International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Analyzation of Patient-Authored Text Data and Extract Depression Symptoms using Lexical Analysis

[1]V.M.R.Krishna Rao, [2]Jalluri Vanmitha, [3]Parasa Amala Ramu, [4]Akurathi Bhuvana, [5] Chalamalasetti Vijay Daniyelu

[1]Assistant Professor, [2,3,4,5] UG Student

Department of CSE-AI & ML, Sri Vasavi Institute of Engineering & Technology, Nandamuru, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 15 April 2024

Published: 25 April 2024

Publication Issue :

Volume 11, Issue 2

March-April-2024

Page Number :

459-466

ABSTRACT

Background: Internet-Delivered Psychological Treatment (IDPT) systems have the potential to provide evidence-based mental health treatments for a far-reaching population at a lower cost. However, most of the current IDPT systems follow a tunnel based treatment process and do not adapt to the needs of different patients'. In this paper, we explore the possibility of applying Natural Language Processing (NLP) for personalizing mental health interventions. Objective: The primary objective of this study is to present an adaptive strategy based on NLP techniques that analyses patient-authored text data and extract depression symptoms based on a clinically established assessment questionnaire, PHQ-9. Method: We propose a novel word-embedding (Depression2Vec) to extract depression symptoms from patient authored text data and compare it with three state-of-the-art NLP techniques. We also present an adaptive IDPT system that personalizes treatments for mental health patients based on the proposed depression symptoms detection technique. Result: Our results indicate that the performance of proposed embedding Depression2Vec is comparable to WordNet, but in some cases, the former outperforms the latter with respect to extracting depression symptoms from the patient-authored text. Conclusion: Although extraction of symptoms from text is challenging, our proposed method can effectively extract depression symptoms from text data, which can be used to deliver the personalized intervention.

Keywords : Internet-Delivered Interventions, NLP, Tailored Intervention, Personalization, Adaptive Treatments, Adaptive Strategies, Adaptive iCBT.

I. INTRODUCTION

Internet-Delivered Psychological Treatments (IDPT) has the potential to offer evidence-based mental health treatments for a larger population using fewer resources. However, despite extensive evidence that Internet Interventions can be an effective means in the treatment of mental health morbidities, many current IDPT systems are tunnel-based, inflexible, and non interoperable. Lack of adaptability results in more dropouts and lower user adherence. Hence, it is relevant to focus on the factors associated with enhancing user adaptation towards such interventions. One way to enhance user adaptation is to make IDPT systems adaptive such that they change behavior according to several factors (user preferences, user needs, user health symptoms, user contexts, etc.). In this study, we aim to build an adaptive IDPT system by extracting depression symptoms from patient-authored text using Natural Language Processing (NLP) techniques.

Our hypothesis is based on the assumption that patients' depression symptoms are reflected in their writing when they communicate about their feelings. Based on this hypothesis, we consider that extracting depression-related symptoms from the patient-authored text should allow us to provide tailored intervention. The proposed method to extract symptoms from the patient-authored text should help people be aware of the significance of their depression and realize if they should seek medical help.

II.RELATED WORK

Funk et al. present a conceptual framework to apply NLP in digital health intervention to support automated analysis of texts authored by patients as well as messages exchanged between therapists and the patients. The study reports applying the framework to predict binge eating disorder and obtaining a result in an area under a curve between 0.57 and 0.72. However, the framework does not show how can we achieve adaptation in an IDPT environment. The feature engineering process used in the study considers the inclusion of several features, including metadata, word usage, topic models, word embedding, parts of speech, sentiment analysis, and others. In contrast, we use several word-embedding techniques and propose our word-embedding Depression2Vec. Yazdavar et al. present a method to detect depressive symptoms, based on the PHQ-9 questionnaire, from Twitter. The study uses a semi-supervised statistical model to evaluate how the duration of these symptoms and their expression on Twitter (in terms of word usage patterns and topic preferences) align with the medical findings reported via the PHQ-9 questionnaire. The work uses two different methods, Latent Dirichlet Allocation (LDA) and a proposed semi-supervised topic modeling over time (ssToT). Several studies have highlighted that the topics learned by LDA are not concrete enough to capture depressive symptom. To empower LDA shortcomings, the authors add supervision to the LDA method by using the terms that are strongly related to the PHQ-9 symptoms as the seeds to the topic clusters and guide the model to aggregate semantically-related terms into the same cluster. Similar to this technique, we use a seed term generation method. The main difference between our seeding model and theirs is that we do not use any dictionary to retrieve synonyms. Instead, we use WordNet to extract not just synonyms but also hypernyms, hyponyms, antonyms. Besides, we apply a different threshold for selecting the words for different methods. Karmen et al. (2015) used a NLP method to detect symptoms of depression from forum texts. While this work focused on keyword density to extract depressive symptoms, we concentrate on finding a more effective approach to obtain depression symptom score from patient-authored text.

