Data61-CSIRO systems at the CLPsych 2016 Shared Task

Sunghwan Mac Kim[1], Yufei Wang[2]*, Stephen Wan[1] and **Cécile Paris[1]**
[1]Data61, CSIRO, Sydney, Australia
[2]School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, Australia
Mac.Kim@csiro.au, Yufei.Wang1@uq.net.au,
Stephen.Wan@csiro.au, Cecile.Paris@csiro.au

Abstract

This paper describes the Data61-CSIRO text classification systems submitted as part of the CLPsych 2016 shared task. The aim of the shared task is to develop automated systems that can help mental health professionals with the process of triaging posts with ideations of depression and/or self-harm. We structured our participation in the CLPsych 2016 shared task in order to focus on different facets of modelling online forum discussions: (i) vector space representations; (ii) different text granularities; and (iii) fine- versus coarse-grained labels indicating concern. We achieved an F1-score of 0.42 using an ensemble classification approach that predicts fine-grained labels of concern. This was the best score obtained by any submitted system in the 2016 shared task.

1 Introduction

The aim of the shared task is to research and develop automatic systems that can help mental health professionals with the process of triaging posts with ideations of depression and/or self-harm. We structured our participation in the CLPsych 2016 shared task in order to focus on different facets of modelling online forum discussions: (i) vector space representations (TF-IDF vs. embeddings); (ii) different text granularities (e.g., sentences vs posts); and (iii) fine- versus coarse-grained (FG and CG respectively) labels indicating concern.

(i) For our exploration of vector space representations, we explored the traditional TF-IDF feature representation that has been widely applied to NLP. We also investigated the use of post embeddings, which have recently attracted much attention as feature vectors for representing text (Zhou et al., 2015; Salehi et al., 2015). Here, as in other related work (Guo et al., 2014), the post embeddings are learned from the unlabelled data as features for supervised classifiers. (ii) Our exploration of text granularity focuses on classifiers for sentences as well as posts. For the sentence-level classifiers, a post is split into sentences as the basic unit of annotation using a sentence segmenter. (iii) To explore the granularity of labels indicating concern, we note that the data includes a set of 12 FG labels representing factors that assist in deciding on whether a post is concerning or not. These are in addition to 4 CG labels.

We trained 6 single classifiers based on different combinations of vector space features, text granularities and label sets. We also explored ensemble classifiers (based on these 6 single classifiers), as this is a way of combining the strengths of the single classifiers. We used one of two ensemble methods: majority voting and probability scores over labels. We submitted five different systems as submissions to the shared task. Two of them were based on single classifiers, whereas the remaining three systems used ensemble-based classifiers. We achieved an F1-score of 0.42 using an ensemble classification approach that predicts FG labels of concern. This was the best score obtained by any submitted system in the 2016 shared task.

The paper is organised as follows: Section 2 briefly discusses the data of the shared task. Section 3 presents the details of the systems we sub-

*This work was performed while Yufei was at CSIRO.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 128–132,
San Diego, California, June 16, 2016.

mitted. Section 4 then shows experimental results. Finally, we summarise our findings in Section 5.

2 Data

The dataset used in the shared task is a collection of online posts crawled from a mental health forum, ReachOut.com[1], collected by the shared task annotators, who then labelled each discussion post with one of 4 CG labels: *Green*, *Amber*, *Red* and *Crisis*, describing how likely a post is to require the attention of a mental health professional. Each post is also annotated with one of 12 FG labels, which are mapped deterministically to one of the 4 CG labels according to the relationships presented in Table 1 (which also provides the frequencies of these relationships). For instance, a post labelled with *Red* could be labelled with one of 4 FG labels: *angryWithForumMember*, *angryWithReachout*, *currentAcuteDistress* and *followupWorse*. As can be seen in the table, the dataset is imbalanced since it contains more *Green* labelled posts than any other post.

The corpus consists of 65,024 posts, and it is subdivided into labelled (947) and unlabelled data (64,077). The final test data contains an extra 241 forum posts. Each post is provided in an XML file and each post file contains metadata, such as the number of "likes" a post received from the online community. The shared task requires each submitted system to predict a label for each of test posts.

In addition to the post data, the data set contains anonymised metadata about post authors, which indicates whether authors were affiliated with ReachOut, either as a community moderator or a site administrator. Specifically, this metadata contains anonymised author IDs and their forum ranking. In total, there were 1,640 unique authors and 20 author rankings on the forums. Each author has one of the 20 rankings. 7 ranking types indicate ReachOut affiliated, whereas 13 author ranking types represent a member of the general public.

3 Systems Description

3.1 Text Pre-processing

We performed several text pre-processing steps prior to feature extraction in order to reduce the noisiness of the original forum posts. We removed HTML special characters, non-ASCII characters and stop words, and all tokens were lower-cased. We used NLTK (Bird et al., 2009) to segment sentences for the sentence-level classifiers, producing 4,305 sentences from the 947 posts.

