Combining Multiple Classifiers Using Global Ranking for ReachOut.com Post Triage

Chen-Kai Wang[1], Hong-Jie Dai[1*], Chih-Wei Chen[2], Jitendra Jonnagaddala[3] and **Nai-Wen Chang[4]**

[1]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan
[2]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan
[3]School of Public Health and Community Medicine, University of New South Wales, Sydney, Australia
[4]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

Abstract

In this paper, we present our methods for the 2016 CLPPsych shared task. We extracted and selected eight features from the corpus consisting of posts from ReachOut.com including the information of the post's source board, numbers of kudos and views, post time, ranks of the authors, unigram of the body and subject, frequency of the used emotion icons, and the topic model features. Two support vector machine models were trained with the extracted features. A baseline system was also developed, which uses the calculated log likelihood ratio (LLR) for each token to rank a post. Finally, the prediction results of the above three systems were integrated by using a global ranking algorithm with the weighted Borda-fuse (WBF) model and the linear combination model. The best F-score achieved by our systems is 0.3 which is based on the global ranking method with WBF.

1 Introduction

The Internet and the WWW (World Wide Web) provide ubiquitous access to the information all around the world, drastically remodeling how humans acquaint facts, comprehend knowledge and communicate with others. For example, at online health communities, patients and their close persons learn diseases and gain insights, seek and offer helps and supports, and become familiar with others with similar conditions (Neal et al., 2006). Physicians and other medical professionals also involves online health communities through content sites, web forums, social media, or other means, providing advices and services (Guseh et al., 2009).

Online board moderators save psychiatric patients from emotional distresses and suicidal attempts (Barak 2007). With proper modulation, even previous self-harm patients become altruistic board members and helpers (Smithson et al., 2011). However, some unmodulated online forums may contain improper posts and messages, influencing and guiding patients' judgements and behaviors in deviant ways. Some self-harm victims report learned behaviors from online forums (Dunlop et al., 2011).

As the messages at online forums are numerous, manual evaluations and responses by broad modulators become tedious and helps may be delayed. With the advances in natural language processing

* Corresponding author.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 176–179,
San Diego, California, June 16, 2016.

(NLP), automatic text categorization become possible and might be integrated into the code base of online forums to assist modulators. The third annual computational linguistics and clinical psychology (CLPsych) workshop, held by Association for Computational Linguistics, focus on language technology applications in mental and neurological health. As a participant of the 2016 CLPPsych shared task, we develop a NLP system to automatically classify posts from ReachOut.com mental health forum into one of the red/amber/green/crisis semaphore that indicates how urgently a post needs moderator attention.

2 Methods

The input of our system is the ReachOut.com forum posts represented in the XML format. The following features are extracted from the structural content and selected by the information gain algorithm using tenfold cross validation (CV) on the training set.

1. Source board: The link of the board contains the post, such as "/boards/id/Something_Not_Right", and "/boards/id/TroughTimes_Hosted_chats" is extracted as a nominal feature.

2. Kudos: The number of kudos (equivalent of up-vote) given to a post is extracted as a numeric feature.

3. Post and the last edit time: The creation and the most recent edit timestamp for the post. In this work, the value was equally discretized into 24 distinct ranges and encoded as a nominal feature to indicate a certain hour of a day.

4. Views: The number of times the post has been viewed is extracted as a numeric feature.

5. Rank of the author: The rank, such as "Mod Squad", and "Frequent scribe", of the author of the post in the forum was extracted and encoded as a nominal feature.

6. Subject and body: The text of the post's subject and the body of the post were extracted. Because the content of the body includes escaped HTML tags, Apache Tika[1] was used to extract all of the plain texts from these HTML tags. Twokenizer (Owoputi et al., 2013) was then used to tokenize the extracted texts. Finally, the normalization process proposed by Lin et al. (2015) was used to normalize all tokens. The unigrams of the normalized texts from both the subject and body were extracted as features.

7. Emotion icon frequency: Based on all of the extract body contents, twelve emotion icon types used in the forum were observed, which include "Happy", "VeryHappy", "Tongue", "Embarassed", "Frustrated", "Wink", "Surprised", "Heart" "LOL", "Indifferent" and "Mad". The frequencies of the occurrences of the above icons were determined by parsing the body content for each post.

8. Topic model: The features were produced in two steps. The first step was to train a topic model using the training set and the second step was to use the trained model to generate features. The type of topic modelling features extracted in this study include (1) the topic distribution weights per instance and (2) the binary features to represent the presence of a keyword term (obtained from the topics generated) in a given instance. Above features were created by using Stanford topic modeling toolbox[2].

