A Test Collection for Research on Depression and Language Use

David E. Losada[1(✉)] and Fabio Crestani[2]

[1] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
david.losada@usc.es

[2] Faculty of Informatics, Università della Svizzera italiana, Lugano, Switzerland
fabio.crestani@usi.ch

Abstract. Several studies in the literature have shown that the words people use are indicative of their psychological states. In particular, depression was found to be associated with distinctive linguistic patterns. However, there is a lack of publicly available data for doing research on the interaction between language and depression. In this paper, we describe our first steps to fill this gap. We outline the methodology we have adopted to build and make publicly available a test collection on depression and language use. The resulting corpus includes a series of textual interactions written by different subjects. The new collection not only encourages research on differences in language between depressed and non-depressed individuals, but also on the evolution of the language use of depressed individuals. Further, we propose a novel early detection task and define a novel effectiveness measure to systematically compare early detection algorithms. This new measure takes into account both the accuracy of the decisions taken by the algorithm and the delay in detecting positive cases. We also present baseline results with novel detection methods that process users' interactions in different ways.

1 Introduction

Citizens worldwide are exposed to a wide range of risks and threats and many of these hazards are reflected on the Internet. Some of these threats stem from criminals such as stalkers, mass killers or other offenders with sexual, racial, religious or culturally related motivations. Other worrying threats might even come from the individuals themselves. For instance, depression may lead to an eating disorder such as anorexia or even to suicide.

In some of these cases appropriate action or intervention at early stages could reduce or minimise these problems. However, the current technology employed to deal with these issues is only reactive. For instance, some specific types of risks can be detected by tracking Internet users, but alerts are triggered when the victim makes his disorders explicit, or when the criminal or offending activities are actually happening. We argue that we need to go beyond this late detection technology and foster research on innovative early detection solutions.

N. Fuhr et al. (Eds.): CLEF 2016, LNCS 9822, pp. 28–39, 2016.
DOI: 10.1007/978-3-319-44564-9_3

Depression is a health problem that severely impacts our society. According to the World Health Organisation[1], more than 350 million people of all ages suffer from depression worldwide. Depression can lead to disability, to psychotic episodes, and even to suicide. However, depression is often undetected and untreated [16]. We believe it is crucial to develop tools and to compile data to shed light on the onset of depression.

Language is a powerful indicator of personality, social or emotional status, but also mental health. The link between language use and clinical disorders has been studied for decades [15]. For instance, depression has been associated with linguistic markers such as an elevated use of first person pronouns. Many studies of language and depression have been confined to clinical settings and, therefore, to analysing spontaneous speech or written essays. A stream of recent work has come from the area of Text and Social Analytics, where a number of authors have attempted to predict or analyse depression [3,4,13,14]. Some of them proposed innovative methodologies to gather textual contents shared by individuals diagnosed with depression. However, there are not publicly available collections of textual data. This is mainly because text is often extracted from social networking sites, such as Twitter or Facebook, that do not allow re-distribution. Another limitation of previous studies is that the temporal dimension has often been ignored. We strongly believe that tracking the evolution of language is crucial and, therefore, a proper sample collection strategy, which facilitates studying depression over time, should be designed.

In this paper we make four main contributions. First, we describe the methodology that we have applied to build a collection of text to foster research on the characteristics of the language of depressed people and its evolution. This methodology could be adopted by others to build collections in similar areas (for example, offensive or deceptive language). Second, we sketch the main characteristics of the collection and encourage other teams to use it to gain insights into the evolution of depression and how it affects the use of language. Third, we propose an early detection task and define a novel effectiveness measure to systematically compare early detection algorithms. This new measure takes into account both the accuracy of the decisions taken by the algorithm and the delay in detecting positive cases. Risk detection has been studied in other areas–e.g. privacy risks related to user's search engine query history [2] or suicidality risks on Twitter [11]–but there is a lack of temporal-aware risk detection benchmarks in the domain of health disorders. Four, we performed some experiments with baseline techniques and we report here their performance. These experiments provide an initial set of early risk detection solutions that could act as a reference for further studies.

