Automatic Triage of Mental Health Online Forum Posts
CLPsych 2016 System Description

Hayda Almeida
Concordia University
Montreal, QC, Canada

Marc Queudot
Université du Québec à Montréal
Montreal, QC, Canada

Marie-Jean Meurs
Université du Québec à Montréal
Montreal, QC, Canada
meurs.marie-jean@uqam.ca

Abstract

This paper presents a system capable of performing automatic triage of forum posts from ReachOut.com, a mental health online forum. The system assigns to each post a tag that indicates how urgently moderator attention is needed. The evaluation is based on experiments conducted on the CLPsych 2016 task, and the system is released as an open-source software.

1 Introduction

This paper describes a system that was presented at the CLPsych Shared Task 2016[1]. The goal of the task is to perform automatic triage of user posts gathered from the ReachOut.com mental health online forum[2]. Posts must be classified into four categories (green, amber, red, and crisis), which indicate how urgently any intervention from forum moderators is required. The automatic triage of ReachOut forum posts is a challenging task. First, the targeted documents - from the amber, red, and crisis classes - are highly underrepresented in the data to be analyzed. Second, forum post content can be highly noisy, since posts commonly present symbols, emoticons, pictures, and mispelled words.

The objective of an automatic triage of ReachOut posts is to allow forum moderators to quickly identify posts that require urgent intervention. Posts labeled as red or crisis could indicate an imminent dangerous or harmful condition, for example, an author that suggests a possibility of self-harm.

To handle the task of ReachOut post automatic triage, we propose a system relying on the combination of two text classification techniques, namely supervised learning and rule-based classification. Our experiments are performed utilizing three classification algorithms, and classification rules designed based on discriminative vocabularies selected from documents of the minority classes. In addition, we studied the use of different feature types and subsets.

This paper is organized as follows: Section 2 describes some related works while Section 3 provides details about our approach, and the system architecture. Experiments and results are reported in Section 4, and we conclude in Section 5.

2 Related Work

The automatic triage of documents can be used to support a variety of data handling processes. It supports professionals and researchers working in the medical (Tuarob et al., 2014; Almeida et al., 2015) or biological fields (Almeida et al., 2014). Data gathered from forum posts have been used in several related classification tasks. In (Huh et al., 2013), the triage supports patients handling several health conditions, while it was used to identify mental health issues in (Saleem et al., 2012), and to recognize user sentiments in (Thelwall et al., 2012).

Designing efficient automatic approaches for textual data triage can be challenging, especially when documents of interest represent a very small part of the entire dataset. Machine learning approaches are impacted by the class distribution, and many classifiers do not perform well in unbalanced contexts. Support Vector Machines (SVM) (Vapnik,

[1] http://clpsych.org/shared-task-2016/
[2] http://au.reachout.com/

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 183–187,
San Diego, California, June 16, 2016.

1995) were previously utilized in forum post triage handling mental health subjects (Saleem et al., 2012). Models using Sequential Minimal Optimization (SMO) (Platt, 1998) for optimizing SVM, were applied to perform sentiment analysis in forum data, outperforming other methods when used on large datasets (Thelwall et al., 2012). Logistic Model Trees (LMT) (Landwehr et al., 2005) were shown to outperform other classification algorithms in tasks that handle (highly) imbalanced data (Charton et al., 2013; Almeida et al., 2014). Previous studies have combined rule-based and supervised classification approaches to handle forum posts (Saleem et al., 2012), patients medical records (Xu et al., 2012), or sentiment in social media (Chikersal et al.,). In these works, combined strategies usually obtained better performance compared to supervised only or rule-based only approaches.

The use of lexical features, such as n-grams, Part-Of-Speech (POS) tags, and lemmas, as well as sentiment dictionaries, were shown to perform well in tasks handling forum posts (Biyani et al., 2014), and mining sentiments or opinion (Thelwall et al., 2012). Feature selection methods have been studied to choose relevant attribute subsets (Liu et al., 2010; Basu and Murthy, 2012). Among these methods, Correlation-based Feature Selection (CFS) selects a subset of attributes that are *highly correlated with the class, yet uncorrelated with each other* (Hall, 1999). Methods to determine relevant vocabulary for specific class labels were previously studied (Melville et al., 2009; Charton et al., 2013). Melville et al. (2009) built a discriminative vocabulary to represent sentiment polarity, while Charton et al. (2013) used one to represent minority classes. In both cases, the use of discriminative vocabularies in the classification models improved performance.

