MULTI-TASK LEARNING ON MENTAL DISORDER DETECTION SENTIMENT DETECTION, AND EMOTION DETECTION

COURAGE ARMAH

IŞIK UNIVERSITY
2024

IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

MULTI-TASK LEARNING ON MENTAL DISORDER DETECTION, SENTIMENT DETECTION AND EMOTION DETECTION

MASTER'S DEGREE PROJECT
DEPEARTMENT OF COMPUTER ENGINEERING
THESIS MASTER OF COMPUTER ENGINEERING DEGREE PROGRAM

COURAGE ARMAH
20COMP5015

SUPERVISOR
ASST. PROFESSOR RAHIM DEHKHARGHANI

IŞIK UNIVERSITY
2024

IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COMPUTER ENGINEERING MASTER'S DEGREE PROGRAM

MULTI-TASK LEARNING ON MENTAL DISORDER DETECTION, SENTIMENT DETECTION AND EMOTION DETECTION

COURAGE ARMAH

APPROVALS:

Asst. Prof. Rahim Dehkharghani (Project Supervisor)	Kadır Has University
Asst. Prof. Emine Ekin	FMV Işık University
Asst. Prof. Ilknur Karadeniz	Özyeğin University

APPROVAL DATE: 25/01/2024

MULTI-TASK LEARNING ON MENTAL DISORDER DETECTİON, SENTIMENT DETECTION AND EMOTION DETECTION

ABSTRACT

Suicidal behavior is a global cause of life-threatening injury and most of the time, death. Mental disorders such as depression, anxiety, and bipolar are prevalent among the youth in recent decades. Social media are popular platforms for individuals to post their thoughts and feelings on. Extracting people's sentiments and

feelings from such online platforms would help detect mental disorders of the users to treat them before it becomes too late.

This thesis investigates the use of multi-task learning systems and single-task learning techniques to estimate behaviors and mental states for early diagnosis. I used data mined from Reddit, one of the popular social media platforms that provides anonymity. Anonymity increases the chances of individuals sharing what they truly feel in their real life. The obtained results by the proposed approaches open new doors to the understanding of how multi-task systems can increase the performance of text classification problems such as depression detection, emotion detection, and sentiment analysis, trained together in a multi-task learning network when compared to their training in isolation in a single-task learning network. We used the SWMH dataset, already labeled by 5 different depression labels (depression, anxiety, suicide, bipolar, and off my chest) and then added emotion and polarity labels to it and made it publicly available for researchers in the literature. The obtained results in this study are also comparable to other approaches in the field.

Keywords: Natural Language Processing, Multi-task Learning, Sentiment Analysis, Deep Learning, Emotion Detection.

ZIHINSEL BOZUKLUK TESPİTİ, DUYGUSALLIK(SENTIMENT) TESPİTİ VE DUYGU TESPİTİ ÜZERİNDE ÇOK GÖREVLİ ÖĞRENİM

ÖZET

İntihar düşüncesi, dünya çapında, ömür boyu tehdit eden yaralanmaların ve çoğu zaman ölümün bir nedenidir. Depresyon, ankseyete bozukluğu ve bipolar gibi zihinsel bozukluklar, son yıllarda gençler arasında yaygındır. Sosyal medya, bireylerin duygu ve düşüncelerini paylaştıkları popüler platformlardır. Sosyal medya platformlardan insanların duygu ve düşüncelerinin çıkarılması, uzmanlar kullanıcıların zihinsel bozukluklarınını tespit edilmesine ve çok geç olmadan tedavi edilmesine yardımcı olacaktır.

Bu tez, erken tanı için davranışları ve zihinsel durumları tahmin etmeye yönelik, çok görevli öğrenme sistemlerinin ve derin öğrenme tekniklerinin kullanımını araştırmaya çalışmaktadır. Anonimlik sağlayan popüler sosyal medya platformlarından biri olan Reddit'in metin verilerini kullandım. Anonimlik, bireylerin gerçek yaşamlarında hissettiklerini paylaşmasına artırır. Önerilen yaklaşımlarla elde edilen sonuçlar, çok görevli sistemlerin, izole eğitimlerine kıyasla birlikte eğitilen depresyon tespiti, duygu tespiti ve duygu analizi gibi metin sınıflandırma problemlerinin performansını nasıl artırabileceğinin anlaşılmasına yeni kapılar açmaktadır. Bu çalışmada elde edilen sonuçlar, alandaki diğer yaklaşımlarla da karşılaştırılabilir niteliktedir.

Anahtar Kelimeler: Doğal Dil İşleme, Duygu Analizi, Duygu Algılama, Derin Öğrenme, Çok Görevli Öğrenme.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Almighty God for the will and strength he granted me throughout my master’s studies.

I also extend my kindest gratitude to my supervisor Asst. Professor Rahim Dehkarghani, for his patience and guidance throughout the projects and experimentations that birth this thesis. I carry with me all the knowledge and skills you imparted to me during this process.

I also wish to acknowledge students, Mr. Özay Özerceli and Mr. Ayham Treter for their help in the labelling and interpretation of the dataset associated with this thesis.

I wish to thank Asst. Prof Emine Ekin, Asst. Professor Dr. Ilknur Karadeniz and the staff of our institute for their help in making this study process and thesis defense process a smooth one.

Last but not the least, I would like to thank my family for their spiritual and moral support throughout my studies.

Courage ARMAH

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

LIST OF ABBREVIATIONS

NLP: Natural Language Processing

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

MTL: Multi-Task learning

SWMH: Reddit Suicide Watch Mental Health Collection

PTSD: Post traumatic stress disorder

DNN: Deep Neural Network

MRC: Machine Reading Comprehension

RAG: Retrieval Augmented Generation

GPU: Graphics Processing Unit

TPU: Tensor Processing Unit

BiLSTM: Bi-directional Long Short-term Memory Layer

Conv1D: 1-dimensional convolution layer

LR: Learning Rate Scheduler

CHAPTER 1

1. INTRODUCTION

1.1 Purpose of Study

Depression is a mental health disorder prevalent among the youth. Mental disorders such as depression, suicidal ideation, anxiety, and bipolar continue to reduce the quality of life of individuals worldwide. Social media posts are an extension of an individual’s thoughts and feelings which makes it possible to mine sentiments from such write-ups. Social media has gained global popularity, especially amongst the youth. People use social media to share their feelings, opinions, and events. These platforms have become safe havens for people suffering from mental disorders to talk about their issues. This has happened because of the stigma associated with mental disorders.

Online social media like Twitter and Facebook are sources of large amounts of textual data in the form of social media posts. These data have been implemented in artificial intelligence research using natural language processing (NLP) techniques to extract meanings from posts (De Choudhury et al.,2013; Homan et al.,2014; Coppersmith et al.,2015). Social media has provided a low-cost medium for detecting mental illnesses. However, the ethical issues of data privacy and anonymity have made it difficult for steady research to be conducted using social media posts in the field of psycho-medicine (Horvitz et al.,2015). These ethical issues have slowed down research in this deep-learning field.

The social media platform, Reddit is becoming well-known as a source of textual data for detecting mental disorders as it provides more anonymity. This advantage makes it possible for individuals to post messages openly without fear (Pirina et al.,2018; Tadesse et al.,2019). Recently, researchers have used deep learning approaches like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) on image, audio, and video data; in creating frameworks that are used to detect mental disorders such as depression (De Melo et al.,2019; Zhang et al.,2020; Balbuena et al.,2021). A hybrid of NLP methods and deep learning techniques have also been employed on textual data to detect depression, by extracting certain features and using these features on deep learning models to provide accurate detection models. The challenge has been to find a simple but robust methodology that can improve overall generalization and data efficiency (Shah et al.,2020). Multi-task learning (MTL) in deep learning was introduced to help solve this challenge.

The main goal of multi-task learning is to combine important encoded information shared between multiple related tasks to improve generalization on all tasks (Caruana et al.,1997). This methodology makes it possible to work on small datasets without compromising accuracy. Even though MTL is relatively new, two main model training methods have been proposed. These are called joint training and multi-step training. The joint training methodology is more popular in MTL because it embodies the main essence of MTL. In Joint training, several tasks are trained simultaneously so that task-specific data can be learned at the same time (Dong et al.,2015; Liu et al.,2015; Kim et al.,2017; Xiao et al.,2018). Multi-step training on the other hand relies on the concept of feedforwarding, where a task’s output or hidden layer is used as the input of another related task (Sogaard et al.,2016; Hashimoto et al,2017; Dinan et al.,2019; Lewis et al, 2020).

This project seeks to introduce and analyze scalable natural language processing and multitask learning techniques that could be leveraged in providing more accessible forms of early mental disorder diagnosis which will curtail the high risks of these mental disorders, especially depression. The project primarily uses the Reddit Suicide Watch and Mental Health Collection (SWMH) dataset as a benchmark dataset. This is a textual corpus mined from Reddit social media platform. The SWMH corpus is a multi-label textual dataset of five communities from Reddit. These are 'Anxiety', 'bipolar', 'depression', ‘Suicide Watch', and 'offmychest'. The original dataset has more than 50,000 individual posts (Ji et al.,2022). We labeled 5,000 instances of the SWMH dataset with negative and positive sentiments, and emotion labels such as anger, fear, hopefulness, hopelessness, joy, sadness, calmness, and disgust. This is to help in the multitask learning process. According to the labels, 1 means there is sentiment or emotion present in the text instance and 0 means the sentiment or emotion is not present in the text instance. We then build custom end-to-end deep neural-based multitask frameworks using the joint training and shared trunk approach (Ghosh et al.,2022). Upon comparison of our framework with other benchmark frameworks, our best-performing framework attains an f1-score of 62% for mental disorder detection, 4% better than its single-task model.

The major contributions of the project are explained further.

- We propose and evaluate custom multi-task frameworks that perform better than their counterparts, especially when it comes to mental disorder detection and sentiment detection.
- We extended 5,000 instances of the benchmark dataset (SWMH) with sentiment labels and emotion labels using majority voting.
- Each instance in the extended SWMH dataset is marked with multilabel sentiment and emotion classes as opposed to the single labeling scheme followed in the introductory work (Ji et al.,2022). This updated dataset provides a benchmark avenue for future extensive studies in the field of psycho-medicine using MTL approaches.

The structure of the rest of the thesis is as follows. Brief explanations of major terms and methods relevant to this research are explained in the first part of Chapter 2. The second part of Chapter 2 describes some previous works in this domain. Next, we discuss the proposed methodologies and metrics relevant to this project in Chapter 3. Chapter 4 shows and explains the Experimental Evaluations obtained from our proposed methodologies. Results pertaining to the experimental evaluations are discussed in Chapter 5. Conclusion and Future works are then shown in next.

CHAPTER 2

2.LITERATURE REVIEW

2. 1 Theoretical Background

One of the main characteristics of a successful machine learning model is the availability of large, labeled datasets. Multi-task learning is a major solution proposed to reduce the dependence on this feature. MTL has been proposed to be implemented in instances where large datasets are not readily available without the model compromising accuracy and efficiency. Generally, this is achieved by jointly learning related tasks or similar tasks to reduce overfitting. Even though an MTL approach can perform multiple tasks simultaneously, the success of an MTL model is measured when at least the accuracy of one of the tasks performed is improved in comparison to the associated single-task model(s) (McCann et al., 2018).

We can infer from the purpose of MTL systems that there are two basic elements of MTL. These are the definitions of tasks and task relatedness. We can define tasks in deep learning through supervised learning methodologies such as classification and regression, unsupervised learning approaches such as clustering, semi-supervised learning, reinforcement learning, etc. Task relatedness on the other hand is based on whether the tasks involved are related to each other to be able to inherently learn from each other. These two factors are considered when choosing a specific methodology to use in MTL systems (Zhang et al.,2018). As stated in the Introduction section, we can categorize MTL into Joint training and multi-step training according to task relatedness or task dependency according to Zhang et al. (2022).

