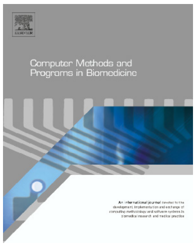
computer methods and programs in biomedicine 120 (2015) 27–36



journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Screening Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods

Christian Karmen[a,*], *Robert C. Hsiung*[b], *Thomas Wetter*[a,c]

[a] *Heidelberg University Hospital, Institute of Medical Biometry and Informatics, Medical Informatics Unit, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany*

[b] *Dr. Bob LLC, PO Box 14579, Chicago, IL 60614-0579, United States*

[c] *University of Washington, Department of Biomedical Informatics and Medical Education, 1959 NE Pacific Street, Seattle, WA 98195-7240, United States*

ARTICLE INFO

Article history:
Received 28 March 2014
Received in revised form
10 March 2015
Accepted 20 March 2015

Keywords:
Depression
Automatic screening
Natural Language Processing
Social Internet communication

ABSTRACT

Depression is a disease that can dramatically lower quality of life. Symptoms of depression can range from temporary sadness to suicide. Embarrassment, shyness, and the stigma of depression are some of the factors preventing people from getting help for their problems. Contemporary social media technologies like Internet forums or micro-blogs give people the opportunity to talk about their feelings in a confidential anonymous environment. However, many participants in such networks may not recognize the severity of their depression and their need for professional help. Our approach is to develop a method that detects symptoms of depression in *free text*, such as posts in Internet forums, chat rooms and the like. This could help people appreciate the significance of their depression and realize they need to seek help. In this work Natural Language Processing methods are used to break the textual information into its grammatical units. Further analysis involves detection of depression symptoms and their frequency with the help of words known as indicators of depression and their synonyms. Finally, similar to common paper-based depression scales, e.g., the CES-D, that information is incorporated into a single depression score.

In this evaluation study, our depressive mood detection system, *DepreSD* (**Depression Symptom Detection**), had an average precision of 0.84 (range 0.72–1.0 depending on the specific measure) and an average F measure of 0.79 (range 0.72–0.9).

1. Introduction

1.1. *Barriers to seeking depression treatment*

Loss of quality of life is a serious matter for everyone. Whatever makes us unhappy or sad for a longer period of time can lower our self-esteem, spirit, and love for life. In the worst case it causes people to commit suicide [1]. Recently, there has been a sharp increase of depression and suicide registered in the US [2] as well as in Europe [3]. It is estimated that about 27% of the adult European Union (EU) population is or has been affected by at least one mental disorder in the previous 12 months [4]. However, only 42.3% of patients with a

[*] *Corresponding author*. Tel.: +49 6221 56 7634; fax: +49 6221 56 4997.
E-mail addresses: christian.karmen@med.uni-heidelberg.de (C. Karmen), thomas.wetter@urz.uni-hd.de (T. Wetter).

http://dx.doi.org/10.1016/j.cmpb.2015.03.008
0169-2607/© 2015 Elsevier Ireland Ltd. All rights reserved.

depressive disorder are correctly diagnosed in primary care [5].

Suffering from clinical depression can be hard for people to recognize and needing professional help can be even harder to accept. Shyness, anxiety about discussing problems, and especially the stigma of having depression can keep people from sharing how they are feeling and from getting help.

Nowadays, many people, especially adolescents, have found a way to share their suffering without being exposed: the Internet [6–8]. Countless Internet forums and other social communication platforms provide an anonymous environment where like-minded people from practically anywhere in the world can exchange experiences and ideas.

Online detection may be a proactive and promising way to identify high risk individuals, may facilitate timely intervention, and may improve public health [9].

1.2. *Computer-aided depression detection*

The use of computer-aided methods to improve depression detection has been pursued previously and many approaches have been developed and published. This section gives a short overview of similar projects.

[10] developed a computer-adaptive test for depression (D-CAT) [10] based on Item Response Theory (IRT). Similar to the classic paper-based depression scales like the *Center for Epidemiological Studies – Depression Scale* (CES-D) [11], the patient is interviewed with an adaptive questionnaire. If an early answer is that the patient is in a good mood and feels positive about the future, he is not later asked about active suicidal ideation. Their work was meant to improve clinical depression detection and rating while reducing the respondent's burden. However, as with paper-based depression scales, an active role is assumed by the clinician who administers the test to patients. Although many questionnaires are freely available and can be used for self-assessment [12], most people are not motivated to test themselves at appropriate intervals, if at all.

Another approach is the work of Lehrmann et al. [13]. These researchers developed a system for *detecting distressed and non-distressed affect states in short forum texts*. Their goal was to detect distress rather than depression (even though the two might be correlated). They used 42 linguistic features (including positive and negative word list matches, affect word list matches, and pronouns) and supervised machine-learning (ML) methods (Naive Bayes, Max Entropy, Decision Tree and Max Vote) to classify posts. Depending on the k in k-fold cross-validation their automatic classification method is generally 20% more accurate than randomly assigning a distress label to a post.

Meta-learners are an interesting method to increase correct classification rates for ML methods, for example with stacked generalization as performed by Dinakar et al. [14]. Here, an online community supporting adolescents under stress was analyzed. First, they performed widely used supervised learning methods for text classifications to categorize into 23 themes: support-vector machine with a linear kernel (SVM-L), a radial basis function kernel (SVM-R) and a stochastic gradient boosted decision trees (GBDT) model. They used a specific set of features (consisting of unigrams, bigrams, POS bigrams and *tf-idf*) in order to get base classifiers for each theme. The output of the base classifiers lead to vectors of predictions as used by Guyon et al. and Friedman [15,16] (1) which together with the topic distribution from the L-LDS model for a given story (2) is then combined into the meta-feature set. They applied SVM-L, SVM-R and GBDT again, now on the outputs as meta-learners and compared the best meta learner's results with each base model. As it turns out, the best meta learner for each theme is in most cases significantly better than any of the base models. Especially poorly categorized themes benefited most.