2 Building a Textual Collection for Depression

Some authors have analysed mental health phenomena in publicly available Social Media [5,6,12]. These studies are often confined to understand language

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

differences between people suffering from a given disorder and a control group (e.g., depressed vs non-depressed, bipolar vs non-bipolar). To the best of our knowledge, no one has attempted to build a dataset where a large chronological sequence of writings leading to that disorder is properly stored and analysed. This is precisely our main objective.

Time is a fundamental factor because appropriate action or intervention at early stages of depression can be highly beneficial. We want to instigate research on innovative early detection solutions able to identify the states of those at risk of developing major depression episodes, and want to stimulate the development of algorithms that computationally treat language as a meaningful tracker of the evolution of depression. These challenging aims can only be achieved with the help of solid evaluation methodologies and benchmarks.

The next section presents the data source selection process performed; Sect. 2.2 reports the method employed to extract a group of depressed individuals; Sect. 2.3 explains the method employed to create a control group of non-depressed individuals; Sect. 2.4 gives details on the submissions extracted from each individual; and, finally, Sect. 2.5 reports the main statistics of the collection built.

2.1 Selection of Data Source

We have studied the adequacy of different types of Internet repositories as data sources to create test collections for research on depression and language use. Within this process, the main aspects that we analysed were: (i) the size and quality of the data sources, (ii) the availability of a sufficiently long history of interactions of the individuals in the collection, (iii) the difficulty to distinguish depressed cases from non-depressed cases, and (iv) the data redistribution terms and conditions (this is important to make the collection available to others). The main sources considered were:

Twitter. Most previous works have focused on microblogs and, in particular, tweets. However, tweets provide little context about the tweet writer. It is therefore difficult to determine when a mention of depression is genuine. Another limitation for us is that Twitter is highly dynamic and only allows to retrieve a limited number of previous tweets per user (up to 3200). In many cases, this is only a few weeks of history. Clearly, this is not enough for collecting a sufficiently long history of previous interactions. Besides, Twitter is highly restrictive about data redistribution.

MTV's A Thin Line (ATL). ATL is a social network launched by the MTV channel in 2010. It is a platform designed to empower distressed teenagers to identify, respond to, and stop the spread of digital abuse. Within this campaign, information is given on how a teenager might cope with issues ranging from sexting to textual harassment and cyberbullying. Young people are encouraged to share their stories publicly and they get feedback, help and advise from the website's visitors. On the ATL platform, posted personal stories have 250 characters or less and other

users can rate the story as "over the line" (i.e., inappropriate and rude), "on the line" (could go either way), or "under the line" (nothing to get uptight about). Dinakar and others [7] obtained a set of 7144 stories posted on ATL over a period of three years from 2010 to 2013 (along with their ratings, comments, the age and gender of posters) and analysed teenage distress language. The dataset, which contains no personally identifiable information of its participants, was obtained through a licensing agreement with Viacom (MTV's parent company). We also contacted Viacom, signed a similar agreement and got access to this collection of data. But there are some limitations that prevent us for using it for creating a benchmark. First, the data cannot be redistributed. Second, the subject identifiers are anonymous and untraceable (i.e., no uniquely identifiable) and, therefore, there is no way to obtain a previous history of interactions.

Reddit. Reddit is an open-source platform where community members (*redditors*) can submit content (posts, comments, or direct links), vote submissions, and the content entries are organised by areas of interests (*subreddits*). Reddit has a large community of members and many of the members have a large history of previous submissions (covering several years). It also contains substantive contents about different medical conditions, such as anorexia or depression. Reddit's terms and conditions allow to use its contents for research purposes[2]. Reddit fulfills all our selection criteria and, thus, we have used it for creating the depression test collection. In the following, we explain how we have used Reddit to create the collection.