As shown in Figure 2.1, a general encoder-decoder architecture can be used to further understand how the two frameworks work (Zhang et al., 2022). An MTL model will have input data and desired output task(s)

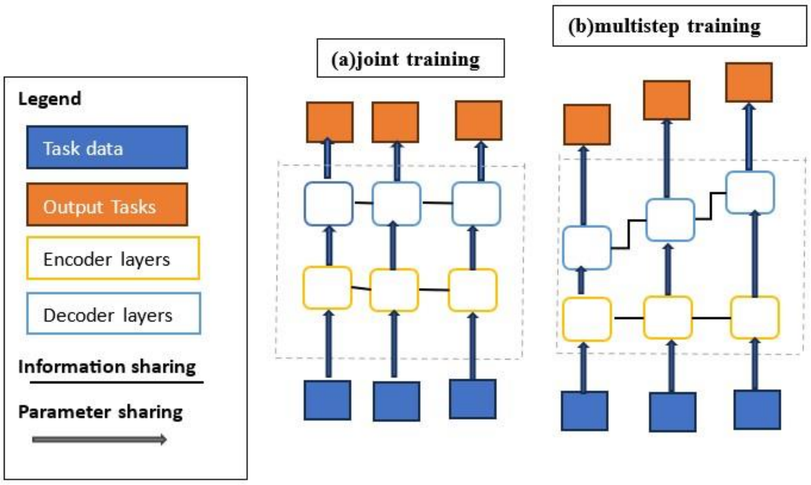


Figure 2.1. Multitask learning frameworks according to relatedness.

In Natural Language Processing (NLP) models, the encoder layer usually consists of but is not limited to the embedding layer. The word-level or sentence-level embedding layer is used to map each word instance or sentence instance to a low-dimensional vector, where each dimension also maps into a particular feature of the word. This is easier to simulate by the model. It is to be noted that the mainly used encoders can be in the form of a neural network called the Bidirectional Long Short-Term Memory (BILSTM) which is a type of Recurrent Neural Network (RNN), or Bidirectional Encoder Representations from Transformers (BERT) layers (Devlin et al., 2018). Parameters can be shared amongst decoder layers.

When tasks are independent at the decoding level, the decoder layer transforms the hidden layer vectors into task-specific outputs for the joint approach. If one task depends on the output from a previous task, the multistep approach is employed. In this project, we use the joint training approach with a shared trunk framework where the encoder layer is shared amongst the tasks for MTL.

2.2 Related Works

Techniques used in detecting mental disorders such as depression have evolved from pure natural language processing (NLP) techniques (Ren et al., 2015) to a hybrid of NLP and deep learning techniques (Ji et al., 2018; Mustafa et al.,2020). De Choudhury et al. (2013) investigated the usage of Twitter to diagnose major depressive disorder in individuals. The study made use of behavioral cues from individual posts such as language, and the mention of anti-depressants to build a statistical classifier that estimates the risk of depression. Results obtained from the research highly suggested that important signals for detecting depression amongst individuals could be measured in their posts. Through a shared task and unshared task competition, Coppersmith et al. (2015) came up with binary classification experiments to classify both depression and Post Traumatic-stress disorder (PTSD). The study primarily realized that simple linguistic features with simple machine learning techniques, provide some classification probabilities for depression detection tasks. It was reiterated in this study that a well-constructed dataset provides the basis for good identification techniques.

It is worth noting that the ethical nature of using legally acquired posts from social media is still a matter of discussion. In a study by Horwitz et al. (2015), it was determined that social media has provided a potential low-cost medium for detecting mental illnesses, but the ethical issues of data privacy and anonymity have made it quite difficult for steady research to be conducted using social media posts in the field of psychomedicine. Even though no major structure has been put in place, researchers are guided by moral and ethical intuitions in how to access and use such data. In recent times, the social media platform, Reddit is gaining grounds in its use as a source of textual or linguistic data for detecting mental illnesses as it provides more anonymity. Tadesse et al. (2019) used posts mined from Reddit to gain insight into the characteristics that show the depressing attitudes of individuals on the social media platform. The study demonstrated that combining natural language processing techniques and machine learning approaches provides the top-performing pipeline in depression detection using textual data.

As a way of reducing the dependence on large datasets, MTL learning has started gaining ground in NLP research and application areas like text classification, information extraction, and machine translation. Liu et al. (2015) proposed a multi-task deep neural network (DNN) architecture for semantic classification and semantic information retrieval tasks on supervised datasets. These tasks were query detection tasks and web-search prioritization respectively. Their approach involved a cross-task system which resulted in significant generalization gains. Dependency parsing is an NLP technique used to evaluate grammar structure in a sentence and find out related words and the relation-type between these words. Dependency parsing is an important part of many NLP systems. But it is very costly in building large and accurate treebanks that are an essential component of dependency parsing. Dong et al. (2015) explored a method that involved sharing related language treebanks to achieve a more accurate parser. Also, according to Dong et al. (2015), structural information from the source, which is a resource-rich language, is added as a pre-requisite in the supervised training of a target language parser with poor resource by the use of a mall treebank. When compared with a supervised single-task model, the benefit increases to 8.7% for just 1000 tokens.

To further reduce the effect of older MTL systems being unable to distinguish between helpful information and helpless information, Xiao et al. (2018) introduced a Convolutional neural network (CNN) based multi-task architecture with allocated private subnets to help share features amongst tasks in a selective way. They called it the Gate Sharing Unit which can identify and select features beneficial to the specific tasks involved. They conducted experiments with several benchmark test classification datasets and obtained results that show that their model has better accuracy than single-task models. Machine Reading Comprehension (MRC) is also another area that has benefitted from MTL innovations. MRC simply aims to teach machines to understand text like a human. Examples of MRC systems are passport and document readers.

Xu et al. (2018) proposed an MTL system that jointly learns and performs several MRC tasks. Their model performed better than even human counterparts when added to semantic representations from pre-trained language system models. Liu et al. (2017) also proposed a type of Adversarial multi-task learning framework that can reduce noise activities amongst shared layers in a framework so that the most important features are retained and shared. Their proposed system divides the task-specific and shared space more accurately than the traditional way of roughly sharing parameters. Based on the lens of Lagrangian duality, Mao et al. (2020) proposed a novel adaptive adversarial MTL system that improves the performance of original adversarial MTL methods. Their model was used to conduct experiments on sentiment analysis and topic classification applications. So far, we have been looking at MTL experiments that generally use the joint training approach in their work.

Sogaard et al. (2016) proposed a multi-step training approach that utilizes complex and large bi-directional RNNs to perform various tasks at different layers. They experimented on syntactic chunking and CCG super tagging, coupled with the additional task of POS tagging. According to their research, it is more advantageous to have POS supervision in the hidden layers rather than the outermost layer. Lewis et al. (2015) experimented on a multi-purpose architecture for retrieval-augmented generation (RAG). RAG models combine pre-trained parametric and non-parametric memory features for language generation. Their architecture consisted of RAG models where the parametric feature set is a pre-trained seq2seq model and the non-parametric feature memory is a vector index of Wikipedia, that are accessed with a pre-trained neural retriever. Results generated from their experiments showed that combining parametric and non-parametric features for the generation of knowledge-intensive-task solutions is better because even humans could not be expected to perform such tasks with no prior external knowledge. An example of such knowledge-intensive tasks is answering trivia questions.

CHAPTER 3

3.SUGGESTED APPROACH

3.1 Data Pre-Processing

Data pre-processing is defined as data manipulation to ensure that the most important form of data is available to be used by a model for certain tasks after data collection. Data pre-processing involves data cleaning and other techniques that help in optimizing system and model efficiency. I performed the following pre-processing techniques for the labeled datasets.

- **Stop word Removal:** This is the removal of commonly used words that do not hold any importance and are just space or memory occupiers. Commonly removed English stop words include but are not limited to ‘the’, ‘a’, ‘but’, ‘and’, etc.
- **Contractions Replacement:** This is the replacement of short forms and contracted words or phrases with their full versions. The SWMH dataset has a lot of contracted words and phrases. Examples of such phrases and words include but are not limited to ‘isn’t’, ‘aren’t’, ‘asap’, ‘doc’, ‘gf’, ‘meds’ etc.
- **Tokenization:** The dataset presented is a sentence-level dataset. For algorithms to further process these sentences easily, we break the sentences down into word tokens, which can be considered as discrete elements and can be represented directly as a vector.
- **Normalization:** This is defined as putting all words on the same level. I performed Normalization techniques such as Case folding and Lemmatization. Case folding is where all words with capital letters are turned into small letters. Lemmatization simply is when the tokens or words are represented by their

lemma or shared root on a semantic level, thus the meaning of the word is kept.

- **Padding:** Since our tokens will be in vector forms, and then later in a matrix format, padding is done so that every review representation is uniform to determine a maximum length input for the CNN. The padding adds zeros to matrix lengths that are not up to the maximum parameter length.

3.2 General Methodology

This project employs a combination of NLP techniques and deep learning methods such as text pre-processing methods, transfer learning, and custom model building. All methods are implemented using Python programming. The proposed pipeline is shown in the figure below.

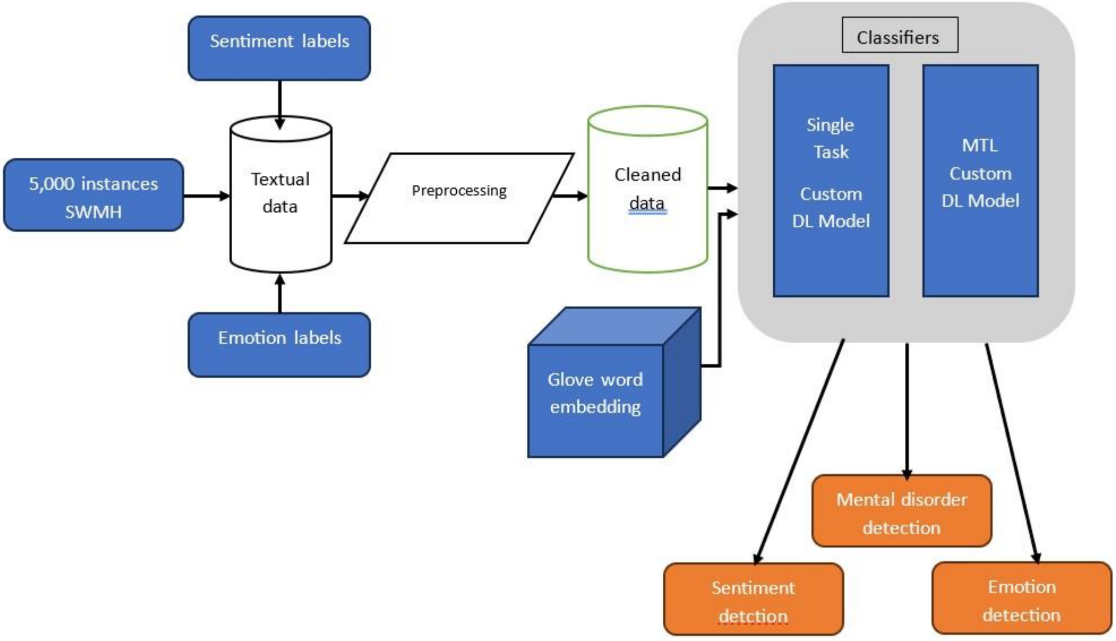


Figure 3.1.The General view of the suggested Methodology.

We implemented general pre-processing functions such as lemmatization, contractions removal, stop word removal, tokenization, and sequence padding. We proposed and compared two baseline models with custom layers for the single tasks (mental disorder detection, emotion detection, and sentiment detection) as shown in Figure 3.2 and Figure 3.3 below.

