Monitoring Tweets for Depression to Detect At-risk Users

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha
School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, ON, Canada, K1N 6N5
{zjami096, diana.inkpen, pkiri056}@uottawa.ca

Kenton White
Advanced Symbolics, Ottawa, ON, Canada, K1N 5S7
kenton.white@advancedsymbolics.com

Abstract

We propose an automated system that can identify at-risk users from their public social media activity, more specifically, from Twitter. The data that we collected is from the #BellLetsTalk campaign, which is a wide-reaching, multi-year program designed to break the silence around mental illness and support mental health across Canada. To achieve our goal, we trained a user-level classifier that can detect at-risk users that achieves a reasonable precision and recall. We also trained a tweet-level classifier that predicts if a tweet indicates depression. This task was much more difficult due to the imbalanced data. In the dataset that we labeled, we came across 5% depression tweets and 95% non-depression tweets. To handle this class imbalance, we used undersampling methods. The resulting classifier had high recall, but low precision. Therefore, we only use this classifier to compute the estimated percentage of depressed tweets and to add this value as a feature for the user-level classifier.

1 Introduction

According to a recent report of the World Health Organization (WHO), mental health is an integral part of health and well-being (WHO, 2004). Mental disorders can affect anyone, rich or poor, male or female, of any age or social group. The experience of mental illness is often described as difficult, especially when associated with demeaning prejudices and lack of understanding. Mental illness is also difficult to diagnose. There is no reliable laboratory test for most forms of mental illness and typically, diagnostic is based on the patient's self-reported experiences, behaviors reported by relatives, and a mental status examination. Unfortunately, mental disorder problems are increasing worldwide.

In the context of mental illness, depression is very common. In Canada, 5.3% of the population had presented a depressive episode in the past 12 months.[1] According to Canadian Mental Health Association (CMHA, 2016), 20% of Canadians belonging to different demographics have experienced mental illness during their lifetime, and around 8% of adults have gone through major depression. Mental Health Commission of Canada (MHCC, 2016) has reported on the broad implications of mental illness, where from nearly 4,000 Canadians that die each year by suicide, 90% of them were identified as having some form of a mental disorder. According to World Health Organization (WHO, 2016), suicide is a preventable health problem and to be successful in preventing suicide; therefore, it is of great importance to identify depression as a first indicator of further problems.

Apart from the severity of mental disorders and their influence on one's mental and physical health, the social stigma or discrimination in the forms of rejection, isolation, abuse and fear of embarrassment have made the individuals with mental disorders to be neglected by the community, as well as to stay away from obtaining the necessary treatments (WHO, 2016). Due to the severity mental disorders can cause to one's life and the impact it has on the entire society, organizations such as Bell Canada have initiated programs to raise funding for mental health programs as well as to create awareness within the society.[2]

The goal of this research is to exploit the mas-

[1] http://www.phac-aspc.gc.ca/cd-mc/mi-mm/depression-eng.php

[2] http://letstalk.bell.ca/en/

sive data issued from Twitter and apply social media mining and sentiment analysis methods to detect users at-risk of depression. It is an open question whether a tweet-level or user-level classifier is best for detecting at-risk people. A tweet-level classifier monitors individual tweets, identifying messages that indicate risk for depression; a user-level classifier looks at the tweet history and determines if a person is at risk from their corpus of messages over a period of time. This paper describes experiments on both classifiers.

Our system can be used by authorities to find a focused group of at-risk users. It is not a platform for labeling an individual as a patient with depression, but only a platform for raising an alarm so that the relevant authorities could take necessary interventions to further analyze the predicted user to confirm his/her state of mental health. We respect the ethical boundaries relating to the use of social media data and therefore do not use any user identification information in our research.

2 Related Work

With the gradual increase in social media usage and the extensive level of self-disclosure within such platforms (Park et al., 2012), research has been conducted to identify mental disorders at an individual as well as at a society level. Researchers have used features such as behavioural characteristics, depression language, emotion and linguistic style, reduced social activity, increased negative affect, clustered social network, raised interpersonal and medical fears and increased expression in religious involvement, use of negative words, in order to determine the cues of major depressive disorder (De Choudhury et al., 2013a; Tsugawa et al., 2015). Tsugawa et al. (2015), also used syntactical features such as bag of words (BOW) and word frequencies to identify the ratio of tweet topics and managed to conclude that topic modeling also adds a positive contribution to the predictive model compared to the use of the bag-of-words model, which could also result in overfitting.