2.2 Depression Group

A fundamental issue is how to determine subjects that have depression. Some studies, e.g. [4], have resorted to standard clinical depression surveys. But relying on self-reported surveys is a tedious process that requires to individually contact every participant. Besides, the quality and volume of data obtained in this way is limited. Coppersmith et al. [5] opted instead for an automatic method to identify people diagnosed with depression in Twitter. We have adapted Coppersmith et al.'s estimation method to Reddit as follows.

Self-expressions of depression diagnoses can be obtained by running specific searches against Reddit (e.g. "I was diagnosed with depression"). Next, we manually reviewed the matched posts to verify that they were really genuine. Our confidence on the quality of these assessments is high because Reddit texts are long and explicit. As a matter of fact, many of the matched posts came from the depression subreddit, which is a supportive space for anyone struggling with depression. It is often the case that redditors go there and are very explicit about their medical condition. Although this method still requires manual intervention, it is a simple and effective way to extract a large group of people that

[2] Reddit privacy policy states explicitly that the posts and comments redditors make are not private and will still be accessible after the redditor's account is deleted. Reddit does not permit unauthorized commercial use of its contents or redistribution, except as permitted by the doctrine of fair use. This research is an example of fair use.

explicitly declare having being diagnosed with depression. The manual reviews were strict. Expressions like "I have depression", "I think I have depression", or "I am depressed" did not qualify as explicit expressions of a diagnosis. We only included a redditor into the depression group when there was a clear and explicit mention of a diagnosis (e.g., "In 2013, I was diagnosed with depression", "After struggling with depression for many years, yesterday I was diagnosed").

2.3 Control Group

This initial set of (depressed) redditors was expanded with a large set of random redditors (control group). Besides random members, we also included in the control group a number of redditors who were active on the depression subreddit but had no depression. There is a variety of such cases but most of them are individuals interested in depression because they have a close relative suffering from depression. These individuals often talk about depression and including them in the control group helps to make the collection more realistic. We cannot rule out the possibility of having some truly depressed individual in the control group, and we cannot rule out the possibility of having some non-depressed individual into the depressed group (an individual's claim about his diagnosis might be false). Still, we expect that the impact of such cases would be negligible and, anyway, other screening strategies (e.g. based on questionnaires) are not noise-free either.

2.4 Texts Extracted

For each redditor, the maximum amount of submissions that we can retrieve is 1000 posts and 1000 comments (Reddit's API limit). We retrieved as many submissions as possible and, therefore, we have up to 2000 submissions from the most active redditors. This included textual contents (posts, comments to posts made by others, links) submitted to any subreddit. Redditors are often active on a variety of subreddits and we collected submissions to any subreddit. We are interested in tracking the redditor's language (regardless of the topic discussed). The collection therefore contains submissions from a wide range of subreddits (e.g., food, videos, news). We organised all these contents in chronological order. The resulting data cover a large time period for most redditors and, thus, enables to study not only the differences in language use between depressed and non-depressed users, but also the evolution of the written text.

We also stored the link to the post where the redditor made the explicit mention to the diagnosis. This information might be useful for further experiments. However, we removed this post from the user's chronology. Otherwise, depression text classifiers would be strongly centered on the specific phrases that we used to manually search for depression diagnosis.

The collection was created as a sequence of XML files, one file per redditor. Each XML file stores the sequence of the redditor's submissions (one entry per submission). Each submission is represented with the submission's title, the submission's text and the submission's date. No other metadata is available.

Table 1. Main statistics of the collection.

	Depressed	*Control*
Num. subjects	137	755
Num. submissions (posts & comments)	49,580	481,873
Avg num. of submissions per subject	361.9	638.2
Avg num. of days from first to last submission	578.3	625.3
Avg num. words per submission	27.4	36.7

Regarding diagnosis dates, we have a variety of cases. Sometimes, the diagnosis is recent (e.g. "Yesterday", "This week") and, therefore, most of the messages retrieved are *pre-diagnosis*. Other times, the diagnosis was a long time before ("In 2010", "3 years ago") and, therefore, most of the redditor's text is *post-diagnosis*. In other cases, retrieved texts contain both *pre-diagnosis* and *post-diagnosis* submissions. There is often some degree of uncertainty about the specific date of the diagnosis but this approximate information about the diagnosis date is still valuable and can be potentially used in a variety of ways.

