Using contextual information for automatic triage of posts in a peer-support forum

Edgar Altszyler
Universidad de Buenos Aires, FCEyN, DC;
CONICET-ICC; Fundación Sadosky
ealtszyler@dc.uba.ar

Ariel J. Berenstein
Hospital de Niños Ricardo Gutierrez
Traslational Bioinformatics Unit
arieljberenstein@gmail.com

David Milne
School of Electrical and Information
Engineering, The University of Sydney
david.milne@sydney.edu.au

Rafael A. Calvo
School of Electrical and Information
Engineering, The University of Sydney
Rafael.Calvo@sydney.edu.au

Diego Fernandez Slezak
Universidad de Buenos Aires, FCEyN, DC;
CONICET-ICC
dfslezak@dc.uba.ar

Abstract

Mental health forums are online spaces where people can share their experiences anonymously and get peer support. These forums, require the supervision of moderators to provide support in delicate cases, such as posts expressing suicide ideation. The large increase in the number of forum users makes the task of the moderators unmanageable without the help of automatic triage systems. In the present paper, we present a Machine Learning approach for the triage of posts. Most approaches in the literature focus on the content of the posts, but only a few authors take advantage of features extracted from the context in which they appear. Our approach consists of the development and implementation of a large variety of new features from both, the content and the context of posts, such as previous messages, interaction with other users and author's history. Our method has competed in the CLPsych 2017 Shared Task, obtaining the first place for several of the subtasks. Moreover, we also found that models that take advantage of post context improve significantly its performance in the detection of flagged posts (posts that require moderators attention), as well as those that focus on post content outperforms in the detection of most urgent events.

1 Introduction

According to the World Health Organization (WHO), 20% of children and adolescents in the world have mental disorders or problems (WHO, 2014). Suicide ranks as the second leading cause of death in the 15-29 years old group and every 40 seconds a person dies by suicide in the world. The WHO pointed early identification and intervention as a key factor in ensuring that people receive the care they need (WHO, 2014). Mental health problems have a strong impact on our society and require the use of new techniques for their study, prevention, and intervention.

In this context, text mining tools are emerging as a powerful channel to study and detect the mental state of the writers (Calvo and Mac Kim, 2013; Bedi et al., 2015, 2014; De Choudhury et al., 2013a,b; Coppersmith et al., 2015). In particular, there is a greater interest in the study and detection of suicidal ideation in texts coming from social networks. In this line, Tong et al. (2014) and O'Dea et al. (2015) developed automatic detection systems to identify suicidal thoughts in tweets, and Homan et al. (2014) studied the network structure of users with suicidal ideation in a forum. Furthermore, the CLPsych 2016 shared task proposed the triage of posts, based on urgency, from a peer-support mental health forum (for a more ex-

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 57–68

haustive review see (Calvo et al., 2017)). In the present article, we build an automatic post triage system and compete in the CLPsych 2017 shared task (Milne et al., 2016). The automatic detection of suicidal ideation in social networks and forums provide a powerful tool to address early interventions in serious situations. Additionally, these techniques allow tracking the prevalence of different suicide risk factors among the population (Jashinsky et al., 2014; Fodeh et al., 2017), which provides valuable information that can be capitalized for the design of prevention plans.

1.1 CLPsych 2017 Shared Task

The CLPsych 2017 shared task involves the triage of posts from an Australian mental health forum, *Reachout.com*, which provides a peer-support online space for adolescents and young adults. *Reachout.com* offers a space to read about other peoples experiences and talk anonymously. Additionally, the forum has trained moderators who intervene in delicate situations, such as when a user is expressing suicidal ideation. There is an escalation process to follow when forum members might be at risk of harm. As the number of forum members increases the reading of all post become impossible, thus an automatic triage that efficiently guides moderator's attention to the most urgent posts result essential (Calvo et al., 2016). The CLPsych 2017 Shared Task consists of identifying each forum post with one of four triage levels: *crisis*, *red*, *amber* and *green* (in decreasing priority). A *crisis* label indicates that the author is in risk so moderators should prioritize this post above all others, while a *green* label indicates that post does not require the attention of any moderator. See Milne et al. (2016) for a detailed description of the annotation process and the ethical considerations.

CLPsych 2017 Shared Task dataset consists of 157963 posts written between July 2012 and March 2017 (see Table 1). Among these posts, 1188 were labeled by 3 annotators in order to train the model (training set), and 400 were selected to form the testing set. Posts in the training set were written between April 2015 and June 2016 while posts in the test set were written between August 2016 and March 2017.