The successful use of computational linguistics techniques in identifying the progress and level of depression of individuals in online therapy could bring greater insights to clinicians, to apply interventions effectively and efficiently. Howes et al. (2014) used 882 transcripts gathered from an online psychological therapy provider to determined that use of linguistic features can be considered as more valuable in predicting the progress of a patient compared to sentiment and topic-based analysis. In contrary to traditional sentiment analysis approaches that use three main polarity classes (i.e., positive, negative, and neutral), Shickel et al. (2016), divided the neutral class into two classes: neither positive nor negative and both positive and negative. With the use of syntactic, lexical, and also by representing words as vectors in the vector space (word embeddings), the authors managed to achieve an overall accuracy of 78% for the four-class polarity prediction.

De Choudhury et al. (2013b) and Schwartz et al. (2014) proposed methods to identify the level of depression among social media users (SMDI: Social Media Depression Index). Schwartz et al. (2014) used a classification model trained with n-grams, linguistic behavior and Latent Dirichlet Allocation (LDA) topics as features for predicting the individuals who are susceptible to having depression. In addition to open-vocabulary analysis and lexicon-based approaches such as Linguistic Inquiry and Word Count (LIWC), Coppersmith et al. (2014a) suggested language models, primarily based on unigrams and character 5-grams to determine the existence of mental disorders.

The Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task (Coppersmith et al., 2015) used self-reported data on Twitter about Post Traumatic Stress Disorder (PTSD) and depression, collected according to the procedure introduced by Coppersmith et al. (2014b). The shared task participants were provided with a dataset of self-reported users on PTSD and depression. For each user in the dataset, nearly 3,200 most recent posts were collected using the Twitter API. Resnik et al. (2015a), whose system ranked first in the CLPsych 2015 Shared Task, created 16 systems based on features derived using supervised LDA, supervised anchors (for topic modeling), lexical TF-IDF, and a combination of all. An SVM classifier with a linear kernel obtained an average precision above 0.80 for all the three tasks (i.e., depression vs. control, PTSD vs. control and depression vs. PTSD) and a maximum precision of 0.893 for differentiating PTSD users from the control group. Preotiuc-Pietro et al. (2015) employed user metadata and textual features from the corpus provided by the CLPsych 2015 Shared Task to develop a linear classifier to predict users

having either one of the mental illnesses. They have used the bag-of-words approach to aggregate word counts, topics derived from clustering methods and metadata (e.g., followers, followees, age, gender) from the users Twitter profile as the main feature categories. With the use of logistic regression and linear SVM in an ensemble of classifiers, the authors managed to obtain an average precision above 0.800 for all the three tasks and with a maximum score of 0.867 for differentiating users in the control group from the users with depression.

The use of the supervised LDA and the supervised anchor model was proven to be highly successful compared to the unsupervised clustering approaches, and even more efficient than using linguistic methods such as the use of n-grams and other lexicon based approaches (Resnik et al., 2015b). Resnik et al. (2015a) proved that such approaches can be successfully used in identifying users with depression, who have self-disclosed their mental illnesses on Twitter. In general, a clear distinction in the lexical and syntactic structure of the language used by individuals with different mental disorders, as well as between individuals within a control group, can be identified throughout the literature mentioned above, as well as from the explorative analysis conducted by Gkotsis et al. (2016). Due to the reliability of the lexical and behavioral features used in many of the models mentioned above, our proposed solution also focused on these feature categories. Even though the dataset we have used is relatively smaller than the ones used by most of the experiments mentioned above, we managed to obtain reliable results in identifying users with mental disorders.

