signs of Depression in Text

David E. Losada · Pablo Gamallo

Received: date / Accepted: date

Abstract While considerable attention has been given to the analysis of texts written by depressed individuals, few studies were interested in evaluating and improving lexical resources for supporting the detection of signs of depression in text. In this paper, we present a search-based methodology to evaluate existing depression lexica. To meet this aim, we exploit existing resources for depression and language use and we analyze which elements of the lexicon are the most effective at revealing depression symptoms. Furthermore, we propose innovative expansion strategies able to further enhance the quality of the lexica.

Keywords Depression Screening · Depression Lexicon · Lexicon Evaluation · Lexicon Expansion · Text Analysis · Natural Language Processing

1 Introduction

Automatic Text Analysis to detect signs of depression is an increasingly important research topic. Depression is a common mental disorder that severely impacts our society. According to the World Health Organization (WHO)[1], more than 300 million people of all ages suffer from this type of mental illness. This is a serious health condition that causes the affected person to suffer greatly, function poorly and, at its worst, it can lead to suicide. Although there exist effective treatments, WHO estimates that fewer than half of those affected by depression receive such treatments. Rates of diagnosing depression have improved over the past decades, but the prevalence of depression

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela, Spain
Tel.: +34 8818-16400
Fax: +34 8818-16405
E-mail: david.losada@usc.es,pablo.gamallo@usc.es

[1] `http://www.who.int/mediacentre/factsheets/fs369/en/`

in our society makes that many cases still remain undetected (Cepoiu et al, 2008). Furthermore, only a low percentage of the detected cases receive adequate treatment (Wang et al, 2005). New forms of intervention are required to enhance treatment initiation.

Symptoms and signs associated with depression are observable on the Internet and automatic text analysis is a promising tool for early detection. For example, individuals whose writings (e.g. posts on a Social Media website or messages in web forums) show elevated depression could be targeted for a more thorough assessment or provided with support and further resources. As argued by Nease and Maloin (2003), social-media based screening may become a valuable stage in a mental health screening strategy as a means to overcome the limitations of short screening inventories. As a matter of fact, multi-step screening strategies can alleviate the low true positive rate and high false positive rate associated with assessments done by non-psychiatric physicians (Cepoiu et al, 2008; Mitchell et al, 2011).

Our modest objective is to analyze and improve current language resources for identifying signs of depression in a text. By no means we aim at designing diagnostic tools. A subject whose writings show signs of depression is not necessarily depressed. This should be determined by trained professionals and such diagnosis is out of the scope of our work. Nevertheless, we strongly believe that Text Analytics may be highly relevant to complement human experts.

Of course, putting this technology into practice is a challenge that requires to consider ethics at all steps. Design considerations need to honour the privacy of the affected subjects and, additionally, appropriate ethical guidelines need to be defined. For example, an automatic tool may be configured to reveal detected risks to the subjects themselves, or to identified contacts (e.g. clinicians or trusted authorities). This must be done under a proper framework that ensures that the intended benefits outweigh the risks. In this work we do not further explore these ethical issues but we acknowledge that we must take them into account in the future exploitation of this research.

A number of studies have attempted to build predictive models that detect depression and other mental illnesses on Social Media (Guntuku et al, 2017). Many of these studies employ textual features extracted from Social Media data. Language is commonly encoded with word-based features. Word weights, computed from frequencies or from more sophisticated pieces of evidence, are typically utilized. In the absence of training data, the availability of domain-specific lexica, such as depression-oriented dictionaries, plays a fundamental role. Lexica can be employed to automatically analyze the level of depression in texts. An accurate lexicon might become a valuable guidance to understand the text's author. With such a tool, we could evaluate whether the author is referring to melancholic or physical symptoms of depression. For example, Neuman et al (2012) developed a lexicon-based method for screening for depression that is fully automatic and unsupervised. The importance of lexica and dictionaries has been also shown in recent challenges on language and depression. For example, Almeida et al (2017a) developed a system for

early detection of depression that combined supervised learning and information retrieval. The method relied on depression-related dictionaries.

However, the creation of depression-specific lexica is challenging. Individuals employ a wide range of linguistic means to express depression. Such variety of linguistic artifacts cannot be identified in advance by a group of experts. Only a few resources are publicly available. This includes the Pedesis lexicon, created by Neuman et al (2012), and the depression lexicon created by Choudhury et al (2013). In our work, we build on these previous resources and provide a thorough evaluation of them. To meet this aim, we propose a new evaluation methodology to test how effective the lexica are to screen for signs of depression. Furthermore, we suggest and evaluate some methods to improve the lexica. These methods expand the lexicon with selected terms following distributional and thesaurus-based models. The result of this addition are enhanced lexica that are publicly available. Our experiments show that the resulting lexica are effective at identifying signs of depression and can be used to analyze text in a non-supervised way. By identifying the most important components of the original lexica and expanding them with related terms, we have gained a more complete picture of the linguistic artifacts people use to express depression.

The rest of this paper is organized as follows. Related work is presented in section 2, section 3 presents our strategy to evaluate depression lexica and reports some experiments performed with two existing lexica for depression. Section 4 discusses our proposal to enhance the lexica through selective expansion and presents additional experiments with the improved lexica. The paper ends with some conclusions and lines of future work.

2 Related Work

Screening for depression through online data is an increasingly important research area (Guntuku et al, 2017). Studies to date have mostly focused on designing automated methods to evaluate the level of depression in texts. Many predictive models use content-based features, such as word frequencies or linguistic variables extracted from general-purpose resources (e.g. LIWC (Ramirez-Esparza et al, 2008)). Very often, the models require training data, which is expensive, and sometimes infeasible to obtain. The availability and improvement of depression lexica is crucial to advance in the development of unsupervised methods that help to identify at-risk individuals.

Neuman et al (2012) developed Pedesis, a system for proactive screening for depression through Text Analysis. Pedesis built a depression lexicon by harvesting the web for methaphorical relations in which terms related to depression are embedded. Choudhury et al (2013) created another depression lexicon using a labeled collection of Twitter posts. This lexicon is a set of unigrams associated to four main dimensions related to depression (namely, symptoms, disclosure, treatment and relationships/life). The authors' approach was based on selecting words that appear with high frequency in the depression

class. This kind of language resources is expensive to obtain because they require training data or some sort of human supervision. We have used these lexica as a core resources in our study. Our experiments help to assess the effectiveness of these lexica at searching for signs of depression. By testing these language resources against recent collections on depression and language use, we have been able to identify its strongest components (i.e. adjectives) and, furthermore, we designed innovative lexicon expansion strategies that led to improved lexica.

Our analysis, improvement strategy and evaluation methodology can be tested in the future with other language resources. For example, Schwartz et al (2014) analyzed depression through Facebook and provided a shortlist of words, phrases and topics associated with depression. Brandt and Boucher (1986) studied depression-type words in eight cultures. The main aim ot the paper was to analyse cultural differences in the manifestation of mental disorder. One of the outcomes of the study was a set of clusters of depression-type words that emerged from different cultural or language groups. Cheng et al (2016) developed a depression lexicon for supporting screening of depression in mobile applications. Words related to depression and its symptoms were gathered from clinical manuals, international classifications of mental disorders, focus discussion groups and interviews with mental health professionals. This approach, which required extensive manual intervention, led to a wide lexicon. At the moment, this valuable language resource is not publicly available.