Fifteen teams took part in CLPsych 2017 shared task, with unlimited submissions per group. Each post of the dataset contains the text of the subject and the body, structured in XML format. Additional metadata is also provided, such as boards, thread, post time, or if the post was written by a moderator or not. The official metrics of the task are:

	crisis	red	amber	green	total
train	40	137	296	715	1188
test	42	48	94	216	400
extra	-	-	-	-	156375

Table 1: Training dataset and extra unlabeled dataset statistics. Crisis, red, amber and green, are the four triage levels and reflects a decreasing priority of required moderator intervention/response. We had access to the test dataset only after the competition have finished

- *Macro-averaged f-score*: the average f1-score among *crisis*, *red* and *amber* labels.
- *F-score for flagged vs. non-flagged*: the average f1-score among flagged (*crisis* + *red* + *amber*) and non-flagged (*green*) labels. This is considered considered by the task organizers as the most important metric, given that it measures the system's capability to identify post that need moderators attention.
- *F-score for urgent vs. non-urgent*: the average f1-score among urgent (*crisis* + *red*) and non-urgent (*amber* + *green*) labels.

The official measures are the f-scores, as accuracy is known to be less sensitive to misclassification of elements in the minority class in highly unbalanced datasets. In this paper, we also analyze the f-score for *crisis* vs. *non-crisis*, which measures the system's capability to identify the most serious cases. This competition is a new version of the CLPsych 2016 Shared Task (Milne et al., 2016), which has the same goal but counts with a smaller dataset. The different approaches used in 2016 competition involved a huge variety of features, such as N-grams, lexicon-based features, word embeddings, and metadata. Most of the models extracted features from the content of posts, but only a few authors took advantage of features extracted from the context of the posts, such as n-grams of previous posts of the thread, or previous author's posts (Malmasi et al., 2016; Cohan et al., 2016; Pink et al., 2016).

In the present work, we extract and test a large variety of new features from both the body of the

posts and the context in which the posts occur, such as: (1) authors' history, (2) adjacent posts, and (3) the authors' interaction network. We hypothesize that the contextual features will be useful to capture new elements that allow building a better profile of the author of the posts. This idea is grounded in Van Orden et al. (2010) observation that suicidal behavior tends to persist over the lifetime, and also De Choudhury et al. (2013b), Homan et al. (2014) studies in which they show that interaction patterns have valuable information about the underlying mental state of the users.

2 Method

To triage posts we apply a supervised classification-based approach. In the present section, we describe the texts preprocessing step, the features that were used, the feature transformation process and the classification method.

First, we preprocessed the body of the post: we removed HTML format and eliminated quotes (HTML quotes tags), we converted ReachOut links, other webpage links, author mentions, and forum's emoticons to tokens such as #reachout_link, #ref_link, #reference, #SmileyHappy respectively. Then we transformed the text to lowercase and word-tokenized it with the happierfuntokenizing.py (World Well-Being Project, 2017), which can handle most common emoticons.

We extracted a total of 2799 features from each post. We organized features in seven main categories, four of them are content based features: (Word2vec - N-grams - Metadata - Body), and the remaining are context-based ones (Interaction features - Adjacent features - Author features).

After the feature extraction process, a Z-score transformation was applied to all features, with the exception of n-grams features in which we performed a TF-IDF weighting. Then, missing values were filled with the mean value of those features in the unlabeled dataset.

Following we present a brief description of each category (see Table 2 for features statistics). In this section, we will use parenthesis to show the number of features.

Given the large number of features, in some categories we built subsets of features, in which we selected the features that we considered the most relevant in each case (see Supplemental Material A.1 for a detailed description of each subset).

Feature	Type	Complete	Subset
Word2vec	content	50	-
N-grams	content	2274	50
Metadata	content	23	7
Body	content	68	23
Interaction	context	155	57
Adjacent	context	152	100
Author's	context	77	50
Total	both	2799	-

Table 2: Features statistics. For each feature category, we show the type, the number of extracted features (noted as complete), and the number of selected features in its subset (Noted as Subset)

2.1 Word2vec representation (50 features)

We used all post bodies in the unlabeled dataset to train a Skip-gram model (Mikolov et al., 2013a) of 50 dimensions. We discarded infrequent tokens, with less than 5 repetitions and very frequent tokens, with a frequency higher than 10^{-3}. We set the window size and negative sampling to 15 (which were found to be maximal in two semantic tasks over TASA corpus (Altszyler et al., 2017)). Word2vec semantic representations were generated with the Gensim Python library (Rehuek and Sojka, 2010). After the training, the resulting Word2vec post's features were computed as the average of all word-embeddings in the post.

