Discovering Latent Depression Patterns in Online Social Media

Esteban A. Ríssola
esteban.andres.rissola@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

David E. Losada
david.losada@usc.es
Universidade de Santiago de Compostela (USC)
Santiago de Compostela, Spain

Fabio Crestani
fabio.crestani@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

ABSTRACT

Mental health disorders are a major concern worldwide. However many cases still go undetected. Due to their increasing popularity, online social media sites became promising means to develop innovative methods for detecting such mental disorders. In this work, we present our research towards building automatic early detection systems based on user-generated content. Our experimental results on a real-world dataset reveal evidence that building such systems is viable and can provide promising results.

CCS CONCEPTS

• **Information systems** → **Data stream mining**; • **Applied computing** → *Psychology*.

KEYWORDS

Online mental state assessment, social media, text mining

1 INTRODUCTION

In the last decade, the recognise importance of mental health has motivated the search for cutting-edge and innovative methods for identifying the onset of mental disorders at early stages. Worldwide more than 350 million people, of different ages and communities, suffer from depression[1]. The lack of a timeliness treatment can lead to disability, psychotic episodes, thoughts of self-harm and, at its worst, suicide.

Many useful cues about an individual's mental state (as well as personality, social and emotional conditions) can be discovered by examining the patterns of language use [4]. Previous research shows that language attributes can be indicators of current mental state [14], personality [12] and even personal values [2]. That is because such latent mental-related variables are manifested in the words that individuals use to express themselves. [13].

On a daily basis, people use online social media websites, like Reddit and Twitter, to share their thoughts, feelings and moods. The increasing popularity of these platforms has resulted in new opportunities for innovative methods for detecting different mental disorders [5], like depression [3], anorexia, or similar disorders. In fact, the availability of such user-generated content can enable an exploratory screening process to automatically identify people who might be struggling with psychological disorders, provide a preliminary assessment and, if needed, call for professional action.

[1]Source: https://www.who.int/mental_health/advocacy/en/#Factsheets

IIR 2019, September 16–18, 2019, Padova, Italy

In this paper, we exploit the latent semantic structure in social media users' textual posts to identify individuals potentially at-risk of depression. Semantic structure analysis has been previously applied to determine a reduction in the semantic coherence in patients who suffered from schizophrenia. Such analysis has been shown to achieve a diagnostic accuracy comparable to clinical ratings [7].

Our initial contributions reported in this paper are the following:

- We build a model to measure the semantic proximity between user's textual posts and a set of words with topical relevance to depression and use such information to early identify the onset of depression.
- We perform various experiments with the proposed model on a real-world dataset and analyse the model's performance. In particular, we study the effect of the model's threshold on effectiveness and we explore the temporal spread of the cues that indicate the development of depression.

The remainder of the paper is organised as follows. Section 2 summarises the related work; Section 3 details how the semantic proximity is computed and used to build a model to identify individuals potentially at-risk of depression; Section 4 describes the collection used to conduct the experiments, the evaluation metrics, and the results analysis; Section 5 concludes the paper.

2 RELATED WORK

Of all the many papers dealing with automatic detection of early signs of depression, we list here only those that most directly contributed to our work.

The Early Risk Prediction on the Internet (eRisk) [9, 10] workshop was one of the first initiatives to bring together many researchers to study the interaction between language and mental disorders in online social media. In particular, the organisers proposed to address the early detection of depression in an automatic way and released a corpus of social media users who suffered from depression. The results of the workshop showed that there is a large spectrum of techniques that can be used to detect this psychological disorder and that the vast majority of the workshop participants preferred to use Machine Learning techniques [11]. In this paper, we use the eRisk corpus to undertake the various experiments to assess the performance of our model.

De Choudhury et al. [3] presented an early work on automatic detection of depression using crowd-sourcing to collect assessments from several Twitter users who reported being diagnosed with depression. They built a depression lexicon containing words that are associated with depression and its symptoms. We use this lexicon as one of the tools for creating the set words with topical relevance to depression.

Bedi et al. [1] quantified semantic and structural facets of speech to assess how mental-state changes can be detected after drug induction. They examined a set of transcribed interviews from individuals who had been administered with different drugs and discovered that effectively speech semantic content is affected after drug intoxication. Furthermore, they concluded that such semantic alterations can be accurately used to discriminate between the drugs tested. In this paper, we examine the semantic content of users' posts to identify individuals potentially affected by depression.

