Automatic Detection of Depressive Users in Social Media

Farah Benamara [1] — Véronique Moriceau [1,2] — Josiane Mothe [1,3] — Faneva Ramiandrisoa [1] — Zhaolong He [1]

[1] Institut de Recherche en Informatique de Toulouse, IRIT, UMR5505 CNRS,
Université de Toulouse
[2] LIMSI-CNRS, Univ. Paris-Sud, Université Paris Saclay, Orsay
[3] ESPE, UT2J

RÉSUMÉ. *La dépression est une affection courante qui concerne environ 350 millions de personnes dans le monde selon les estimations de l'Organisation Mondiale de la Santé. La détection de ce trouble est donc un enjeu majeur de santé publique. Plusieurs recherches en psychologie ont démontré l'existence d'un lien fort entre l'état dépressif d'un individu et son expression langagière. Dans cet article, nous proposons de repérer automatiquement ces indices linguistiques dans le but de détecter les comportements dépressifs à partir de messages postés sur les réseaux sociaux. Notre approche est supervisée et se base sur un ensemble de traits d'apprentissage allant de traits standards tels que les sacs de mots ou les traits de surface, à des traits plus sémantiques. L'approche a été évaluée sur des données issues du réseau social Reddit et appliquée sur deux tâches : (a) étant donné les posts d'un utilisateur, détecter si l'auteur est dépressif ou non; (b) étant donné un fil de posts d'un même utilisateur présentés chronologiquement, détecter au plus tôt les signes de la dépression. Nos résultats montrent l'intérêt de notre approche pour ces deux tâches.*

ABSTRACT. *According to the World Health Organization, 350 million people worldwide suffer from depression. Detecting this trouble constitutes thus a challenge for personal and public health. Research in psychology has shown a strong correlation between the psychological state of an individual and its language use. In this paper, we propose to leverage such linguistic features to automatically detect depressive users on social media posts. Our approach is supervised and relies on a set of features going from standard bag of words and surface features to more linguistically informed features. This approach has been evaluated on Reddit social media posts and applied on two tasks: (a) Given user's posts, detect whether the author is depressive or not, (b) Given a user's history of writings, early detect signs of depression. Our results show that our approach is reliable on both tasks.*

MOTS-CLÉS : *Recherche d'information, Dépression, Médias sociaux, Traitement automatique des langues.*

KEYWORDS: *Information retrieval, Depression, Social media, Natural language processing.*

1. Introduction

Depression is a common mental disorder. The 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, a non-profit organization providing support groups for people with depression or bipolar disorder, 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”[2]. Detecting this trouble is thus crucial and constitutes a challenge for personal and public health.

Many studies in the literature have been devoted to this challenge (France *et al.*, 2000 ; Low *et al.*, 2011 ; Ozdas *et al.*, 2004). While there are clinical factors that can help for early detection of patients at risk for depression (Sagen *et al.*, 2010), there are also some language usages that are specific to depressive states (Pennebaker *et al.*, 2003 ; Rude *et al.*, 2004). Indeed, depression was found to be associated with distinctive linguistic patterns, such as the excessive use of personal pronouns, past tense or negative emotions. People’s writing can thus be used to capture their psychological states.

In recent years, the emergence of social media platforms like Facebook, Twitter or Reddit allows people to share their personal experiences, ideas or thoughts in a more straight way. It becomes possible to analyse posts on these platforms using linguistic indicators to detect depressive users. Most state of the art approaches on depression detection in social media employ supervised learning methods trained on manually annotated datasets. Several groups of features have been used, such as : emoticons (Wang *et al.*, 2013), posting time (Choudhury *et al.*, 2013), use of sear words (Schwartz *et al.*, 2014), and topic modelling (Resnik *et al.*, 2015).

In this paper, we propose a supervised model to automatically detect depressive users on social media. Our approach relies on groups of features going from standard bag of words and surface features to more linguistically informed features. Some of these features have already been used in past studies while others are new. The approach we developed has been evaluated on Reddit posts and, as far as we know, applied for the first time on two tasks : (a) Given a user’s writings, detect if the corresponding user is depressive or not, (b) Given a user’s history of writings, early detect signs of depression. Our results show that our approach is reliable for both tasks.

The remaining of this paper is organized as follows. Section 2 is an overview of state-of-the-art approaches for depression detection in social media posts. Section 3 presents the data collection. Section 4 details the features we used as well as the mo-

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

[2] `http://www.dbsalliance.org`

dels and method developed. Section 5 reports on the experiments and the results obtained on both tasks. Section 6 concludes this paper.

2. Related work

2.1. *Existing social media datasets for depression detection*

Overall, there is a lack of publicly available data for conducting research on the interaction between language and depression (Losada et Crestani, 2016). Among the few available resources, we can distinguish between two types. The first one focuses on language differences between people suffering from a given disorder and a control group (e.g., depressed vs. non-depressed, bipolar vs. non-bipolar). Tweets collection used recently in the CLPsych (Computational Linguistics and Clinical Psychology Workshop)[3](Coppersmith *et al.*, 2015) shared tasks series is an example of such a dataset. The second type of resource attempts, in addition, to capture the evolution of the language used by depressed individuals by analysing a large chronological sequence of writings leading to that disorder. Time is considered as a fundamental factor in building such a resource because appropriate action or intervention at early stages of depression can be highly beneficial (Losada et Crestani, 2016). Detecting depression at early stage was the main objective of eRisk shared task [4](Losada et Crestani, 2016) in its 2017 and first edition, where Reddit posts have been used.