The retrieval of submissions was done with Reddit's Python API[3] and all redditors with less than 10 submissions were removed, as we think there would be not enough history to be able to track the evolution of the depression.

2.5 Resulting Collection

The statistics of the resulting collection are reported in Table 1[4]. Following our strategy, we have been able to collect a reasonably high number of subjects and a large number of submissions. The average period of time between the redditor's first submission and the redditor's last submission covers more than a year. There is a high variance in the length of the submissions. Some submissions are short replies to an existing post (comments), while other submissions–typically posts–are lengthy. The average submission length is relatively low (around 30 words after pre-processing) because the number of submitted comments is higher than the number of submitted posts.

We have the firm intention to support research on these topics. The collection is available for research purposes under proper user agreements[5].

3 Early Prediction Task

In this section we present a task of detection of early traces of depression and propose a new metric to measure the effectiveness of early alert systems. Of course, these systems can never become substitutes of trained medical practitioners and, additionally, the widespread adoption of technologies for analysing

[3] https://praw.readthedocs.org/en/v3.1.0/.

[4] The number of terms per submission are counted after pre-processing the texts with the scikit-learn Python toolkit, *scikit-learn.org*. This was configured with no stopword processing and no vocabulary pruning based on document frequency.

[5] http://tec.citius.usc.es/ir/code/dc.html.

health-related publicly shared data has to be dictated by a legal framework. Still, we think it is important to bring the possibilities of such predictive technologies to the front and stimulate discussion on their role in enhancing public health.

The challenge consists of sequentially processing pieces of evidence and detect risk cases as soon as possible. Texts should be processed in the order they were created. In this way, we can simulate systems that monitor social media evidence as it appears online. Let us consider a corpus of documents written by p different individuals ($\{I_1, \ldots, I_p\}$). For each individual I_l ($l \in \{1, \ldots, p\}$), the n_l documents that he has written are provided in chronological order (from the oldest text to the most recent text): $D_{I_l,1}, D_{I_l,2}, \ldots, D_{I_l,n_l}$. Given these p streams of messages, we define the following early risk detection task:

- An early risk detection system (ERDS) has to process every sequence of messages (following the order in which the messages are produced). At some point k ($k \in \{1, \ldots, n_l\}$) the system has to make a binary decision on whether or not the individual might be a positive case of depression.
- It is desirable to detect positive cases as soon as possible. But there is a tradeoff between making early decisions and making *more informed* decisions (as we gain more evidence on the subjects, the system's estimations can be more accurate).

This task can be regarded as a new form of data stream classification where systems not only have to assign a class for the stream, but also have to decide when to make the assignment.

3.1 Evaluation Metric

Standard classification measures, such as the F-measure, could be employed to assess the system's output with respect to golden truth judgments that inform us about what subjects are really positive cases. However, standard classification measures are time-unaware and, therefore, we need to complement them with new measures that reward early alerts.

An early risk evaluation metric needs to take into account the correctness of the (binary) decision and the delay taken by the system to make the decision. The delay is measured here by counting the number (k) of distinct textual items seen before giving the answer. Another important factor is that, in many application domains, data are unbalanced (many more negative cases than positive cases). Hence, we also need to weight different errors in a different way.

