Mental Health Severity Detection from Psychological Forum Data using Domain-Specific Unlabelled Data

Braja Gopal Patra, PhD¹, Reshma Kar, MTech², Kirk Roberts, PhD³, Hulin Wu, PhD¹,³
¹Department of Biostatistics and Data Science, School of Public Health,
²School of Biomedical Informatics,
The University of Texas Health Science Center at Houston, Houston, TX
³Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India

Abstract

Mental health has become a growing concern in the medical field, yet remains difficult to study due to both privacy concerns and the lack of objectively quantifiable measurements (e.g., lab tests, physical exams). Instead, the data that is available for mental health is largely based on subjective accounts of a patient’s experience, and thus typically is expressed exclusively in text. An important source of such data comes from online sources and directly from the patient, including many forms of social media. In this work, we utilize the datasets provided by the CLPsych shared tasks in 2016 and 2017, derived from online forum posts of ReachOut which have been manually classified according to mental health severity. We implemented an automated severity labeling system using different machine and deep learning algorithms. Our approach combines both supervised and semi-supervised embedding methods using corpus from ReachOut (both labeled and unlabelled) and WebMD (unlabelled). Metadata, syntactic, semantic, and embedding features were used to classify the posts into four categories (green, amber, red, and crisis). The developed systems outperformed other state-of-the-art systems developed on the ReachOut dataset and obtained the maximum micro-averaged F-scores of 0.86 and 0.80 for CLPsych 2016 and 2017 test datasets, respectively, using the above features.

Introduction

Around 970 million people currently suffer from mental or neurological disorders, placing mental disorders among the leading causes of ill-health and disability worldwide. The World Health Organization reported that 800,000 people die by suicide every year due to mental illness. Lack of awareness of mental health causes individuals to associate it with social stigma, which prevents early help-seeking behavior. Consequently, chances of a complete or partial recovery in patients may be reduced. Effective and evidence-based interventions can be implemented at population, sub-population, and individual levels to prevent suicide and suicide attempts.

With recent technological advances, mental health forums are being used to share the experiences anonymously and relevant information which is helpful for users with similar mental health conditions. These forums are a simple way to facilitate such peer-support online, but when they involve vulnerable people and sensitive matter, they require careful cultivation. Social media driven peer-support groups like ReachOut help patients to feel better by allowing them to discuss their experiences freely with other patients who have been through similar situations. ReachOut is a non-profit organization initially started in Australia and later extended to the USA, Ireland, and Canada. It is an online forum, where psychologically affected young people and their parents can get advice from professionals or other users recovered from similar issues or problems. However, it requires the supervision of moderators to provide quick support to users expressing suicidal indentation. The human moderation for these online forums raises concerns about cost and scalability. The increased number of users in such forums makes the task of moderators unmanageable without the help of automated systems. The recent progress in text mining and natural language processing (NLP) offers a solution to this problem by identifying urgent posts so that the human moderator can be assigned sooner. It was observed that NLP can be employed to reduce the average response time of the human moderators for urgent posts.

More recently, several NLP techniques have been proposed to decipher psychological information or psychiatric symptoms from electronic health records (EHRs), psychiatric notes, tweets, support groups, and so on. Dif-

*https://ourworldindata.org/mental-health
†https://www.who.int/news-room/events/detail/2019/10/10/default-calendar/world-mental-health-day-2019-focus-on-suicide-prevention
‡https://au.reachout.com
ifferent supervised and semi-supervised methods were employed to extract psychiatric information from the text using many machine learning algorithms. Further, deep learning techniques were popularly used to improve performance of these systems.

In this paper, we present a system that we developed to identify the severity of posts based on which moderators can be assigned to serious patients. Latent Semantic Analysis (LSA), word2vec, Doc2Vec, and other handcrafted features were implemented to convert posts into feature vectors. Further, embedding methods were built using labeled and unlabelled corpora of the Computational Linguistics and Clinical Psychology (CLPsych) shared tasks as well as a large unlabelled corpus collected from the WebMD psychiatry forum. Finally, different machine learning and deep learning algorithms were used for classifying posts into four categories. The systems were trained on CLPsych 2016 and 2017 training datasets and tested on CLPsych 2016 and 2017 test datasets, respectively. The CLPsych datasets are unbalanced (different number of instances for each class with significantly less number of instances in red and crisis categories) and thus we propose a multi-level classification approach for better system performance. A multi-level system classified the posts into urgent (red and crisis) and non-urgent (green and amber), then classified them into respective categories.

