Mental Distress Detection and Triage in Forum Posts: The LT3 CLPsych 2016 Shared Task System

Bart Desmet and **Gilles Jacobs** and **Véronique Hoste**
LT3, Language and Translation Technology Team
Department of Translation, Interpretation and Communication
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
bart.desmet@ugent.be, gillesm.jacobs@ugent.be, veronique.hoste@ugent.be

Abstract

This paper describes the contribution of LT3 for the CLPsych 2016 Shared Task on automatic triage of mental health forum posts. Our systems use multiclass Support Vector Machines (SVM), cascaded binary SVMs and ensembles with a rich feature set. The best systems obtain macro-averaged F-scores of 40% on the full task and 80% on the green versus alarming distinction. Multiclass SVMs with all features score best in terms of F-score, whereas feature filtering with bi-normal separation and classifier ensembling are found to improve recall of alarming posts.

1 Introduction

The 2016 ACL Workshop on Computational Linguistics and Clinical Psychology included a shared task focusing on triage classification in forum posts from ReachOut.com, an online service for youth mental health issues. The aim is to automatically classify an unseen post as one of four categories indicating the severity of mental distress. ReachOut staff has annotated a corpus of posts with *crisis*/*red*/*amber*/*green* semaphore labels that indicate how urgently a post needs moderator attention.

The system described in this paper is based on a suicidality classification system intended for Dutch social media (Desmet and Hoste, 2014). Therefore, we approach the current mental distress triage task from a suicide detection standpoint.

2 Related Work

Machine learning and natural language processing have already shown potential in modelling and detecting suicidality in the arts (Stirman and Pennebaker, 2001; Mulholland and Quinn, 2013) and in electronic health records (Haerian et al., 2012). However, work on computational approaches to the automatic detection of suicidal content in online user-generated media is scarce.

One line of research focuses on detecting suicidality in individuals relying on their post history: Huang et al. (2007) aim to identify Myspace.com bloggers at risk of suicide by means of a keyword-based approach using a manually collected dictionary of weighted suicide-related terms. Users were ranked by pattern-matching keywords on their posts. This approach suffered from low precision (35%) and the data does not allow to measure recall, i.e. the number of actually suicidal bloggers that are missing from the results. Similarly, Jashinsky et al. (2014) manually selected keywords by testing search queries linked to various risk factors in a user's Twitter profile. In order to validate this search approach, users posting tweets that match the suicide keywords were grouped by US state for trend analysis. The proportion of at-risk tweeters vs. control-group tweeters were strongly correlated with the actual state suicide rates. While this methodology yields a correct proportion of at-risk users, it is unclear how many of those tweets are false positives and how many at-risk tweets are missing.

Going beyond a keyword-based approach, Guan et al. (2015) performed linear regression and random forest machine learning for Chinese Weibo.com microbloggers. Suicidality labels were assigned to users in the data set by means of an online psychological evaluation survey. As classification features

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 148–152,
San Diego, California, June 16, 2016.

they took social media profile metadata and psychometric linguistic categories in a user's post history. Results showed that Linear Regression and Random Forest classifiers obtain similar scores with a maximum of 35% F-score (23% precision and 79% recall) being the highest performance.

As in the CLPsych 2016 Shared Task, another line of research aims to classify suicidality on the post level, rather than the level of user profiles. Desmet and Hoste (2014) proposed a detection approach using machine learning with a rich feature set on posts in the Dutch social media platform Netlog. Their corpus was manually annotated by suicide intervention experts for suicide relevance, risk and protective factors, source origin, subject of content, and severity. Two binary classification tasks were formulated: a relevance task which aimed to detect posts relevant to suicide, and a threat detection task to detect messages that indicate a severe suicide risk. For the threat detection task, a cascaded setup which first filters irrelevant messages with SVM and then predicts the severity with k-Nearest Neighbors (KNN) performed best: 59.2% F-score (69.5% precision and 51.6% recall). In general, both KNN and SVM outperform Naive Bayes and SVM was more robust to the inclusion of bad features. The system presented in this paper is for the most part an extension and English adaptation of this suicidal post detection pipeline.

