Identifying depression on Reddit: the effect of training data

Inna Pirina
Department of Linguistics
University of Tübingen
inna.pyrina@gmail.com

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

Abstract

This paper presents a set of classification experiments for identifying depression in posts gathered from social media platforms. In addition to the data gathered previously by other researchers, we collect additional data from the social media platform Reddit. Our experiments show promising results for identifying depression from social media texts. More importantly, however, we show that the choice of corpora is crucial in identifying depression and can lead to misleading conclusions in case of poor choice of data.

1 Introduction

Clinical depression, also referred to as major depressive disorder, is a serious mental condition that can interfere with normal daily life activities. One of the many risks of clinical depression is suicide – research has indicated that approximately two-thirds of people who die by suicide were dealing with depression at the time of death (Richards and O'Hara, 2014). Meanwhile, according to the World Health Organization, nearly 50% of people with clinical depression worldwide remain untreated. One of the main reasons why the disorder is ignored is believed to be under-diagnosis (Sheenan, 2004).

One of the many ways that the condition could be manifested is in the way people write: the words they choose and the general tone of the produced text are affected by the disorder (Reece et al., 2017). Due to the stigma around clinical depression, people tend to turn to the Internet, thus, making the data gathered from social media a valuable source of literary cues that could help identify depression from texts. Moreover, the ease of obtaining data makes Internet an attractive source for the purpose.

Early research on the relation between language and depression has been mostly theoretical, mainly focusing on the linguistic features that are manifested in 'depressed language', such as negatively-valenced words Beck et al. (1987), and frequent use of first-person pronouns Pyszczynski et al. (1987). These observations have been verified using corpus studies (Rude et al., 2004; Pennebaker et al., 2008), indicating, indeed, certain aspects of linguistic output is related to the speaker's or author's mental state.

A challenge in investigating the link between depression and the linguistic output is obtaining suitable data. And, one of the easy (and fruitful) data source for this purpose has been the Internet, in particular social media platforms (Ramirez-Esparza et al., 2008; Coppersmith et al., 2015).

Most of the earlier works have been focused on analyzing the language used by depressed individuals, and/or finding linguistic correlates of the depression. A more applicable approach to monitoring public or individual mental health requires explicit identification of depression from the linguistic samples. Such an application can complement the conventional diagnosis methods, and, if proven successful, it can be useful for diagnosis where conventional methods are not applicable. Similar to some of the recent studies (Coppersmith et al., 2015; Yates et al., 2017; Lynn et al., 2018), our aim in this paper is identifying depression from linguistic data. Using (mainly) corpora we gather from the social media platform Reddit, we experiment with a number of different classification models. Our focus here is on selection of corpora for reliable and generalizable analysis or identification of depression from the social media data.

2 Methods

2.1 Data

In part of our experiments, we use the data collected by Ramirez-Esparza et al. (2008), which

consists of 400 forum posts by depressed individuals. Ramirez-Esparza et al. (2008) used a similarly sized data set from a breast cancer forum as the 'control group' in their analyses. Since the control data was not available to us, we report results using an alternative set of documents collected by Gorbunova (2017) as the negative class in our classification experiments.

Our data was gathered from Reddit, which hosts over 10 000 online communities (also known as 'subreddits') of anonymous users united by common interests or discussion topics. In all data sets described below, we only collect the original posts, 'submissions', not the comments.

As an approximation to the data collection method of Ramirez-Esparza et al. (2008), we collect data from a relatively large subreddit that is devoted to depression, where authors seek support from the community. Similar to Ramirez-Esparza et al. (2008), we also collect the number of posts from subreddit devoted to breast cancer discussion, as the control set (or negative class). Since the differences between depression and breast cancer may involve serious topical differences, we also collect yet another set of posts from 'family' and 'friendship advice' subreddits, which we presume is topically more similar to depression subreddit.