**Background**

Mental health issues may often have devastating impact on patients and their family. However, it is unclear how mental impairment manifests in common forms of communication, for example, social media communications. Hence, it is worthwhile to analyze subjects’ text for symptoms of mental illnesses. In addition to diagnosing diseases, detecting its severity is also important in the context of social peer support groups, to detect the patients who require urgent attention. This section broadly reviews the progress in text analytics for mental health. Mainly, three types of tasks have been performed in recent years depending upon the datasets: a) identifying symptoms or type of diseases from the medical notes of psychiatric patients or EHR data; b) identifying the type of psychological diseases from social media posts such as Twitter or blogs; c) identifying the severity of mental health issues from psychological peer support forums.

**Psychological Analysis of Medical Notes and EHRs:** Electronic health records (EHRs) facilitate the storage of narrative interpersonal communications and observations of patients. Psychiatric symptoms are the key information in such narratives, which often consist of a subjective description with details of the patient’s experience. Psychiatric symptoms play an important role in prevention, diagnosis, and intervention of mental problems. Psychiatric symptom detection from psychiatric notes may serve as decision support systems for doctors. There has been work on both supervised and unsupervised methods for EHR data and labeling psychiatric notes. Tran and Kavuluru implement the former task by applying deep learning and obtained average F-scores of 63.14 and 61.90 using convolutional neural networks (CNNs) and recurrent hierarchical networks with attention, respectively. In Zhang et al., the authors present an unsupervised technique for labeling psychiatric notes by comparing symptoms in psychiatric notes to certain ‘seed terms’ present in medical knowledge databases, namely MedlinePlus, Mayo Clinic, and the American Psychiatric Association. This study emphasizes on the embeddings created using different knowledge bases which is useful for many tasks in mental health.

**Psychological Analysis of Social Media Data:** Most of the work on social media text is based on data collected from Twitter. The common objectives are either a) automatic labeling of mental health issues or b) predicting the onset of diseases. For automatic labeling of mental health issues, the common approach was to label Twitter data manually and then train a classification algorithm. Gkotsis et al. first identified mental health related posts (accuracy 91.8%) and then classified the posts into 11 themes (accuracy 71.3%). The best result was obtained using a CNN deep learning model. Coppersmith presented the results of a shared and an unshared task (restricted within John Hopkins University). Their primary objective was to identify patients with PTSD and depression who shared their mental health issues on social media. The major drawback in this work was that it primarily considered precision as an evaluation metric and did not consider recall. Choudhury et al. recruited patients with depression and evaluated

---

1. [https://medlineplus.gov/]
2. [https://www.mayoclinic.org/]
3. [https://www.psychiatry.org/]

---

488
their Twitter history. The authors found certain behavioral metrics to be related to depression and the system obtained an accuracy of 70.35% in predicting depression. Since mental illness presents symptoms which are common among otherwise healthy subjects, proactive detection of diseases from social media is an important research area. However, recent concerns about the privacy issues created concerns in the study of mental health from social media.

