Comparison of machine learning models for early depression detection from users' posts

Josiane Mothe, Faneva Ramiandrisoa, Md Zia Ullah

Contents

List of Contributors

Josiane Mothe
ESPE, Univ. Toulouse Jean-Jaurès, Univ. de Toulouse, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, Toulouse, France, e-mail: josiane.mothe@irit.fr

Faneva Ramiandrisoa
Univ. de Toulouse, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, Toulouse, France, e-mail: r.faneva.mahery@gmail.com

Md Zia Ullah
Univ. de Toulouse, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, Toulouse, France, e-mail: mdzia.ullah@irit.fr

Part I
The Best of eRisk Five Years Labs

Chapter 1
Comparison of machine learning models for -early- depression detection from users' posts

Josiane Mothe, Faneva Ramiandrisoa, Md Zia Ullah

Abstract With around 300 millions people worldwide suffering from depression, the detection of this disorder is crucial and a challenge for individual and public health. As with many diseases, early detection means better medical management; the use of social media messages as potential clues to depression is an opportunity to assist in this early detection by automatic means. This chapter is based on the participation of the CNRS IRIT laboratory in the early detection of depressive people (e-Risk) task at the CLEF evaluation forum. Early depression detection differs from depression detection in that it considers temporality; the system must make its decision about a user's possible depression with as little data as possible. In this chapter we re-evaluate the models we have developed for our participation at e-Risk over the years on the different collections, to obtain a more robust comparison. We also add new models. We use well-established classification methods, such as Logistic regression, Random forest, and Support vector machine. The input data of the users, the system should detect if they are depressed, are represented as vectors composed of (a) various task-oriented features including depression related lexicons and (b) word and document embeddings, extracted from the users' posts. We perform an ablation study to analyze the most important features for our models. We also use BERT deep learning architecture for comparison purposes, both for depression detection and early depression detection. According to our results, well-established machine learning models are still better than more modern models for -early- detection of depression.

Josiane Mothe
ESPE, Univ. Toulouse Jean-Jaurès, Univ. de Toulouse, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, Toulouse, France e-mail: josiane.mothe@irit.fr

Faneva Ramiandrisoa
Univ. de Toulouse, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, Toulouse, France e-mail: r.faneva.mahery@gmail.com

Md Zia Ullah
Univ. de Toulouse, IRIT, UMR5505 CNRS, 118 Rte de Narbonne, Toulouse, France e-mail: mdzia.ullah@irit.fr

1.1 Introduction

According to the World Health Organization (WHO), the number of people with mental disorders worldwide is increasing day by day; the most common mental disorders are depression and anxiety. Globally, the number of people suffering from depression is more than 300 million over the world (See Figure 1.1); this corresponds to an increase of more than 18% between 2005 and 2015[1]. The WHO has also found that depression affects more women than men. Using the same data, the BMC Medicine journal found that France is the most affected country with a rate of 21% followed by the United States (19.2%)[2]. In France, it is estimated that nearly one person in five suffered or will suffer from depression during their lifetime. Thus, the detection of this disorder is crucial and constitutes a challenge for individual and public health.

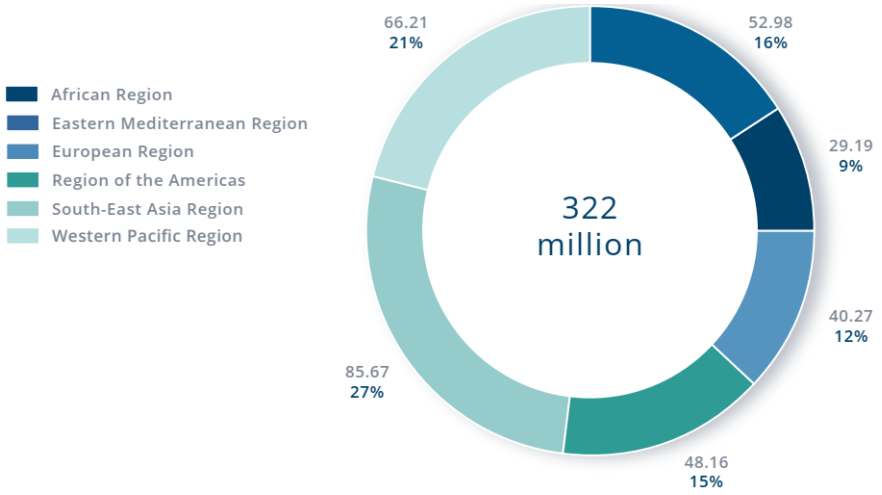


Fig. 1.1: Cases of depressive disorder by region (continent) in 2017 according to WHO [30].

Clinical factors are key to detect patients at risk. Some studies have also shown that depressed people use particular linguistic expressions [12, 24, 32]. Depressed people would use some specific linguistic patterns such as excessive use of personal pronouns, past tense or negative emotions. People's writings can therefore be used as cues to try to detect their psychological state and if they are possibly depressed.

In recent years, the emergence of social networks such as Facebook, Twitter or Reddit has allowed people to share their personal experiences, ideas or thoughts in

[1] https://www.la-depression.org/, accessed January 28, 2021

[2] http://www.doctissimo.fr/psychologie/news/la-france-pays-le-plus-touche\\-par-la-depression, accessed January 28, 2021

a simple way. According to Bhavani and Kulkarni [19], people prefer to express themselves online rather than offline. This phenomenon has generated a lot of data which is an opportunity for many research topics, and therefore also for the medical field [8]. Besides, a study by Marriott and Buchanan [26] reports there is no significant difference between an individual's online and offline personality in terms of authenticity. Social media is thus a good source for studying the ability of an automatic system to help on the detection of depressed people. It becomes possible to study the writings of these users in order to try to detect depressed users based on linguistic indicators.

Most approaches in the literature on detecting depression on social networks use supervised learning methods trained on manually annotated datasets. Different groups of features have been used such as the use of emoticons [47], the time of publication on the social network [5] and the themes mentioned [39]. Different machine learning techniques have been applied to this type of problem, from well-established categorization techniques to more recent deep learning models [16]. Nonetheless, these studies consider different datasets and there is no detailed analysis to compare different models and different types of features for early detection of depression.

Early depression detection implies that time is considered; the idea is to detect the early signals in the texts. Social media posts are generally time-stamped and thus early detection is technically possible. The challenge then is to re-consider the models and features used for depression detection for the task of early detection. Early detection means less information from each user to train the models; this may impact more on some models than others. In the experiments, this is introduced through chunks of texts where the first chunk contains the first part of the users' posts. This notion is further detailed in Other book chapter CH2

In this chapter, we present a supervised learning approach for -early- detection of depression. This work has started on the e-Risk task datasets of the 2017 and 2018 editions presented in other chapter from this book, CH2 **??**. On the one hand, we used well-established classifiers including Random forest and Support vector machines. On the other hand, we used the more recently developed BERT modeling [10]. With the first type of models, we need to define the features that will be used in the vectors that will represent the information. While we have re-used some features from the literature, we also consider new features in our models. We also investigate the use of word and document embeddings in this chapter. BERT learns the semantic relationships between the words in a sentence to create language representations which are later used as features (instead of manually defining a set of features, which implies a certain level of domain knowledge). Considering the impressive results of BERT-based models, we hypothesize that BERT will be better on the condition that it has enough information to be fine-tuned, while well-established models will better answer the early detection of depression. This chapter presents the multiple models we developed and evaluated on the task of depression detection and early depression detection. We also deeply analyze these results in order to provide useful insights into the advantages of the different versions of the models.

This chapter is organized as follows: Section 1.2 presents features and models used to identify signs of depression from social media in related work. Section 1.3 details the data modeling, that is to say, the features we use to build the models to be trained for early detection of depression. Section 1.4 describes the different machine learning models that we use to detect depression. The experimental framework, collection statistics, and the evaluation metrics are stated in Section 1.5. Section 1.6 presents the results on 2017 and 2018 e-Risk challenge datasets and includes a deep analysis of the results obtained by the different machine learning models as well as the most important features for feature-based models. Section 1.8 concludes this chapter.

1.2 Related work

Most of the work on the identification of depression in social media attempts to make the diagnosis process automatic. Many of these works are based on supervised learning methods and use techniques such as statistical methods and natural language processing.

Approaches based on statistical methods collect statistical elements by quantifying activity on social networking platforms, social isolation, number of friends, user network based on users' interests or mutual connections, etc. As for approaches that use natural language processing techniques, they are mainly based on linguistic/semantic analysis of posts such as analysis of emotions expressed, sentence structure, etc. Some studies combine both types of features [5, 4].