Our paper is also related to other studies that built and evaluated lexica for specific domains. For example, Abdaoui et al (2017) elaborated and evaluated a lexicon for Sentiment Analysis. Their elaboration method also included an expansion stage, which incorporated synonyms of the English NRC Word Emotion Association Lexicon (NRC-EmoLex). In our study, one of the expansion methods utilized Wordnet. Wordnet has had a profound influence on research on Computational Linguistics and a wide range of applications. A recent survey on Wordnet and relations can be found in (Piasecki et al, 2013).

The test collections that we employed have recently been the main benchmarks in eRisk 2017, an early risk detection challenge (Losada et al, 2017b,a). The challenge, which ran as a lab under the CLEF evaluation campaign, consisted of a pilot task on early risk detection of depression. The participants had to sequentially process the individuals' writings and detect early traces of depression. For each individual, his collection of writings was divided into 10 chunks: the first chunk contains the oldest 10% of the writings, the second chunk contains the second oldest 10%, and so forth. The pilot task consisted of 10 sequential releases of data (at different dates). The first release consisted of the 1st chunk of data (oldest writings of all individuals), the second release consisted of the 2nd chunk of data (second oldest writings of all individuals), and so forth. After each release, the participants had one week to process the data and, before the next release, each participating system could emit a decision about the individual (depressed or non-depressed) or opt to make no decision (i.e. wait to see more chunks). The performance measure combined the effectiveness of the decisions with a penalty related to the delay in emit-

ting decisions. The performance results of the participating teams (e.g. best classification performance –F1– was 64%) showed that there is a need to further improve the language resources to screen for depression. We worked with these collections but we ignored the chronology of the writings and considered each sequence of writings (by a given individual) as a single unit of text. In the future, the lessons learned with our evaluation of lexica could be applied to support algorithms, such as those proposed in eRisk 2017, that iteratively process chunks of writings. Such application of our results is promising because all eRisk 2017 teams implemented some sort of supervised learning technology. The existence of effective lexica to extract signs of depressions could lead to the development of unsupervised solutions that require no training stage and are applicable to a wider range of domains and types of texts. Furthermore, among the systems evaluated in eRisk 2017, the algorithms that combined search with learning were highly effective (Almeida et al, 2017b). The need of search-based components for detecting signs of depression suggests that depression lexica can play a role in the future of these search technologies. Additionally, learning-based components could be enhanced by incorporating features based on depression lexica. These intriguing topics will be further explored in future research.

Some of the techniques we used for estimating signs of depression or for enhancing lexica were based on distributional models. The existing distributional methods for estimating word similarity differ in, at least, the way the word space model is built. A number of alternatives have been proposed in the literature, including counting explicit contexts, neural-based or predicted embeddings. There is some controversy about the performance of different types of word space models when they are applied on specific NLP tasks (Gamallo, 2017). Some authors claim that neural embeddings outperform traditional count-based models to compute word similarity (Baroni et al, 2014; Mikolov et al, 2013). Other researchers argue that there are no significant differences between them (Lebret and Collobert, 2015; Levy and Goldberg, 2014b; Levy et al, 2015), and claim that both embeddings and explicit models have actually succeeded in capturing word similarities. Some studies report heterogeneous results , where the relative effectiveness of the two models varies with the task performed (Blacoe and Lapata, 2012; Huang et al, 2012; Gamallo, 2017). In the distributional literature, little attention has been paid to context filtering within count-based and transparent approaches. Most traditional approaches mainly focused on converting sparse matrices into dense ones by dimensionality reduction (Landauer and Dumais, 1997). However, some works attempted to reduce the raw matrix following simple filtering strategies that select the most salient contexts for each word (Bordag, 2008; Gamallo and Bordag, 2011; Biemann et al, 2013; Padró et al, 2014). In the experiments reported in Section 3, we employed distributional approaches based on the most salient contexts and standard embeddings. None of them outperformed the baseline strategy based on just counting matches of PoS tagged lemmas.

Studying the role of Part-of-Speech (PoS) in text-based depression analysis is another contribution of our research. To the best of our knowledge, this is the

first study that analyses the effect of different PoS components for depression screening. PoS tags have been used in a number of text analytics tasks. For example, PoS-based evidence is considered a good indicator for sentiment and affective language and has been employed to classify opinions (Chenlo and Losada, 2014). Among different PoS tags, special attention has been paid to adjectives. Turney showed that adjectives are important indicators of opinions (Turney, 2002) and, in Benamara et al (2007), Parts-of-Speech taggers are used to select adjectives followed or preceded by adverbs because that combination of PoS tags is considered to be polarity-sensitive.

3 Evaluating Depression Lexica

In this section, we propose a search-based method for evaluating depression lexica. The main idea is to exploit existing resources for depression and language use (Losada and Crestani, 2016). Given a corpus containing texts written by depressed and non-depressed individuals, we define three search strategies that work with the lexicon and the corpus to produce a ranking of individuals in decreasing order of estimated level of depression. Such a ranking may be highly convenient, for example, to public health departments seeking automatic means to screen for signs of depression. Unlike automatic classifiers, ranking tools do not set pre-defined thresholds on their output. This gives users freedom to inspect as many ranked elements as they wish. Depending on the user's task and the available resources, users might want to see a few top ranked items (high precision task) or a larger subset of the corpus (high recall task). Lexica will therefore be evaluated in terms of how well they support this search for signs of depression, and the ranked lists will be evaluated with precision-oriented and recall-oriented effectiveness metrics. Such variety of measures allows to analyze lexica under different possible use cases. For example, a narrow and detailed lexicon might be highly effective at placing depression cases at the top positions of the ranking. However, such precision-oriented lexicon might miss many depression cases and, thus, it would not be a solid tool when recall is a must.

We propose three evaluation methods that follow different strategies in representing the contents of the lexicon and estimating signs of depression. All these methods rely on a domain-specific lexicon consisting of terms identified by experts. In our case, the terms of the lexicon are indicative of a mental disorder, namely depression. All terms were lemmatized and PoS tagged. The same lemma can be associated to multiple PoS tags. Given the lexicon and a document, which contains all text written by any given individual, a score of estimated level of depression is produced as follows.

3.1 Lemma-PoS approach

A basic strategy consists of counting the number of occurrences within the document of lemmas-PoS from the depression lexicon and then dividing the

count by the total number of lemmas in the document. Given a document d, the Lemma-PoS depression score, $ScoreDepr_{lp}$, is computed as:

$$ScoreDepr_{lp}(d) = \frac{\sum_{(t_i,PoS_i)\in Lex} freq((t_i, PoS_i), d)}{len_d} \quad (1)$$

where (t_i, PoS_i) is the *i-th* entry in the lexicon (t_i is a lemmatized term and PoS_i is its PoS tag in the lexicon), $freq((t_i, PoS_i), d)$ stores the number of occurrences of the lemma-PoS in the document, and len_d represents the total number of lemma-PoS in the document. Note that $ScoreDepr_{lp}(d) \in [0, 1]$ (equals 0 when no depression entry occurs in the document, and equals 1 when all the lemma-PoS of the document are elements of the depression lexicon).

3.2 Word Embedding-based approach

Terms in a particular corpus are known to follow Zipf's law: there is a small vocabulary of common words and a large vocabulary of individually rarer words. Such imbalance makes counting-based methods biased towards frequent words. Furthermore, exact matching between documents and lexicon ignores fundamental associations between words. For example, there are many ways to refer to the same concept and, additionally, many words have more than one meaning. Traditional bag of words (or Bag of Lemma-PoS, such as the method sketched above) represent words as unique entities with no association between them. This makes that *sad* is as distinct from *fast* as it is from *pessimistic*. Recent Text Mining models, instead, employ *distributed* representations of words. Every word is represented by a vector that captures contextual and semantic information. These vectors are commonly referred to as embeddings and are often learned using neural-network models over large corpora. It is possible to obtain a distributed representation of the depression lexicon. The key advantage is to enable inexact matching between lexicon and documents in the embedding space.