2.2. *Main approaches*

Most of the approaches from the literature are based on supervised learning (Wang *et al.*, 2013 ; Choudhury *et al.*, 2013 ; Nakamura *et al.*, 2014 ; Cook *et al.*, 2016 ; Mowery *et al.*, 2016 ; Losada *et al.*, 2017). These studies have shown that people suffering from depression tend to :

– talk more about relationship and life (friends, home, dating, death ...),

– show their personality (openness, extraversion/introversion, ...),

– become more self-concerned (more first person pronoun used),

– use more emoticons, negative emotion words (anger, anxiety, ...) and negation words,

– use more verbs and adverbs, exclamation and question marks,

– frequently use of semantic words (swear, ...),

– retrospect their past and are concerned about their future.

From these observations, several features have been proposed : n-grams (mostly unigrams, bigrams and trigrams) (Choudhury *et al.*, 2013 ; Schwartz *et al.*, 2014 ; Farias-Anzaldua *et al.*, 2017 ; Almeida *et al.*, 2017), dedicated lexicons to account for depression symptoms, drug names, and medical words (Choudhury *et al.*, 2013 ; Trotzek *et al.*, 2017 ; Sadeque *et al.*, 2017 ; Almeida *et al.*, 2017), topic models (e.g.

[3] `http://clpsych.org/shared_task/`

[4] `http://early.irlab.org/`

Latent Dirichlet Allocation) (Resnik *et al.*, 2013 ; Resnik *et al.*, 2015 ; Armstrong, 2015), sentiment or emotion lexicons (Choudhury *et al.*, 2013 ; Schwartz *et al.*, 2014 ; Mowery *et al.*, 2016 ; Resnik *et al.*, 2015 ; Sadeque *et al.*, 2017). In addition to these bags of words and surface features, other studies rely on more semantic features such as first person pronoun (Trotzek *et al.*, 2017), temporal indicators (Wang *et al.*, 2013 ; Farias-Anzaldua *et al.*, 2017), or users online behaviours in the social media at the post level (Choudhury *et al.*, 2013 ; Almeida *et al.*, 2017 ; Farias-Anzaldua *et al.*, 2017 ; Shen *et al.*, 2017). Temporal and users behaviour features have proven to be significant factors in detecting depressive troubles since the irregular activities of users are the direct reflect of their mind state.

We developed a supervised learning approach based on several state of the art features as well as new features to predict both the depressive state of a user given a set of posts and the early traces of depression of a user given a chronological order of his postings (i.e. identify the post where to make a decision). We developed different features and models, and evaluate them models on eRisk 2017 Reddit data. The work presented in this paper extends the one presented in (Malam *et al.*, 2017). While using feature-based machine learning to treat the problem is not novel, the features we used are.

3. Data

3.1. *The eRisk 2017 Reddit dataset*

The dataset we consider is composed of posts from the Reddit[5] social media platform. Contents in Reddit are organized by areas of interest called "subreddits". Users can post comments, or respond back and forth in a conversation-tree of comments. Posts and comments are represented by a user ID, a posting time and a textual content. The dataset used in this study is the one used at CLEF eRisk 2017 task[6], that aims at detecting early traces of depression by analysing users' writings that are provided as a simulated data flow. To build the CLEF eRisk dataset, Losada and Crestani collected a maximum number of submissions (posts and comments) from any subreddits for each user and those who have less than 10 submissions were excluded. In this dataset, users are annotated as *depressed* and *non-depressed*. To consider that a user is depressed, s/he must have posts/comments that matched self-expressions of depression diagnoses such as "I was diagnosed with depression", and then the organizers manually verified that it was really genuine. These posts/comments with self declaration were discarded from the dataset to avoid making the detection trivial. On the other hand users whose posts/comments in depression subreddits do not contain any posts with declaration of depression were considered as non-depressed. Some users and their posts were also selected from random subreddits and considered as non-depressed. The dataset is described in detail in (Losada et Crestani, 2016). For each user, the text collection is a sequence of writings sorted in chronological order. It has been divided into 10 chunks in CLEF eRisk 2017 task, where chunk 1 contains the first (oldest) 10% of a user's

