Mental Health Classification with Social Media

Lei Xian, Jiahao Xu, Yang Shi
Department of Computer Science
University of Georgia
Athens, GA 30602
{lei.xian,jiahaoxu,yang.atrue}@uga.edu

Abstract

In this project, we focus on using machine learning models to classify the potential mental health patients from the posts in social media. Samples are scrapped from Reddit, and manually labeled for both detection classification of potential patients. The testing accuracy of 91.74% in two class classification (detection), and 80.17% in four class classification are achieved through our experiments. One model that we apply in this project features the attention mechanism, which helps researchers with an importance on words when the model makes the classifications. We also provide a word cloud analysis for a higher level comparison for the key words in different potential mental illness and normal posts. With more understandings of the features and attentions of the mental health related posts, we look forward to bring the clinical natural language processing research area into a more detailed discussion in the usage of weighted words and attentions in the future.

1 Introduction

Mental health has been a major concern for both individuals and the society through recent years. According to the Suicide Facts report [1] published by U.S. Centers for Disease Control and Prevention, suicide has been one of the highest ten death causes of the country in 2013. Among the suicide cases, 80% of the people are with mental health issues [2]. The diagnosis, comparison and data analysis of the mental health issue would be an important and meaningful research topic to look into for this course. As there is a limitation for the patient privacy and thus the acquirement of the real diagnosis data is not possible, we use the social media data as the data resource of this project.

The target of this project is to help prevention and early detection of mental health issues on reddit posts. Specifically, Depression [3], Post traumatic stress disorder (PTSD) [4], and bipolar disorder [5]. Note that there are no evidence that the posts shown in the forum are from the patients, nor the identity of the posts are confirmed by any medical agencies. However, since there is a severe shortage of psychiatrists and mental health care providers in the United States [6], it would be hard to discover the potential patients through formal diagnosis. We find that there are some existing works focusing on detecting depressions on Reddits [7], and assume that the detecting techniques is able to help the early prevention of mental health issues on Internet. We have the contributions of the project listed below:

- We introduce attention mechanism [8] in mental health issue detection model, and achieve a classification accuracy of over 90% in detecting the illness and an accuracy of over 80% in diagnosing the illness type.
- By applying the attention mechanism to the model, we reveal the key words of each type of illness and potentially serve as a higher level of features in future research.
- Several baseline models are compared with the attention based model to evaluate the performance of the detection and diagnosis.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

- We also generate word clouds to visualize the hot words of normal posts and illness-suspected posts for another high level comparisons on the data distributions, which will also benefit related future works.

This report is formatted as below: related works is introduced in Section 2, and following with the data introduction in Section 3. Then in Section 4, we discuss the feature extraction methods we used and the embedding we applied on the data. In Section 5, we provide the attention based model that we used, and also the other base models for comparison. We then evaluate the results in Section 6, and this report concludes in Section 7.

2 Related work

In recent years, as the social media becomes more and more active especially in teenagers, it has been recognized to be a powerful insights in to psychological area and health of individuals.One of the most relevant paper also used the data collected from Reddit, found that there is a significant transfer from mental heath subreddit to suicide subreddit. This finding is a strong indication that mental health identification is contributed to suicide early detection.[9] Since social media is recorded in the present and preserved, it minimizes the hindsight bias sometimes induced by retrospective analyses. The rich repository of social media data also allows for the discovery, tracking, and perhaps forecasting of risk attributes longitudinally. Not only for the observation and insight, social media may also provide support may be extended to vulnerable communities.

Another paper also presents the strong evidence between mental health and suicide. Contacted with primary care providers in the time leading up to suicide is common. While three of four suicide victims had contact with primary care providers within the year of suicide, approximately one-third of the suicide victims had contact with mental health services.About one in five suicide victims had contact with mental health services within a month before their suicide.[10]

Based on previous analysis work, in this project, we will further analyze the social media data to (1) generate a word cloud as a key analysis feature and also a visualization for different health issue groups, (2) diagnose mental illness based on the social media data, and (3) study the commodity of different mental health issues.

3 Data

Data will be collected from Reddit (`https://www.reddit.com`) via the Reddit official developer API (`https://www.reddit.com/dev/api/`). Posts in Reddit are organized by "subreddits", which are areas of interests. We are going to obtain posts and comments data from subreddits **"r/depression", "r/bipolar", "r/ptsd"**, which are corresponding to the depression disease, bipolar disorder, and post traumatic stress disorder (ptsd) respectively. A user is labeled to be suffering from a disease if he/she sent two or more posts in the corresponding subreddits. In addition, a control group data for individuals who did not sent posts to the above-mentioned subreddits will be collected as well.

Since Reddit does not enforce the real name rule, we refer "users" to "user accounts" in our work [9].

3.1 Data Collection

The data is collected through *praw* (Python Reddit API Wrapper) package [11]. Using the API access of Reddit, it can scrape data from forums in reddit.

The data is from the subforums of *depression*, *pstd*, *bipolar*. From each of the subreddits, we collected 200 posts and manually labeled them. The target of the labeling is to ensure that none of the comforting messages are counted towards a post from patients.

To make a comparison, posts from normal and unrelated subreddits were collected and labeled as normal. The posts we are collecting are from *learnprogramming*, *movies*, *politics*, *fun*, *nba*, and *finance*. After manually labeling, we selected 100 samples from each of the subreddits of patients and 50 samples from each of the unrelated subreddits.

As none of us is a licensed clinical psychologist, there could be systemetic bias in our dataset.

4 Feature Extraction

4.1 Data Cleaning

First, we converted the data to lowercase to make the text to be consistent. Then, the lemmatization technique was applied, which could reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. It was implemented by WordNetLemmatizer package provided by nltk [12]. Afterward, we removed the noise, i.e. the punctuations, url links, and stop words.

4.2 Features

4.2.1 Embedding

The embedding is achieved through *GloVe* embedding vector with 300 dimension [13]. Embedding vectors enable the word representations to sit in a latent space that has a meaningful distance for each of the word. For example, Fig. 1 shows the distance between pairs of words projected in a 2-D space.

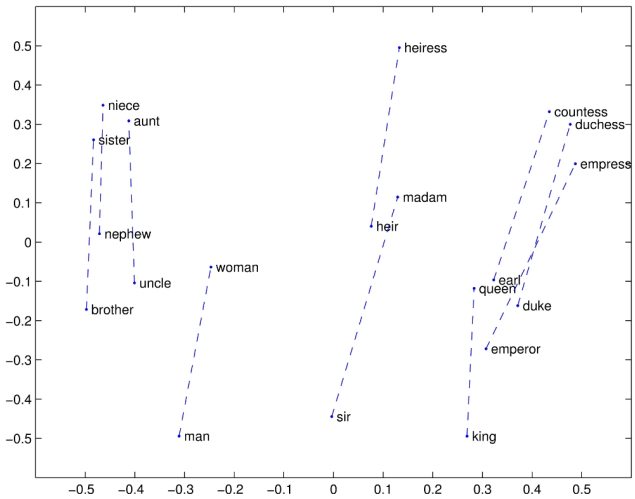


Figure 1: The distance between pairs of words projected in a 2-D space by using *GloVe* embedding vector.

4.2.2 Count Vector

Count vector converts a collection of text documents into a matrix of token counts. It is a matrix representation of a dataset, where each row denotes a document from the corpus, each column denotes a token (word) from the corpus, and each cell denotes the frequency count of a particular token (word) in a particular document. In this work, we implemented the *CountVectorizer* API provided by sklearn.

4.2.3 TF-IDF

Term frequency-inverse document frequency (TF-IDF) [14] reflects the relative importance of a word to the document in a collection or corpus. TF-IDF score is composed by two parts: the first part is the normalized term frequency (TF), the second part is the inverse document frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

$$TF(w) = \frac{\text{Number of occurrence word } w \text{ appears in a document}}{\text{Total number of terms in the document}} \tag{1}$$

$$IDF(w) = \ln\left(\frac{\text{Total number of documents}}{\text{Number of documents with word } w \text{ in it}}\right) \tag{2}$$

$$TF\text{-}IDF(w, j) = TF(w \text{ in } j^{th} \text{ document}) \times IDF(w) \tag{3}$$

TF-IDF can not only be applied on a word level, it can also be applied to a combination of n words together. And this is called n-gram. In this work, we implemented the *TfidfVectorizer* API provided by sklearn. As for n-gram, parameter *ngram_range=(2,3)* was used, which represents 2-gram and 3-gram are extracted.

5 Methods and Models

5.1 Label Embedding Attentive Model (LEAM)

Label embedding attentive model (LEAM) [8] is a model that uses label information to calculate an attention score for assisting the classification process. The details of the model is elaborated in Fig. 2.

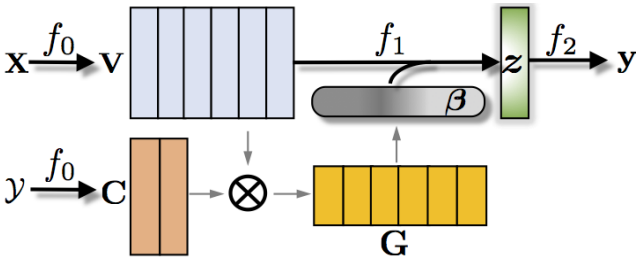


Figure 2: LEAM model structure.

In the model, the input, denoted as $\boldsymbol{x}$ and the label names, denoted as y are embedded through the same function f_0, into the same latent space, namely $\boldsymbol{V}$ and $\boldsymbol{C}$ respectively. Then the compatibility of word-label pairs are measured through cosine similarity $\boldsymbol{G}$, calculated through $\boldsymbol{G} = (\boldsymbol{C}^T\boldsymbol{V}) \oslash \hat{\boldsymbol{G}}$, where $\oslash$ represents element-wise division. The element of the row k column l in $\hat{\boldsymbol{G}}$ is calculated as $\hat{g}_{kl} = ||c_k||||v_l||$. Then a local window of $\boldsymbol{G}$ is fed into a linear layer with ReLU activation, then softmax $\boldsymbol{\beta}$ is calculated to get the aggregated representation $\boldsymbol{z}$ of the input $\boldsymbol{x}$.

5.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (ConvNets or CNNs) [15] are a category of Neural Networks that have proven very effective in areas such as image recognition and classification. ConvNets have been successful in identifying faces, objects, and traffic signs apart from powering vision in robots and self-driving cars.

Using CNN in NLP problem is the same way that a 3×3 filter can look over a patch of an image, a 1×2 filter can look over a 2 sequential words in a piece of text, i.e. a bi-gram. In this CNN model we will look at the bi-grams (a 1×2 filter), tri-grams (a 1×3 filter) and 4grams (a 1×4 filter) within the text. We can then use a filter that is $[n \times$ emb_dim$]$. This will cover n sequential words entirely, as their width will be emb_dim dimensions. Consider Fig. 3, with our word vectors are represented in the bottom. Here we have 4 words with 7-dimensional embeddings, creating a $[4 \times 7]$ "image" tensor. A filter that covers two words at a time (i.e. bi-grams) will be $[2 \times 7]$ filter, shown in convolution layer, and each element of the filter with have a weight associated with it. The output of this filter (shown in 8 feature maps) will be a single real number that is the weighted sum of all elements covered by the filter. The next step is to use pooling (specifically max pooling) on the output of the convolutional layers. The idea here is that the maximum value is the "most important" feature for determining the sentiment of the review, which corresponds to the "most important" n-gram within the review. The last layer is a dense layer with the softmax activation function which do the final classification. The output of this figure3 example is 2 classes which is a binary classification problem.

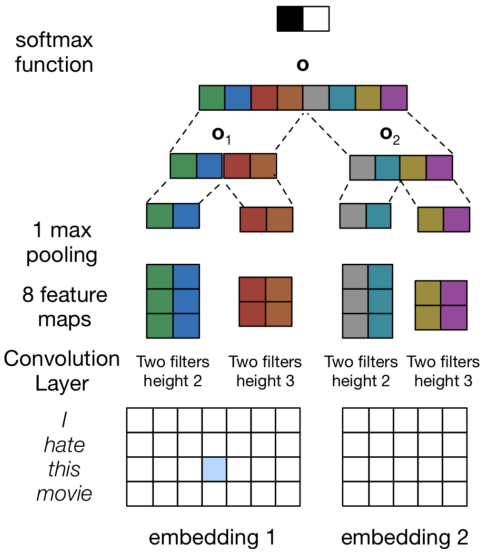


Figure 3: Schematic of CNN structures [16].

In this project, we use 3 conv layers with [3, 4, 5] three different filter sizes followed by three max pooling layers and one dense layer which means we examined 3-gram, 4-gram, and 5-grams in our model.

5.3 Long Short-Term Memory (LSTM)

Long Short Term Memory networks (LSTMs) are a special kind of recurrent neural network (RNN), which is capable of learning long-term dependencies. They were first introduced by Hochreiter and Schmidhuber [17].

Generally, RNN looks like a chain of repeating modules. Such repeating modules have a very simple structure, e.g. a single tanh layer. However, LSTMs have a similar chain-like structure, but the structure of the repeating module is different. Instead of a single neural network layer, there are 4 layers (gates), interacting in a special way, as shown in Fig. 4.

In Fig. 4, each arrow carries a vector. The flow of each arrow is from the output of one node to the input of others. The circles in pink color represent pointwise operations, e.g. vector addition. And the yellow boxes means the learned neural network layers. Merging lines denote concatenation, and a forking line denotes its content being copied and the copied information is going to different locations.

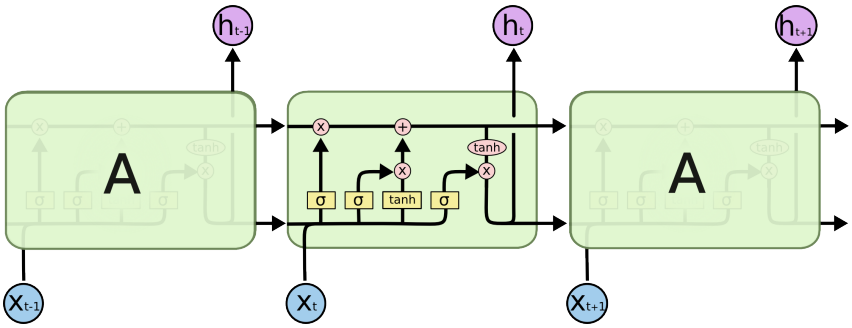


Figure 4: The repeating module in an LSTM contains four interacting layers. (Graph is taken from [18].)

Inspired by Google Translate model [19], which implements a Seq2Seq structure, consisting of multiple layers of bidirectional LSTM, we believe the bidirectional LSTM is good candidate model for this project.

Bidirectional LSTM splits the neurons of a regular RNN into two directions. One direction is for positive time (forward states), while the other direction is for negative time (backward states). Imagine there is a sentence consisting of 9 words and we want to get the information from the 5th word. As for a regular structure, it would extract the information based on the first 4 words, while the bidirectional one would extract the information from both directions, the first 4 words and the last 4 words. Therefore, we believe the bidirectional LSTM would be more informative. In our work, we implemented two bidirecional LSTM layers in our model to implement the classification.

5.4 Random Forest

Random forest [20] is an ensemble learning method for classification by constructing multiple decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees.

In general, a decision tree is built using the whole dataset considering all features, while in the random forest a fraction of the number of rows is selected at random and a particular number of features are selected at random to train on and a decision tree is built on this subset. As for the output, voting and stacking are methods to combine multiple models generated through different techniques. Voting takes the majority vote of different models. For stacking, different types of learners are combined by using a higher level meta-learner [21].

5.5 Naive Bayes

Naive Bayes [22] is a conditional probability model based on the Bayes' Theorem, and it returns the label with the highest probability as our prediction. It assumes that a particular feature in a class is independence on the presence of any other feature.

Given a problem instance to be classified, represented by a vector $\vec{x} \in X$, where $\vec{x} = \{x_1, x_2, ..., x_d\}$. and d is the number of features, we would like to classify it into one of the K classes, $y_k \in Y$

$$P(Y = y_k|X = \vec{x}) = \frac{P(Y = y_k)P(X = \vec{x}|Y = y_k)}{P(X = \vec{x})} \tag{4}$$

In Naive Bayes, we assume that each feature is an independent variable, and none of them is correlated with each other. Thus,

$$P(Y = y_k|X = \vec{x}) \propto P(Y = y_k) \prod_{i=1}^{d} P(x_i|Y = y_k) \tag{5}$$

To perform the classification, we can compute the above probability for each class y_k, and then return the label with the highest probability as our prediction [23].

$$\hat{y} = \text{argmax}_{k \in 1,2,...,K} P(Y = y_k) \prod_{i=1}^{d} P(x_i|Y = y_k) \tag{6}$$

5.6 Support Vector Machine

A Support Vector Machine (SVM) [24] is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

Whereas the original problem is stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher dimensional feature space, presumably making the separation easier in that space. The mappings used by SVM schemes are defined in terms of a kernel function [25], which is selected to suit the problem.

6 Results

In this section, we first discuss about the training process of the models, using LEAM as an example. Then the classification for 2 class and 4 class classification are introduced in the next subsection. We also provide the attention results and the word cloud results to give a better look into the project.

6.1 Training result

In this work, we implemented two classifications: detect if a user is suffering from an illness or not, and detect which class a user falls in (normal, depression, bipolar, and PTSD).

As for the binary classification training, we found that LEAM perform the best as for the validation accuracy. Fig. 5 (Left) shows that LEAM learns and the training accuracy increases dramatically up to 45 epochs (10 steps in 1 epoch). Afterward, it tends to converge and it almost approaches to 100%. In the meantime, the validation accuracy increases as the training one increases. After 25 epochs, the validation accuracy is systematically lower than the training one, which indicates that there exists some degree of overfitting. Additionally, the training loss for LEAM drops quickly at the first 20 epochs, and converge to almost 0.

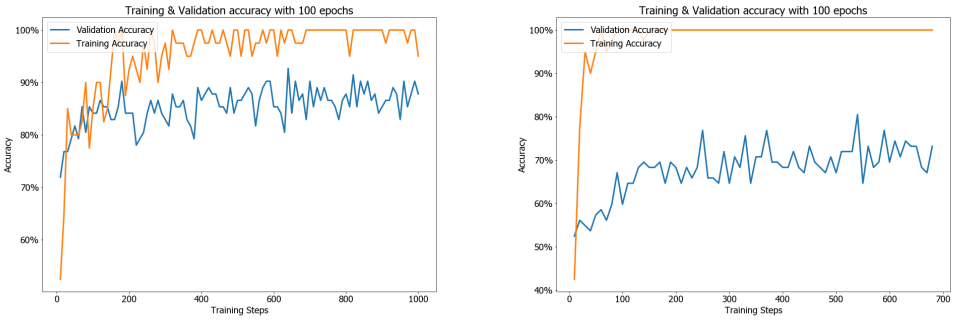


Figure 5: Training and validation accuracy. (Left) Binary classification (whether or not is suffering from an illness) for LEAM. (Right) 4-class classification (normal, depression, bipolar, or PTSD) for CNN.

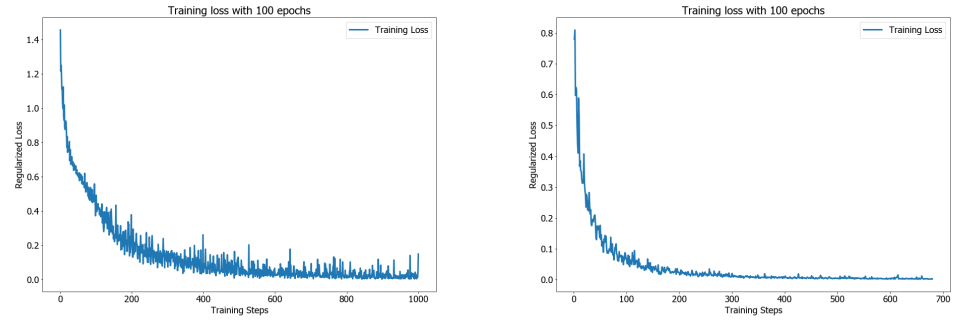


Figure 6: Training loss. (Left) Binary classification for LEAM. (Right) 4-class classification for CNN.

For the 4-class classification CNN outperforms other models, and its training accuracy reach 100% after 150 epochs, as shown in Fig. 5 (Right). However, the validation accuracy refuse approach 80% even for a long enough run. It tends to converge at 75%. We argue that there could be two main reasons for the overfitting. First, none of our models are sophisticated enough to extract the information and learn the classification in a highly accurate way. Second, there could be some

inevitable bias in our dataset, as each sample (post from Reddit) was labeled manually and none of us is a licensed clinical psychologist. If a more powerful classifier was built or a better dataset should we have, we believe the classification would perform better with a less degree of overfitting.

In addition, we notice that the loss of CNN converges to 0 substantially faster than LEAM and LSTM, we think this may be because CNN extract the feature informaiton more effecitively. But this is still an open question, and we would like to explore more in the future study.

6.2 Testing result

Table 1: Testing result for 2 classes

Model	# of epochs	optimizer	learning rate	accuracy
LEAM	100	Adam	3.00E-04	0.917355
RF-TFIDF	-	-	-	0.917355
CNN	50	RMSprop	1.00E-04	0.909091
LEAM	50	Adam	3.00E-04	0.909091
LSTM	50	RMSprop	5.00E-04	0.909091
LEAM	100	RMSprop	3.00E-04	0.900826
LSTM	50	Adam	5.00E-04	0.900826
NB-TFIDF	-	-	-	0.884298
CNN	70	Adam	1.00E-04	0.876033
CNN	32	RMSprop	1.00E-04	0.859504
NB-ngram	-	-	-	0.818182
RF-ngram	-	-	-	0.776860
RF-wc	-	-	-	0.586777
SVM-wc	-	-	-	0.578512
NB-wc	-	-	-	0.479339
SVM-TFIDF	-	-	-	0.479339
SVM-ngram	-	-	-	0.479339

Below shown in Table 1 and able 2 are the testing results for the models we evaluated for 2 class and for class diagnosis. In each of the neural network models, different epochs and optimizers are used for classifications.

Table 2: Testing result for 4 classes

Model	# of epochs	optimizer	learning rate	accuracy
CNN	68	RMSprop	1.00E-04	0.801653
RF-TFIDF	-	-	-	0.785124
CNN	43	Adam	1.00E-04	0.768595
CNN	100	RMSprop	1.00E-04	0.768595
CNN	100	Adam	1.00E-04	0.760331
LEAM	100	Adam	3.00E-04	0.752066
LEAM	50	Adam	3.00E-04	0.727273
LEAM	100	RMSprop	3.00E-04	0.719008
LSTM	50	Adam	5.00E-04	0.719008
LSTM	50	RMSprop	5.00E-04	0.710744
RF-ngram	-	-	-	0.611570
NB-TFIDF	-	-	-	0.487603
NB-wc	-	-	-	0.479339
SVM-TFIDF	-	-	-	0.479339
SVM-ngram	-	-	-	0.479339
SVM-wc	-	-	-	0.479339
NB-ngram	-	-	-	0.462810
RF-wc	-	-	-	0.380165

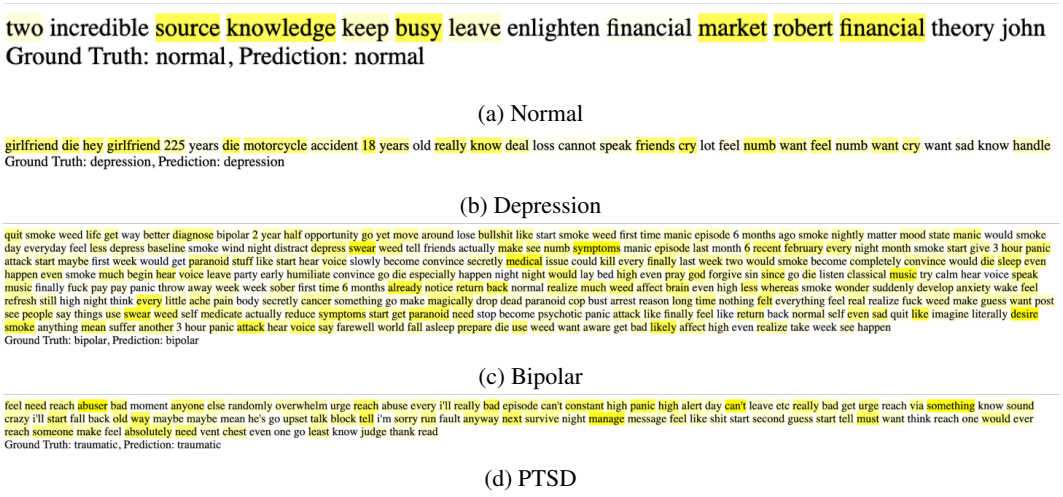
In 2 class classifications, we find that the LEAM model performs the best among the models with Random Forest model applied on TFIDF features. LEAM performs the same as a base model, which is a bit of surprise. We interpret this performance comparison as a trade-off between the application of attention information and not using TFIDF features. By sacrificing the features, attention mechanism helped the model to have a better interpretability for researchers to understand the importance in making the classification decisions. We also see a clear watershed in accuracies between most neural network based models and the base models, since the introduction of non-linearity of networks give more freedom to the parameters of the models to adjust with.

In 4 class classifications, we are not surprised to see an overwhelming performance for CNN models over other neural network models, since the versatility of the CNN based models. We still see a high performance of the combination of random forest and TFIDF in the classifications. The classification rate is about 80%, and we also see a difference between the network based models and the base models. The experiments show a competence of neural networks over base models on the project tasks.

6.3 Attention result

Fig. 7 shows some of correct predictions made by the four class LEAM model and the attention words. The darker color overlaid on the word, the more attention and weight the model gives in making the classification. The model generally makes sense that in normal, most heavy attention words are non-related words, while in depression, the model grasped the main reason that makes the depression, which is the first two words of the sentence. In bipolar disorder classifications, the words getting more attention are more likely to be a attitude or sentiment. PTSD shows a favor in sentences with words such as "abuser", which overlaps with the trauma characteristic of the illness.

It should be noticed that we only have our observations from non-experts for counseling, we only make possible interpretations for the results. This also leads to a pitfall that some classifications are not accurate, not because of the models themselves, but from the vague definitions of the label words, or the label is not accurate, as shown in Fig. 8.



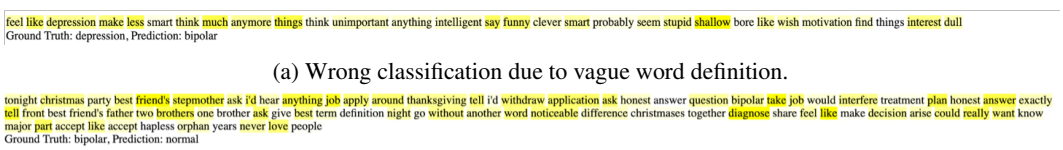
(a) Normal

(b) Depression

(c) Bipolar

(d) PTSD

Figure 7: Attention words from correct classifications.



(a) Wrong classification due to vague word definition.

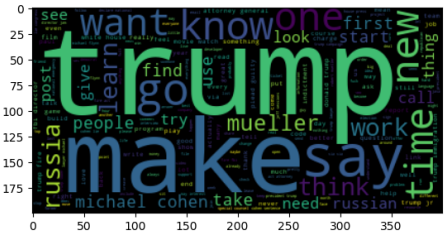
(b) Wrong classification due to label inaccuracy.

Figure 8: Attention words from wrong classifications.

6.4 Word cloud

Fig. 9 shows the word clouds for text data for each class. The larger of a word, the more frequent it occurs in the corresponding class. We observe that users suffering form an illness tend to express their emotion more often, since words "feel", "think" appear noticeably in the Fig. 9b 9c 9d. Also, "people" is a high weighted word in Fig. 9b 9c 9d. It may indicate that users with illness hope to connect with "people", or want to stay away from "people", or have some issues dealing with other "people".

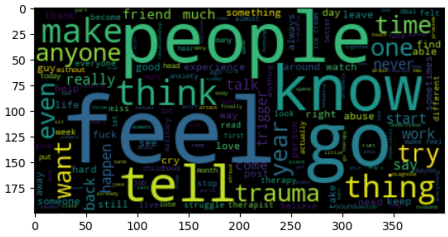
In Fig. 9b, "life", "depression", "want" are some frequent words for depression class. As for bipolar class, "take", "really", and some extreme words occur in Fig. 9c. Also, Fig. 9d shows the PTSD case, where "trauma", "try" are some key words. We have to notice that there are some background words, which are similar to the noise. They may contain some information, or they may be some "stop" words.



(a) Normal

(b) Depression

(c) Bipolar



(d) PTSD

Figure 9: Word clouds for text data corresponding to each class.

7 Discussions and Conclusion

In this study, we addressed the problem of automatically classifying content on the social media platform Reddit for the case of mental health conditions. We manually labeled data originating from several subreddits. The derived label of posts into different classes was further evaluated by applying six different algorithms. We then applied two different classification strategies, a binary classification to determine whether or not a post contains mental health related content, and multiclass classification to identify the mental health condition a post is referring to. Our results show that by applying a LEAM approach in binary classification task, we achieve an accuracy of 91.74%. Further to that, we can identify Reddit posts as belonging to one of the 4 different classes with an overall precision of 80.17%. Taken in conjunction, these results suggest that we can reliably identify mental health content and determine the mental health type that is further referred to in the post.

Our results for the both classification tasks are satisfactory, and provide some interesting insights. Especially the attention words gives some interesting insights to know more about how does people behave in social media. And give us more information about the differentiation between different mental health conditions. One limitation of our approach is the manually labeled data need to be further verified by some experts. Since none of us has public health background.

In conclusion, our suggested method is applicable to the identification of posts relevant to a mental health subreddit as well as the identification of the actual mental health conditions they relate to. In

future work, we aim to further improve the classification into mental health illness types (multiclass classification), address inter-related types, increase the coverage of mental health types. Add more syntactic features especially in binary classification with Random Forest combined TF-IDF model approach. Since the importance in sentence structure in mental health posts is already been proved by many researchers [26].

Acknowledgments

We would like to thank Dr. Quinn and the Department of Computer Science at the University of Georgia for the research project opportunity in the course of CSCI 8360 Data Science Practicum.

References

[1] Facts at a glance 2015. 4:17–0, 2013.

[2] Steven John Stack. Mental illness and suicide. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, pages 1618–1623, 2014.

[3] Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

[4] Frank W Weathers, Brett T Litz, Debra S Herman, Jennifer A Huska, Terence M Keane, et al. The ptsd checklist (pcl): Reliability, validity, and diagnostic utility. In *annual convention of the international society for traumatic stress studies, San Antonio, TX*, volume 462. San Antonio, TX., 1993.

[5] Myrna M Weissman, Roger C Bland, Glorisa J Canino, Carlo Faravelli, Steven Greenwald, Hai-Gwo Hwu, Peter R Joyce, Eile G Karam, Chung-Kyoon Lee, Joseph Lellouch, et al. Cross-national epidemiology of major depression and bipolar disorder. *Jama*, 276(4):293–299, 1996.

[6] Kathleen C Thomas, Alan R Ellis, Thomas R Konrad, Charles E Holzer, and Joseph P Morrissey. County-level estimates of mental health professional shortage in the united states. *Psychiatric Services*, 60(10):1323–1328, 2009.

[7] Judy Hanwen Shen and Frank Rudzicz. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65, 2017.

[8] Wenlin Wang Yizhe Zhang Dinghan Shen Xinyuan Zhang Ricardo Henao Lawrence Carin Guoyin Wang, Chunyuan Li. Joint embedding of words and labels for text classification. In *ACL*, 2018.

[9] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM, 2016.

[10] Jason B Luoma, Catherine E Martin, and Jane L Pearson. Contact with mental health and primary care providers before suicide: a review of the evidence. *American Journal of Psychiatry*, 159(6):909–916, 2002.

[11] `https://pypi.org/project/praw/`.

[12] `https://www.nltk.org/api/nltk.stem.html#module-nltk.stem.wordnet`.

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 2014.

[14] A. Rajaraman and J.D Ullman. *Mining of Massive Datasets*. Cambridge University Press, Cambridge, 1st edition, 2011.

[15] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, May 2016.

[16] Ye Zhang, Stephen Roller, and Byron Wallace. Mgnc-cnn: A simple approach to exploiting multiple word embeddings for sentence classification. *arXiv preprint arXiv:1603.00968*, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[18] `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`.

[19] Yonghui Wu et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[20] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, Aug 1995.

[21] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition, 2011.

[22] M. E. Maron. Automatic indexing: An experimental inquiry. *J. ACM*, 8(3):404–417, July 1961.

[23] Project 1 document in Data Science Practicum class instructed by Dr. Shannon Quinn.

[24] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[25] William H., Saul A. Teukolsky, Vetterling, William T., and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York, 3rd edition, 2007.

[26] Paul D Freund. Professional role (s) in the empowerment process:" working with" mental health consumers. *Psychosocial Rehabilitation Journal*, 16(3):65, 1993.