A document can be represented by the sum or average of the vectors (embeddings) corresponding to the document 's terms (Mitra and Craswell, 2017). Following a similar approach, a lexicon can be represented by the sum or average of the embeddings of the words in the lexicon. Given a depression lexicon, we compute the average vector (AV) by adding the embeddings associated with all terms of the lexicon divided by the size of the lexicon. This vector represents the semantics of the depression lexicon. Next, the $ScoreDepr_{emb}$ of a document d is computed as the (Cosine) similarity between the two vectors:

$$ScoreDepr_{emb}(d) = Cosine(AV(Lex), AV(d)) \quad (2)$$

where $AV(Lex)$ and $AV(d)$ are the AV of the lexicon and document, respectively.

3.3 Explicit Context-Based Approach

Neural-based embeddings is a type of distributional semantic models that allow scalable and unsupervised training of dense vectors from very large corpora. Although embeddings have received increasing attention, there are also good reasons to consider alternative formulations of distributional models. For example, some alternative models keep sparse and explicit representations and, thus, lead to interpretable solutions (Biemann, 2016; Gamallo, 2017). Explicit distributional representations identify and select the most relevant contexts of a given word. For example, given the distributional information extracted from Wikipedia[2], the three most relevant syntactic contexts of the word *"despair"* are the following:

(NOUN, at, prospect)	e.g. *in despair at prospect of approaching sorrow*
(cry, of, NOUN)	e.g. *a cry of despair*
(shake, in, NOUN)	e.g. *her death made me shake in despair*

Under these models, words are represented by their p most salient lexico-syntactic contexts. Saliency is measured by frequency or by a statistical measure that prefers frequent co-occurrence (e.g. point-wise mutual information or log-likelihood). This approach can be seen as a filtering strategy to deal with explicit distributional models (Gamallo and Bordag, 2011; Gamallo, 2017; Biemann, 2016).

Each word w has a set of salient contexts:

$$SC(w) = \{(sc_1, w_1), \ldots, (sc_p, w_p)\} \tag{3}$$

where sc_i is an identifier of i-th salient context and w_i is the weight of sc_i for word w.

Explicit and salient contexts not only may be used to represent words, but also lists of words. For example, an entire lexicon or the bag-of-words of a document can be effectively represented using contexts. In order to get the salient contexts of the depression lexicon, we just need to add the weights of the salient contexts of the words in the lexicon:

$$SC(Lex) = \{(sc_1, w_1), \ldots\} \tag{4}$$

$$\forall j : w_j = \sum_{w \in Lex} \sum_{(sc_j, w_i) \in SC(w)} w_i \tag{5}$$

and the most salient contexts of the lexicon, $MSC(Lex)$, are obtained by extracting the top p contexts (p pairs of $SC(Lex)$ with the highest weights).

A similar method is employed to extract the most salient contexts of the words in a document, $MSC(d)$. For example, a document written by an individual suffering from depression might contain salient contexts related to different warning signs of depression (e.g., feeling down, staying asleep).

[2] `https://www.wikipedia.org/`

In a similar spirit to the embedding case, the final $ScoreDepr_{expl}$ of a document d given a lexicon Lex is computed as a normalized sum of the matching contexts (contexts that belong to both $MSC(d)$ and $MSC(Lex)$):

$$ScoreDepr_{expl}(d) = \frac{\sum_{(sc_j,w_l)\in MSC(Lex),(sc_j,w_d)\in MSC(d)} w_l \cdot w_d}{\sqrt{\sum_{(sc_l,w_l)\in MSC(Lex)} w_l^2} \cdot \sqrt{\sum_{(sc_d,w_d)\in MSC(d)} w_d^2}} \quad (6)$$

3.4 Experiments

Our first set of experiments tested two lexica following the search-based methods sketched above. To meet this aim, we obtained the *Pedesis* lexicon, produced by Neuman et al (2012), and the depression lexicon created by Choudhury et al (2013).

3.4.1 Lexica

Pedesis is a system for building depression lexica that harvests the web for metaphorical relations in which depression is embedded and extracts relevant concepts related to depression. The original Pedesis lexicon is available from the authors upon request. We ran experiments with the original lexicon, which contains some multiword entries, and also experimented with the following subsets of the lexicon: unigrams only, adjectives only, verbs only and nouns only.

A second language resource that we evaluated was the depression lexicon available in (Choudhury et al, 2013). This lexicon, in the following De Choudhury et al. lexicon, is a set of unigrams associated to four main dimensions related to depression (namely, symptoms, disclosure, treatment and relationships/life). The lexicon was created by De Choudhury and colleagues by selecting words that appear with high frequency in the depression class of a training collection consisting of Twitter posts. With this lexicon we also produced four (sub)lexica (unigrams only, adjectives only, verbs only and nouns only) according to the PoS tags of the words.

We got the lemmas-PoS of each word using Linguakit's lemmatizer (Gamallo and Garcia, 2017). Ambiguous terms, which were assigned several PoS tags, were added to more than one lexicon. The PoS of each word can be crucial. For example, in Sentiment Analysis, it has been shown that adjectives are indicators of opinions (Liu, 2012). Thus, specific lexica have been created and analyzed based on PoS (e.g. sentiment adjectives only) (Devitt and Ahmad, 2013). In Text Analytics for depression, there is a lack of studies that specifically analyze what PoS tags are more effective at identifying signs of depression. In this regard, we hope that our experiments help to shed light on how depressive individuals express their feelings (for example, "are adjectives or nouns more important than verbs?").

3.4.2 Document corpus

The document corpus utilized in our experiments is the test collection of depression and language use built by Losada and Crestani (2016). It is a collection of writings (posts or comments) from a set of Social Media users. The collection was obtained from Reddit. Losada and Crestani (2016) studied the adequacy of different data sources, including Twitter, MTV's A Thin Line and Reddit, to create a collection for research on depression and language use. The main dimensions analyzed were: the quality and size of each source, the availability of a long history of user submissions, the difficulty to distinguish depressed and non-depressed users, and the terms and conditions of the data sources. The authors concluded adopting Reddit. Reddit is an open-source social network that has a large community of active users. For each user, the available set of submissions is typically large (covering several years) and Reddit has group of users devoted to talk about different medical conditions, such as anorexia or depression. The resulting corpus is unique in a number of ways. First, it is publicly available[3] while most previous studies worked with data that cannot be shared. For example, some researchers focused on tweets written by depressed users (Choudhury et al, 2013), but their experiments cannot be reproduced. Second, the sizes of the classes (depressed vs non-depressed) obtained by Losada and Crestani (2016) are comparable to those used by previous studies, but the collection has a much richer user representation (on average, more than 500 submissions per user, and user submissions cover a wide range of dates). Losada and Crestani (2016) obtained two classes of individuals, depressed and non-depressed, following the method proposed by Coppersmith et al (2014). The extraction of the depression class consisted of first identifying self-expressions of depression diagnoses (e.g., "Yesterday, I was diagnosed with depression") and, next, doing a manual review of the matched submissions (to verify that these expressions of diagnosis look really genuine). The selection of users for the non-depression class consisted of random sampling from Reddit. This approach to identify these two classes of users has proved to be effective in several past studies (Coppersmith et al, 2014; Losada and Crestani, 2016; Losada et al, 2017b). After assigning the users' classes, all available submissions from each user was retrieved from Reddit (up to 2000 submissions, including post and comments submitted to any Reddit community). As a consequence, the average user is represented in the corpus with a large sequence of submissions. This collection has become a standard for evaluation of early risk technologies and it has been used in well-known evaluation campaigns (Losada et al, 2017b). The collection contains two splits, which will

[3] https://tec.citius.usc.es/ir/code/dc.html

	DLU16A	DLU16B
Num. individuals (depressed/non depr.)	83/403	54/352
Num. writings	295,103	236,479
Avg num. of writings per subject	607.2	582.5

Table 1 Main statistics of the datasets.

be referred here to as DLU16A and DLU16B[4]. The main statistics of these datasets are reported in Table 1.