2.2 N-grams (2274 features)

We extracted unigrams and bigrams from all body posts, and kept the 3000 most frequent N-grams in the training corpus (following (Brew, 2016)) and applied a TF-IDF transformation. As the training and test sets contain posts from different time periods, the language patterns may have changed during this time lapse. In order to eliminate most different N-grams, we have excluded all N-grams with a frequency lower than 5.10^{-5} in the posts form unlabeled dataset in the period Aug 2016 - Mar 2017 (726 N-grams where eliminated in this way).

2.3 Metadata features (23 features)

We included several non-linguistic features derived from post's metadata and removed all features showing lack of variability in our training set (std = 0). The selected features are: week day (7),

board (5), whether the author is a moderator or not (1), whether the author created the thread (1) and time since the last edition (1). Additionally, We subdivided the day in 8 timeslots of 3 hours, and create *post time* features, consisting of 8 dummy variables to identify the timeslot of the post (8).

2.4 Body content features (68 features)

These features aim to characterize the emotional and psychological state of the author of the post. We employed several well-established lexicons, such as Emolex (Mohammad and Turney, 2010) (10), Hedonometer (Dodds et al., 2011)(1), DAL (Whissell, 1989) (3), Warriner's Norms (Warriner et al., 2013) (3), Age of Adquisition (Kuperman et al., 2012)(1), Bristol familiarity and imaginary norms (Stadthagen-Gonzalez and Davis, 2006) (2), and WWBP lexicons (Schwartz et al., 2016, 2013; World Well-Being Project, 2017) which includes: PERMA (10), OCEAN (5), time-oriented (3) and affect-intensity lexicons (2). We also used MentalDisLex (Zirikly et al., 2016) (1), profanity word-list (Smedt and Daelemans, 2012) (1), Von Ahn offensive lexicon (Von Ahn, 2016) (1), subjectivity and sentiment analysis (Smedt and Daelemans, 2012) (2), fraction of first person singular and second person pronouns (2), determiners (1), word counts (1), mean word length (1), number of webpage links (1), lexical diversity (1)(mean fraction of different words among 100 random subsamples of 10 words) and the fraction of words semantically similar to several keywords[1] (8). The semantic similarity was measure with word2vec pre-trained vectors (Mikolov et al., 2013b) and the threshold to identify a word as similar was set to 0.3.

We also included categorical features, such as predefined forum emoticons (4), references to helplines[2] (1), references to advisors[3] (1) and self-harm expressions were present or not [4] (2). We only take into account self-harm expressions in which only appears first-person pronouns and did not appears negations in a window of 15 or 50 words around the regexp.

Missing values in lexicon-based features which have a neutral value were filled by the neutral value (for example in DAL, pleasantness range from 1 (unpleasant) to 3 (pleasant), thus we replaced missing values with 2). All features showing lack of variability (std = 0) in our training set were removed.

2.5 Interaction features (155 features)

We believe that the interaction patterns between users hold valuable information about the underlying intention and emotions of the posts. To this end, we built a *directed mention graph* where a node (post), n_i has an incoming edge from n_j if n_j has mentioned the n_i post author within a 10 post temporal windows, or an outgoing edge if n_i has mentioned n_j author in the same period. First, we take advantage of this network to extract seven basic network structural features such as: in/out degree, number of in/out edge from different authors, number of loops, number of post from the author in the window, out degree of the author mentioned in the post.

Then, we define on this graph node attributes based on some set of *Body* and *Word2Vec* features, namely f_a. After that, for the k-th node (post) in our network, we define a new set of interaction-based features $Fint_a$, by averaging the feature f_a across the neighborhood of the post (Nei). It is:

$$Fint_a := \frac{1}{|Nei|} \sum_{k \in Nei} f_a \qquad (1)$$

74 features were extracted from incoming edges and 74 from outgoing edges. The extracted features consist on, Word2vec (50), WWBP lexicons (20), Hedonometer (1), pronouns (2) and semantic coherence (1), which is measured as the cosine similarity between the word2vec embedding of the node and the central post.

Missing values that use Word2vec similarity were filled by the mean similarity between successive posts in the unlabeled dataset, missing values

[1] Word2vec keywords: depression, suicide, fear, mental_health, suicidal_ideation, antidepressant, hopelessness and anxiety

[2] helplines keywords: kidshelpline, eheadspace, helpline, kidshelp, khl, counselling, headspace, helplines, mensline, www.eheadspace.org, 1800respect, beyondblue, lifeline, callback, lifeline's, scbs, catt, triage, suicideline

[3] advisors keywords: supervisor, supervisors, mentor, manager, tutor, manager, casemanager, managers, manager's, psych, pysch, psychiatrist, gp, gp's, counsellor, counsellor, counselor

[4] Self-harm regular expressions: "suicid\w*","kill\w* myself", "kill\w* my self", "cut\w* myself", "cut\w* my self", "hurt\w* myself", "hurt\w* my self", "harm\w* myself", "harm\w* my self", "I want\w* to die" , "I don't want to live", "end my life" (\w* refers to 0 or more alphabetic letters. The selected self-harm expressions where inspired in posts from the subreddit *suicidewatch*. In keyword spotting, it is important not to be influenced by the train data in order to avoid overfitting.

which count outgoing edges were filled by -1 and missing values in $Fint_a$ features were filled with the mean value of the feature a in the unlabeled dataset.

