BT5153 Group Project Report
Detecting Depression Users in Chinese Microblogs

Zhangliang Chen [*1] **Lichen Dong** [*1] **Wanbing Li** [*1] **Xiaokun Zhang** [*1]

Abstract

Depression has become a more and more serious social issue. People are more likely to express their felling on social media, such as Weibo, than the psychological counseling agencies. In this project, we designed a method to detect depressed users based on Weibo data. We conducted feature engineering on the user text and compared the performance of 7 shallow machine learning models and 2 deep learning models. Results have shown that ensemble models tend to have better performance in this task. Finally, we chose XGBoost as our model and analyzed the feature importance in terms of accuracy and interpretability. This model has potential commercial value for hospitals, Weibo platform and social psychological counselling agencies. The code can be found via the link: https://github.com/bingaaaa/BT5153-group-project

1. Introduction

1.1. Background

Rapid social change in the past 30 years has witnessed China's tremendous development, but it is also likely to bring about a general increase in mental pressure and disease. A study of 185,787 people illustrated that the overall prevalence of depression among Chinese university students was shown to be 28.4 percent (Gao & Wang, 2020). Apart from the educated young people, across the whole society, the prevalence of depressive disorder (i.e., feel depressed most of the time for most days of the week) in China is reported to be 3.59 percent (XinhuaNet, 2017). Considering the large population in China, as well as the negative consequences of depression on daily life and even physical health, some identification and intervention work should be done to reduce the resulting harm.

One of the important places to do the work is on social media. For those who suffer from depression, it tends to be difficult for them to seek help from psychological counselling agencies actively and promptly due to personality, economic or other reasons, but they will confide their feelings on social media, such as Weibo. Studies have shown a high positive correlation between talking about suicide on Weibo and the actual risk of suicide (Cheng & Yip, 2015). In other words, when users talk about negative or even suicide-related emotions on Weibo, it is probably not a joke.

1.2. Problem Statement

Unfortunately, having posted such emotions online, the author would not be comforted immediately because the method of finding these statements relies mainly on Weibo users stumbling across them occasionally in a flood of microblogs. Last October, for example, it was after a graduate student in Dalian committed suicide that his suicide note was found on Weibo. Therefore, this project aims to build a real-time and proactive model that uses machine learning and deep learning algorithms to understand the post content on Weibo and identify Weibo users with depressive tendencies, thus enabling subsequent interventions if necessary.

1.3. Related Work

Recent literature introduced several methods related to this topic. Burnap et al. (Burnap & Scourfield, 2017) classified text relating to suicide on Twitter based on lexical and other psychological features using rotation forest algorithm. Liu et al. (Liu & Zhu, 2019) integrated risk identification and crisis management into one system called 'Proactive Suicide Prevention Online' (PSPO). Its core identification algorithm is a binary classification involving Support Vector Machine, Decision Tree and Logistical Regression based on Weibo comments. Cao et al. (Cao L., 2019) built effective suicide-oriented word embeddings to better understand the implicit meanings of words in users' posts. Huang et al. (Huang X., 2014) collected 53 Weibo users who were later committed suicide and integrated NLP techniques using HowNet word resource to build the depressed posts detection model. In this project, we acquired a dataset of over 20,000 tagged Weibo tweets and carried out cleaning and feature engineering work, which was partially inspired by the relevant methodology. Following that, we trained seven shallow models and two deep models and evaluated their performance. The innovation and application of this project will be given in the last section.

2. Dataset

2.1. Main Dataset

The main dataset is for training and validation purpose. We obtained it from Weibo User Depression Detection Dataset from Github(Li et al., 2020). According to Li et al., the dataset was crawled from Weibo using official Weibo Crawling API and has gone through the pre-processing process which deleted additional words such as '查看原图(view original image)'. The dataset contains 10,325 depressed users and 10,748 normal users, and in each row it recorded one user's basic account information and all their posts. An overview of the variables is shown in Table 1.