3 SEMANTIC PROXIMITY

Meaning can be understood as emerging from mutual dependencies of words within the language. Semantically related words co-occur in texts with coherent topics at a higher frequency than unrelated words. Using this property, the similarity between two words can be quantitatively measured by the frequency of their co-occurrence. Latent Semantic Analysis (LSA) [6] is an associative model that captures the meaning of words by means of linear representations in a high-dimensional semantic space. The semantic content of a word is encoded as a vector and this vectorial representation can be used to estimate how *similar* other words are.

Following [1], we use LSA to compute the similarity of a set of depression-related words with respect to every word in a user's collection of posts. When the similarity is above a threshold of 0.1 it is converted to 1 and when it is below the threshold to 0, producing a binary trace. Subsequently, the mean value of this trace is computed for each depression-related word in the set. We use an LSA model trained on the TASA corpus, that is a collection of educational materials compiled by Touchstone Applied Science Associates. TASA is comprised of general reading texts believed to be common in the US educational system up to college, including a wide variety of short documents from novels, newspapers, and other sources. It includes 37, 651 documents and 12, 190, 931 words, from a vocabulary of 77, 998 distinct words. In particular, we make use of the freely available[2] TASA 4 LSA model developed by Stefanescu et al. [16].

To build the set of depression-related words we use the words from the depression lexicon created in [3]. Additionally, we extend this set by looking for related concepts by running the query with the word "depression" in the well-known lexical database WordNet[3]. The final set is comprised of 96 words. These are all the words closely connected to the concept of "depression", including *anxiety*, *withdrawal*, *delusions*, *blues*, *megrims*, among others.

We use the 96 resulting values from the semantic proximity computation as features to train a Support Vector Machine classifier (SVM). We also consider the total number of words in a user's collection of posts as a feature. The rationale behind this is that previous studies has shown that increases in word count were positively associated with depression [15].

4 EVALUATION

In this section we outline the evaluation framework followed to assess the effectiveness of the model trained on the semantic proximity features. We first describe the dataset used for the experiments, then we describe the performance metrics assessed, and finally we report our results with a preliminary analysis.

4.1 Dataset

To conduct the various experiments we use the eRisk 2017 dataset [9]. This publicly available corpus consists of a set of documents posted by users of Reddit and includes two groups of users, namely depressed and non-depressed. Following the methodology proposed by Coppersmith et al. [5], users of the positive class (*i.e.*, depression) were gathered by retrieving self-expressions of depression diagnoses (*e.g.*, the sentence "I was diagnosed with depression") and verifying if they truly contained a statement of diagnosis. Non-depressed users were collected by randomly sampling from the large set of available users in the platform.

For each user, up to their most recent 2, 000 submissions were retrieved and included in the corpus. In order to make the corpus more realistic, users who were active on the "depression subreddit[4]" but had no depression were included in the non-depressed class. These were mostly people concerned about depression because they had a close relative suffering from the disorder. The resulting corpus comprised of 531, 394 submissions from 887 unique users. A summary of the dataset, including the train/test splits provided by the workshop organisers, is shown in Table 1.

Table 1: Summary of eRisk 2017 dataset.

	Train		Test	
	Positive	Control	Positive	Control
# of Subjects	83	403	52	349
# of Documents	30, 851	264, 172	18, 706	217, 665
Avg. # of Documents/Subject	371.7	655.5	359.7	623.7
Avg. # of Words/Document	27.6	21.3	26.9	22.5
Avg. Activity Period (days)	573.23	627.17	608.8	623.7

4.2 Evaluation Metrics

As it was previously stated, the eRisk 2017 collection is divided into a train and a test split. The train split contains the full history of a set of training users and, for example, it allows you to build predictive technology (*e.g.*, text classifiers) from the entire threads of posts written by the training users. In the test split, each collection of user's posts is divided into ten chunks (in chronological order, based on the time each post was written).

Any system implementing an early risk detection algorithm is given one chunk of user data (oldest posts are given first) and it has to decide whether to classify a user or to wait for the next chunk. Based on this decision, performance is assessed in terms of recall, precision and F_1. Additionally, the organisers developed an error metric called ERDE [8]

which penalises late decisions even when they are correct. As a measure of error, the ultimate goal is to minimise it.

Let U be the set of users in the collection. Let d be a binary decision taken by a system for user u with delay k. Given the

[2] Available at: www.semanticsimilarity.org/
[3] See: https://wordnet.princeton.edu
[4] Titled forums on Reddit are denominated *subreddits*. They include postings on specific topics.

ground truth, d can be one of the following: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) or *False Negative* (FN). Taking into account these four cases, *ERDE* is defined as:

$$ERDE_o(u) = \begin{cases} c_{fp} & \text{if } d \text{ is FP} \\ c_{fn} & \text{if } d \text{ is FN} \\ c_{tp} \cdot lc_o(k) & \text{if } d \text{ is TP} \\ 0 & \text{if } d \text{ is FN} \end{cases}$$

The factor $lc_o(k)$ encodes a cost associated with the delay taken in spotting a TP, and it is defined as a monotonically increasing function of k:

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

The parameter o governs the point in which the cost starts to rapidly increase. Somehow, it represents a sort of "urgency" in detecting positive cases, and the lower the value of o, the higher is the urgency in identifying TP cases.

