Identification of Depression Strength for Users of Online Platforms: A Comparison of Text Retrieval Approaches

Ayan Bandyopadhyay[1], Linda Achilles[2], Thomas Mandl[2], Mandar Mitra[1], and Sanjoy Kr. Saha[3]

[1] Indian Statistical Institute, India
bandyopadhyay.ayan@gmail.com
mandar@isical.ac.in
[2] University of Hildesheim, Germany
mandl@uni-hildesheim.de
achilles@uni-hildesheim.de
[3] Jadavpur University, India
sks_ju@yahoo.co.in

Abstract. Social media became one of the most popular platform to express feelings and thoughts in the world of digital information sharing. Facebook, Snapchat, Instagram, QQ, Weibo, Twitter, Tumblr, Reddit and LinkedIn are among the most popular social networks. They are used to share, spread and create new information, receive and spread news locally, globally or privately. Many citizens share their feelings and thoughts in social media, consequently mining of emotions and psychological states from social media posts has become an active research area. In the CLEF 2019 eRisk task 3, the goal is to detect how strong a user of social media is suffering from depression. The ground truth is obtained by asking persons a set of standardised questions. This paper shows how a variety of ad-hoc retrieval approaches can be adopted to perform this task. The results do not reach a high level of accuracy, but compare to supervised classification approaches. In the discussion section, the adequacy of measures for the task is reflected.

Keywords: Text Classification · Depression Detection · Social Media · Information Retrieval.

1 Introduction

The classification of text documents has seen great progress in recent years. Meanwhile research is approaching complex problems like gender attribution,

content reliability as well as different quality attributes of text (e.g. helpfulness [7] [27]). The advances in deep learning technologies have contributed to the expansion of classification tasks. Word embeddings as a latent model of the content of words are representations which are learned by a system during the processing. The training items are constructed typically as n-grams of words of subsequent text. Word embeddings as a representation model have often achieved very good results in recent years. One assumption behind many computation tasks in the psychological domain is that text tells a lot about the writer. Consequently, the prediction of psychological traits of people based on text has become an important research area. The base is often a collection of texts from social media due to the large amount of text that can be found and the ease of availability. Researchers have tried to predict the personality of a person based on the Big-5 model ([5]. More recently, the prediction of mental health issues has been seen as a task for classification systems. First collections have been developed for analysis (e.g. [26]). The eRisk task (Early risk prediction on the Internet) at the Conference and Labs of the Evaluation Forum (CLEF) became a venue for comparative analysis of depression detection. In 2019, eRisk moved to predicting the level of depression of persons based on their social media postings. This paper reports on heterogeneous experiments for this task and reviews some technologies for depression detection. Often, there are few data samples available due to the high level of the required confidentiality. As a consequence, we test mainly methods based on string similarity and matching techniques instead of supervised approaches.

2 Related Work

2.1 Depression and depression detection

Traditionally, depression is diagnosed in a therapy in which a therapist checks whether depression symptoms appear during a period of time in the behavior of the patient or not. These symptoms are, for instance, described in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) [2]. The current fifth edition replaces the now outdated fourth edition.
Another instrument in this field is the *Beck's Depression Inventory* (BDI) [9]. The BDI is a questionnaire consisting of 21 questions assessing the patient's mental state regarding feelings like sadness, pessimism, loss of energy and similar. The following example shows the first question of the BDI:

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

A different questionnaire was developed by Radloff [22]. It consists of 20 questions, dealing with the frequency of various symptoms of depression. This ques-

tionnaire is called the *CES-D Scale* (Center for Epidemiologic Studies Depression Scale). This self-report depression scale has been revised in 2004 (DESD-R) [12]. Instead of relying on self-report, Eichstaedt et al. [13] used medical codes from an electronic medical report (EMR) of a patient to establish the depression diagnosis [13]. The researchers then analysed the patients' Facebook posts that were created before the diagnosis in the EMR. Besides the textual post content, they also used the post length, the frequency of posting, the temporal posting patterns, as well as the demographic information to predict the future diagnosis of depression in the EMR. Overall, language features outperformed all other features considered. They could also show, that their approach resulted in a prediction accuracy comparable to validated self-report depression scales.