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

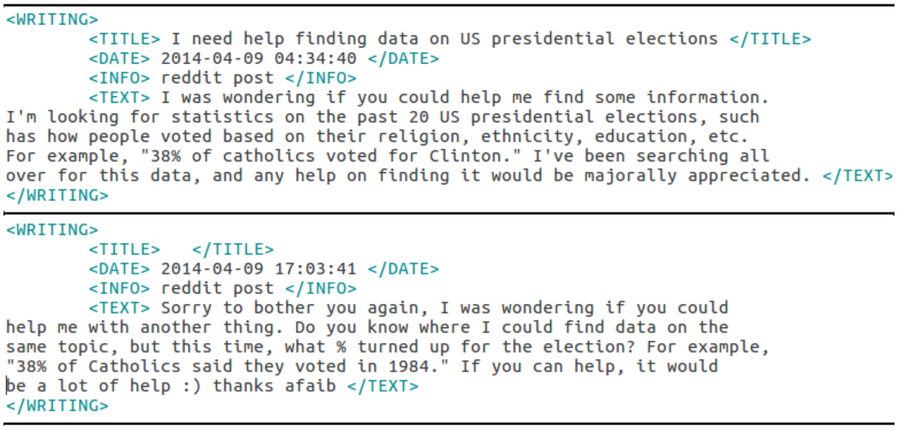
[6] `http://erisk.irlab.org/2017/index.html`

writings, chunk 2 contains the second 10% and so on. Figure 1 shows an example of content posted by a user annotated as "depressed".

> I was feeling much better, myself harm stopped/suicidal thoughts stopped, I became more social, I could focus on school/get things done. Then my mom noticed that I wasn't eating as much as I used to, and decided to do some research about the medication I was on. She made me stop taking it immediately afterwards...

Figure 1. *Example of a depressive user's text*

Each chunk consists in XML files (one file per user) that store : the user's identifier and a collection of his or her writings (posts or comments). Each writing further contains : the title of the post, the posting time, and finally the user's textual content. If the title is empty, a writing is considered as a comment, otherwise it is a post. Figure 2 shows an example of a post and a comment in the eRisk collection.



```
<WRITING>
        <TITLE> I need help finding data on US presidential elections </TITLE>
        <DATE> 2014-04-09 04:34:40 </DATE>
        <INFO> reddit post </INFO>
        <TEXT> I was wondering if you could help me find some information.
I'm looking for statistics on the past 20 US presidential elections, such
has how people voted based on their religion, ethnicity, education, etc.
For example, "38% of catholics voted for Clinton." I've been searching all
over for this data, and any help on finding it would be majorally appreciated. </TEXT>
</WRITING>
```

```
<WRITING>
        <TITLE>   </TITLE>
        <DATE> 2014-04-09 17:03:41 </DATE>
        <INFO> reddit post </INFO>
        <TEXT> Sorry to bother you again, I was wondering if you could
help me with another thing. Do you know where I could find data on the
same topic, but this time, what % turned up for the election? For example,
"38% of Catholics said they voted in 1984." If you can help, it would
be a lot of help :) thanks afaib </TEXT>
</WRITING>
```

Figure 2. *Example of a post (top) and a comment (bottom) in the eRisk 2017 collection.*

3.2. *Statistics*

The dataset is split into a training and a test sets, as described in Table 1. The training set consists of 83 depressive users and 486 non-depressed, while the repartition of users in the test set is 52 vs. 349 respectively. For depressed users there are 4,911 posts and 25,940 comments in the train set, much less that for non-depressed users. We can see that the ratio between depressed and non-depressed users is not perfectly the same for train (0.21) and test (0.15), but this splitting was given by the organizers.

Table 2 shows the mean number of words, posts and comments for each chunk and each user. If we look at the mean per chunk, we observe that the posts from depressive users contain about 20 times less words than the non-depressive ones and about 6 times less for the comments. The data is thus (fortunately) extremely unbalanced, which makes our task more difficult.

	Train		Test	
Number of	Depressed	Non depressed	Depressed	Non depressed
Users	83	403	52	349
Posts	4,911	91,381	1,928	65,735
Comments	25,940	172,791	16,778	151,930
Writings per user (Avg)	371.7	655.5	359.7	623.7

Table 1. *Distribution of training and testing data on eRisk 2017 data collection.*

	Depressive users' statistics			Non-depressive users' statistics		
Mean per	Words	Posts	Comments	Words	Posts	Comments
Chunk	142,913	491	2,808	868,968	9,138	18,265
User	17,218	59	338	21,563	227	453

Table 2. *Statistics (round up) about users' posts according to the class the users belong to.*

4. Supervised learning to detect depression from social media

In this section, we present our supervised approach for automatic depression detection on eRisk data. We first detail the set of features we rely on, then present the models we have built.

4.1. *Features used*

We represent user writings (all posts and comments) with a vector composed of seven groups of features : Bag of words, Language Style, User behaviour in social media, Self-Preoccupation, Reminiscence, Symptoms and drugs, and finally Sentiment and emotion. Some of them have already been used in past studies while others are new (the latter are put in bold font).

Bag of words

We selected from the depressive users' writings in the training set, the top 50 most frequent unigrams according to their term frequency[7]. Among them, we only kept 18 unigrams according to a Chi-squared filter. To set the number of unigrams to keep, we conducted a preliminary study, using various numbers of unigrams. Eighteen was the best trade-off between the number and the accuracy of the results. The resulting selection is as follows : *feel, im, really, things, help, ive, know, someone, life, time, going, like, even, much, day, though, work, people*. These eighteen simple features are used as a baseline.

