Text Classification to Automatically Identify Online Patients Vulnerable to Depression

Taridzo Chomutare(✉)

University Hospital of North Norway, Tromsø, Norway
taridzo.chomutare@telemed.no

Abstract. Online communities are emerging as important sources of support for people with chronic illnesses such as diabetes and obesity, both of which have been associated with depression. The goal of this study was to assess the performance of text classification in identifying at-risk patients. We manually created a corpus of chat messages based on the ICD-10 depression diagnostic criteria, and trained multiple classifiers on the corpus. After selecting informative features and significant bigrams, a precision of 0.92, recall of 0.88, f-score of 0.92 was reached. Current findings demonstrate the feasibility of automatically identifying patients at risk of developing severe depression in online communities.

Keywords: Online communities · Text classification · Mood disorders

1 Introduction

The number of users in health online communities has grown exponentially in the past decade. Online psychosocial support, as a mediator of health outcomes, is now an even more compelling concept. The continued growth may be indicative of new roles online communities play in self-care, although the nature of the roles is still not well-understood, and this has existed as a problem for several years [2]. Further, little is yet known about the relationship between online participation and health outcomes. More recently however, studies have emerged that attempt to explain this relationship [5]. Alternative approaches to physiological measurements, such as analysis of psychosocial elements is warranted, especially since depression has been shown to co-occur with both diabetes and obesity.

In this study we use text classification as a tool for identifying online participants who may be at risk of developing severe depression. To date, text classification has been applied on varying biomedical datasets, including sentiment analysis [4]. Candid online conversations encode data about the mental state of the participants, and machine learning tools are suitable for automatically detecting mood disorders from these conversations. Text characteristics such as length and vocabulary have been shown to be important elements of any biomedical corpus [3]. The objective of this study was twofold: (i) to develop a categorized text corpus for depressive symptoms based on online chat messages, and (ii) train and test classifiers on the corpus, and test on unlabelled data.

P. Cipresso et al. (Eds.): MindCare 2014, LNICST 100, pp. 125–130, 2014.
DOI: 10.1007/978-3-319-11564-1_13

2 Methods

An overview of the method used in this study is illustrated in Fig. 1, where the study starts off with obtaining the relevant data. We selected an opportunistic convenience sample from two online communities; one for diabetes, and the other for obesity. The two communities are vibrant, with several million posts and a combined user base of more than 200k, which makes them a suitable sample to demonstrate automatic methods for the 'big data'.

A python program was developed to crawl and parse only the publicly available HTML data. Next, a categorized text corpus of depressive symptoms was developed using the ICD-10 criteria *(F32, F33)*. There is need to evaluate other criteria since the current choice of ICD-10 encoding was arbitrary, based on its simplicity. Three classifiers were trained on the corpus and tuned; first by selecting informative features, and then by using significant bigrams. Using only the features with high information gain can enhance classifier performance, while bigrams can be important in cases with large amounts of text features. An additional tuning point was combining multiple classifiers with voting, and this can sometimes enhance performance.

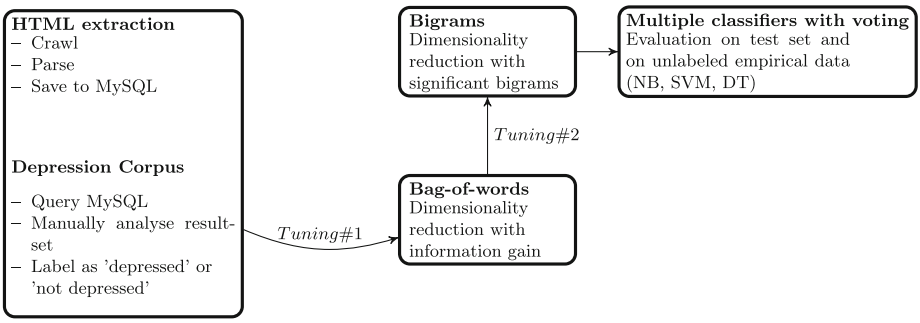

Fig. 1. The flow of the method; showing the chronology of the steps.

2.1 Diagnosis Criteria for Depression

Using the ICD-10 depression diagnostic criteria, the main symptoms and co-occurring secondary symptoms are shown in Table 1. Severity of depression is specified – no depression if less than 4 symptoms, mild if there are 4 symptoms, moderate with 5 or 6, and severe depression with more than 7 symptoms. However, we disregarded the time constraints specified by the diagnostic criteria. It is concievable that patients who are below the diagnostic threshold still suffer distress, therefore ignoring the time constraint does not necessarily diminish the value of our work. In addition, since the study cannot make claims about clinical diagnosis or evaluation of patients, the objective is only to identify patients who may be vulnerable or susceptible to depression – as a proof of concept.