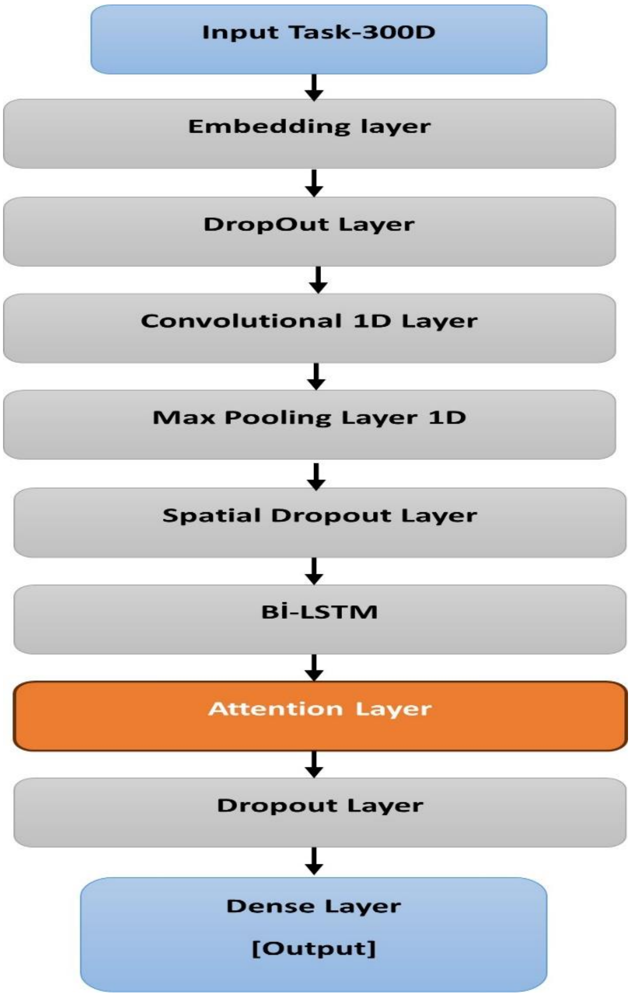


Figure 3.2. Custom single-task model with attention layer.

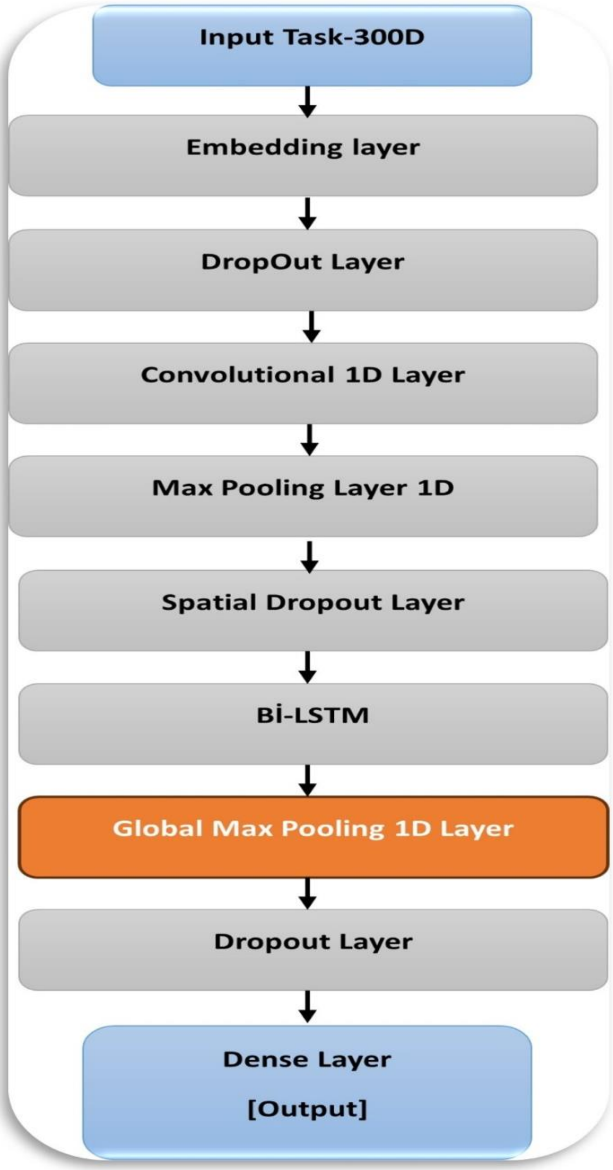


Figure 3.3. Custom single-task model without attention layer.

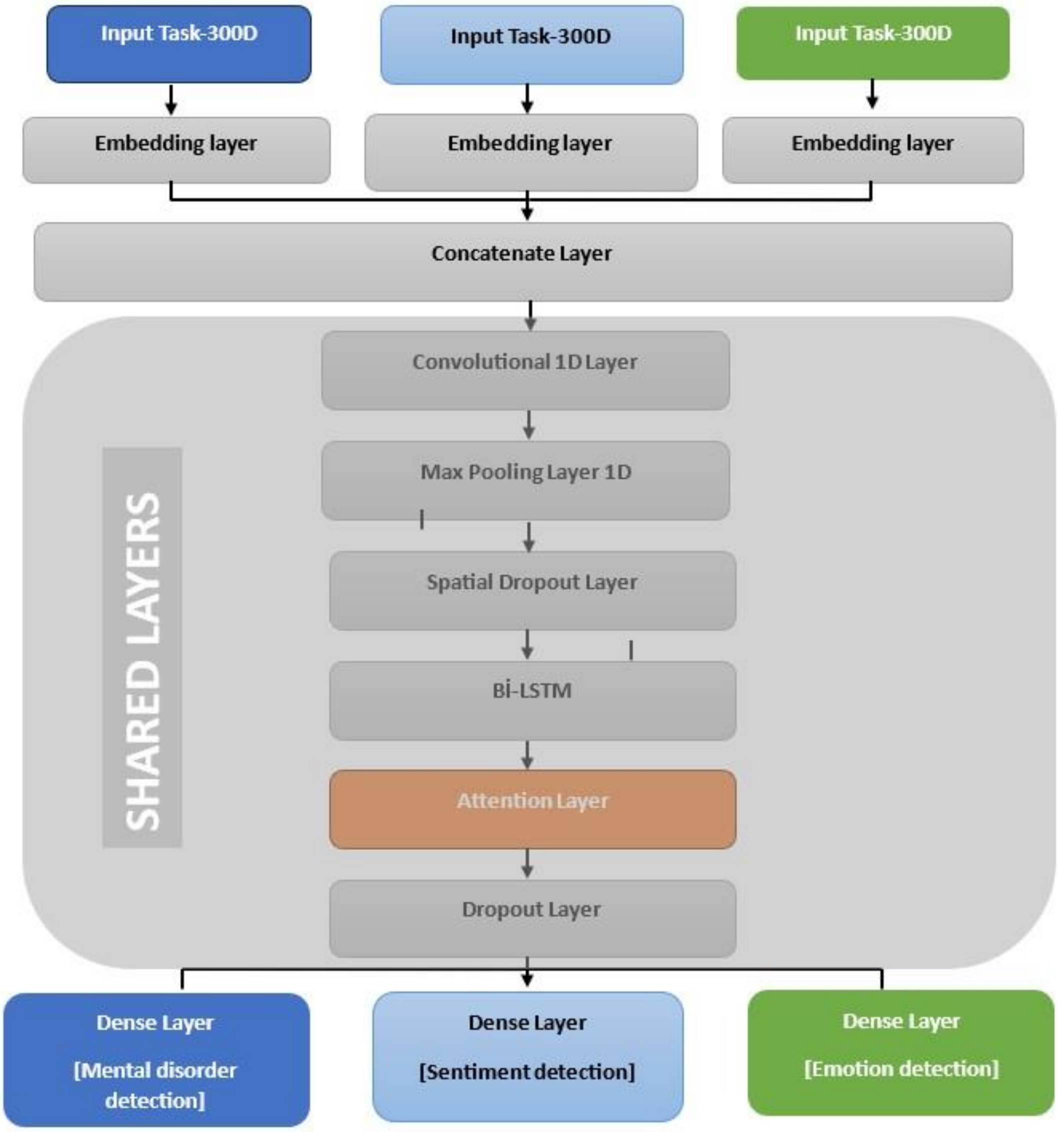


Figure 3.4. The proposed MTL model.

All main layers of the MTL custom model shown in Figure 3.4 are shared except the input and the last dense layers for the output or classification. This ensures that each task is classified separately after possible generalization from the shared layers. Another advantage is that both tasks have the potential to use the shared layers and parameters which has increased volume for better generalization. To further understand the reasons for the use of each layer we explain what the layers consist of with their hard parameters.

Word Embedding

Word embedding is converting the tokens into constituent vectors following certain rules. The word embedding vectors are then converted into word embedding matrices before being used as input for our custom model. Instead of training features from our dataset, we download a pre-trained word embedding and use it as the basis for our model. This process is a form of transfer learning. For transfer learning, I downloaded and implemented a pre-trained 300 vector-dimension Glove embedding (Pennington et al., 2014). Word embedding occurs when the word vectors are passed through a non-trainable embedding layer that outputs feature for every word in the text used. All tokens from our datasets were also used as features for our custom deep-learning model. Another advantage is that both tasks have the potential to use the shared layers and parameters which has increased volume for better generalization. To further understand the reasons for the use of each layer we explain what the layers consist of with their hard parameters.

Dropout Layers

It is easy for deep neural networks to overfit on a training dataset with few instances. This results in the model memorizing statistical noises in the training data.As a result, the model will perform poorly when introduced to a new dataset during evaluation. The Drop out layer is a layer introduced to help improve the regularization of the training model so it can generalize better on the test data. When training a model, sample layer outputs are randomly forgotten. With this mechanism, the model is forced to see every layer as a new form of the old one to reduce memorization.

Convolution Layer 1D

Convolution is simply a mathematical operation where you reduce a matrix or a vector into a smaller one. If your input matrix is one dimensional then you summarize along that dimension, if the matrix is n dimensions, you could do the reduction along n dimensions. Two-dimension convolutional neural networks (CNN) have been very popular in image and video processing for deep learning applications due to their ability to filter out unnecessary portions of an image per direction. This idea is how 1-dimensional CNNs were conceived. They act like 2-dimensional CNN but are in one dimensional space which is suitable for time series and texts. Since my project is a text classification problem, I employ the 1-dimensional CNN(Conv1D) layer. Conv1D layers have two main parameters. The first is the filter. As the name suggests, the filter is responsible for capturing features or patterns from input data. It is worth noting that just increasing the number of filters of a CONV1D layer increases the complexity of the model, thereby increasing the potential for overfitting. It is thus important to find a balance through intuition and trials. For this project, the CONV1d layer has 300 filters. The second parameter of the CONV1D layer is the kernel size. The kernel can be described as the lens the layer uses to scan the vector data to find important features. The kernel size is very important since it can affect the performance of the model. A large kernel size means the lens of the layer looks at a larger area and might not be able to capture fine details in the input data. For the project, the optimum kernel size is 5.

Max-Pooling Layer 1D

Max pooling is a type of filtering operation that's typically added to CNNs after each convolutional layer. The main purpose of these type of pooling layers is to breakdown the dimensionality of the output of the convolution layer by scanning through the input data plane according to certain parameters such as pool size and the strides it can take during the scanning process. With that idea in place, one-dimensional max-pooling layers are typically used just after a one-dimensional convolutional layer to further learn features from a text vector input.

Spatial Dropout Layer 1D

The one-dimensional spatial dropout performs the same functions as its two-dimensional spatial dropout. The difference is one is used in one-dimensional spaces like text and the other is used in two-dimensional spaces like images. The one-dimensional spatial dropout was adopted in this model to help improve regularization of the activations from the previous layers by dropping entire 1-d feature maps it does not deem important. This mechanism helps promote independence between feature maps the model is learning from, thereby reducing memorization.

Bi-directional LSTM

Bidirectional Long Short-Term Memory (BiLSTM) is a recurrent neural network with two LSTM networks used primarily in natural language processing because of the ability of the network to forward and backpropagate and keep information while doing so. BiLSTM contains two LSTM's. These inadvertently increase the information size the layers of the model can use, thereby increasing the ability of the model to learn the context for better generalization. In simple terms, BiLSTM helps the model figure out words that precede each other in a sentence to help the semantic generalization of the model.

Global Max-Pooling Layer 1D

One-dimensional Global max-pooling layers are usually used at the last stages of a model just before classification is done to downsample its input space. It takes the maximum values in each vector space which is equivalent to the most important features in that vector space. This type of neural layer produces one feature map for its equivalent category of the classification in the output layer. The resulting vector is then fed into either a sigmoid or softmax function before classification.

Optimization Algorithm

For training the model, we use Adam's optimization algorithm for Gradient Descent using a stochastic Gradient Descent(localization).

Callbacks are functions that are implemented when an epoch ends during training. We use the following callbacks:

- Learning Rate Scheduler (LR): This function is used to change a Learning Rate at a specific epoch to get more accurate results. In the single-task models, the learning rate exponentially decreases by a 0.5 factor if validation loss continues to fall after the third epoch. For the MTL model, the learning rate decreases by 0.1 as hard parameters.
- Early Stopping: This function hard-stops the model training and assigns

the weights of the most accurate generalization that has occurred during the training so far to the trained model.