In the case of postpartum depression (a mood disorder that can affect mothers after childbirth), De Choudhury et al. [4] defined several measures to characterize the differences between mothers with postpartum depression and those without, that they evaluated on Facebook posts of 165 users from which 28 with postpartum depression. The authors defined 49 characteristics that they grouped into 4 categories: *user characteristics* which include characteristics that measure social network activity such as the number of posts per day, *social capital* which includes measures on the interaction with others such as the number of mentions of the word *love* for friends' posts, *content characteristics* that are computed from the content of publications such as emotion analysis and *linguistic style* that measures behavioral changes based on the use of linguistic styles in publications. The authors found that 35 of the 49 characteristics significantly (considering t-test) distinguish between the two groups. They developed several regression models to predict whether a mother is at risk of postpartum depression. On the prenatal data, the model using all characteristics achieved the best result, however, their best model overall was the one based on prenatal data combined with postnatal data. In our work, we were inspired or adapted some characteristics from their *content characteristics* and *linguistic style* categories.

In another study, De Choudhury *et al.* [5] defined several features and feature categories to characterize Twitter users concerning depression. The authors crowd-

sourced 476 volunteer Twitter user' accounts of which 171 had a high CES-D[3] score used as a cue to automatically detect depression. The authors defined 43 characteristics that they grouped into the following categories: *engagement*, *self-centered social graph (self-centered network)*, *emotion*, *language style* and *depression language*. Four more demographic characteristics were considered: *age*, *gender*, *level of education* and *income*. The authors then analyzed the data to try to distinguish between depressed and non-depressed behaviors. The authors found that those with depression engage in less social activity, display more negative emotions, pay more attention to themselves, show an increase in relational and medical concerns, and express more religious thoughts. They also found that even though their egocentric networks are small, depressed users appear to belong to closely grouped networks and are generally very connected to their network. De Choudhury *et al.* [5] also created several models, using different combinations of characteristics and different classification methods to predict whether a user is depressed or not based on their tweets. They found that the SVM classifier trained with features, reduced in size with the principal component analysis method, provided the best result with an accuracy of 0.70 and a precision of 0.74. In our work, we adapted some of the characteristics from the groups *emotion*, *commitment* and *depression language*.

The work carried in the e-Risk task (cf. Chapter 2 **??** of this book) are closer to ours than the ones from De Choudhury *et al.* [5]. E-Risk differs from the above-mentioned work in that it aims to detect as early as possible signs of depression from users' texts, in this case, Reddit [4] forum posts. For early detection, the temporality is considered by dividing the set of chronologically ordered posts of a given user into 10 partitions (also called chunks) containing 10% of the user's texts each. Hence, early depression detection is to predict whether a user is depressed by using as few partitions as possible. Early detection is measured by ERDE_x which has been defined for the e-Risk task (see section in Chapter 2 Section **??** for more details) in addition to recall, precision, and F_1 measures.

Trotzek *et al.* [43] developed five models for early depression detection for e-Risk 2017. All these models use linguistic meta-information extracted from each user's texts and combine them with classifiers based on a bag-of-words, paragraph vector, latent semantic analysis, and recurrent neural networks. Their bag of word model performed the best considering the F_1 measure while their model based on paragraph vectors performed best considering the precision and the ERDE_5. In e-Risk 2018, the same authors used four machine learning models, two are based on convolutional neural network; the other two are based on features computed from the user's texts either considering bags of words or a combination of linguistic metadata and bags of words. They also considered an ensemble model that combines the results obtained from three (two models based on convolutional neural network and one features based model) of these four models [45]. The model that considers bags of words only performed the best according to the ERDE_{50} and F_1 measure in 2018. They also obtained the best result on the ERDE_5 and ERDE_{50} for e-Risk 2017 dataset after the

[3] CES-D stands for Center for Epidemiologic Studies Depression who provides a questionnaire that can be used to detect depression [36].

[4] Reddit is a social news aggregation, web content rating, and discussion website (`reddit.com`)

competition, with another ensemble model which combines a convolutional neural network-based model using fastText as word embedding [17] trained on Wikipedia articles and a model based on linguistic features [43].

Funez *et al.* [13] proposed a model based on concise semantic analysis [22] which extracts a few concepts from category labels, here depressed vs. non-depressed. Each term is then represented as a vector in this concept space and a text or a set of texts is represented as the centroid of the vectors terms it is composed of. This is used to represent each user considering their posts. For early depression detection, the authors considered the temporal difference of the representation of the users by using previous posts (first chunk) and current posts. They named their model Temporal Variation of Terms (TVT). TVT model obtained better performance for $ERDE_5$ and $ERDE_{50}$ measures than the standard bag-of-words model on e-Risk 2017.

Later, Funez *et al.* [14] proposed a variant of TVT called FTVT (Flexible Temporal Variation of Terms) where the number of posts used is flexible. Their other model, SIC, incrementally reads documents and estimates the evidence that document words provide depressive and/or non-depressive labels. SIC classifies a user as depressive as soon as the accumulated evidence of the depressive class surpasses the evidence of the other one.

For their participation to e-Risk 2018, Funez *et al.* submitted five models, three of which are variants of their FTVT approach (hyper-parameter values are changed), from which one obtained the best $ERDE_5$, and two are variants from their SIC approach.

Burdisso *et al.* [2] proposed the SS3 model which builds for each category (depressed vs. non-depressed) a dictionary of words with their frequency on the training set. The dictionary is incrementally updated as new posts are considered. During the classification phase, SS3 calculates a score that measures the relationship between a word w and a category c based on these word frequencies. For classification, SS3 first divides the text into blocks (e.g., paragraphs), then each block is further divided into smaller blocks until the word level is reached. Then, SS3 calculates the $\overrightarrow{gv}$ vector for each word. These vectors are aggregated (e.g., sum, maximum, etc.) to generate the vectors for the upper block in a recursive process up to the input text. Finally, the texts are classified using these vectors according to the category with the highest value. The authors obtained the best result by considering the $ERDE_5$ measure on e-Risk 2017. They also proposed $SS3^{\triangle}$, which differs from SS3 on the classification policy. With $SS3^{\triangle}$, they got the best result considering the $ERDE_{50}$ measure. Burdisso *et al.* improved SS3 with $\tau-$SS3 [1] which dynamically recognizes useful patterns on text streams, i.e., it can learn and recognize n-grams of variable lengths. $\tau-$SS3 performed the best considering $ERDE_{50}$ on the 2018 edition.

Table 1.1 presents the results participants obtained for the early depression detection task at e-Risk (marked-up with a "*") as well as other results that have been obtained using the e-Risk datasets, but after the competition ends.

Table 1.1: Results on the e-Risk task from participants -marked with *- as well as post challenge results. Bold font highlights the best run for the considered collection and measure for an official participation. Best models are different according to the collection and measure.

Models		$ERDE_5$	$ERDE_{50}$	F_1	P	R
2017						
Trotzek [43]	FHDOA*	12.82 %	9.69 %	**0.64**	0.61	0.67
	FHDOB*	**12.70** %	10.39 %	0.55	**0.69**	0.46
UArizona* [41]		17.93 %	12.74 %	0.34	0.21	**0.92**
UNSL* [46]		13.66 %	**9.68** %	0.59	0.48	0.79
TVT [13]		13.13 %	8.17 %	0.54	0.42	0.73
SS3 [2]		12.60 %	8.12 %	0.52	0.44	0.63
τ−SS3 [1]		12.60 %	7.70 %	0.55	0.43	0.77
Trotzek [43]	(a)	12.13 %	8.77 %	0.71	0.71	0.71
	(b)	13.52 %	7.29 %	0.55	0.41	0.85
	(c)	13.32 %	11.33 %	0.73	0.77	0.69
2018						
FTVT* [14]		**8.78** %	7.39 %	0.38	0.48	0.32
BCSGB* [45]		9.50 %	**6.44** %	**0.64**	0.64	0.65
RKMVERIC* [33]		9.81 %	9.08 %	0.48	**0.67**	0.38
UDCB* [3]		15.79 %	11.95 %	0.18	0.10	**0.95**
τ−SS3 [1]		9.48 %	6.17 %	-	-	-

1.3 Information modeling

For depression detection, the objective is to detect whether users are predicted to be depressed giving their posts. For early depression detection, the objective is close to the previous one, but the detection should be based on a minimum of posts considering their chronological order. To tackle this problem, we rely on supervised machine learning techniques whose principle is to learn a decision function from a set of labeled examples. In the case of depression detection, an example is a user described by a list of features and the binary label, indicating whether the user is depressed or not.