3 Methodology

To tackle the task of automatic triage of forum posts, the proposed system combines rule-based and machine learning based classification. Our approach makes use of several feature types, such as n-grams, POS tags, and a sentiment dictionary generated from two sentiment libraries. Various features subsets were filtered using the CFS feature selection method. In the following sections we explain with more details the system pipeline, and the methods utilized in each step.

	Training		Test	
Class	# posts	ratio (%)	# posts	ratio (%)
Green	549	57.49	166	68.88
Amber	249	26.30	47	19.50
Red	110	11.61	27	11.20
Crisis	39	4.18	1	0.42
Total	947	100	241	100

Table 1: Statistics on the CLPsych Shared Task 2016 dataset

3.1 CLPsych Dataset

The CLPsych corpus consists of 65024 publicly available posts gathered from the ReachOut forum, which have been posted between July 2012 and May 2015. Among these posts, 1188 posts were manually annotated with class labels, then split into a training and a test set. The training set is composed of 947 posts while the test set contains 241 posts. The class distribution on the training and the test data is shown in Table 1.

3.2 Feature Extraction and Selection

Prior to performing feature extraction, the forum posts were pre-processed by normalization procedures, which included normalizing HTML characters, symbols, punctuation, smiley pictures, and smiley symbols. Each smiley was replaced by a corresponding word extracted either from the picture URL, or from a concise mapping containing the smiley textual meaning (e.g., `:)` or `=]` or `:D` are all replaced by `happy`). The features used in our experiments were of type bigrams, POS tags, and sentiments. Extraction of POS tags was performed using the POSTaggerAnnotator from the Stanford CoreNLP suite (Manning et al., 2014). POS features are composed of forum post words annotated with discriminative POS tags, which were adjective (JJ*), nouns (NN*), predeterminer (PDT), particle (RP), and verbs (VB*). The selection of discriminative POS tags was based on experimental results. Sentiment features are dataset lemmas found within a sentiment dictionary. The dataset lemmas were extracted using the Stanford CoreNLP suite. We built a sentiment dictionary based on a list of feeling words used in mental status exams (see `http://psychpage.com/learning/library/assess/feelings.html`), and a conceptual feature

Feature type	# features	# CFS features
Bigrams	35,442	73
POS	5,828	43
Sentiments	2,387	45

Table 2: Number of unique features in CLPsych dataset

map from SenticNet (Cambria et al., 2014). Stopwords were not removed from the data, since they seem to carry relevant discriminative power for the task, as previously demonstrated by (Saif et al., 2014). All feature lists were separately filtered by the CFS method. Feature distributions by type before and after CFS filtering are reported in Table 2.

3.3 Classification Algorithms

We performed experiments utilizing three classification algorithms: Bayesian Network (BN) (Pearl, 1988), SMO, and LMT. A BN is a probabilistic directed acyclic graph, in which nodes are random variables with arcs representing their conditional dependencies. BN was used as a baseline classifier. SMO-SVM were previously applied in similar tasks as described in Section 2. SMO (Platt, 1998) is an optimization algorithm for training SVMs. SMO is an iterative algorithm that solves the quadratic programming problem of SVM training by breaking it into smaller sub-problems easier to solve. As described in Section 2, LMT previously demonstrated good performance in classification tasks on imbalanced datasets. LMT is an algorithm that produces decision trees with linear logistic models at the leaves.