3 Datasets

For this research, we prepared a dataset consisting of tweets from users who participated in #BellLetsTalk 2015 campaign. #BellLetsTalk is a campaign created by Bell Canada to help reduce stigma and promote awareness and understanding of mental health issues. Canadians opened up the dialogue on mental health, contributing more than 122 million tweets, texts, calls and social media shares on #BellLetsTalk Day, helping to raise more than $6.1 million for mental health initiatives.[3]

We collected data for the year 2015 and we limited it to Canadian users. 156,612 tweets were obtained from 25,362 users. Only data made public by users was collected for this task. To clean the dataset, we used LDA (Grün and Hornik (2011)), to obtain topics from tweets. Prominent topics included "campaign publicity", "mental health awareness", "raising donations", "facts about mental health". If a tweet contained two or more keywords from any of the mentioned topics, it was removed from the dataset. Additionally, retweets, tweets beginning with a mention (@), short tweets (less than 5 words), and URLs were removed. We then used words like "depressed", "suffer", "attempt", "suicide", "battle", "struggle", "diagnosed", in addition to first person pronouns, to identify a subset of tweets where users are talking about depression. A human annotator reviewed these tweets to verify whether the user is disclosing their own depression or talking about a friend or family member. Using this method we identified 95 users who disclosed their own depression. For these 95 users we collect all tweets from 2015 and refer to these as "self-disclosed" set. All remaining users were considered as control users. Similarly, for control users, all tweets from 2015 are collected and referred to as "control" set.

To prepare a dataset to label at tweet-level, we selected 60 users who had between 100 and 300 tweets. 30 users were selected from self-disclosed set, and 30 from control set. We asked two annotators to label 10 users with depression level 0-1, where 0 indicates no depression and 1 indicates some depression.[4] We found that most tweets fell into the "no depression" class. Since annotation is an expensive and a time-consuming task, we looked for tweets that could be removed without losing relevant tweets. Our first intuition was to remove tweets containing positive words, but this intuition proved to be false as many of the tweets labeled as depressed contained positive words. Next we looked for neutral tweets. Most neutral tweets were labeled as "no depression" and hence we decided to remove these from our dataset. The list of positive and negative words was obtained

[3] http://www.ctvnews.ca/health/bell-let-s-talk-breaks-records-raises-more-than-6m-for-mental-health-1.2211607

[4] Our annotators were not experts, though one of them is a student in Psychology. We would like to have the annotations verified by an expert, in the future.

from Hansen et al. (2011). The final dataset consisted of 8,753 tweets. We refer to this dataset as **60Users**.[5] The annotators were then asked to label the remaining 50 users. The Kappa value for 2-annotator agreement was found to be 0.67. If a tweet was labeled as depressed by at least one annotator, the tweet was considered as depressed.[6]

We prepared a larger dataset to be labeled at user-level. This dataset consists of 80 users from self-disclosed set and 80 control users. It included the 60 users annotated above at tweet-level. We refer to this dataset as **160Users**.[7] For fast annotation at user-level, we provided an undersampled version of the dataset to annotators. It was undersampled using our tweet-level classifier discussed in section 4. Nonetheless, for our experiments, we used all tweets from 160 users. The dataset was annotated by two annotators as "depressed" and "not-depressed" user. The conflicts were resolved by a third annotator. The following guidelines were provided for the task:

- Depressed: The user shows clear signs of depression, or shows signs that could result in depression in near future. There is enough reason for a public health member or doctor to investigate further. Additionally, users who self-disclose depression but there are no other tweets indicative of depression, are also labeled as depressed.[8]

- Not-depressed: the user does not show any signs of depression.

A third dataset is obtained from CLPsych shared task 2015 (Coppersmith et al., 2015). The dataset consists of 1,746 users. The training set consists of 327 depression users, 246 PTSD users, and, for each, an age and gender matched control user. The test set consists of 150 depression users, 150 PTSD users, but we cannot use it because the labels for the test set are not available. For our task, we use the depression and control users from the training set. We refer to this as the **CLPsych2015** dataset.

The 60Users dataset was split to contain one-third of the tweets for testing (2,971 tweets) and two-thirds for training purposes (5,782 tweets). In the case of 160Users and the CLPysch2015 datasets, we split each dataset into 70% training and 30% test set. Each model was trained on the training set using 10-fold cross validation and then tested on a held out test set.

4 Tweet-level Classifier

For the tweet-level classification, a preliminary experiment was performed on 60Users dataset using BOW as features and SVM classifier. This gave a very high accuracy because it classified all the tweets in the majority class. This was due to class imbalance. The dataset consisted of 95% not depressed tweets and 5% depressed tweets. To deal with the class imbalance, we then experimented with re-sampling methods including undersampling (randomly removing examples from the majority class) and with oversampling, in particular with adding examples for the minority class using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). For evaluation, we will look at recall, precision, and F-measure for the class of interest (depression), instead of accuracy.

The goal of training a tweet-level classifier is to predict whether a given tweet indicated depression or not. For this, we perform two sets of experiments. The first set of experiments uses 7 features derived from tweet's text. These include polarity words, depression words, first person pronoun, and second person pronoun counts. These are referred to as **initial features**. Polarity words include counts of very negative words, negative words, positive words and very positive words. The list of polarity words was obtained from AFINN (Hansen et al., 2011). Depression related terms are obtained from Maigrot et al. (2016). The second set of experiments uses unigrams (BOW), in addition to the 7 initial features.

Each set consists of 3 experiments performed on the 8,753 tweets from the 60Users dataset.

1. Linear SVM trained on the original dataset

2. Linear SVM trained on the dataset balanced using SMOTE[9]

[5]The 60Users dataset annotated at tweet-level will be made available on request for further research

[6]Considering a tweet as depressed only when both annotators agreed that the tweet was depressed reduced the amount of positive training samples, but did not impact performance

[7]The 160Users dataset will be made available on request for further research.

[8]The users who self-disclose depression, but do not have other tweets indicative of depression in the dataset are marked as depressed in order to maximize the number of at-risk users predicted by the classifier.

[9]For oversampling, we use the SMOTE function from the

3. Linear SVM trained on the dataset balanced by undersampling[10]

5 User-Level Classifier

The goal of training a user-level classifier is to predict if a given user is at-risk of suffering from depression. For this, we train models on the 160Users dataset. For user-level classification, we start by making tweet-level predictions using the best model obtained from experiments described in Section 4. The initial features are generated as a requirement for the tweet-level classifier. The tweet-level predictions are then used to compute the percentage of depressed tweets for each user. Next, the text of all the tweets for each user is merged, and the initial features are summed.[11]

During data annotation of the #BellLetsTalk users, we noticed that several users disclosed depression, but their tweets, at least those included in our dataset, did not indicate depression. Although these users were labeled as depressed, we noticed that removing such users from training set helps us to improve our models. For this we compute an additional feature called IsSelfReported for each user. The percentage of depressed tweets (hereafter called %DT) along with isSelfReported is used to decide whether a user should be removed from the training set. If IsSelfReported is True AND %DT is less than 10%, only then, the user is removed from the training set.

Several sets of experiments are performed for this task. An initial baseline experiment is performed using 7 initial features. The second set of experiments uses 8 features (the initial features + %DT). The third set of experiments uses 9 features (the initial features + %DT + isSelfReported). The fourth set of experiments uses a total of 115 features. The purpose for this was to identify whether increasing the number of features has a significant impact on performance. These **additional-features** include LIWC features, sentiment features, emoticon counts, text readability (SMOG, Flesh, Kincaid), and community features such as favorite counts, replies, mentions, retweets, in addition to initial features, %DT, and IsSelfReported.

DMwR package (Torgo, 2010) with default values. This implementation is based on (Chawla et al., 2002)

[10]For undersampling we used the "downSample" function in the CARET package (Kuhn et al., 2012) with default values

[11]The data is centered and scaled during model training

Each set includes three experiments performed on 160Users dataset.

1. Linear SVM trained on the original dataset
2. Linear SVM trained on the dataset balanced using SMOTE
3. Linear SVM trained on the dataset balanced by undersampling

Unlike 60Users dataset that was highly imbalanced, 160Users dataset had a relatively smaller degree of imbalance. The 160Users dataset consisted of 43% positive class and 57% negative class samples. The reason for using re-sampling methods at user-level was to investigate if performance can be improved by training a model on a fully balanced dataset.