At about the same time as our research De Choudhury and et al. [17,18] have also applied ML for inferences about depression from segments of natural language text. However, their approach and the populations differ in various important aspects: First, their subjects are a priori diagnosed with depression and NLP methods try to replicate that diagnosis. In contrast, our approach starts without prior assumptions and tries to classify posts as to whether they do or do not indicate depression. Second, their subjects are users of general social media. Therefore, the prevalence of depression is presumably higher in our sample. Third, the being aware of a depression context of our users versus a neutral context of their users may unconsciously affect the expressiveness of depression related utterances in both directions between the two situations without that he had a clue which of the two: from a communicational perspective participants in general media maybe clearer and more detailed because they assume that the readers do not know a lot about depression. Conversely, from a perspective of social desirability subjects may well be less explicit because they unconsciously want to not showcase the full depth of their problems. Fourth, De Choudhury establishes *ground truth* through apparently approved instruments – CES-D and BDI – the validation and calibration of which for presentation online is still due.

Still in the context of depression in unstructured text in social media, Coppersmith et al. [19] found a method to quantify mental health signals in micro-blog data (Twitter). Their work uses the Linguistic Inquiry Word Count (LIWC), a validated tool for psychometric analysis [20]. LIWC quantifies different categories of words and determines the degree of positive or negative emotions, self-references, causal words, and many other language dimensions. Coppersmith conducted an analysis of a set of mental illnesses (post-traumatic stress disorder, depression, bipolar disorder and seasonal affective disorder) to measure deviations of several LIWC categories in each illness group (1238 users) from the control group (5728 users). They used two language models (LM): a traditional unigram LM to examine the probability of each whole word (1) and a character 5-gram LM to examine sequences of up to 5 characters (2). When calibrated to a false alarm rate of 0.1, the precisions of their results range from 0.42 to 0.67 for the four mental illness categories, and from 0.65 to 0.82 with a false alarm rate of 0.2.

Another comparable, but different kind of approach to detect stress is the usage of a deep neural network (DDN), as done by Lin et al. [21]. They analyzed data from four micro-blog platforms and compared their 4-layered DDN with "traditional" classifiers, such as SVM, Naive Bayes and Random Forest. To measure performance they used three pooling models (max pooling, mean-over-instance and mean-over-time) for each model. Depending on the pooling method, each model performed slightly better or worse. Overall best results were achieved using the DNN with mean-over-time pooling with an achieved precision of 0.79 and *F* measure of 0.84.

A different approach in NLP depression analysis was developed by Neuman et al. [22]. They introduced a system for metaphor-based automatic screening for depression. To achieve this, they developed a methodology, *Pedesis*, to crawl websites, based on web search engine results, and to extract patterns of the form *depression is like* *, where * is a wildcard. With the NLP method of Dependency Parsing they extracted what they called *domains*: words or phrases that were metaphorical expressions of *depression*. These elements were further extended with the help of first- and second-degree synonyms and formed a list they called a *depression lexicon*. They evolved three different depression measures, the most enhanced of which was *Dep_Scale**, and performed an evaluation study of each.

2. Methods

2.1. Concept

The diagnosis of depression is more complex than it appears from the ICD-10 definition. The clinician listens for all hints of symptoms as the patient describes his condition and asks him questions to elicit additional information. A standardized questionnaire [23] like the CES-D or the WHO-5 Well-Being Index (WHO-5) [24] may be used as a clinical tool to aid in the diagnosis. For example, by presenting a tool to the patient in advance, the clinician can be better prepared for the patient and respond to red flags. These questionnaires have basically the same schema:

1. Detecting depression symptoms (e.g., insomnia)
2. Determining the frequency of each symptom
3. Calculating a depression score by summing the frequencies of the symptoms
4. Using thresholds to classify the intensity of depression (none, mild, moderate, severe)

We used a similar approach to generate a depression score: Raw scores for symptoms and their frequencies were multiplied, and the sum of those products gave an overall score for the present severity of depression. We assumed that the more symptom terms there were in a forum text and the higher the frequency of those symptoms, the higher the likelihood that the mood of the author at that moment was depressed.

Social online communication is usually characterized by colloquial language [25]. Psychiatric jargon is rarely used and thus hardly detectable. Clinical meta-thesaurus and classification systems such as UMLS and ICD-10 do not cover lay language expressions. Therefore, to capture phrases such as *(had a) bad sleep* or *(cannot get to) lala-land*, thesaurus generated synonyms for the clinical terms for depression symptoms were used. Since language changes continuously and some patients may use rare or even bizarre words, we further extended the word list with synonyms of synonyms or *second-degree synonyms*. *Terrible* is an example. It is not an obvious symptom but may be used as such and demonstrates the fuzziness of the synonym of synonym approach to building lexicons.

In a nutshell the whole concept fits into four steps:

- Symptom lexicon generation: Create lexicon files with depression symptom terms and generate as many reasonable synonyms as possible
- Frequency lexicon generation: Create lexicon files with frequency terms in four categories (never, sometimes, often, always) by generating as many reasonable synonyms as possible
- NLP processing: Use NLP methods on texts to determine depression symptoms, associated frequencies, personal pronouns, and negation
- Score calculation: Generate a depression score

Fig. 1 illustrates the concept. We assumed that each post reflected the current state of the author's mood.

