Using Topic Extraction on Social Media Content for the Early Detection of Depression

Diego Maupomé and Marie-Jean Meurs

Université du Québec à Montréal, Montréal, QC, Canada
maupome.diego@courrier.uqam.ca
meurs.marie-jean@uqam.ca

Abstract. As part of the eRisk2018 shared task on depression, which consists in the early assessment of depression risk in social media users, we implement a system based on the topic extraction algorithm, Latent Dirichlet Allocation and simple neural networks. The system uses uni-gram, bi-gram and tri-gram frequency to extract 30 latent topics in an unsupervised manner. Once transformed onto this feature space, the users are given a diagnostic probability by a Multilayer Perceptron. Finally a decision algorithm based on an absolute threshold of probability, which shrinks with time, classifies every user.

Keywords: Topic extraction · Depression assessment · Multilayer perceptron.

1 Introduction

Depression is a major cause of morbidity worldwide. Although prevalence varies widely, in most countries, the number of persons that would suffer from depression in their lifetime falls between 8 and 12% [6]. Access to proper diagnosis and care is overall lacking because of a variety of reasons, from the stigma surrounding seeking treatment [10] to a high rate of misdiagnosis [11]. These obstacles could be mitigated in some way among social media users by analyzing their output on these platforms, and assessing their risk of depression or other mental health afflictions. To promote such analyzes that could lead to the development of tools supporting practitioners and moderators, the research community has put forward shared tasks like CLPsych [2] and the CLEF eRisk pilot task [1,7]. These tasks provide participants with annotated data and a framework for testing the performance of their approaches.

In the context of the CLEF eRisk 2018 task, which is aimed toward using as little content as possible from each user before assessing the risk of depression, we implemented a simple system based on unsupervised topic extraction and neural networks.

2 Dataset

The dataset used for eRisk 2018 consists of the written production of `reddit` [3] English-speaking users. Both training and test sets are divided into a total of

	Training dataset	Test dataset
# users	887	820
# writings	531,188	544,447
# no-risk users	752	741
# risk users	135	79
# no-risk writings	481,631	503,782
# risk writings	49,557	40,665

Table 1. Statistics on the eRisk 2018 pilot task dataset

10 chunks each, chronologically organized. Each chunk represents a sequence of writings for a given user in a period of time. Table 1 presents some statistics on the task datasets, which are further described hereafter.

The training set was built using the writings of 887 users, and was provided in whole at the beginning of the task. Users in the RISK class have admitted in separate outlets to being diagnosed with depression; NO_RISK users have not. It should be noted that the users' writings (in XML format) are divided into separate individual writings, or posts, which may originate from different separate discussions on the website. The individual writings, however, are not labelled. Only the user as a whole is labelled as RISK or NO_RISK. Furthermore, the focus of the task being on *early* assessment, each user's production is divided into 10 separate *chunks*. Each one of these corresponds to approximately 10% of a user's production. This proportion is computed on the total number of individual writings, as opposed to the total number of words or the total time frame for these. The two classes of users are highly imbalanced in the training set with the positive class only counting 135 users to 752 in the negative class.

The test set was built using the writings of 820 users. To assess the capacity of a model to predict risk of depression as early as possible, the test data were also divided into chunks in the same manner. During ten weeks, a chunk was released every week, with participants submitting for each user either a decision (RISK or NO_RISK) or no decision. Once a decision was made, it could not be changed. A decision had to be taken for each user after the final chunk.

3 Methodology

As the chunks accumulate, the total textual output of users can become quite large, with a few users having up to 2000 total writings. In addition to our previous analysis of the dataset [5], this motivated us to use approaches that would summarize the writings of a user in a manner that would be easily translatable to emotion analysis. We opted for topic extraction as, intuitively, the topics of discussion in which a person engages would be telling of their mental state.

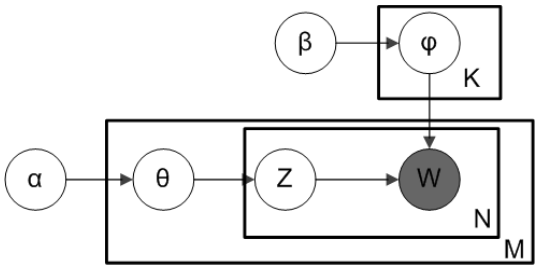

Fig. 1. Latent Dirichlet Allocation (LDA) in plate notation

Therefore, we conceived a simple system that begins by extracting topics using LDA [4].

3.1 LDA

LDA is a statistical generative model that posits documents (users in our case) as resulting from a mixture of topics, with each topic having its own word distribution. The model is presented in plate notation in Figure 1. Both the topics and words have a Dirichlet prior distribution, respectively, with α being the parameter of the per-document Dirichlet prior on the topics, and β being the parameter of the per-word Dirichlet prior on the words. θ_m is the topic distribution for document m. ϕ_k is the word distribution for topic k. z_{nm} is the topic for the n^{th} word in the m^{th} document. w_{nm} is the actual n^{th} word in the m^{th} document.

