The GW/UMD CLPsych 2016 Shared Task System

Ayah Zirikly
George Washington University
Washington, DC
ayaz@gwu.edu

Varun Kumar
University of Maryland
College Park, MD
varunk@cs.umd.edu

Philip Resnik
University of Maryland
College Park, MD
resnik@umd.edu

1 Introduction

Suicide is the third leading cause for death for young people, and in an average U.S. high school classroom, 30% have experienced a long period of feeling hopeless, 20% have been bullied, 16.7% have seriously considered suicide, and 6.7% of students have actually made a suicide attempt.[1] The 2016 ACL Workshop on Computational Linguistics and Clinical Psychology (CLPsych) included a shared task focusing on classification of posts to ReachOut, an online information and support service that provides help to teens and young adults (aged 15-24) who are struggling with mental health issues.[2] The primary goal of the shared task is to identify posts that require urgent attention and review from the ReachOut team (i.e. moderators).

2 System Overview

We use Stanford CoreNLP (Manning et al., 2014) for preprocessing (tokenization, lemmatization, POS tagging) and a supervised learning approach for classification. Section 2.1 describes the features we use, and Section 2.2 describes our classifiers.

2.1 Features

The features used in our model range from simple unigrams to more complex features such as syntactic, sentiment, psychological, and other data-driven features.

- Unigram features: We choose the *n* most important unigrams based on their TF-IDF values, restricting attention to unigrams appearing in between 2 and 60% of documents.
- Part-of-speech features: We use part-of-speech (POS) tag counts for adverbs, pronouns, and modal auxiliaries (e.g. can, cannot, couldn't, might).
- Sentiment features: For every post we generate three sentiment features, calculated as follows: i) split the post into sentences; ii) tag each sentence as one of {positive, negative, neutral} using Stanford CoreNLP; iii) as three document-level features, include the number of sentences that are tagged as *negative*, *positive*, and *neutral*.
- ReachOut meta-data features: From the meta-data of the posts, we use: number of views, time of day of the post, and the board on which the post appeared. The *time* feature is bucketed into eight categories, where each category represents a three hour window. (This feature is based on observations in the literature showing that depressed users tend to be more active on social media at night (Choudhury et al., 2013).) The *board* is represented as six binary features, one each for *Everyday_life_stuff*, *Feedback_Suggestion*, *Getting_Help*, *Intros*, *Something_Not_Right*, and *mancave*. For any post in the test set where the board is not among these, the six board features is set to zero.
- Emotion features: We use the count of emotion words occurring in the post, based on the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013). The emotions included are anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. To expand the number of matches, we do lookups in the NRC for words, tokens, and lemmas and use the maximum value.

[1] http://us.reachout.com/about-us/what-we-do/; see also Centers for Disease Control and Prevention 2015

[2] http://us.reachout.com

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 166–170,
San Diego, California, June 16, 2016.

- Linguistic Inquiry and Word Count (LIWC): We include the category for each LIWC category (Tausczik and Pennebaker, 2010) using the post's lemmas.

- Positive/negative counts: In non-green posts, some users list "positives" and "negatives" associated with the issue or situation the user is facing. For example, a user might say *Negative: Everything is going wrong in my life, I feel so depressed and worthless. Positive: I need to appreciate small things and be grateful to what I have.* We include the total number of such positive or negative lists as a single feature whose value is the frequency of any of the following tokens: *(positive:, negative:, pos:, neg:)*. In the above example the value of the feature would be 2.

- Mention features: As the *mention* feature, we use the count of explicit user mentions (identified using @) within the post.

- The *word_count* feature is the number of words in the post.

In work after the the shared task was completed, we also experimented with additional features that were not part of our official submission.[3]

- ReachOut author: This binary feature is enabled when the user is ReachOut-affiliated (e.g. moderator, staff). This feature is a cue that the post is *green* (no further follow-up is needed).