In all of the cases above, however, the posts in the both positive and negative classes are topically specific. In practice, we would like to identify depression from everyday language, not necessarily language used for talking about depression, and seeking community support. As our more realistic example, we collected a number of posts following a protocol similar to Coppersmith et al. (2015) and Yates et al. (2017). First, we looked for expressions like 'I was just diagnosed with depression', on the depression subreddit. Unlike Yates et al. (2017), we do not manually check the sampled texts. As a result, a certain number of false positives are expected. For each author mentioning a diagnosis, we searched for the postings of the same author within a month of the original post in other subreddits, excluding some potentially related ones like 'Anxiety', 'mentalhealth' and 'depression_help'. The resulting posts are written by authors with (likely) depression, and to a large extend topically different than that of depression subreddit. To keep the training set size the same as the other data sets we use, we randomly sample 400 posts obtained in this manner for training, and another 400 posts for testing. Another difference from Yates et al. (2017) is that our training and test instances are the documents, not the authors. We also sample randomly the same amount of posts as our texts with authors without depression, from the same set of subreddits, but excluding the authors that posted in the depression subreddit during the time period we used for our investigation. For each setting, we pick only one document for each author.

In sum, we experiment with 8 data sets:

- `DSF` Posts from Depression Support Forums (Ramirez-Esparza et al., 2008)
- `DND` Posts from Non Depression Forums (Gorbunova, 2017)
- `DS` Posts from Depression Support subreddit
- `BC` Posts from Breast Cancer subreddit
- `FF` Posts from subreddits related to Family and Friends
- `DO` Posts from authors with (probable) Depression posted on Other forums
- `ND` Posts from authors with (probably) No Depression

All data sets have 400 training set items, and the data sets `DO` and `ND` also have additional 400 posts used as a reasonable test set.

2.2 Classifiers and tuning

We have experimented with a relatively large number of classification methods, including logistic regression and recurrent neural networks, in a number of different settings. In our experiments, the support vector machines (SVMs) with a combination of character and word n-grams of various sizes performed the best. We only report the experiments with SVM models.

In all cases we used linear SVMs with bag-of-n-grams features. SVMs are known to work well in a number of other text classification problems in this setting. The character and word n-grams of various sizes are extracted from the texts, and weighted using BM25 algorithm (Robertson et al., 2009). We optimized maximum order of character and word n-grams as well as the SVM margin parameter C through random search. A 5-fold cross-validation is performed for each parameter setting explored. For each experiment we report the setting where average scores over the 5-fold cross validation is the highest. The BM25 parameters 'k1' and 'b' were not optimized, and set to 0.75 and 2.0 respectively. For experiments with class imbalance, we

Model	5-fold F1	Test set F1
DSF–NDF	94.75	64.05
DS–BC	98.62	56.88
DS–FF	92.25	55.62
DS–ND	91.75	56.48
DO–ND	68.12	67.49
allD–allND	91.40	58.28

Table 1: Best 5-fold CV results obtained on each classification setting, together with the performance of the system on the test set. The model descriptions list data used for positive and negative class respectively. The data sets are explained in Section 2.1. The last row combined all 'positive' and 'negative' data sets, except the DO and ND sets. The scores are percentages.

used class weights during training to overcome the class imbalance problem.

All models were implemented in Python programming language, using scikit-learn package (Pedregosa et al., 2011). The source code for the classification models and data collection scripts are available at https://github.com/Inusette/Identifying-depression.

2.3 Evaluation

For evaluating the models, we report the standard measures of F_1 score (harmonic mean of precision and recall). We use the 'binary' version of the scores with positive class being text from authors with depression.

2.4 Experiments and results

We train 6 SVM classifiers, using different data sets descried in Section 2.1. Table 1 presents the performance comparison of the classifier on a number of different settings.

Each row in Table 1 presents F1-score of a binary SVM classifier on the data set as well as the performance of the same system on the test set consisting of DO and ND. Since each model is tuned for F1-score, the precision and recall values are rather balanced, and are not reported in Table 1. In general, 5-fold cross validation results are rather high, especially if both data sets are specific. Best results are obtained when both data sets are very specific, as in DS–BC case. The success of the classifier goes down as the texts belonging to the negative class comes from less specific domains. And in fact, the worse in-dataset results are obtained in our target setting, during which the classification of documents written by authors with diagnosed depression in non-depression related topics (DO) against the documents on the similar topics written by (presumably) healthy authors (ND). The gap between all other settings and DO–ND setting is rather large.