III.PROPOSED SYSTEM

We propose a novel word-embedding (Depression2Vec) to extract depression symptoms from patient authored text data and compare it with three state-of-the-art NLP techniques. We also present an adaptive IDPT system that personalizes treatments for mental health patients based on the proposed depression symptoms detection technique. We are getting personalized treatments for mental health patients.

3.1 MODULES DESCRIPTION

NLTK: NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

Lemmatization: Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

Tokenization: Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

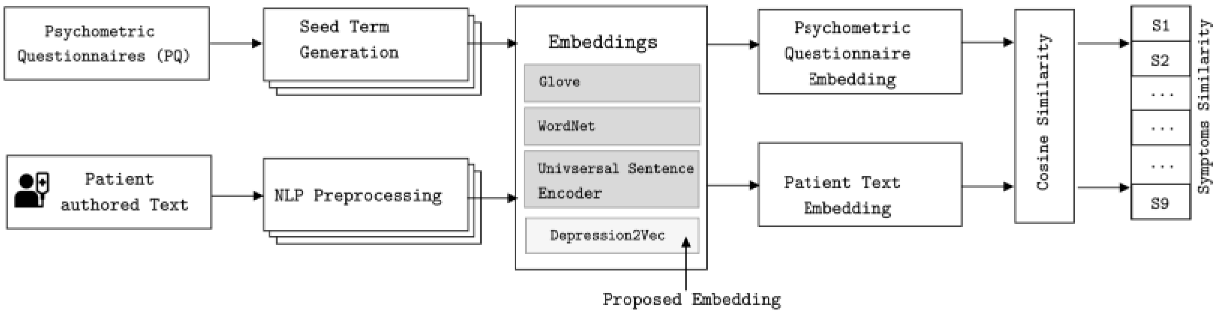
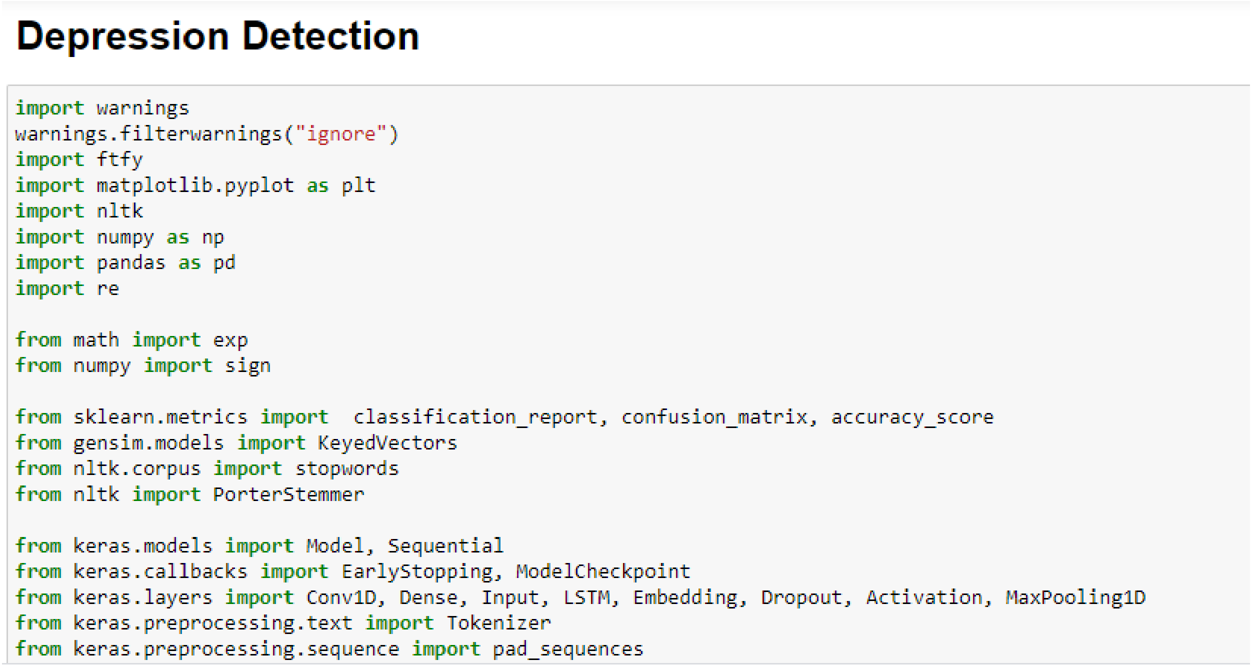