- Mental Disease Lexicon *mentalDisLex*: This feature is a count tokens in the post that match entries in a mental disease lexicon.[4]

- Word shape: We include two binary features that reflect the occurrence of words that either have character repetitions like "hmmm" or all capitalized letters like "DIE".

- Word embeddings: We use word2vec to generate word embeddings as described in (Mikolov et al., 2013).[5] The post's document-level embedding is calculated as the average of all the words' vectors.

[3]For the rest of the document, when we mention features, we mean the above features that were used in the official runs, unless otherwise stated.

[4]http://mental-health-matters.com/psychological-disorders/alphabetical-list-of-disorders

[5]http://word2vec.googlecode.com/svn/trunk/

2.2 Framework

We experimented with a diverse set of multi-class balanced supervised classifiers.

2.2.1 Lexically based classifier

In this setup we used both the SVM *(uniSVM)* and logistic regression *(uniLR)* classifiers. We use unigrams as binary features. We pick the top n unigrams based on their TF-IDF weighting scores and combine them with the other features.

2.2.2 Non-lexical classifier

In this setup *(nonLexLR)*, we incorporate all features (Section 2.1) except the unigram features and classify using the logistic regression classifier.

2.2.3 Two-stage classifier

This setup *(2stage)* is based on an ensemble supervised learning approach as depicted in Figure 1. The first stage is a support vector machine clasifier (Cortes and Vapnik, 1995) using lexical features with TF-IDF weighting. The second stage is a logistic regression classifier which uses the output probabilities of the SVM classifier, along with the features described in Section 2.1.

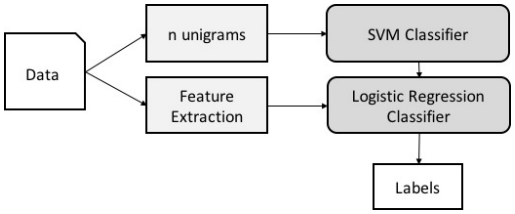

Figure 1: Two-stage classifier

Ensemble methods have proven to be more effective than individual classifiers when the training data is significantly small (as shown in Table 1) and not a good representative of the classes (Polikar, 2006).

2.2.4 Majority vote classifier:

In this setting *(maj1)*, we use the majority vote based on the *uniSVM*, *uniLR*, and *nonLexLR* classifiers.

3 Experiments

3.1 Dataset

The shared task dataset contains posts annotated with four classes (green, amber, red, and crisis), and the main goal is to correctly classify the posts

that belong to the last three classes. Table 1 shows the number of posts per class.

Subset	green	amber	red	crisis	total
Train	549	249	110	39	947
Test	166	47	27	1	241

Table 1: Dataset Train-Test Stats

3.2 Metrics

For evaluation, we used the script provided by the shared task organizers, which does not include the green labels.[6] The evaluation metrics are precision, recall, and F1-score for each of the three classes (amber, red, crisis), in addition to the macro F1 (official score).

3.3 Results & Discussion

During the system building phase, we experimented with the models in Section 2.2 using 5-fold cross validation (CV) on the training data, making use of all the features mentioned in Section 2.1 except word shape, author ranking, mental disease lexicon, and word embedding features. For the *uniLR, uniSVM*, and *2stage* classifiers, we empirically choose $n = 300$ as the number of most-important unigrams based on best results of the 5-fold CV.

Table 2 depicts the models' performance on the test data. Although they were not included in the official submissions, Table 4 also includes the extra features we explored.

Model	Test data
uniLR	0.32
uniSVM	0.34
nonLexLR	0.34
2stage	**0.36**
maj1	0.32

Table 2: Macro F1-Scores on Test Data

Two key challenges in this shared task turned out to be the highly imbalanced data and the extremely small number of *crisis* and *red* posts, with just 39 crisis posts in the training data and one (!) crisis post in the test set. We addressed the imbalanced dataset problem by using multi-class balanced classifiers, and using five-fold cross validation on training data (941 posts) helped to avoid design choices based on a particularly lucky or unlucky training/test split (Khoshgoftaar et al., 2007). However, in order to tackle the second issue, we need a feature set that is capable of capturing red and crisis posts, which are the most important classes since they require immediate action from ReachOut's moderators and/or administrators.