Considering e-Risk data, we have no information on the users for anonymity reasons, which is often the case when considering personal data, but we know their posts. User features have thus to be extracted from a series of user's posts, either all the posts for depression detection, or a subset of them for early depression detection. To extract the users' features, we considered two approaches: feature-based and word/document embedding based. We then successively use the two representations on well-establish machine learning algorithms (see Sub-section 1.4.1) to learn the model for predicting depression.

1.3.1 Feature-based representation

The features we considered are many (256 in total). They come both from the literature [5, 4, 43] or from our own developments [25, 38, 37, 16].

We categorized these features into six meta-categories (META1 to META6) each of which are divided in turn into categories (numbered with roman numerals from I to XVI). This hierarchical representation of features will be used for deep analysis of the impact of features on the results. The six meta-categories are as follows: representation of full texts, lexicons on depression, temporal aspects, writing style, emotion, and lexical categories. These meta-categories, the associated categories, and features are detailed in Table 1.2.

Table 1.2: Description of the features used to represent the users' posts. Similar types of features are grouped into categories; and similar categories are further grouped into meta-categories.

Cat.	Number	Name	Hypothesis or tool/resource used
		META$_1$: REPRESENTATION OF FULL TEXTS	
I	1-18	Bag-of-words	18 most frequent uni-grams in the e-Risk 2017 training set [16]
II	19-22	Part-Of-Speech frequency	Higher usage of adjectives, verbs, adverbs, and lower usage of nouns [5]
		META$_2$: LEXICONS ON DEPRESSION	
III	23	Depression symptoms and related drugs	From Wikipedia list[5] and [5]
	24	Relevant 3-grams	25 3-grams [6]
	25	Relevant 5-grams	25 5-grams [6]
IV	26	Frequency of "depress"	Depressed people talk often about the depression
	27	Related words to depression	Words: "sleep," "depress," and "sad"
V	28	Drugs name	The chemical and brand names of antidepressants from WebMD available in United States[6] [43]
		META$_3$:TEMPORALITY	
VI	29	Temporal expressions	High use of words that refer to past: last, before, ago, ...[29]

[5] `http://en.wikipedia.org/wiki/List\_of\_antidepressants`, accessed on 23/02/2017

[6] `http://www.webmd.com/depression/guide/depression-medications-antidepressants`, accessed on 10/01/2018

VII	30	Ratio of Posting Time	High frequency of publications in deep night (00 pm - 07 am) [5]
	31-34	Season of a year (4 seasons in total)	Frequency of publications in season (one season corresponds to 3 months)
VIII	35	Past frequency	Combination of temporal expressions and past tense verbs
	36	Past tense verbs	Depressive people talk more about the past
	37	Past tense auxiliaries	Same motivation as above

META$_4$: WRITING STYLE

IX	38	Average number of posts	Depressed users have a much lower number of posts
	39	Average number of words per post	Posts of depressed user are more longer
	40	Minimum number of posts	Generally depressive users have a lower value
	41	Average number of comments	Depressed users have a much lower number of comments
	42	Average number of words per comment	Comments of depressed and non depressed users have different means
X	43	Gunning Fog Index	Estimate of the years of education that a person needs to understand the text at first reading [43]
	44	Flesch Reading Ease	Measure how difficult to understand a text is [43]
	45	Linsear Write Formula	Developed for the U.S. Air Force to calculate the readability of their technical manuals [43]
	46	New Dale-Chall Readability	Measure the difficulty of comprehension that persons encounter when reading a text. It is inspired from Flesch Reading Ease measure [43]
XI	47-51	First person pronouns	High use of : *I, me, myself, mine, my*
	52	*I* subject of *be*	High use of *I'm* [40]
	53	All first person pronouns	Sum of frequency of each first pronoun [47]
	54	*I* in subjective context	Depressive users refers to themselves frequently (all *I* targeted by an adjective)
XII	55	Over-generalization	Depressed users tend to use intense quantifiers and superlatives [29]

XIII	56	Negation	Depressive users use more negative words like: *no, not, didn't, can't*, ..., etc.
XIV	57	Capitalized	Depressive users tend to put emphasis on the target they mention
	58	Punctuation marks	! or ? or any combination of both tend to express doubt and surprise [47]
	59	Emoticons	Another way to express sentiment or feeling [47]

META$_5$: EMOTION

XV	60	Emotions	Frequency of emotions from specific categories: anger, fear, surprise, sadness and disgust
	61-62	Sentiment	Use of NRC-Sentiment-Emotion-Lexicons[7][28] to trace the polarity in users writings

META$_6$: LEXICAL CATEGORIES

XVI	63-256	Empath	All 194 Empath categories [8] [11]

Whatever the feature is, its value is calculated by averaging the number of occurrences found at the post level over all the user' posts to consider; values are normalized.

1.3.1.1 META$_1$: REPRESENTATION OF FULL TEXTS

Group I: These features are extracted from posts known to be written by people annotated as depressed in e-Risk 2017 training set. We extracted the 50 most frequent uni-grams from the depressed users' texts and learnt a first model (Random forest based) whose task was to detect depressed users. We then selected the 18 uni-grams that the chi-squared ranking filter considers as the most important. Although these terms may be collection-dependent, we kept them across the models.
Group II: They correspond to the frequency of Part-Of-Speech categories (adjectives, verbs, adverbs, and nouns). According to Choudhury *et al.*, people who want to commit suicide were characterized with a higher use of verbs, adjectives, and adverbs, but lower use of nouns [5]. We used these cues as detection characteristics

[7] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm, accessed on 23/02/2017

[8] http://empath.stanford.edu

for depression even though a depressed person is not necessarily suicidal. Although the World Health Organization has shown that depression is one of the major causes of suicide in the world[9].

1.3.1.2 META$_2$: LEXICONS RELATED TO DEPRESSION

Group III: Depressed users often mention depression symptoms and antidepressants names in their texts. We used the list of depressive symptoms and antidepressant names obtained from [5] and Wikipedia[10]. The value of this characteristic is the normalized frequency of the uni-grams from the lists. Depressed users use group of words such as “want to die" (tri-grams) or “to take my own life" (5-grams), etc. We used the 25 3-grams and 25 5-grams which were the most frequent expressions of a potentially suicidal thought, from a collection of tweets by Colombo *et al.* [6].
Group IV: The first feature of this group corresponds to the “depression" term and its syntactic variants. The next feature is the frequency of the words: “sleep," “depress," and “sad."
Group V: A list of antidepressants from WebMD available in United States[11] [43].

1.3.1.3 META$_3$: TEMPORALITY

Group VI: Depressed users often talk about the past [29]. We used a list of English words referring to the past: yesterday, last, before, ago, past, back, earlier, later, nostalgia.
Group VII: Depressed users tend to post publications late at night [5]. We capture the ratio of publications between 12am (midnight) and 7am. We also calculate the characteristic seasons of a year by dividing the 12 months of the year into 4 seasons (season 1: December, January, and February; season 2: March, April, and May; etc.). Each season corresponds to a characteristic, thus considering the seasonality of depression [31].
Group VIII: The features in this group (past frequency, past tense verbs, and past tense auxiliaries) make allusion to the past.

1.3.1.4 META$_4$: WRITING STYLE

Group IX: This group gathers different features related to the number of posts and the number of words per post.
Group X: These features measure the readability of the posts, thus the complexity of the sentences and texts.

---

[9] `https://www.who.int/news-room/fact-sheets/detail/depression`

[10] `http://en.wikipedia.org/wiki/List\_of\_antidepressants`, accessed on 23/02/2017

[11] `http://www.webmd.com/depression/guide/depression-medications-antidepressants`, accessed on 10/01/2018

Group XI: These features are related to the first person use in the posts.
Group XII: Depressed persons are prone to over-generalize [29]. For example, instead of criticizing someone who has hurt them, a depressed person may write "all men/women are xx."
Group XIII: When suffering from depression, people usually express with much more negative feeling in their writings than healthy individuals. We therefore decided to consider this characteristic.
Group XIV: Depressed people are more likely to focus on the target they mention. The use of capitalized words is one way of expressing this emphasis. Their writings are a true reflection of their moods. In this case, special punctuation marks, such as question marks or exclamation marks, tend to encourage users to express their doubts or surprises towards a target [48]. Emoticons are another way of expressing how people feel. Xinyu Wang *et al.* [47] have shown that taking emoticons into account is important in detecting depression.

1.3.1.5 META$_5$: EMOTION

Group XV: Depression is closely related to emotions and sentiments. Depressed people tend to be more subjective about what they mention or write, which is why we have assumed that it is useful to trace the polarity and emotions of their texts.