The examples above show that getting meaningful data can be a difficult and time, labor and cost consuming task, which also relates to the sensitivity of the topic. This becomes apparent in the study of Eichstaedt and colleagues, for which they asked 11,224 patients of an emergency department of a hospital of which only 1,175 agreed to participate fully in the study [13]. However, Shen et al. made the point that the DSM, for instance, took over a decade to evolve from fourth to fifth edition and is so relatively slow in updating depression criteria, especially those that are conveyed by the behavioral patters in social media [26]. Automatically analyzing the online behavior and language on social media therefore can help in early detection of mental disorders like for instance depression.

2.2 Early risk prediction on the Internet (eRisk)

The eRisk task is an evaluation lab as part of the CLEF initiative. Its main objective is to examine evaluation methodologies, effectiveness and performance metrics, as well as practical applications and the building of test collections related to early risk detection on the internet. Technologies that can detect disorders at an early stage can be applied to variety of different cases and can be especially useful in those associated with safety and health. For instance, notifications can be sent when sex offenders start interacting with children. Besides potential paedophiles other examples encompass stalkers, or persons with suicidal thoughts or those with tendencies to depression or other mental disorders [16].
In 2018, two tasks were organized by the lab: 1) Early Detection of Signs of Depression and 2) Early Detection of Signs of Anorexia. The lab in 2019 organized three tasks: 1) Early Detection of Signs of Anorexia (continuation of eRisk 2018's T2 task), 2) Early Detection of Signs of Self-harm (this is a new task in 2019) and 3) Measuring the Severity or Strength of the Signs of Depression (this is a new task in 2019).
The test collections for task one and two of both years have the same format as described in the overview paper [15]. They consist of writings (post and comments) from social media authors.
For evaluating the performance of the systems in the tasks, standard measures

like F_1, Precision and Recall have been used. They do not take the decision making time into account, so that the organizers proposed the ERDE (early risk detection error) measure [15]. Early detection is rewarded, meaning the fewer posts required to detect e.g. anorexia the better the system is considered to be. The measure is parameterised to control the place in the X axis where the cost (the delay in detecting true positives) grows more quickly. $ERDE_5$ therefore is very demanding with decision delays, because if a system needs more than 5 writings the value for $ERDE_5$ quickly decreases. However, $ERDE_{50}$ is less strict with decision delays [16]. The ERDE measure is in the range [0, 1] [15]. In 2018, the best results for $ERDE_5$ were achieved by flexible temporal variation of terms (FTVT) and sequential incremental classification (SIC) [14]. In case of $ERDE_{50}$ as well as F_1 word embeddings and linguistic metadata led to the best results [28]. The highest precision was achieved by using effective machine learning algorithms (a bag of words model has been used to perform ada boost, random forest, logistic regression and support vector machine classifiers) [20]. Fidel and colleagues obtained the highest recall by applying two independent models (one trained to predict depression cases, the other one to predict non-depression cases) with two variants: Duplex Model Chunk Dependent (DMCD) and Duplex Model Writing Dependent (DMWD) [10].

3 Measuring the Severity or Strength of the Signs of Depression (eRisk 2019 task 3)

The third task in eRisk 2019 is an exploratory new task in eRisk. Participants of the challenge have to build an algorithm that estimates the level of depression of a user based on a history of postings. Depending on these, the participants of the eRisk lab have to fill in the questionnaire BDI for each user. This means that the task consists of predicting how a user would fill in the questionnaire given her or his texts [17].

3.1 Data Set

The data set consists of BDI questionnaires that were filled in by social media users along with each user's history of writings. After submitting the BDI, the user's writings were extracted right after. These original questionnaires are the ground truth data for task 3 and were used to evaluate the performance of the lab participants' systems. The participants were given a data set of 20 social media authors' writing history. They were then asked to develop an algorithm that produces the following structure:

```
username1 answer1 answer2 .... answer21
username2 ....
....
```

Each line identifies the author and the estimated answers to the questions in the BDI. The ground truth data has the same format [17].