Let us consider a binary decision d taken by a ERDS at point k. Given golden truth judgments, the prediction d can lead to one of the following cases: true positive (TP), true negative (TN), false positive (FP) or false negative (FN). Given these four cases, we propose and *early risk detection error* (ERDE) measure defined as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d = \text{positive AND ground truth} = \text{negative (FP)} \\ c_{fn} & \text{if } d = \text{negative AND ground truth} = \text{positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d = \text{positive AND ground truth} = \text{positive (TP)} \\ 0 & \text{if } d = \text{negative AND ground truth} = \text{negative (TN)} \end{cases}$$

How to set c_{fp} and c_{fn} depends on the application domain and the implications of FP and FN decisions. We will often face detection tasks where the number of negative cases is several orders of magnitude greater than the number of positive cases. Hence, if we want to avoid building trivial classifiers that always say no, we need to have $c_{fn} >> c_{fp}$. For instance, we can fix c_{fn} to 1 and set c_{fp} according to the proportion of positive cases in the data (e.g. if the collection has 1 % of positive cases then we set c_{fp} to 0.01). The factor $lc_o(k)(\in [0,1])$ encodes a cost associated to the delay in detecting true positives. In domains where late detection has severe consequences we should set c_{tp} to c_{fn} (i.e. late detection is equivalent to not detecting the case at all). The function $lc_o(k)$ should be a monotonically increasing function of k. Inspired by the TREC temporal summarization track [1], which incorporated a latency discount factor (sigmoid function) to penalize late emission of relevant sentences, we propose the following cost function that grows with k:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{1}$$

The function is parameterised by o, which controls the place in the X axis where the cost grows more quickly (Fig. 1 plots $lc_7(k)$ and $lc_{20}(k)$).

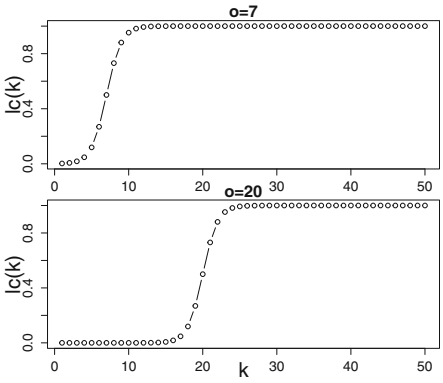

Fig. 1. Latency cost functions: $lc_7(k)$ and $lc_{20}(k)$

Observe that the latency cost factor was introduced only for the true positives. We understand that late detection is not an issue for true negatives. True negatives are non-risk cases that, in practice, would not demand early intervention. They just need to be effectively filtered out from the positive cases. Algorithms should therefore focus on early detecting risk cases and detecting non-risk cases (regardless of when these non-risk cases are detected).

According to the formulas above, if all cost weights are in $[0,1]$ then ERDE would also be in the range $[0,1]$. Since we have p unique individuals in the

collection, systems would have to take p decisions (one for each subject, after analysing the subject's stream of submissions). The overall error would be the mean of the p ERDE values.

4 Baseline Experiments

We implemented several relatively simple early detection strategies and ran a series of experiments to evaluate their performance. These experiments aim to provide a pool of depression detection solutions that could be used by others as a reference for comparison.

First, we randomly split the collection into a training and a test set. The training set contained 486 users (83 positive, 403 negative) and the test set contained 406 users (54 positive, 352 negative). Some of the methods described below require a training stage, which consists of building a depression language classifier. To meet this aim, each training user was represented with a single document, consisting of the concatenation of all his writings. After vectorising these 486 documents[6], we built a depression language classifier as follows. We considered a Logistic Regression classifier with L1 regularisation as our reference learning method. This classification approach, which simultaneously selects variables and provides regularisation, has state-of-the-art effectiveness on a range of text categorisation tasks [8]. The resulting models are sparse (many variables are assigned a weight equal to 0). This improves human interpretability and reduces computational requirements at prediction time. Furthermore, this type of sparse models has shown to be superior to other regularised logistic regression alternatives [8].