**Psychological Analysis of Peer Support Forums:** In 2016 and 2017, the CLPsych Shared Task presented a challenge of marking the severity of patient’s posts. The severity was graded into four labels with green signifying the least severe and crisis signifying the most severe. The best performing team in the 2016 task obtained an F-score of 0.85 with ensemble classification using three classifiers. Multiple variations of n-gram features were used for the above experiment. Interestingly, the classifiers were trained on 12 fine labels and then adjusted to 4 coarse labels. In the CLPsych 2016 shared task, SVM or logistic regression was a common choice for teams. In the CLPsych 2017 shared task, Xia and Liu’s system obtained the maximum macro averaged F-score of 0.467 using voting based classification. Altszyler et al. used different features such as syntactic, semantic, n-gram, and embedding features with ensemble based machine learning algorithms and the system obtained 0.462 macro-averaged F-score. The major drawback of the CLPsych shared tasks was selecting macro-averaged F-score as one of the official metrics for ranking the participant’s systems. Macro-average is good for evaluation when the number of instances in each category is balanced whereas both of the shared tasks do not have an equal number of instances. In another sense, this is a useful metric for emphasizing red and crisis categories with fewer instances. Also, some participants did not consider some instances from green and amber categories to reduce the imbalance in the datasets. In the above CLPsych shared tasks, the participants used mainly machine learning algorithms with stylistic, metadata, and word embeddings. However, domain knowledge and deep learning algorithms have not been explored for this task to date.

**Materials and Methods**

**Dataset**

The datasets used in our experiments were provided by the organizers of the CLPsych tasks in 2016** and 2017†. These datasets contain posts which were annotated with four categories (green, amber, red, and crisis) and were collected from ReachOut. The four categories are:

- **Crisis** indicates that the author (or someone they know) of the post is in imminent risk of being harmed, or harming others. Such posts should be prioritized above all so that a moderator can be assigned immediately.

- **Red** indicates that a moderator should respond to the post as soon as possible.

- **Amber** indicates that a moderator should address the post at some point, but they need not do so immediately.

- **Green** identifies posts that do not require direct input from a moderator and can safely be left for the wider community of peers to respond to.

The training datasets consist of 947 and 1,188 posts, and test datasets consist of 241 and 400 posts for 2016 and 2017, respectively. The CLPsych 2016 training and test datasets are a subset of the CLPsych 2017 training and test datasets, respectively. The organizers also provided 140,892 unlabeled posts. The statistics for training and test datasets for each category are provided in Table 1. It can be observed from Table 1 that the datasets are not evenly distributed among all the categories. The number of instances in crisis category is lesser than other categories.

Each instance contains useful information related to users such as login id, response status, post time, edit time, etc. with a subject and a comment. It also contains the number of kudos for post by other users in the forum. For the current study, we did not consider kudos because any relevant statistic concerning all categories was not found. However, we do not undermine the importance of kudos for other tasks, where a high number of kudos after the post is posted may have significance and it may be considered as one of the key information for that post. Example of posts from four categories are presented in Table 2. Separate files were provided with gold annotations, and the annotation details can be found in Milne et al.

---

**References**

**http://clpsych.org/shared-task-2016/**

††http://clpsych.org/shared-task-2017/
Table 1: Training and test datasets of CLPsych 2016 and 2017 with class-wise distribution.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Green</th>
<th>Amber</th>
<th>Red</th>
<th>Crisis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training 2016</td>
<td>549</td>
<td>249</td>
<td>110</td>
<td>39</td>
<td>947</td>
</tr>
<tr>
<td>Training 2017</td>
<td>715</td>
<td>296</td>
<td>137</td>
<td>40</td>
<td>1188</td>
</tr>
<tr>
<td>Test 2016</td>
<td>166</td>
<td>47</td>
<td>27</td>
<td>1</td>
<td>241</td>
</tr>
<tr>
<td>Test 2017</td>
<td>216</td>
<td>94</td>
<td>48</td>
<td>42</td>
<td>400</td>
</tr>
</tbody>
</table>

There are many online resources for mental health patients such as MedlinePlus, Mayo Clinic, and the American Psychiatric Association. These are descriptions of mental health and are different from user posts in forums. Hence, we crawled the WebMD community forum containing comments associated with different diseases. Keywords including posts for anxiety, panic, bipolar, bipolar disorder, depression, mental health, etc. were selected and used for embeddings along with CLPsych datasets (both labeled and unlabelled). The filtered WebMD dataset consists of 13,607 comments related to psychological experiences.

Features

This sub-section lists the features used for experimental purposes. Linguistic features describing syntax and style of posts, and non-linguistic features describing metadata of posts are tested separately and in combination. Features are grouped into four categories for simplified comparison in experiments. Feature selection is performed as follows: testing is started using a single feature group as input, then a new feature group is added for each experiment, to find out the best set of features for the given experiment.