3.2 Evaluation Measures

The task employs a variety of evaluation metrics to measure the success of algorithms. Losada et al. [17] define them as follows:

Hit Rate (HR) HR determines how often the prediction was correct, compared to the real questionnaire and gives the ratio. For instance, a prediction where 5 of the 21 questions of the BDI for correct get an HR value of 5/21.

Average Hit Rate (AHR) AHR is HR, but averaged over all users.

Closeness Rate (CR) CR considers the ordinal scale underlying the questions in the BDI. For each question an absolute difference (ad) between the actual answer and the predicted one. A system that is farther away from the answer than a second system should be penalized for this greater distance. For that the measure is build like this:

$$CR = \frac{(mad - ad)}{mad} \tag{1}$$

Here, *mad* stands for the maximum absolute difference (number of possible answers minus one).

Average Closeness Rate (ACR) ACR is CR, but averaged over all users. However, the questions #16 and #18 have seven possibilities to answer, where for answers 1 to 3 two possible options (a and b) are available. However, those options were considered equal, since they represent the same level of depression.

Difference between overall depression levels (DODL) This measure does not take into account the system's correct predictions on question-level, but gives the overall depression level based on the sum of all answers for the real and system generated BDI. Furthermore, the absolute difference (ad overall) between the real and the predicted depression score is calculated.

A depression level is an integer between 0 and 63. These numbers are derived from adding the numbers of the answers from the BDI. For example, considering question #1 (see section 2.1), if the answer was option 1, the depression level integer is raised by 1. This way, the following four categories are associated with the respective depression levels:

1. Minimal depression (depression levels 0-9)
2. Mild depression (depression levels 10-18)

3. Moderate depression (depression levels 19-29)
4. Severe depression (depression levels 30-63)

These levels are widely accepted in the psychological literature [8].

The DODL measure is finally normalized into $[0, 1]$ in the following way:

$$DODL = \frac{(63 - \text{ad overall})}{63} \tag{2}$$

Average DODL (ADODL) ADODL is DODL, but averaged over all users.

Depression Category Hit Rate (DCHR) DCHR computes the fraction of cases, in which the system generated BDI led to the same depression category obtained from the real author's questionnaire.

4 Processing Approaches

We experimented with several heterogeneous ad-hoc information retrieval approaches for depression prediction. That way, a variety of parameter settings can be explored. An important research question is, whether such processing without additional resources can compete with deep learning approaches for a domain with relatively little text volume.

4.1 Ad-hoc Retrieval Approaches

We considered the posts given for each user and the BDI as a document corpus and as traditional ad-hoc information retrieval queries. Each answer of a BDI question is treated as a query. Each set of user posts is treated as a document collection and indexed. This allows to retrieve (compute a query document similarity score) documents and produce the result as quickly as possible. The main concept behind our approach is as follows: The post "p_i" ($i = 1, 2..k$, k is total number of posts by user "u") of an user "u" which is returned with the maximum similarity value for a BDI answer with number $1.j$ (j=0,1,2,3 here. See example query number 1) from a question set "1" determines the answer. For the user "u", "j" is the result of query set 1. In the example, question number 1 is concerned with the concept "sadness", so for user "u" j is the "sadness" label predicted.

This approach allows the use of information retrieval technology for the task. It also enables a completely unsupervised approach which does not require additional resources.

Due to the nature of text on social media microblogs, it seems unclear whether stop word removal and stemming as traditional pre-processing methods are beneficial for the task. Consequently, we conducted experiments with and without both techniques. documents by

– stemming and stop word removal, and

– no stemming and no stop word removal

The following experiments with different retrieval models and parameter settings were carried out with Lucene as the basic search engine:

– TF-IDF
– BM25 [24][25] [23] (3 ISIKol-bm25-1.2-0.75-5000-Dtac-Qtac): BM25 model with parameter settings as follows: $k1 = 1.2$ and $b = 0.75$
– Language Model - Divergence from Randomness with second normalization model (DFR) [4]
– LM-dir (3 ISIKol-lm-d-1.0-5000-Dtac-Qtac): Language model with Dirichlet prior smoothing with $\mu = 1.0$.
– Multi-Similarity (3 ISIKolmultiSimilarity-5000-Dtac-Qtac): This experiment represents a fusion appraoch with the combined sum of a Language model with Dirichlet prior smoothing (LM-d) with $\mu = 1.0$, Language model with Jelinek-Mercer Smoothing (LM-jm) with $\lambda = 0.5$, DFR with second normalization model (DFR) [4] and a BM25 model with $k1 = 1.2$ and $b = 0.75$.