From these experiments, we identify the model with highest performance. The set of features, and re-sampling method identified in relation to this model are then used in further experiments. These experiments include training further models using the CLPsych2015 dataset instead of the 160Users dataset. We also merge the 160Users dataset and CLPsych2015 dataset to investigate whether using a larger training data improves the performance.

6 Experimental setup

For this research, all the development is done in R version 3.3 (R Development Core Team, 2008) using the Rstudio IDE (RStudio Team, 2015). Data preparation, feature extraction, and classification tasks are performed using a variety of R packages. All classifiers were used from R's Caret package (Kuhn et al., 2012). Classifiers were trained using 10-fold cross validation to avoid over-fitting and then tested on a held-out test set. The results presented in Section 7 are those obtained on the held-out test set.

7 Results

For both tasks, tweet-level classification and user-level classification, we report precision, recall & F-measure for the positive class (depression), as performance measures. Precision and recall are more informative than accuracy, due to the data being imbalanced. For example, baseline experiments for tweet-level classification returns an accuracy of 95% by classifying all samples as majority class, which is not a true reflection of classifier's performance.

Model	training set	features
Tweet-level		
baseline.tweet	60Users	BOW
exp1	60Users	Initial features (polarity word counts, depression word count, pronoun counts)
exp2	60Users	Initial features + BOW
User-level		
baseline.user	160Users	Initial features
exp3	160Users	Initial features + %DT
exp4	160Users	Initial features + %DT + IsSelfReported
exp4 + additional features	160Users	Initial features + %DT + isSelfReported + community features + LIWC features + NRC sentiment feat. + emoticon features + readability features
exp5	CLPsych2015	Initial features + %DT + isSelfReported
exp6	160Users + CLPsych2015	Initial features + %DT + isSelfReported

Table 1: Datasets and features used for tweet-level and user-level experiments

For measuring performance at user level, we think that recall is somewhat more important for the task, therefore we aim at achieving high recall. This can be justified by keeping in mind the problem we are attempting to solve. In the context of detecting depression, a false positive (FP) is defined as a user who is predicted to have depression but does not actually suffer from depression. A false negative (FN) is defined as a user who is actually depressed but is predicted to not have depression. A classifier detecting more false positives would result in lower precision, the cost of which is that the state would need to invest more money to help users who are not actually depressed. On the other hand, a classifier detecting more false negatives would result in lower recall, the cost of which is that users suffering from depression will not get the help they need on time, which could lead to serious consequences, like suicide. So low recall could lead to loss of human life.

At the same time, we are trying to find a balance of precision and recall. A perfect recall of 1, with a very low precision (e.g., 0.2) is also not an acceptable outcome. In such cases, we look at F-measure, which combines both precision and recall. In particular, we look at the precision, recall, and F-measure of the positive class, obtained on the held-out test sets.

7.1 Tweet-level Classifier

Table 7.1 shows the results obtained for the tweet-level classification experiments. Performance is reported on a held-out testset obtained from 60Users dataset. None of the classifiers performed well on the task of identifying depressed tweets. The best performing model (exp1-Undersample) is identified in bold. This is a Linear SVM classifier trained on an undersampled training set and uses 7 initial features without BOW. We obtain a precision of 0.1237 and a recall of 0.8020, with F1 of 0.2144.[12]

The poor performance of all models indicates the complexity of the task and the fact that one tweet is not sufficient to detect depression.

7.2 User-level Classifier

Table 7.2 shows results obtained for user-level classification experiments. Performance is reported on a 30% held-out test set obtained from 160Users dataset. For exp3, the results improved a lot over the baseline with initial features. This shows that the features %DT computed with the tweet-level classifier helps. The best performing model (exp5) is identified in bold. This is a Linear SVM classifier trained on a balanced dataset (CLPsych2015) and uses 9 features (Initial features + %DT + isSelfReported). We obtain a precision of 0.7083, a recall of 0.85, and F1 of 0.7727.

From exp3 and exp4 in Table 7.2, we observe that the dataset balanced using re-sampling methods provide better recall. For this reason, when we train models on the combined dataset (exp6), we continue to balance the datasets using SMOTE and undersampling. The CLPsych2015 dataset (exp 5) is perfectly balanced and therefore does not require balancing using re-sampling methods.