Table 1. Main Dataset Variable Description

VARIABLE	TYPE	EXAMPLE
LABEL	INT	1
NICKNAME	OBJECT	迷失路径
GENDER	OBJECT	男(MALE)
PROFILE	OBJECT	此人严重丧，不是绕行，谢谢
BIRTHDAY	OBJECT	1993-12-10
NUM_OF_FOLLOWER	INT	1
NUM_OF_FOLLOWING	INT	0
ALL_TWEET_COUNT	INT	10
ORIGINAL_TWEET_COUNT	INT	10
REPOST_TWEET_COUNT	INT	0
TWEETS	OBJECT	[{'TWEET_CONTENT':'有点累想休息了', 'POSTING_TIME':'2019-11-19 18:09:21', 'POSTED_PICTURE_URL':'无', 'NUM_OF_LIKES':0, 'NUM_OF_FORWARDS':0, 'NUM_OF_COMMENTS':1, 'TWEET_IS_ORIGINAL':'TRUE'}, ...]

2.2. Additional Testing Dataset

Although in the modelling section, we trained the data and tested the accuracy of the model on the validation set, we doubted that since the depression users posted highly negative posts, it's intuitively easy to classify the two types of the users in the main dataset. Therefore, additional test set should be applied to test whether the trained model is good enough to classify more general Weibo users. We crawled two test sets using the weibo-crawler code from Github, and we only collected the users' posts after Jan 1st, 2020 (Chen, 2021). The criteria to select the depressed and normal users are specified below.

Test 1 contains 200 depressed users and 200 normal users. The depressed users are from "抑郁症(depression)" supertopic on Weibo. Only diagnosed depressed patients who has high engagement index in the "抑郁症(depression)" supertopic can post, therefore it's confident to say the users that we obtained are depressed users. We found the normal users under the Weibo trending hashtag '五一国内出游有望达2亿人次(Domestic travel is expected to reach 200 million on May Day)' and we excluded the marketing accounts. This topic was quite close to the daily life, so that the user who discussed under the hashtag tend to share their personal life or thoughts instead of solely reposting Weibo. We deleted their posts containing this hashtag in order to avoid the model classifying based on detecting whether it has this hashtag in the posts.

Since many depressed users from Test 1 use the account solely for posting in the '抑郁症(depression)' supertopic, we crawled 100 additional depressed users from a Tree hole post to test how many depressed users can be successfully detected by our model. The depressed users that we obtained in Test 2 specified under the comments that they have depressive disorder, but most of them didn't use their account specifically for expressing their negative feelings either because their engagement index was not high enough for posting on the supertopic by the time they commented or they don't want to expose their situation. This test could help us to figure out how well the trained model is when the target user is not an obvious depressed user, and whether the trained model can sort out the depressed words out of all the information.

Figure 1. User example for each Test set

3. Data Pre-processing

Since our project aims on text classification, we use only post content and text related features. To extract, clean and generate desired features to fit into the model, following steps are performed on raw data set.

3.1. Data Cleaning

Extract Post Content. As described in the data structure from above section, post contents are stored in a nested variable *tweets*. So the first step is to extract post content from all the posts of a user and join them together as a flattened array.

Keep only Chinese. To remove unnecessary characters, such as alphanumeric, emojis and special characters, Uni-

code range *\u4e00 - \u9fa5* are applied to keep only Chinese characters.

Text Segmentation. To cut sentence into words, we make use of *Jieba Chinese Text Segmentation Tool* with paddle mode (Sun, 2020). 'Jieba' (Chinese for 'to stutter') is an open source Chinese word segmentation project based on python. It provided four modes of word segmentation and in this project paddle mode is adopted, which use the *PaddlePaddle* deep learning framework, training sequence annotation (bidirectional GRU) network model to achieve word segmentation. It also supports part-of-speech tagging.

Remove Stopwords. We use *stopwordsiso* Python package, which provides collection of stopwords for multiple languages, to remove stopwords. Besides that, we also add on customized stopwords by observing the dataset, e.g. 转发(retweet), 微博(post).

3.2. Feature Engineering

3.2.1. POS Count

Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text in correspondence with a particular part of speech, depending on the definition of the word and its context. The role of the word describe the characteristic structure of lexical terms within a sentence or text, therefore, we can use them for making assumptions about semantics.

In this project, we use *posseg* from *Jieba* to cut the text into labeled word. It supports 24 POS tags in paddle mode. Frequency of each POS tag in the text are calculated as new features.

Table 2. Sample POS tags supported by Jieba

TAG	POS	TAG	POS
N	NOUN	PER	PERSON
A	ADJECTIVE	LOC	LOCATION
V	VERB	ORG	ORGANIZATION
R	PRONOUN	TIME	TIME

3.2.2. Sentiment Count

Sentiment analysis, more specifically, the usage of positive and negative words, helps to understand the mental health condition of user. In this project, we use *BosonNLP Sentiment Dictionary* from *BosonNLP* (Author, 2014), a Chinese-based emotional-words resource, to tag each word in the posts as positive and negative and then calculate frequency.

Words in BosonNLP Sentiment Dictionary are collected and constructed from millions of tagged data in Chinese social media. Therefore, it covers not only formal languages but also slang and Internet languages, which will help us to better understand the text from Weibo posts.

Table 3. Sample from BosonNLP Sentiment Dictionary

WORD	SCORE	WORD	SCORE
最尼玛 (YOUR MAMA, A DIRTY WORD)	-6.704	富婆团 (RICH WOMEN GROUP)	6.375
真无语 (SPEECHLESS)	5.604	福如东海 (HAPPINESS AS IMMENSE AS THE EASTERN SEA)	5.389
救命 (HELP)	-4.165	幸运儿 (LUCKY DOG)	4.259
郁闷 (DEPRESSED)	-3.029	行大运 (STRANGE LUCK)	3.376

As shown in above table, the greater the absolute value, the stronger the emotion. In order to keep only words showing strong positive or negative emotion. We applied threshold 2 to filter out unnecessary words.

3.3. Exploratory Data Analysis

3.3.1. Frequent Word Analysis

The Word Cloud displays the top 15 frequent words that the normal users and the depressed users used in the posts. Normal users mentioned the social media related activities or tools often such as "拍(play)", "视频(video)", and they included the positive words such as "喜欢(like)","爱(love)" frequently. Depressed users mentioned negative terms such as "痛苦(pain)", and they recognized and mentioned the depressive disorder term "抑郁症(depression)","抑郁(depressed)" often in their posts. This indicates that they might record their experiences and feelings for battling with depression more in their Weibo posts.

Normal User Depressed User

Figure 2. Word Cloud for Normal and Depressed User

3.3.2. Sentiment Word Frequency Analysis

Apart from scanning through the frequent words, we conducted the sentiment word frequency analysis to analyze the frequency difference of the positive and negative words used by the two types of users. Figure 3 shows that both users posted more negative words while the normal users posted more positive words than the depressed users, and the depressed users posted more negative words than the normal users. The major difference lies in that the normal

users posted a lot more positive words than the depressed users. Figure 4 addresses a more direct comparison between the positive and negative word frequency distribution within the same type of users. It shows the positive word distribution in depressed users has a very low standard deviation. Besides, compared to normal users, the distribution of the positive and negative words between depressed users are more separate, which means depressed users posted more extreme posts than the normal users.

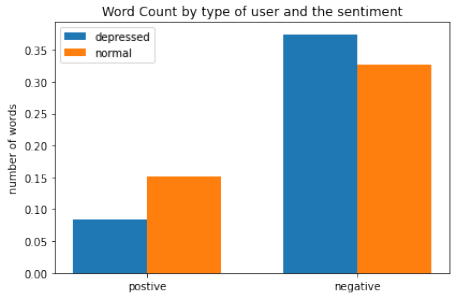

Figure 3. Sentiment Word Frequency by Types of User

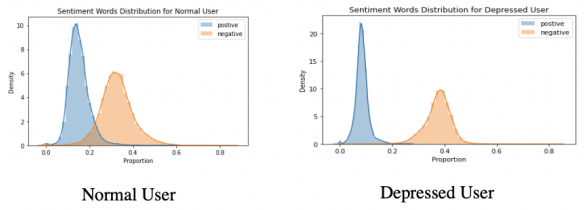

Figure 4. Sentiment Word Frequency Distribution by Types of User

4. Methodology

Fist different shallow machine learning models, convolutional neural network and BERT model are trained on training data. Then these models are evaluated on validation set to make a preliminary model selection. According to the metrics, top 3 models are selected and are further optimized.

4.1. Vector Representation

In order to convert words into numerical features, we use both sparse and dense vector representations. For sparse vector representations, we try Bag of Words and Tf-idf Vectorization, and compare the performance. Bag of Words is an algorithm that counts how many times a word appears in a document, which is only concerned with the occurrence of the word and not the order in bag. In this approach, we set 'max df' as 0.6 and min df' as 5, which means it ignore terms that have a document frequency strictly beyond the given range when building the vocabulary. The other sparse vectorization is Tf-idf vectorization and Tf-idf is short for term frequency-inverse document frequency. 'min df' is set to be 10, which means terms with absolute counts less than 10 are ignored. As for dense vector representation, we pre-train and used Word2Vec, a shallow two-layered neural network, which can capture a large number of precise syntactic and semantic word relationship by learning from a text corpus in a standalone way. The benefit of the method is that it can produce high-quality word embeddings very efficiently, in terms of space and time complexity. Here, with Word2Vec a dictionary is created for 100-dimension vectors for vocabularies that at least occur twice in the corpus.

In the following modelling, we choose Tf-idf Vectorization in terms of performance to integrate with shallow machine learnings together with other features generated from text. The Word2Vec embedding matrix is created to replace the original layer of CNN.

4.2. Modelling

4.2.1. Shallow Machine Learning

We applied various supervised machine learning to this classification problem, including statistical models and ensemble models.

Logistic Regression: Logistic regression, an extension of the linear regression model for classification problems, models the probabilities for classification problems with two possible outcomes.

Decision Tree: Decision tree builds the classification model in the form of a tree structure, breaking down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes, so it has good interpretability.

Random Forest: Random forest is an ensemble learning method for classification based on decision tree, constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes.

AdaBoost: AdaBoost, short for Adaptive Boosting, is a statistical classification, which is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms.

XGBoost: XGBoost is an optimized distributed gradient boosting designed to be highly efficient, flexible and portable, which implements machine learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting that can classification problems in a fast and accurate way. It arguably dominates structured or tabular datasets on classification problems, as it is the go-to algorithm for competition winners on the Kaggle. Thus, we

choose this ensemble method.

SGD: SGD, short for stochastic gradient descent, is an iterative method for optimizing an objective function with suitable smoothness properties, and is one of the most popular algorithms to perform optimization and by far.

KNN: KNN, short for k-nearest neighbors algorithm, is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve classification problem.

4.2.2. Convolutional Neural Network

First, the pre-trained Word2Vec is loaded to extract word vectors from. Then as the neural network model will expect all the data to have the same dimension, but in case of different sentences, they will have different lengths. Thus, the input sentences have been padded, and the embedding matrix is created and used to replace the original word embedding in CNN. Next, a concatenation of filters are slided over these embeddings to find convolutions and these are further dimensionally reduced in order to reduce complexity and computation by the Max Pooling layer. Lastly, they are followed by fully connected layers and the activation function on the outputs that will give probabilities for each class. The structure of CNN is shown in the following figure.

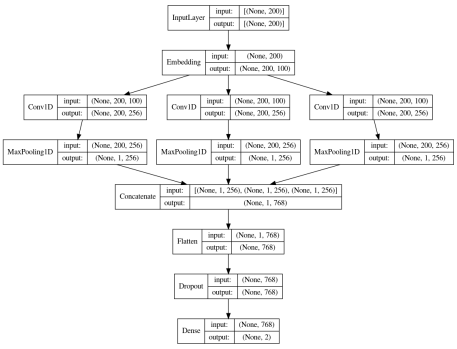

Figure 5. CNN

4.2.3. BERT

Bidirectional Encoder Representations from Transformers (BERT) has been proven to achieve obvious improvements across various NLP tasks. In this project, we also make a tentative attempt to apply BERT to do the classification task.

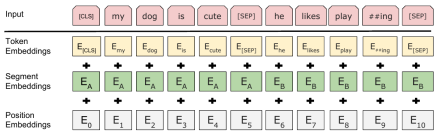

Figure 6. Input representation of the model

In this project, we use an adapted pre-trained model, i.e., BERT Whole Word Masking (BERT-wwm)(Cui et al., 2020). The major difference of this pre-trained model is that it changes the masking strategy. As shown in Figure 7, for BERT model, it divides Chinese characters into single character, which is similar to English. However, for Chinese language, usually it only has clear meaning when two or several characters make up a 'phrase'. The right part in Figure 7 comes from Baidu ERNIE, yet it adopts the similar ideas so that it is more suitable for Chinese language.

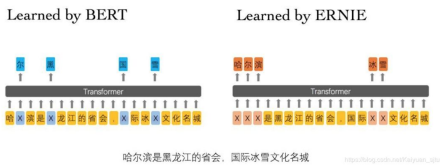

Figure 7. Input representation of the model

Due to time constraint, we do not apply fine tuning method, but to follow the programme of the classification file and define a new class to predict the probabilities of positive and negative scenarios. We use the key parameters recommended by the BERT-wwm program, with the optimum learning rate of 2e-5. In addition, we set the maximum sequence of length to be 128 and train batch size to be 32. Because the model does not require manual text segmentation, we only need to use initial cleaned data with label.

4.3. Model Selection

The training set is split into the training data and the validation data at the ratio of 0.8:0.2. We test the trained models on the validation set and made a preliminary comparison to select the candidate models. The results of how models perform on validation data are obtained and shown in the below table.

Table 4. Optimal Hyper-Parameter Values

MODEL	RECALL	PRECISION	ACCURACY	F1 SCORE
XGBoost	0.966	0.967	0.966	0.966
AdaBoost	0.954	0.955	0.954	0.954
Logistic Regression	0.952	0.954	0.952	0.952
Stochastic Gradient Descent	0.949	0.952	0.950	0.950
Random Forest	0.947	0.949	0.947	0.947
Decision Tree	0.934	0.935	0.935	0.934
CNN	0.920	0.920	0.920	0.920
K Nearest Neighbor	0.892	0.893	0.893	0.892

In this case, label 1, positive sample, represents the depressed user. To capture as many positives as possible and minimise the occurrence of false negatives, we use recall (Recall = (TP)/(TP+FN)) to evaluate modelling and prediction performance. From the table, it can be seen that XGBoost, AdaBoost and Logistic Regression are models with top 3 performance on recall. Thus, we choose these 3 models for further model optimization and evaluation with

testing data.

The CNN and the BERT models do not achieve expected performance partially due to the fact that the fine tuning method has not been applied. This also results in the very slow convergence of the BERT model. In addition, the informal language style in microblogs may not fit the pre-trained BERT model very well.

Through the results, it can be found that the ensemble models show high effectiveness in this problem. Another finding is that Logistic Regression, a statistical model with simple algorithm, is also suitable for this kind of classification with a number of features. The selected models are all shallow learnings, but with better interpretability and able to used for machine learning interpretability, such as feature importance, which is valuable for the application of machine learning in healthcare.

4.4. Hyper-parameter Tuning

The preliminary results show high accuracy and recall scores on validation set, but to better the classifiers' performance on unseen data, hyper-parameter tuning is conducted to find the optimal model architecture. The 5-fold grid search is done on the combined data set of train and validation set, as grid search is the most basic hyper-parameter tuning method, in which we simply build a model for each possible combination of all of the hyper-parameter values provided. The results of hyper-parameter tuning for XGBoost, Adaboost, and Logistic Regression are shown in the following table.

MODEL	OPTIMAL HYPER-PARAMETER VALUE
XGBOOST	LEARNING$_r ate = 0.5, n_e stimater = 75, max_d epth = 3, objective =' binary : logistic', random_s tate = 3$
ADABOOST	N$_e stimators = 500, random_s tate = 3$
LOGISTIC REGRESSION	SOLVER='LIBLINEAR', C=0.1, PENALTY='L2', RANDOM$_s tate = 3$

5. Evaluation

5.1. Performance on New Dataset

We applied the tuned XGBoost, AdaBoost and Logistic Regression model on the Test 1 Dataset that we crawled from Weibo, and the test metrics are as follows:

Table 5. Performance on Test Set 1

MODEL	RECALL	PRECISION	ACCURACY	F1 SCORE
XGBOOST	0.843	0.843	843	0.843
ADABOOST	0.830	0.831	0.830	0.830
LOGISTIC REGRESSION	0.763	0.773	0.763	0.760

The result shows that XGBoost and AdaBoost achieved high performance of over 0.83, and the two models are good enough to classify the identified and active depressed user accounts and the normal user accounts.

We applied the same models on the Test 2 Dataset that we crawled from weibo, and the test metrics are as follows:

Table 6. Performance on Test Set 1

MODEL	ACCURACY
ADABOOST	0.620
XGBOOST	0.610
LOGISTIC REGRESSION	0.460

Both AdaBoost and XGBoost only achieve around 0.60 of accuracy when predicting whether the user is depressed or not in the Test 2. This means further model should be refined if our future purpose is to sort out the depressed users who are more implicit on weibo.

5.2. Feature Importance

As there are 17,069 features in our model, studying the feature importance can sort out the features that are more significantly affect the model prediction and whether they increase the probability of being predicted as depressed users. We therefore selected the XGBoost model which has high test performance, and conducted Global Shap Value and Local Shap Value analysis to provide insights about the important features in this model.

5.2.1. SHAP VALUE

Shap Value shows how much the feature changes the prediction compared to the prediction when the feature is at the baseline value. The relationship betwee shap value and prediction probability is:

$$\begin{aligned} Odds &= exp(ShapValue) \\ p &= odds/(1+odds) \end{aligned} \tag{1}$$

We computed global shap value on the validation set to study the top 10 features that return the highest shap value. The result in Figure 8 shows that the word "共青团员","划破" are the most important features. Percentage of positive words in the posts also significantly influence the prediction, while percentage of negative words is not in the list. There are several rare words as the important features such as the philosophical related words '清俾类钞' and medical related words '震眼'. This could be the dataset includes the repost weibo which contains special terms. The summary plot displays how the value of each feature affects the prediction. The blue dots mean the specific feature value is lower and the red dots mean the feature value is higher. Figure 9 shows when the user didn't mention "共青团员" or "划破", or the percentage of positive words mentioned is high, the prediction probability value is largely decreased. However, though the lower frequency of the word "共青团员" can significantly impact the model to predict the users as normal, there are quite number of blue dots when the shap value is positive. This means that there are users who rarely mention '共青团员' and it makes the model to pre-

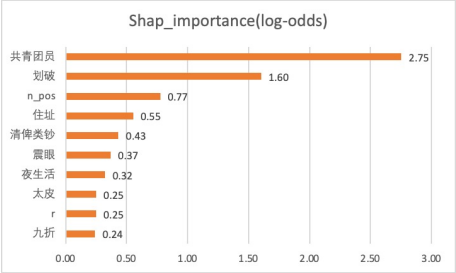

Figure 8. Top 10 Important features

dict it as a depressed user. This might be that this feature has correlation with other features which impact the model prediction. Therefore, more dependence analysis should be done in the future.

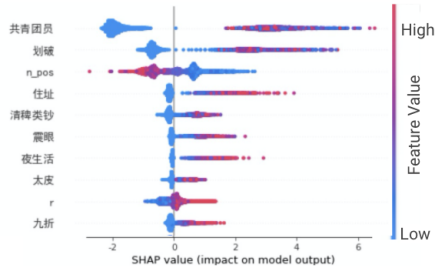

Figure 9. Summary Plot for Validation Set

Besides Global Shap Value Analysis, we conducted Local Shap Value Analysis to screen through how the features affect the model to predict specific user. We selected two particular users and their feature importance graphs are shown below.

Figure 10. Local Shap Value Analysis for Depressed and Normal User

The red block pushes the predicted probability higher and the blue block pushes it lower. The width of the block represents the importance of the feature for the specific row. The wider the block is, the more important the feature is to push the predicted probability.

The model predicted the case 1 user correctly as a depressed user. Their mention of "划破" and low percentage of positive words mentioned largely affect the model to predict it as a depressed user.

The model predicted the case 2 user correctly as a normal user. The fact that they didn't mention "共青团员" and "划破" and the high value of percentage of positive words mentioned push the model to predict it as a normal user.

To sum up the feature importance section, different from the assumption and the word frequency analysis in the exploratory data analysis part, frequents words mentioned by depressed users such as "痛苦","抑郁" are not the important features in the XGBoost model.

However, there are some observations that are aligned with the intuition and can provide insights besides serving as a predicted model to detect the depression users. The feature percentage of positive words in the posts are more important than the percentage of negative words in the posts. This means that based on the Bosonnlp Dataset that was crawled from the weibo and other forums, all the users might include some negative words in their posts, but normal users tend to have recognizable more positive words than the depressed user.

6. Application

This project has constructed a system to detect people's depression based on social media content. There are not many studies that specifically focus on people with depression. If any, they usually study serious depression associated with suicide. However, we believe it is even more of a concern of the common depression. This affects an even wider group of people, and these people still have a lot of potential to be cured and improve their situation, especially compared with those committed suicide.

To summarize, our project has four characteristics in terms of innovation. The first one is our topic. As mentioned above, we focus on a large group of people who are unfocused and vulnerable. Secondly, we make full use of high-quality data from different sources, which ensures the adaptability of our model. Thirdly, we try various methods of feature engineering and construct multidimensional features that can better reflect the nature of the content. Last but not least, our model can evaluate the degree of depression shown in microblogs based on the output probabilities, rather than a binary classification. This is useful for graded diagnosis and different recommendations and treatments for patients.

With regard to application, we hope it can be part of a public welfare program supported by the government, while it also has promising commercial prospect. Firstly, for hospitals and research institutions, this model as a supplementary diagnosis tool, can help mental health professionals to identify symptoms of depression more comprehensively and accurately. Secondly, it can become a part of AI treatment program, integrated with AI chatbot to provide early active

intervention. Due to the uneven distribution of resources, not everyone has access to professional counseling and help. However, AI psychotherapists have made great progress in recent years, and the techniques for chatbot based on human-computer interaction, which have been employed by Facebook and Wechat, are getting better and better. It is a chance for Microblog to develop the built-in AI counselling service to detect depressed users and give suitable recommendations. In terms of our findings, compared to the normal users, the depressed users more tend to post blogs at midnight, so the chatbot can provide timely intervention and accompany to users, which can be a supplement to traditional psychotherapists. In return, this service may also enhance the customer stickiness for Weibo. Lastly, for the social profit-making psychological counselling institutions, this system can help provide targeted advertisement, which will benefit to both patients and counselling institutions. The patients can be aware of their mental problems timely and find the prompt solutions. For institutions, they can be accessed to more patients, and, therefore, increase the revenue. Thus, the system based on the model is of both social and business values.

For further improvement, we can modify parameters and features based on different business application scenarios. For example, the features, such as user age, user sex, the demography of followers and followings, and the frequent posting time period, can be added to predict based on users, while the posting time can be used to make a real-time post analysis and depression detection.

7. Conclusion

Motivated by the large population of Chinese who suffer from depression, this project focuses on the identification of depressed Weibo users based on content of Weibo posts. Having finished feature engineering, we try different shallow machine learning models and deep learning models. Results have shown that on the basis of good feature engineering, shallow models can have better performance and interpretability. XGBoost is chosen as the final model which can achieve a recall over 96%. The output probabilites could be applied in depression detection in the healthcare sector; meanwhile, the model has business values for Weibo chatbot and counselling institutions.

Despite the accurate model and social and business values in various application scenarios, there are still limitations. The major concern is the privacy of Weibo users. Even if the data is voluntarily open to public, users may still be anxious and upset by the accurate analysis of their personality traits. Especially for patients with depressive disorder, they often do not want others to know their depression. But what we can do is that the analysis results need to be kept confidential and provided only to a professional authority, in order to protect the user's privacy as much as possible.

References

Author, N. N. Bosonnlp, 2014. URL http://static.bosonnlp.com/dev/resource.

Burnap, P., C. G. A. R. H. A. and Scourfield, J. Multi-class machine classification of suicide-related communication on twitter. *Online Social Networks and Media, 2, pp.32-44*, 2017.

Cao L., Zhang H., F. L. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Chen, L. , 2021. URL https://github.com/dataabc/weibo-crawler.

Cheng, Q., K. C. Z. T. G. L. and Yip, P. Suicide communication on social media and its psychological mechanisms: An examination of chinese microblog users. *International Journal of Environmental Research and Public Health, 12(9), pp.11506-11527*, 2015.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. Revisiting pre-trained models for chinese natural language processing. *A Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Gao, L., X. Y. J. C. and Wang, W. Prevalence of depression among chinese university students: a systematic review and meta-analysis. *Scientific Reports, 10(1)*, 2020.

Huang X., Zhang L., C. D. L. T. L. X. Z. T. Detecting suicidal ideation in chinese microblogs with psychological lexicons. *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, 2014.

Li, C., Zhang, Y., Wang, Y., and Wang, Z. Weibo user depression detection dataset, 2020. URL https://github.com/aidenwang9867/Weibo-User-Depession-Detection-Dataset.

Liu, X., L. X. S. J. Y. N. S. B. L. Q. and Zhu, T. Proactive suicide prevention online (pspo): Machine identification and crisis management for chinese social media users with suicidal thoughts and behaviors. *Journal of Medical Internet Research, 21(5), p.e11705*, 2019.

Sun, J. , 2020. URL https://github.com/fxsjy/jieba/blob/master/README.md.

XinhuaNet. Prevalence of depression in china stands at 3.59 pct, 2017. URL http://www.xinhuanet.com/english/2017-04/07/c_136190779.htm.