2.6 Adjacent features (152)

For each post, we extract 76 features from the previous post in the same thread and 76 features from the previous post produced by the same author in the thread. The extracted features consist on: Word2vec (50), WWBP lexicons (20), Hedonometer (1), pronouns (2) semantic coherence between the post and the previous post (1), post day of the previous post (1), time between posts (1).

2.7 Authors' features (77 features)

We replicated and extended Shickel *et* al's (Shickel and Rashidi, 2016) idea of deriving attributes from the history of the authors. For each post, we computed the mean value of several features for all the previous posts written by the same author. These features provide a baseline for the authors, which may allow the machine learning algorithm to identify when a post differs from the typical behavior of its author. The extracted features consist on, Word2vec (50), WWBP lexicons (20), Hedonometer (1), pronouns (2), post day (1)

Additionally, we added other features to identify more general authors behavior, such as, entropy in thread and board participation (2) and median time between posts measured in log-scale, $log(\#minutes + 1)$ (1).

2.8 Models

We used Support Vector Machine classifiers (SVM) with linear kernels and Radial Basis Function (RBF) kernels. Each model was trained on different combinations of features, and the hyperparameter C was selected with a grid search scheme for each model. In the grid search, the performance metric was the macro f-score with a 10-fold Cross-Validation (CV). The C hyperparameters were varied among {0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100} for the SVM-RBF models and among {0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1} for the SVM-linear models. As the training dataset is highly imbalanced, both SVM models were trained with class weights inversely proportional to class frequencies in the training dataset. We also tested XGBoost and Random Forest models which underperformed the SVM models, and a feature selection process which did not produce significant performance improvements in the SVM RBF and SVM linear models (see Supplemental Material A.3). All the models were implemented in python with Sklearn or XGBoost packages, and all other parameters, not included in the grid search, were set to their default values.

We have built nine collections of features composed of different categories and features subsets (for a full description of the collections see Supplementary Materials A.2). With this features collections, we trained 18 SVM models, half with an RBF kernel and half with a linear kernel. Additionally, we implemented four ensemble models composed by SVM's combined with a majority voting method. We used ensembles with four and seven SVMs with RBF and linear kernels and the differences within the voting SVM's are their training features (see Supplementary Materials A.4 for a full description of the voting SVM's features). In case of a tie between classes, the post is classified as the most urgent class.

3 Results

Table 3 shows the top performing models of the CLPsych 2017 challenge divided by metric, in which only the best model of each team is showed. We have obtained the **2nd position** in the *Macro-averaged f-score* with an ensemble of 4 SVM-linear models, the **1st position** in the *flagged vs non-flagged* f-score with an ensemble of 7 SVM-RBF models, the **1st position** in the *urgent vs non-urgent* f-score with a SVM-RBF trained with Word2vec + N-grams + subset of body features, and the **1st position** in *crisis vs non-crisis* f-score with a SVM-RBF trained with Word2vec + N-grams + subset of metadata features.

In Table 4 we show our model's results ordered by the performance in the flagged vs non-flagged metric, which is considered by the organizers as the most relevant metric, as it measures the system's capability to identify posts that need moderators attention.

It is worth noting that there is not a universal best model, however, our approach obtains very good results in all performance measures. In particular, our models tend to outperforms other team's models in the flagged vs non-flagged f-scores, where nine of the top ten models are from our team (see bold scores in Table 4).

Among our models, those that take advan-

metric	pos	team	model	f-score
macro averaged f-score	1st	Xia and Liu	voteing_submission	0.467
	2nd	*Our team*	*ensemble_4models_linear*	*0.462*
	3rd	Nair et al.	Run23	0.461
flagged vs non-flagged	*1st*	*Our team*	*ensemble_all_rbf*	*0.905*
	2nd	Yates et al.	mpid5_cl17out_20_v3	0.883
	3rd	French et al.	15	0.877
urgent vs non-urgent	*1st*	*Our team*	*body_rbf*	*0.686*
	2nd	Yates et al.	53	0.673
	3rd	French et al.	C13/SH2	0.624
crisis vs non-crisis	*1st*	*Our team*	*metadata_rbf*	*0.484*
	2nd	Xia and Liu	jxufe-lda-svm	0.480
	3rd	Nair et al.	Run23	0.468

Table 3: Official results for the CLPsych 2017 shared task. We also show the f-scores for *crisis* vs. *non-crisis*, which measures the system's capability to identify the most serious cases. Flagged refers to *crisis* + *red* + *amber*, while urgent to *crisis* + *red*. For each metric, only the best model of each team is showed

model	category	N	CV macro	macro	flagged	urgent	crisis
ensemble_all_rbf	content+context	575	0.537	0.442	**0.905**	0.586	0.328
ensemble_4models_rbf	content+context	514	0.580	0.392	**0.905**	0.472	0.241
selection_linear	content+context	337	0.490	0.400	**0.887**	0.508	0.416
metadata_rbf	only content	107	0.479	0.445	**0.887**	**0.655**	**0.484**
selection_rbf	content+context	337	0.549	0.436	**0.881**	0.618	0.400
content_rbf	only content	130	0.489	0.436	**0.881**	0.618	0.400
metadata_linear	only content	107	0.452	0.442	**0.881**	**0.677**	**0.476**
ensemble_4models_linear	content+context	514	0.540	**0.462**	**0.881**	0.598	**0.452**
ensemble_all_linear	content+context	575	0.506	**0.453**	**0.880**	0.637	0.410
all_features_linear	content+context	2799	0.512	0.394	0.879	0.497	0.299
ngrams_filtered_linear	content+context	575	0.512	0.393	0.879	0.494	0.294
content_linear	only content	130	0.448	0.423	0.878	0.645	0.381
all_features_rbf	content+context	2799	0.535	0.384	0.876	0.453	0.246
ngrams_filtered_rbf	content+context	575	0.556	0.291	0.866	0.326	0.081
base_linear	only content	100	0.425	0.429	0.859	**0.680**	0.444
body_rbf	only content	123	0.464	**0.456**	0.859	**0.686**	**0.476**
base_rbf	only content	100	0.453	**0.456**	0.859	**0.686**	**0.476**
word2vec_rbf	only content	50	0.446	0.422	0.857	0.647	0.450
word2vec_linear	only content	50	0.423	0.435	0.852	**0.677**	**0.460**
body_linear	only content	123	0.419	0.437	0.852	**0.667**	0.395
ngrams_only_rbf	only content	2274	0.440	0.446	0.825	0.542	0.305
ngrams_only_linear	only content	2274	0.436	0.445	0.803	0.540	0.281

Table 4: Our models' scores, ordered by the performance in the flagged vs non-flagged metric. We show in bold the scores of the models that are within the top ten among the 251 models that have participated in the shared task

tage of contextual features tend to obtain better *flagged vs non-flagged* f-scores (p-value= 4.09E-09, Wilcoxon rank sum test). *Amber* class includes posts where the author is following up on their own previous *red* or *crisis* post (Milne et al., 2016), thus, the inclusion of contextual features is essential to capture these situations. On the other hand, complex models with many features may learn the particularities and details of the authors present in the training set, thus decreasing the predictive capability in posts from authors never seen before (89% of the authors in training set are not in the test set). This overfitting effect in complex models can be observed in the corre-

lation between the number of features (column *N* in Table 4) and the differences in f-scores between the Cross-Validation and the test set (column *CV macro* - column *macro* in Table 4), Spearman correlation of 0.523 with a p-value=0.012. Also, this effect may explain the good performance obtained by less complex models, such as the SVM-linear trained with only 50 word2vec features. Furthermore, it can be seen that models that use only content features tend to obtain better results in *urgent vs non-urgent* and *crisis vs non-crisis* metrics (p-value= 4.17E-09 and p-value= 4.15E-09 respectively, Wilcoxon rank sum test).

We propose that training with a greater amount of data with more users diversity will avoid this overfitting, thus boosting the performance of the models that use more number of features.

Finally, we extract the 25 most relevant features given by the random forest importance measure when it is trained with the training dataset and all the 2799 features (see Table 5). Within the most important features, 10 came from the *Interaction* category, 8 from the *Body*, 4 from *Word2vec*, 2 from *author's* and 1 from *N-grams*.

Furthermore, Table 5 shows that *crisis* posts tend to exhibit more negative PERMA elements, negative sentiment, first person reference and less happiness than *non-crisis* posts (p-value¡0.5e-6 in each comparison with a Wilcoxon rank sum test). Depict Word2vec dimensions have not a straightforward interpretation, it can be seen that there is no shared Word2vec components within the relevant interaction features and the selected Word2vec features extracted from posts text. These results show that content of severe posts and their interacting posts provide different features which result useful in the post triaging task.

4 Conclusion

Mental health forums, such as ReachOut.com, are online spaces where users can share their experiences and get peer support. The large increase in the number of users makes the task of the moderators considerably difficult. This ends in the loss of critical messages that would require immediate attention. In this context, an automatic triaging system is a valuable tool to guide moderators effort.