Table 1. Symptoms based on the ICD-10 depression diagnostic criteria

Main symptoms	Secondary co-symptoms
persistent sadness or low mood; and/or	disturbed sleep, low self-confidence
loss of interests or pleasure	poor concentration or indecisiveness
fatigue or low energy	poor or increased appetite
	suicidal, agitation, guilt

2.2 Corpus of Depressive Symptoms

SQL queries were ran to create chat message profiles for each patient. Each profile is based on a set of messages the patient has written; both through creating their own message threads or commenting on (or responding to) existing threads. The query is constructed so that is satisfies the diagnostic criteria for co-occurrence of the primary and secondary symptoms. Ultimately, SQL queries become complex and unable to properly capture some contexts, e.g., chat lingo such as *'steamed up'* to mean *'agitated'*. However, using text classification, we can train models that learn more complex situations, such as someone jokingly saying *'. . . how sad am I!'*, when in fact they do not think of themselves as being actually sad.

The occurrences of depressive symptoms in the message profiles were manually examined to determine if they meet the threshold for the diagnosis. Given $f(x)$ as the mood disorder state, and a threshold number of symptoms T, then:

$$f(x) = \begin{cases} \textit{depressed}, & \text{if } x \geq T \\ \textit{not depressed}, & \text{otherwise} \end{cases}$$

Thus the profiles were classified into either 'depressed' or 'not depressed', and anything higher than a mild depression threshold ($x \geq 4$) is considered a depression state. Each patient profile of messages is stored as a separate text file in the respective corpus class directory structure.

2.3 Experimental Setup and Evaluation

The developed corpus was then used to train and test three classifiers; Naive Bayes (NB), Support Vector Machine (SVM) with a linear kernel and Decision Trees (DT). With two additional classifiers, a total of five classifiers are combined with voting. The evaluation metrics considered were (i) precision, (ii) recall, and (iii) the F-1 measure. Additional comparison is made between performance on unlabelled message profiles for patients in diabetes and obesity communities. The Natural Language Toolkit (NLTK) [1] and scikit-learn [6], both machine learning libraries in Python, were used in the experiments.

3 Results and Discussion

After selecting informative features, significant bigrams, and using multiple classifiers with voting, the maximum precision of 0.92, recall of 0.88, f-score of 0.92 was obtained using a linear kernel SVM. The developed corpus had a total of 100 message profiles, that is, representing a set of chat messages from 100 patients – where half of the corpus was categorized as 'depressed' and the other half as 'not depressed'. We started off with a corpus of just 20 profiles, but that proved insufficient for reasonable performance.

3.1 Training and Testing on Corpus

We trained the models with 75 % of the data and tested with the rest. We average the results of repeated random splits between training and testing data. Although this certainly increases the reliability of the models, the more rigorous k-fold cross-validation might have been a better alternative. The results in Table 2 show the outcomes before and after the tuning, as discussed next.

Dimensionality Reduction. The first point for performance tuning was to reduce the dimensionality of the feature space, because removing potentially noisy data such as stop-words, non-alpha characters and low information features can add clarity to the models. For example, words such as *'sad'* or *'depressed'*, or even more subtle, *'gulp'*, would likely occur on profiles for depressed patients, just as *'yeah!'* or *'happy'* would occur in non-depressed patients. Such words that likely occur more in one class, and not so much in the other, have high information gain.

Bag-of-Words Model and Significant Bigrams. One of the used classifiers, the Naive Bayes, has an assumption that the features are independent; using the *bag-of-words* model, where each feature is assumed to be independent. This has some shortcomings because of the tendency to use negated statements in chat lingo, such as, *'so not happy'*. In this instance, the classifier would take *not* and *happy* as two independent words, when clearly they are not. Uniting a sequence of two adjacent words (bigrams) can reduce the effect of these negated statements. By extension, *'not so happy'* may require analysing ngrams to yield even better performance, but this case is not considered in this study. Determining important bigrams implies taking account of the frequency distribution to check if they occur more frequently in one class than the other. By scoring each bigram, we can determine the significant ones to be used in the training of the model.

The results in Table 2 show that SVM with a linear kernel was the best classifier for the datasets. However, we find the DT classifier improving the most after tuning. The NB classifier had poor performance, which worsened after tuning, perhaps partially because of the underlying assumption of feature independence. Next, the best performing classifiers are applied to unlabelled data.