A. Metadata Features

Several non-linguistic features from a post’s metadata were included as features. The selected metadata features are days of the week (1-7), whether the author is a moderator or not (1 or 0), the time difference between post and last edited time in minutes, and the post time of day. The posting time is considered as eight time slots of 3 hours as in Altszyler et al., and we also considered the same as it gave better performance compared to 24 features.

B. Stylistic Features

These are the important features and have been used extensively in several text classification tasks. The stylistic features used in the experiments are described below.

Number of question marks: It was observed that most of the users from red and crisis categories were not confident and asked several questions while writing their posts. Thus, the number of question marks were counted.

Table 2: Class wise instances from CLPsych datasets.

<table>
<thead>
<tr>
<th>Class</th>
<th>Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crisis</td>
<td>@Sans-RO I don’t know what’s gotten over me. Increased hopelessness, increased desperation. I’m spiralling downwards, I feel like I can’t hold on. I just want to give up. I feel like my life is over. I’m curled up in bed, thinking about how I can finish.</td>
</tr>
<tr>
<td>Red</td>
<td>my dad doesn't know how i have been feeling i have been writing and listening to music most of the day today it just isn't doing much right now. i just everything hurts so much inside and i want to scream all the time i cant even explain it i am so agitated as well and i keep having panic attacks and i just feel really bad.</td>
</tr>
<tr>
<td>Amber</td>
<td>I woke up with nightmares about people from placement hurting me/ torturing me. I hate sleep. The good thing was I woke up at a sensible time so I turned on the light before getting up. I also felt safe being at my family's place and it's quiet here atm (not heaps of people like at Christmas)</td>
</tr>
<tr>
<td>Green</td>
<td>Ah no it was kind of a joke (have you seen the ‘I am forcibly removed from the premises’ meme?) But for real though internal screaming is just my Christmas setting unfortunately It’ll be over soon.</td>
</tr>
</tbody>
</table>
Number of first person pronouns: When users are severely affected by the psychological disorder, they tend to write more about themselves and used words like ‘I’, ‘me’, etc. Thus, the number of pronouns in posts were counted.

Number of words: It was observed that users from crisis and red categories wrote more than other users. In fact, some users from green category used only one or two words to post. Thus, the number of words were used as a feature.

Number of stop words: We counted the number of stop words present in each of the posts. The nltk English stop word list is considered for the present study.

C. Semantic Features

Number of lexicon matches: There are relatively fewer instances of red and crisis categories, hence it is challenging for machine learning algorithms to detect these. Therefore, we manually collected two lexicons from training datasets for red and crisis categories containing 56 and 69 words/phrases, respectively. We counted the number of words matched in each post with the red and crisis lexicons and used the number of matches as features.

Number of positive and negative sentences: The organizers provided sentences annotated with positive and negative sentiment in the posts for the CLPsych 2016, 2017 training datasets, and the test datasets were not annotated with sentiment. For all the datasets, we used VADER Sentiment Intensity Analyzer‡‡ to calculate the number of sentences with positive and negative sentiments automatically (for both training and test datasets) and used these as features. We also used Text Blob Sentiment Analysis tool§§ to identify whether the post is positive or negative or neutral and used it as a feature.

Number of positive, negative words: Sentiment word count features were proved important in the state-of-the-art severity detection systems. Therefore, several sentiment and emotion lexicons such as SentiWordNet, WordNet Affect, SentiSense Synset, Effect WordNet, AFINN dictionary, NRC Word-Emotion Association Lexicon, Taboada adjective list and Whissell dictionary were used to identify positive and negative words in posts and are used as features.

Number of positive and negative emoticons: Similarly, a happy and sad emoticon list was created manually containing 58 and 59 emoticons, respectively. The number of positive and negative emoticons present in the posts were used as features.

D. Embedding Features

Embedding methods have gained immense popularity in the last few years and it has been successfully used in several text mining and NLP tasks. These are easy to implement and capture the semantic information of words or documents. For the present study, we considered both word and document embeddings.