3.4.3 Search task

The identification of individuals in risk of depression can be approached as a document search task where each individual is associated to a document, which contains all his writings. Each writing is 32.1 words on average and each individual produced an average of about 600 writings. This leads to an extensive representation of each individual (about 19,000 words per individual). Ranking of individuals was done with the three search methods sketched above and we considered the following effectiveness measures: Average Precision (AP), Normalized Discount Cumulative Gain (NDCG), R-Precision, P@5 and P@10. These five metrics are commonly employed in retrieval experiments. The challenge consists of searching for users in risk of depression and, thus, P@k is here the fraction of the top k individuals that are depressed. R-Precision is the proportion of the top-R retrieved individuals that are depressed, where R is the number of depressed individuals in the collection. This means that R-Precision is the precision at the position of the ranking where a perfect system would have already identified all depressed individuals. AP summarizes the ranking of individuals by averaging the precision values from the rank positions where a depressed individual was retrieved. NDCG goes a step further and introduces an increasingly high discount for depressed individuals located at low rank positions. More specifically, NDCG counts the *cumulative gain* from traversing the ranked list. This gain represents how much total gain the user has if he/she examines all individuals. The contribution of gain of the individuals in the ranking is weighted by their position. This captures the intuition that a lowly ranked depressed individual does not contribute as much gain as a highly ranked depressed individual. The top 1 individual is not discounted because it is assumed that the user always sees this individual. The gain of the rest of individuals is discounted by dividing by a logarithm of its position in the list. Formally, these measures are defined as follows:

$$P@k = \frac{\sum_{i=1}^{k} dep_i}{k} \tag{7}$$

[4] The collection was divided into two halves because the early risk challenge proposed in (Losada et al, 2017b) promoted the development of supervised learning solutions. DLU16A was the training split and DLU16B was the test split. We are concerned here with unsupervised (search-based) methods and, therefore, we use these two splits as independent test corpora.

$$R\text{-}Precision = P@R \tag{8}$$

$$AP = \frac{\sum_{i \in Pos\_Dep} P@i}{R} \tag{9}$$

$$NDCG = \frac{dep_1 + \sum_{i=2}^{n} \frac{dep_i}{\log_2 i}}{IDCG} \tag{10}$$

where dep_i represents the depression label of the individual at position i ($dep_i = 1$ for depressed and $dep_i = 0$ for non-depressed), Pos_Dep is the set of positions of the depressed individuals in the ranking, n is the size of the ranking, R is the overall number of depressed individuals in the collection, and $IDCG$ is a normalization factor that divides the discounted cumulative gain (numerator) by the ideal gain achieved by a perfect ranking (in this way, NDCG ranges in $[0, 1]$ and equals 1 for an ideal ranked list).

These five metrics give a complete picture of search performance. For example, P@k should be preferred for high precision applications (i.e. avoid false positives), while the other three measures are more oriented to recall (find all depressed individuals). All metrics range in $[0, 1]$ (1 means perfect performance, while 0 means that no depressed individual was found). For example, P@5 equals 1 when the top 5 individuals in the ranking are depressed and equals 0 when the top 5 individuals are non-depressed. More details about these performance measures can be found elsewhere (Baeza-Yates and Ribeiro-Neto, 2011).

Our evaluation design included experiments with two baselines: i) a *random* baseline, which consists of ranking the available subjects in a random way, and ii) a query-based baseline. The first method is naïve –it does not make any attempt to search for signs of depression– and we expect it to perform poorly. In any case, its performance is a good reference to understand the effect of the competing methods. The second baseline is a method that searches for signs of depression using a short query (extracted from a previous study on personality disorders). More specifically, Neuman et al (2015) conducted a study on automatic text analysis of school shooters. The authors' approach was based on vectorial semantics and measured the similarity between texts –written by school shooters– and word vectors representing four personality disorder traits. The vector associated to depression contains four words (sad, lonely, hopeless and worthless) that were identified by the authors based on diagnostic manuals of mental disorders and existing methods for personality assessment. We used these four words to search for signs of depression in our corpora. We expect this baseline to be more effective than the random baseline. Furthermore, it is useful to compare the performance of the lexicon-based methods against the performance of this method, which searches using a few selected terms. The two baselines are referred to as *random* and *depression query* and their performance figures are shown in the upper block of Tables 2, 3, 4 and 5.

In the experiments, the AVs were created from the embeddings described by Levy and Goldberg (2014a). These embeddings were generated using `Word2vec`[5] and are publicly available[6]. The MSCs were created from the most salient contexts described by Gamallo (2017), which also are publicly available[7]. The two distributional models, based on embeddings or explicit contexts, were learnt from the same resource, namely English Wikipedia (August 2013 dump[8]).

3.4.4 Experimental results

Table 2 and 3 present the results of these experiments. A first conclusion that we can draw from these initial experiments is that the lemma-PoS approach is the best performing search strategy. Most cases of lexica-collection get their highest effectiveness when search is based on lemma-PoS matching. This suggests that sophisticated matching based on embeddings or explicit contexts does not provide added value here. Embeddings led to improved performance in a couple of cases. For example, in DLU16A, embeddings applied on De Choudhury et al. lexicon verbs or nouns, got to P@5 equal to 1 (the top 5 ranked subjects are depression cases). However, the improvements in performance from embeddings are not consistent across lexica and collections. This outcome might be due to the quality and richness of the original lexica. Pedesis contains several hundred entries (see Table 6, upper block) in all sublexica. De Choudhury et al. lexicon is smaller (see Table 6, third block) but still does not benefit from a treatment based on embeddings or contexts. Embeddings and explicit contexts implement different forms of inexact matching and, according to our results, such an approach harms performance. We will therefore adopt the lemma-PoS as our reference method for making the most of the lexica.

The Pedesis lexicon and its sublexica seem more consistent than the De Choudhury et al. sublexica. This makes sense because Pedesis contains many more entries and, thus, it improves recall of depression cases. De Choudhury et al. sublexica, instead, are much narrower and only work well for some high precision metrics (e.g. DLU16A, P@5/P@10).

By analyzing the lemma-PoS results, we can observe that most lexica perform relatively well at placing depressed individuals at the top positions of the ranking. The results of P@5 show that the percentage of depressed individuals in the top 5 positions tends to be around 50% and the results of NDCG are around 70%. NDCG strongly favors systems that put depressed individuals at high ranking positions, and the high NDCG figures suggest that the lexica are able to populate the top positions with depressed individuals. AP and R-Prec performance is lower, suggesting that some depressed individuals are placed at low rank positions (recall is weak).