1.3.1.6 META$_6$: LEXICAL CATEGORIES

Group XVI: The idea here is to analyze the publication through lexical categories. For this, we used the empath python library[12] that analyzes texts through lexical categories.

1.3.1.7 Embedding representation

A user is represented by a vector built from the history of posts available that each of them has. Here, we use embeddings based representation both at the word level from Word2Vec [27] and at the sentence level with a combination of two Doc2Vec [20] variants.

Word embedding refers to the representation of a word in a semantic space as a vector of numerical values. Words that are semantically and syntactically similar tend to be close in this embedding space. Document embedding refers to the representation of a document's content using a vector of numerical values. The goal of Doc2Vec is to create a vector representation of a document/paragraph/sentence, disregarding of its length. The Doc2Vec model can work very well when trained on a small corpus compared to Word2Vec [44].

[12] `http://empath.stanford.edu`

To represent users from their posts/comments at the word level, we use the pre-trained Word-vector which was trained on GoogleNews corpus using Word2Vec model [27] following the Skip-gram architecture with negative sampling and a window size of 10 words. The pre-trained model contains 300-dimensional vectors for 3 million words and phrases. We first extract the word vector from Word2Vec model for each word in a post. When we could not find a word in the model, we represent it with a zero vector of 300 dimension. We average the word vectors of every word from a post in row-wise. Then, for a given user, we aggregate all the vectors from that user's posts by averaging the corresponding vectors. This average aggregated vector is considered as the vector representation for that user.

To represent users from their posts/comments at the sentence level, we use the Sentence-vector which is a combination of two Doc2Vec variants trained using Doc2Vec [20] model based on Word2Vec. First, we extract a sentence embedding vector for each post from two Doc2Vec model variants, (a) distributed memory model and (b) distributed bag-of-words model (see Figure 1.2). Then for a given user, we aggregate all the vectors from that user's posts/comments by averaging the vectors; we thus calculate the centroid vector for each user.

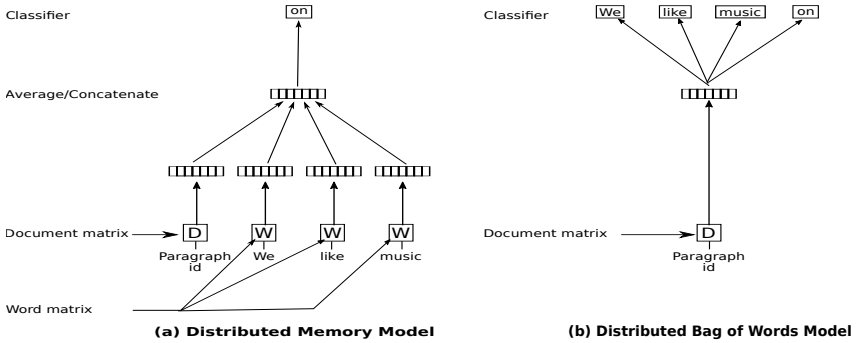


Fig. 1.2: Doc2Vec models inspired from [20].

More specifically, we trained the two following Doc2Vec model variants (see Figure 1.2) with the dataset of 2017 and 2018 separately, thus creating 4 models (see Section 1.5):

- A distributed memory model which is an extension of Word2Vec CBOW model: instead of using only words to predict the next word, this Doc2Vec model adds another input feature vector which is unique to the paragraph (ID of the paragraph in the Figure 1.2 (a)). During training, the word matrix W and the paragraph matrix D are trained (i.e., change their weight), and at the end, a column in the paragraph matrix D represents a unique vector of a paragraph.
- A distributed bag-of-words model which is similar to the Word2Vec skip-gram model but uses the paragraph matrix to classify all the words from the document (see Figure 1.2 (b)). Instead of predicting the next word, it uses the paragraph

matrix to classify all the words in a paragraph. Specifically, during the training phase, a classifier is trained to decide whether the words belong to the paragraph or not. This model consumes less memory, as there is no need to save word vectors as in the distributed memory model.

Mikolov *et al.* [20] recommend using a combination of the two models (i.e. distributed memory model and distributed bag-of-words model) to represent a text/document; although distributed memory model generally achieves state-of-the-art performances for most tasks [7]. To represent a sentence/post, we concatenate the two vectors of 100 dimension representing the sentence/post, obtained by the two Doc2Vec model variants, thus forming a single vector of 200 dimensions.

1.4 Machine learning models

1.4.1 Well-established machine learning models

We consider the following well-established machine learning (ML) algorithms from different categories:

- Random Forest (RF): a Random forest classifier with 100 decision trees,
- SGD_Logloss (SGD-log) logistic regression: a stochastic gradient descent classifier trained using the “log" loss function,
- SVM_Linear (SVM-$linear$): a support vector machine classifier with the linear kernel,
- SVM_RBF (SVM-rbf): a support vector machine classifier with the RBF kernel,
- Neural network (NN): a multi-layer perceptron classifier where the configuration is one hidden layer with 100 units, the Relu activation function, and the “log" loss function using LBFGS optimizer,
- K-nearest-neighbor (KNN): a 3-nearest neighbors classifier.

Unless otherwise stated, we kept the default values of the respective hyper-parameters of all machine learning algorithms from Scikit-learn (version 0.22.1) [34]. To produce the results of all machine learning algorithms, we used the random_state=42 (the scikit-learn parameter that handles the random seed for the models and which is required for RF for sampling of the features and the randomness of the bootstrapping of the samples, SGD for shuffling the data, SVM for probability estimate, NN controls).

These classical classifiers are not designed to solve detection problems at the earliest possible time while we deal with such problems in this work. Early detection problems integrate a time dimension in the classification, in other words, they use data streams as input. That is, at a given instant t, only a partition of the data is available for classification; the further forward in time, the more data is available. The objective is to be able to decide accurately using the minimum amount of data, i.e. as quickly as possible over time.

To allow classifiers to deal with the early detection problem, we adopt a classification process where we integrate the time dimension outside the classifiers. This process consists of training classifiers without taking into account the time dimension, in other words, using all the data available in the training corpus, and consider the time dimension during the test stage. More precisely during the test stage, when only the first posts are considered (i.e., chunk 1), and for a given user, a trained classifier makes its decision on the user using only the data available at this stage. If the user is predicted as depressed then the classification process stops, otherwise the classifier predicts the user again using the data available at the first stage plus the one available at the second stage (i.e. chunk 1 and chunk 2). This process is repeated and in the end, if the user is still not predicted as depressed when using all the data, then the user is considered as non-depressed. In e-Risk, one stage (or chunk) contains partial data of a user, only 10% of users' writings, chronologically ordered.

1.4.2 BERT-based model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language representation model [10]. BERT model architecture is a multi-layer bidirectional transformer encoder, which considers both left and right contexts in all of its layers. BERT is pre-trained on large corpus using a masking technique: given a sentence, 15% of the input words are masked and during training, the goal is to predict these words. This technique overcomes the limitation of unidirectional processing and is also superior to language models that combine right-to-left and left-to-right processing, see [10, 35].

Pre-trained Google BERTs are freely available on github[13]. The BERT_{Base} model contains 12 layers of size 768, 12 self-attention heads and 110M parameters, while the BERT_{Large} model contains 24 layers of size 1,024, 16 self-attention heads and 340M parameters. BERT can be fine-tuned to answer specific downstream tasks. BERT_{Base} model has two variants: $\text{BERT}_{Base-cased}$ which is case sensitive and makes a difference between for example "english" and "English" words, and $\text{BERT}_{Base-uncased}$ which is not case sensitive.

In this work, we fine-tune Google's $\text{BERT}_{Base-uncased}$ pre-trained model on the training data using an 80/20 subdivision where 20% percent of the training set is used as the validation set. This 80/20 subdivision of training set is performed following the stratified fashion using Scikit-learn's "train_test_split" function where the random_state is 42. We add two additional dense layers (512 and 2 units, respectively) at the end of the 12-layer pre-trained network. During the fine-tuning, we freeze all parameters of the first 12-layers of the network (pre-trained part) but allow learning the new added two layers. We use cross-entropy loss function, AdamW optimizer, $1e{-}5$ learning rate, 32 batch size, 10 epochs, and a maximum of 128 tokens per sentence. Training is repeated on the entire training dataset across epochs. We store

[13] `https://github.com/google-research/bert`, accessed on 02/02/2021

the model checkpoint if the validation loss is lower than training loss after each epoch. We name this BERT's fine-tuned model as "BERT-fine-tuned-natural."

