IRIT at e-Risk

Idriss Abdou Malam [1] Mohamed Arziki [1] Mohammed Nezar Bellazrak [1]
Farah Benamara [2] Assafa El Kaidi [1] Bouchra Es-Saghir [1] Zhaolong He[2]
Mouad Housni [1] Véronique Moriceau [3] Josiane Mothe [2] Faneva Ramiandrisoa[2]

(1) IRIT, UMR5505, CNRS & ENSEEIHT, France,
(2) IRIT, UMR5505, CNRS & Univ. Toulouse, France,
(3) LIMSI, CNRS, Univ. Paris-Sud, Universit Paris-Saclay, France

Abstract. In this paper, we present the method we developed when participating to the e-Risk pilot task. We use machine learning in order to solve the problem of early detection of depressive users in social media relying on various features that we detail in this paper. We submitted 4 models which differences are also detailed in this paper. Best results were obtained when using a combination of lexical and statistical features.

1 Introduction

The WHO (World Health Organization) reports that "the number of people suffering from depression and/or anxiety increased by almost 50% from 416 million to 615 million" from 1990 to 2013[1]. Depression and Bipolar Support Alliance also estimates that "major depressive disorder affects approximately 14.8 million American adults" and "annual toll on U.S. businesses amounts to about $70 billion in medical expenditures, lost productivity and other costs" (http://www.dbsalliance.org).

Depression detection is crucial and many studies are devoted to this challenge [7]. While there are clinical factors that can help for early detection of patients at risk for depression [10], in this paper we present our approach to help early depression detection from social media analysis, as part of our participation to CLEF e-risk 2017 pilot task [6].

Recent related work focus on people communication and social media post analysis to detect depression. Rude's study shows that depressed people tend to use the personal pronoun ("I") more intensively than others [9]. Other features have also been noticed. For example, De Choudhury *et al.* noticed that the depressive people show less activity during the day and more activity during the night [3]. Schwartz *at al.* reported that depressive people tend to use swear words and talk more about the past [11].

These previous studies show that some cues and features extracted from social media posts can be related to depression. In this paper, we report our investigations on using various features in order to answer the e-risk challenge

[1] `http://www.who.int/mediacentre/news/releases/2016/depression-anxiety-treatment/fr/`

as described in [6]. The e-risk pilot task aims to detect a depressive person as soon as possible by analysing her or his posts in Reddit[2] that are provided as a simulated data flow.

In our participation runs, the features we used to characterize posts are of two types: lexicon-based (extracted using NLTK toolkit[3]) and numerical features. These features are used in a machine learning method using Weka.

The remaining of this paper is organized as follows: Section 2 provides an overview of the model we used. Section 3 details the different features we implemented to train different models. In Section 4 we detail the 4 runs we have submitted and the underlying models and present the results. In Section 5 we discuss the results and depict future work.

2 Model overview

Figure 1 presents an overview of the model we use.