Attention Layer

Till recently, machine learning models found it difficult to classify long sentences which are transformed as large sequences of vector data, due to the inability to create semantics from their related vectors. The attention mechanism for NLP was introduced by Luong et al. (2015) and Bahdanau et al. (2016). The attention mechanism of the attention layer was introduced to help neural networks in memorizing large sequences of data. There are different variations of the attention mechanism which depend on the purpose of the model involved. In this project, we introduce and implement a self-attention variation of the attention mechanism.The attention layer has three main parameters called Queries, values, and keys respectively. A relation can be described for the three parameters algebraically. Every input vector is compared to other vectors to do three main things; to get its vector output we can call Query or y_i, to get the j-th output we can call the Key or y_j, and to calculate the Value(V) of each output matrix vector after the corresponding weights are known. To obtain these parameters, three weight matrices(W) will be needed to compute three linear transformations for each input vector we call x_i. These matrices are generally called Q, K, and V, and can be calculated using the following operations.

- $k_i = W_k x_i$
- $q_i = W_q x_i$
- $v_i = W_v x_i$

The attention scores are then calculated using the Q, K, and V matrices obtained using the following formula.

- $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})\mathrm{V}$

The input consists of queries and keys of dimension dk and values of dimension dv. We compute the dot product of the query with all keys, divide each by the square root of dk, and apply a softmax function to obtain the weights on the values. After applying softmax, the resulting value is multiplied by a value matrix V to help the attention mechanism keep its focus on important word vectors while gradually diminishing irrelevant word vector.

CHAPTER 4

4.EXPERIMENTAL EVALUATION.

4.1 Dataset

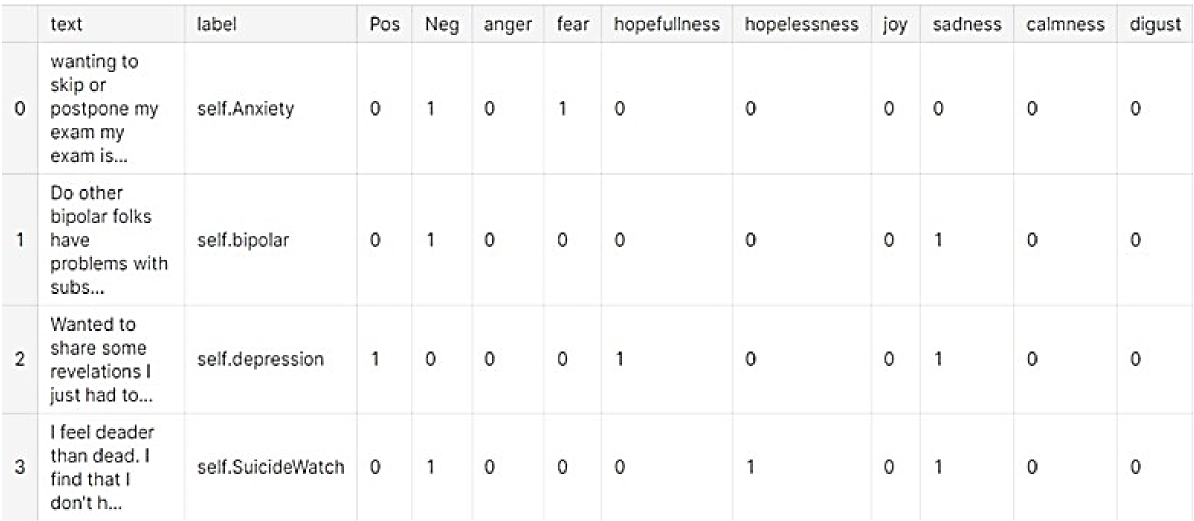
The dataset used for this project is a Reddit-based dataset curated by Ji et al. (2022). The original dataset contains more than 54,000 posts mined from depression and suicidal ideation-related communities such as ‘suicide watch’, ‘anxiety’, ‘bipolar’, ‘offmychest’, and ‘depression’. These five (5) communities became the bases for mental disorder labeling. These posts were mined using Reddit’s official mining API. The 10th edition of the International Statistical Classification of Diseases (ICD-10) helps describe the main mental disorders associated with the project. According to Chapter V of the ICD-10, Suicide includes suicidal tendencies and intentional self-harm. Depression can also be separated into mild, moderate, or severe depression. Characteristics of depression include patients having very low moods, reduction in patient energy, and reduction in patient activity levels. Bipolar on the other hand is characterized by a patient having two or more episodes of mood swings and significant disturbances in daily activities. Anxiety can be categorized into phobic and other anxiety disorders. Phobic anxiety disorders result from the fear of uncertainties. It is to be noted that ‘offmychest’ is not a mental disorder. It is, however, a class mined from Reddit to help in the mental disorder detection process by Ji et al. (2022). Table 4.1 shows a random sample of the original SWMH dataset.

Table.4.1. A subset of original dataset.

Text	**label**
Suicidal Thoughts and Venting I would first like to start off by saying that, if I could, I would not want to die or cease my existence I would want to be happy and not feel like things are crashing down on me, but I can’t	self.SuicideWatch
I really REALLY envy those who are liked and loved and respected	self.depression
19 years and counting Its the last month of 2017 and I feel so bitter at everything.Christmas is coming and even though its supposed to be a period of festive cheer, everyday I wonder if my parents can provide for the family.	self.offmychest
Any tips on easing a nervous stomach before eating? When I get anxious, I tend to lose my appetite. Happens a lot before meals and events, just looking for some ways to ease the feeling and get myself to eat a little bit so as to not make others uncomfortable.	self.Anxiety

Feeling so behind in life How do I deal with the feelings that everyone around me is years ahead in life? I'm 26 and still finishing undergrad/haven't had a job yet, while everyone else is years into a career and/or graduate education. I feel like a complete failure and know I'll always have and be less than everyone else.	self.bipolar

From the original data, we realized there were a lot of contracted words and phrases such as can’t, haven’t, etc. We introduced the replacement of contracted words with their original forms to help in the generalization process of the model. Since the main aim of the project is to investigate how we can leverage the advantages of multi-task learning to mainly improve mental disorder detection using emotion detection and sentiment detection, we labeled 5,000 instances of the swmh dataset with negative and positive sentiments, and emotion labels such as anger, fear, hopefulness, hopelessness, joy, sadness, calmness, and disgust (Dehkharghani et al., 2023). Three (3) field experts labeled the SWMH dataset with intuitive sentiment and emotion labels. According to the rules of labeling, each labeler gave a value of 1 to any sentiment (positive and negative) or emotion detected in the text. 0 were given to sentiments or emotions that are not detected in a specific text. After that, majority voting or hard voting was implemented on the results obtained by the three labelers. An agreed value was assigned to each text if two out of the three labelers agreed. If an agreement was not reached, a fourth labeler labeled that instance or text with his or her relevant sentiment or emotion. Figure 4.1 shows the extended SWMH dataset after majority voting was done. Appendix A.1 and Appendix A.2 show a more detailed look at the textual data instances. The first instance of Figure 4.1 has an original label of Anxiety. According to our majority voting system, that data exhibits the feature of negative sentiment and fear emotion, thus we gave it a 1. Other sentiments and emotions are not recognized in the first instance hence a 0 is given to them. For the second instance labeled as bipolar disorder, we concluded that it also had a negative sentiment and a more recognized emotion of sadness. We therefore labeled them as 1 and the other instances which were not significant were labeled as 0. The third instance with a depression tag was labeled with a positive sentiment and was concluded to also have hopefulness and sadness emotions as significant. We can understand from the labeling format that an instance of the dataset can have one or more emotion labels or sentiments. This helps enrich the feature set recognition for the models after probability implementations are made. The model will then benefit from the relatively large feature set made available from the labeling instances even though the extended dataset is just 5000 instances.



	text	label	Pos	Neg	anger	fear	hopefullness	hopelessness	joy	sadness	calmness	digust
0	wanting to skip or postpone my exam my exam is...	self.Anxiety	0	1	0	1	0	0	0	0	0	0
1	Do other bipolar folks have problems with subs...	self.bipolar	0	1	0	0	0	0	0	1	0	0
2	Wanted to share some revelations I just had to...	self.depression	1	0	0	0	1	0	0	1	0	0
3	I feel deader than dead. I find that I don't h...	self.SuicideWatch	0	1	0	0	0	1	0	1	0	0

Figure 4.1. Subset of Extended SWMH dataset after labelling by polarity and emotion labels.

Due to the novel nature of MTL for text classification, it is quite difficult in getting other baseline methodologies from other research. The sparse nature of the data, 5000 instances affects accuracy greatly. Also, the data is very imbalanced. Some instances of the sentiment or emotion labels have less than 50 instances out of the 5000 which makes the data very skewed and can affect general performance and generalization. We can see the significant effect of MTL with these data limitations.

4.2 Used Programming Languages, Platforms and Tools

All methods are implemented using Python programming. Kaggle Cloud platform with Python and machine learning packages was the main platform used for programming. In addition to Kaggle having deep learning-dependent tools such as TensorFlow and Keras packages, it also provides users with powerful processing units on the cloud and permits registered users to use not more than 30 hours of GPU and not more than 20 hours of TPU per week. Researchers and enthusiasts in the data science and artificial intelligence field can use Kaggle.

Microsoft Excel was the main software used to process the majority voting scores for the newly added sentiment and emotion labels to the SWMH dataset.

Python programming language was also chosen for this project as it is the programming language with the most widely used artificial intelligence dependencies and packages. It is also a very dynamic and easy-to-understand programming language.

4.3 Metrics for Evaluation

The process of detecting whether a post has elements of a mental disorder or not is a binary classification problem. The same can be said for sentiment detection and emotion detection problems. Due to this, we create a classification report from the results we obtain after our trained custom model is applied to the test dataset. The table below shows the basic elements and explanations that are needed to further understand the metrics used to evaluate the experimental results.

Table 4.2. Illustrative table that shows basic elements to explain evaluation metrics.

CLASSIFICATION	**Classification matrix for custom model**	
	positive	negative
positive	True Positive (TP)	False Positive (FP)
negative	False Negative (FN)	True Negative (TN)

A **True Positive (TP)** is a probabilistic outcome where a system correctly predicts the positive instances. In the same vein, a True Negative (TN) is a probabilistic outcome where a system correctly predicts the negative instances. Also, A False-positive (FP) is a probabilistic outcome where a system incorrectly predicts the positive instances. A False-negative (FN) is a probabilistic outcome where a system incorrectly predicts the negative instances. It is to be noted that in mental disorder detection, not correctly predicting whether someone has a mental disorder or not can lead to catastrophic results, thus the need to create models that reduce the False-negative as much as possible. The abovementioned elements are used in the following metrics to help evaluate our model.

- **Accuracy:** This is the percentage of predictions our prediction system got correctly. In terms of binary classification:

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

- **Precision(p):** This unit of measurement evaluates what percentage of positive identifications was correctly made.

$$\text{Precision(p)} = \frac{TP}{TP+FP}$$

- **Recall(R):** This unit of measurement also evaluates what percentage of actual positive identifications was made correctly.

$$\text{Recall(R)} = \frac{TP}{TP+FN}$$

- **F1-Score(F1):** F1-score is simply a way of adding the precision and recall of the model. The F-score is commonly used by deep learning researchers as a common unit of measurement for machine learning models, especially in natural language processing. I will be using the f-score mostly during the evaluation of this project.

$$\text{F1} - \text{Score(F)} = \frac{2(TP+FP)(TP+FN)}{TP}$$

Keras is one of the high-level frameworks based on TensorFlow used for machine learning implementations. The use of Keras has provided machine learning researchers with simple ways to understand the complex mathematics behind neural networks. The training and testing dataset is split into 80-20 ratios for both datasets respectively. The optimizer used to train the models was Adam optimizer. Small datasets for deep learning are difficult to implement for models because small training datasets encourage memorization in deep neural networks. Because of this, training errors will tend to be very low while test errors on the test datasets are increased significantly. To help mitigate the problem of memorization, I employ the services of glove word embedding and padding to add more weights to the input dataset.