Language Style

The aim here is to capture the language style adopted in a user's writing. Eight features are used, all of them are normalized frequencies of :

[7] We also experiment with bigrams and trigrams but the results were not conclusive.

– adjectives, verbs, nouns and adverbs. The intuition is that depressive users, as suicidal people, are characterized by a higher use of verbs and adverbs, but lower use of nouns (Choudhury *et al.*, 2013). We used the NLTK toolkit[8] to extract POS categories,

– **negation**. This is a new feature that captures the fact that depressed people use much more negative words in their writings. We use a small lexicon of English negative words like : *no, not, didn't, can't*, ...

– **capitalized words**. We observed that depressive users are more likely to put emphasis on the target they mention, like in : ''*I'm the UNLUCKIEST man in the world!*",

– Punctuation marks (! or ? or any combination of both), and emoticons (Wang *et al.*, 2013). Indeed, punctuation marks tend to express doubt and surprise while emoticons are another way for users to express their sentiment or their feeling.

User behaviour in social media

This group of features represents the user's behaviour in writing a post/comment and its posting time (Choudhury *et al.*, 2013), and is therefore dependent of the social media used (Reddit in our case). We believe however that equivalent behaviours can be easily found in most social media platforms. We used five features defined as follows :

– at the post level : average number of words per post and average number of posts of each user. We counted the posts/words in posts for each chunk, then divide it by the total number of chunks (recall that we have 10 chunks per user – cf. Section 3),

– at the comment level : average number of words per comment and average number of comments of each user, computed in the same way as the previous two features,

– **at the posting time level** : since the sleeping habits of depressive users may not be regular, (Choudhury *et al.*, 2013) assume that they tend to post message late at night. We extended this feature to capture the **ratio of late posting times**. We thus split a day into 4 segments : "morning" (7am-12am), "afternoon" (12am-18pm), "night" (18pm-00pm), "deep night" (00pm-7am), and then counted the number of posts posted in the "deep night" and divided it by the total number of posts for this user to normalize the results.

Self-Preoccupation

Self-preoccupation features deal with user's psychological aspect and capture to what extent users are self-preoccupied, by excessive use of first personal pronouns to refer to themselves or the tendency to over-generalize.The nine features of this group are the normalized frequency of :

– first person pronouns (*I, me, myself, mine, my*) and the pronoun *I* when subject of *be* (e.g., *I'm*) (Rude *et al.*, 2004),

[8] `http://www.nltk.org/`

– all first person pronouns as the sum of frequency of each first pronoun (Wang *et al.*, 2013),

– **pronoun *I* in subjective context**, focusing in particular on all *I* targeted by an adjective, as this grammatical category is often used to convey subjective meaning regarding its target. The aim here is to capture how often a user expresses sentiments or emotions when he talks about himself. To extract this feature, we rely on specific lexico-syntactic patterns such as : `I'm NEG ADJ` (e.g., *I'm not attractive*), and `I'm ADV ADJ` (e.g., *I'm very nervous*),

– over-generalization, including intense quantifiers and **superlatives**. (Mowery *et al.*, 2016) noticed that a depressive person is inclined to over-generalization by using intense quantifiers, like *all, everything, nothing, anymore, etc*. For example, instead of criticizing a specific person, he may write *all men/women are bad*. We extended this features to account, in addition, for superlatives like *worst*.

Figure 3 is an example of a depressed user's text that illustrates some of these features.

> **I'm** struggling right now with **all** **my** relationships. **I** just broke up with **my** girl because **my** heart wasn't in it and it was the right thing to do [„,] The problem is **I** underestimated the friendship **I** had with **my** gf. It wasn't perfect but at least **I** had someone who was obligated to put **me** before **everyone** else. At least **I** had someone to vent to. Now that **I'm** single and depressed **nobody** is around. **I** have friends, but **I** always feel like **I'm** bothering them and they **all** have other priorities whether it's kids or a significant other. So this leaves **me** by **myself**. Who am **I** supposed to talk to ? **I** feel like **nobody** understands me.

Figure 3. *Example of a depressive user's text with first person pronouns in red, "I" subject of "be" in blue and over-generalization in brown.*

Reminiscence

Mowery *et al.* (2016) showed that depressive users tend to make reference to past more frequently than non-depressive users. We defined four *reminiscence* features to capture the reference to past as the normalized frequency of :

– temporal expressions referring to past (*yesterday, last, before, ago, past, back, earlier, later*) (Mowery *et al.*, 2016),

– **past tense verbs**,

– past tense auxiliaries (*was, were*),

– a combination of the two previous features.

Examples of occurrences of these features can be seen in Figure 4.