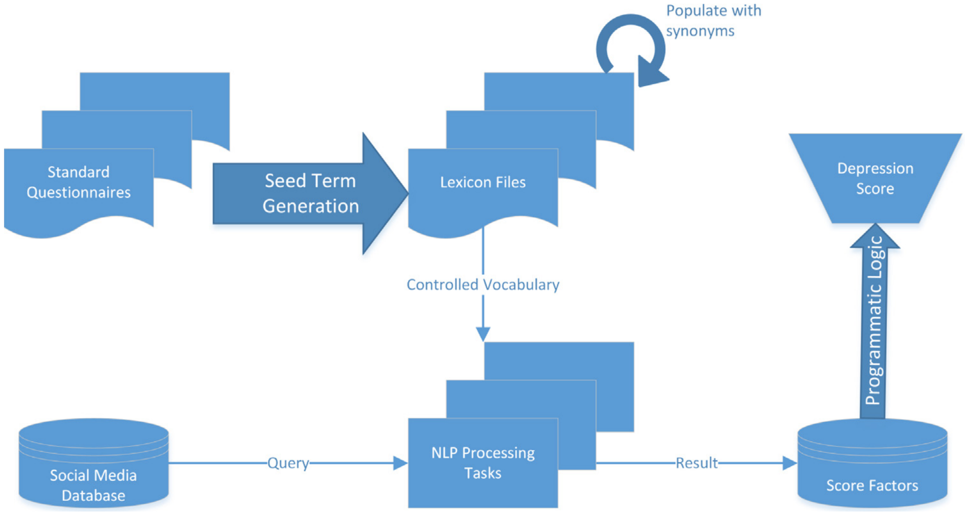


Fig. 1 – Graphical representation of the concept.

2.2. *Architecture*

2.2.1. *Lexicon generation*

In this section, we describe the creation of a generic list of depression symptoms, from now on called seed terms or the symptom lexicon. It contained a hand-selected list of depression symptoms from conventional sources for psychiatrists. This list was critical since it was the base for the synonym finder and thus errors would propagate into the synonyms.

As mentioned earlier, there are different lists of depression symptoms in different classification systems. Those make use of either clinical or colloquial symptom terms, depending on whether the questionnaire is directed to the patient or the clinician. A good overall base list is assumed by merging the two major classification systems for depression, the DSM-IV and the ICD-10, as well as the most commonly used depression scales:

- Center for Epidemiologic Studies Depression Scale (CES-D) [11]
- Beck Depression Inventory II (BDI-II) [26]
- Geriatric Depression Scale (GDS) [27]
- Hamilton Depression Scale (HAM-D) [28]
- Montgomery–Asberg Depression Rating Scale (MADRS) [29]
- Zung Self-Rated Depression Scale (SDS) [30]
- World Health Organization-Five Well-being Index (WHO-5) [24]

Since the questionnaires for patients do not contain a straight symptom list but rather corresponding practical situations, manual (and thus subjective) rephrasing was necessary.

After the manual selection of symptoms from the above sources, the word base of each item was extended with the corresponding noun, verb and adjective whenever reasonable, e.g., *depressed*, *depressive*, *depression*. By doing so, more hits for the lexicon-matching algorithm were expected. This symptom lexicon was discussed with and reviewed by the psychiatrist author.

Having the seed terms for depression symptoms, the generation of synonyms was started. We developed a synonym finder tool that parsed a publicly available web-based dictionary service: *Thesaurus.com* from *Dictionary.com, LLC* [31]. It contained more than one million words [32], including derivatives and inflections. On a small sample of seed words it delivered more than 30 synonyms per word. The occasional metaphorical synonyms were especially suitable for finding words and phrases used in colloquial texts.

Many off-topic, unrelated synonyms were generated. After manual tracking of individual seed terms and their synonyms, the nature of this issue was found: ambiguous synonyms like homonyms (*blue* = color; mood) and antagonyms (*bad* = *very good*) as well as wrong or fuzzily interpreted synonyms.

We solved this problem by creating individual blacklists containing all redundant, misinterpreted and fuzzily interpreted words for each degree of synonym. This way unnecessary processing of unwanted phrases was avoided. Each blacklist was reviewed by the psychiatrist author to preserve the psychological value of the synonyms.

The found symptoms in texts were matched with their frequencies. We adopted the frequency gradations from the CES-D, which are similar to those of other depression scales:

- Level 0: Rarely or none of the time
- Level 1: Some or a little of the time
- Level 2: Occasionally or a moderate amount of time
- Level 3: Most or all of the time

Since the word variation in the frequency lexicons was very limited compared to the symptom lexicons, some phrases were manually added. This way a considerable number of corresponding synonyms was found and enriched the word base. Overlapping, redundant and ambiguous phrases were manually removed to yield a unique set of distinctive items.

As a last step, the lexicon items were lemmatized. This was necessary because some items were used in conjugated form in texts. For example, the phrase *this is depressing* does not match the item *depressed*, but after lemmatization, *this is depress* does match *depress*. Lemmatization further reduced the size of the lexicon lists because some items were initially present in multiple forms. Table 1 shows the impact of lemmatization on the number of items per lexicon.

Another issue in parsing English grammar are paradigms, grammatically related forms of words. According to Oxford English Grammar [33], paradigms can be established for verbs, nouns, adjectives and in some cases adverbs. There are a number of conjugations depending on the pronoun and tense. With nouns there are variations in declensions, with adjectives, inflections for comparison (e.g. *new*, *newer* and *newest*). Once again, lemmatization demonstrated its importance in NLP by resolving these paradigms.

2.2.2. *NLP processing*

We used our symptom and frequency lexicons as controlled vocabularies to implement a grammar-oriented processing. The steps were

1. Preprocessing. We removed unwanted text such as formatting, HTML links and quotations of earlier comments [34].
2. Boundary detection. We used sentence detection by the Stanford Parser [35] because it is a relatively reliable method in NLP. The detection of phrases was more

Table 1 – Lexicon items: Lemmatization statistics.

Lexicon	# of items (not lemmatized)	# of items (lemmatized)	Reduction
Symptoms – Seed terms	197	180	8.63%
Symptoms – Synonyms 1st degree	1873	1757	6.19%
Symptoms – Synonyms 2nd degree	8738	7942	9.09%
Overall	10,816	9887	**8.59%**

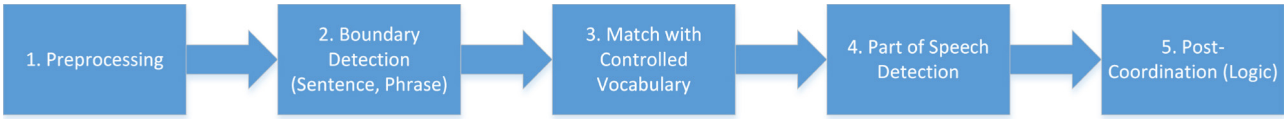

1. Preprocessing
2. Boundary Detection (Sentence, Phrase)
3. Match with Controlled Vocabulary
4. Part of Speech Detection
5. Post-Coordination (Logic)

Fig. 2 – Graphical representation of the NLP processing chain.

complicated. We developed an algorithm that detected phrases and split the sentences into smaller units. Ideally, the units were atomic statements consisting only of one subject and its predicate.

