Predicting Post Severity in Mental Health Forums

Shervin Malmasi[1,2] **Marcos Zampieri**[3,4] **Mark Dras**[1]
[1]Macquarie University, Sydney, NSW, Australia
[2]Harvard Medical School, Boston, MA 02115, USA
[3]Saarland University, Germany
[4]German Research Center for Artificial Intelligence, Germany
{first.last}@mq.edu.au, marcos.zampieri@dfki.de

Abstract

We present our approach to predicting the severity of user posts in a mental health forum. This system was developed to compete in the 2016 Computational Linguistics and Clinical Psychology (CLPsych) Shared Task. Our entry employs a meta-classifier which uses a set of of base classifiers constructed from lexical, syntactic and metadata features. These classifiers were generated for both the target posts as well as their contexts, which included both preceding and subsequent posts. The output from these classifiers was used to train a meta-classifier, which outperformed all individual classifiers as well as an ensemble classifier. This meta-classifier was then extended to a Random Forest of meta-classifiers, yielding further improvements in classification accuracy. We achieved competitive results, ranking first among a total of 60 submitted entries in the competition.

1 Introduction

Computational methods have been widely used to extract and/or predict a number of phenomena in text documents. It has been shown that algorithms are able to learn a wide range of information about the authors of texts as well. This includes, for example, the author's native language (Gebre et al., 2013; Malmasi and Dras, 2015a), age and gender (Nguyen et al., 2013), and even economic conditions such as income (Preoţiuc-Pietro et al., 2015). These tasks are often considered to be a part of a broader natural language processing task known as authorship profiling (Rangel et al., 2013).

More recently, such approaches have been applied to investigating psychological factors associated with the author of a text. For practical purposes most of the applications that deal with clinical psychology use social media data such as *Twitter, Facebook*, and online forums (Coppersmith et al., 2014). Examples of health and psychological conditions studied using texts and social media are: suicide risk (Thompson et al., 2014), depression (Schwartz et al., 2014), autism (Tanaka et al., 2014; Rouhizadeh et al., 2015), and schizophrenia (Mitchell et al., 2015).

In this paper we propose an approach to predict the severity of posts in a mental health online forum. Posts were classified into for levels of severity (or urgency) represented by the labels *green, amber, red*, and *crisis* according to indication of risky or harmful behavior by users (e.g. self-harm, suicide, etc.). This kind of classification task serves to provide automatic triage of user posts in order to help moderators of forums and related online communities to respond to urgent posts. Our approach competed in the CLPsych 2016 shared task and achieved the highest accuracy among submitted systems.

2 Task and Data

The dataset of the CLPsych shared task was compiled from the *ReachOut.com*[1] forums. *ReachOut.com* is an online youth mental health service that provides information, tools and support to young people aged 14-25.

The corpus consists of a total 65,024 posts formatted in XML and including metadata (e.g. time stamp, thread, post id, user id, etc.). Each post in

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

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 133–137,
San Diego, California, June 16, 2016.

the labeled sets was manually annotated with a label representing how urgent a post should be handled by one of the *ReachOut.com* moderators.

Data Sets	Posts
Labeled Train	977
Labeled Test	250
Unlabeled	63,797
Total	65,024

Table 1: CLPsych Corpus Divided by Data Set

According to the shared task organizers, these labels were attributed according to the following criteria:

- Green: a moderator does not need to prioritize addressing this post.
- Amber: a moderator needs to look at this and assess if there are enough responses and support from others or if they should reply.
- Red: a moderator needs to look at this as soon as possible and take action.
- Crisis: the author (or someone they know) might hurt themselves or others (a red instance that is of urgent importance).

Participating systems should be trained to predict these labels, with evaluation on the test set.

3 Feature Extraction

We used three categories of features: lexical, syntactic, and metadata features. These features and our preprocessing method are outlined here.

3.1 Preprocessing

The following preprocessing was performed on the texts: HTML removal was performed, with links and anchor text being preserved. Smileys and emoticons were converted to text tags, e.g. #SmileySad and #SmileyHappy. Quotes from previous posts (e.g. the one being replied to) were also removed so as not to mix features from distinct messages.[2]

[2]This was facilitated by the fact that such quotations were labeled as such using the HTML blockquote tag.