4.3 Experimental Results

The dataset is built for the time-dependent identification of the early signs of depression. Using the entire history of posts written by each of the training users we trained an SVM based on the 97 features outlined in Section 3. This trained classifier is then used at test time for the detection of depression on each separate chunk of data corresponding to a time slot. When a chunk of data is received, our model classifies a user as depressed only when the estimated class membership probability exceeds a certain threshold. Otherwise, the decision is delayed until the next chunk arrives. The probabilities estimates are calibrated using Platt scaling. In essence, it works by fitting a logistic regression model to the SVM's scores.

We run different experiments to analyse how the value of the threshold affects the performance of the early detection. The values we evaluate range from 0.5 to 0.9. It is noteworthy that, at this initial stage, we use the same threshold for every user, while it is obvious that it should be different. Furthermore, given that the classes in the corpus are not balanced, we study the effect of undersampling the minority class (reducing the size of the non-depressed user set at training time). Table 2 presents the results obtained after conducting the various experiments. A first observation reveals that when either the threshold becomes larger or the number of negative examples in the training set gets smaller recall improves. Conversely, precision diminishes. However, it should be noted that ERDE starts to grow.

As the threshold becomes more conservative and stringent decisions are taken in the latter chunks. This delay is highly penalised by ERDE and highlights the trade-off between taking *early* decisions at the risk of making more mistakes or waiting to receive more data to take more informed decisions. To better understand this, Figure 1 depicts a boxplot based on the number of chunks that a particular configuration needed to take a decision on a certain user. Whereas under a low threshold (such as 0.5) most of the decisions are made in the first chunks, a larger one (such as 0.9) forces the system to wait until the lasts chunks.

Table 2: ERDE and F_1 analysis

Size	$threshold = 0.5$			$threshold = 0.7$		
	$ERDE_5$	$ERDE_{50}$	F1	$ERDE_5$	$ERDE_{50}$	F1
403	**12.62** %	11.67 %	0.19	**13.11** %	**11.67** %	0.32
200	12.63 %	**9.7** %	**0.45**	14.68 %	12.35 %	**0.45**
100	15.22 %	10.85 %	0.34	27.88 %	24.05 %	0.39
50	20.08 %	14.48 %	0.24	28.61 %	22.48 %	0.26

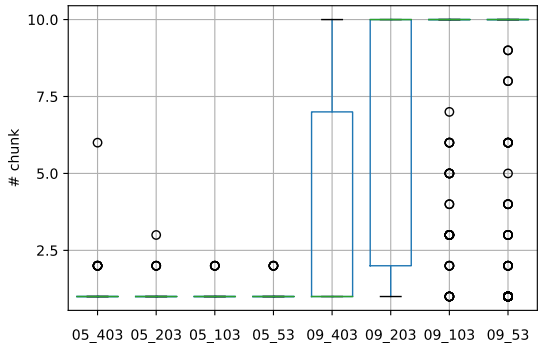

Figure 1: Number of chunks that a particular system configuration needed to take a decision on a certain user (*i.e.*, to classify him/her as depressed). For instance, 05_400 **refers to a configuration where the non-depressed class in the training set comprises** 403 **examples and the threshold is** 0.5**.**

Finally, we explore the temporal spread of the cues that indicate the onset of depression. Figure 2 depicts two examples of this analysis. Each figure presents a comparison of two different users, all of them affected by depression. We observe that the evidence that some users show is very close to the threshold (black dotted-lined) but does not surpass it until a breaking point (blue dots). This means that the user might show some signs of depression, but these are not severe yet and could easily disappear. Therefore, more chunks need to be processed in order to identify the true onset of depression. Conversely, there are other cases where from the processing of the first chunk we can see that there is already strong enough evidence to conclude that the user is rapidly developing depression (orange dots).