With Pedesis, the original lexicon seems slightly inferior to the lexicon composed of adjectives only. In DLU16A, the performance of both lexica is roughly

[5] `https://code.google.com/archive/p/word2vec/`

[6] `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`

[7] `http://fegalaz.usc.es/~gamallo/resources/count-models.tar.gz`

[8] `http://dumps.wikimedia.org/enwiktionary`

	DLU16A				
	AP	NDCG	R-prec	P@5	P@10
			baselines		
random	.177	.636	.181	.200	.100
depression query	.335	.769	.349	.400	.500
Pedesis lexicon (w. multiwords)	.380	.758	.470	.400	.400
			lemma-PoS		
Pedesis lexicon (unigrams)	**.393**	**.780**	**.434**	**.600**	.400
Pedesis lexicon (adj)	.390	.773	**.434**	.200	.400
Pedesis lexicon (vrb)	.306	.744	.301	**.600**	.500
Pedesis lexicon (noun)	.256	.701	.253	.200	.300
			word embeddings		
Pedesis lexicon (unigrams)	.232	.721	.193	**.600**	**.600**
Pedesis lexicon (adj)	.193	.655	.205	.200	.200
Pedesis lexicon (vrb)	.198	.684	.157	.200	.400
Pedesis lexicon (noun)	.242	.729	.253	.400	.500
			explicit contexts		
Pedesis lexicon (unigrams)	.205	.652	.217	.200	.100
Pedesis lexicon (adj)	.218	.703	.217	.400	.400
Pedesis lexicon (vrb)	.185	.635	.217	.000	.000
Pedesis lexicon (noun)	.216	.656	.229	.000	.100
			lemma-PoS		
De Choudhury et al. lexicon	**.346**	**.755**	.361	.600	**.700**
De Choudhury et al. lexicon (adj)	.307	.715	.325	.200	.300
De Choudhury et al. lexicon (vrb)	.302	.732	.325	.400	.300
De Choudhury et al. lexicon (noun)	.321	.728	**.398**	.400	.400
			word embeddings		
De Choudhury et al. lexicon	.246	.736	.253	.600	.500
De Choudhury et al. lexicon (adj)	.251	.720	.229	.800	.600
De Choudhury et al. lexicon (vrb)	.240	.734	.181	**1.00**	.600
De Choudhury et al. lexicon (noun)	.330	.791	.325	**1.00**	**.700**
			explicit contexts		
De Choudhury et al. lexicon	.165	.613	.145	.000	.000
De Choudhury et al. lexicon (adj)	.278	.735	.229	.800	**.700**
De Choudhury et al. lexicon (vrb)	.174	.641	.181	.400	.300
De Choudhury et al. lexicon (noun)	.207	.694	.205	.400	.200

Table 2 Evaluation of two lexica lexica and three subsets of the lexica (adjectives only, verbs only and nouns only) for the DLU16A collection. For each lexicon, the best results are bolded.

the same, but the adjectives-only lexicon is clearly superior in DLU16B[9]. With the De Choudhury et al. lexicon, the adjectives-only sublexicon is inferior to both the original lexicon and the nouns-only lexicon. This might be due to the low number of adjectives in this lexicon (it has 19 adjectives, 85 nouns and the total number of unigrams is 149; see Table 6, third block). However, observe that the few adjectives available still lead to competitive performance. These initial experiments suggest that adjectives are important indicators of signs of depression. Adjectives appear to be valuable indicators of expressions related to depression. The adjectives-only lexica can therefore become useful tools to guide the identification of signs of depression.

[9] Observe that these experiments involve a single search for depressed individuals and, thus, we cannot perform tests of statistical significance over the differences found.

	DLU16B				
	AP	NDCG	R-prec	P@5	P@10
			baselines		
random	.169	.616	.167	.400	.300
depression query	.261	.682	.296	.200	.400
Pedesis lexicon (w. multiwords)	.254	.671	.259	.400	.300
			lemma-PoS		
Pedesis lexicon (unigrams)	.237	.661	.241	.400	.300
Pedesis lexicon (adj)	**.269**	**.694**	**.278**	**.600**	**.400**
Pedesis lexicon (vrb)	.207	.649	.222	.400	.300
Pedesis lexicon (noun)	.186	.615	.204	.000	.200
			word embeddings		
Pedesis lexicon (unigrams)	.129	.557	.111	.000	.000
Pedesis lexicon (adj)	.148	.583	.167	.000	.100
Pedesis lexicon (vrb)	.128	.552	.111	.000	.000
Pedesis lexicon (noun)	.129	.558	.111	.000	.000
			explicit contexts		
Pedesis lexicon	.144	.570	.204	.000	.000
Pedesis lexicon (unigrams) (adj)	.131	.563	.093	.200	.100
Pedesis lexicon (vrb)	.165	.608	.204	.200	.100
Pedesis lexicon (noun)	.156	.580	.204	.000	.000
			lemma-PoS		
De Choudhury et al. lexicon	**.260**	**.675**	.278	**.200**	**.300**
De Choudhury et al. lexicon (adj)	.256	.672	**.315**	**.200**	.200
De Choudhury et al. lexicon (vrb)	.210	.635	.241	.200	**.200**
De Choudhury et al. lexicon (noun)	.261	**.675**	.278	**.200**	.200
			word embeddings		
De Choudhury et al. lexicon	.129	.557	.111	.000	.000
De Choudhury et al. lexicon (adj)	.148	.583	.167	.000	.100
De Choudhury et al. lexicon (vrb)	.128	.552	.111	.000	.000
De Choudhury et al. lexicon (noun)	.129	.558	.111	.000	.000
			explicit contexts		
De Choudhury et al. lexicon	.117	.546	.074	.000	.100
De Choudhury et al. lexicon (adj)	.174	.614	.222	**.200**	**.300**
De Choudhury et al. lexicon (vrb)	.115	.550	.093	**.200**	.000
De Choudhury et al. lexicon (noun)	.203	.550	.185	**.200**	**.300**

Table 3 Evaluation of the two lexica and three subsets of the lexica (adjectives only, verbs only and nouns only) for the DLU16B collection. For each lexicon, the best results are bolded.

Observe also that search performance is lower in DLU16B than in DLU16A. This is a natural consequence of DLU16B having a lower percentage of depressed individuals (see Table 1, DLU16A has 20% depressed individuals, while DLU16B has 15% depressed individuals). The difficulty of DLU16B was already shown in Losada and Crestani (2016).

4 Lexicon Enhancement

In an attempt to reduce the cost of manual annotation, some researchers have explored automatic methods of expanding and re-building lexica. Lexicon expansion can be performed following two main strategies: corpus-based

or thesaurus-based. Corpus-based expansion uses distributional similarity induced from distributional semantic models (e.g., embeddings or explicit contexts) that are learnt from large corpora (Wang and Xia, 2017). Thesaurus-based methods require lexical resources such as WordNet (Fellbaum, 1998), which make use of lexical relationships such as synonymy (synsets) to build domain-specific resources, e.g., sentiment lexicon (Baccianella et al, 2010).

In this section, we describe our endeavors to improve the depression lexica following corpus-based (or distributional-based) and thesaurus-based strategies. In the corpus-based approach, the lexicon is expanded with new terms that are selected among the most similar words associated to each term of the lexicon. To meet this aim, distributional similarity is computed following the explicit-based semantic model that relies on the most salient contexts extracted from Wikipedia (as described in subsection 3.3). This expansion approach will be referred to as distributional-based expansion (DE expansion). For example, the adjective *desperate* has the following 10 most salient lexico-syntactic contexts: {*(desperate attempt), (desperate need), (desperate gamble), (desperate strait), (desperate to find), (desperate to escape), (desperate gambit), (desperate plea), (desperate to get), (desperate plight)*}. Other words have similar contexts and, in this case, the most similar terms are: *insupportable*, *unaccompanied* and *heartbreaking*. These three terms are included into the expanded lexica. In the thesaurus-based approach, new terms are selected from the synsets associated to each lemma of the original lexicon. In the case of desperate, the Wordnet-based expansion (WE expansion) led to the following new entries: *despairing*, *dire*, *heroic*, and *do-or-die*.