The depression dataset is unbalanced since the majority of the posts/comments are not depressive individually and collectively over all users. To tackle the unbalance issue, we balance the classes when fine-tuning the Google's $\text{BERT}_{Base-uncased}$ pre-trained model. The balancing is done via weighting the classes in the loss function. Therefore, we first compute the class weights for the labels in the training set and pass these weights to the loss function [18]. We keep all the hyper-parameters exactly the same as "BERT-fine-tuned-natural." We name this fine-tuned model as "BERT-fine-tuned-balanced."

During testing stage, we evaluate both "BERT-fine-tuned-natural" and "BERT-fine-tuned-balanced" models. We use our BERT-fine-tuned model (either "BERT-fine-tuned-natural" or "BERT-fine-tuned-balanced") to obtain the predicted probability for the testing samples. We assume that all the posts/comments written by a user labeled as depressed are also depressed. In the end, we need to classify the user as depressed or not. To tackle the early detection problem for a user, we follow a cumulative approach and also consider two aggregation functions (i.e., maximum or mean) to calculate the probabilities. We apply the first aggregation function (maximum or mean) on the predicted probabilities for the posts/comments of each chunk by a user. Then, when combining two or more chunks, we again apply the second aggregation function (max or mean) on the first aggregated values of two or more chunks by a user. Moreover, to tackle the detection problem for a user, we apply the same aggregation strategy that we use for early detection.

1.5 Experimental framework

We developed a series of experiments whose objective is to compare different models on the e-Risk -early- depression task. More precisely we aim at comparing well-established machine learning methods with the more recent BERT model. With regard to well-established machine learning, we also investigate the variety of features we described in Section 1.3. For this, we used e-Risk collection that is briefly recall here as well as different evaluation measures also presented shortly below. More details can be found in this book, Chapter xx

1.5.1 Collections

In the evaluation stage, we use the two collections available for early depression detection: CLEF e-Risk 2017 and 2018 collections. E-Risk 2018 has more users and more posts for both depressed and non-depressed users than e-Risk 2017. The data is naturally non-balanced considering depressed vs. non-depressed users. The

proportion of users belonging to the two classes on the training and testing datasets are similar.

Table 1.3: Distribution of training and test data from the e-Risk 2017 and 2018 collection on early depression detection task.

Number		Training		Test	
		Depressed	Non-depressed	Depressed	Non-depressed
2017	Users	83	403	52	349
	Posts	4,911	91,381	1,928	65,735
	Comments	25,940	172,791	16,778	151,930
2018	Users	135	752	79	741
	Posts	6,839	157,116	7,672	169,930
	Comments	42,718	324,721	32,993	333,852

1.6 Results and discussion

We first evaluate the models considering a standard depression detection task. In that case, we are not considering early detection, but rather, the system has to predict depression considering the entire set of texts written by the user (Sub-section 1.6.1). Then, we consider early depression detection where the moment in which decisions are taken is crucial (Sub-sections from 1.6.2 to 1.6.4 and Section 7).

1.6.1 Depression detection

In this section, we present the results for depression detection, that is to say without considering its early detection. From the set of user's posts, the system has to predict whether the user is depressed or not. We thus trained each model on the training set and run them on the entire test set, without considering post chunks or partitions.

Table 1.4 presents the results where machine learning models use:

- 256 features and well-established machine learning algorithms (RF, SGD-log, SVM-$linear$, SVM-rbf, Neural network (NN), KNN) (first block),
- Doc2Vec based embeddings with the same well-established machine learning algorithms (second block),
- Word2Vec based embeddings with the same well-established machine learning algorithms (third block), and
- BERT model (fourth block).

We can observe that some of the well-established algorithms perform strangely bad when considering the representation with 256 features. We suspect this is due

Table 1.4: Well-Established machine learning (ML) methods for detecting depression using the 256 features listed in Table 1.2, Word2Vec, Doc2Vec, and BERT-based representations for e-Risk 2017 and 2018 collections. We considered the threshold on the output probability of the classification algorithm as 0.5 to decide whether the user is depressed or not.

		2017			2018		
	ML model	P	R	F_1	P	R	F_1
256	RF	**.73**	**.46**	**.56**	**.71**	**.46**	**.55**
	SGD-log	.17	.10	.12	.32	.27	.29
	SVM-$linear$	**.73**	.15	.25	.67	.23	.34
	SVM-rbf	.00	.00	.00	.62	.10	.17
	NN	.00	.00	.00	.00	.00	.00
	KNN	.23	.17	.20	.15	.10	.12
Word2Vec	RF	.63	.37	.46	.54	.34	.42
	SGD-log	**.73**	.15	.25	**.61**	.32	.42
	SVM-$linear$	.48	.31	.38	.56	.35	.43
	SVM-rbf	.47	.31	.37	**.61**	.35	**.45**
	NN	.46	**.50**	**.48**	.36	.41	.38
	KNN	.43	.38	.40	.48	**.43**	**.45**
Doc2Vec	RF	**.60**	.35	.44	.71	.38	.50
	SGD-log	.41	**.75**	**.53**	.76	.41	**.53**
	SVM-$linear$	.57	.23	.33	**.84**	.33	.47
	SVM-rbf	.57	.23	.33	**.84**	.33	.47
	NN	.49	.58	**.53**	.42	**.65**	.51
	KNN	.53	.46	.49	.52	.46	.49
BERT (Fine-tuned)	Natural-max-max	.25	.79	.38	.25	.82	.38
	Natural-max-mean	**.42**	.19	.26	.62	.19	.29
	Natural-mean-max	.33	.02	.04	**.80**	.10	.18
	Natural-mean-mean	.00	.00	.00	.00	.00	.00
	Balanced-max-max	.13	**1.0**	.23	.10	**1.0**	.18
	Balanced-max-mean	.14	.98	.24	.10	**1.0**	.19
	Balanced-mean-max	.20	.94	.33	.18	.94	.31
	Balanced-mean-mean	.33	.79	**.47**	.32	.70	**.44**

to the sparsity of some of the features that lead to model overfitting. We can also observe that when considering the threshold as 0.5 on the output probability of the classification algorithm, precision is favored compared to recall for those machine learning models. Both the 256 feature models and Word2Vec reach a precision of around 0.7 on the 2017 collection. For the 2018 collection, the 256 features model and Doc2Vec yielded the best precision (0.71 and 0.84, respectively). When considering F_1 measure, that combines recall and precision, RF with 256 features is consistently the best. Overall, RF with the 256 feature representation is more robust across the collections and thus we would recommend it when considering well-established machine learning models. Moreover, on the e-Risk 2017 dataset, KNN classifier achieves the second best F_1.

Now, when considering BERT variants, the best F_1 measure is obtained for Balanced-mean-mean and the second best by Natural-max-max. Whatever the ver-

sion, recall is generally favored (except for one configuration) compared to precision when considering the threshold on the output probability of the classification algorithm as 0.5. Recall is up to 1 for balanced versions on both collections, however, in those cases precision is low. As a result when considering F_1 measure, still the RF using 256 features remains the best. BERT results are rather disappointing. It might be because not enough data is available. Considering these results, we do not expect BERT to perform well when considering early detection, where not enough data is available for proper fine-tuning.

On the variants of BERT we analyzed, we can see that the trade-off between recall and precision is slightly changed for the different versions. The consistently best across collections is the “Balanced-mean-mean" version for balanced distribution (eighth row of fourth block). Since the model is trained on a balanced distribution by weighting the classes, the predicted score is balanced towards both the positive and control classes. That’s why, if we apply the mean aggregation function on the predicted probabilities of posts/comments for each chunk by a user, we obtained a robust estimate on the positive class. When combining multiple chunks by applying the mean function again, we obtained the best overall estimate across chunks. If we rely on the mean of predicted probabilities for the positive class for the balanced model, it produces the best performance across collections for the BERT variants.

Regarding the balanced distribution models, we can see that the greedy max function on the predicted probabilities of posts/comments of each chunk by a user (first max in the names of Table 1.4, fourth block) makes the aggregated score higher than the threshold (0.5). That means, even if the output probability of a single post/comment of a chunk by a user is higher than 0.5, the user is predicted as depressive. Then, whatever the aggregation function applied when combining multiple chunks for a user (fifth and sixth rows of block 4; Balanced-max-max and Balanced-max-mean), it makes the recall closer to 1, but the precision closer to 0.1.