In the present paper, we present a machine learning approach for the automatic triage of posts from ReachOut.com forum. Our models participated in the CLPsych 2017 Shared Task competition, obtaining very good results along with all official metrics.

The CLPsych 2017 Shared Task is the second part of the 2016 edition, but with more training data and a more balanced test set. Most of approaches used in CLPsych 2016 Shared Task extract features from the content of the posts, but only a few took advantage of features extracted from the posts context. In the present paper we focused on the development and implementation of a large variety of new features from both, the content and the context of posts. The content-based features consist on N-grams, Word2vec, metadata and other features from the body of the posts, while the context-based features extract attributes from the content and structure of the user history, other post in the conversation and the interaction network.

Our implementation obtained the first position on several official metrics. In particular, we obtained the best performance in the flagged vs non-flagged measure, which tests the system's capability to identify posts that require attention from moderators.

We found that exploitation of contextual features tend to improve the detection of posts that require attention from moderator. On the other hand, complex models with many features may learn the particularities and details of the authors present in the training set, thus decreasing the predictive capability in posts from authors never seen before. To avoid this overfitting effect we propose to feed the models with a greater amount of training data with more diversity of users. This can be easily solved with the use of online classifiers (Bordes et al., 2005; Calvo et al., 2016), in which the model can continuously learn from the manual classifications made by the moderators, ensuring that the system is kept up-to-date.

A feature importance analysis emphasize the importance of the interactions among users and the content of the interacting post. In this respect we showed that the content of *crisis* posts and theirs interacting posts provide different elements which result useful in the post triaging task. These analysis also highlighted the predictive capabilities of new open-source psycholinguistic measures designed by the world Well-Being Project group (WWBP), specially the ones related to well-being elements (PERMA).

feature	category	crisis	red	amber	green
Word2vec_2	Word2vec	**0.96 +/- 0.04**	0.90 +/- 0.03	0.78 +/- 0.02	0.14 +/- 0.03
neg_E (PERMA)	Body	**0.97 +/- 0.09**	0.80 +/- 0.05	0.60 +/- 0.03	-0.11 +/- 0.02
neg_P (PERMA)	Body	**1.05 +/- 0.09**	0.93 +/- 0.06	0.56 +/- 0.03	-0.07 +/- 0.03
neg_M (PERMA)	Body	**0.96 +/- 0.09**	0.87 +/- 0.05	0.62 +/- 0.03	-0.10 +/- 0.03
neg_A (PERMA)	Body	**1.17 +/- 0.10**	1.01 +/- 0.05	0.71 +/- 0.04	-0.05 +/- 0.03
incoming_edge_second_pron	Interaction	0.67 +/- 0.22	0.78 +/- 0.12	**1.42 +/- 0.08**	-0.04 +/- 0.04
author_sing_first_pron	Author's	0.96 +/- 0.21	1.15 +/- 0.12	**1.28 +/- 0.06**	0.13 +/- 0.04
incoming_edge_w2v_20	Interaction	-0.44 +/- 0.20	-0.45 +/- 0.08	-0.87 +/- 0.07	**0.17 +/- 0.04**
incoming_edge_w2v_41	Interaction	0.68 +/- 0.14	0.60 +/- 0.09	**1.12 +/- 0.06**	0.01 +/- 0.03
Word2vec_36	Word2vec	0.55 +/- 0.06	**0.58 +/- 0.03**	0.44 +/- 0.02	-0.03 +/- 0.03
happiness (Hedonometer)	Body	-0.56 +/- 0.06	-0.55 +/- 0.04	-0.40 +/- 0.03	**0.16 +/- 0.03**
incoming_edge_w2v_16	Interaction	-0.9 +/- 0.18	-0.37 +/- 0.09	-1.05 +/- 0.07	**0.27 +/- 0.04**
neuroticism (OCEAN)	Body	-0.48 +/- 0.06	-0.45 +/- 0.05	-0.24 +/- 0.02	**0.08 +/- 0.02**
Word2vec_37	Word2vec	**0.81 +/- 0.06**	0.76 +/- 0.03	0.49 +/- 0.03	0.04 +/- 0.03
incoming_edge_w2v_45	Interaction	1.08 +/- 0.21	0.56 +/- 0.10	**1.23 +/- 0.07**	-0.08 +/- 0.04
author_w2v_2	Author's	0.70 +/- 0.23	0.76 +/- 0.11	**0.96 +/- 0.06**	0.01 +/- 0.04
incoming_edge_w2v_47	Interaction	-0.80 +/- 0.26	-0.03 +/- 0.07	-0.57 +/- 0.06	**0.31 +/- 0.04**
Word2vec_8	Word2vec	-0.61 +/- 0.03	-0.55 +/- 0.02	-0.52 +/- 0.02	**-0.13 +/- 0.02**
incoming_edge_w2v_11	Interaction	-0.72 +/- 0.18	-0.45 +/- 0.10	-0.79 +/- 0.06	**0.20 +/- 0.04**
incoming_edge_w2v_25	Interaction	-0.44 +/- 0.16	-0.13 +/- 0.07	-0.72 +/- 0.06	**0.16 +/- 0.04**
sing_first_pron	Body	**1.18 +/- 0.06**	1.10 +/- 0.05	0.88 +/- 0.04	0.12 +/- 0.04
i	N-gram	**0.15 +/- 0.01**	**0.15 +/- 0.01**	0.13 +/- 0.01	0.06 +/- 0.00
incoming_edge_w2v_9	Interaction	0.70 +/- 0.13	0.40 +/- 0.08	**0.84 +/- 0.05**	-0.19 +/- 0.04
negative (EmoLex)	Body	**1.22 +/- 0.19**	0.89 +/- 0.08	0.41 +/- 0.06	-0.05 +/- 0.03
incoming_edge_w2v_32	Interaction	0.65 +/- 0.14	0.51 +/- 0.08	**0.99 +/- 0.06**	0.00 +/- 0.04