From Table 4, we observe that the mental disease lexicon feature set was the one capable of capturing the single instance of *crisis* in the test data; additionally, it improved the recall of *red* and precision of *amber*. This results in our best system performance, an unofficial post-shared-task macro-F1 score of 0.45, which improves on the best shared-task official score of 0.42. The LIWC features also provide a major boost in performance (on both CV and test data) which aligns with the results in Table 2; there a feature set that does not include any lexical features (0.34) performs equally to a single classifier using a combination of lexical and non-lexical features.

4 Conclusion & Future Work

We have presented a collaborative effort between George Washington University (GW) and University of Maryland (UMD) to tackle the CLPsych 2016 ReachOut shared task. Using a 2-stage ensemble classification approach, our best official submission yielded 0.36% macro-F1, which is 6% short of the best system. However, further feature experimentation after the conclusion of the shared task yielded a macro F1 score of 0.45%. In future work, we plan to experiment with an extended ReachOut meta-data feature set and to expand LIWC features using word embeddings.

[6] https://github.com/clpsych-2016-shared-task/ro-evaluation

Features	F1 CV	macro-F1	Accuracy	crisis			red			amber		
				Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
U=unigrams	0.12	0.33	0.75	0	0	0	0.5	0.41	0.45	0.48	0.64	0.55
U+POS	0.11	0.29	0.73	0	0	0	0.42	0.3	0.35	0.46	0.62	0.53
X=U+sentiment	0.12	0.33	0.75	0	0	0	0.5	0.41	0.45	0.48	0.62	0.54
+LIWC	0.36	0.36	0.75	0	0	0	0.59	0.48	0.53	0.49	0.62	0.55
+emotion	0.33	0.35	0.74	0	0	0	0.57	0.44	0.5	0.48	0.62	0.54
+meta_data	0.35	0.36	0.77	0	0	0	0.59	0.48	0.53	0.51	0.60	0.55
+positive/negative counts	0.34	0.36	0.76	0	0	0	0.57	0.48	0.52	0.51	0.62	0.56
Y=..+mention	0.36	***0.36***	0.77	0	0	0	0.56	0.52	0.54	0.51	0.60	0.55
Z=..+word_count	0.356	0.36	0.76	0	0	0	0.54	0.48	0.51	0.52	0.62	0.56
Y+wordShape	0.356	0.36	0.77	0	0	0	0.59	0.48	0.53	0.5	0.57	0.53
Y+authorRanking	0.36	0.36	0.78	0	0	0	0.57	0.48	0.52	0.52	0.60	0.55
Y+mentalDisLex	0.364	0.44	0.78	0.12	1	0.22	0.57	0.48	0.52	0.54	0.6	0.57
Y+word2vec	0.356	0.36	0.78	0	0	0	0.59	0.48	0.53	0.52	0.62	0.56
Z‘=Y+mentalDisLex+authorRanking	**0.37**	**0.45**	0.78	0.12	1	0.22	0.58	0.52	0.55	0.56	0.60	0.58
Z‘+wordShape+word2vec	**0.37**	**0.43**	0.78	0.11	1	0.20	0.57	0.48	0.52	0.54	0.6	0.57
All	0.35	0.35	0.77	0	0	0	0.52	0.48	0.50	0.51	0.57	0.54

Table 3: Features’ impacts on system performance, using 5-fold cross-validation and evaluation on test data

References

Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. 2013. Predicting depression via social media. AAAI, July.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September.

T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse. 2007. An empirical study of learning from imbalanced data using random forest. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 310–317, Oct.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

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

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods.