Text-based experiments for predicting mental health emergencies in online web forum posts *

Hector Franco-Penya
hector.franco@dit.ie
Dublin Institute of Technology

Liliana Mamani Sanchez
mamanisl@tcd.ie
Trinity College Dublin

Abstract

This article explores how to build a system for detecting users in a need of attention on *ReachOut.com* forums. The proposed method uses Tree Kernels over binary Support Vector Machines classification and linear regression, comparing these two machine learning techniques. Predictions from one of these systems were submitted to the CLPsych 2016 Shared Task. Nonetheless, results indicate that it is possible to build an accurate system using only text features without the use of other meta data.

1 Introduction

Online communities such as web forums have become places where people participate according to common interests with other members of such communities. Language and interaction analysis may be done in these forums as to test hypothesis related to participation. Particularly, in web forums where the main topic of conversation is about issues related to their mental health, analysis may help address some situations where the well being of participants is compromised.

One of the duties of web forums moderators is to detect abnormal behaviour and take action over it. In the case of mental health web forums, the moderator should detect conversations that reveal a seemingly dangerous situation for the participants. For instance, conversations that might reveal that one of the participants wants to commit self-harm. The CLPsych Shared Task 2016 has the goal of evaluating systems that address the identification of web forum posts that reveal this kind of risk situations.

In order to assist moderators, this shared task consists on creating a system to automatically label posts, so moderators can identify where to focus their attention with more ease.

This report is structured as follows: Section 2 briefly describes the task and dataset, Section 3 presents all the details about the systems we built, Section 4 summarizes the results, Section 5 presents the discussion of these results, and finally we conclude with Section 6.

2 Task and dataset

The system has to classify each post into four categories that indicate how urgently a post needs the moderator's attention: green, amber, red or crisis. According to the annotation procedure carried on by the task organizers, those labels may be subdivided into twelve fine-grained categories shown in Table 1. This table also shows how many examples are present on the training dataset for each fine-grained category. For our experiments we only used the dataset of posts that have a label.

3 Systems description

Our systems are based on two machine learning techniques: 1) linear regression, and 2) three-step binary classification. For each technique, two types of features were extracted: grams (unigrams and bigrams), and grammatical tree structures. The system we submitted to the official CLPsych shared task is a

* Both authors contributed equally to the contents and experiments described in this paper.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 193–197,
San Diego, California, June 16, 2016.

Label	Fine-grained	Samples
green	allClear	366
	supporting	166
	followupBye	16
amber	followupOk	165
	currentMildDistress	40
	underserved	34
	pastDistress	10
red	currentAcuteDistress	87
	followupWorse	20
	angryWithReachout	2
	angryWithForumMember	1
	crisis	39

Table 1: Fine-grained distribution of labels in the training dataset.

gram-based linear regression system. From now on wards, we will refer to this as *baseline* system.

3.1 Pre-processing of web forum posts

In order to prepare the data for training a classifier system, text normalization was performed over two kinds of elements in posts: a) quoted text, and b) emoticons.

The inclusion of quoted text in post is frequent as it serves the purpose of clarifying which statements the post's writer is replying to. Since we are aiming to develop a text-based classification of posts into distinct categories, it is important to identify what is original post content and what is not. We consider quoted text cannot be deemed as original content, and can lead to missclassification. Therefore, we replaced quotations with a wilcard term.

Emoticons are signals of emotion expressed by using pictorial elements, or made up mostly of punctuation characters. We consider emoticons are essential on determining the writer's mood and are language independent to some extent. In the dataset provided, there is a large variation of emoticons instances that may convey similar mood, e.g. happy-smiley and very-happy-smiley. We reduced the possible set of emoticon labels and replaced them by wildcards. This approach is similar to the one followed in (Vogel and Mamani Sanchez, 2012) as they work with a dataset of pictorial emoticons extracted from the same web forum platform.

Other types of standardization were applied such as replacing HTTP links by wildcards.

3.2 Feature extraction

We describe here the linguistic and non-linguistic features that were extracted. Linguistic features were extracted after normalization.

N-grams Our baseline system uses unigrams and bigrams to create binary features to indicate if those grams occur in a post or not.