The extracted features trained with the support vector machine (SVM) (Cortes et al., 1995). Two SVM models were created. One used features one to seven and the other used all eight features. In addition to the supervised learning method, a baseline system based on the log likelihood ratio (LLR) was developed. In this system, we ranked the tokens observed in the training dataset based on their values calculated by using LLR and selected the tokens with positive values to compile a keyword list for each triage label. The compiled lists were then used to rank a given post. The triage label with the highest LLR value is selected as the output for the post.

[1] https://tika.apache.org
[2] http://nlp.stanford.edu/software/tmt/tmt-0.4/

Run	Official F-score	Accuracy
1	0.29	0.76
2	0.26	0.63
3	0.22	0.66
4	0.3	0.73
5	0.28	0.69

Table 1: Official F-score and accuracy of the submitted runs.

Run	Non-green vs. green macro F-score	Non-green vs. green accuracy
1	0.8	0.87
2	0.66	0.74
3	0.62	0.76
4	0.76	0.83
5	0.69	0.79

Table 2: Non-green vs. green F-score and accuracy of the submitted runs.

Finally, we merge all results of above three systems by using a global ranking method based on two data fusion algorithms (Dai et al., 2010). First of all, the outputs of all three systems were collected and their performance on the tenfold CV training set were determined. The simple weighting scheme based on the weighted Borda-fuse (WBF) model was employed, which multiply the points assigned to a semaphore determined by a system by the F-score of that system. The second fusion algorithm is the linear combination (LC) model which multiplies the predictions probability of a semaphore determined a system by the F-score of that system.

3 Results

We submitted five runs. Both the first and the second runs are based on SVM. As mentioned in the Methods section, the first run includes features one to seven, while the second run further adds the topic model feature. The third run is the baseline system based on LLR. The fourth and fifth runs are created by using the global ranking method with WBF and LC, respectively. Table 1 and 2 shows the results of the submitted runs.

As shown in Table 1, the best run of our system achieves an F-score of 0.3, which is based on the global ranking with WBF. The second best run is the SVM model w/o topic model features. However, the difference between the two runs may not be significant.

4 Discussion

Here we only focus on the comparison between the run 1 and 2. A manual inspection of the keyword terms within the topics generated from the training set shows that the topics didn't quite capture the themes correctly, as the words within the topic don't belong to a particular theme. For example, if we see the top keyword terms within the topics for 4-topic model, negative sentiment is not captured effectively though the dataset had several negative sentiment themed topics. In addition, words which stand for positive and negative sentiments are grouped under the same topic. Also, the analysis of the document topic distribution shows that in almost all the instances, one particular topic is having the most weight, making it hard for our classifier to discriminate themes and sentiments.

5 Conclusion

This work selected eight features and studied their impact for the triage task of ReachOut.com posts. The global ranking algorithm is then used to combine the generated results from three systems. In the future work, we consider to apply ensemble classifiers and compare the results with the ranking-based method.

References

Azy Barak. 2007. *Emotional support and suicide prevention through the Internet: A field project report*, Computers in Human Behavior 23(2): 971-984.

Corinna Cortes and Vladimir Vapnik. 1995. *Support-vector networks*, Machine Learning 20(3): 273-297.

Hong-Jie Dai, Po-Ting Lai, Richard Tzong-Han Tsai and Wen Lian Hsu. 2010. *Global Ranking via Data Fusion*, Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 223-231.

Sally M. Dunlop, Eian More and Daniel Romer Dunlop. 2011. *Where do youth learn about suicides on the Internet, and what influence does this have on suicidal ideation?*, Journal of Child Psychology and Psychiatry 52(10): 1073-1080.

James S. Guseh, Rebecca W. Brendel and D. H. Brendel 2009. *Medical professionalism in the age of online social networking*, Journal of Medical Ethics 35(9): 584-586.

Wei-San Lin, Hong-Jie Dai, Jitendra Jonnagaddala, Nai-Wun Chang, Toni Rose Jue, Usman Iqbal, Joni Yu-Hsuan Shao, I-Jen Chiang and Yu-Chuan Li Lin. 2015. *Utilizing Different Word Representation Methods for*

Twitter Data in Adverse Drug Reactions Extraction, Proceedings of the 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI). Tainan, IEEE**:** 260-265.

Lisa Neal, Gitte Lindgaard, Kate Oakley, Derek Hansen, Sandra Kogan, Jan Marco Leimeister, and Ted Selker. 2006. *Online health communities*, CHI '06 Extended Abstracts on Human Factors in Computing Systems. ACM**:** 444-447.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. *Improved part-of-speech tagging for online conversational text with word clusters*, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics.

Janet Smithson, Siobhan Sharkey, Hewis Elaine, Ray B Jones, Tobit Emmens, Tamsin Ford and Christabel Owens. 2011. *Membership and Boundary Maintenance on an Online Self-Harm Forum*, Qualitative Health Research.