In the case of the model trained on natural distribution (“imbalanced model"), the predicted class is biased to the control class and only a few posts/comments are classified as positive class. If we apply the greedy maximum function on the predicted probabilities of posts/comments for each chunk by a user, we obtained the maximum probability for the rarely predicted positive class. When combining multiple chunks if we apply the maximum function again, we obtained the best predicted probability across chunks. Since the network is biased towards the control class if we rely on the maximum function both across the posts/comments in a chunk and across chunks for the rare positive class, it produces a comparatively better performance (first row of block 4; Natural-max-max). We can also see that the mean function on the predicted probabilities of posts/comments for each chunk makes the mean aggregated score lower, perhaps lower than the threshold (0.5). This indeed makes the recall closer to 0 whatever the aggregation function (mean/max) is applied later to combine multiple chunks (3rd and 4th rows of block 4 in Table 1.4; Natural-mean-max and Natural-mean-mean).

The weak performance of BERT compared to other models on both collections could be linked to the missing annotations at the posts/comments level, e.g., if a user only became ill during the course of their history. In this case, all posts are still

annotated as depressive while only the later posts/comments should be annotated as depressive. This annotation problem may have an impact on the BERT fine-tuning. We keep this problem for future study.

1.6.2 Early depression detection

In this section, we present the results for early depression detection. We trained the models on the entire training set and run the obtained models by steps on the chunks of the test set: first using the first chunk only, then the two first chunks, then the first three chunks, and until the system uses the entire test set. The system decides at each step. At each step, if the system is confident enough on the predicted score, then it can decide to predict the user as depressed (1), otherwise, it waits for more information. When the system gets all the data (last step) of test set it has to decide for all the remaining users depressed (1) or non-depressed (2).

Table 1.5 presents the results with the same blocks as in Table 1.4. We can observe that when compared to Table 1.4 and keeping the same prediction probability 0.5, precision slightly decreases while recall slightly increases, resulting in a slightly improved F_1 measure; this holds for both collections and for most of the models using well-established machine learning, whatever the representation is (256 features, Word2Vec or Doc2Vec). The fact that precision decreases is due to the fact that some decisions are taken with the limited number of posts processed on the user. As expected from previous results in Table 1.4, Random forest with 256 features is the most stable across collections and thus would be the preferred model. It also has among the smallest ERDE values, showing that it is effective for early detection. With regard to BERT, the results are consistent with the one we obtained for -not early- depression detection. That means, for balanced distribution, “Balanced-mean-mean" or “Balanced-mean-max" version (seventh or eighth row of the fourth block) achieves good performance whereas ‘Natural-max-max" or “Natural-max-mean" version produces better for natural distribution (first or second row of the fourth block).

1.6.3 Simplified models

In this section, we focused on simplified models trained using a smaller number of significant features. In Table 1.2, we can see a list of 256 individual features, 16 groups, and 6 meta-groups of features. We hypothesize that some of those features could be important features to detect early depression while other features could on the contrary be noisy. To extract the important features, we applied the Chi-square (Chi2) feature selection [21] on the training set of each collection individually. We found that 126 features are selected for 2017 collection while 156 features are selected for 2018 collection. With these features only, we then trained the models

Table 1.5: Performance of the early depression detection using the well-established machine learning (ML) methods using the 256 features listed in Table 1.2, Word2Vec, Doc2Vec, and BERT based representation for e-Risk 2017 and 2018 collections. We considered the threshold on the output probability of the classification algorithm as 0.5 to decide whether the user is depressed or not.

		2017					2018				
	ML model	P	R	F_1	$ERDE_5$	$ERDE_{50}$	P	R	F_1	$ERDE_5$	$ERDE_{50}$
256	RF	**.63**	**.62**	**.62**	**12.71%**	**9.48%**	**.65**	**.53**	**.58**	**9.41%**	**7.19%**
	SGD-log	.16	.29	.21	15.46%	13.25%	.28	.38	.32	9.53%	8.33%
	SVM-$linear$	.21	.31	.25	14.61%	13.45%	.57	.27	.36	9.82%	9.21%
	SVM-rbf	.00	.00	.00	13.10%	13.10%	.62	.10	.17	9.69%	9.69%
	NN	.03	.04	.04	14.78%	14.53%	.04	.08	.05	11.21%	11.21%
	KNN	.21	.50	.30	14.91%	11.09%	.20	.46	.28	10.45%	8.40%
Word2Vec	RF	**.44**	.58	**.50**	13.24%	**9.20%**	.33	.46	.38	9.36%	7.17%
	SGD-log	.41	.35	.38	**12.61%**	9.82%	.35	.41	.38	**8.77%**	7.12%
	SVM-$linear$	.36	.46	.40	12.98%	9.62%	.32	.49	.39	8.98%	6.94%
	SVM-rbf	.37	.44	.40	12.88%	9.53%	**.39**	.52	**.45**	8.87%	**6.35%**
	NN	.27	.67	.39	14.19%	9.78%	.22	**.68**	.33	10.59%	7.55%
	KNN	.34	**.81**	.48	14.29%	8.85%	.23	.57	.33	10.04%	7.91%
Doc2Vec	RF	**.53**	.50	**.51**	12.73%	9.73%	.56	.51	**.53**	8.92%	6.58%
	SGD-log	.27	**.87**	.41	15.05%	9.40%	.40	.56	.46	8.92%	6.64%
	SVM-$linear$	.42	.37	.39	**12.29%**	10.07%	.49	.43	.46	8.74%	6.89%
	SVM-rbf	.45	.38	.42	12.37%	9.76%	**.58**	.48	.52	**8.51%**	**6.55%**
	NN	.30	.71	.42	14.32%	10.76%	.23	**.81**	.36	10.52%	7.41%
	KNN	.34	.62	.44	13.69%	9.83%	.36	.67	.46	9.50%	6.83%
BERT (Fine-tuned)	Natural-max-max	.25	.79	.38	16.59%	11.86%	.25	.82	.38	11.45%	8.65%
	Natural-max-mean	.32	.42	.36	14.16%	12.24%	.39	.46	**.42**	10.11%	8.83%
	Natural-mean-max	.33	.02	.04	12.79%	12.78%	.80	.10	.18	**9.32%**	8.68%
	Natural-mean-mean	**1.0**	.02	.04	**12.72%**	12.72%	**1.0**	.01	.02	9.60%	9.51%
	Balanced-max-max	.13	**1.0**	.23	21.64%	14.99%	.10	**1.0**	.18	16.30%	12.41%
	Balanced-max-mean	.13	**1.0**	.24	21.30%	14.64%	.10	**1.0**	.18	16.08%	12.13%
	Balanced-mean-max	.20	.94	.33	16.60%	10.45%	.18	.94	.31	11.64%	8.18%
	Balanced-mean-mean	.27	.90	**.41**	14.55%	**8.87%**	.22	.76	.34	10.35%	**7.24%**

on the training set and tested the obtained models on the chunks of the test set in a cumulative approach for early depression detection. We present the results in Table 1.6.

For 2017 collection, we can see that the model with feature selection (Chi2) does not improve the performance compared to the model with 256 features. However, on 2018 collection the feature selection model improves the F_1 and $ERDE_{50}$ measures slightly.

Table 1.6: Performance of the early depression detection using the well-established machine learning (ML) methods based on the features selected from Table 1.2 using Chi2 technique for e-Risk 2017 and 2018 collections. We considered the threshold on the output probability of the classification algorithm as 0.5 to decide whether the user is depressed or not.

		2017					2018				
	ML model	P	R	F_1	$ERDE_5$	$ERDE_{50}$	P	R	F_1	$ERDE_5$	$ERDE_{50}$
256	RF	**.63**	**.62**	**.62**	**12.71**%	**9.48**%	**.65**	.53	.58	**9.41**%	7.19%
Chi2	RF	.60	**.62**	.61	12.76%	9.54%	.63	.56	**.59**	**9.41%**	**6.65%**

1.6.4 Ablation analysis

In the ablation analysis, we considered Random forest with feature-based models. On the one hand, we consider the reference model, the one that uses all the 256 features, to which we removed some features (exclusion). On the other hand, we considered models where a limited set of features are considered (inclusion). We conducted the analysis at the meta-level as well as at the category-level.

1.6.4.1 Meta level analysis

In the *exclusion analysis*, from the reference model (first row, Table 1.7), we removed one meta category of features, considering successively each of the six meta categories. This approach helps in understanding which are the less and most influencing meta categories. If the results decrease when removing a meta category of features, that means that the corresponding features are important. If the results are stable when removing a category of features, that means either that other features cover well the modeled phenomenon or that the features are not important. To be able to distinguish between these two options, we add the inclusion analysis. In the *inclusion analysis*, we build a model that contains the features of a single meta category each time.