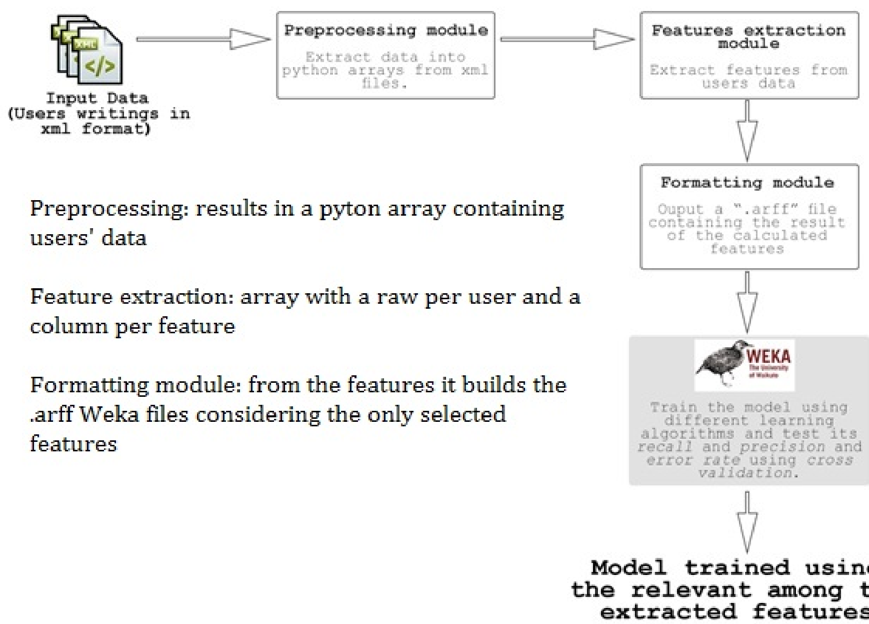

Fig. 1. Overview of the model used by the e-risk IRIT system.

[2] `https://www.reddit.com/`

[3] NLTK is a platform for building Python programs for natural language processing that interfaces easily with text processing and machine learning libraries (www.nltk.org)

The model is composed of three modules. In the first one, we pre-process the XML files that contain the users' posts. The second module aims at extracting the features. Notice that while some features capture information from any textual parts, others focus either on the Title part (corresponding to the initial post) or on the Text part which corresponds to comments on the initial post. The feature extraction module is extensible: while we developed some features, new features can easily be added. Then, in the formatting module, we select a subset of the features to be used in the model.

3 Features and models

We developed different types of features. Some have linguistic foundation while others are more statistically-based. We distinguish lexicon-based features from other numerical features.

For lexicon-based features, we rely either on previous observations on depressive subjects' behaviour [3, 11] or on hypothesis that we wanted to evaluate.

Table 1 presents the features that rely on a lexicon. Each feature is calculated as follows: (a) we extract the considered lexicon words from each user's post, (b) for a given user, each lexicon word is weighted by the normalized word frequency (division of the frequency by the total number of words in the user's posts), (c) we then create one feature by averaging the obtained weights over the lexicon words.

Features are calculated for each user as follows: we first calculate the feature value for each of his or her post or comment, then we average the value over his or her posts in the chunk ; when several chunks are used, we average the feature values obtained for each chunk for the considered user.

We also used some other numerical features that are described in Table 2. The details of the features are described in [8].

We submitted 4 runs corresponding to 4 models. The features that were used for each model are listed in Table 3. While we used model GPLA to start with, the other models were introduced later on. The second column of Table 3 indicates the chunk number when each model was introduced. The 4 runs corresponding to our 4 models were performed with the Random Forest learning algorithm under the Weka platform using the default parameters.

In order to decide whether to issue a decision for a subject or wait for more chunks, we used the prediction confidence rate that Weka generates for each prediction. We set a threshold (estimated using samples of depressive subjects) and we only issued decisions that had a prediction confidence that exceeds the selected threshold. The evolution of the threshold for each model through the runs and according to the chunks can be tracked using Table 4, a threshold of 0.5 basically means that all predictions are considered.

Num	Name	Hypothesis or tool/resource used
1	Self-Reference	High frequency of self-reference words.
2	Over generalization	Depressive users use words like: "everyone", "everywhere", "everything" a lot.
3	Sentiment	Use of Vader analyser [4] for assigning a polarity score to users' posts: - Negative < -0.05 and Positive > 0.05 - Neutral otherwise
4	Emotion	High frequency of emotionaly negative words Used WordNet-Affect [12], to assign a label to each word: *Negative, Positive or Ambiguous* we then calculated the frequency of each category
5	Depression symptoms & related drugs	From De Choudhury *et al.* [3] and Wikipedia list[4].
6	Past words	High frequency of past words.
7	Specific verbs	High frequency of "were" and "was", "like" "have", "being"
8	Targeted "I"	Depressive people tend to target themselves more in subjective context expecially using adjectives
9	Negative words	High frequency of negative words Used SentiWordNet [1] to detect negative words in texts
10	Part-Of-Speech frequency	Higher usage of verbs and adverbs and lower usage of nouns
11	Relevant 3-grams	Higher frequency of 3-grams described by Gualtiero B. *et al.* [2] and suggested ones
12	Relevant 5-grams	Higher frequency of 5-grams described by Gualtiero B. *et al.* [2] and suggested ones
13	Relevant 1-grams	Higher frequency of 1-grams described by Gualtiero B. *et al.* [2] and suggested ones

Table 1. Details of the features based on lexicons.