3.2 Lexical Features

We represent words in the texts using different features based on characters, word forms and lemmas. We summarize the lexical features used in our system as follows:

- **Character n-grams:** we extracted n-grams of order 2–8.
- **Word n-grams:** words were represented as 1–3 grams.
- **Word skip-grams:** To capture the longer distance dependencies not covered by word n-grams we also used word skip-grams as described in Guthrie et al. (2006). We extract 1, 2 and 3-skip word bigrams.
- **Lemma n-grams:** we used a lemmatized version of the texts and extract lemma n-grams of order 1–3.
- **Word Representations:** To increase the generalizability of our models we used word representation features based on Brown clustering as a form of semi-supervised learning. This was done using the method described by Malmasi et al. (2015a). We used the clusters generated by Owoputi et al. (2013). They collected From 56 million English tweets (837 million tokens) and used it to generate 1,000 hierarchical clusters over 217 thousand words.

3.3 Syntactic Features

We used a set of (morpho-)syntactic features for deeper linguistic analysis, using the Stanford CoreNLP system for extracting these. The intuition is that structural or syntactic patterns present in posts might reveal relevant information regarding the psychological condition of writers.

- **Part-of-Speech (POS) n-grams:** these features rely on POS annotation and they are used to represent morphosyntactic patterns. We use POS tags modeled as 1–3 grams.
- **Dependencies:** we use dependency relations between constituents of sentences as features. They provide good indication of syntactic patterns in the data.

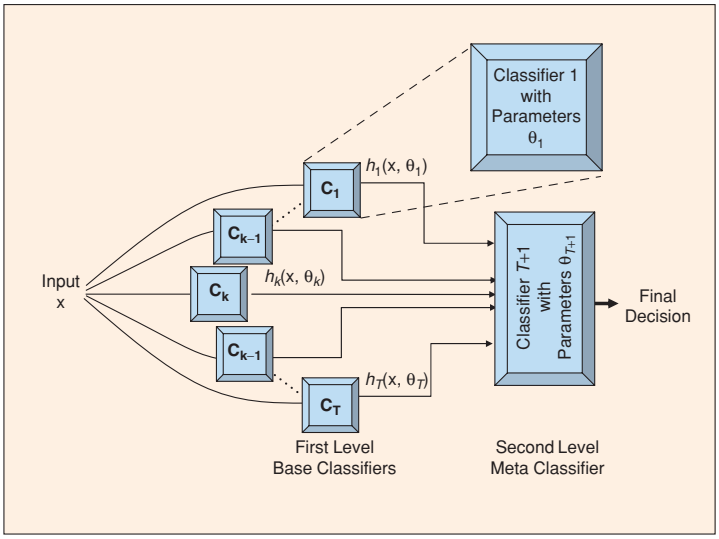

Figure 1: An illustration of a meta-classifier architecture. Image reproduced from Polikar (2006).

- **Production Rules:** similar to dependency relations, production rules capture the overall structure of grammatical constructions.

3.4 Metadata and Other Features

Finally, the third type of features used in our system relies on metadata. We used two feature groups taking advantage of the information present in the corpus about the forum itself and the user.

- **Board ID:** The forum is divided into individual boards according to topic. The ID of the board to which a post belongs is used as a feature.
- **User details:** The user information of a post's author, including the number of posts and affiliation status were used as features. This helps with the correct classification of messages from moderators and veteran users.
- **Subject:** The subjects of the postings were too short and unvaried for training a classifier. Instead, we applied the LIWC lexicon (Pennebaker et al., 2015) as a proxy measure of the subject's sentiment. These lexicon features were used to train a classifier.

3.5 Feature Contexts

Our features were extracted from several contexts, including the post itself in isolation, the last 1-2 recent posts by the author, the last 2-5 recent posts in the thread and the next 1-2 posts by the author (where available).

4 Methodology and Systems

We employed a meta-classifier for our entry, also referred to as classifier stacking. A meta-classifier architecture is generally composed of an ensemble of base classifiers that each make predictions for all of the input data. Their individual predictions, along with the gold labels are used to train a second-level meta-classifier that learns to predict the label for an input, given the decisions of the individual classifiers. This setup is illustrated in Figure 1. This meta-classifier attempts to learn from the collective knowledge represented by the ensemble of local classifiers. The first step in such an architecture is to create the set of base classifiers that form the first layer. We describe this process below.

4.1 Ensemble Construction

Our ensemble was created using linear Support Vector Machine classifiers.[3] We used the features listed in Section 3 to create our ensemble of classifiers. A single classifier was trained for each feature type and context, resulting in an ensemble of over 100 classifiers. Each classifier predicts every input and assigns a continuous output to each of the possible labels.

[3]Linear SVMs have proven effective in many text classification tasks (Malmasi and Dras, 2014; Malmasi et al., 2015b; Malmasi and Dras, 2015b).