3. Matching. The phrases were matched with the lexicons. Each lemma in the phrase was compared with all lemmas in the three symptom lexicons (seed terms and first- and second-degree synonyms) and the four frequency lexicons (Fig. 2).
4. Pronouns and negation. With the help of the part of speech (POS) tags generated by the Stanford Parser, various negation words were detected. Under the assumption that the phrases were minimal, the simple occurrence of a negation word was sufficient to indicate inversion of the phrase.
5. Post coordination. We defined a score as a measure of the four previously detected factors (pronoun, negation, matched symptom and frequency) in each phrase.

2.2.3. *Score calculation*

We now had all the ingredients for a depression score. The *DepreSD*-score itself is a sum of products. Each product represents a phrase in the post. Four characteristics of the phrase are converted into four factors. The product of the four factors is the score of the phrase. The sum of the scores of the phrases is the score of the post.

$$DepreSD = \sum_{(phrases)} w_{pronouns} \cdot w_{negation} \cdot f_{frequencies} \cdot f_{symptoms} \qquad (1)$$

The first two factors, for pronouns and negation, are simple weights while the third and fourth, for frequencies and symptoms, are weighted sums.

In a phrase no pronoun at all (*How depressing*), or only first person pronouns, or only other person pronouns, or first and other person pronouns mixed (*I'm furious about what you said* or *I made you furious*) may be found. In the first three cases we choose $w_{pronouns} = 0.9$ and in the last $w_{pronouns} = 0.1$. This reflects our observations of the text material.

When no negation is found $w_{negation} = 1.0$. Since in at least 10% of identified negations their contextual referents were not the patient or a depression symptom of his we then set $w_{negation} = 0.1$.

To obtain the frequencies factor we add expressions $w_{level} \cdot n_{level}$ for each frequency lexicon. n_{level} is the number of words in each frequency lexicon that match the phrase. In the absence of any better knowledge or results we choose weights $w_0 = 0.1, w_1 = 1/3, w_2 = 2/3, w_3 = 0.9$ for the four levels of frequency. As a special case we set $w_{none} = 0.5$ and $n_{none} = 1$ so that if no frequency phrase matches, $f_{frequencies} = 0.5$.

$$f_{frequencies} = \begin{cases} 0.5, & \textit{no frequency word found} \\ \sum_{level=0}^{3} w_{level} \cdot n_{level}, & \textit{else} \end{cases} \qquad (2)$$

Similarly, the symptoms factor is the sum of $w_{lexicon} \cdot n_{lexicon}$ for each symptom lexicon. We choose $w_{seed} = 1.0, w_{1stsyn} = 0.85, w_{2ndsyn} = 0.35$.

$$f_{symptoms} = \sum_{lexicon \in \{seed, 1stsyn, 2ndsyn\}} (w_{lexicon} \cdot n_{lexicon}) \qquad (3)$$

Example: *depressed* is a symptom (seed) term and *always* is a level 3 frequency term. The post is a single sentence: "I always feel depressed." It consists of a single phrase. The pronoun factor is 0.9 since only a first person pronoun is found. The negation factor is 1.0 since negation is not found. The frequency factor is 0.9 because the frequency term (*always*) matches 1 time and has weight 0.9 (level 3). The symptom factor is 1.0 because the symptom term (*depressed*) matches 1 time and has weight 1.0 (seed term). Therefore $DepreSD = 0.9 \cdot 1.0 \cdot 0.9 \cdot 1.0 = 0.81$.

We also applied two modifications to the raw *DepreSD*-score: a *DepreSD* sentences ratio (*DepreSD*/# ofsentencesinpost) as well as a *DepreSD* words ratio (*DepreSD*/# ofwordsinpost). This way we could also consider hit "concentration". That gave us a total of three measures.

2.3. *Material*

The psychiatrist author moderates a public forum called *Psycho-Babble* [36] for people to have conversations about a number of topics – mainly related to depression and other mental health issues [37]. Since it opened in 1998, over 20,000 users have registered and over 1,000,000 messages have been posted. This dataset is ideal as a text corpus for the research of depression symptoms in non-standardized text data (often sloppily called *free text* [38,39]) as it provides a large number of posts from people who actually have or have had some form of depression. Only a small percentage of the posts are written by others like people simply interested in the topic or Hsiung himself in his moderator role.

2.4. *Selection from the material*

We chose the forum domain *Psycho-Babble Grief* because its topic promised relatively high scores. Some key features of this forum are:

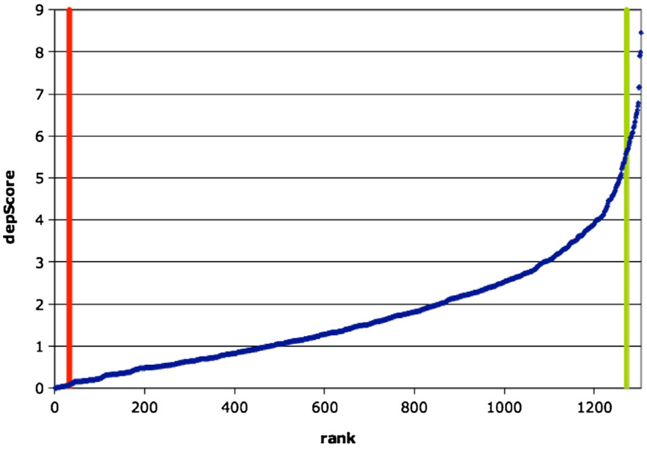


Fig. 3 – (Raw) *DepreSD*-score distribution. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

- 2,110 posts in total
- 1,304 posts within a 20–200 word interval that contained
  - 9,096 sentences
  - 22,576 phrases
  - 107,814 words

The gold standard for the detection of depressive symptoms in social media posts is manual review by experts. To keep the number of posts to be manually assessed realistic we decided on an contrast group design. We designed a presumably robust filtering of the available posts to identify extreme cases.

3. Results

We applied our algorithm to the selected *Psycho-Babble* subset and compared the results to independent experts' ratings. Each of the three measures delivered a ranking of the 1304 posts. From each ranking, we selected the 33 highest (presumably most severe) and 33 lowest (presumably least severe). Figs. 3, 4, and 5 illustrate the distribution of the measures as well as the thresholds (green line = high end; red line = low end). In the collected 198 posts 76 congruent ones were detected and thus removed, so the final number of unique posts was 122.

The experts, one male psychiatrist and one female nurse practitioner, each with several years of clinical experience, were blind to the purpose of the study. We asked them to use their clinical judgment to rate the posts according to this 4-point Likert scale:

- 0 – not depressed
- 1 – mildly depressed
- 2 – moderately depressed
- 3 – severely depressed

Theses categories corresponded to the those in the DSM-IV. One expert rated all 122 posts, the other 61, half from each measure's highest rankings and half from each measure's lowest. In this study, the focus is on the algorithm's performance in detecting depressive evidence without the determination of its severity. Thus, after the expert's judgments, their results for categories 1-3 were aggregated to one single category (1), indicating a depressive statement. We considered ratings that disagreed to be unreliable and excluded them from further analysis. The total ratings per measure were:

- 53 for *DepreSD*-score (13 disagreements)
- 46 for *DepreSD* sentence ratio (20 disagreements)
- 60 for *DepreSD* word ratio (6 disagreements)

A statistical measure of inter-rater agreement between two raters is *Cohen's kappa coefficient* (*κ*) [40] and is generally considered more robust than the percent agreement, because it considers the agreement by chance. For the expert ratings

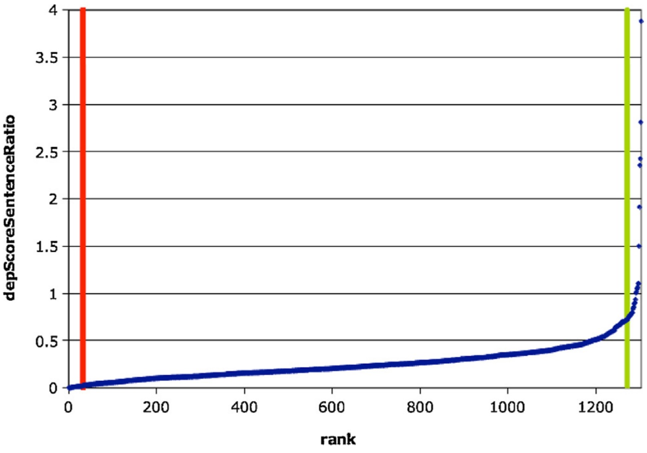


Fig. 4 – *DepreSD*-score sentence-ratio distribution. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

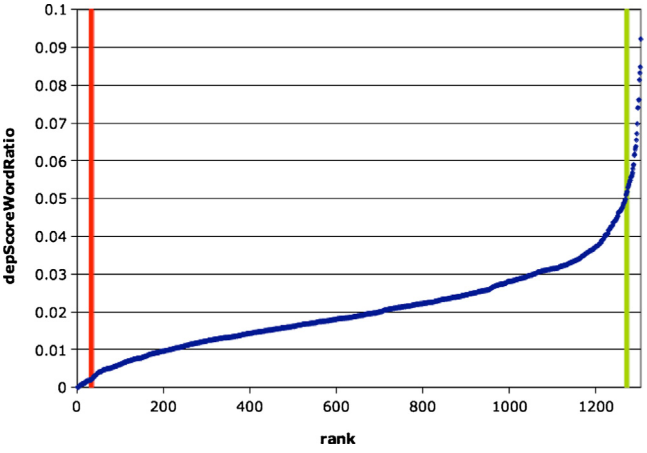


Fig. 5 – *DepreSD*-score word-ratio distribution. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

Table 2 – Expert evaluation agreement statistics.

Expert A	Expert B		Total
	Not depressed (0)	Depressed (1)	
Not depressed (0)	20	27	47
Depressed (1)	0	14	14
Total	20	41	61

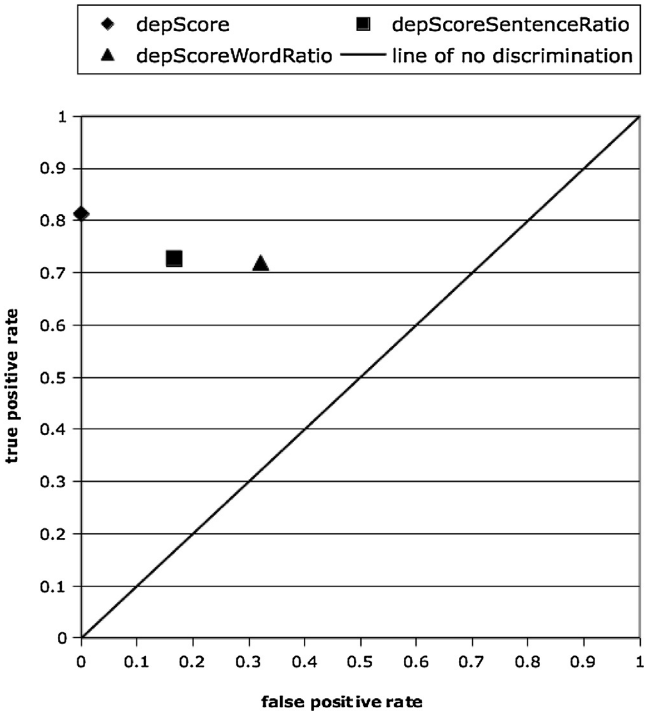


Fig. 6 – *DepreSD*-score results in TP-FP relation.

from Table 2 κ is 0.37. Due to Landis [41] this value is categorized as *Fair* (0.21-0.40) for the strength of agreement.