3 System Overview

We investigated a supervised classification-based approach to the mental distress triage task using SVMs. Below, we describe the data and features that were used, and the way classifiers were built, optimized and combined.

3.1 Data

Labeled data sets: 1/8th of the manually annotated training data was sampled as a held-out development set ($n = 118$ with at least 4 instances of each class), the remainder ($n = 829$) was used for training. In the results section, we also report on the held-out test set ($n = 241$).

Reddit background corpus: In order to perform terminology extraction and topic modelling, we collected domain-relevant text from Reddit.com, a predominantly English social news and bulletin board website. We used the title and body text from all opening posts in mental health and suicide-related boards posted between 2006 and 2014, resulting in a 82.7 million token corpus of over $270,000$ posts. The selected boards mainly contain user-generated discussion on mental health, depression, and suicidal thoughts, similar to the ReachOut forums.

Tokenization and preprocessing: All textual data was tokenized and lower-cased to reduce variation. For topic modelling, emoji and punctuation were removed. Pattern (De Smedt and Daelemans, 2012) was used for lemmatization.

3.2 Features

We aimed to develop a rich feature set that focused on lexical and semantic information, with fine-grained and more abstract representations of content. Some syntactic and non-linguistic features were also included.

Bag-of-words features: We included binary token unigrams, bigrams and trigrams, along with character trigrams and fourgrams. The latter provide robustness to the spelling variation typically found in social media.

Term lists: Domain-specific multiword terms were derived from the Reddit background corpus, using the TExSIS terminology extraction tool (Macken et al., 2013). One list was based on suicide-specific boards (*/r/SuicideWatch* and */r/suicidenotes*, 2884 terms), the other included terms only found in other mental health boards (1384 terms).

Lexicon features: We computed positive and negative opinion word ratio and overall post sentiment using both the MPQA (Wilson et al., 2005) and Hu and Liu's (2004) opinion lexicons. We added positive, negative and neutral emoji counts based on the BOUNCE emoji sentiment lexicon (Kökciyan et al., 2013). We also included the relative frequency of all 64 psychometric categories in the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2007). LIWC features have proven useful in (Stirman and Pennebaker, 2001) for modelling suicidality in literary works. Furthermore, we included diminisher, intensifier, negation, and "allness" lexica because of their significance in suicide notes analysis (Osgood and Walker, 1959; Gottschalk and Gleser, 1960; Shapero, 2011).

Topic models: Using the gensim topic modelling library (Řehůřek and Sojka, 2010) we trained several LDA (Blei et al., 2003) and LSI (Deerwester et al., 1990) topic models with varying granularity ($k = 20, 50, 100, 200$). A similarity query was done on each model resulting in two feature groups: k topic similarity scores and the average similarity score. This should allow the classifier to learn which latent topics are relevant for the task, and to what extent the topics align with the ones in the Reddit background corpus. In line with Resnik et al. (2015), we used topic models to capture latent semantic and syntactic structure in the mental health domain. However, we did not include supervised topic models.

Syntactic features: Two binary features were implemented indicating whether the imperative mood was used in a post and whether person alternation occurred (i.e. combinations of first and second person pronouns).

Post metadata: We furthermore included several non-linguistic features based on a post's metadata: the time of day a post was made (expressed in three-hour blocks), the board in which it was posted, whether the post includes a subject line or a URL, the role of the author and whether he or she is a moderator, whether the post is the first in a thread, whether there are (moderator) reactions or kudos (i.e. thumbs-up votes).

When applied to the training data, this resulted in 59 feature groups and $107,852$ individual features, the majority of which were bag-of-words features (almost 96%).

3.3 Classifiers

Using SVMs, we tested three different approaches to the problem of correctly assigning the four triage labels to the forum posts. We considered detection of posts with a high level of alarm (*crisis* or *red*) to be the priority. Where possible, recall of the priority labels was promoted, since false negatives are most problematic there.

With **multiclass SVMs**, one model is used to predict all four labels at once. We hypothesized that distinguishing green from non-green posts would require different information than detecting the more alarming categories. We therefore also tested cascades of three **binary SVMs**, in which each classifier predicts a higher level of alarm: green vs. rest; red or crisis vs. rest; and crisis vs. rest. The binary results are combined in a way that the label with the highest level of alarm is assigned. This essentially sacrifices some precision on lower-priority classes for better high-priority recall.