4.2 Deep Representations for Matching

Recently, deep representations based on word embedding have received much attention, in particular for supervised learning. Based on our approach described above, further experiments were done with word embedding representations. For that, we used the word2vec pre-trained model [19][18] and represented a document as a vector using Equation-3. In this case, $\overrightarrow{d}$ is the document vector of document d, $\overrightarrow{w_{id}}$ is the vector for the i^{th} word (or term) from document d.

Equation-4, describes how query and document similarities were calculated. This method was used by Bandyopadhyay et al. [6] in a retrieval approach for tweet classification during natural disasters. In Equation-4 $\overrightarrow{q}$ is the query vector of a query q. $CosSim(\overrightarrow{q}, \overrightarrow{d})$ is cosine similarity of $\overrightarrow{q}$, $\overrightarrow{d}$.

$$\overrightarrow{d} = \sum_{i=1}^{|W_d|} \overrightarrow{w_{id}} \tag{3}$$

$$Sim(q, d) = CosSim(\overrightarrow{q}, \overrightarrow{d}) = \frac{\overrightarrow{q} \cdot \overrightarrow{d}}{||\overrightarrow{q}|| \cdot ||\overrightarrow{d}||} \tag{4}$$

We used Google's pre-trained word2vec vectors[1] and the GloVe pre-trained [21](Table 1) word vectors to compute our document vectors using Equation-3 formula.

Table 1. Results for the experiments with stemming and without stemming as well as with stop word removal and without stop word removal.

Results	no stemming, no stop word removal				stemming, stop word removal			
	AHR	ACR	ADODL	DCHR	AHR	ACR	ADODL	DCHR
BM25 $k1 = 1.2$, $b = 0.75$	29.29%	59.37%	73.02%	25.0%	**32.38%**	60.00%	72.38%	20.0%
TF-IDF	32.14%	63.10%	74.59%	**40.0%**	30.48%	59.76%	71.98%	20.0%
DFR-I(n)-L-2	30.48%	60.16%	73.65%	25.0%	32.10%	60.95%	73.02%	25.0%
LM-d $\lambda = 1.0$	29.52%	61.67%	75.48%	25.0%	31.67%	**61.03%**	74.68%	25.0%
LM-jm $\mu = 0.5$	28.33%	61.11%	74.92%	10.0%	31.43%	60.95%	74.44%	20.0%
Multi Similarity	30.71%	61.75%	74.71%	25.0%	31.14%	60.35%	73.84%	25.0%
Google	23.10%	55.00%	77.22%	05.00%	19.29%	62.38%	70.79%	40.00%
GloVe	25.71%	59.76%	80.24%	30.00%	20.16%	61.35%	76.41%	30.00%

5 Results

This section shows the results of our experiments and compares them to the outcomes of the submitted runs for the task at CLEF eRisk.

The experiment LM-d $\lambda = 1.0$ returns the best value for the measures ACR. BM25 $k1 = 1.2$, $b = 0.75$ is best for ACR and TF-IDF for DCHR. The language model was used for experiments with query expansion (QE). In Table-2 query expansion results are given. In Table-2 "D"= number of top docs used in QE. "T"= number of top terms used in QE and "RM3" [3] = value of qmix used in RM3 QE.

Table 2. Experiments with RM3 query expansion based on the baseline LM-d model.