> **Back** in my days, it **was** pretty embarassing to admit that you're doing something just because "it's cool". It **was** only cool if you **had** a better reasons for it than that. Good ol' times of 2010 :/

Figure 4. *Example of a depressive user's text with reference to past time in red font.*

Symptoms, drugs and relevant depression vocabularies

These five features capture the frequency of :

– depression symptoms and antidepressant drugs, obtained from (Choudhury *et al.*, 2013) and Wikipedia,

– **depress word and its morphological variations** (*depressing, depressed, depression, depressive, etc.)*,

– 25 trigrams and 25 5-grams from (Colombo *et al.*, 2016) expressing depressive feelings (e.g. *to kill myself, want to die right now, have nothing to live for*, etc.),

– **words related to sleep**. We noted that depressed users tend to tell more about their sleeping in their writings, by using words such as *sleep*.

Sentiment and Emotion.

The last group of features concerns the use of sentiment and emotion words as sentiment analysis is important in depression detection as observed in (Wang *et al.*, 2013). We rely on NRC-Sentiment-Emotion-Lexicons[9] (Mohammad et Turney, 2013), freely available subjective lexicons, from which we extracted three features

– frequencies of negative and of positive sentiment,

– frequency of emotions from specific categories that may be linked to depressive person's feelings : anger, fear, surprise, sadness and disgust.

4.2. *Models*

We built 4 models that each uses a sub-set of the above features. We trained/tested these models using machine learning algorithms and compared them to the baseline. To evaluate and tune our models during the training stage, we used 10-fold cross validation.

– Baseline uses the bag of words features, i.e. the normalized frequency of each unigram (18 unigrams),

– Model 1 uses the following features : bag of words, features from Language style and from User behaviour,

– Model 2 uses all the Model 1 features plus the features from Self-Preoccupation,

– Model 3 uses all Model 2 features plus the features from Reminiscence and Symptoms/drugs/relevant vocabularies,

– Model 4 uses all the Model 3 features plus the features from Sentiment and Emotion. In other words, all the features presented in section 4.1.

5. Experiments and results

In this section, we report the results obtained when using a supervised classifier on the models listed above. We tested four classifiers : SMO (Sequential Minimal

[9] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

Optimization), Naive Bayes, Random Forest and Logistic regression as implemented in Weka toolkit [10]. We report here the best results only, obtained using Random Forest and the default parameters. We tested different parameters for Random Forest but the default ones work best.

The models we developed were respectively trained and evaluated on training and test sets of eRisk 2017 data collection (section 3) and applied on two tasks : (a) Given a user's writings, detect if the author is depressive or not ; we use all the test data without temporal consideration (b) Given a user's writings in chronological order, early detect signs of depression. Task (b) is similar to the eRisk challenge : test data is considered chunk by chunk in chronological order and a decision is made for each chunk in order to detect early trace of depression.

5.1. *Task (a) : Detection of depressive users given a user's writings*

Table 3 reports the results on the testing dataset with Random Forest (default parameters) after training on the eRisk training data set.

Model	Non depressed		Depressed		Macro	Accu-
	Precision	Recall	Precision	Recall	F-score	racy
Baseline	0.916	0.966	0.636	0.404	0.717	89.3%
Model 1	0.911	0.963	0.594	0.365	0.694	88.5%
Model 2	**0.917**	0.954	0.579	**0.423**	0.696	88.5%
Model 3	0.916	**0.971**	**0.677**	0.404	**0.725**	**89.8%**
Model 4	0.916	0.966	0.636	0.404	0.717	89.3%
Mod45 (45 features)	**0.932**	**0.974**	**0.750**	**0.519**	**0.783**	**91.52%**

Table 3. *Results with Random Forest. Bold values are higher than the baseline.*

According to the results of the four models (section 4.2), we note that precision for class *Depressed* is not stable, probably due to the high number of features. After feature selection with Chi-squared ranking, only 45 features were selected and used to train a Random Forest classifier. These features are :

– the 18 features from *Bag of words*,

– the 8 features from *Language style*,

– the 5 features from *User behaviour*,

– 7 features from *Self-preoccupation* : all features except frequency of *mine* and over-generalization,

– 1 feature from *Reminiscence* : frequency of past tense auxiliaries,

– the 5 features from *Symptoms, Drugs and relevant vocabularies*,

– 1 feature from *Sentiment-emotion* : frequency of words belonging to one of the five selected emotion categories,

[10] `http://www.cs.waikato.ac.nz/ml/weka/`

Six out of 45 features we kept for the model Mod45 are new features we proposed in this paper : negation, capitalized words, ratio of late posting times, pronoun *I* in subjective context, depress word and its morphological variations, and words related to sleep.

We can see from Tables 3 that this latter model, with 45 features only, outperforms the previous four models and the baseline. The differences between all models are not statistically significant (McNemar's test, $p < 0.05$). Both recall and precision are increased for the depressed class as well as for the non-depressed class. This model is thus used for task (b).

We analysed the correlation among the 10 best of the 45 features from Mod45 model. We found out that most of them are not correlated or have low correlations which show there are complementary for the model.