We optimised the penalty parameter, C ($C > 0$), and the class weight parameter w ($w \geq 1$). C is the parameter associated to the error term of the optimisation formula of the L1-penalised Logistic Regression classifier. C controls the trade-off between the training error and the complexity of the resulting model. If C is large we have a high penalty for training errors and, therefore, we run the risk of overfitting. If C is small we may instead underfit. Another important issue is that our classification problem is unbalanced. When dealing with unbalanced problems, discriminative algorithms may result in trivial classifiers that completely ignore the minority class [10]. Adjusting the misclassification costs is a standard way to deal with this problem. We set the majority class ("non-depression") weight to $1/(1+w)$ and the minority class ("depression") weight to $w/(1+w)$. If $w = 1$ both classes have the same weight (0.5). As w grows, we give more weight to the minority class and, therefore, the learner will penalise more the errors of classifying a depression case as a non-depression case. Following standard practice [9], we applied a grid search on the tuning parameters, with exponentially growing sequences ($C = 2^{-10}, 2^{-4}, ..., 2^9$ and $w = 2^0, 2^1, ..., 2^9$). Model selection was done by 4-fold cross-validation on the training data (optimising F1 computed with respect to the minority class). $C = 16$ and $w = 4$ was the parameter configuration with the

[6] We employed sklearn library, version 0.16.1, for Python. Vectorisation was done with the TfidfVectorizer–with a standard stoplist and removing terms that appear in less than 20 documents–and classification was done with the LogisticRegression class.

highest F1 (avg 4-fold performances: $F1 = .66$, $Precision = .65$, $Recall = .67$). We finally proceeded to fix this parameter setting and built a depression language classifier from the whole training data.

We experimented with different strategies to process the stream of texts written by each user in the test split. Some strategies employ the depression language classifier described above and other strategies do not require a text classifier. More specifically, we implemented and tested the following methods:

- **Random**. This is a naïve strategy that emits a random decision ("depression"/"non-depression") for each user. It does not use the depression language classifier and it emits its random decision right after seeing the first submission from every user[7]. This method is therefore fast–delay equal to 1–but we expect it to have poor effectiveness. We include it here as a baseline for comparison.
- **Minority**. This is another naïve strategy that emits a "depression" decision for each user. It does not use the depression language classifier and it also emits its decision right after seeing the first submission from every user[8]. This method is also fast–delay equal to 1–but we expect it to have poor effectiveness. Observe that we do not include here the alternative strategy (majority, always "non-depression") because it does not find any depression case and, therefore, it would score 0 on all our effectiveness metrics.
- **First** n. This method consists of concatenating the first n texts available from each user (first n submissions written by the subject) and making the prediction–with the depression language classifier–based on this text. The delays are therefore fixed to n. If n is larger than the maximum number of submissions per user then the strategy is gonna be slow (it waits to see the whole sequence of submissions for every user) but it makes the decisions with all the available data (we label this particular instance as "All" in Table 2).
- **Dynamic**. The dynamic method does not work with a fixed number of texts for each user. Instead, it incrementally builds a representation of each user, passes this text to the depression language classifier, and only makes a "depression" decision if the depression language classifier outputs a confidence value above a given threshold (thresholds tested: 0.5, 0.75 and 0.9). Otherwise, it keeps concatenating more texts. If the stream of user texts gets exhausted then the dynamic method concludes with a "non-depression" decision.

Not surprisingly, random and minority are the worst performing methods in terms of F1 and ERDE. The fixed-length strategies score well in terms of F1 but their ERDE results show that they are perhaps too slow at detecting positive cases. The dynamic methods, instead, can make quicker decisions. Overall, the

[7] This strategy does not make any text analysis and, therefore, it does not make sense to wait any longer to make the decision.

[8] Again, this strategy does not make any text analysis and, therefore, it does not make sense to wait any longer to make the decision.