Word Embeddings: Pre-trained word embedding model Google N-Grams (News corpus) and GloVe (dimension = 200) were used for our experiment. Additionally, a word2vec model (dimension = 100) was built on corpus collected from WebMD, labeled and unlabelled training CLPsych forum data. The word vectors were created by taking the normalized summation of the vectors of each word in the posts (other than stop words and punctuation), which were present in the vocabulary of the pre-trained model ($\vec{v}_p^t = \frac{1}{N} \sum_{i=1}^{N} \vec{w}_i^t$). Here, $\vec{v}_p^t$ is a vector for each post, $w_i^t$ is the $i^{th}$ word that post; $N$ is the number of words present in GloVe or Google n-gram or word2vec. The gensim¶¶ package was used to build all word and document embeddings.

Document Embeddings: Document embeddings are popular in the NLP community and have been used in several NLP tasks. We used Doc2Vec (dimension = 100) implemented in gensim. Doc2Vec model was trained on same datasets as word2vec.

Latent Semantic Analysis: LSA is a technique for creating a vector representation of a document, similar to document embeddings or Doc2Vec. It has been successfully used for developing several applications in NLP. We used the LSA

‡‡http://www.nltk.org/howto/sentiment.html
§§https://textblob.readthedocs.io/en/dev/
¶¶https://radimrehurek.com/gensim/
System Architecture

We combined the subject and body of each post. The posts were pre-processed before creating word/document embeddings. The pre-processing step includes tokenization, removal of stop words and punctuation, identifying lemma of each word. During the development phase, we divided our training set into 80-20% split for tuning the parameters. Finally, two types of systems were developed based on the datasets, a) the classifier trained on CLPysch 2017 training dataset and tested on CLPysch 2017 test dataset, b) the classifier trained on CLPysch 2016 training dataset and tested on CLPysch 2016 test dataset.

Previously, most tasks on CLPysch had been performed using SVM or Random Forest, and SVM outperformed all other classifiers\(^{14}\). Hence, we implemented SVM for the present task. Deep learning models are quite popular these days and outperformed in many tasks in NLP. Deep learning was not implemented on CLPsych datasets until now. Thus, we implemented a few popular deep learning algorithms as mentioned next. We developed different systems based on the different feature combinations using SVM and identified the best performing feature set. The best performing feature set was used with different classification techniques such as SVM, ensemble based voting, convolutional neural network (CNN), and long short-term memory (LSTM). The training and test datasets were not well distributed among different classes. For imbalanced datasets, it was observed that the multi-level classification performs well\(^{27}\). We also used multi-level ensemble based voting method where at first, we classified urgent (combined red, crisis) and non-urgent (combined amber, green); then classified urgent and non-urgent results into each separate categories.

LSTM and CNN models are trained using the combined features. The main goal of this experiment was to examine how LSTM and CNN are performing on handcrafted features. These models are trained on CLPsych 2016 and 2017 separately and tested on the corresponding test datasets. However, deep learning models are typically trained with raw data rather than providing handcrafted features. Thus, we implemented a deep learning model using just the word and character embeddings. Word and character embeddings were trained separately using LSTM, and then both outputs are concatenated. The concatenated output is then fed to a BiLSTM followed by a densely connected layer with four ReLU activation representing each of the four classes. Finally the output is passed through a softmax activation function. The detailed BiLSTM architecture is provided in Figure 1.

Experimental Details

Each feature was normalized by dividing it with the maximum value of the corresponding feature input range. Polynomial kernel of degree three was used in SVM. For ensemble voting, the ensemble based voting classifier is constructed using polynomial SVM (degree=3), Random Forest (maximum depth=3) and Logistic Regression. For deep learning algorithms, the adam optimizer was used to minimize the categorical cross entropy loss function. The training data was divided into a batch size of 50 samples. Each of the neural networks was trained for 25 epochs. For LSTM and BiLSTM, the models are initialized with 200 hidden layers and dropout of 0.3. The ReLU activation function was used for each layer and softmax was performed in the output layer for classification.
For CNN, two CNN layers followed by maxpooling over $2 \times 2$ grid and a dropout of 0.5 were inputted to a ReLU activation function. Finally, softmax was performed in the output layer for classification. The deep learning models are implemented using Keras.

**Result Analysis and Discussion**

We have considered different evaluation metrics for comparing our results. **Micro-averaged F-score**: the F-score of all categories as well as single categories. **Macro-averaged F-score**: the average F-score among crisis, red and amber labels. **Flagged vs. non-flagged F-score**: the F-score of flagged (crisis + red + amber) and non-flagged (green) labels. **Urgent vs. non-urgent F-score**: the F-score of urgent (crisis + red) and non-urgent (amber + green) labels.

**Results**

Initially, several SVM-based systems were developed using different feature combinations. Detailed comparisons between various features for CLPsych 2016 and 2017 datasets are provided in Table 3. A maximum micro-average F-score of 0.71 was obtained for CLPsych 2017 test datasets using all feature categories (metadata, stylistic, semantic, and word2vec). The word2vec embeddings trained on domain-specific features performed better compared to other embeddings, and this word2vec embedding was created using the training dataset, unlabeled corpus provided by organizers, and WebMD forum data. For CLPsych 2016 test datasets, the maximum F-score of 0.76 was obtained using similar features. It may be observed that combining embedding features to metadata, stylistic and semantic feature improved performances drastically (by at least 11% in terms of micro-average F-score) for CLPsych 2017 test dataset, while not much improvement was observed for the CLPsych 2016 test dataset. The metadata, stylistic, and semantic features performed well in the case of the CLPsych 2016 datasets and achieved micro-average F-score of 0.70.

There was only one instance from crisis category in CLPsych 2016 test dataset and most of the features were not able to identify it. The GloVe and word2vec feature based systems were able to identify the instance from the crisis category. However, these systems also mis-identified another instance from red category as crisis. Though there is

| **Table 3**: System performances of using different feature sets and SVM (M: Metadata feature, St: Stylistic feature, Se: Semantic feature, E: Embedding feature, GN: Google N-gram) |
|---|---|---|---|---|---|---|---|
| **F-score** | **features** | **micro-** | **macro-** | **flagged vs.** | **urgent vs.** | **crisis** | **red** | **amber** | **green** |
| **CLPsych 2017 test dataset** | | **average** | **average** | **non-flagged** | **non-urgent** | | | | |
| M | 0.36 | 0.29 | 0.52 | 0.58 | 0 | 0.11 | 0.25 | 0.52 |
| M + St | 0.52 | 0.33 | 0.56 | 0.77 | 0 | 0.12 | 0.27 | 0.66 |
| M + St + Se | 0.56 | 0.45 | 0.61 | 0.79 | 0.08 | 0.14 | 0.31 | 0.70 |
| M + St + Se + E (LSA) | 0.70 | 0.55 | 0.82 | 0.81 | 0.20 | 0.48 | 0.70 | 0.82 |
| M + St + Se + E (GN) | 0.67 | 0.52 | 0.80 | 0.81 | **0.25** | 0.42 | 0.59 | 0.82 |
| M + St + Se + E (word2vec) | **0.71** | 0.56 | **0.82** | **0.84** | **0.23** | **0.51** | **0.65** | **0.84** |
| M + St + Se + E (GloVe) | 0.68 | **0.61** | 0.81 | 0.83 | **0.25** | 0.45 | 0.63 | 0.83 |
| M + St + Se + E (Doc2Vec) | 0.67 | 0.50 | 0.71 | 0.82 | 0.23 | 0.46 | 0.50 | 0.79 |
| **CLPsych 2016 test dataset** | | | | | | | | | |
| M | 0.53 | 0.19 | 0.59 | 0.62 | 0 | 0.13 | 0.14 | 0.69 |
| M + St | 0.65 | 0.28 | 0.73 | 0.80 | 0 | 0.13 | 0.18 | 0.80 |
| M + St + Se | 0.70 | 0.46 | 0.75 | 0.81 | 0 | 0.46 | 0.40 | 0.81 |
| M + St + Se + E (LSA) | 0.72 | 0.49 | 0.78 | 0.91 | 0 | 0.56 | 0.47 | 0.83 |
| M + St + Se + E (GN) | 0.71 | 0.47 | 0.76 | 0.89 | 0 | 0.55 | 0.49 | 0.82 |
| M + St + Se + E (word2vec) | **0.76** | **0.67** | **0.82** | **0.92** | **0.67** | **0.60** | **0.54** | **0.86** |
| M + St + Se + E (GloVe) | 0.73 | 0.59 | 0.79 | 0.90 | **0.67** | 0.48 | 0.35 | **0.86** |
| M + St + Se + E (Doc2Vec) | 0.73 | 0.36 | 0.77 | 0.90 | 0 | 0.47 | 0.27 | 0.85 |
Table 4: Performances of different systems using fixed features and embeddings (word and character) (EV: Ensemble voting, EVML: Ensemble voting multi-level)