4 Results

The evaluation takes into account not only the correctness of the output of the system (i.e. whether or not the user is depressed) but also the delay taken to emit its decision. To this aim, the ERDE (*Early Risk Detection Error*) metric proposed in [5] is used. This measure rewards early alerts and the delay taken by the system to make its decision is measured by counting the number of distinct textual items seen before giving the answer.

Our best results when considering ERDE measures are obtained using model GPLC which does not use POS results nor the most frequent n-grams. Including them in the model slightly improves F1 measure mainly because of higher recall

Num	Name	Hypothesis or tool/resource used
14	Variation of the number of posts	For depressive people, the variation of the number of posts is generally small.
15	Average number of posts	Depressive users have a much lower number of posts.
16	Average number of words per post	The two groups of users have different means.
17	Minimum number of posts	Depressive users have a lower value in general.
18	Variation of the number of comments	For depressive people, the variation of the number of comments is generally small.
19	Average number of comments	Depressive users have a much lower number of comments.
20	Average number of words per comment	The two groups of users have different variances.

Table 2. Details of the other numerical features.

Name	Details	Used for the first time in chunk
GPLA	Features: 1, 7-9, 14-19	**1**
GLPB	Features: 1, 6 (verbs only), 7-9 , 14-19	**2**
GLPC	Features: 1-9, 14-19	**3**
GLPD	All features	**10**

Table 3. Features used in each run.

Model / Chunk	1	2	3	4 - 8	9	10
GPLA	0.9	0.9	0.9	0.9	0.6	0.5
GPLB	-	0.8	0.8	0.8	0.6	0.5
GPLC	-	-	0.8	0.8	0.6	0.5
GPLD	-	-	-	-	-	0.5

Table 4. Evolution of the decision threshold for the 4 models according to the considered chunk

.

(0.60 against 0.50) (see Table 5). Our run GPLB had the 2nd best Recall (0.83) across participants and GPLA the 5^{th}.

5 Conclusion and Future Work

In the runs we submitted we consider 19 features. However, some additional features are worth studying. In future work, we aim at considering temporal features such as the date of the posts, part of the day, etc. Moreover, we would like to modify the way features are calculated : in the case of lexicon-based

Name	$ERDE_5$	$ERDE_{50}$	F1	P	R
GPLA	17.33%	15.83%	0.35	0.22	0.75
GPLB	19.14%	17.15%	0.30	0.18	0.83
GPLC	**14.06%**	**12.14%**	0.46	0.42	0.50
GPLD	14.52%	12.78%	**0.47**	0.39	0.60

Table 5. Results for our 4 runs.

features, each lexicon item would be a distinct feature. By this way, we would obtained a richer representation of each user and potentially a better detection.

References

1. S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Istituto di Scienza e Tecnologie dell Informazione Consiglio Nazionale delle Ricerche Via Giuseppe Moruzzi 1, 56124 Pisa, Italy*, 2010.
2. G. B. Colombo, P. Burnap, A. Hodorog, and J. Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, 2015.
3. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, page 2, 2013.
4. C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014*, 2014.
5. D. E. Losada and F. Crestani. A test collection for research on depression and language use. In *Conference Labs of the Evaluation Forum*, page 12. Springer, 2016.
6. D. E. Losada, F. Crestani, and J. Parapar. eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations. In *Proceedings Conference and Labs of the Evaluation Forum CLEF 2017*, Dublin, Ireland, 2017.
7. L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescents speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2011.
8. I. A. Malam, M. Arziki, M. N. Bellazrak, A. El Kaidi, B. Es-Saghir, and M. Housni. Automatic detection of depression in social networks. Technical report, Universit de Toulouse, France, 07 2017.
9. S. Rude, E.-M. Gortner, and J. Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.
10. U. Sagen, A. Finset, T. Moum, T. Mørland, T. G. Vik, T. Nagy, and T. Dammen. Early detection of patients at risk for anxiety, depression and apathy after stroke. *General hospital psychiatry*, 32(1):80–85, 2010.
11. H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, 2014.

12. C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086*, 2004.