Fig 1: Flow chart for the proposed system

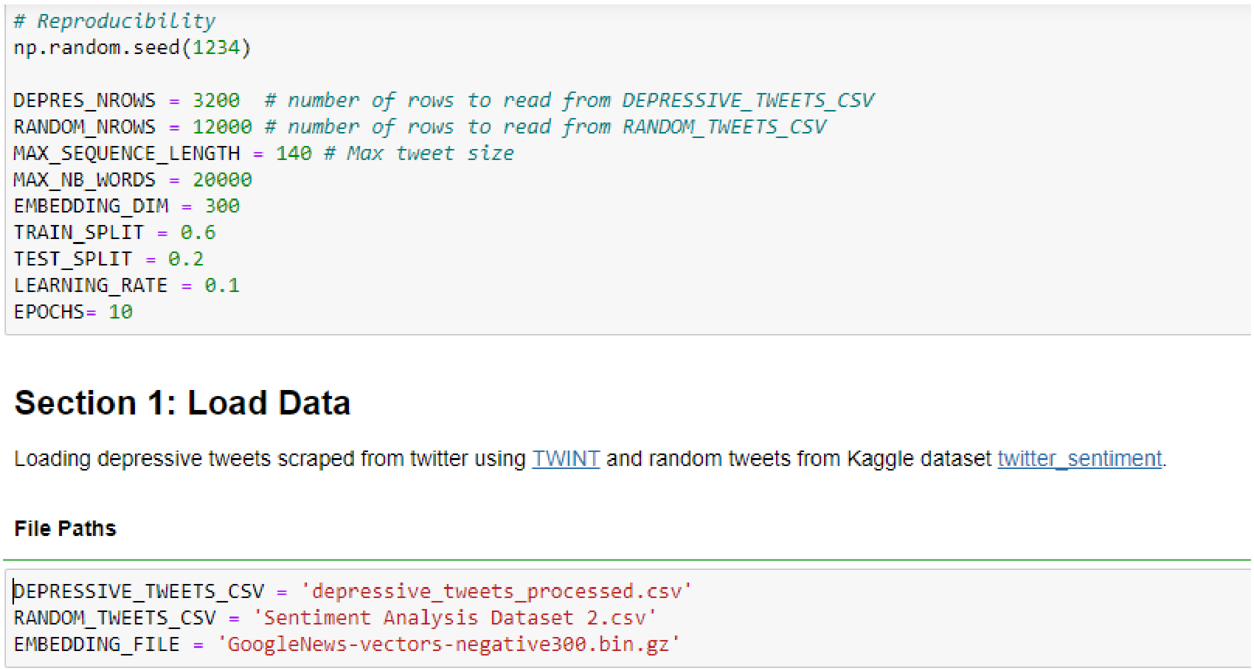
IV. RESULTS AND DISCUSSION



Depression Detection

```
import warnings
warnings.filterwarnings("ignore")
import ftfy
import matplotlib.pyplot as plt
import nltk
import numpy as np
import pandas as pd
import re

from math import exp
from numpy import sign

from sklearn.metrics import  classification_report, confusion_matrix, accuracy_score
from gensim.models import KeyedVectors
from nltk.corpus import stopwords
from nltk import PorterStemmer

from keras.models import Model, Sequential
from keras.callbacks import EarlyStopping, ModelCheckpoint
from keras.layers import Conv1D, Dense, Input, LSTM, Embedding, Dropout, Activation, MaxPooling1D
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
```

Fig 2. Results Screenshot



```
# Reproducibility
np.random.seed(1234)

DEPRES_NROWS = 3200  # number of rows to read from DEPRESSIVE_TWEETS_CSV
RANDOM_NROWS = 12000 # number of rows to read from RANDOM_TWEETS_CSV
MAX_SEQUENCE_LENGTH = 140 # Max tweet size
MAX_NB_WORDS = 20000
EMBEDDING_DIM = 300
TRAIN_SPLIT = 0.6
TEST_SPLIT = 0.2
LEARNING_RATE = 0.1
EPOCHS= 10
```

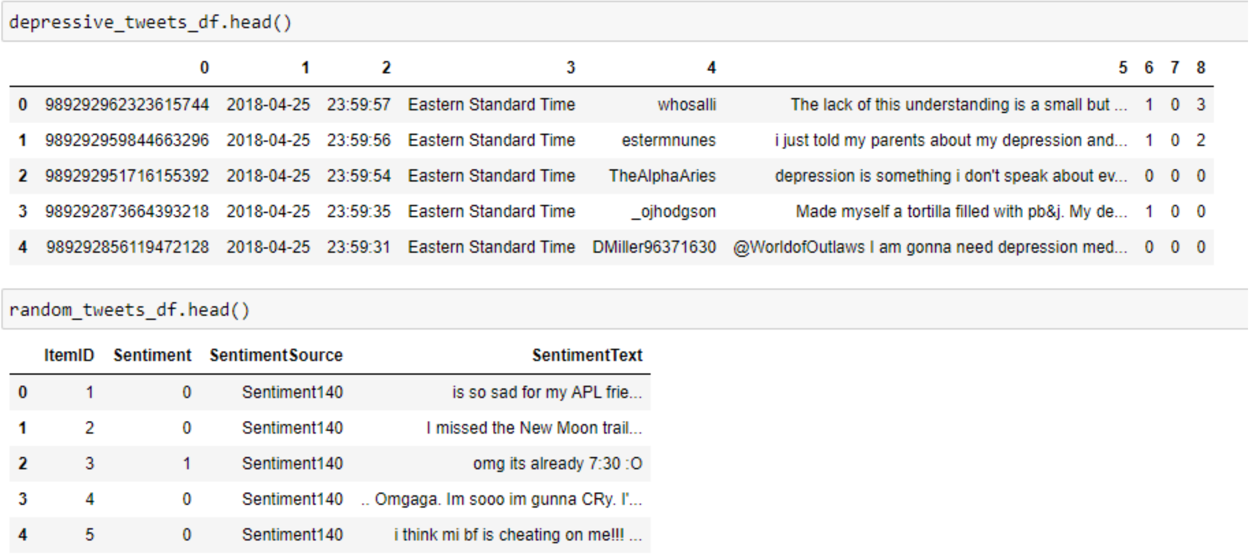
Section 1: Load Data

Loading depressive tweets scraped from twitter using TWINT and random tweets from Kaggle dataset twitter_sentiment.

File Paths

```
DEPRESSIVE_TWEETS_CSV = 'depressive_tweets_processed.csv'
RANDOM_TWEETS_CSV = 'Sentiment Analysis Dataset 2.csv'
EMBEDDING_FILE = 'GoogleNews-vectors-negative300.bin.gz'
```

Fig 3. Results Screenshot



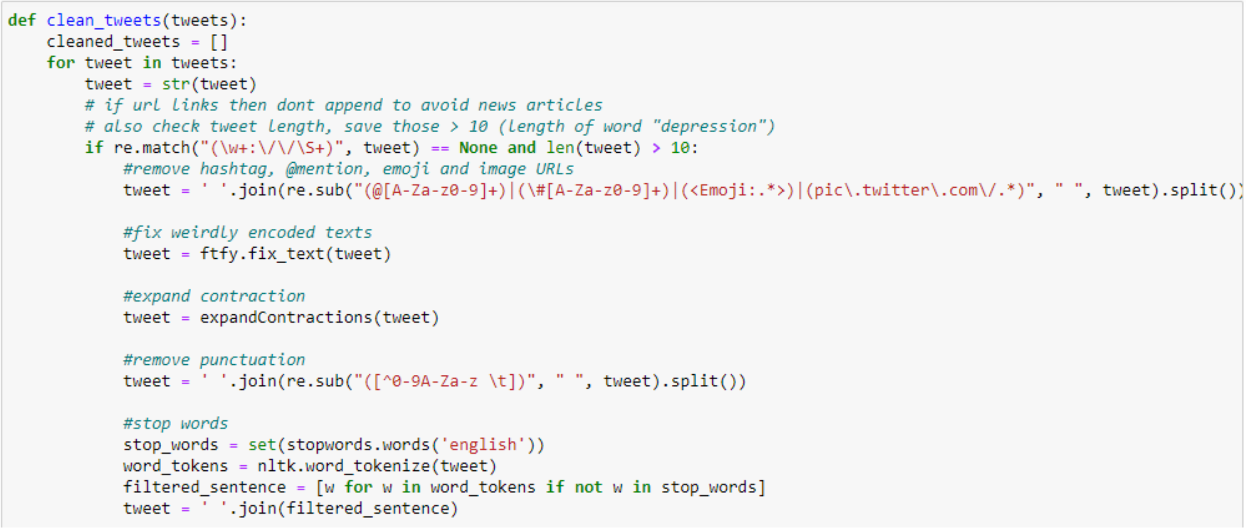
```
depressive_tweets_df.head()
```

	0	1	2	3	4	5	6	7	8
0	989292962323615744	2018-04-25	23:59:57	Eastern Standard Time	whosalli	The lack of this understanding is a small but ...	1	0	3
1	989292959844663296	2018-04-25	23:59:56	Eastern Standard Time	estermnunes	i just told my parents about my depression and...	1	0	2
2	989292951716155392	2018-04-25	23:59:54	Eastern Standard Time	TheAlphaAries	depression is something i don't speak about ev...	0	0	0
3	989292873664393218	2018-04-25	23:59:35	Eastern Standard Time	_ojhodgson	Made myself a tortilla filled with pb&j. My de...	1	0	0
4	989292856119472128	2018-04-25	23:59:31	Eastern Standard Time	DMiller96371630	@WorldofOutlaws I am gonna need depression med...	0	0	0

```
random_tweets_df.head()
```

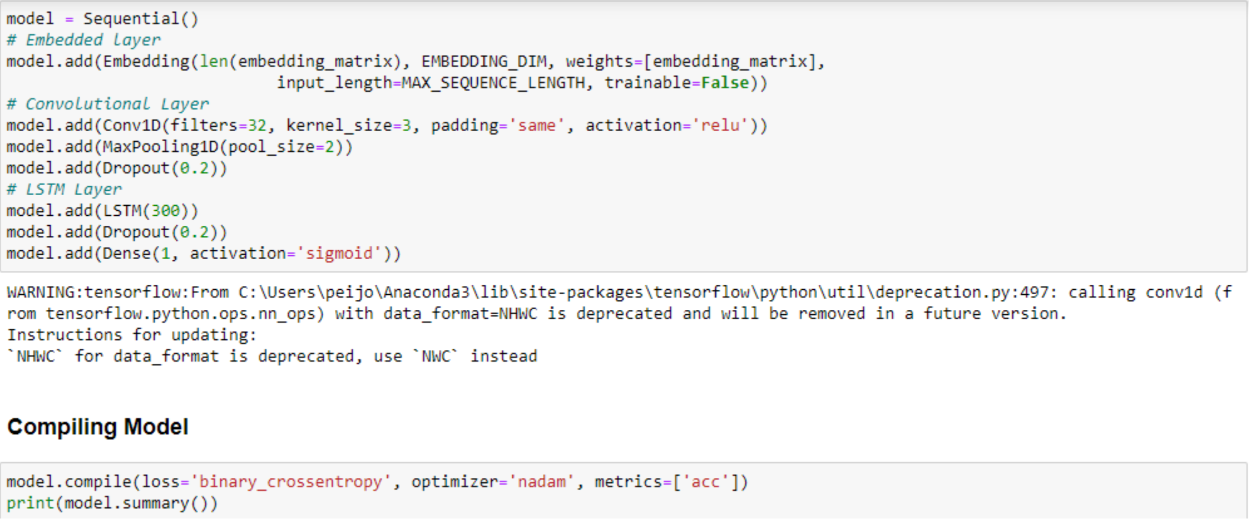
	ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140	is so sad for my APL frie...
1	2	0	Sentiment140	I missed the New Moon trail...
2	3	1	Sentiment140	omg its already 7:30 :O
3	4	0	Sentiment140	.. Omgaga. Im sooo im gunna CRy. I'...
4	5	0	Sentiment140	i think mi bf is cheating on me!!! ...

Fig 4. Results Screenshot



```
def clean_tweets(tweets):
    cleaned_tweets = []
    for tweet in tweets:
        tweet = str(tweet)
        # if url links then dont append to avoid news articles
        # also check tweet length, save those > 10 (length of word "depression")
        if re.match("(\w+:\/\/\S+)", tweet) == None and len(tweet) > 10:
            #remove hashtag, @mention, emoji and image URLs
            tweet = ' '.join(re.sub("(@[A-Za-z0-9]+)|(\#[A-Za-z0-9]+)|(<Emoji:.*>)|(pic\.twitter\.com\/.*)", " ", tweet).split()

            #fix weirdly encoded texts
            tweet = ftfy.fix_text(tweet)

            #expand contraction
            tweet = expandContractions(tweet)

            #remove punctuation
            tweet = ' '.join(re.sub("([^0-9A-Za-z \t])", " ", tweet).split())

            #stop words
            stop_words = set(stopwords.words('english'))
            word_tokens = nltk.word_tokenize(tweet)
            filtered_sentence = [w for w in word_tokens if not w in stop_words]
            tweet = ' '.join(filtered_sentence)
```

Fig 5. Results Screenshot



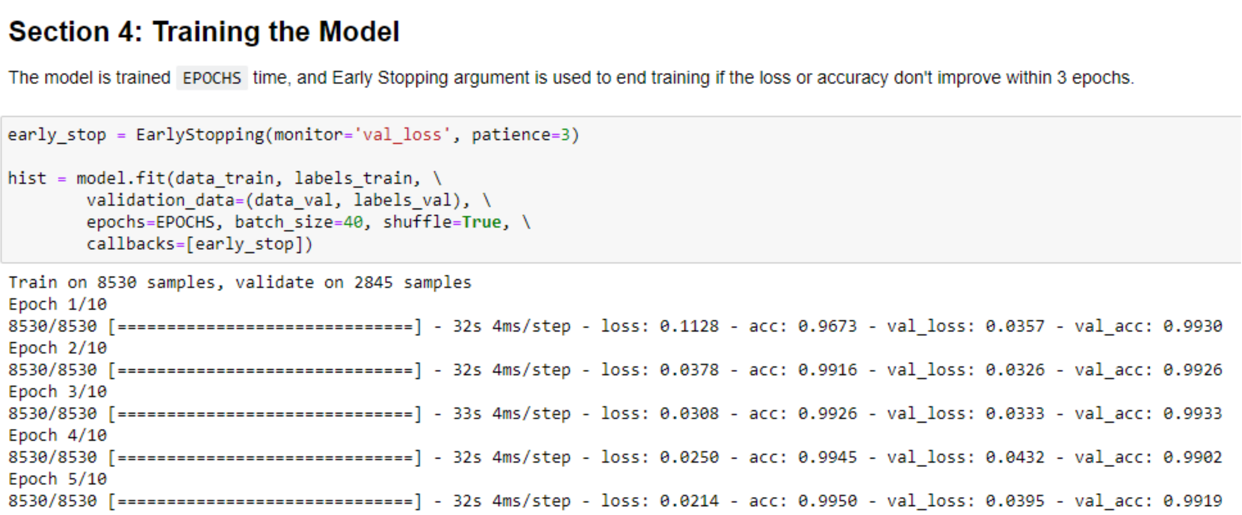
```
model = Sequential()
# Embedded Layer
model.add(Embedding(len(embedding_matrix), EMBEDDING_DIM, weights=[embedding_matrix],
                    input_length=MAX_SEQUENCE_LENGTH, trainable=False))
# Convolutional Layer
model.add(Conv1D(filters=32, kernel_size=3, padding='same', activation='relu'))
model.add(MaxPooling1D(pool_size=2))
model.add(Dropout(0.2))
# LSTM Layer
model.add(LSTM(300))
model.add(Dropout(0.2))
model.add(Dense(1, activation='sigmoid'))
```

```
WARNING:tensorflow:From C:\Users\peijo\Anaconda3\lib\site-packages\tensorflow\python\util\deprecation.py:497: calling conv1d (f
rom tensorflow.python.ops.nn_ops) with data_format=NHWC is deprecated and will be removed in a future version.
Instructions for updating:
`NHWC` for data_format is deprecated, use `NWC` instead
```

Compiling Model

```
model.compile(loss='binary_crossentropy', optimizer='nadam', metrics=['acc'])
print(model.summary())
```

Fig 6. Results Screenshot



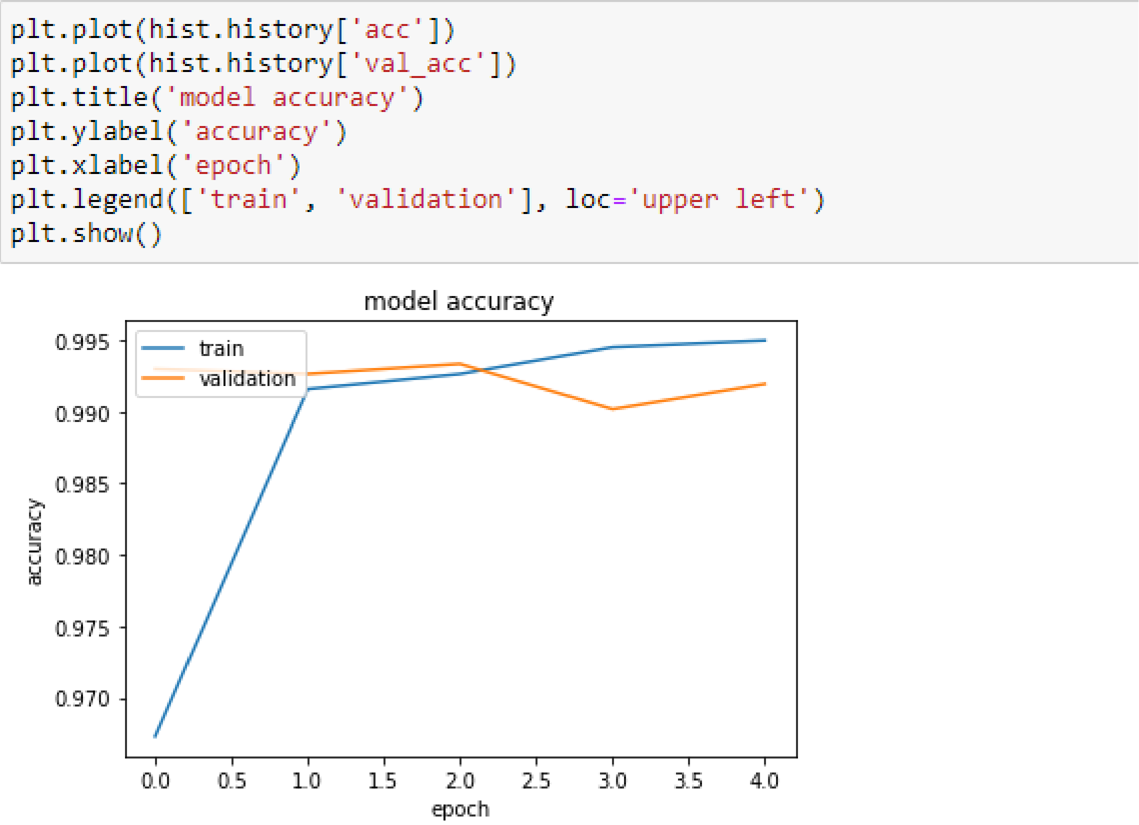
Section 4: Training the Model

The model is trained `EPOCHS` time, and Early Stopping argument is used to end training if the loss or accuracy don't improve within 3 epochs.

```
early_stop = EarlyStopping(monitor='val_loss', patience=3)

hist = model.fit(data_train, labels_train, \
        validation_data=(data_val, labels_val), \
        epochs=EPOCHS, batch_size=40, shuffle=True, \
        callbacks=[early_stop])
```

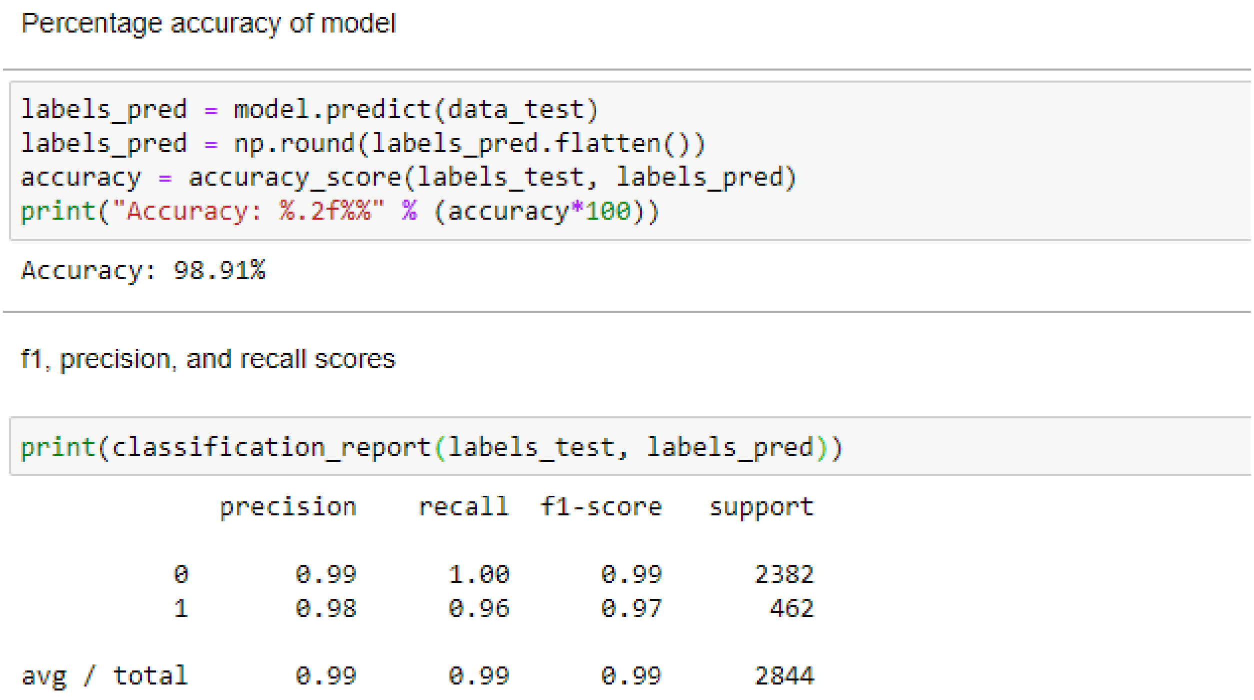
```
Train on 8530 samples, validate on 2845 samples
Epoch 1/10
8530/8530 [==============================] - 32s 4ms/step - loss: 0.1128 - acc: 0.9673 - val_loss: 0.0357 - val_acc: 0.9930
Epoch 2/10
8530/8530 [==============================] - 32s 4ms/step - loss: 0.0378 - acc: 0.9916 - val_loss: 0.0326 - val_acc: 0.9926
Epoch 3/10
8530/8530 [==============================] - 33s 4ms/step - loss: 0.0308 - acc: 0.9926 - val_loss: 0.0333 - val_acc: 0.9933
Epoch 4/10
8530/8530 [==============================] - 32s 4ms/step - loss: 0.0250 - acc: 0.9945 - val_loss: 0.0432 - val_acc: 0.9902
Epoch 5/10
8530/8530 [==============================] - 32s 4ms/step - loss: 0.0214 - acc: 0.9950 - val_loss: 0.0395 - val_acc: 0.9919
```

Fig 7. Results Screenshot



```
plt.plot(hist.history['acc'])
plt.plot(hist.history['val_acc'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'validation'], loc='upper left')
plt.show()
```

Fig 8. Results Screenshot



Percentage accuracy of model

```
labels_pred = model.predict(data_test)
labels_pred = np.round(labels_pred.flatten())
accuracy = accuracy_score(labels_test, labels_pred)
print("Accuracy: %.2f%%" % (accuracy*100))
```

```
Accuracy: 98.91%
```

f1, precision, and recall scores

```
print(classification_report(labels_test, labels_pred))
```

```
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      2382
          1       0.98      0.96      0.97       462

avg / total       0.99      0.99      0.99      2844
```

Fig 9. Results Screenshot

V. CONCLUSION

We pursued an NLP approach to perform adaptation in IDPT systems for two main reasons: a) IDPT deals with a significant amount of texts in the form of computerized exercises for psycho-education, b) NLP method can provide an elegant way to adapt the intervention on the one hand

and provide personalized feedback on the exercises. Several studies have reported that lack of personalized feedback on their interventions was one of the leading causes of high dropouts. Hence, in this paper, we present an NLP based adaptive strategies to adapt intervention based on the symptoms exhibited by the text data. To extract symptoms from text data, we evaluate three different state-of-the-art NLP techniques and proposea novel technique. The results show that both Word2vec and the proposed embedding Depression2Vec captures depressive symptoms better than other methods. However, there are several challenges, as presented by the study associated with the detection of symptoms from text data. Outlining all complexities and challenges is beyond this paper's scope and is kept as one of the immediate future work.

VI.FUTURE WORK

In this study, we incorporated text crawled from several forums/websites related to mental health intervention. While prescreening of the documents was done manually by the authors, it was not validated with the domain experts. Hence, one of the immediate future work is to verify the initial corpus with domain experts such as psychiatrists and linguistics. Another improvement in the embedding would be to incorporate Internet slang, correct spellings, and abbreviations before creating embedding. We did not attempt to validate and study the effects of complex negation within our initial study's time limits. Hence, one of the enhancements of this study is to verify and detect the presence of negation. We expect to improve the performance of the proposed embedding by identifying conditional sentences, uncertain sentences, and NLP dependency in the next phase of our study.

II. REFERENCES

[1]. A. e. a. Konrad, "Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction," in Conference on Human Factors in Computing Systems - Proceedings, vol. 2015-April, (New York, New York, USA), pp. 3829–3838, Association for Computing Machinery, 4 2015.

[2]. S. K. Mukhiya, F. Rabbi, K. I. Pun, and Y. Lamo, "An architectural design for self-reporting e-health systems," in Proceedings – 2019 IEEE/ACM 1st International Workshop on Software Engineering for Healthcare, SEH 2019, pp. 1–8, Institute of Electrical and Electronics Engineers Inc., 5 2019.

[3]. D. e. a. Cer, "Universal Sentence Encoder," EMNLP 2018 – Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings, pp. 169–174, 3 2018.

[4]. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014.

[5]. G. Miller, C. Fellbaum, J. Kegl, and K. Miller, "WordNet: An Electronic Lexical Reference System Based on Theories of Lexical Memory," Revue qu ́eb ́ecoise de linguistique, vol. 17, pp. 181–212, 5 2009.

[6]. B. e. a. Funk, "A Framework for Applying Natural Language Processing in Digital Health Interventions.," Journal of medical Internet research, vol. 22, p. e13855, 2 2020.

[7]. A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-Supervised Approach to

Monitoring Clinical Depressive Symptoms in Social Media," in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017, 2017.

[8]. D. Andrzejewski and X. Zhu, "Latent Dirichlet Allocation with Topicin- Set Knowledge *," 2009.

[9]. D. Ramage, C. D. Manning, and S. Dumais, Partially Labeled Topic Models for Interpretable Text Mining. 2011.

[10]. C. Karmen, R. C. Hsiung, and T. Wetter, "Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods," Computer Methods and Programs in Biomedicine, vol. 120, pp. 27–36, 6 2015.

[11]. K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure," Journal of General Internal Medicine, vol. 16, no. 9, pp. 606–613, 2001.

[12]. N. e. a. Americans, "The ICD-10 Classification of Mental and Behavioural Disorders," IACAPAP e-Textbook of child and adolescent Mental health, vol. 55, no. 1993, pp. 135–139, 2013.

[13]. X. Liu, J. Meehan, W. Tong, L. Wu, X. Xu, and J. Xu, "DLI-IT: A deep learning approach to drug label identification through image and text embedding," BMC Medical Informatics and Decision Making, 2020.

[14]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, International Conference on Learning Representations, ICLR, 2013.

[15]. W. G. Charles, "Contextual correlates of meaning," Applied Psycholinguistics, vol. 21, no. 4, pp. 505–524, 2000.

[16]. S. K. Mukhiya, J. Wake, Y. Inal, and Y. Lamo, "Adaptive Systems for Internet-Delivered Psychological Treatments (In review).," IEEE Access, 2020.

[17]. B. M. Bewick, K. Trusler, B. Mulhern, M. Barkham, and A. J. Hill, "The feasibility and effectiveness of a web-based personalised feedback and social norms alcohol intervention in UK university students: A randomised control trial," Addictive Behaviors, vol. 33, pp. 1192–1198, 9 2008.

[18]. Z. Hilvert-Bruce, P. J. Rossouw, N. Wong, M. Sunderland, and G. Andrews, "Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders," Behaviour Research and Therapy, vol. 50, pp. 463–468, 8 2012.

[19]. C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," 5 2019.

[20]. S. Perera, A. Sheth, K. Thirunarayan, S. Nair, and N. Shah, "Challenges in understanding clinical notes: Why NLP engines fall short and where background knowledge can help," in International Conference on Information and Knowledge Management, Proceedings, (New York, New York, USA), pp. 21–26, ACM Press, 2013.