<table>
<thead>
<tr>
<th>models</th>
<th>micro-average</th>
<th>macro-average</th>
<th>flagged vs. non-flagged</th>
<th>urgent vs. non-urgent</th>
<th>crisis</th>
<th>red</th>
<th>amber</th>
<th>green</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.71</td>
<td>0.56</td>
<td>0.82</td>
<td>0.84</td>
<td>0.23</td>
<td>0.51</td>
<td>0.65</td>
<td>0.84</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.75</td>
<td>0.61</td>
<td>0.82</td>
<td>0.86</td>
<td>0.40</td>
<td>0.61</td>
<td>0.76</td>
<td>0.85</td>
</tr>
<tr>
<td>CNN</td>
<td>0.73</td>
<td>0.58</td>
<td>0.84</td>
<td>0.85</td>
<td>0.35</td>
<td>0.55</td>
<td>0.67</td>
<td>0.84</td>
</tr>
<tr>
<td>EV</td>
<td>0.78</td>
<td>0.68</td>
<td>0.85</td>
<td>0.88</td>
<td>0.44</td>
<td>0.68</td>
<td>0.77</td>
<td>0.85</td>
</tr>
<tr>
<td>EVML</td>
<td><strong>0.80</strong></td>
<td><strong>0.69</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.50</strong></td>
<td>0.68</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.78</td>
<td>0.68</td>
<td>0.87</td>
<td>0.88</td>
<td>0.44</td>
<td>0.68</td>
<td>0.80</td>
<td>0.85</td>
</tr>
</tbody>
</table>

CLPsych 2017 test dataset

<table>
<thead>
<tr>
<th>models</th>
<th>micro-average</th>
<th>macro-average</th>
<th>flagged vs. non-flagged</th>
<th>urgent vs. non-urgent</th>
<th>crisis</th>
<th>red</th>
<th>amber</th>
<th>green</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.76</td>
<td>0.67</td>
<td>0.82</td>
<td>0.92</td>
<td>0.67</td>
<td>0.60</td>
<td>0.54</td>
<td>0.86</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.85</td>
<td>0.75</td>
<td>0.82</td>
<td>0.92</td>
<td>0.67</td>
<td>0.78</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td>CNN</td>
<td>0.79</td>
<td>0.71</td>
<td>0.82</td>
<td>0.92</td>
<td>0.67</td>
<td>0.67</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>EV</td>
<td>0.85</td>
<td>0.78</td>
<td>0.84</td>
<td>0.92</td>
<td>0.67</td>
<td>0.78</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td>EVML</td>
<td><strong>0.86</strong></td>
<td><strong>0.79</strong></td>
<td><strong>0.87</strong></td>
<td><strong>0.96</strong></td>
<td><strong>0.67</strong></td>
<td>0.84</td>
<td>0.77</td>
<td>0.90</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.85</td>
<td>0.79</td>
<td>0.85</td>
<td>0.93</td>
<td>0.67</td>
<td>0.82</td>
<td>0.75</td>
<td>0.90</td>
</tr>
</tbody>
</table>

CLPsych 2016 test dataset

misclassification, it solved our purpose by identifying red category as crisis category as moderators are assigned to urgent posts as soon as possible. Metadata and stylistic features were not enough for identifying the crisis categories for both datasets. The dominance of the green category may be observed from the individual F-score for each category. For all of the situations, the F-scores are higher for the green category than others.

Further, we implemented different machine and deep learning algorithms with metadata, stylistic, semantic, word2vec embedding features on CLPsych 2016 and 2017 datasets. The detailed system performance is provided in Table 4. The maximum micro average F-scores of 0.80 and 0.86 were obtained for CLPsych 2017 and 2016 test datasets, respectively using multi-level ensemble voting classifier. The BiLSTM with raw input (word and character embeddings) achieved equivalent scores as compared to the ensemble voting multi-level (EVML) classifier. For CLPsych 2017 dataset, the BiLSTM achieved better F-score for the amber category.

Discussion

Our study proposes an NLP system based on a multi-level ensemble voting classifier to identify severity in forum data. The system achieved better performance compared to other approaches such as LSTM, CNN, BiLSTM, and ensemble voting. We find that the overall performance of combined features and trained word2vec embedding outperforms other feature combinations. Most state-of-the-art systems were not able to identify urgent posts due to the imbalanced training datasets. The lexicon-based features helped our systems to identify more urgent (red and crisis) posts, which increased the number of false negatives. However, this system is more useful as no urgent posts should be missed.

The BiLSTM with only embeddings input performed fairly compared to multi-level ensemble voting approach. One of the main reasons may be that there is fewer training instances. The deep learning is known to be a data-hungry architecture and requires sufficient data to support its complex architecture fully. Deep learning-based systems may perform better if more data is provided in the future.

Our current system on CLPsych 2016 test dataset outperformed the system of Kim et al.\textsuperscript{12}, which achieved the maximum F-score of 0.85 using TD-IDF features and SVM. In the case of CLPsych 2017 datasets, the maximum macro averaged F-score of 0.462 was achieved by the system of Altszyler et al.\textsuperscript{14} using TF-IDF of weighted unigrams, word2vec (trained only on CLPsych training data), and ensemble-based classifiers. Our proposed system outperformed the above system in terms of macro averaged F-score. However, the F-score of flagged vs. non-flagged is an important evaluation criterion considered by the task organizers because it measures the system’s capability to identify a post that needs moderators attention. The system developed by Altszyler et al.\textsuperscript{14} achieved maximum flagged vs.
non-flagged F-score of 0.905 and our proposed system achieved 0.89.

The key contributions of this paper are listed below.

- Unlabelled psychological forum corpus from WebMD was extracted and then used to enhance embeddings with CLPsych labeled and unlabelled datasets.
- Multiple machine and deep learning algorithms were trained to predict the severity of posts in CLPsych data.
- Multi-level ensemble based voting classification was performed to remove bias due to the fewer number of posts with high severity. This system outperformed all other systems including state-of-the-art systems in terms of micro averaged F-score.

Conclusion

In this paper, we designed an NLP based system to detect patient mental health severity from forum data extracted from ReachOut. We also applied different embedding features and found that \textit{word2vec} performed well among all other embeddings. A multi-level ensemble-based voting classifier performed well for classifying the \textit{flagged} posts. We achieved better results in classifying the severity in both CLPSych 2016 and 2017 datasets compared to the all state-of-the-art systems developed to date. We believe that this system may be used to identify psychological issue severity from posts in other social networks like Facebook, Twitter, etc. In the future, we will explore different techniques for removing bias in the datasets. Different advanced deep learning algorithms and character embeddings may be explored for these datasets in further studies.

Acknowledgements

This project is mainly supported by the Center for Big Data in Health Sciences (CBD-HS) at School of Public Health, University of Texas Health Science Center at Houston (UTHealth), and partially supported by the Cancer Research and Prevention Institute of Texas (CPRIT) project RP170668 (KR, HW) as well as the Patient-Centered Outcomes Research Institute (PCORI) project ME-2018C1-10963 (KR). We also thank the organizers of CLPsych 2016 and 2017 for providing datasets.

References