3.4 Discriminative Vocabulary Rules

For the red and the crisis classes, a discriminative vocabulary was utilized to develop classification rules. The discriminative vocabulary was extracted from red and crisis labeled documents. The extraction of the discriminative vocabulary was implemented with the approach described in (Charton et al., 2013). The relative frequency of each word is computed for each class. Then, the average difference of word frequencies between the red/crisis classes and the green and amber classes is computed. Each word for which the average difference is above an experimentally set threshold is added to the discriminative vocabulary of a given class. After defining the discriminative vocabularies for the red and the crisis classes, we utilized up to the five best ranked vocabulary terms to build classification rules based on the appearance of these words in a forum post. The rules were applied on top of the predictions made by the supervised classifiers.

model		LMT & rules (5 words)	
class	Precision	Recall	F-measure
crisis	0.22 (15/69)	0.38 (15/39)	0.28
red	0.24 (36/150)	0.33 (36/110)	0.28
amber	0.26 (51/196)	0.20 (51/249)	0.23
accuracy		0.45	
macro-averaged F-score		0.26	
model		SMO & rules (5 words)	
class	Precision	Recall	F-measure
crisis	0.25 (14/56)	0.36 (14/39)	0.29
red	0.24 (33/136)	0.30 (33/110)	0.27
amber	0.25 (51/169)	0.17 (42/249)	0.20
accuracy		0.47	
macro-averaged F-score		0.25	
model		BN	
class	Precision	Recall	F-measure
crisis	0.08 (2/26)	0.05 (2/39)	0.06
red	0.09 (9/98)	0.08 (9/110)	0.09
amber	0.27 (55/205)	0.22 (55/249)	0.24
accuracy		0.44	
macro-averaged F-score		0.13	

Table 3: Results obtained on training set

4 Experiments and Results

We performed a set of experiments to evaluate the usage of different classifiers, feature sets (combining different feature types), as well as the use of CFS, and finally the integration of classification rules to the supervised approach. The system pipeline is implemented as follows:

1. Dataset pre-processing and normalization
2. POS and lemma annotation
3. Feature extraction (POS tags, bigrams, sentiments)
4. CFS filtering of feature sets
5. Generation of documents versus features matrix using selected feature subsets
6. Output of predictions by machine learning based classifiers
7. Re-evaluation of predictions using classification rules

Run 1	model	LMT & rules (5 words)	
class	Precision	Recall	F-measure
crisis	0.00 (0/10)	0.00 (0/1)	0.00
red	0.33 (8/24)	0.30 (8/27)	0.31
amber	0.49 (20/41)	0.43 (20/47)	0.45
accuracy		0.72	
macro-averaged F-score		0.26	
Run 2	model	LMT & rules (3 words)	
class	Precision	Recall	F-measure
crisis	0.00 (0/9)	0.00 (0/1)	0.00
red	0.36 (9/25)	0.33 (9/27)	0.35
amber	0.49 (20/41)	0.43 (20/47)	0.45
accuracy		0.72	
macro-averaged F-score		0.27	
Run 3	model	SMO & rules (5 words)	
class	Precision	Recall	F-measure
crisis	0.00 (0/8)	0.00 (0/1)	0.00
red	0.43 (10/13)	0.37 (10/27)	0.40
amber	0.59 (19/32)	0.40 (19/47)	0.48
accuracy		0.74	
macro-averaged F-score		0.29	
Run 4	model	LMT only	
class	Precision	Recall	F-measure
crisis	0.00 (0/6)	0.00 (0/1)	0.00
red	0.46 (6/13)	0.22 (6/27)	0.30
amber	0.45 (21/47)	0.45 (21/47)	0.45
accuracy		0.75	
macro-averaged F-score		0.25	

Table 4: Results obtained on test set

Our results run	official macro F-m	accuracy	non-green v. green macro F-m	non-green v. green accuracy
run 1	0.26	0.72	0.72	0.83
run 2	0.27	0.72	0.72	0.83
run 3	0.29	0.74	0.68	0.82
run 4	0.25	0.75	0.75	0.85

Table 5: Official results for our system

Overall summary	max	min	median (all runs)	median (team bests)
official score	0.42	0.13	0.32	0.335
accuracy	0.85	0.42	0.77	0.775
non-green v. green macro F-m	0.87	0.58	0.765	0.77
non-green v. green accuracy	0.91	0.60	0.85	0.85

Table 6: Overall summary results for all teams

On the CLPsych training data, the best results were obtained by LMT and SMO algorithms trained on bigrams, sentiment features, and specific POS features. Rule-based classification was applied on the predictions, using a subset of 5 discriminative words from the vocabularies of each red and crisis classes. Table 3 presents the results obtained on the training data while Table 4 shows the results obtained on the test data. We submitted 4 runs using the models that performed best on the training data, namely LMT with and without rules (using 5 or 3 words), and a SMO with rules (5 words). None of our approaches found the unique crisis post present in the test. Posts from the crisis class are indeed the most difficult to find since they are rare, but we also explain this by the difference between crisis ratio in the training set (4.18%) and the test set (0.42%). The system performed consistently on the other classes. Our official results are presented in Table 5, and official results for the 16 teams that participated in the task are provided in Table 6.