From Table 1.7, we can see that the model that includes all the features is among the best for both collections. Each measure aims at providing different insights on the model capability. Recall and precision are inversely related while F_1 measure combines them. When focusing on F_1 measure, we can observe that excluding META_6 (lexical categories) is not consistently problematic: it is for 2017 dataset, where both recall and precision decrease, but not for the 2018 dataset where precision only decreases but recall increases. Overall, when considering F_1 measure, recall and precision, removing one of the meta categories does not affect the results. This shows that the features from different categories have some kind of redundancy although there are not of the same nature. The same type of conclusion can be drawn when analyzing the early detection measures. META_5 (emoticons) does not help much on

early detection: when removed (Exclusion), results improved for both collections. We can also see that in general not considering the full-text representation (Exclusion META$_1$) hurts the results the most, for almost all the measures. That means that focused features such as lexicons, writing style and emoticons, are not enough, still the entire text is useful.

Now when considering a limited number of features (Inclusion), we can see that some meta categories are more helpful than others and that they can be complementary. For example, META$_3$ (temporality) features help for recall for both collections but are not enough to obtain acceptable precision. On the other hand, META$_1$ features are key for early detection (lowest ERDE for the model that uses these features) while META$_3$ features are not (high ERDE). We can also see that model with only META$_6$ (lexical categories) produces the best overall result for Precision and F_1 measures for the 2017 collection. With the META$_6$ features, the inclusion result is also quite convincing for the 2018 collection.

Table 1.7: Ablation study of the meta group of features both in exclusion and inclusion strategy using the Random forest classifier. We considered the threshold on the output probability of the classification algorithm as 0.5 to decide whether the user is depressed or not.

	Features	Exclusion					Inclusion				
		P	R	F_1	$ERDE_5$	$ERDE_{50}$	P	R	F_1	$ERDE_5$	$ERDE_{50}$
2017	All	.63	**.62**	**.62**	**12.71%**	**9.48%**	.63	**.62**	.62	**12.71%**	9.48%
	- META$_1$	.60	**.62**	.61	12.91%	9.78%	.50	.52	.51	13.45%	10.35%
	- META$_2$	.61	.60	.60	12.73%	9.75%	.39	.50	.44	13.51%	11.30%
	- META$_3$	**.68**	.58	**.62**	12.79%	9.56%	.12	**.62**	.20	18.47%	14.95%
	- META$_4$	.61	.58	.59	13.08%	10.09%	.26	.58	.36	14.80%	11.79%
	- META$_5$	.67	.58	**.62**	12.99%	10.46%	.20	.23	.21	14.52%	13.77%
	- META$_6$	.60	.58	.59	12.81%	10.12%	**.70**	**.62**	**.65**	12.80%	**9.19%**
2018	All	**.65**	.53	.58	9.41%	7.19%	**.65**	.53	**.58**	9.41%	7.19%
	- META$_1$	.65	.49	.56	9.62%	7.08%	.49	.47	.48	**9.38%**	**7.04%**
	- META$_2$	.57	.54	.56	9.40%	**6.60%**	.32	.66	.43	10.64%	8.38%
	- META$_3$	.63	.53	.58	9.44%	6.88%	.10	**.68**	.18	13.32%	11.06%
	- META$_4$	.62	.54	.58	9.43%	6.89%	.21	.53	.30	10.13%	7.62%
	- META$_5$	**.66**	.57	**.61**	9.29%	6.73%	.08	.09	.09	10.55%	10.06%
	- META$_6$	.60	**.63**	**.61**	**9.12%**	6.74%	.59	.49	.54	9.76%	7.15%

1.6.4.2 Category level analysis

When considering the exclusion of a category of features, we can see that Part-Of-Speech features (category II) could be omitted, either results improved like on 2017 or they are stable like on 2018 data. The results are consistently better when omitting the readability features (category X) as well. When considering only category II or only category X, we can also see that the resulting models are poorly performing.

Table 1.8: Ablation study of the category of features defined in Table 1.2 considering both exclusion and inclusion strategy using the Random forest classifier. We considered the threshold on the output probability of the classification algorithm as 0.5 to decide whether the user is depressed or not.

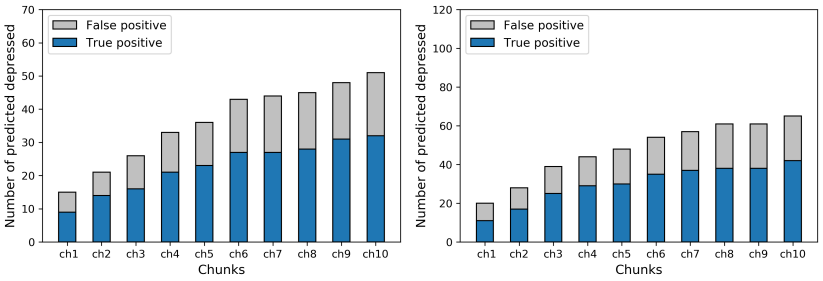
	Features	Exclusion					Inclusion				
		P	R	F_1	$ERDE_5$	$ERDE_{50}$	P	R	F_1	$ERDE_5$	$ERDE_{50}$
2017	All	.63	**.62**	.62	12.71%	9.48%	.63	.62	.62	**12.71**%	9.48%
	- I	.62	.54	.58	12.97%	10.03%	.47	.54	.50	13.17%	9.98%
	- II	**.70**	**.62**	**.65**	12.72%	9.37%	.38	.58	.46	13.86%	10.12%
	- III	.66	.60	.63	12.80%	9.87%	.18	.52	.27	15.85%	13.14%
	- IV	.67	**.62**	.64	12.68%	**9.13**%	.50	.62	.55	13.53%	11.50%
	- V	.62	.56	.59	12.70%	10.13%	.55	.12	.19	13.13%	12.63%
	- VI	.67	**.62**	.64	12.82%	9.19%	.44	.13	.21	12.84%	11.79%
	- VII	.65	**.62**	.63	12.72%	9.41%	.13	.73	.23	18.88%	14.62%
	- VIII	.62	.58	.60	12.69%	9.93%	.25	.65	.36	15.29%	10.99%
	- IX	.64	.58	.61	12.82%	9.66%	.19	.44	.26	15.27%	12.24%
	- X	**.70**	**.62**	**.65**	**12.55**%	9.25%	.25	.56	.34	14.59%	11.94%
	- XI	.66	.56	.60	12.58%	9.41%	.41	.52	.46	14.17%	11.73%
	- XII	.65	**.62**	.63	12.84%	9.22%	.14	**.90**	.24	21.66%	17.54%
	- XIII	.66	.60	.63	12.61%	10.12%	.15	.83	.25	20.73%	17.67%
	- XIV	.63	**.62**	.62	12.94%	9.47%	.27	.50	.35	14.77%	11.27%
	- XV	.67	.58	.62	12.99%	10.46%	.20	.23	.21	14.52%	13.77%
	- XVI	.60	.58	.59	12.81%	10.12%	**.70**	.62	**.65**	12.80%	**9.19**%
2018	All	.65	.53	.58	9.41%	7.19%	**.65**	.53	**.58**	**9.41**%	7.19%
	- I	.59	.49	.54	9.15%	7.02%	.50	.46	.48	9.70%	7.25%
	- II	.65	.52	.58	9.34%	6.84%	.29	.44	.35	9.82%	7.41%
	- III	.60	.56	.58	9.39%	6.82%	.17	.63	.27	11.88%	9.83%
	- IV	.62	.56	.59	9.44%	**6.66**%	.40	.63	.49	10.21%	8.14%
	- V	.60	.51	.55	9.49%	**6.66**%	.47	.28	.35	9.89%	9.32%
	- VI	.58	.51	.54	9.50%	6.80%	.06	.01	.02	9.73%	9.70%
	- VII	**.71**	.49	.58	9.53%	7.02%	.09	.68	.16	14.40%	11.79%
	- VIII	.58	.51	.54	9.40%	7.17%	.23	.57	.33	10.73%	8.71%
	- IX	.65	.51	.57	9.58%	7.33%	.16	.41	.23	10.39%	8.73%
	- X	.66	.53	.59	9.38%	6.97%	.22	.52	.31	10.23%	8.07%
	- XI	.62	.53	.57	9.34%	7.13%	.37	.49	.42	9.81%	7.61%
	- XII	.60	.52	.56	9.56%	6.78%	.10	.76	.18	15.61%	13.29%
	- XIII	.60	.53	.56	9.44%	6.79%	.10	**.80**	.18	15.59%	13.36%
	- XIV	.59	.52	.55	9.58%	6.79%	.22	.39	.28	10.38%	8.60%
	- XV	.66	.57	**.61**	9.29%	6.73%	.08	.09	.09	10.55%	10.06%
	- XVI	.60	**.63**	**.61**	**9.12**%	6.74%	.59	.49	.54	9.76%	**7.15**%

Reversely, we can observe that the lexical features (category XVI) are very useful: without this category, the results decrease both for precision and recall on e-Risk 2017 and for precision on e-Risk 2018. While the results are already quite high when these features are the only considered (inclusion study), the best overall result on 2017 collection and strong performance for 2018 collection. When considering the inclusion analysis, we can also observe that the textual representation (bag-of-words

and POS (category I)) is also quite useful just by itself. With regard to the other categories, things are less definite.