As can be seen in Table 3, the precision of the raw *DepreSD*-score was 1.0 and its recall 0.81. In a diagnostic environment the term *sensitivity* can also be used instead of the statistical term *recall* – our focus however is on the technical aspect of the algorithm. The other measures were a little less accurate. The sentence ratio had precision 0.8 and recall 0.73, the word ratio precision and recall both 0.72. Fig. 6 shows that all measures were clearly higher than random results.

4. Discussion

4.1. *Rationale for an NLP approach*

Other approaches to characterizing possibly depressed persons based on raw data may come to mind. We pursued an NLP approach because of the nature of the disease of *depression* and our lack of access to the post authors.

Depression has been clinically characterized for 150 years. Standardized instruments such as the CES-D and the BDI-II were developed half a century ago. The frequency of symptom occurrence is a major determinant of severity, and patients with equal severity may present with widely different emotional, somatic, behavioral, or other symptoms.

The variety of symptoms largely invalidates the assumption of many text mining approaches that semantic proximity, i.e., similar meaning, is substantiated through syntactic proximity, i.e., similar terms. For depression it is not true that similarly severe cases are necessarily represented by similar texts. In terms of support vector machines: similarly severe cases do not necessarily populate connected regions in a multidimensional space. The variety of symptoms also invalidates the assumption of decision trees that there is a single overarching symptom about which to ask first and by which the patients can neatly be separated into meaningful major categories which can then be further subdivided using the next level of discriminatory symptoms.

Frequency as an indicator of severity seems to favor automatic methods such as *tf-idf* [43,44]. *tf-idf* determines the uniqueness and character of a text through *elite* terms: terms used frequently in a text vs. rarely in the overall corpus. Text are regarded as pertaining to the same topic if their elite terms overlap. The multifaceted nature of depression may, however, lead to some equally severe cases being associated with similar elite terms while other equally severe cases are associated with different ones. *tf-idf* would detect how writers of texts are depressed, not how depressed they are. Hence *tf-idf* would require high computational effort to achieve a differentiation not required here while at the same time missing the subtle cues of different words for equally strong determinants of frequency. In our approach words such as *rarely*, *often*, *always*, etc., capture the frequency of symptoms. Since they are common words they would likely escape the elite word search and we would lose that key information.

Negation and pronouns call for an NLP rather than an ML approach. The mere presence of a term that standard ML methods would make use of is not enough. We have to be certain that the term is not negated and should not be attributed to another person. What may appear to be linguistic

Table 3 – Result overview.

Measure	Total	TP	FP	TN	FN	Spec.	Prec.	Prec. CI	Recall	Recall CI	F
DepreSD-score (Raw)	53	26	0	21	6	1.00	1.00	[0.87;1.0]	0.81	[0.65;0.91]	0.90
DepreSD sentence ratio	46	16	4	20	6	0.83	0.80	[0.58;0.92]	0.73	[0.52;0.87]	0.76
DepreSD word ratio	60	23	9	19	9	0.68	0.72	[0.55;0.84]	0.72	[0.55;0.84]	0.72

TP, true positive; FP, false positive; TN, true negative; FN, false negative; CI, confidence interval [42].

preprocessing is actually essential data quality assurance. Not considering negation and pronouns would create data so noisy that even large samples and highly sophisticated ML methods would have difficulty.

A disadvantage of ML approaches (e.g., Lehrmann [13]) is their limitation to a couple of features (distinctive words) to determine distress. The linguistic body for mood description is certainly not covered by a couple of terms (Lehrmann: 42 terms). And as Lehrmann et al. already mentioned: *Choosing the right feature set remains a difficult, poorly understood process*. The nature of classification methods leads to an extensive (usually even exponential) computational complexity per feature and thus limits its number.

Another aspect is that ML approaches focus on correlations rather than semantics. This means their results are based on statistics. A clinician may have difficulty interpreting statistical correlations. On the contrary, a grammatical approach, as used in our work, can be understood more intuitively because it is based on linguistic methods.

Our approach sets out from the assumption that there is considerable value in the 150 years of research on depression and the instruments such as PHQ-9 where this research has found a linguistic manifestation. We want to bring to bear this knowledge with a new type of material by using the approved indicators of depression and their synonyms and frequencies to detect depression. We do not want to re-invent depression. We are not in search of new symptoms although this is a legitimate goal as well. In other medical fields such as oncology this has been successfully applied: potential new side effects of chemotherapies have been identified by automatically detected similarities of descriptions provided by numerous patients [45].

4.2. *Comparison to other NLP approaches*

The *DepreSD* algorithm is able to match phrases that relate to negative moods. This study's generated lexicon, with more than 10,000 phrases overall, can be considered quite substantial compared to other NLP approaches. Because of the high number of lexicon items, the matching frequency will be relatively high even compared to regular manual annotation by experts. To detect depression in social media posts, the lexicon must be extensive enough to achieve high recall.

Manual annotation by experts of a related corpus is the current gold standard for creating mood lexicons. Depending on the size and type (for instance, reports by psychiatrists or diaries of patients) of the corpus, the results can be very different. Furthermore, many annotations focus on particular aspects, for instance only on words, adjectives, phrases or metaphors. Only a subset of depressive statements will then match. In contrast, we automatically generated mood lexicons. The domain knowledge of clinical experts was "imported" via standard classifications systems (DSM-IV, ICD-10) and common depression scales (CES-D, BDI-II, etc.). Instead of training a classifier on a manually selected set of social media posts (e.g., Lehrmann [13]), symptom phrases were "learned" with the help of a comprehensive thesaurus. By doing so, the synonym engine included colloquial and sometimes even metaphorical phrases. A comparison of the lexicons in this work and the one generated by Neuman [22], generously provided by those authors, revealed a 41.2% overlap after lemmatizing.