Table 2. Early risk classifiers

	F1	P	R	$ERDE_5$	$ERDE_{50}$
Random	.19	.12	.48	13.0 %	13.0 %
Minority	.23	.13	1	11.6 %	13.0 %
First 10	.31	.50	.22	11.1 %	10.9 %
First 100	**.62**	.64	.59	7.0 %	6.7 %
First 500	**.62**	.59	.65	6.6 %	6.2 %
All	.59	.55	.65	6.7 %	6.4 %
Dynamic 0.5	.53	.40	.78	**6.0 %**	**5.3 %**
Dynamic 0.75	.58	.57	.59	7.0 %	6.6 %
Dynamic 0.9	.56	.71	.46	8.1 %	7.8 %

results suggest that if we are only concerned about the correctness of the decisions (F1 measure) then we should go for a fixed-length strategy that analyses the first 100/500 messages. However, this fixed-length strategy is suboptimal in terms of ERDE. The dynamic method with the default threshold (Dynamic 0.5) is the best performing method when we want to balance between correctness and time. Anyway, there is substantial room for improvement and we expect that these results instigate others to design innovative and more effective early detection solutions.

5 Conclusions

In this paper, we presented a new test collection to foster research on depression and language use. We have outlined the methodology followed to build a test collection that includes a series of textual interactions written by depressed and non-depressed individuals. The new collection not only encourages research on differences in language between depressed and non-depressed people, but also on the evolution of the language use of depressed users.

We started working on a suitable evaluation methodology to accompany the collection. We also started working on baseline methods to detect early symptoms of depression and we provided an initial report on the effectiveness of these preliminary solutions.

Acknowledgements. This research was funded by the Swiss National Science Foundation (project "Early risk prediction on the Internet: an evaluation corpus", 2015). The first author also thanks the financial support obtained from "Ministerio de Economía y Competitividad" of the Goverment of Spain and FEDER Funds under the research project TIN2015-64282-R.

References

1. Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., Sakai, T.: TREC temporal summarization track overview. In: Proceedings of the 23rd Text Retrieval Conference, Gaithersburg (2014)
2. Biega, J., Mele, I., Weikum, G.: Probabilistic prediction of privacy risks in user search histories. In: Proceedings of the First International Workshop on Privacy and Security of Big Data, PSBD 2014, pp. 29–36. ACM, New York (2014)
3. Choudhury, M.D., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Davis, H.C., Halpin, H., Pentland, A., Bernstein, M., Adamic, L.A. (eds.) WebSci, pp. 47–56. ACM (2013)
4. Choudhury, M.D., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Kiciman, E., Ellison, N.B., Hogan, B., Resnick, P., Soboroff, I. (eds.) ICWSM. The AAAI Press (2013)
5. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: ACL Workshop on Computational Linguistics and Clinical Psychology (2014)
6. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: CLPsych: depression and PTSD on Twitter. In: NAACL Workshop on Computational Linguistics and Clinical Psychology (2015)
7. Dinakar, K., Weinstein, E., Lieberman, H., Selman, R.L.: Stacked generalization learning to analyze teenage distress. In: Adar, E., Resnick, P., Choudhury, M.D., Hogan, B., Oh, A. (eds.) ICWSM. The AAAI Press (2014)
8. Genkin, A., Lewis, D., Madigan, D.: Large-scale bayesian logistic regression for text categorization. Technometrics **49**(3), 291–304 (2007)
9. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University (2003)
10. Nallapati, R.: Discriminative models for information retrieval. In: Proceeding of ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71 (2004)
11. O'Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H.: Detecting suicidality on Twitter. Internet Interventions **2**(2), 183–188 (2015)
12. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in Twitter. In: 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD) Workshop on Health Informatics (HI-KDD) (2012)
13. Park, M., McDonald, D.W., Cha, M.: Perception differences between the depressed and non-depressed users in Twitter. In: Kiciman, E., Ellison, N.B., Hogan, B., Resnick, P., Soboroff, I. (eds.) ICWSM. The AAAI Press (2013)
14. Paul, M.J., Dredze, M.: You are what you Tweet: analyzing Twitter for public health. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S., (eds.) ICWSM. The AAAI Press (2011)
15. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. **54**(1), 547–577 (2003)
16. Saeb, S., Zhang, M., Karr, C., Schueller, S., Corden, M., Kording, K., Mohr, D.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J. Med. Internet Res. **17**(7), e175 (2015). http://www.jmir.org/2015/7/e175/