Tree kernels We used the Stanford parser (Klein and Manning, 2003b; Klein and Manning, 2003a) to generate constituent trees for all sentences from a single post. This generates a collection of trees, which where co-joined to have a tree representing the entire post. This structure was used thereafter in a tool that extract subtrees from such a tree and uses them as features to train a Support Vector Machine. For this purpose, we used the SVM-light implementation by (Joachims, 1999) and SubSet Tree kernel (SST) computation tool (Moschitti, 2006).

To our knowledge, SVMs over grammar trees for entire documents have not been explored before. Tree kernels are usually used to classify single sentences but not large pieces of text that could contain multiple paragraphs. This is due to the quadratic complexity of computing this kind of kernels.

Additional meta features In addition to text-based or linguistic features, we consider some additional features extracted from a post metadata. This metadata comprises the board name, a flag indicating if a posts is the first one in the thread or not, the rank (user category) of the post's author, and the base 10 logarithm plus one of number of views and the number of kudos. Names for our systems that used these additional features are suffixed with "full", while those that only use text features are suffixed with "textOnly". This naming convention is used in results in Table 3.

Table 2 shows the 20 user ranks labels and the number of users per rank. This table shows an unbalanced distribution of user across ranks: the first four categories ("Rookie scribe", "Casual scribe", "Rookie" and "Visitor") make 80% of the total of users, this produces a perplexity value of 7.3 (far from the value of 20 that could be reached if users were uniformly distributed across user categories).

rank	members
Rookie scribe	420
Casual scribe	402
Rookie	351
Visitor	151
Frequent scribe	90
Super frequent scribe	64
Youth Ambassador	39
Special Guest Contributor	24
Star contributor	20
Frequent Visitor	12
Staff	12
Contributor	11
Post Mod	11
Mod Squad	8
Community Manager	6
Mod	5
Uber contributor	5
Reachout.com Crew	4
Mod In Training	3
Super star contributor	2

Table 2: Author ranking

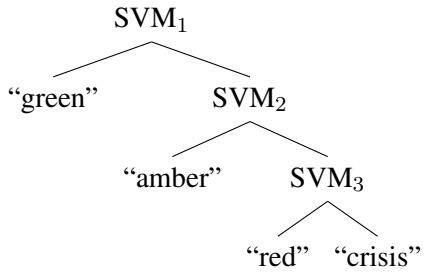

Figure 1: SVM classification

System	non-green		all-labels	
	acc	F1	acc	ma-F1
baseline	60%	.58	42 %	.13
reg_tree_full	**89%**	**.85**	73%	.28
reg_tree_textOnly	**89%**	**.85**	**78%**	**.38**
3s_tree_full	85%	.76	**78%**	.32
3s_tree_textOnly	77%	.67	69%	.29

Table 3: Results in terms of accuracy and F1 measures for green vs non-green classification, and for green vs all the other labels classification.

3.3 Architecture design

3.3.1 Linear regression systems

For the linear regression models, labels for the training set posts were mapped to an ordinal scale according to how urgently a post needs attention: "green" was mapped to 0, "amber" to 1, "red" to 2 and "crisis" to 3.

Then SVM-light software was used to create the model. In the evaluation stage, the predicted values for the test set were used to rank the posts according to their need of attention, for which the higher values where labelled as "crisis", then "red", "amber" and "green" following the same distribution as in the training set: "crisis" 4.1%, "red" 11.7%, "amber" 26.3% and "green" 57.9%. Linear regression systems are prefixed with "reg".

3.3.2 Three step binary classification systems

The three-step binary classification systems are developed as decision trees of three nodes. Decisions in each node are calculated according to classification performed by a Support Vector Machine (SVM). The first SVM decides if the post has "green" or "non-green" as a label. If the example is labelled as "non-green", the second SVM decides if the posts is labelled "amber" or "non-amber". If the example is labelled as "non-amber", the third SVM decides if the label is "red" or "crisis". Figure 1 illustrates this procedure.

The training set for each SVM only contains relevant examples for the specific step. This means that the first SVM is trained with all examples that have a "green" label as negative samples, and the remaining examples are deemed positive examples. The examples labelled as "green" are not used to train the second and third SVMs. Three-step binary classification systems are prefixed with "3s".