4.3. *Rationale for the validation method*

The typical starting point for building and testing a classifier is to have subjects assigned to classes *a priori* according to some superior gold standard judgment and to try to replicate the superior judgment with minimum likelihood of error. However, our assignment to classes is tentatively achieved by application of *DepreSD* and *a posteriori* confirmed or refuted by human experts who use the same text material. This is the only workable way to handle the situation because we do not know the gold standard (clinical professional) diagnoses of the authors. Some of them may never have received a professional diagnosis. If we had a clinical professional assess them, Hawthorne or subject expectancy effects might severely bias or blur any results we then collected. The material we used was, however, fully authentic.

4.4. *Conclusions*

The automatic generation of a depression symptom lexicon enriched by synonyms has the potential to replace expert manual annotation. In addition, it could be applied with different or additional dictionaries to cover domain-specific jargon, slang, metaphors, etc. It also could be extended to use a more fine-grained distance of related phrases; in our study the only difference in closeness to seed terms was the synonym degree. Other methods could include intersections of a range of dictionaries or semantic vocabularies such as *WordNet*® [46].

Experts can disagree about weighting. This happened when we created the blacklists for the automatically generated synonyms. The second-degree synonyms (almost 10,000 unique phrases) were especially controversial. This was because the semantic distance to the original 197 seed terms was so great that in many cases they could be interpreted as indicating depressed, neutral, or even elevated mood. Appropriate assessment remains a subjective task unless a metric is available to determine semantic word distance. In such cases a reasonable threshold would also need to be determined.

The weights of the factors have a significant impact on the precision of the measures. Calibration based on a fair number of manually rated posts could help. Alternatively, the weights could be treated as parameters and adjusted to optimize precision and recall. In the current version, the focus was on high precision (rather than on high recall). This is appropriate for detecting gradual onsets which eventually render above threshold scores and still leave the therapist sufficient time to intervene when he has confirmed manually that a false positive is unlikely. To detect sudden onsets and risks of life threatening crises the system would have to be optimized to focus on high recall. The scoring mechanism could also be considered a symptom finder rather than a depression scale. If, for instance, an author complains about bad sleep over and over, his score will increase with each insomnia-related phrase.

5. Future work

In this project, the symptom lexicons included all depression symptoms defined by the DSM-IV and ICD-10. However, the depth of the disorder depends on the number of different categories of symptoms (sleep, appetite, suicidal ideation, etc.) experienced by the patient. If the compiled depression symptom lexicon were split into separate symptom category lexicons, the measures could be modified to take into account the number of categories detected. Two matches in different categories could have more predictive value than two matches in the same category.

Another improvement in the lexicons would be the incorporation of other mood related items such as emoticons and common *Internet Slang*. Their translation into phrases would enable the NLP methods to consider them and could lead to further evidence of depressed mood. A similar effect could be expected from resolving common abbreviations.

The word basis of the seed terms was expanded by creating related nouns, verbs and adverbs of symptom terms. It is possible that some forms of modification were omitted. An automated method to generate related POS elements could be implemented. For instance, the word stem *mood* could be extended to *moody* (adjective) and *moodiness* (noun).

Another method to increase the matching rate could be to apply spell correction to the texts. Ideally, the spell-corrected substitutions would be those closest in word distance.

Our negation detection used POS tagging, which detected linguistic categories, but in complex sentences, might not have interpreted negation precisely. NLP Dependency Parsing, by detecting the relationships between words and following the dependencies of a negation to their root, could greatly increase the reliability of negation status. Dependency Parsing can be complex. Within the time limits of our initial implementation we did not attempt such advanced negation detection. Furthermore, Dependency Parsing could increase the accuracy of the mapping of symptoms to actors and to frequencies. Another improvement could be the analysis of additional linguistic means of intensifying symptoms, for instance degree modifiers (e.g., *high, higher, highest*) or qualifiers (e.g., *very* or *hardly*). We expect more precise *DepreSD*-score results with such techniques.

Our setting of offering a fully anonymous membership to the Pycho-Babble forum is a prerequisite for open spontaneous patient contribution, but at the same time precludes for patient histories to be taken into account. We acknowledge that knowing our subjects patient histories would allow to more validly judge a subject as depressed or not depressed. Thus, a next step towards validation can be a trial where volunteers are enrolled who keep using the forum but give consent for their identity to be known and used to verify intensity of depressive episode in-person. Great care has to be applied, though, because once such a fall back structure is in place there may be an ethical obligation to actually intervene when a severe episode become detected. We maintain, however, that the presentation of the material blinded forces the judges to focus exactly on the material that the automated procedure sees and avoids all kinds of confounding presuppositions.

6. Mode of availability

A free online demonstration of the *DepreSD*-score is available on the project's homepage at http://babble.imbi.uni-heidelberg.de/.

REFERENCES

[1] D.S. Vandivort, B.Z. Locke, Suicide ideation: its relation to depression, suicide and suicide attempt, Suicide Life Threat. Behav. 9 (1979) 205–218.

[2] E. Sullivan, J.L. Annest, F. Luo, T. Simon, L. Dahlberg, Suicide among adults aged 35–64 years – United States, 1999–2010, Center for Disease Control and Prevention, Morbidity and Mortality Weekly Report, 2013.

[3] H. Ellyatt, Depression – Suicides Rise as Euro Debt Crisis Intensi?es, 2013 http://www.cnbc.com/id/48883704 (last visited: 04.11.13).

[4] H.-U. Wittchen, F. Jacobi, Size and burden of mental disorders in Europe – a critical review and appraisal of 27 studies, Eur. Neuropsychopharmacol. 15 (2005) 357–376.

[5] M. Cepoiu, J. McCusker, M.G. Cole, M. Sewitch, E. Belzile, A. Ciampi, Recognition of depression by non-psychiatric physicians – a systematic literature review and meta-analysis, J. Gen. Inter. Med. 23 (2008) 25–36.