The table below shows the micro-average f1-scores of the single-task models and their equivalent multi-task learning models when mental disorder detection, sentiment detection, and emotion detection are run simultaneously.

Table 1.3. Experimental results of f1-score from test dataset compared with other research in the field.

	Mental disorder detection	**Sentiment detection**	**Emotion detection**
Single Task-with attention layer (5000 instances)	**0.59**	**0.88**	**0.68**
Single Task-without attention layer (5000 instances)	**0.60**	**0.88**	**0.66**
MTL Task-with attention layer (5000 instances)	**0.60**	**0.88**	**0.62**
MTL Task-without attention layer (5000 instances)	**0.62**	**0.87**	**0.63**
Single Task-Ji et al (RN model)2021(SWMH dataset-54000 instances)	**0.65**	**-**	**-**

We noticed an increase in the f1-score of mental illness detection from 60% to 62% after multitask learning is employed using the MTL model with no attention layer. This shows a 2% increase in the f1-score. Sentiment detection remains the same and there is a significant reduction in the f1-score of emotion detection from 66% to 63%. With the attention layer MTL model, the f1-score increases for mental illness detection from 59% to 60%. Sentiment detection remains the same and there is a drop in the f1-score from 68% to 62%. Compared to state-of-the-art methodologies our models seem to be more accurate as shown in the table above. We also noticed a high sentiment detection f1-score of 88% for all four custom models experimented on. These are significantly higher than other baseline models from other papers as shown in Table 4.3. For an MTL model to be successful, at least one of the target tasks must show an increase in its performance. I believe the proposed model has proved to be a success in this field. I further investigate the f1-scores by looking at the classification report for each of the tasks below.

4.4 Single Task with Attention Layer vs MTL with Attention Layer

The classification report for the single tasks of mental disorder detection, sentiment detection, and emotion detection using the custom model with an attention layer are shown in the figures below respectively. We also show the classification report for the MTL custom model with an attention layer.

	precision	**recall**	**F1-score**	**support**
Self.Anxiety	0.80	0.72	0.76	179
Self.bipolar	0.84	0.53	0.65	132
Self.depression	0.65	0.51	0.57	326
Self.suicidewatch	0.74	0.32	0.45	154
Self.offmychest	0.60	0.46	0.52	209
Micro avg	0.70	0.51	0.59	1000
Macro avg	0.72	0.51	0.59	1000
Weighted avg	0.70	0.51	0.59	1000
Samples avg	0.51	0.51	0.51	1000

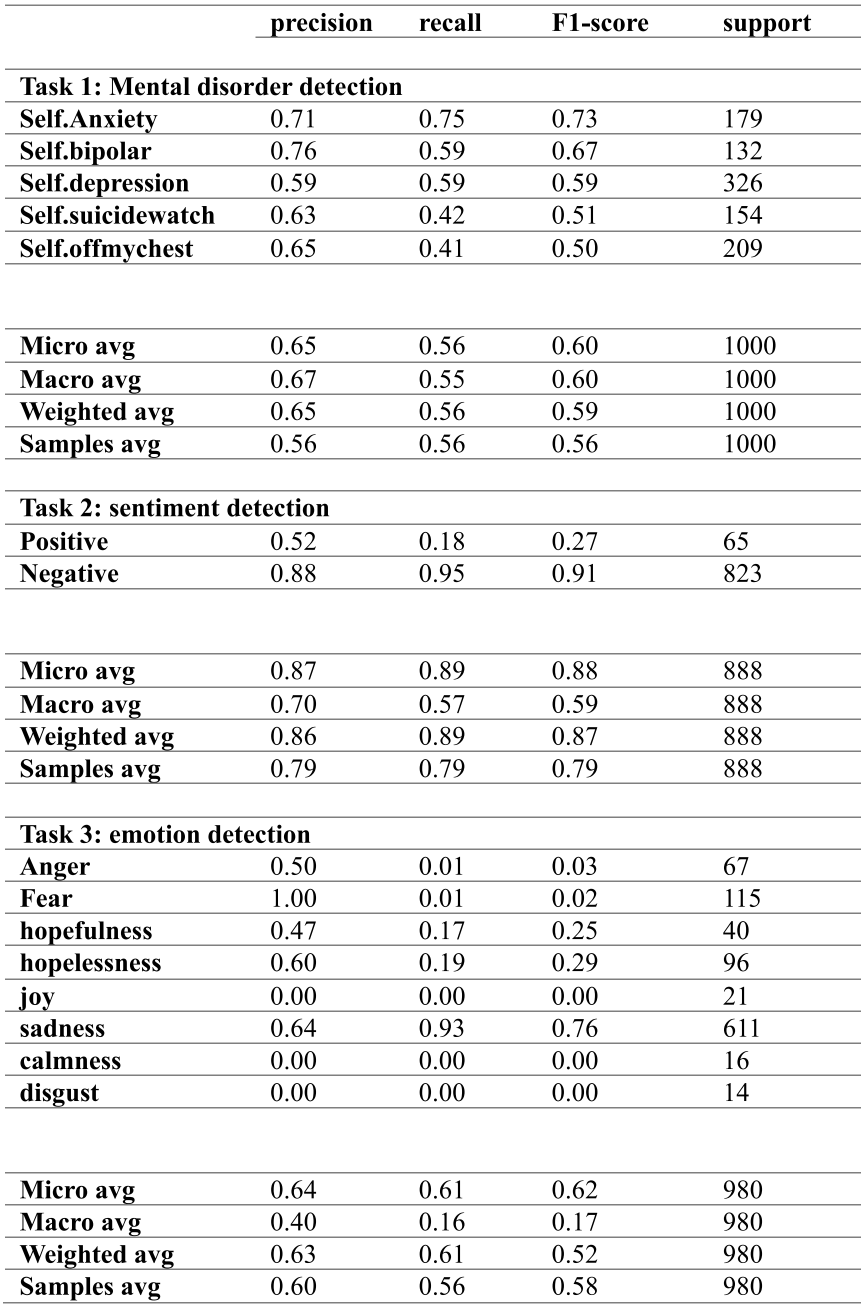
Figure 4.2. Classification report of mental disorder detection using custom model that has an attention layer.

	precision	recall	F1-score	support
Positive	0.58	0.11	0.18	65
Negative	0.88	0.94	0.91	823
Micro avg	0.88	0.88	0.88	888
Macro avg	0.73	0.52	0.55	888
Weighted avg	0.86	0.88	0.86	888
Samples avg	0.78	0.78	0.78	888

Figure 4.3. Classification report of sentiment detection using a custom model that has an attention layer.

	precision	recall	F1-score	support
Anger	0.55	0.27	0.36	67
Fear	0.78	0.43	0.56	115
Hopefulness	0.25	0.03	0.05	40
Hopelessness	0.61	0.39	0.47	96
joy	0.00	0.00	0.00	21
sadness	0.73	0.86	0.79	611
calmness	0.00	0.00	0.00	16
disgust	0.00	0.00	0.00	14
Micro avg	0.71	0.64	0.68	980
Macro avg	0.36	0.25	0.28	980
Weighted avg	0.65	0.64	0.63	980
Samples avg	0.60	0.60	0.59	980

Figure 4.4. Classification report of emotion detection task using a custom model that has an attention layer.



	precision	recall	F1-score	support
Task 1: Mental disorder detection				
Self.Anxiety	0.71	0.75	0.73	179
Self.bipolar	0.76	0.59	0.67	132
Self.depression	0.59	0.59	0.59	326
Self.suicidewatch	0.63	0.42	0.51	154
Self.offmychest	0.65	0.41	0.50	209
Micro avg	0.65	0.56	0.60	1000
Macro avg	0.67	0.55	0.60	1000
Weighted avg	0.65	0.56	0.59	1000
Samples avg	0.56	0.56	0.56	1000
Task 2: sentiment detection				
Positive	0.52	0.18	0.27	65
Negative	0.88	0.95	0.91	823
Micro avg	0.87	0.89	0.88	888
Macro avg	0.70	0.57	0.59	888
Weighted avg	0.86	0.89	0.87	888
Samples avg	0.79	0.79	0.79	888
Task 3: emotion detection				
Anger	0.50	0.01	0.03	67
Fear	1.00	0.01	0.02	115
hopefulness	0.47	0.17	0.25	40
hopelessness	0.60	0.19	0.29	96
joy	0.00	0.00	0.00	21
sadness	0.64	0.93	0.76	611
calmness	0.00	0.00	0.00	16
disgust	0.00	0.00	0.00	14
Micro avg	0.64	0.61	0.62	980
Macro avg	0.40	0.16	0.17	980
Weighted avg	0.63	0.61	0.52	980
Samples avg	0.60	0.56	0.58	980

Figure 4.5. Classification report of MTL using a custom model with attention layer.

We noticed that there is a 1% to 6% increase in the f1-score for individual mental disorder detection when using the MTL custom model with an attention layer. For example, Bipolar detection increased from 65% to 67% f1-scores while depression detection increased from 57% to 59% f1-scores respectively. Suicide ideation detection increased significantly from 45% to 51% while there was a slight decrease in Anxiety detection from 76% to 73% f1-score respectively. There was an increase in the f1 score for sentiment detection too as the positive sentiment increased from 18% to 27% f1-score while the negative sentiment score remained the same. It was realized that emotion detection scores suffered decreases as one increase was not noticed. It is noticed also that generally, task instances with very small support test data suffered significantly when it came to detection. Task instances with only 10,20 or 30 instances were difficult to detect whiles tasks with more data instances obtained detection scores.

4.5 Single Task without Attention Layer vs MTL with no Attention Layer

The classification report for the single tasks of mental disorder detection, sentiment detection, and emotion detection using the custom model with no attention layer are shown in the figures below respectively. We also show the classification report for the MTL custom model with no attention layer.

	precision	**recall**	**F1-score**	**support**
Self.Anxiety	0.82	0.72	0.77	179
Self.bipolar	0.83	0.54	0.65	132
Self.depression	0.55	0.64	0.59	326
Self.suicidewatch	0.65	0.38	0.48	154
Self.offmychest	0.69	0.35	0.47	209
Micro avg	0.66	0.54	0.60	1000
Macro avg	0.71	0.53	0.59	1000
Weighted avg	0.68	0.54	0.59	1000
Samples avg	0.54	0.54	0.54	1000

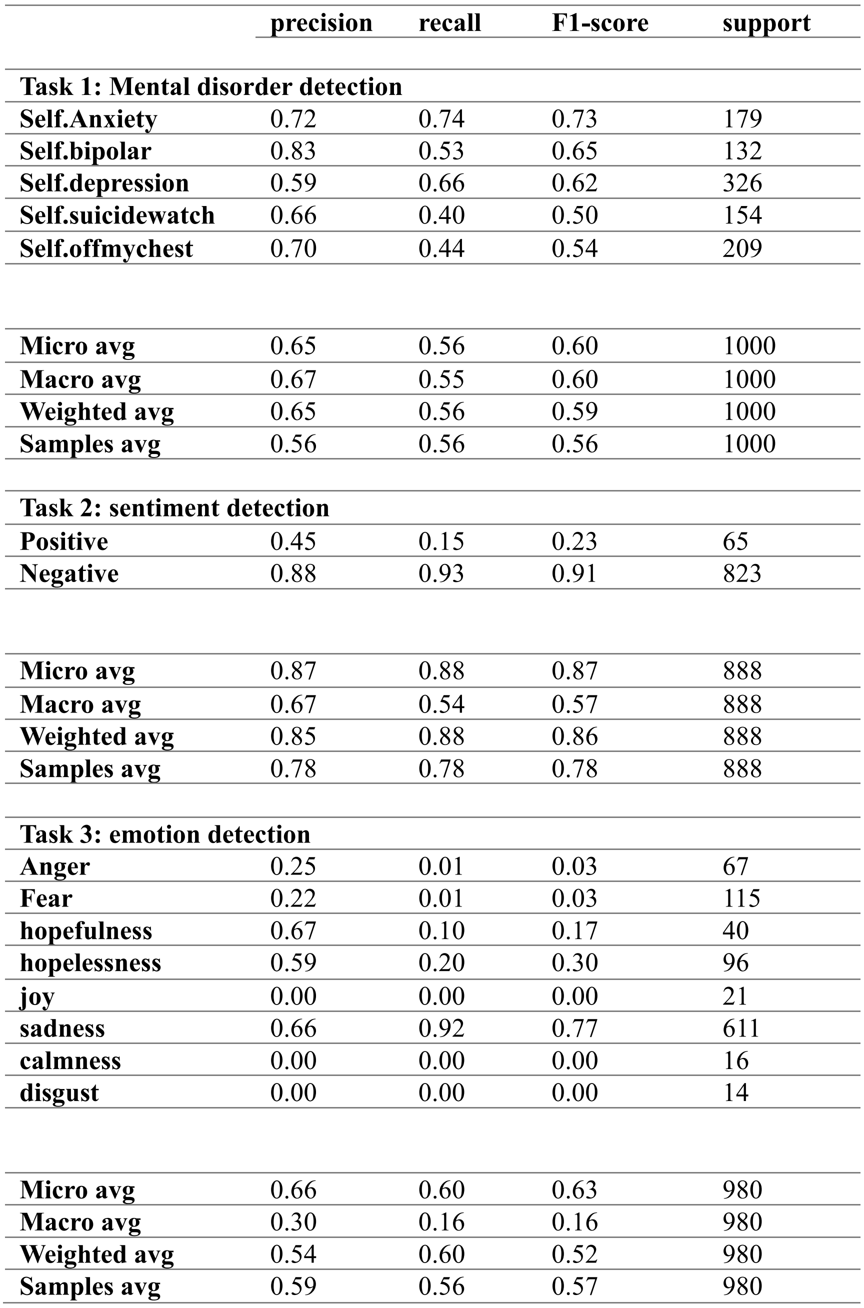
Figure 4.6. Classification report of mental disorder detection using the custom model without an attention layer.

	precision	recall	F1-score	support
Positive	0.75	0.05	0.09	65
Negative	0.88	0.93	0.91	823
Micro avg	0.88	0.87	0.87	888
Macro avg	0.82	0.49	0.50	888
Weighted avg	0.87	0.87	0.85	888
Samples avg	0.77	0.77	0.77	888

Figure 4.7. Classification report of sentiment detection using the custom model without an attention layer.

	precision	recall	F1-score	support
anger	0.44	0.16	0.24	67
fear	0.73	0.33	0.46	115
hopefulness	1.00	0.03	0.05	40
hopelessness	0.60	0.30	0.42	96
joy	0.00	0.00	0.00	21
sadness	0.73	0.84	0.78	611
calmness	0.00	0.00	0.00	16
disgust	0.00	0.00	0.00	14
Micro avg	0.71	0.61	0.66	980
Macro avg	0.44	0.21	0.24	980
Weighted avg	0.67	0.61	0.60	980
Samples avg	0.57	0.56	0.56	980

Figure 4.8. Classification report of emotion detection using the custom model without an attention layer.



	precision	recall	F1-score	support
Task 1: Mental disorder detection				
Self.Anxiety	0.72	0.74	0.73	179
Self.bipolar	0.83	0.53	0.65	132
Self.depression	0.59	0.66	0.62	326
Self.suicidewatch	0.66	0.40	0.50	154
Self.offmychest	0.70	0.44	0.54	209
Micro avg	0.65	0.56	0.60	1000
Macro avg	0.67	0.55	0.60	1000
Weighted avg	0.65	0.56	0.59	1000
Samples avg	0.56	0.56	0.56	1000
Task 2: sentiment detection				
Positive	0.45	0.15	0.23	65
Negative	0.88	0.93	0.91	823
Micro avg	0.87	0.88	0.87	888
Macro avg	0.67	0.54	0.57	888
Weighted avg	0.85	0.88	0.86	888
Samples avg	0.78	0.78	0.78	888
Task 3: emotion detection				
Anger	0.25	0.01	0.03	67
Fear	0.22	0.01	0.03	115
hopefulness	0.67	0.10	0.17	40
hopelessness	0.59	0.20	0.30	96
joy	0.00	0.00	0.00	21
sadness	0.66	0.92	0.77	611
calmness	0.00	0.00	0.00	16
disgust	0.00	0.00	0.00	14
Micro avg	0.66	0.60	0.63	980
Macro avg	0.30	0.16	0.16	980
Weighted avg	0.54	0.60	0.52	980
Samples avg	0.59	0.56	0.57	980

Figure 4.9. MTL classification report using the custom model without an attention layer.

From the classification reports, we realize that there is a 1% to 3% increase in the f1-score for individual mental disorder detection when using the MTL custom model with no attention layer. For example, depression detection increased from 59% to 62% f1-scores while bipolar detection remained the same at 65% f1-scores respectively. There was an increase in suicide ideation detection from 48% to 50% f1-score while there was a slight decrease in anxiety detection from 77% to 73%. There was an increase in the f1 score for sentiment detection too as the positive sentiment increased from 9% to 23% f1-score while the negative sentiment score remained the same. It was realized that emotion detection scores suffered decreases as one increase was not noticed. We can see from the figures that the f1-score for suicide ideation, depression, bipolar and offmychest detection increased significantly while Anxiety detection reduced slightly when the MTL model with no attention layer was employed. There was also an increase in the positive sentiment of the sentiment detection tasks. Emotion detection decreased significantly with only sadness remaining the same.

CHAPTER 5

5.DISCUSSION ON RESULTS.

The architecture pipeline for the suggested models is very important to the success of the models implemented. The quality of a dataset is an important factor in how a model performs on that dataset. Performing pre-processing techniques such as contraction replacements, lemmatization, stop word removals, and padding is meant to further remove noise elements from the extended dataset and thus improve the performance and accuracy of the suggested models.

The use of word embeddings to further improve the quality of generalization by the suggested models was well planned. Word embeddings are known to help models choose quality feature sets through a process of transfer learning. These embeddings are vector assignments of the words in a text, that takes into consideration the context of the specific word and other neighboring words. These help with the semantics of how models understand sentences. I chose pre-trained glove embedding as it has been extensively used and is known to be one of the quality pre-trained word embedding datasets available.

The single task with attention layer had f1-scores of 59% for mental disorder detection, 88% f1-score for sentiment detection, and 68% f1-score for emotion detection respectively. These scores when compared to other single tasks and MTL tasks research available in the field show an improvement in the scores. The same can be said for the single-task model with no attention layer which had f1-scores of 60% for mental disorder detection, 88% for sentiment detection, and 66% for emotion detection respectively.

The success criteria of an MTL model are measured when at least the accuracy of one of the tasks performed is improved in comparison to the associated single-task model(s) (McCann et al., 2018).

We saw a 1% improvement in the mental disorder detection task for the suggested MTL model with attention layer when compared to its single-task model. The f1-score increased from 59% to 60%. I also noticed that the sentiment detection task scores for both the single task and MTL model task for that experiment remained the same and the emotion detection task f1-score reduced from 68% to 62%. Even though these evaluations were noticed the success criteria for classifying an MTL model were met since the suggested MTL model with attention layer had a significant increase.

We also noticed a 2% increase in the mental disorder detection task for the suggested MTL model with no attention layer when compared to its single-task model. The f1-score increased from 60% to 62%. The sentiment detection score also remained the same at 88% f1-score. There was also a decrease in the f1-score for emotion detection from 66% for the single task to 63% for the suggested MTL model with no attention layer. This model can also be said to pass the success criteria of a MTL model as at least one of the tasks involved showed a significant increase in its accuracy.

When compared to the performance of other suggested pipelines in the field, I believe the suggested models in this research performed well and can be used as the basis for further studies in the field. The curated dataset which is also novel can be the basis for other variations of research in the field not limited to multi-task learning.

5.1 Error Analysis

We conduct error analysis in this section taking into consideration how mental disorder detection and sentiment detection were affected in both the proposed single task model and the multitask model systems.

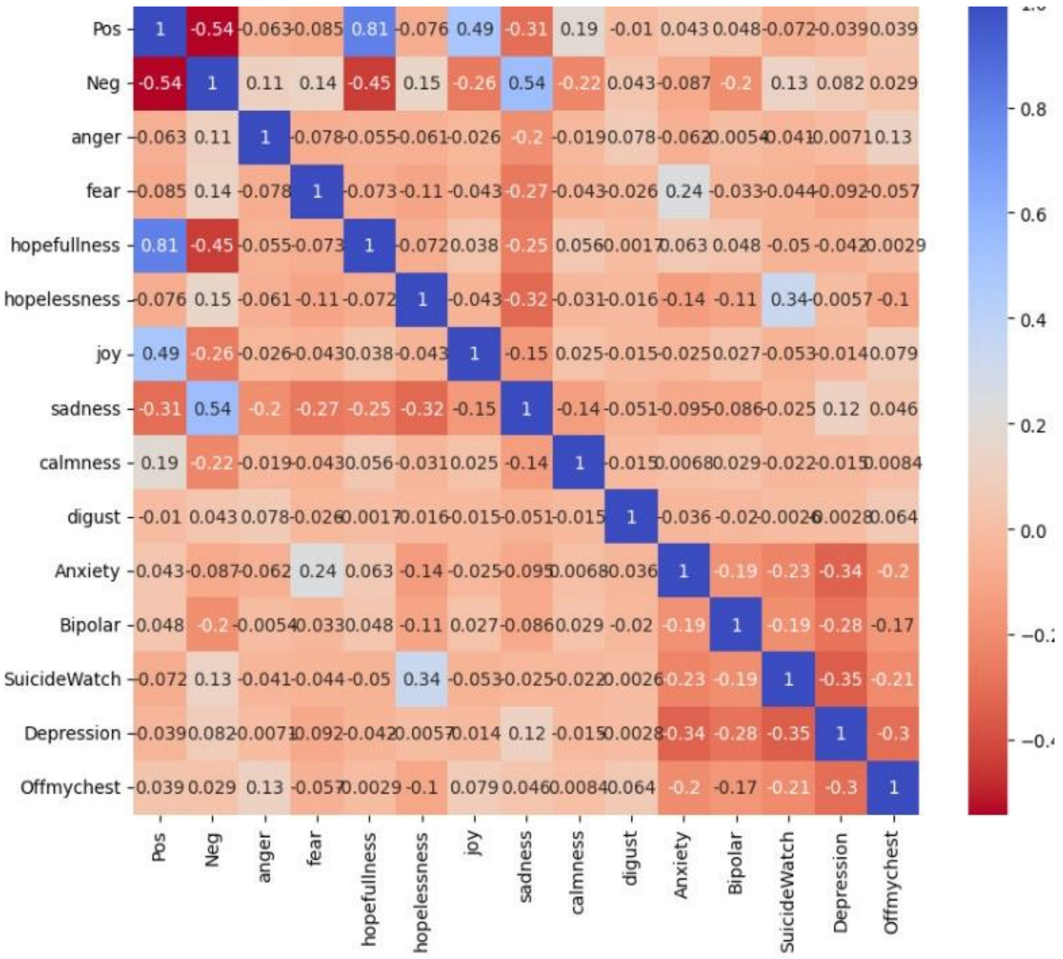


Figure 5.1. Correlation matrix showing the relation between mental disorder labels, sentiment and emotion labels.