We also observe a very sharp drop of performance between the 5-fold cross validation results and the results on the test set. Interestingly, the most successful model (except DO–ND) on the test data is the forum data which is expected to be rather different from the all others which came from Reddit.

The last row of Table 1 reports the performance of a model where positive/negative instances of all other (except DO and ND) settings are combined. The resulting model is trained on more data, however, its data sources are not as harmonized as in other settings. As a result, it performs comparably, but worse than other specific models. However, the non-specificity seems to slightly help in the test set, resulting in better than all others (except DSF–NDF setting).

3 Discussion and conclusions

In this paper we reported a number of experiments on detecting depression from language samples collected from social media. Being able to detect depression from linguistic material is interesting both theoretically, and due to its potential applications as a diagnostic aid or for monitoring of public or individual mental health. These goals are viable only if we can identify depression to a successful degree. There has been a number promising results for detecting depression from the writing samples, particularly from social media texts (Ramirez-Esparza et al., 2008; Coppersmith et al., 2015; Reece et al., 2017; Lynn et al., 2018).

In this study, our focus has been the selection of sources for successful detection of depression from social media text. Our results clearly show that careful selection of sources is important for not obtaining illusionary results. This is particularly important if one intends to use the resulting systems in practical applications. However, it may be equally important, not to get wrong conclusions for more theoretically oriented research as well.

Another important contribution of our works is the use of Reddit for the purpose. There has been relatively few studies using Reddit for investigating linguistic aspects of mental health (De Choudhury and De, 2014; Yates et al., 2017). We believe Reddit's emphasis on anonymity is useful for ob-

taining less biased results. Reddit corpora also has the advantage of availability,[1] which can help reproducing the earlier results. Furthermore, not having length limitation like Twitter, a popular choice in other studies, may also be important in some cases. The F_1-scores we obtained on Reddit corpus, although higher than the results reported in Yates et al. (2017), is lower than the earlier results on Twitter (Coppersmith et al., 2015). This could potentially be due to the small number of training instances in our study. However, further investigation is needed for understanding the differences.

In this study we only reported results from linear classifiers, using simple character and word bag-of-n-gram features. These models are simple, fast, language independent, and performed better than other systems we experimented with, including a number of deep learning architectures (this is in line with some earlier work where same models and methodolgy is used on similar tasks, e.g., Çöltekin and Rama, 2016, 2018). Furthermore, although our focus in this paper has been their performance, the linear models are also more open to analysis, allowing investigation of (types) of features that are useful for the task.

References

Aaron T Beck, A John Rush, Brian F Shaw, and Gary Emery. 1987. *Cognitive Therapy of Depression*. Guilford Press.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear svms and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs at emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39. Association for Computational Linguistics.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media (ICWSM)*, Ann Arbor, Michigan, USA.

[1]The corpus we use is publicly available at `https://files.pushshift.io/reddit/submissions/`.

Anastasia Gorbunova. 2017. Predicting depression from online communication: comparison of three classification techniques. Master's thesis, University of Tübingen, Tübingen, Germany.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. Clpsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Cindy K. Chung, Ewa Kacewicz, and Nairan Ramirez-esparza. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *ICWSM*.

Tom Pyszczynski, Kathleen Holt, and Jeff Greenberg. 1987. Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. *Journal of Personality and Social Psychology*, 52(5):994–1001.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *International Conference on Weblogs and Social Media*, pages 102–108.

Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. 2017. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*.

C Steven Richards and Michael W O'Hara. 2014. *The Oxford Handbook of Depression and Comorbidity*. Oxford Library of Psychology. Oxford University Press.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

DV Sheenan. 2004. Depression: underdiagnosed, undertreated, underappreciated. 13.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.