4 Results

Table 3 reports results for the systems accuracy and macro F1 measures. The first two columns report the results of predicting posts that need attention, where all the labels but "green" were unified into a single category "non-green". The last two columns report results for all labels. The macro-F1 measure is low mainly because all systems failed to identify the single "crisis" post. This lead to a F1 value of zero for prediction of "crisis", this drags down the macro accuracy value since all labels have the same weight.

It is puzzling, that the system that which produces best results is the tree kernel based linear regression based uniquely on the text of the posts, as our in-

	positives	negatives	n/p ratio
SVM_1	42.1%	57.9%	1.375
SVM_2	37.5%	62.5%	1.666
SVM_3	25.9%	74.1%	2.861

Table 4: Positives and negatives per SVM step

tuition suggests this should have been outperformed by the variation that includes metadata, which is the case when comparing the two tree kernel systems based on three binary classification steps. Also, the regression models seem to outperform the other systems in the detection of non-green labels. The success of the linear regression systems could be related to the fact that the regression models do have a quota of predictions for each type of labels.

Due to time limitations only the baseline system was submitted on time for the public evaluation.

5 Discussion and future work

The tree model shown in Section 3.3.2 was designed as a three-step decision tree based on machine learning classifiers. These steps decide first the label in growing order, this way each machine learning step has a fairly balanced training set, which gets more unbalanced as the labels involved in the decision have higher priority than in the first step. Figure 4 illustrate this observation. Any other combination of steps would lead to more unbalanced training sets; it would be necessary to use balancing techniques.

Another possible design would involve the use of the eleven binary classification steps as described in the annotation procedure document provided by the organizers. Therefore, the classifier systems should be designed to mimic this annotation procedure. As a final step, the eleven fine-grained labels should be converted back the original four-label range used in the competition. This system would had been substantially more complex, the first step would have had to classify a sample as a "crisis" or "non-crisis". In such case, the first machine learning classifier would had dealt with a very unbalanced training set as only 4.1% of samples are labelled as "crisis".

Some sparse fine-grained labels would had been very difficult to predict such as "angryWithForumMember" (1 example in the training set), or "angryWithReachout" (2 examples in the training set).

The prediction of the labels: "followupBye", "followupOk", and "followupWorse" could benefit from analysing and labelling previous posts in a thread as they only exist as following posts labelled as "red" or "crisis", and features extracted from these posts may not help the prediction of other labels.

These observations suggest a major change on the design of the system in which all posts of a thread should be labelled and re-labelled based on the previous posts in the thread and according to author roles. We consider this fine-grained model as future work. The linear regression model proposed in Section 3.2 only requires one step of machine learning classification. However, it requires to map ordinal data into numerical to create the training set and numerical into ordinal to interpret the predictions. For the proposed system, labels are mapped into consecutive numbers, this assumes that the difference between consecutive labels are the same. Which may not be the case, perhaps "crisis" posts should be mapped to a much larger value than "red" posts. Perhaps the mapping function should be related to the percentile in which the (mapped) values appear, or some other feature. The problem of mapping ordinal data into numerical is another open research topic outside the scope of this experiment. Tuning of the mapping procedure is left for future work.

6 Conclusions

We have described the basic setup for systems that address the CLPsych 2016 Shared Task. Our systems do not reach top positions in the ranking for this competition, however they provide some opportunities to explore ideas on how to deal with this kind of classification task. The main principle followed on designing these systems was to make them as portable as possible and independent of exogenous features to the post's contents. There is several aspects to improve if the goal is to build system for post classification that are uniquely based on text. Besides our goals summarized in the section for future work, one issue to explore further is to determine how noisy text affects classification.

Overall, we also have to explore the corresponding caveats of relying only on text for building classifier systems.

References

Thorsten Joachims. 1999. Making large scale svm learning practical. Technical report, Universität Dortmund.

Dan Klein and Christopher D. Manning. 2003a. Accurate Unlexicalized Parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, 1:423–430.

Dan Klein and Christopher D Manning. 2003b. Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems*, 15:3–10.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *EACL*, volume 113, page 24.

Carl Vogel and Liliana Mamani Sanchez. 2012. Epistemic signals and emoticons affect kudos. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 517 –522, Dec.