5 Conclusion

We presented a system capable of performing automatic triage of forum posts from a mental health online forum. The system assigns to each post a tag that indicates how urgently moderator attention is needed. The evaluation is based on experiments conducted on the CLPsych 2016 task, and the system is available as an open-source software in the following repository: https://github.com/BigMiners/CLPsych2016_Shared_Task

Acknowledgments

The presented system was developed as a Master's student's project. We would like to thank Adrien Appelgren, Antoine Briand, and Sara Zacharie for their contribution.

References

Hayda Almeida, Marie-Jean Meurs, Leila Kosseim, Greg Butler, and Adrian Tsang. 2014. Machine Learning for Biomedical Literature Triage. *PLOS ONE*, 9(12), 12.

Hayda Almeida, Marie-Jean Meurs, Leila Kosseim, and Adrian Tsang. 2015. Supporting HIV Literature Screening with Data Sampling and Supervised Learning. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015*, pages 491–496, Washington, USA, November, 2015. IEEE.

Tanmay Basu and C.A. Murthy. 2012. Effective Text Classification by a Supervised Feature Selection Approach. In *Proceedings of the IEEE 12th International Conference on Data Mining Workshops (ICDMW)*, pages 918–925, Brussels, Belgium, December 2012. IEEE.

Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, and Prasenjit Mitra. 2014. Using Non-lexical Features for Identifying Factual and Opinionative Threads in Online Forums. *Knowledge-Based Systems*, 69:170–178.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: a Common and Common-sense Knowledge Base for Cognition-driven Sentiment Analysis. In *28th AAAI conference on artificial intelligence*, Quebec City, Canada, July 2014.

Eric Charton, Marie-Jean Meurs, Ludovic Jean-Louis, and Michel Gagnon. 2013. Using Collaborative Tagging for Text Classification. *Informatics 2014*, pages 32–51.

Prerna Chikersal, Soujanya Poria, and Erik Cambria. SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, pages 647–651, Denver, Colorado, June 2015.

Mark Hall. 1999. *Correlation-based Feature Selection for Machine Learning*. Ph.D. thesis, The University of Waikato.

Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. 2013. Text Classification for Assisting Moderators in Online Health Communities. *Journal of Biomedical Informatics*, 46(6):998–1005.

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic Model Trees. *Machine Learning*, 59(1-2):161–205.

Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao. 2010. Feature Selection: An Ever Evolving Frontier in Data Mining. In *Proceedings of the 4th Workshop on Feature Selection in Data Mining*, pages 4–13, Hyderabad, India, June 2010.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1275–1284, New York, NY, USA. ACM.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

John Platt. 1998. Sequential Minimal Optimization: a Fast Algorithm for Training Support Vector Machines. *Microsoft Research Technical Report MSR-TR-98-14*, April 1998.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In *Proceedings of 9th of the Language Resources and Evaluation Conference (LREC)*, pages 810–817, Reykjavik, Iceland, May 2014.

Shirin Saleem, Rohit Prasad, Shiv Naga Prasad Vitaladevuni, Maciej Pacula, Michael Crystal, Brian Marx, Denise Sloan, Jennifer Vasterling, and Theodore Speroff. 2012. Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts. In *COLING*, pages 2375–2388.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Suppawong Tuarob, Conrad S Tucker, Marcel Salathe, and Nilam Ram. 2014. An Ensemble Heterogeneous Classification Methodology for Discovering Health-related Knowledge in Social Media Messages. *Journal of Biomedical Informatics*, 49:255–268.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Yan Xu, Kai Hong, Junichi Tsujii, I Eric, and Chao Chang. 2012. Feature Engineering Combined with Machine Learning and Rule-based Methods for Structured Information Extraction from Narrative Clinical Discharge Summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832.