CG label	Frequency	FG label	Frequency
Green	549	allClear	367
		followupBye	16
		supporting	166
Amber	249	underserved	34
		currentMildDistress	40
		followupOk	165
		pastDistress	10
Red	110	angryWithForumMember	1
		angryWithReachout	2
		currentAcuteDistress	87
		followupWorse	20
Crisis	39	crisis	39

Table 1: CG and FG label sets. Their frequencies represent the number of posts in the labelled dataset.

3.2 Features

We used two types of feature representations for the text: TF-IDF and post embeddings. The TF-IDF feature vectors of unigrams were generated from the labelled dataset, whereas the embeddings were obtained using both labelled and unlabelled dataset using sent2vec (Le and Mikolov, 2014). We obtained the embeddings for the whole post directly instead of combining the embeddings for the individual words of the post due to the superior performance of document embeddings (Sun et al., 2015; Tang et al., 2015).

In our preliminary investigations, we explored various kinds of features such as bi- and trigrams, metadata from the posts (such as the number of views of a post or the author's affiliation with ReachOut) and orthographic features (for example, the presence of emoticons, punctuation, etc.), but we did not obtain any performance benefits with respect to intrinsic evaluations on the training data.

3.3 Classifiers

For the text classifiers, we trained a MaxEnt model using scikit-learn's SGDClassifier (Pedregosa et al., 2011) with the log loss function and a learning rate of 0.0001 as our classifier for all experiments. In the training phrase, the weights of SGDClassifier are

[1] http://forums.au.reachout.com/

optimised using stochastic gradient descent (SGD) through minimising a given loss function, and L2 regularisation is employed to avoid overfitting. The log loss function in SGDClassifier allows us to obtain the probability score of a label at prediction time.

We developed classifiers for two granularities of text: (i) entire posts, and (ii) sentences in posts. For the latter, we post-processed the predicted sentence-level labels to produce post-level labels (to be consistent with the shared task). We obtained distributions of probabilities for the label sets for each sentence, and then summed the distributions for all sentences in a post. This provided a final distribution of probabilities for labels for a post. The label with the highest probability was then taken as the inferred label for the post.

To perform the post-processing steps above, we used the distributions for labels produced by the MaxEnt model. That is, the model can be used to provide estimates for the probabilities of:

- CG labels given a post, $P(CG\ label|post)$;
- CG labels given a sentence, $P(CG\ label|sentence)$;
- FG labels given a post, $P(FG\ label|post)$; and
- FG labels given a sentence, $P(FG\ label|sentence)$.

We also developed classifiers for the CG and FG label sets. In the case of the FG set, we again performed post-processing steps to produce CG labels. In this case, we deterministically reduced the predicted 12 labels to the 4 CG labels, using the mapping presented in Table 1.

This allowed us to experiment with different combinations of the 3 facets, described in Section 1. We built 6 classifiers based on the combination of the configurations described so far as follows:

C1. post-level TF-IDF classifier using 4 labels
C2. post-level embedding classifier using 4 labels
C3. sentence-level TF-IDF classifier using 4 labels
C4. post-level TF-IDF classifier using 12 labels
C5. post-level embedding classifier using 12 labels
C6. sentence-level TF-IDF classifier using 12 labels

3.4 Ensembles

One reason why the ensemble approaches may work well is that, even if a classifier does not pick the correct label, the probabilities for all labels can still be taken as input to the ensemble approach. For example, although a classifier may have chosen a label incorrectly, the correct label could have had the second highest probability score, which when combined with information from other classifiers may lead to the correct label being assigned.

System	Training results	Official test results
Post-tfidf-4labels	0.25	0.39
Sent-tfidf-12labels	0.35	0.37
Ensb-6classifiers-mv	**0.37**	0.37
Ensb-3classifiers-4labels-prob	0.35	0.35
Ensb-3classifiers-12labels-prob	**0.37**	**0.42**

Table 2: F1 results for 5-fold cross-validation on training data and the official test results from the shared task.

As mentioned in Section 1, the outputs of the ensemble models were produced using one of two ensemble methods: majority voting and probability scores over labels. In the majority voting method, each classifier votes for a single label, and the label with highest number of votes is selected for the final decision. The second ensemble method uses an estimate of the posterior probability for each label from individual classifiers, and the label with highest sum of probabilities is chosen for the final prediction. Neither ensemble method requires any parameter tuning.