5.2. *Task (b) : Early detection of depressive users given 10 sequential releases of user writings*

As in eRisk challenge, for each sequential chunk, the system has to make a decision about the user : whether he or she is depressed or not; alternatively, the system can wait for more writings (more chunks) to make its decision. To solve this problem, we set a threshold for the prediction confidence score generated by our models for each prediction. This threshold has been estimated using samples of depressive subjects. A user is assigned to the target class if he had a prediction confidence score that exceeds the selected threshold. We have tested all our models with different thresholds but report the best results only : the model with 45 selected features (named Mod45 in the Table 3) with a threshold of 0.50 (i.e prediction confidence > 0.50).

Table 4 reports the results for each chunk. To evaluate the results for each chunk, we used all writings received up to the current chunk and applied the measures used in the eRisk challenge[11] which are : $ERDE$, F-score, Precision and Recall. $ERDE$ (Early Risk Detection Error), defined in (Losada et Crestani, 2016; Losada *et al.*, 2017), takes into account the accuracy of the decisions and the delay of these decisions (i.e. given a chunk, if a system does not emit a decision then it has access to the next chunk but the system gets a penalty for late decision). It is defined as follows :

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

Where :

– $c_{fn} = c_{tp} = 1$;

– $c_{fp} = 0.1296$ (proportion of positive cases in the test data);

– $d =$ binary decision taken by a system with delay k;

[11] The organizers provided scripts written in python to evaluate the results.

– $lc_o(k) = \frac{1}{1+e^{k-o}}$;
– o is a parameter and equal 5 for $ERDE_5$ and equal 50 for $ERDE_{50}$.

The lower the ERDE, the better the system (while for the other measures, the closest to 1, the better). $ERDE_5$ measures the error after 5 writings of a user (it promotes systems that emit few but quick decisions), $ERDE_{50}$ after 50 writings.

If we had participated to eRisk challenge with Mod45 system, our official results would be the results given in chunk 10. Our model would have achieved the second F-score and precision, the fourth when considering $ERDE_5$ and the fifth when considering $ERDE_{50}$. We also reported results on other chunks in order to show more detailed results. Table 4 reports the results when the system gives its decision for all the users at the reported chunk. At each chunk a three-level scale (depressive, non depressive, no decision (wait for more writings)) is used and at chunk 10, a two-level scale (depressive, non depressive) is used. We can observe that the results are similar in terms of $ERDE_5$ whatever the chunk is. Results are a little less stable with regard to $ERDE_{50}$. F-Score increases with the number of chunks and thus with the number of texts we rely on for the decision for each user.

	ch 1	ch 2	ch 3	ch 4	ch 5	ch 6	ch 7	ch 8	ch 9	ch 10
$ERDE_5$	12.76	12.59	**12.43**	12.47	12.50	12.56	12.56	12.60	12.60	12.69
$ERDE_{50}$	11.82	11.35	10.42	10.45	10.48	10.05	10.05	10.08	10.08	**9.93**
F-score	0.17	0.25	0.37	0.41	0.44	0.51	0.51	0.53	0.55	**0.61**
Precision	0.62	0.67	0.68	0.68	0.68	0.68	0.68	0.68	**0.69**	0.67
Recall	0.10	0.18	0.25	0.29	0.33	0.40	0.40	0.44	0.46	**0.56**

Table 4. *Results chunk by chunk of our RF-based model Mod45 with 45 features.*

Table 5 reports also the official participants' results from eRisk challenge 2017 : FHDOA, FHDOB (Trotzek *et al.*, 2017), UNLSA (Villegas *et al.*, 2017), UArizonaC (Sadeque *et al.*, 2017). Results are reported according to decreasing F-Score. The team who proposed UNLSA model extended the analysis of their model after eRisk challenge and show that their new model named TVT (Errecalde *et al.*, 2017) achieved better $ERDE_5$ value (12.30) and better $ERDE_{50}$ value (8.17) but not better F-Score. These values were obtained with different configurations of TVT.

Compared to these models, our model (Mod45) achieved the second F-score (chunk 10) and the fifth $ERDE_{50}$ (chunk 10). If we have a look to partial results (when not all the chunks are delivered), we achieve the second $ERDE_5$ after TVT (chunk 3) and the best precision (chunk 9) as FHDOB (see Table 5).

5.3. *Error analysis*

Although we did not yet analysed in an exhaustive way all the errors our algorithm made, we started to analyse the early detection of false positives, i.e. users who are detected as depressed from the first chunk while they are not (Task (b)).

	$ERDE_5$	$ERDE_{50}$	F-score	Precision	Recall
FHDOA*	12.82	9.69	**0.64**	0.61	0.67
Mod45 (ch 10)	*12.69*	*9.93*	*0.61*	*0.67*	*0.56*
UNSLA*	13.66	9.68	0.59	0.48	0.79
TVT_1	**12.30**	8.95	0.56	0.54	0.58
Mod45 (ch 9)	*12.66*	*10.08*	*0.55*	***0.69***	*0.46*
FHDOB*	12.70	10.39	0.55	**0.69**	0.46
TVT_2	13.13	**8.17**	0.54	0.42	0.73
Mod45 (ch 3)	*12.43*	*10.42*	*0.37*	*0.68*	*0.25*
UArizonaC*	17.93	12.74	0.34	0.21	**0.92**