We note that the model trained on CLPysch2015 dataset performs better than the model trained on the 160Users dataset when using the same features. This could be due to larger training data. On the other hand, performance (in terms of recall) drops when the dataset size is increased further by combining the 160Users and CLPsych2015 datasets and

[12]For the tweet-level and user-level classifiers, we experimented with other SVM kernels, but the results were worse. The same for other classifiers than SVM.

ModelName	Accuracy	Precision	Recall	F1
baseline	0.9469	1.0000	0.0111	0.0219
exp1-Original	0.9337	NA	0.0000	NA
exp1-SMOTE	0.7816	0.1706	0.5939	0.2650
exp1-Undersample	**0.6102**	**0.1237**	**0.8020**	**0.2144**
exp2-Original	0.9303	0.2222	0.0203	0.0372
exp2-SMOTE	0.7711	0.1124	0.3553	0.1707
exp2-Undersample	0.6143	0.1219	0.7766	0.2107

Table 2: Performance of tweet-level classifiers on the test set

balanced using SMOTE, but remains constant when balanced using undersampling.

Upon investigation as to why undersampling performs better than SMOTE, we discovered that SMOTE oversamples minority class instances, but does not fully balance the training data, whereas undersampling balances the training data. Hence, models trained on a balanced training set result in better performance.

It is interesting to see that models trained on 160Users (exp3 and exp4) perform better on CLPsych2015 dataset, while the model trained on CLPsych2015 dataset (exp5) performs better on the 160Users dataset.

The results for exp4+additionalFeatures are not reported because they are not significantly different from exp4 (though further investigations will need to be done in future work).

In terms of comparing the tweet-level classification task and the user-level classification task, we conclude that user-level models perform much better even with a small number of features.

7.3 Comparison to Related Work

Resnik et al. (2015a) and Preotiuc-Pietro et al. (2015) reported good performance on the dataset made available through the CLPsych2015 shared task, as mentioned in Section 2. We ran our top-performing user-level classifiers on the training set of CLPsych2015 shared task data. Results are provided in Table 7.3. We report only the SMOTE versions of the classifiers since they obtained better results. The feature %DT helps a lot on this dataset (according to exp3). We note that exp5 that gave the highest performance on the 160Users dataset performs consistently well on the CLPsych users, even though performance is slightly lower in comparison.

These results are not comparable with those reported by (Resnik et al., 2015a) and (Preotiuc-Pietro et al., 2015), for two reasons. First, in comparison to Resnik et al. (2015a) and Preotiuc-Pietro et al. (2015), who report performance on a different test set. We report performance on the 30% of the training users provided to us, that we kept aside for testing, because of the unavailability of the labels for the test users from the shared task. Second, the shared task uses precision at a certain recall level as the main performance measure, while we report standard precision and recall, and we selected our model to have a high recall.

8 Conclusion and Future Work

In conclusion, we proposed models for tweet-level classification and used them to compute the percentage of depressed tweets for each user. We also proposed models for user-level classification. We experimented with many features, including the percentage of depressed tweets, which was shown to help improve the performance of the user-level classifier. We annotated our own dataset from the #BellLetsTalk campaign, but we also experimented with the existing dataset from CLPsych2015.

In future work, we plan to study depression among groups of users based on their age, gender, locations and other demographic attributes. We also plan to look into identifying other kinds of mental disorders, and detecting suicidal ideation.

Acknowledgments

We appreciate the helpful comments of the anonymous reviewers. We thank the annotators (Bryan Paget and Sameen Salim) for their time and expertise. We thank the organizers of CLPsych 2015 for providing us access to their datasets. This research is funded by Natural Sciences and Engineering Research Council of Canada (NSERC).