Run	Official Score	Accuracy	F-score (NG vs. G)	Accuracy (NG vs. G)	Rank
Run 1	0.37	0.80	0.83	0.89	11th
Run 2	0.38	0.80	0.83	0.89	9th
Run 3	**0.42**	0.83	**0.87**	**0.91**	1st
Run 4	**0.42**	**0.84**	**0.87**	**0.91**	1st
Run 5	0.40	0.82	0.85	0.90	6th

Table 2: Official CLPsych scores. Best results in bold. Rankings are out of the 60 systems submitted.

Classifiers ensembles have proven to be an efficient and robust alternative in other text classification tasks such as language identification (Malmasi and Dras, 2015a), grammatical error detection (Xiang et al., 2015), and complex word identification (Malmasi et al., 2016).

4.2 Meta-classifier

For our meta-classifier, We experimented with three algorithms: Random Forests of decision trees, a linear SVM just like our base classifiers and a Radial Basis Function (RBF) kernel SVM. The inputs to the meta-classifier are the continuous outputs from each base SVM classifier in our ensemble, along with the original gold label. For the Random Forest classifiers, the final label is selected through a plurality voting process across all decision trees in the forest.

All were found to perform well, but the linear SVM was was outperformed by its RBF-kernel counterpart. This could be because the RBF-kernel SVM is more suitable for data with a smaller number of features such as here and can provide non-linear decision boundaries. Accordingly, we did not use the linear SVM for our entry due to the 5 run limit.

4.3 Systems

Using the methods described so far, we created five different systems for the CLPsych shared task:

- **System 1:** Our first system used the RBF-kernel SVM meta-classifier.
- **Systems 2–5:** The other four systems were based on Random Forests. This is because we noted some performance variation between different Random Forest classifiers, likely due to the randomness inherent to the algorithm.

5 Results

Submissions were evaluated on the unlabeled test set. The official evaluation metric is the F-score over all non-green labels. The results obtained by our 5 systems are shown in in Table 2. We report the official score by the organizers and the ranking among all submitted systems. According to the the organizers a total of 60 runs were submitted.

The meta-classifier approach proved to be robust and appropriate for this task. We observed that all five runs submitted were ranked in the top half of the table (four of them in the top 10). Systems 3 and 4 were ranked first according to the official score, achieving 84% accuracy for all four classes and 91% accuracy in discriminating between *green* and *non-green* posts.

The Random Forest meta-classifiers all outperformed their SVM counterpart. The differences in results among the four different Random Forest classifiers highlights the randomness that is inherent to their training.

6 Conclusion and Future Work

We presented an approach to predict severity of posts in a mental health forum. We proposed the use of a meta-classifier and three types of features based on words, syntax, and metadata presented in Section 3. We submitted five runs to the CLPsych shared task and all of them were ranked in the top half of the table. Our best system achieved 84% accuracy for all four classes and 91% accuracy in discriminating between *green* and *non-green* posts. Our approach was ranked first in the shared task.

Acknowledgments

The authors would like to thank the organizers for proposing this interesting shared task.

References

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the BEA Workshop*.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of LREC*.

Shervin Malmasi and Mark Dras. 2014. Chinese Native Language Identification. In *Proceedings of EACL*.

Shervin Malmasi and Mark Dras. 2015a. Language identification using classifier ensembles. In *Proceedings of the LT4VarDial Workshop*.

Shervin Malmasi and Mark Dras. 2015b. Multilingual Native Language Identification. In *Natural Language Engineering*.

Shervin Malmasi, Hamed Hassanzadeh, and Mark Dras. 2015a. Clinical Information Extraction using Word Representations. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 66–74, Sydney, Australia, 12.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015b. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *PACLING 2015*, pages 209–217.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of SemEval*.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*.

Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, and Theo Meder. 2013. "how old do you think i am?" a study of language and age in twitter. In *Proceedings of ICWSM*.

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. In *Proceedings of NAACL*.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*.

Robi Polikar. 2006. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9).

Francisco Rangel, Efstathios Stamatatos, Moshe Moshe Koppel, Giacomo Inches, and Paolo Rosso. 2013. Overview of the author profiling task at PAN 2013. In *Proceedings of CLEF*.

Masoud Rouhizadeh, Richard Sproat, and Jan van Santen. 2015. Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*.

Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*.

Paul Thompson, Craig Bryan, and Chris Poulin. 2014. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*.

Yang Xiang, Xiaolong Wang, Wenying Han, and Qinghua Hong. 2015. Chinese grammatical error diagnosis using ensemble learning. In *Proceedings of the NLP-TEA Workshop*.