3.5 Submitted Systems

Five different systems were adopted for our submissions to the shared task. Two were based on a single MaxEnt classifier, whereas the remaining three systems used ensemble-based classifiers. The two single classifiers were as follows:

1. a single classifier C1 (*Post-tfidf-4labels*)
2. a single classifier C6 (*Sent-tfidf-12labels*)

And the three ensemble classifiers are:

3. an ensemble classifier combining all six C1-C6 by majority voting (*Ensb-6classifiers-mv*)
4. an ensemble classifier combining C1, C2, C3 by posterior probabilities (*Ensb-3classifiers-4labels-prob*)
5. an ensemble classifier combining C4, C5, C6 by posterior probabilities (*Ensb-3classifiers-12labels-prob*)

The *Post-tfidf-4labels* system uses a standard approach predicting 4 CG labels with respect to posts using TF-IDF feature representation. The *Sent-tfidf-12labels* system predicts 12 fined-grained labels for sentences using the same feature representation method. The *Ensb-6classifiers-mv* system combines all judgements of the six MaxEnt classifiers described in Section 3.3 through majority voting. The

System	F1	Accuracy	Filter F1	Filter Accuracy
Post-tfidf-4labels	0.39	0.81	0.82	0.88
Sent-tfidf-12labels	0.37	0.80	0.81	0.88
Ensb-6classifiers-mv	0.37	0.83	0.81	0.90
Ensb-3classifiers-4labels-prob	0.35	0.82	0.80	0.89
Ensb-3classifiers-12labels-prob	**0.42**	**0.85**	**0.85**	**0.91**

Table 3: Results for the test set. The filter decides whether the label of a forum post is *green* or not (non-green vs. green).

		P	R	F1
Ensb-3classifiers-4labels-prob	Amber	**0.60**	0.57	0.59
	Red	0.69	**0.33**	0.45
	Crisis	0.00	0.00	0.00
Ensb-3classifiers-12labels-prob	Amber	**0.71**	0.53	0.61
	Red	0.68	**0.63**	0.65
	Crisis	0.00	0.00	0.00

Table 4: Comparison results on the test dataset in terms of precision, recall and F1.

remaining two systems, *Ensb-3classifiers-4labels-prob* and *Ensb-3classifiers-12labels-prob*, use the sum of label probabilities estimated from individual classifiers to select the most probable label. The main difference between the two systems is the estimation of probability scores in different level of label granularities (CG labels vs. FG labels).

4 Experimental Results

In this section, we present two evaluation results: the cross-validation results and the final test results. We performed 5-fold cross-validation on the training set (947 labelled posts). We also report the shared task evaluation scores for the five systems on the test set of 214 posts. These are shown in Table 2 where scores are computed for three labels: *Amber*, *Red* and *Crisis* (but not Green), since this is the official evaluation metric in the shared task.

We observe that two of the ensemble systems (*Ensb-6classifiers-mv* and *Ensb-3classifiers-12labels-prob*) show higher F1-scores than the others in the cross-validation experiments. In particular, *Ensb-3classifiers-12labels-prob* performs best both in the cross-validation experiment (0.37) and the main competition (0.42).

Somewhat surprisingly, the first system, *Post-tfidf-4labels*, gave us an F1-score of 0.39 on the test data, while its F1-score was the lowest in the cross-validation experiment. This result indicates that good performance is possible on the test dataset using a "textbook" TF-IDF classifier but further investigation is required to understand why the official test result differs from our cross-validation result.

Table 3 shows the superior performance of the *Ensb-3classifiers-12labels-prob*, with respect to the other systems in terms of F1 and accuracy. It achieved the highest accuracy (0.85) for the three labels. Furthermore, it is a robust system for identifying the non-concerning label, *Green*.

It is interesting to see that the F1-score was improved by performing the hard classification task of 12 labels compared to 4-label classification. We compare the performance of the *Ensb-3classifiers-4labels-prob* and *Ensb-3classifiers-12labels-prob* systems on the test data per label, as shown in Table 4 to shed light on why the 12-labelling system has superior performance. Both systems were unable to detect any *Crisis*-labelled posts. A notable difference between the two systems is that the *Ensb-3classifiers-12labels-prob* system produces significantly higher recall (0.63) than the *Ensb-3classifiers-4labels-prob* system (0.33). In addition, the *Ensb-3classifiers-12labels-prob* system has a higher precision for finding *Amber* posts. These results consequently led to overall better F1 as shown in Table 3, and suggest that identifying *Green* and *Amber* posts for a user-in-the-loop scenario may be one way to help moderators save time in triaging posts.

5 Conclusion

We applied single and ensemble classifiers to the task of classifying online forum posts based on the likelihood of a mental health professional being required to intervene in the discussion. We achieved an F1-score of 0.42 with a system that combined post and sentence-level classifications through probability scores to produce FG labels. This was the best score obtained by any submitted system in the 2016 shared task. The experimental results suggest that identifying *Green* and *Amber* posts for a user-in-the-loop scenario may be one way to help moderators save time in triaging posts.

Acknowledgments

We would like to thank the organisers of the shared task for their support.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. JMLR Workshop and Conference Proceedings.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado, May–June. Association for Computational Linguistics.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 136–145, Beijing, China, July. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal, September. Association for Computational Linguistics.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 250–259, Beijing, China, July. Association for Computational Linguistics.