Finally, we tested **ensembles** of various multiclass and binary systems. Predictions were combined with two voting methods: normal majority voting (reported as *ensemble-majority*), and crisis-priority voting (*ensemble-priority*) where the most alarming label with at least 2 votes is selected.

3.4 Optimization

Typically, the performance of a machine learning algorithm is not optimal when it is used with all implemented features and with the default algorithm settings. SVMs are known to perform well in the presence of irrelevant features, but dimensionality reduction can still be beneficial for classification accuracy and resource usage. In this section, we describe the methods we tested for feature selection and hyperparameter optimization.

With **feature filtering**, a metric is used to determine the informativeness of each feature, given the training data. Yang (1997) found that Information Gain (IG) allows aggressive feature removal with minimal loss in accuracy. Forman (2003) corroborates this finding, but remarks that IG is biased towards the majority class, unlike the Bi-Normal Separation (BNS) metric, which typically achieves better minority class recall. In the results, we compare both filtering methods (*-ig* and *-bns*) to no filtering (*-nf*). IG was applied with a threshold of 0.005 (92-97% reduction), BNS with threshold 3 (79-93% reduction for binary tasks, no multiclass support).

We also applied **wrapped optimization**, where combinations of selected feature groups and hyperparameters are evaluated with SVM using three-fold crossvalidation. Exhaustive exploration of all combinations was not possible, so we used genetic algorithms to approximate an optimal solution (Desmet et al., 2013). In the results section, all reported systems have been optimized for feature group and hyperparameter selection, except for *multiclass-unopt* (baseline without filtering or optimization) and *multiclass-hyper* (only hyperparameter optimization, no feature filtering or selection).

4 Results and discussion

In Table 4, we report the four-label classification results of all systems. Most systems perform well in comparison to the shared task top score of 42% macro-averaged F-score, with the *multiclass-nf* submission scoring highest at 40%. This indicates that the implemented features and approach are within the current state of the art.

	dev		test	
system	**F**	**acc**	**F**	**acc**
multiclass-unopt	0.00	0.64	0.00	0.69
multiclass-hyper	0.36	0.75	0.41	0.80
multiclass-nf *	0.50	0.75	0.40	0.80
multiclass-ig	0.36	0.74	0.35	0.78
binary-nf *	0.39	0.69	0.36	0.74
binary-ig	0.36	0.75	0.32	0.77
binary-bns *	0.38	0.64	0.19	0.54
ensemble-majority *	0.54	0.79	0.35	0.77
ensemble-priority *	0.51	0.75	0.37	0.78

Table 1: Results for four-label classification (F = macro-averaged F-score, acc = accuracy). The 5 systems submitted for the shared task are indicated with an asterisk.

Arguably, macro-averaged F-score is a harsh metric for this task: it treats the three alarming categories as disjunct, although confusion between those classes can be high and the distinction may not matter much from a usability perspective. Since the test set only contained one *crisis* instance, failing to detect it effectively limits the ceiling for macro-averaged F-score to 67%. This partly explains the low scores in Table 4. For comparison, we list F-score, precision and recall for the *green* vs. *alarming* distinction in Table 4. Alarming posts can be detected with $F = 80\%$ and recall up to 89% (*ensemble-priority*).

	dev				test			
system	**F**	**P**	**R**	**acc**	**F**	**P**	**R**	**acc**
multicl-unopt	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.69
multicl-hyper	0.76	0.78	0.74	0.83	0.78	0.77	0.79	0.86
multicl-nf	0.75	0.76	0.74	0.82	0.80	0.77	0.84	0.87
multicl-ig	0.78	0.77	0.79	0.84	0.76	0.75	0.77	0.85
binary-nf	0.72	0.70	0.74	0.79	0.79	0.72	0.88	0.85
binary-ig	0.81	0.78	0.84	0.86	0.75	0.75	0.76	0.85
binary-bns	0.73	0.62	0.88	0.76	0.63	0.50	0.87	0.68
ensemble-maj	0.82	0.79	0.86	0.86	0.77	0.73	0.81	0.85
ensemble-prior	0.77	0.67	0.91	0.81	0.80	0.73	0.89	0.86