Experiment	Accuracy	Precision	Recall	F1
baseline	0.617	1.0000	0.1000	0.1818
exp3-Original	0.6383	1.0000	0.1500	0.2608
exp3-SMOTE	0.6809	0.8571	0.3000	0.4444
exp3-Undersample	0.7021	0.7500	0.4500	0.5625
exp4-Original	0.6809	0.7778	0.3500	0.4828
exp4-SMOTE	0.766	0.7647	0.6500	0.7027
exp4-Undersample	0.766	0.7143	0.7500	0.7317
exp5-Original	**0.7872**	**0.7083**	**0.8500**	**0.7727**
exp6-SMOTE	0.7872	0.8571	0.6000	0.7059
exp6-UnderSample	0.7872	0.7083	0.8500	0.7727

Table 3: Performance of user-level classifiers on 160Users test set

Experiment	Accuracy	Precision	Recall	F1
exp3-SMOTE	0.6198	0.5966	0.7396	0.6605
exp3-Undersample	0.625	0.5984	0.7604	0.6697
exp4-SMOTE	0.5885	0.5895	0.5833	0.5864
exp4-Undersample	0.5885	0.5876	0.5938	0.5907
exp5-Original	**0.6094**	**0.5827**	**0.7708**	**0.6637**
exp6-SMOTE	0.6146	0.5902	0.7500	0.6606
exp6-UnderSample	0.6094	0.5827	0.7708	0.6637

Table 4: Performance of user-level classifiers on the CLPsych2015 test set

References

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

CMHA. 2016. Facts about mental illness. http://www.cmha.ca/media/fast-facts-about-mental-illness/#.WHbdMRsrK00.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Measuring Post Traumatic Stress Disorder in Twitter. In *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).*. volume 2, pages 23–45.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014b. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 51–60. http://www.aclweb.org/anthology/W/W14/W14-3207.

Glen Coppersmith, Mark Dredze, Craig Harman, Hollingshead Kristy, 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.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference* pages 47–56. https://doi.org/10.1145/2464464.2464480.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting Depression via Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. volume 2, pages 128–137. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351.

George Gkotsis, Anika Oellrich, Tim J P Hubbard, Richard J B Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, San Diego, CA, USA, pages 63–73.

Bettina Grün and Kurt Hornik. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40(13):1–30. https://doi.org/10.18637/jss.v040.i13.

Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good friends, bad news-affect and virality in twitter. *Future information technology* pages 34–43.

Christine Howes, Matthew Purver, and Rose McCabe.

2014. Linguistic Indicators of Severity and Progress in Online Text-based Therapy for Depression. In *Workshop on Computational Linguistics and Clinical Psychology*. 611733, pages 7–16.

Max Kuhn, Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, and Allan Engelhardt. 2012. *caret: Classification and Regression Training*. R package version 5.15-044. http://CRAN.R-project.org/package=caret.

Cédric Maigrot, Sandra Bringay, and Jérôme Azé. 2016. Concept drift vs suicide : How one can help prevent the other? In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*. Konya, Turkey.

MHCC. 2016. http://www.mentalhealthcommission.ca/sites/default/files/mhcc_annualreport2015_en.v7-ebook_0.pdf.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive Moods of Users Portrayed in Twitter. In *ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*. pages 1–8.

Daniel Preotiuc-Pietro, Maarten Sap, H. Andrew Schwartz, and Lyle Ungar. 2015. Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 40–45.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015a. The University of Maryland CLPsych 2015 Shared Task System. In *CLPsych 2015 Shared Task System*. c, pages 54–60.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-an Nguyen, and Jordan Boyd-graber. 2015b. Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. volume 1, pages 99–107.

RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

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: From Linguistic Signal to Clinical Reality*. pages 118–125. http://www.aclweb.org/anthology/W/W14/W14-3214.

Benjamin Shickel, Martin Heesacker, Sherry Benton, Ashkan Ebadi, Paul Nickerson, and Parisa Rashidi. 2016. Self-Reflective Sentiment Analysis. In *Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, pages 23–32. http://www.aclweb.org/anthology/W16-0303.

L. Torgo. 2010. *Data Mining with R, learning with case studies*. Chapman and Hall/CRC. http://www.dcc.fc.up.pt/ ltorgo/DataMiningWithR.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. pages 3187–3196. https://doi.org/10.1145/2702123.2702280.

WHO. 2004. Promoting mental health: Concepts, emerging evidence, practice: Summary report .

WHO. 2016. Mental health: a state of well-being. http://www.who.int/features/factfiles/mental_health/en/.