Table 2. Performance evaluation pre- and post-tuning of the classifiers.

	Naive Bayes		Linear SVM		Decision Tree		Voting
	Pre	Post	Pre	Post	Pre	Post	Post
Precision	0.63	0.59	0.88	0.92	0.81	0.89	0.92
Recall	0.58	0.56	0.79	0.88	0.64	0.79	0.88
F-1 score	0.60	0.57	0.88	0.92	0.68	0.88	0.92

3.2 Tests on Unlabeled Empirical Datasets

We randomly sampled unlabelled profiles for 10k patients in each diabetes and obesity community, and found there were generally more depressed people in the diabetes community, as shown in the pie charts in Fig. 2. If these findings were investigated further, together with literature related to clinical depression and chronic illnesses, there could be some interesting new insights into the comorbidities.

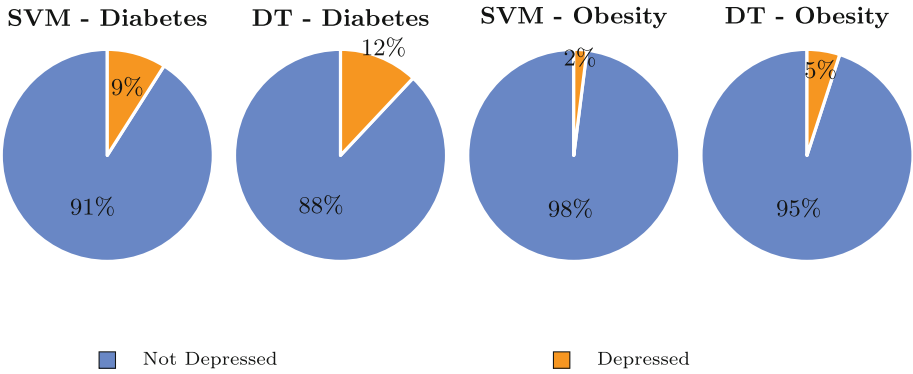

Fig. 2. Percentage of depression profiles in unlabeled message profiles.

The tests further revealed quite interesting gender and experience variations, although not tested for significance. We also identified several at-risk mothers of children with diabetes, one with a 12-year old son; likening the new diagnosis of her son to an *'axe falling on a loved one'*, describing herself as being *'numb'* and *'crying every morning'*, feeling like a morbidly obese man was *'sitting on her chest'*, and having an overwhelming *'sense of guilt'*.

3.3 Limitations

One limitation may be that we only used one diagnostic criterion, the ICD-10 encoding, and it is likely that there may be some differences with other diagnostic

criteria such as the DSM-IV. Another limitation may be that we did not consider how some depression symptoms such a fatigue relate to diabetes physiology such as hypoglycemia. Current work did not evaluate performance on unlabelled data, and this is left as future work.

4 Conclusion

Our results demonstrate the feasibility of a depression corpus based on online chat messages; something that was generalized to diabetes and obesity communities. Automatically identifying at-risk patients enables online communities to provide targeted help to patients who might otherwise be unaware of the impending problem. Although our approach produced promising results, further work is required before mood disorder behaviours in online communities can be more clearly understood. Future work will be based on multidisciplinary team effort to refine the corpus.

Acknowledgment. This work was supported in part by the Research Program for Telemedicine (HST), Helse Nord RHF, Norway.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python, 1st edn. O'Reilly Media Inc., New York (2009)
2. Chang, T., Chopra, V., Zhang, C., Woolford, S.J.: The role of social media in online weight management: systematic review. J. Med. Internet Res. **15**(11), e262 (2013)
3. Figueroa, R.L., Zeng-Treitler, Q.: Text classification performance: is the sample size the only factor to be considered? In: Lehmann, C.U., Ammenwerth, E., Nøhr, C. (eds.) MedInfo. Studies in Health Technology and Informatics, vol. 192, p. 1193. IOS Press, Amsterdam (2013)
4. Huh, J., Yetisgen-Yildiz, M., Pratt, W.: Text classification for assisting moderators in online health communities. J. Biomed. Inform. **46**(6), 998–1005 (2013)
5. Hwang, K.O., Ning, J., Trickey, A.W., Sciamanna, C.N.: Website usage and weight loss in a free commercial online weight loss program: retrospective cohort study. J. Med. Internet Res. **15**(1), e11 (2013)
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)