[6] T.K. Houston, L.A. Cooper, H.T. Vu, J. Kahn, J. Toser, D.E. Ford, Screening the public for depression through the Internet, Psychiatr. Serv. 52 (2001) 362–367.

[7] P. Leiberich, J. Nedoschill, M. Nickel, T. Loew, K. Tritt, Selbsthilfe und Beratung im Internet, Medizinische Klinik 99 (2004) 263–268.

[8] R. Savolainen, Requesting and providing information in blogs and Internet discussion forums, J. Doc. 67 (2011) 863–886.

[9] T.M. Li, M. Chau, P.W. Wong, P.S. Yip, A hybrid system for online detection of emotional distress, in: Intelligence and Security Informatics, Springer, 2012, pp. 73–80.

[10] H. Fliege, J. Becker, O.B. Walter, J.B. Bjorner, B.F. Klapp, M. Rose, Development of a computer-adaptive test for depression (D-CAT), Qual. Life Res. 14 (2005) 2277–2291.

[11] L.S. Radloff, The CES-D scale a self-report depression scale for research in the general population, Appl. Psychol. Meas. 1 (1977) 385–401.

[12] R.C. Hsiung, Babbleometer, 2007 http://www.dr-bob.org/babble/ometer/ (last visited: 04.04.13).

[13] M.T. Lehrman, C.O. Alm, R.A. Proaño, Association for computational linguistics, in: Proceedings of the Second Workshop on Language in Social Media, 2012, pp. 9–18.

[14] K. Dinakar, E. Weinstein, H. Lieberman, R. Selman, Stacked generalization learning to analyze teenage distress, in: International AAAI Conference on Weblogs and Social Media, 2014.

[15] I. Guyon, B. Boser, V. Vapnik, Automatic capacity tuning of very large VC-dimension classifiers, Adv. Neural Inf. Process. Syst. (1993) 147–155.

[16] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[17] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: ICWSM, 2013.

[18] M. De Choudhury, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 47–56.

[19] G.C.M.D.C. Harman, Quantifying mental health signals in twitter, ACL 2014 (2014) 51.

[20] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic inquiry and word count: LIWC 2001, Lawrence Erlbaum Associates, Mahway, 2001, pp. 2001.
[21] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, L. Feng, User-level psychological stress detection from social media using deep neural network, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 507–516.
[22] Y. Neuman, Y. Cohen, D. Assaf, G. Kedma, Proactive screening for depression through metaphorical and automatic text analysis, Artif. Intell. Med. 56 (2012) 19–25.
[23] S.M. Gilbody, A.O. House, T.A. Sheldon, Routinely administered questionnaires for depression and anxiety: systematic review, Br. Med. J. 322 (2001) 406–409.
[24] World Health Organization (WHO), (Five) well-being index (1998 version), 1998 http://www.who-5.org (last visited: 18.03.14).
[25] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteský, A. Holzinger, Opinion mining on the web 2.0-characteristics of user generated content and their impacts, in: Human–Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Springer, 2013, pp. 35–46.
[26] A.T. Beck, C.H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Arch. Gen. Psychiatry 4 (1961) 561.
[27] J. Yesavage, T. Brink, T. Rose, et al., Geriatric depression scale (GDS), Handbook of psychiatric measures, American Psychiatric Association, Washington DC, 2000, pp. 544–546.
[28] M. Hamilton, A rating scale for depression, J. Neurol. Neurosurg. Psychiatry 23 (1960) 56.
[29] S.A. Montgomery, M. Asberg, A new depression scale designed to be sensitive to change, Br. J. Psychiatry 134 (1979) 382–389.
[30] W.W. Zung, C.B. Richards, M.J. Short, Self-rating depression scale in an outpatient clinic: further validation of the SDS, Arch. Gen. Psychiatry 13 (1965) 508.
[31] Dictionary.com LLC, Find Synonyms and Antonyms of Words at Thesaurus.com, 2013 http://thesaurus.com/ (last visited: 11.04.13).
[32] Dictionary.com LLC, Dictionary.com – DevCenter API, 2013 http://content.dictionary.com/api (last visited 04.11.13).
[33] S. Greenbaum, The Oxford English Grammar, Oxford University Press Oxford, 1996.
[34] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, S.M. Winkler, S. Schaller, A. Holzinger, On text preprocessing for opinion mining outside of laboratory environments, in: Active media technology, Springer, 2012, pp. 618–629.
[35] The Stanford Natural Language Processing Group, The Stanford NLP (Natural Language Processing) Group, 2013 http://nlp.stanford.edu/software/corenlp.shtml (last visited: 04.04.13).
[36] R.C. Hsiung, Psycho-Babble, 2007 http://www.dr-bob.org/babble/ (last visited: 04.08.13).
[37] R.C. Hsiung, The best of both worlds: an online self-help group hosted by a mental health professional, CyberPsychol. Behav. 3 (2000) 935–950.
[38] A. Holzinger, P. Yildirim, M. Geier, K.-M. Simonic, Quality-based knowledge discovery from medical text on the web, in: Quality Issues in the Management of Web Information, Springer, 2013, pp. 145–158.
[39] M. Kreuzthaler, M. Bloice, K.-M. Simonic, A. Holzinger, Navigating through very large sets of medical records: an information retrieval evaluation architecture for non-standardized text, Springer, 2011.
[40] J. Cohen, et al., A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[41] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.
[42] E.B. Wilson, Probable inference, the law of succession, and statistical inference, J. Am. Stat. Assoc. 22 (1927) 209–212.
[43] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, J. Doc. 60 (2004) 503–520.
[44] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF* IDF, LSI and multi-words for text classification, Expert Syst. Appl. 38 (2011) 2758–2765.
[45] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C.E. Leonard, J.H. Holmes, Identifying potential adverse effects using the web: a new approach to medical hypothesis generation, J. Biomed. Inform. 44 (2011) 989–996.
[46] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (1995) 39–41.