Table 5: Statistics of the 25 most relevant features for the triage task, ordered by the random forest importance measure when it is trained with all features. The numbers are showing the mean value and the standard deviation of the mean for each feature in each triage level. For each feature, we have highlighted in bold the highest mean value among the different groups. Sing_first_pron refers to the fraction of words that are first-person pronouns, such as I, me, myself,etc.

References

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition* .

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia* 1:15030.

Gillinder Bedi, Guillermo A Cecchi, Diego F Slezak, Facundo Carrillo, Mariano Sigman, and Harriet De Wit. 2014. A window into the intoxicated mind? speech as an index of psychoactive drug effects. *Neuropsychopharmacology* 39(10):2340–2348.

Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. 2005. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research* 6(Sep):1579–1619.

Chris Brew. 2016. Classifying reachout posts with a radial basis function svm. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* .

Rafael A Calvo, M Sazzad Hussain, David Milne, Kjartan Nordbo, Ian Hickie, and P Danckwerts. 2016. Augmenting online mental health support services. *Integrating Technology in Positive Psychology Practice* page 82.

Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence* 29(3):527–543.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering* pages 1–37.

Arman Cohan, Sydney Young, and Nazli Goharian. 2016. Triaging mental health forum posts. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic*

Signal to Clinical Reality, San Diego, California, USA, June. volume 16.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 31–39.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, pages 47–56.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In *ICWSM*. page 2.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* 6(12):e26752.

Samah Fodeh, Joseph Goulet, Cynthia Brandt, and Al-Talib Hamada. 2017. Leveraging twitter to better identify suicide risk. In *Medical Informatics and Healthcare*. pages 1–7.

Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014. Social structure and depression in trevorspace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, pages 615–625.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis* .

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods* 44(4):978–990.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* .

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)* http://arxiv.org/pdf/1301.3781v3.pdf.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013b. Pre-trained word2vec representation. *https://code.google.com/archive/p/word2vec/* .

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums .

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, pages 26–34.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions* 2(2):183–188.

Glen Pink, Will Radford, and Ben Hachey. 2016. Classification of mental health forum posts. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* .

Radim Rehuek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

HA Schwartz, Sap M, ML Kern, Eichstaedt JC, A Kapelner, Agrawal M., E Blanco, L Dziurzynski, G Park, D Stillwell, M Kosinski, M Seligman, and Ungar LH. 2016. Predicting individual well-being through the language of social media pages 516–527.

Benjamin Shickel and Parisa Rashidi. 2016. Automatic triage of mental health forum posts. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* .

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research* 13(Jun):2063–2067.

Hans Stadthagen-Gonzalez and Colin J. Davis. 2006. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods* 38(4):598–605. https://doi.org/10.3758/BF03193891.

Christopher M Homan Ravdeep Johar Tong, Liu Megan Lytle Vincent Silenzio Cecilia, and O Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. *ACL 2014* page 107.

Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. 2010. The interpersonal theory of suicide. *Psychological review* 117(2):575.

Luis Von Ahn. 2016. web-page: http://www.cs.cmu.edu/~biglou/resources/. Accessed: September 2016.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods* 45(4):1191–1207.

Cynthia Whissell. 1989. The dictionary of affect in language. *Emotion: Theory, research, and experience* 4(113-131):94.