Results			AHR	ACR	ADODL	DCHR
D	T	RM3				
10	10	0.5	30.48%	60.87%	71.19%	30.0%
20	10	0.3	30.00%	60.71%	70.87%	20.0%
20	10	0.7	31.43%	61.59%	72.38%	20.0%
20	10	0.9	31.90%	61.51%	74.21%	30.0%
20	15	0.9	31.90%	61.90%	74.76%	**35.0%**
20	20	0.9	**32.38%** **+7.9%**	**61.98%** **+6.9%**	74.68% +0.7%	**35.0%** **+40.0%**
30	10	0.9	31.90%	61.51%	74.21%	30.0%

Table 3. Results of participants in the submitted runs for the task.

Run	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	34.05%	66.43%	77.70%	25.00%
BiTeM	32.14%	62.62%	72.62%	25.00%
CAMHGPTnearestunsupervised	23.81%	57.06%	81.03%	45.00%
CAMHGPTsupervised.181features.58hr	35.47%	68.33%	75.63%	20.00%
CAMHGPTsupervised.769features.55hr	36.43%	67.22%	72.30%	20.00%
CAMHGPTsupervised.949features.75hr	36.91%	69.13%	75.63%	15.00%
CAMHLIWCsupervisedSVM	35.95%	66.59%	75.48%	25.00%
Fazl	22.38%	56.27%	72.78%	5.00%
Illinois	22.62%	56.19%	66.35%	40.00%
ISIKolmultiSimilarity-5000-Dtac-Qtac	**29.76%**	**57.94%**	74.13%	**25.00%**
ISIKol-bm25-1.2-0.75-5000-Dtac-Qtac	**29.76%**	**57.06%**	72.78%	**25.00%**
ISIKol-lm-d-1.0-5000-Dtac-Qtac	**30.00%**	**57.94%**	73.02%	**15.00%**
Kimberly	38.33%	64.44%	66.19%	20.00%
UNSLA	37.38%	67.94%	72.86%	30.00%
UNSLB	36.93%	70.16%	76.83%	30.00%
UNSLC	41.43%	69.13%	78.02%	40.00%
UNSLD	38.10%	67.22%	78.02%	30.00%
UNSLE	40.71%	71.27%	80.48%	35.00%

Table-2 shows, that these experiments show slightly better results.

6 Discussion

The results of our experiments are not far behind the supervised approaches submitted at CLEF. This shows that straightforward approaches using only IR technologies currently perform almost as good as advanced algorithms.
The measure DODL and ADODL need to be interpreted with care. They are a very useful measure as they consider the depression level of one user overall. However, it can even out bad results from individual questions. An approach to trick ADODL would give results in the middle of the answer range. In this case, ADODL would be 50 per cent for an even distribution. Consider that this would be better then all submitted experiments which have higher (worse) values. For an uneven or highly skewed distribution, even better (lower) values could be obtained by appropriate guessing. In a realistic scenario, such a classification would probably need to find out the few cases with depression from many users. In such a case, the set of individual with and without depression are likely to be highly imbalanced. This needs to be taken into consideration when developing classifiers for realistic scenarios.

7 Conclusion

Traditional IR methods including query expansion do not perform best for the eRisk depression severity detection. However, the performance is not much worse

when compared to the submitted runs.
In order to improve performance, we need to further analyze why IR methods are not doing well. One of the reason might be the BDI question length. Average question length is 8.45 (in words) when no stemming is used or no stop words are removed. When we remove stop words and stems (porter) BDI query, the average query length becomes 3.57 (in words).
There are many directions for future research. It is necessary to obtain on the one hand a better understanding of the models for professionals in the field and reach some sort of transparency for them. The type of transparency and how it can be reached is a new research area. Maybe the performance of different sub-classes of depression can be a first step towards that goal.
On the other hand, experts need to be able to feed their expertise into the systems and improve their performance. The society overall needs to find ethical ways to handle such technology. It seems important that citizens are more aware of the information they are providing to readers by writing online text which can be analyzed easily. Basically, they might reveal much about their psychological traits without being aware of it. One important tool would be a classifier available to everyone, such that citizens can test the predictions gained from their texts. This gives users back some of their informational autonomy.

8 Acknowledgements

This work was carried out during a stay of the first author at the University of Hildesheim in Germany. The work was partially sponsored by the federal state Niedersachsen and the Institute of Information Science and Language Technology (IWIST) at the University of Hildesheim.