Table 5. *Best results for each evaluation measure from eRisk challenge, ordered by F-score. Official runs are marked-up with a *, and our model with 45 features in italic font.* TVT_1 *and* TVT_2 *were released after the competition.*

There are three users that Mod45 misclassified when using the first chunk. Figure 5 displays the median values (computed using the first chunk of test data only) of the ten features that have been considered by Random Forest classifier as the most important to detect depression. We plot the values for one of the misclassified users (green circles), users who are actually annotated as depressed in the ground truth (orange squares) and users who are non-depressed (blue crosses). From this figure, we can see that the green circles (misclassified user) and the orange squares (depressed) dots are very close (while they should rather be close to blue crosses) which illustrates why this user has been misclassified. Moreover, from this figure we can see that the second feature – (b) frequency of personal pronouns – is the main cause of misclassification. We also plotted the same type of figure for the two other misclassified users (false positive) and same comment holds for them.

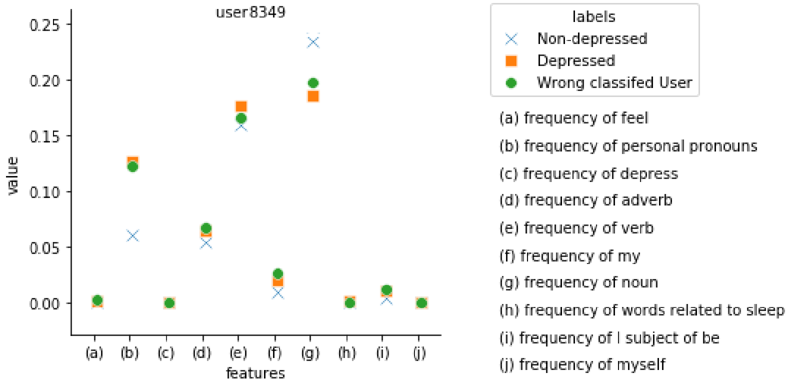


Figure 5. *Median values for one misclassified user (user8349), depressed and non-depressed users on the ten most important features.*

Interestingly enough, we found that this user misclassified after chunk 1 in Task (b) is correctly classified in Task (a) when considering all his writings.

6. Conclusion and Future Work

In this paper, we proposed a supervised model based on linguistic features to automatically detect depressive users on social media. This approach has been evaluated on Reddit social media posts and applied on two tasks : (a) Given user's posts, detect whether the author is depressive or not, (b) Given a user's history of writings, early detect signs of depression.

For Task (a), after feature selection (with 45 features), the best results are achieved with Random Forest which detects the depressive users with a precision of 75%, a recall of 51.9%, and an accuracy of 91.5%. This model is then used for Task (b). If we had the same condition as eRisk challenge, our result would be the Mod45 (chunk 10) where we obtained the second F-score and Precision compared to the participants, the fourth $ERDE_5$ and the fifth when considering $ERDE_{50}$. We also analysed the results in different chunks in order to know what would have been the results if we have decided to submit final decision at this time. We obtained good results when compared to the participants : second $ERDE_5$ (chunk 3) and the best precision (chunk 9).

In future work, we would like to go deeper in failure analysis, specifically related to false negative, users who are indeed depressed but that our system did not detected. On the other hand, we will analyse what the best features are and if the model fit other reference collections in the domain. For example, we would like to use transfer learning on the 2016 CLPsych shared task[12] dataset. CLPsych is a workshop focusing on language technology applications in mental health. We would also like to enrich the model by using topic models (Steyvers et Griffiths, 2007) to represent the post content or word embedding (Baroni *et al.*, 2014). Finally, we would like to develop a model based on deep learning in order to avoid the feature engineering step and to give insights on how well such approach could capture the discriminating features and reinvest our previous work on keyword extraction (Mothe *et al.*, 2018).

Acknowledgements : This work is supported by ***Ministère des Affaires étrangères et du Développement international*** under scholarship ***EIFFEL-DOCTORAT 2017/n°P707544H*** for the PhD thesis of Faneva Ramiandrisoa.

7. Bibliography

Almeida H., Briand A., Meurs M., « Detecting Early Risk of Depression from Social Media User-generated Content », *Working Notes of CLEF*, 2017.

Armstrong W., Using Topic Models to Investigate Depression on Social Media, Technical report, University of Maryland, USA, 2015. Scholarly paper.

[12] `http://clpsych.org/shared-task-2016/`

Baroni M., Dinu G., Kruszewski G., « Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors », *Proceedings of the 52nd Annual Meeting of ACL*, 2014.

Choudhury M. D., Gamon M., Counts S., Horvitz E., « Predicting Depression via Social Media », *Proceedings of the Seventh International Conference on Weblogs and Social Media*, The AAAI Press, 2013.

Colombo G. B., Burnap P., Hodorog A., Scourfield J., « Analysing the connectivity and communication of suicidal users on Twitter », *Computer Communications*, 2016.

Cook B. L., Progovac A. M., Chen P., Mullin B., Hou S., Baca-Garcia E., « Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid », *Comp. Math. Methods in Medicine*, 2016.