WHO. 2014. *Preventing suicide: a global imperative*. World Health Organization.

WWBP World Well-Being Project. 2017. web-page: http://www.wwbp.org/data.html. Accessed: September 2017.

Ayah Zirikly, Varun Kumar, and Philip Resnik. 2016. The gw/umd clpsych 2016 shared task system.

A Supplemental Material

A.1 Features subsets

We build subsets of features, in which we selected the ones that we consider the most relevant in each category:

- *Subsets of body features (23):* self-harm regular expression (1), MentalDisLex (1), advisor and helplines keywords (2), negative PERMA features (5), neuroticism from OCEAN (1), affect lexicon from WWBP (1), pronouns (2), Hedonometer (1), negative lexicon from EmoLex (1) and word2vec semantic similarity to keywords (8).
- *Subsets of metadata features (7):* A selection of 5 boards (ToughTimes_Hosted_chats, Everyday_life_stuff, Intros, Something_Not_Right, Getting_Help), whether the author is a moderator or not (1), and whether the author created the thread (1).
- *Subsets of interaction features (57):* number of in/out edges from different authors (2), number of loops (1), number of authors post in the window (1), out degree of the author mentioned in the post (1), mean pronouns from incoming edges (2) and mean word2vec from incoming edges (50).
- The *subsets of author and adjacent features* (50 and 100 features respectively) consist of the subsets of features that consider Word2vec representations.
- *Subsets of N-grams (50):* We performed a random forest feature importance procedure over word2vec and N-grams, in which we kept the 100 most relevant features. The selected features consist of all the 50 Word2vec features and 50 N-grams, thus this procedure led us only to a discarding of N-grams.

A.2 Features collections

Starting from the set of all the features, we progressively discarded some of them, thus generating nine collections of features of decreasing quantity. Each collection was used to train SVM-linear and SVM-RBF models, resulting in 18 of our 22 models (the other four are ensembles).

The collections are:

- all_features (2799): all 2799 features
- ngrams_only (2274): the complete set of 2274 N-grams
- ngrams_filtered (575): all features but using the subset of N-grams instead of the complete set
- selection (337): Word2vec + N-grams subset + metadata subset + body subset + author subset + adjacent subset + interaction subset
- content (130): Word2vec + N-grams subset + metadata subset + body subset
- metadata (107): Word2vec + N-grams subset + metadata subset
- body (123): Word2vec + N-grams subset + body subset
- base (100): Word2vec + N-grams subset
- word2vec (50): Word2vec

A.3 Models comparison

In table 6 we compare the macro f-scores of different models in a 10-fold cross-validation scheme with the training set and the 337 features of the *selection* collection (described in section A.2). The models were implemented with sklearn or xgboost python packages. For each model, a grid search was applied to select the best parameters. For the SVM-RBF model the hyperparameter C was varied among {0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100}, for the SVM-linear among {0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1}, for the XGBoost the max_depth was varied among [2, 4, 6, 8] and the learning_rate among [0.001, 0.01, 0.1, 0.3] and for the Random Forest the max_features was varied among [10,20,40,60,80,100,120,140,160,200]. All other parameters were set to their default values. Among the models, the SVM classifiers outperformed the tree-based models. Given the large number of features (337), we also try a feature

model	CV macro f-score
SVM-RBF	0.549
SVM-linear	0.490
XGBoost	0.486
Random Forest	0.442

Table 6: Macro f-scores of different models in a 10-fold cross-validation scheme with the training set and the 337 features of the *selection* collection.

selection stage using the importance measure of a random forest classifier. In the grid search, not only the parameter C was varied but also the number of selected features, taking values among [50,100,150,200,250,300]. The best SVM RBF model obtained f-score=0.518 with the selection of the best 300 features, while the SVM linear model obtained f-score=0.514 with the selection of the best 250 features. Since the feature selection process did not produce significant performance improvements, it was not included in the contest models.

A.4 Ensemble models

We implemented four ensemble models composed by SVM's combined with a majority voting method.

The features sets of the voting models which compose the ensembles architectures are:

1. X = Word2vec + body subset + metadata subset (80)
2. X + N-grams subset (130)
3. Word2vec + metadata + body (141)
4. X + interaction (235)
5. X + adjacent (232)
6. X + author (157)
7. X + N-grams subset + author subset + adjacent subset + interaction subset (337)

We used two different structures of ensemble:

- ensemble_all: in which, each of the 7 features sets are used to train a SVM
- ensemble_4models: in which, only features sets 4, 5, 6 and 7 are used to train a SVM. These features sets are selected because are the ones that produce the best macro f-score in the Cross-Validation (data not showed).

These two ensemble structures implemented with SVMs-linear and SVMs-RBF result in our four ensemble models