Lexicon expansion can introduce valuable terms in the lexicon, but it often introduces noise that may degrade performance. One of the main causes of noise is word polysemy. Many words of the depression lexicon have senses that are far away from the psychological domain. For instance, the noun *depression* is in the original lexicon because it can refer to a mental condition. However, it can also mean a time with very little economic activity, or a mass of air that has low pressure. By automatically expanding *depression* we can wrongly generate new words associated with the senses that are not related to depression as a mental health condition. A simple solution to this problem consists of only expanding non-ambiguous words, that is, words that have only one synset in WordNet. In the experiments described below, we built several lists of depression terms. More specifically, we considered the four combinations of thesaurus-based expansion/distributional-based expansion on all lemmas/non-ambiguous lemmas.

4.1 Experiments

All these experiments were performed with search based on lemma-PoS matching. As argued above, this is the most effective way to take advantage of the depression lexica. In the corpus-based approach, each term was expanded with the three most similar terms. In the thesaurus-based approach, each term was

	DLU16A				
	AP	NDCG	R-prec	P@5	P@10
			baselines		
random	.177	.636	.181	.200	.100
depression query	**.335**	**.769**	**.349**	**.400**	**.500**
Pedesis lexicon (w. multiwords)	.380	.758	.470	.400	.400
Pedesis lexicon (adj)	.390	.773	.434	.200	.400
Pedesis lexicon (adj)+WE	.397	.802	.398	.600	.600
Pedesis lexicon (adj)+DE	.391	.788	.434	.400	.300
Pedesis lexicon (non ambiguous adj)+WE	.435	.809	.410	**.800**	.700
Pedesis lexicon (non ambiguous adj)+DE	**.496**	**.835**	**.518**	**.800**	**.900**
Pedesis lexicon (vrb)	.306	**.744**	.301	**.600**	**.500**
Pedesis lexicon (vrb)+WE	.296	.731	.277	.400	.400
Pedesis lexicon (vrb)+DE	**.339**	.742	**.398**	.400	.400
Pedesis lexicon (non ambiguous vrb)+WE	.245	.686	.277	.200	.100
Pedesis lexicon (non ambiguous vrb)+DE	.312	.714	.386	.200	.200
Pedesis lexicon (noun)	.256	.701	.253	.200	**.300**
Pedesis lexicon (noun)+WE	.288	.707	**.325**	.200	**.300**
Pedesis lexicon (noun)+DE	**.292**	**.739**	.313	**.400**	**.300**
Pedesis lexicon (non ambiguous noun)+WE	.226	.677	.265	.200	.200
Pedesis lexicon (non ambiguous noun)+DE	.253	.686	.289	.200	.200
De Choudhury et al. (adj)	.307	.715	.325	.200	.300
De Choudhury et al. (adj)+WE	.323	.752	.361	.200	.200
De Choudhury et al. (adj)+DE	.307	.715	.325	.200	.300
De Choudhury et al. (non ambiguous adj)+WE	**.455**	**.824**	**.506**	**.400**	**.500**
De Choudhury et al. (non ambiguous adj)+DE	.427	.791	.482	**.400**	**.500**
De Choudhury et al. (vrb)	.302	.732	**.325**	.400	.300
De Choudhury et al. (vrb)+WE	.245	.686	.277	.400	.300
De Choudhury et al. (vrb)+DE	.334	.756	.373	.400	.400
De Choudhury et al. (non ambiguous vrb)+WE	.232	.681	.229	.200	.300
De Choudhury et al. (non ambiguous vrb)+DE	**.338**	**.764**	.301	**.600**	**.600**
De Choudhury et al. (noun)	.321	.728	.398	.400	.400
De Choudhury et al. (noun)+WE	.338	.733	.386	.000	.400
De Choudhury et al. (noun)+DE	**.420**	**.788**	.422	.400	**.500**
De Choudhury et al. (non ambiguous noun)+WE	.269	.695	.313	.200	.200
De Choudhury et al. (non ambiguous noun)+DE	.398	.784	**.458**	**.600**	.400

Table 4 DLU16A collection. Effect of WordNet-based expansion (WE) and Distributional-based expansion (DE) on the lexica. For each block, the best performance is bolded.

expanded with all terms from its Wordnet synsets. In the case of expansion of non ambiguous terms the new terms come from a single synset.

Tables 4 and 5 report the results of these experiments. In general, the three sublexica tested here (adjectives, verbs and nouns) are improved after expansion. Most configurations led to performance figures that are higher that those obtained with the original (non expanded) lexica. This provides evidence to support the claim that these forms of lexicon expansion are effective as a means to improve the original lexica.

Let us now analyze the relative merits of WE and DE with ambiguous and non-ambiguous expansion. With verbs and nouns, the results are a mixed bag. In some cases, focusing expansion on non-ambiguous terms works well but, in other cases, it does not give any added value. Similarly, the relative merits of WE and DE with verbs and nouns does not show a clear pattern. Summing

	DLU16B				
	AP	NDCG	R-prec	P@5	P@10
			baselines		
random	.169	.616	.167	**.400**	.300
depression query	**.261**	**.682**	**.296**	.200	**.400**
Pedesis lexicon (w. multiwords)	.254	.671	.259	.400	.300
Pedesis lexicon (adj)	.269	.694	.278	.600	.400
Pedesis lexicon (adj)+WE	.290	.720	.296	**.800**	.500
Pedesis lexicon (adj)+DE	.270	.680	.315	.400	.300
Pedesis lexicon (non ambiguous adj)+WE	.298	.711	.296	.600	.500
Pedesis lexicon (non ambiguous adj)+DE	**.329**	**.743**	**.352**	**.800**	**.600**
Pedesis lexicon (vrb)	.207	.649	.222	.400	.300
Pedesis lexicon (vrb)+WE	.208	.647	.167	.200	.300
Pedesis lexicon (vrb)+DE	.263	.680	.278	**.600**	.300
Pedesis lexicon (non ambiguous vrb)+WE	.201	.640	.204	.200	.200
Pedesis lexicon (non ambiguous vrb)+DE	**.280**	**.707**	**.296**	.400	**.400**
Pedesis lexicon (noun)	.186	.615	.204	.000	**.200**
Pedesis lexicon (noun)+WE	**.203**	**.631**	**.222**	**.200**	**.200**
Pedesis lexicon (noun)+DE	.171	.606	.185	**.200**	**.200**
Pedesis lexicon (non ambiguous noun)+WE	.151	.583	.185	.000	.100
Pedesis lexicon (non ambiguous noun)+DE	.191	.612	**.222**	.000	.000
De Choudhury et al. (adj)	.256	.672	.315	.200	**.200**
De Choudhury et al. (adj)+WE	.216	.632	.222	.000	**.200**
De Choudhury et al. (adj)+DE	.256	.671	.315	.200	**.200**
De Choudhury et al. (non ambiguous adj)+WE	.300	.729	**.333**	**.400**	**.200**
De Choudhury et al. (non ambiguous adj)+DE	**.310**	**.734**	**.333**	**.400**	**.200**
De Choudhury et al. (vrb)	.210	.635	.241	.200	.200
De Choudhury et al. (vrb)+WE	.190	.615	.185	.200	.200
De Choudhury et al. (vrb)+DE	.237	.649	**.296**	.200	.100
De Choudhury et al. (non ambiguous vrb)+WE	.189	.603	.241	.000	.000
De Choudhury et al. (non ambiguous vrb)+DE	**.239**	**.664**	.259	**.400**	**.400**
De Choudhury et al. (noun)	.261	.675	.278	.200	.200
De Choudhury et al. (noun)+WE	.236	.654	.259	.200	.100
De Choudhury et al. (noun)+DE	**.339**	**.722**	**.389**	**.400**	**.500**
De Choudhury et al. (non ambiguous noun)+WE	.193	.613	.204	.000	.200
De Choudhury et al. (non ambiguous noun)+DE	.260	.708	.278	**.400**	.200

Table 5 DLU16B collection. Effect of WordNet-based expansion (WE) and Distributional-based expansion (DE) on the lexica. For each block, the best performance is bolded.

up, it is good to expand nouns and verbs but these experiments do not provide a clear recommendation about which expansion technique should be chosen.