Coppersmith G., Dredze M., Harman C., Hollingshead K., Mitchell M., « CLPsych 2015 Shared Task : Depression and PTSD on Twitter », *Proceedings of CLPsych@NAACL-HLT*, 2015.

Errecalde M. L., Villegas M. P., Funez D. G., Ucelay M. J. G., Cagnina L. C., « Temporal Variation of Terms as Concept Space for Early Risk Prediction », *Working Notes of CLEF*, 2017.

Farias-Anzaldua A. A., Montes-y-Gómez M., López-Monroy A. P., González-Gurrola L. C., « UACH-INAOE participation at eRisk2017 », *Working Notes of CLEF*, 2017.

France D. J., Shiavi R. G., Silverman S. E., Silverman M. K., Wilkes D. M., « Acoustical properties of speech as indicators of depression and suicidal risk », *IEEE Trans. Biomed. Engineering*, 2000.

Losada D. E., Crestani F., « A test collection for research on depression and language use », *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2016.

Losada D. E., Crestani F., Parapar J., « eRISK 2017 : CLEF Lab on Early Risk Prediction on the Internet : Experimental Foundations », *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association*, vol. 10456 of *Lecture Notes in Computer Science*, Springer, 2017.

Low L. A., Maddage N. C., Lech M., Sheeber L., Allen N. B., « Detection of Clinical Depression in Adolescents' Speech During Family Interactions », *IEEE Trans. Biomed. Engineering*, 2011.

Malam I. A., Arziki M., Bellazrak M. N., Benamara F., Kaidi A. E., Es-Saghir B., He Z., Housni M., Moriceau V., Mothe J., Ramiandrisoa F., « IRIT at e-Risk », *International Conference of the CLEF Association, CLEF 2017 Labs Working Notes, Dublin, Ireland, September, 11/09/2017-14/09/2017.*, vol. 1866 of *ISSN 1613-0073*, CEUR Workshop Proceedings, 2017.

Mohammad S., Turney P. D., « Crowdsourcing a Word-Emotion Association Lexicon », *Computational Intelligence*, 2013.

Mothe J., Ramiandrisoa F., Rasolomanana M., « Automatic Keyphrase Extraction using Graph-based Methods », *ACM Symposium on Applied Computing (SAC)*, ACM, 2018.

Mowery D. L., Park A., Bryan C., Conway M., « Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health », *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@COLING*, 2016.

Nakamura T., Kubo K., Usuda Y., Aramaki E., « Defining Patients with Depressive Disorder by Using Textual Information », *AAAI Spring Symposium Series, North America*, 2014.

Ozdas A., Shiavi R. G., Silverman S. E., Silverman M. K., Wilkes D. M., « Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk », *IEEE Trans. Biomed. Engineering*, 2004.

Pennebaker J. W., Mehl M. R., Niederhoffer K. G., « Psychological aspects of natural language use : Our words, our selves », *Annual review of psychology*, 2003.

Resnik P., Armstrong W., Claudino L. M. B., Nguyen T., Nguyen V., Boyd-Graber J. L., « Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter », *Proceedings of CLPsych@NAACL-HLT*, The Association for Computational Linguistics, 2015.

Resnik P., Garron A., Resnik R., « Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students », *Proceedings of EMNLP*, 2013.

Rude S., Gortner E., Pennebaker J., « Language use of depressed and depression-vulnerable college students », *Cognition & Emotion*, 2004.

Sadeque F., Xu D., Bethard S., « UArizona at the CLEF eRisk 2017 Pilot Task : Linear and Recurrent Models for Early Depression Detection », *Working Notes of CLEF*, 2017.

Sagen U., Finset A., Moum T., Mørland T., Vik T. G., Nagy T., Dammen T., « Early detection of patients at risk for anxiety, depression and apathy after stroke », *General hospital psychiatry*, 2010.

Schwartz H. A., Eichstaedt J. C., Kern M. L., Park G. J., Sap M., Stillwell D., Kosinski M., Ungar L. H., « Towards assessing changes in degree of depression through Facebook », *Proceedings of CLPsych@NAACL-HLT*, 2014.

Shen G., Jia J., Nie L., Feng F., Zhang C., Hu T., Chua T., Zhu W., « Depression Detection via Harvesting Social Media : A Multimodal Dictionary Learning Solution », *Proceedings of IJCAI*, 2017.

Steyvers M., Griffiths T., « Probabilistic topic models », *Handbook of latent semantic analysis*, 2007.

Trotzek M., Koitka S., Friedrich C. M., « Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression », *Working Notes of CLEF*, 2017.

Villegas M. P., Funez D. G., Ucelay M. J. G., Cagnina L. C., Errecalde M. L., « LIDIC - UNSL's Participation at eRisk 2017 : Pilot Task on Early Detection of Depression », *Working Notes of CLEF*, 2017.

Wang X., Zhang C., Ji Y., Sun L., Wu L., Bao Z., « A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network », *Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2013 International Workshops*, vol. 7867 of *Lecture Notes in Computer Science*, Springer, 2013.