References

1. Google pre-trained word vector https://code.google.com/archive/p/word2vec/
2. American Psychiatric Association: Diagnostic and statistical manual of mental disorders : DSM-5. American Psychiatric Association Arlington, VA, 5th ed. edn. (2013)
3. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, O., Larkey, L., Li, X., Smucker, M.D., Wade, C.: Umass at trec 2004: Novelty and hard. In: Proceedings of TREC-13 (2004)
4. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. **20**, 357–389 (October 2002). https://doi.org/http://doi.acm.org/10.1145/582415.582416, http://doi.acm.org/10.1145/582415.582416
5. Bai, S., Hao, B., Li, A., Yuan, S., Gao, R., Zhu, T.: Predicting big five personality traits of microblog users. vol. 1, pp. 501–508 (11 2013). https://doi.org/10.1109/WI-IAT.2013.70
6. Bandyopadhyay, A., Ganguly, D., Mitra, M., Saha, S.K., Jones, G.J.: An Embedding Based IR Model for Disaster Situations. Information Systems Frontiers **20**(5), 925–932 (October 2018). https://doi.org/10.1007/s10796-018-9847-6

7. Basu, M., Ghosh, S., Ghosh, K.: Overview of the FIRE 2018 track: Information retrieval from microblogs during disasters (irmidis). pp. 1–5 (12 2018). https://doi.org/10.1145/3293339.3293340
8. Beck, A.T., Steer, R.A., Carbin, M.G.: Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. Clinical psychology review **8**(1), 77–100 (1988)
9. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. Archives of general psychiatry **4**(6), 561–571 (1961)
10. Cacheda, F., Fernandez, D., Novoa, F.J., Carneiro, V.: Analysis and experiments on early detection of depression. In: Cappellato et al. [11], pp. 10–21, http://ceur-ws.org/Vol-2125/paper_69.pdf
11. Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. No. 2125 in CEUR Workshop Proceedings, Avignon (2018), http://ceur-ws.org/Vol-2125/
12. Eaton, W.W., Smith, C., Ybarra, M., Muntaner, C., Tien, A.: Center for Epidemiologic Studies Depression Scale: review and revision (CESD and CESD-R). (2004)
13. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences **115**(44), 11203–11208 (2018)
14. Funez, D.G., Ucelay, M.J.G., Villegas, M.P., Burdisso, S.G., Cagnina, L.C., Montes-y Gómez, M., Errecalde, M.L.: UNSLs participation at eRisk 2018 Lab. In: Cappellato et al. [11], http://ceur-ws.org/Vol-2125/paper_137.pdf
15. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 28–39. Springer (2016)
16. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In: Cappellato et al. [11], http://ceur-ws.org/Vol-2125/invited_paper_1.pdf
17. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019)
18. Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: NAACL HLT 2013 (2013)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proc. NIPS '13. pp. 3111–3119 (2013)
20. Paul, S., Kalyani, J.S., Basu, T.: Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In: Cappellato et al. [11], pp. 10–21, http://ceur-ws.org/Vol-2125/paper_182.pdf
21. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
22. Radloff, L.S.: The CES-D scale: A self-report depression scale for research in the general population. Applied psychological measurement **1**(3), 385–401 (1977)
23. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information

Retrieval. pp. 232–241. SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA (1994), http://dl.acm.org/citation.cfm?id=188490.188561
24. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (Apr 2009). https://doi.org/10.1561/1500000019, http://dx.doi.org/10.1561/1500000019
25. Robertson, S., Zaragoza, H., Taylor, M.: Simple bm25 extension to multiple weighted fields (2004)
26. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.S., Zhu, W.: Depression detection via harvesting social media: A multimodal dictionary learning solution. In: IJCAI. pp. 3838–3844 (2017)
27. Trienes, J., Balog, K.: Identifying unclear questions in community question answering websites. CoRR **abs/1901.06168** (2019), http://arxiv.org/abs/1901.06168
28. Trotzek, M., Koitka, S., Friedrich, C.M.: Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia. In: Cappellato et al. [11], http://ceur-ws.org/Vol-2125/paper_68.pdf