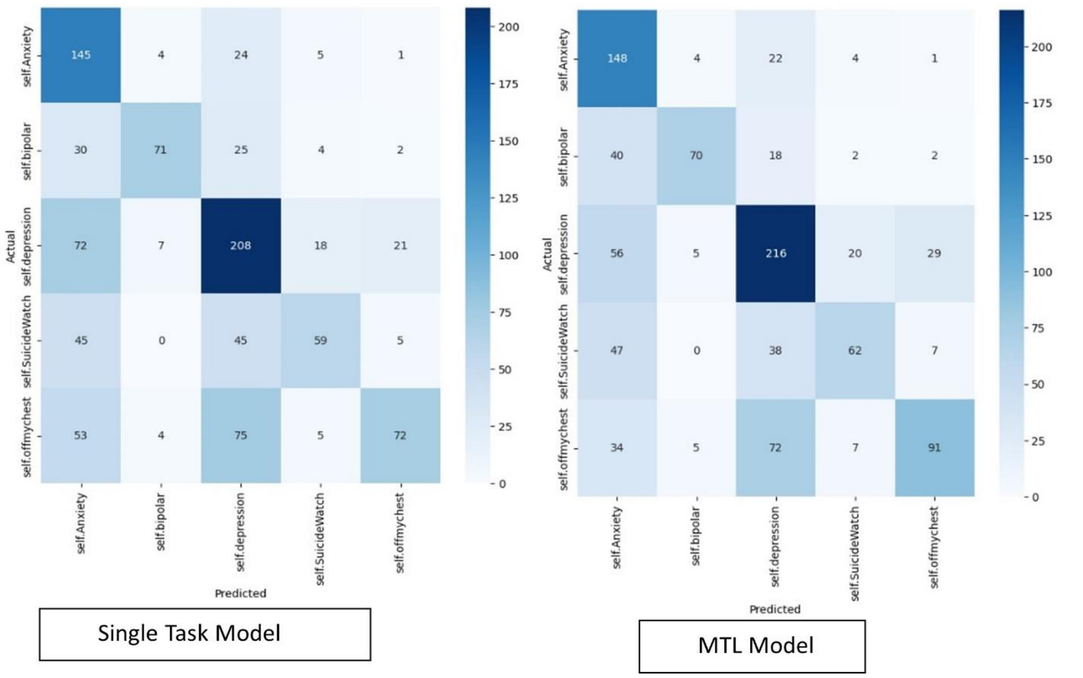


Figure 5.2. Confusion matrix showing mental disorder evaluations for both single task and MTL models respectively.

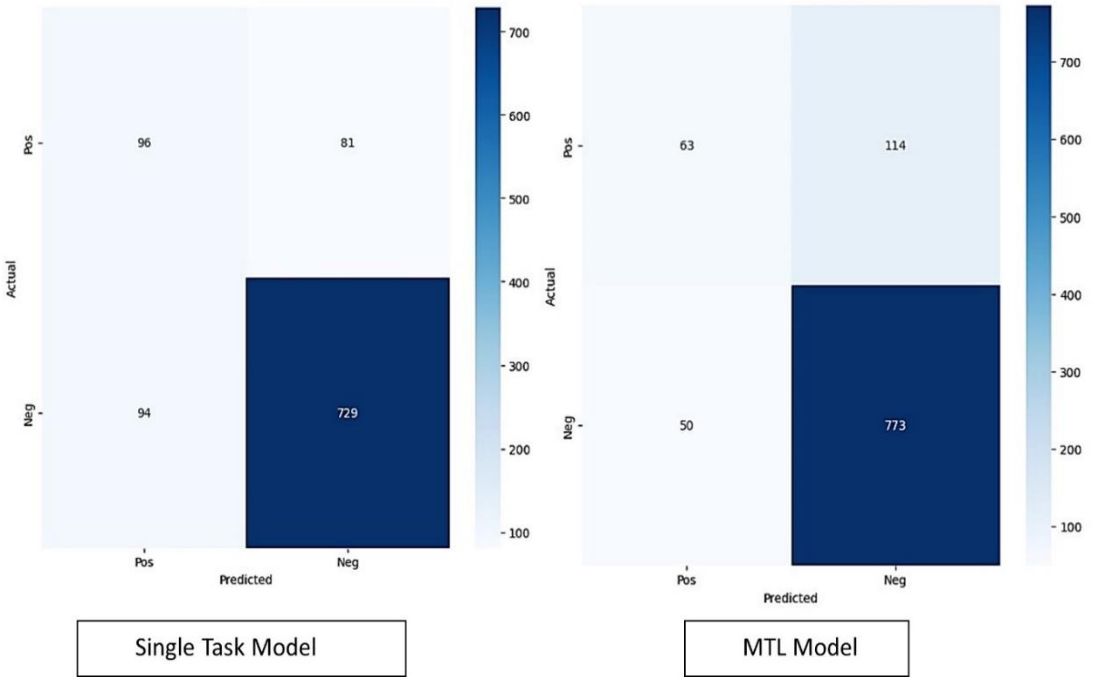


Figure 5.3. Confusion matrix showing sentiment detection evaluations for both single task and MTL models respectively.

The correlation matrix shows a strong correlation between positive sentiment and emotions such as hopefulness and joy. There is also a weak positive correlation between calmness and positive sentiment which means that calmness is more likely to be associated with positivity than negativity. There is also a strong correlation between sadness and negative sentiment. Other emotions such as anger, fear, hopelessness, and disgust are also correlated with negative sentiment but not as strong as the emotion, sadness.

From Figure 5.2, we could see significant improvements in the correct classifications of mental disorders from the proposed Single Task model to the proposed MTL model. Depression detection for instance, had 8 more correct classifications by the MTL model than the Single Task model. We could also notice from the confusion matrices above that, the other mental disorders such as bipolar, and suicide ideation were mostly mis-classified as either Anxiety or depression. Which means that, even though the MTL model performed better that its Single Task model, both models seem to identify elements of Anxiety or depression in instances that were actually labeled as bipolar or suicide ideation. Most instances are predicted as either belonging to the Anxiety class or to the depression class. We could also notice that some of the largest number of wrong predictions stemmed from depression being identified as Anxiety by the single task model. This was reduced significantly when it came to the MTL model. Figure 5.3 also shows that the MTL model performed better in classifying the negative sentiment than its single task model. However, the MTL model had relatively less accuracy when classifying the positive sentiment. We recognize this pattern from the top-view evaluation also in Table 4.3 where, even though there was an increase in f1-score for mental disorder detection when using the MTL model, there was a decrease in the accuracy for the emotion detection model. The table below shows some sample predictions by both single task and MTL task models.

Table 2.1. Sample predictions of custom single task and proposed MTL models.

Text	**True label**	**Single Task Prediction**	**MTL Prediction**
Weird habits? Does anyone else ever make food and then just throw it all away for no reason? I thought I had gotten past doing that but here I am, still hungry from throwing out my dinner because I suddenly felt upset at it.	Bipolar Negative sentiment Sadness	Depression Negative Sentiment Sadness	Bipolar Negative Sentiment Sadness
Im scared to get a job because of my depression and anxiety. I cant even function properly at home. How will I function properly in a working environment?	Depression Negative sentiment fear	Anxiety Negative sentiment Sadness	Anxiety Negative sentiment Sadness
im so tired I have no friends. i havent made a friend in five years. i tried making friends online but no one really talks back to me. i been looking for a job and going to interviews but no one will hire me. im failing all my college classes. i dont feel passionate about my interest anymore. i cant drive. i cant even take care of myself and im too dependent on my parents. im about to be twenty in january.	Depression Negative sentiment Hopelessness Sadness	Anxiety Negative sentiment Sadness	Suicidewatch Negative sentiment Hopelessness Sadness

The proposed MTL model can classify sentences implicitly in some instances. The single task model finds it difficult to make classifications in any implicit instance. For the first sentence, bipolar is insinuated implicitly. In other implicit instances, where two or more explicit words that describe the mental disorder are used, we realize that both single task and MTL model choose one of the explicit words. For example, in the second sentence, explicit words such as scared, depression and anxiety are used. The true classification is depression. But both the single task and MTL task models classified that instance as Anxiety. This means that both models associate the sentence more with anxiety than with depression. This is a major challenge in the classification process as some elements of Anxiety are generally found in suicide ideation and depression. This becomes a confusion for the model when two or mor explicit words that describe several mental disorders are used in one sentence or instance. It was also noticed that when sentences become very long, both proposed models misclassify the mental disorder as shown in the last instance.

Lastly, we also noticed that both models classified the sentiment correctly most of the time when it is negative and when the emotion is sadness. For example, in the last instance in Table 5.1, even though the single task and MTL model misclassified the mental disorder due to ambiguity, they were both able to classify the sentiment correctly. The single task model was only able to detect one of the emotions which is sadness whiles the MTL model detected both emotions as sadness and hopelessness. Partial detections occur and this can be associated with the MTL model recognizing the strong correlation between a negative sentiment and sadness as proven by the correlation matrix. The low classification accuracy of the positive sentiment can be associated with the imbalance nature of sentiment classifications which had just 65 instances for the test. The same can be said for the emotion labels where some emotion instances had as little as 12 support instances like the disgust emotion. Adding more labels for future work can increase the accuracy of models.

CHAPTER 6

6.CONCLUSION AND FUTURE WORKS.

In this project, I proposed and evaluated custom multi-task frameworks that performed better than their counterparts especially when it comes to mental disorder detection and sentiment detection. The performance of the custom models on the emotion detection instances decreased significantly. We also extended 5,000 instances of the benchmark dataset (SWMH) with sentiment labels which are positive and negative sentiment and emotion labels which are fear, hopefulness, hopelessness, anger, sadness, joy, calmness, and disgust.

Each instance in the extended SWMH dataset is marked with multilabel sentiment and emotion classes as opposed to the single labeling scheme followed in the introductory work. This provides an opportunity to optimize the use of MTL models as more features can be mined from the sentiment and emotion labels from the extended models we curated.

The dataset contained skewed data that could affect generalization negatively. The mental disorder labels are not balanced. The sentiment and emotion labels are imbalanced too. I introduced padding to the dataset to help mitigate the problem of memorization during training of the MTL and single task models. Even though 5,000 instances of a dataset are significantly low for deep complex models, the proposed custom models performed well in the detection tasks that were assigned. Even though the main aim of the project is to investigate MTL systems on the dataset, I also received important insights on the performance of single-task systems on the dataset. Further performance analysis could be carried out on the results from the single-tasks systems to receive more insights.

In the future, the updated dataset provides a benchmark avenue for future extensive studies in the field of psycho-medicine using MTL approaches. More instances of the SWMH dataset will be labeled to provide better accuracy for the custom models for future real-world implementations such as health recommender systems and text diagnostic systems and digital chatbot systems in the mental health domains.

Future work experiments could also be conducted on the extended dataset we curated using other variations of MTL systems such as multi-step approaches. This will help investigate the effects of multi-step architecture systems on multi-label datasets such as the extended SWMH.

Research into the use of artificial intelligence in the mental health space in Turkey is very limited. There is an opportunity to extend this research to other languages such as the Turkish language to create the basis for mental disorder detection applications in the Turkish language. This research can be the basis for curating Turkish-oriented datasets to help kickstart interest in using mental health diagnostic systems with a Turkish language base. This can aid mental health professionals in Turkey in better identifying and mitigating the effects of such mental disorders.

REFERENCES

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016, March). End-to-end attention-based large vocabulary speech recognition. *In 2016 IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 4945-4949). IEEE.

Balbuena, J., Samamé, H., Almeyda, S., Mendoza, J., & Pow-Sang, J. A. (2021). Depression detection using audio-visual data and artificial intelligence: A systematic mapping study. *In Proceedings of Fifth International Congress on Information and Communication Technology*: ICICT 2020, London, Volume 2 (pp. 296-306). Springer Singapore.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28, 41-75.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. *In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 31-39).ACL Anthology.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *In Proceedings of the international AAAI conference on web and social media* 7(1), 128-137.

Dehkarghani, R., Armah, C., Ozcerli, O., & Tartra, A. (2023). Extended swmh [Data set]. https://github.com/flightmaestro/extendedswmh

De Melo, W. C., Granger, E., & Hadid, A. (2019, September). Depression detection based on deep distribution learning. *In 2019 IEEE international conference on image processing (ICIP)* (pp. 4544-4548). IEEE.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dinan, E., Humeau, S., Chintagunta, B., & Weston, J. (2019). Build it break it fix it for dialogue safety: Robustness from adversarial human attack. arXiv preprint arXiv:1908.06083.

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015, July). Multi-task learning for multiple language translation. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 1723-1732).

Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, 1-20.

Hashimoto, K., & Tsuruoka, Y. (2017). Neural machine translation with source-side latent graph parsing. arXiv preprint arXiv:1702.02265.

Homan, C., Johar, R., Liu, T., Lytle, M., Silenzio, V., & Alm, C. O. (2014, June). Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. *In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 107-117). ACL Anthology.

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, 349(6245), 253-255.

Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34(13), 10309-10319.

Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Kim, S., Hori, T., & Watanabe, S. (2017, March). Joint CTC-attention based end-to-end speech recognition using multi-task learning. *In 2017 IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 4835-4839). IEEE.

Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S., & Zettlemoyer, L. (2020). Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33, 18470-18481.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742.