1.7 Visualization of early detection

The number of depressed predictions (true positives and false positives) per chunk is shown in Figure 1.3 for the 256 feature model using Random forest. These are the users for which the system has made a decision on a given chunk. We mention both true positives (detected depressed users) and false positives (users detected as depressed while they are not).



(a) 2017 (b) 2018

Fig. 1.3: The number of correctly predicted depressed users increases - 256 features model with Random forest on e-Risk data: (a) 2017 and (b) 2018.

1.8 Conclusion

In this chapter, we compared different models for learning from annotated examples to detect possible depression of users. Both the examples to train the models and the examples to test the prediction are composed of users' textual posts.

We first compared different ways of representing users: either considering some intuitive and domain-related features or considering straightforward word or document embeddings which do not need any domain expertise. We found that the results are better using domain-specific features although document embeddings are close. Document embeddings are most of the cases better than word embedding.

We also compared six well-established machine learning methods. We found that Random forest is the best. This is consistent with the results we obtained in different contexts such as for predicting the diffusion of information on social media using

domain-specific features [15] or predicting the best search engine to use for a given query [9].

We compared depression detection and early depression detection. While our hypothesis was that BERT would outperform the well-established machine learning methods for depression detection, this is not the case. The data available is not enough for proper fine-tuning, even when not considering early detection to fine-tune properly the models.

The feature ablation analysis we conducted informed us on the counter-intuitive result that emotion features are not efficient and can be omitted, this is probably due to the fact they are too sparse and not much used by the users -at least in the dataset we used for evaluation. Choudhury *et al.* [4] also noticed that emotion is not significant for depression detection on Facebook data but their other study [5] where they use Twitter data, emotion feature is significant for depression detection, especially negative emotion.

Acknowlegments

This work is partially supported by the PREVISION project, which has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 833115 (`https://cordis.europa.eu/project/id/833115`). The paper reflects the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

References

1. Sergio G. Burdisso, Marcelo Errecalde, and Manuel Montes-y-Gómez. t-ss3: a text classifier with dynamic n-grams for early risk detection over text streams. *CoRR*, abs/1911.06147, 2019.
2. Sergio G. Burdisso, Marcelo Errecalde, and Manuel Montes-y-Gómez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Syst. Appl.*, 133:182–197, 2019.
3. Fidel Cacheda, Diego Fernández Iglesias, Francisco Javier Nóvoa, and Victor Carneiro. Analysis and experiments on early detection of depression. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.
4. Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Computer Supported Cooperative Work, CSCW '14, Baltimore, MD, USA, February 15-19, 2014*, pages 626–638, 2014.
5. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media*, 2013.
6. Gualtiero B. Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, 73:291–300, 2016.
7. Andrew M Dai, Christopher Olah, and Quoc V Le. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015.
8. Clément Dalloux, Vincent Claveau, Marc Cuggia, Guillaume Bouzillé, and Natalia Grabar. Supervised learning for the ICD-10 coding of french clinical narratives. In *Digital Personalized Health and Medicine - Proceedings of MIE 2020, Medical Informatics Europe, Geneva, Switzerland, April 28 - May 1, 2020*, pages 427–431, 2020.
9. Romain Deveaud, Josiane Mothe, Md Zia Ullah, and Jian-Yun Nie. Learning to adaptively rank document retrieval system configurations. *ACM Transactions on Information Systems (TOIS)*, 37(1):1–41, 2018.
10. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
11. Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 4647–4657, 2016.
12. Daniel J. France, Richard G. Shiavi, Stephen E. Silverman, Marilyn K. Silverman, and D. Mitchell Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomed. Engineering*, 47(7):829–837, 2000.
13. Darío Gustavo Funez, Marcelo Luis Errecalde, Maria Paula Villegas, Maria José Garciarena Ucelay, and Leticia Cecilia Cagnina. Temporal variation of terms as concept space for early risk prediction. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
14. Darío Gustavo Funez, Maria José Garciarena Ucelay, Maria Paula Villegas, Sergio Burdisso, Leticia C. Cagnina, Manuel Montes-y-Gómez, and Marcelo Errecalde. Unsl's participation at erisk 2018 lab. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.
15. Thi Bich Ngoc Hoang and Josiane Mothe. Predicting information diffusion on twitter–analysis of predictive features. *Journal of computational science*, 28:257–264, 2018.
16. RAMIANDRISOA Iarivony Faneva. *Extraction et fouille de données textuelles: application à la détection de la dépression, de l'anorexie et de l'agressivité dans les réseaux sociaux*. PhD thesis, Université de Toulouse, 2020.
17. Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431, 2017.

18. Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.
19. Akshay Bhavani Kumar Kulkarni. Early detection of depression. Master's thesis, University of Houston, 2018.
20. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.
21. Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6), December 2017.
22. Zhixing Li, Zhongyang Xiong, Yufang Zhang, Chunyong Liu, and Kuan Li. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, 32(3):441–448, 2011.
23. David E Losada, Fabio Crestani, and Javier Parapar. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 346–360. Springer, 2017.
24. Lu-Shih Alex Low, Namunu C. Maddage, Margaret Lech, Lisa Sheeber, and Nicholas B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Trans. Biomed. Engineering*, 58(3):574–586, 2011.
25. Idriss Abdou Malam, Mohamed Arziki, Mohammed Nezar Bellazrak, Farah Benamara, Assafa El Kaidi, Bouchra Es-Saghir, Zhaolong He, Mouad Housni, Véronique Moriceau, Josiane Mothe, and Faneva Ramiandrisoa. IRIT at e-risk. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
26. Tamsin C. Marriott and Tom Buchanan. The true self online: Personality correlates of preference for self-expression online, and observer ratings of personality online and offline. *Computers in Human Behavior*, 32:171–177, 2014.
27. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
28. Saif Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 2013.
29. Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. Towards automatically classifying depressive symptoms from twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 182–191, 2016.
30. World Health Organization et al. Depression and other common mental disorders: global health estimates. 2017. *Geneva: WHO*, 2017.
31. Simon Øverland, Wojtek Woicik, Lindsey Sikora, Kristoffer Whittaker, Hans Heli, Fritjof Stein Skjelkvåle, Børge Sivertsen, and Ian Colman. Seasonality and symptoms of depression: A systematic review of the literature. *Epidemiology and psychiatric sciences*, 29, 2020.
32. Asli Ozdas, Richard G. Shiavi, Stephen E. Silverman, Marilyn K. Silverman, and D. Mitchell Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Trans. Biomed. Engineering*, 51(9):1530–1540, 2004.
33. Sayanta Paul, Sree Kalyani Jandhyala, and Tanmay Basu. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.
34. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
35. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

36. LS Radloff. A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1:413–449, 2015.
37. Faneva Ramiandrisoa and Josiane Mothe. Early detection of depression and anorexia from social media: A machine learning approach. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020*, 2020.
38. Faneva Ramiandrisoa, Josiane Mothe, Farah Benamara, and Véronique Moriceau. IRIT at e-risk 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.
39. Philip Resnik, William Armstrong, Leonardo Max Batista Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan L. Boyd-Graber. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of CLPsych@NAACL-HLT*, 2015.
40. Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.
41. Farig Sadeque, Dongfang Xu, and Steven Bethard. Uarizona at the CLEF erisk 2017 pilot task: Linear and recurrent models for early depression detection. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
42. H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Gregory J. Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle H. 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.
43. Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
44. Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601, 2018.
45. Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.
46. Maria Paula Villegas, Darío Gustavo Funez, Maria José Garciarena Ucelay, Leticia Cecilia Cagnina, and Marcelo Luis Errecalde. LIDIC - unsl's participation at erisk 2017: Pilot task on early detection of depression. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
47. Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. A depression detection model based on sentiment analysis in micro-blog social network. In *Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers*, pages 201–213, 2013.
48. Yuanyuan Xue, Qi Li, Li Jin, Ling Feng, David A. Clifton, and Gari D. Clifford. Detecting adolescent psychological pressures from micro-blog. In *Health Information Science - Third International Conference, HIS 2014, Shenzhen, China, April 22-23, 2014. Proceedings*, pages 83–94, 2014.