Let us now focus on expansion of adjectives. As argued above, adjectives seem to be an effective component of the depression lexica and it is important to see the effect of the expansion strategies on adjectives. The results reveal that the adjective-only lexica are the most effective and, furthermore, the expansion experiments show that the adjectives-only lexica can be further improved with expansion. The expansion of non-ambiguous adjectives with DE looks more consistent than the expansion of non-ambiguous adjectives with WE. In any case, the resulting lexica –non ambiguous adj+WE and non ambiguous adj+DE– tend to be superior to the original lexica, composed of all adjectives, nouns and verbs, and superior to those obtained with the original

	original
Pedesis lexicon (w. multiwords)	1638
Pedesis lexicon (unigrams)	1122
Pedesis lexicon (adj)	204 (153 non-ambiguous + 51 ambiguous)
Pedesis lexicon (vrb)	326 (149 non-ambiguous + 177 ambiguous)
Pedesis lexicon (noun)	397 (101 non-ambiguous + 296 ambiguous)
	expanded
Pedesis lexicon (adj)+WE	761
Pedesis lexicon (adj)+DE	455
Pedesis lexicon (non ambiguous adj)+WE	576
Pedesis lexicon (non ambiguous adj)+DE	312
Pedesis lexicon (vrb)+WE	1220
Pedesis lexicon (vrb)+DE	809
Pedesis lexicon (non ambiguous vrb)+WE	613
Pedesis lexicon (non ambiguous vrb)+DE	393
Pedesis lexicon (noun)+WE	1513
Pedesis lexicon (noun)+DE	1052
Pedesis lexicon (non ambiguous noun)+WE	240
Pedesis lexicon (non ambiguous noun)+DE	266
	original
De Choudhury et al. lexicon (unigrams)	146
De Choudhury et al. lexicon (adj)	19 (7 non-ambiguous + 12 ambiguous)
De Choudhury et al. lexicon (vrb)	42 (12 non-ambiguous + 30 ambiguous)
De Choudhury et al. lexicon (noun)	85 (22 non-ambiguous + 63 ambiguous)
	expanded
De Choudhury et al. lexicon (adj)+WE	97
De Choudhury et al. lexicon (adj)+DE	39
De Choudhury et al. lexicon (non ambiguous adj)+WE	16
De Choudhury et al. lexicon (non ambiguous adj)+DE	13
De Choudhury et al. lexicon (vrb)+WE	272
De Choudhury et al. lexicon (vrb)+DE	149
De Choudhury et al. lexicon (non ambiguous vrb)+WE	112
De Choudhury et al. lexicon (non ambiguous vrb)+DE	46
De Choudhury et al. lexicon (noun)+WE	420
De Choudhury et al. lexicon (noun)+DE	274
De Choudhury et al. lexicon (non ambiguous noun)+WE	50
De Choudhury et al. lexicon (non ambiguous noun)+DE	67

Table 6 Main statistics (# words) of the resulting lexica.

(adjectives-only) lexica. This superiority holds for both lexica, Pedesis and De Choudhury et al. lexicon.

Table 6 reports the main statistics of all lexica (original and expanded). These statistics, together with the performance metrics of all lexica, suggest that it is useful to produce lexica with a few dozens of selected words. As a matter of fact, one of the most effective expansion methods (non ambiguous adjectives+DE), leads to lexica whose sizes (312 and 13, respectively) are much smaller than the competing lexica.

An important outcome of these experiments is the lexicon obtained from Pedesis with non ambiguous adjectives and expanded with DE. This lexicon, which consists of 312 terms, is highly effective at identifying individuals in risk of depression. For example, its P@5 and P@10 performance is quite high (about .800). But this lexicon is not merely a high precision device. According to AP,

accelerate adsorb affect alleviate anger ask avoid beat bestow blotched bruise cancel capture carry cause cdot characterise characterize clinch collapse colour confront conquer convert convince cry decline defeat define delay denote depopulate derive destroy detect devastate devote diminish disappear disappoint divide elongate emit encircle enclose encourage enlarge erode evaporate evoke evolve exacerbate exclude exercise extract facilitate fade fill finish flank flatten fleck focus foil forward grab grieve halt hamper hawthorn heal hinder hope impede imply impress induce infuse inject innervate invade ionize isolate kill leach metabolize minimize opt orange-red outflank outrage overhang owe oxidise oxidize pacify peasantry penetrate pertain plan postpone pray prepare present prevent protrude ravage react refer relate remove repel repulse reschedule respond revere reward satisfy schedule seedling seep send separate sharpen shock shower slate soothe speckle stop streak strive subdue subjugate submit surprise surround swell taper tell thwart ting transform traverse treat tremble turn urinate vaporize venerate vine vomit wait wane wield win wish worship yearn

Table 7 New words included in the Pedesis lexicon by the DE expansion method

NDCG and R-precision, the lexicon also acts as an effective mechanism for ranking depressed individuals above non-depressed individuals. Summing up, this new resource adds another valuable tool to systems that screen for signs of depression. The good behavior in terms of precision-oriented and recall-oriented metrics makes the lexicon usable under a wide variety of search and depression screening tasks. The new lexica are publicly available (along with the source code) in GitHub[10]. To further understand the effect of the expansion on the original set of non ambiguous adjectives, Table 7 shows the new words included into this lexicon by the most effective expansion method (DE). The list of expanded words includes many words that are potentially indicative of signs of depression (e.g. anger, bruise, collapse or grieve).

The De Choudhury et al. lexica obtained with non ambiguous adjectives and expanded with DE or WE are also effective. As a matter of fact, these are the best performing lexica obtained from the original De Choudhury et al. lexicon. Although the original lexicon had only 7 non-ambiguous adjectives, its expansion led to resources that are superior to those obtained from nouns or verbs.

5 Conclusions and Future Work

The prevalence of some mental health disorders, such as depression or anxiety, in our society puts severe constraints on the ability of health systems to provide adequate diagnosis and treatment. The growing popularity of Social Media introduces new opportunities for automatic screening of depression. Although the assessment of medical experts has no technological substitute, new automatic solutions should be considered. Automatic screening tools can complement the human labor and assist physicians in early identifying signs of mental disorders. In this context, the availability of training data in the form of annotated textual corpora is scarce and, thus, lexical resources for supporting unsupervised analysis of text are crucial.

[10] `https://github.com/gamallo/depression_classification`

In this paper, we proposed new methods for evaluating depression lexica and showed that recent advances in natural language processing can further enhance the quality of lexical resources. A lexicon composed of non-ambiguous adjectives and expanded following a distributional strategy leads consistently to the best results. Our experiments demonstrate that large-scale automatic screening is a near-future opportunity. We contributed by adding additional resources and evaluation methods to the repertoire of technologies currently available for studying depression and language use.