Liu, X., Gao, J., He, X., Deng, L., Duh, K., & Wang, Y. Y. (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval. *Deep Learning Group*. Microsoft Research.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. (2015). Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025.

Mao, Y., Liu, W., & Lin, X. (2020, November). Adaptive adversarial multi-task representation learning. *In International conference on machine learning* (pp. 6724-6733). PMLR.

McCann, B., Keskar, N. S., Xiong, C., & Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730.

Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., & Gelbukh, A. (2020). A multiclass depression detection in social media based on sentiment analysis. *In 17th International Conference on Information Technology–New Generations (ITNG 2020)* (pp. 659-662). Springer International.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pirina, I., & Çöltekin, Ç. (2018, October). Identifying depression on reddit: The effect of training data. *In Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task* (pp. 9-12).

Ren, F., Kang, X., & Quan, C. (2015). Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE journal of biomedical and health informatics*, 20(5), 1384-1396.

Shah, F. M., Ahmed, F., Joy, S. K. S., Ahmed, S., Sadek, S., Shil, R., & Kabir, M. H. (2020, June). Early depression detection from social network using deep learning techniques. *In 2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 823-826). IEEE.

Søgaard, A., & Goldberg, Y. (2016, August). Deep multi-task learning with low level tasks supervised at lower layers. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 231-235).

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7, 44883-44893.

Xiao, L., Zhang, H., Chen, W., Wang, Y., & Jin, Y. (2018). Mcapsnet: Capsule network for text with multi-task learning. *In Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4565-4574).

Xu, Y., Liu, X., Shen, Y., Liu, J., & Gao, J. (2018). Multi-task learning with sample re-weighting for machine reading comprehension. arXiv preprint arXiv:1809.06963.

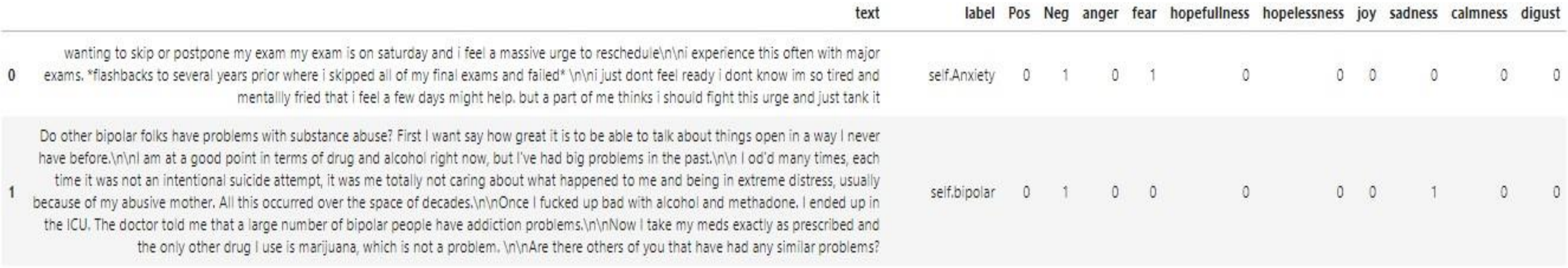
Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, 5(1), 30-43.

Zhang, Z., Lin, W., Liu, M., & Mahmoud, M. (2020, November). Multimodal deep learning framework for mental disorder recognition. *In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 344-350). IEEE.

Zhang, Z., Yu, W., Yu, M., Guo, Z., & Jiang, M. (2022). A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. arXiv preprint arXiv:2204.03508.

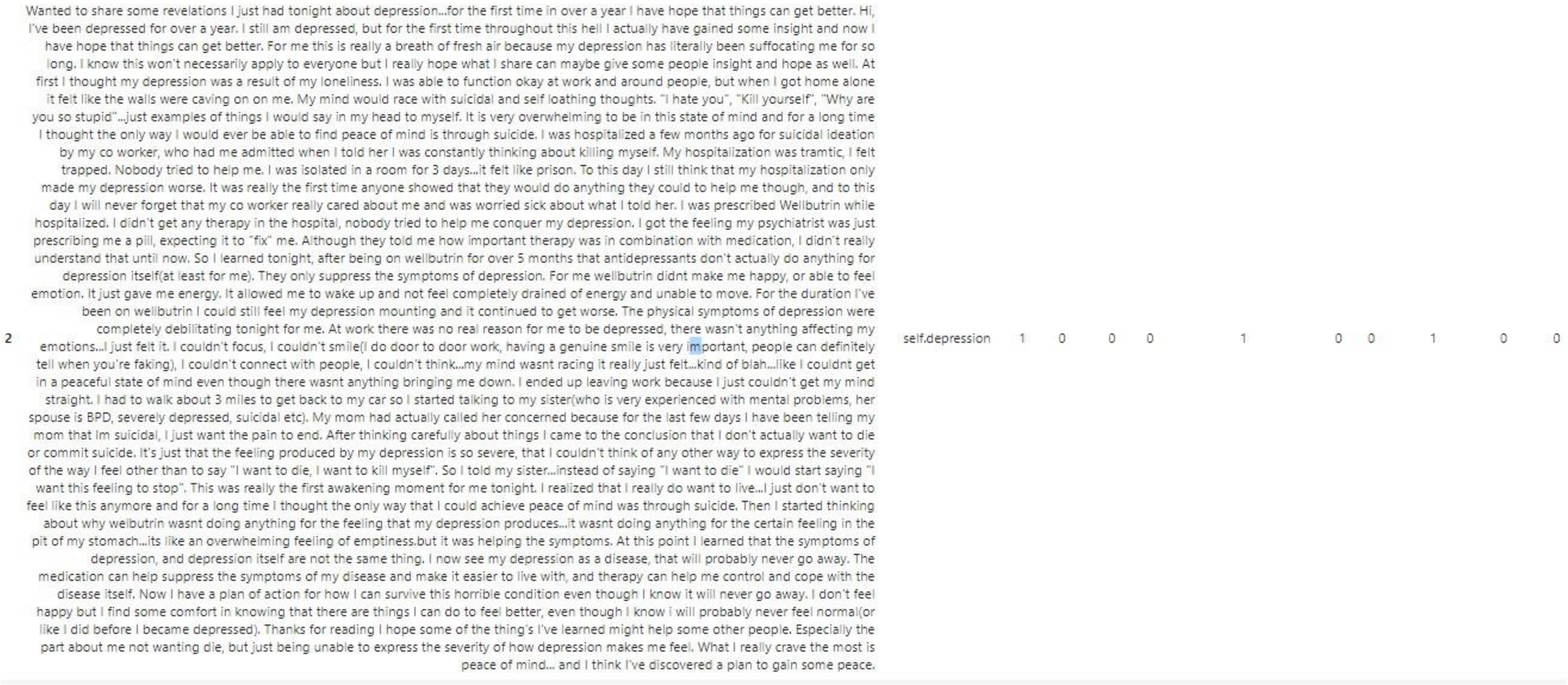
APPENDICES

A.1 Detailed subset of extended SWMH dataset after majority voting



	text	label	Pos	Neg	anger	fear	hopefullness	hopelessness	joy	sadness	calmness	digust
0	wanting to skip or postpone my exam my exam is on saturday and i feel a massive urge to reschedule\n\ni experience this often with major exams. *flashbacks to several years prior where i skipped all of my final exams and failed* \n\ni just dont feel ready i dont know im so tired and mentallly fried that i feel a few days might help. but a part of me thinks i should fight this urge and just tank it	self.Anxiety	0	1	0	1	0	0	0	0	0	0
1	Do other bipolar folks have problems with substance abuse? First I want say how great it is to be able to talk about things open in a way I never have before.\n\nI am at a good point in terms of drug and alcohol right now, but I've had big problems in the past.\n\n I od'd many times, each time it was not an intentional suicide attempt, it was me totally not caring about what happened to me and being in extreme distress, usually because of my abusive mother. All this occurred over the space of decades.\n\nOnce I fucked up bad with alcohol and methadone. I ended up in the ICU. The doctor told me that a large number of bipolar people have addiction problems.\n\nNow I take my meds exactly as prescribed and the only other drug I use is marijuana, which is not a problem. \n\nAre there others of you that have had any similar problems?	self.bipolar	0	1	0	0	0	0	0	1	0	0

A.2 Detailed subset of extended SWMH dataset after majority voting(continuity)



2	Wanted to share some revelations I just had tonight about depression...for the first time in over a year I have hope that things can get better. Hi, I've been depressed for over a year. I still am depressed, but for the first time throughout this hell I actually have gained some insight and now I have hope that things can get better. For me this is really a breath of fresh air because my depression has literally been suffocating me for so long. I know this won't necessarily apply to everyone but I really hope what I share can maybe give some people insight and hope as well. At first I thought my depression was a result of my loneliness. I was able to function okay at work and around people, but when I got home alone it felt like the walls were caving on on me. My mind would race with suicidal and self loathing thoughts. "I hate you", "Kill yourself", "Why are you so stupid"...just examples of things I would say in my head to myself. It is very overwhelming to be in this state of mind and for a long time I thought the only way I would ever be able to find peace of mind is through suicide. I was hospitalized a few months ago for suicidal ideation by my co worker, who had me admitted when I told her I was constantly thinking about killing myself. My hospitalization was tramtic, I felt trapped. Nobody tried to help me. I was isolated in a room for 3 days...it felt like prison. To this day I still think that my hospitalization only made my depression worse. It was really the first time anyone showed that they would do anything they could to help me though, and to this day I will never forget that my co worker really cared about me and was worried sick about what I told her. I was prescribed Wellbutrin while hospitalized. I didn't get any therapy in the hospital, nobody tried to help me conquer my depression. I got the feeling my psychiatrist was just prescribing me a pill, expecting it to "fix" me. Although they told me how important therapy was in combination with medication, I didn't really understand that until now. So I learned tonight, after being on wellbutrin for over 5 months that antidepressants don't actually do anything for depression itself(at least for me). They only suppress the symptoms of depression. For me wellbutrin didnt make me happy, or able to feel emotion. It just gave me energy. It allowed me to wake up and not feel completely drained of energy and unable to move. For the duration I've been on wellbutrin I could still feel my depression mounting and it continued to get worse. The physical symptoms of depression were completely debilitating tonight for me. At work there was no real reason for me to be depressed, there wasn't anything affecting my emotions...I just felt it. I couldn't focus, I couldn't smile(I do door to door work, having a genuine smile is very important, people can definitely tell when you're faking), I couldn't connect with people, I couldn't think...my mind wasnt racing it really just felt...kind of blah...like I couldnt get in a peaceful state of mind even though there wasnt anything bringing me down. I ended up leaving work because I just couldn't get my mind straight. I had to walk about 3 miles to get back to my car so I started talking to my sister(who is very experienced with mental problems, her spouse is BPD, severely depressed, suicidal etc). My mom had actually called her concerned because for the last few days I have been telling my mom that Im suicidal, I just want the pain to end. After thinking carefully about things I came to the conclusion that I don't actually want to die or commit suicide. It's just that the feeling produced by my depression is so severe, that I couldn't think of any other way to express the severity of the way I feel other than to say "I want to die, I want to kill myself". So I told my sister...instead of saying "I want to die" I would start saying "I want this feeling to stop". This was really the first awakening moment for me tonight. I realized that I really do want to live...I just don't want to feel like this anymore and for a long time I thought the only way that I could achieve peace of mind was through suicide. Then I started thinking about why welbutrin wasnt doing anything for the feeling that my depression produces...it wasnt doing anything for the certain feeling in the pit of my stomach...its like an overwhelming feeling of emptiness.but it was helping the symptoms. At this point I learned that the symptoms of depression, and depression itself are not the same thing. I now see my depression as a disease, that will probably never go away. The medication can help suppress the symptoms of my disease and make it easier to live with, and therapy can help me control and cope with the disease itself. Now I have a plan of action for how I can survive this horrible condition even though I know it will never go away. I don't feel happy but I find some comfort in knowing that there are things I can do to feel better, even though I know i will probably never feel normal(or like I did before I became depressed). Thanks for reading I hope some of the thing's I've learned might help some other people. Especially the part about me not wanting die, but just being unable to express the severity of how depression makes me feel. What I really crave the most is peace of mind... and I think I've discovered a plan to gain some peace.	self.depression	1	0	0	0	1	0	0	1	0	0

CURRICULUM VITAE