Table 2: Results for binary classification: *green* vs. all other classes (F = F-score, P = precision, R = recall, acc = accuracy)

We tested three classifier configurations, and find that a multiclass approach performs as well as or better than more complex systems. On the development data, ensemble systems perform best, although this is not confirmed by the four-label test results, possibly due to paucity of *crisis* instances. It appears that ensembles are a sensible choice especially if recall is important. This may be due to the inclusion of the high-recall *binary-bns* cascade, the low precision of which is offset by ensemble voting. Overall, the aim of improving recall with cascaded and ensemble classifiers seems to have been effective: compared to multiclass systems, they all favour recall over precision more, both on development and test data.

The unoptimized *multiclass-unopt* acts as a majority baseline that always predicts *green*, indicating that hyperparameter optimization is essential. Feature selection, on the other hand, does not yield such a clear benefit. On the held-out test data, the *nf* systems consistently outperform their *ig* and *bns* counterparts in terms of F-score. On the development data, feature filtering has a positive effect on recall, particularly when BNS is applied. In summary, the applied feature selection techniques are sometimes successful in removing the bulk of the features without harming performance, although the results suggest that they may remove too many or cause overfitting.

5 Conclusion

This paper discussed an SVM-based approach to the CLPsych 2016 shared task. We found that our systems performed well within the state of the art, with macro-averaged F-scores of 40% on the full task, and 80% for the distinction between green and alarming posts, suggesting that confusion between the three alarming classes is high. Multiclass systems performed best, but ensemble classifiers and feature filtering with BNS perform comparably and are better suited when high recall is required.

Acknowledgments

We would like to thank the organizers for an interesting shared task. This work was carried out in the framework of AMiCA (IWT SBO-project 120007), funded by the Flemish government agency for Innovation by Science and Technology (IWT).

References

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Bart Desmet and Véronique Hoste. 2014. Recognising suicidal messages in dutch social media. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 830–835.

Bart Desmet, Véronique Hoste, David Verstraeten, and Jan Verhasselt. 2013. Gallop documentation. *LT3 Technical report*, pages 13–03.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305.

Louis Gottschalk and Goldine Gleser. 1960. An analysis of the verbal content of suicide notes. *British Journal of Medical Psychology*, 33(3):195–204.

Li Guan, Bibo Hao, Qijin Cheng, Paul SF Yip, and Tingshao Zhu. 2015. Identifying chinese microblog users with high suicide probability using internet-based profile and linguistic features: Classification model. *JMIR mental health*, 2(2):17.

Krystl Haerian, Hojjat Salmasian, and Carol Friedman. 2012. Methods for identifying suicide or suicidal ideation in EHRs. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 1244–1253. American Medical Informatics Association.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Yen-Pei Huang, Tiong Goh, and Chern Li Liew. 2007. Hunting suicide notes in web 2.0-preliminary findings. In *Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on*, pages 517–521. IEEE.

Jared Jashinsky, Scott Burton, Carl Hanson, Josh West, Christophe Giraud-Carrier, Michael Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through Twitter in the US. *Crisis*, 35(1):51–59.

Nadin Kökciyan, Arda Çelebi, Arzucan Özgür, and Suzan Üsküdarl. 2013. BOUNCE: Sentiment Classification in Twitter using Rich Feature Sets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 554–561, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Lieve Macken, Els Lefever, and Veronique Hoste. 2013. Texsis: bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology*, 19(1):1–30.

Matthew Mulholland and Joanne Quinn. 2013. Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using nlp. In *IJCNLP*, pages 680–684.

Charles Osgood and Evelyn Walker. 1959. Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*, 59(1):58.

James Pennebaker, Roger Booth, and Martha Francis. 2007. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The university of maryland clpsych 2015 shared task system. *NAACL HLT 2015*, pages 54–60.

Jess Jann Shapero. 2011. *The language of suicide notes*. Ph.D. thesis, University of Birmingham.

Shannon Wiltsey Stirman and James W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Yiming Yang and Jan Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.