These observations suggest that performance could be enhance if the threshold is defined on a user-dependent basis, to capture the very subjective behaviour of each user. For instance, users with similar characteristics could be grouped in order to create different *stereotypes* or *profiles*. In this way, each group of users would have its own threshold. When a new user arrives he/she would be associated with the group that better suits his/her characteristics and the corresponding threshold would be selected . However, different issues need to be addressed in this case, such as how to set the

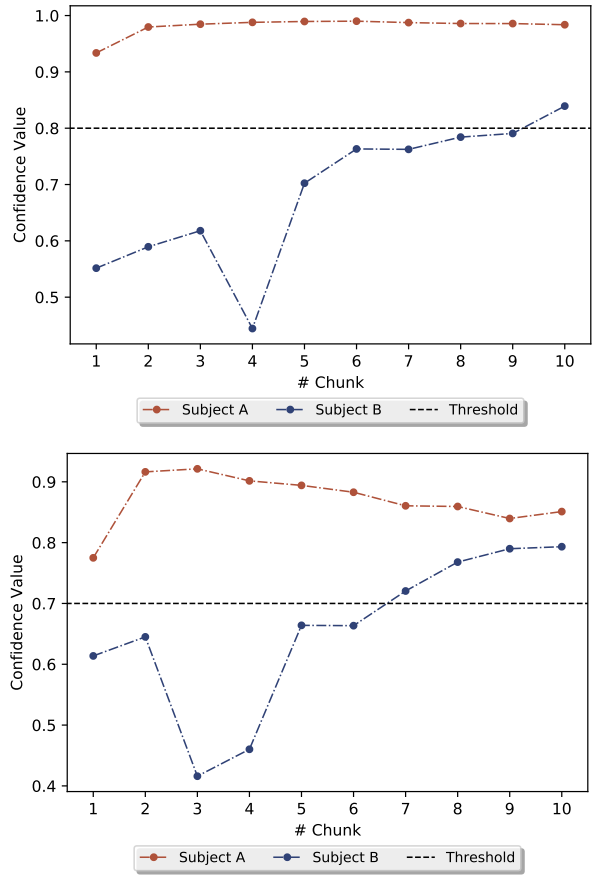

Figure 2: Temporal spread of the cues that indicate the onset of depression. Each figure is comparing different users affected by depression. While for some users the evidence is strong enough already after the first chunk (orange dots), in other cases more information is needed (*i.e*, more chunks) to determine the onset of depression (blue dots).

thresholds for each group of users or how to deal with the cold-start problem. A more thorough analysis on the best way to define the threshold strategy is left as a future work.

5 CONCLUSIONS

By leveraging user-generated content, language-based technologies have a great potential to provide low-cost unobtrusive mechanisms for early screening of mental disorders. In this work, we presented an initial evaluation of how the latent semantic structure of textual posts can be exploited to identify evidence that could suggest the onset of depression. The results of this evaluation highlight the value of developing automatic screening assistants to aid mental health practitioners by providing prognostic information about individuals at-risk of mental disorders, like depression. Moreover, such systems should be customised for each user in order to provide more personalised and precise services.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for the constructive suggestions. This work was supported in part by the Swiss Government Excellence Scholarships and Hasler Foundation.

REFERENCES

[1] Gillinder Bedi, Guillermo A. Cecchi, Diego F. Slezak, Facundo Carrillo, Mariano Sigman, and Harriet de Wit. 2014. A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology* 39, 10 (2014).

[2] Ryan L. Boyd, Steven R. Wilson, James W. Pennebaker, Michal Kosinski, David J. Stillwell, and Rada Mihalcea. 2015. Values in Words: Using Language to Evaluate and Understand Personal Values. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, Oxford, UK.*

[3] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, USA.*

[4] Cindy Chung and James Pennebaker. 2007. The Psychological Functions of Function Words. *Frontiers of social psychology. Social communication* (2007).

[5] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.* Baltimore, USA.

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990).

[7] Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research* 93, 1 (2007).

[8] David E. Losada and Fabio Crestani. 2016. A Test Collection for Research on Depression and Language use. In *Conference Labs of the Evaluation Forum.* Springer.

[9] David E. Losada, Fabio Crestani, and Javier Parapar. 2017. CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. In *Conference and Labs of the Evaluation Forum.* CEUR-WS.org.

[10] David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk Early Risk Prediction on the Internet. In *Conference and Labs of the Evaluation Forum.* CEUR-WS.org.

[11] David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Early Detection of Risks on the Internet: An Exploratory Campaign. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany.*

[12] Yair Neuman. 2016. *Computational Personality Analysis: introduction, practical applications and novel directions.* Springer.

[13] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology* 54, 1 (2003).

[14] Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, , and Lyle Ungar. 2015. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.* Denver, USA.

[15] Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. 2017. Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports* 7, 1 (2017).

[16] Dan Stefanescu, Rajendra Banjade, and Vasile Rus. 2014. Latent Semantic Analysis Models on Wikipedia and TASA. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland.*