In the future, we plan to extend this analysis to other lexical resources. For example, we plan to create new depression lexica from depression questionnaires or structured interviews that are commonly used by psychiatric physicians. Another line of future work consists of applying depression lexica for sequentially processing user's writings (in a chronologically ordered way). This iterative form of analyzing depression has received increasing attention and it is the main topic of recent evaluation campaigns such as the eRisk challenge that runs under CLEF since 2017 (Losada et al, 2017a).

Acknowledgements This work has received financial support from i) the "Ministerio de Economía y Competitividad" of the Government of Spain and FEDER Funds under the research project TIN2015-64282-R, ii) a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, iii) a TelePares (MINECO, ref:FFI2014-51978-C2-1-R) project, and iv) Xunta de Galicia – "Consellería de Cultura, Educación e Ordenación Universitaria" and FEDER Funds through the following 2016-2019 accreditation: ED431G/08.

References

Abdaoui A, Azé J, Bringay S, Poncelet P (2017) Feel: a french expanded emotion lexicon. Language Resources and Evaluation 51(3):833–855

Almeida H, Briand A, Meurs MJ (2017a) Detecting early risk of depression from social media user-generated content. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop proceedings

Almeida H, Briand A, Meurs MJ (2017b) Detecting early risk of depression from social media user-generated content. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop proceedings

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/769\_Paper.pdf`

Baeza-Yates R, Ribeiro-Neto B (2011) Modern Information Retrieval: The Concepts and Technology Behind Search. Addison Wesley

Baroni M, Dinu G, Kruszewski G (2014) Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In:

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, pp 238–247

Benamara F, Cesarano C, Picariello A, Reforgiato D (2007) Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In: In Proceedings of ICWSM conference

Biemann, C, M R (2013) Text: Now in 2d! a framework for lexical expansion with contextual similarity. Journal of Language Modelling 1(1):55–95

Biemann C (2016) Vectors or graphs? on differences of representations for distributional semantic models. In: Proceedings of the Workshop on Cognitive Aspects of the Lexicon, Osaka, Japan, pp 1–7

Blacoe W, Lapata M (2012) A comparison of vector-based representations for semantic composition. In: Empirical Methods in Natural Language Processing - EMNLP-2012, Jeju Island, Korea, pp 546–556

Bordag S (2008) A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. In: 9th CICLing, pp 52–63

Brandt M, Boucher J (1986) Concepts of depression in emotion lexicons of eight cultures. International Journal of Intercultural Relations 10(3):321 – 346, DOI https://doi.org/10.1016/0147-1767(86)90016-7, URL http://www.sciencedirect.com/science/article/pii/0147176786900167

Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A (2008) Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. Journal of General Internal Medicine 23(1):25–36

Cheng FPG, Ramos MR, Bitsch ÁJ, Jonas MS, Ix T, See QPL, Wehrle K (2016) Psychologist in a pocket: Lexicon development and content validation of a mobile-based app for depression screening. JMIR Mhealth Uhealth 4(3):e88, DOI 10.2196/mhealth.5284, URL http://mhealth.jmir.org/2016/3/e88/

Chenlo JM, Losada DE (2014) An empirical study of sentence features for subjectivity and polarity classification. Information Sciences 280:275–288

Choudhury MD, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. In: Kiciman E, Ellison NB, Hogan B, Resnick P, Soboroff I (eds) ICWSM, The AAAI Press, URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2013.html#ChoudhuryGCH13

Coppersmith G, Dredze M, Harman C (2014) Quantifying mental health signals in Twitter. In: ACL Workshop on Computational Linguistics and Clinical Psychology

Devitt A, Ahmad K (2013) Is there a language of sentiment? an analysis of lexical resources for sentiment analysis. Language Resources and Evaluation 47(2):475–511

Fellbaum C (1998) A semantic network of English: The mother of all WordNets. Computer and the Humanities 32:209–220

Gamallo P (2017) Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. Language Resources and Evaluation 51(3):727–743

Gamallo P, Bordag S (2011) Is singular value decomposition useful for word simalirity extraction. Language Resources and Evaluation 45(2):95–119

Gamallo P, Garcia M (2017) Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. Linguamática 9(1)

Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences 18(Supplement C):43 – 49, sI: 18: Big data in the behavioural sciences (2017)

Huang E, Socher R, Manning C (2012) Improving word representations via global context and multiple word prototypes. In: ACL-2012, Jeju Island, Korea, pp 873–882

Landauer T, Dumais S (1997) A solution to Plato's problem: The Latent Semantic Analysis theory of acquision, induction and representation of knowledge. Psychological Review 10(2):211–240

Lebret R, Collobert R (2015) Rehabilitation of count-based models for word vector representations. In: Gelbukh AF (ed) CICLing (1), Springer, Lecture Notes in Computer Science, vol 9041, pp 417–429

Levy O, Goldberg Y (2014a) Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, pp 302–308

Levy O, Goldberg Y (2014b) Linguistic regularities in sparse and explicit word representations. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014, pp 171–180

Levy O, Goldberg Y, Dagan I (2015) Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3:211–225

Liu B (2012) Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers

Losada DE, Crestani F (2016) A test collection for research on depression and language use. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016, Evora, Portugal

Losada DE, Crestani F, Parapar J (2017a) CLEF 2017 eRisk overview: Early Risk Prediction on the Internet: Experimental foundations. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop proceedings

Losada DE, Crestani F, Parapar J (2017b) eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. In: 8th International Conference of the CLEF Association, Springer Verlag, pp 346–360

Mikolov T, Yih Wt, Zweig G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, pp 746–751

Mitchell AJ, Rao S, Vaze A (2011) International comparison of clinicians' ability to identify depression in primary care: meta-analysis and meta-regression of predictors. British Journal of General Practice 61(583):e72–e80

Mitra B, Craswell N (2017) An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval (to appear)

Nease D, Maloin J (2003) Depression screening: a practical strategy. The Journal of family practice 52(2):118–124

Neuman Y, Cohen Y, Assaf D, Kedma G (2012) Proactive screening for depression through metaphorical and automatic text analysis. Artificial Intelligence in Medicine 56(1):19 – 25

Neuman Y, Assaf D, Cohen Y, Knoll JL (2015) Profiling school shooters: Automatic text-based analysis. Frontiers in Psychiatry 6:86, DOI 10.3389/fpsyt.2015.00086, URL https://www.frontiersin.org/article/10.3389/fpsyt.2015.00086

Padró M, Idiart M, Villavicencio A, Ramisch C (2014) Nothing like good old frequency: Studying context filters for distributional thesauri. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp 419–424

Piasecki M, Szpakowicz S, Fellbaum C, Pedersen BS (2013) Introduction to the special issue: On wordnets and relations. Language Resources and Evaluation 47(3):757–767

Ramirez-Esparza N, Chung CK, Kacewicz E, Pennebaker JW (2008) The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. In: In Proc. ICWSM 2008

Schwartz HA, Eichstaedt J, Kern ML, Park G, Sap M, Stillwell D, Kosinski M, Ungar L (2014) Towards assessing changes in degree of depression through facebook. ACL Workshop on Computational Linguistics and Clinical Psychology pp 118–125

Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pp 417–424

Wang L, Xia R (2017) Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pp 502–510, URL https://aclanthology.info/papers/D17-1052/d17-1052

Wang P, Lane M, Olfson M, Pincus H, Wells K, Kessler R (2005) Twelve-month use of mental health services in the United States: Results from the national comorbidity survey replication. Archives of General Psychiatry 62(6):629–640