3.2 Pipeline

The LDA model is applied on a term-document matrix of the users, where the element at position ij is the relative frequency of term i in document j. The LDA model then outputs a topic-document matrix, representing the relative importance of each topic in each document. Finally, this representation is fed to a Multilayer Perceptron (MLP), which produces a predicted label for each user.

We restricted the term-document matrix to the 3000 most frequent n-grams of length 1 to 3, removing all stop words. We experimentally found that the LDA model works best on the validation set when limited to 30 topics and fitted with posts as documents rather than users. The MLP has two intermediate layers of 60 and 30 units with no special activation function for these. Again, these setting yielded the best results in validation.

4 Related approaches

Topic extraction has been used in the detection of mental health disorders with success because of the reasons previously mentioned: it allows to summarize what is potentially lengthy text, and its results are very interpretable. Resnick

	$ERDE_5$	$ERDE_{50}$	F1	P	R
FHDO-BCSGB	9.50%	**6.44%**	**0.64**	0.64	0.65
UNSLA	**8.78%**	7.39%	0.38	0.48	0.32
RKMVERIC	9.81%	9.08%	0.48	**0.67**	0.38
UDCB	15.79%	11.95%	0.18	0.10	**0.95**
UQAMA (ours)	10.04%	7.85%	0.42	0.32	0.62

Table 2. Results for top systems for each metric ($ERDE_5$, $ERDE_{50}$, F1-score, precision and recall)

et al. [9] applied regular LDA and variants, most notably supervised LDA [8], to detect depression in Twitter users. It should be noted, however, that in order to perform classification with unsupervised LDA, a clinical psychologist assessed the relevance to depression of the once-extracted topics from the training data. While they showed promising results, the positive instances in the data were users who self-described as having been diagnosed with depression. This could present a bias as people who openly discuss their diagnostics could potentially be more likely to openly discuss their state of mind.

5 Experiments and Results

The training data were split, using 80% of the users for actual training and saving the remaining 20% for validation. The n-grams were extracted solely from the training subset. The LDA model and the MLP were also only fitted on said subset. The last part of the system, which consists in a decision procedure based on the prediction probabilities output by the classifier was determined on the validation. We found that we obtained the best results by setting an absolute threshold on the prediction, which we shrank by a fixed ratio at every chunk. The initial probability threshold we selected was 0.85, as was the shrinking ratio. Thus, the threshold at chunk i, T_i, was given by $T_i = 0.85^i * 0.85$. This resulted in an $ERDE_5$ measure of 10.04% and an $ERDE_{50}$ of 7.85%. We also tested prediction probability convergence over chunks to no avail.

In testing, all decisions had been taken by the system by chunk 5, resulting in moderate results, presented in Table 2. Our system tends to favor quick decisions for negative samples, resulting in a low ERDE metric. The shrinking threshold forces then a conservative decision, resulting in a relatively high recall. Despite the small size of the dataset, the MLP outperforms a similar system we implemented in the early stages of development, which consisted of one LDA model per class. The decision procedure for this system was based on the perplexity of each model for every new sample.

6 Conclusion and Future Work

We put together a simple and intuitive system for depression detection based on topic extraction with the LDA model. We achieved moderate results, which may be explained by the unsupervised nature of the topic extraction. The limited number of users greatly hinders the predictive power of the MLP and may also be at fault. In future work, we will implement a supervised variant of LDA to compare with these results.

Reproducibility. To ensure full reproducibility and comparisons between systems, our source code is publicly released as an open source software in the following repository: `https://github.com/BigMiners/eRisk2018`.

References

1. CLEF eRisk pilot task. `http://early.irlab.org/`, Accessed July 6, 2018
2. CLPsych Shared Task. `http://clpsych.org/shared-task-2017/`, Accessed July 6, 2018
3. Reddit. `https://www.reddit.com/`, Accessed July 6, 2018
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
5. Briand, A., Almeida, H., Meurs, M.J.: Analysis of Social Media Posts for Early Detection of Mental Health Conditions. In: Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31. pp. 133–143. Springer (2018)
6. Kessler, R., Berglund, P., Demler, O., et al: The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r). JAMA **289**(23), 3095–3105 (2003)
7. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Avignon, France (2018)
8. Mcauliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in neural information processing systems. pp. 121–128 (2008)
9. Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd-Graber, J.: Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 99–107 (2015)
10. Rodrigues, S., Bokhour, B., Mueller, N., Dell, N., Osei-Bonsu, P.E., Zhao, S., Glickman, M., Eisen, S.V., Elwy, A.R.: Impact of stigma on veteran treatment seeking for depression. American Journal of Psychiatric Rehabilitation **17**(2), 128–146 (2014)
11. Vermani, M., Marcus, M., Katzman, M.A.: Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study. The primary care companion to CNS disorders **13**(2) (2011)